



HAL
open science

From signal representation to representation learning: structured modeling of speech signals

Nicolas Obin

► **To cite this version:**

Nicolas Obin. From signal representation to representation learning: structured modeling of speech signals. Sound [cs.SD]. Sorbonne Université, 2023. tel-04223614

HAL Id: tel-04223614

<https://hal.science/tel-04223614v1>

Submitted on 30 Sep 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Soutenu par



**MÉMOIRE D'HABILITATION À DIRIGER DES RECHERCHES
SORBONNE UNIVERSITÉ - FACULTÉ DES SCIENCES ET INGÉNIEURIE**

Présenté par

Nicolas OBIN

Maître de Conférences à Sorbonne Université
STMS (IRCAM, CNRS, Sorbonne Université, Ministère de la Culture)
Paris, France

**DE LA REPRÉSENTATION DU SIGNAL À L'APPRENTISSAGE DE REPRÉSENTATIONS :
MODÉLISATION STRUCTURÉE DE SIGNAUX DE PAROLE**

Habilitation soutenue le 12/09/2023 devant le jury composé de :

M. Thomas HUEBER, Directeur de recherche CNRS, GIPSA-lab	Rapporteur
M. Emmanuel VINCENT, Directeur de recherche INRIA, MultiSpeech	Rapporteur
M. Bjorn SCHULLER, Professeur, Imperial College London	Rapporteur
M. Gérard BIAU, Professeur, Sorbonne Université	Examineur
M. Jean-François BONASTRE, Directeur de Recherche INRIA, Défense et Sécurité	Examineur
Mme Catherine PELACHAUD, Directrice de recherche CNRS, ISIR	Examinatrice
M. Axel ROEBEL, Directeur de recherche, IRCAM	Examineur
Mme Isabel TRANCOSO, Professeure, INESC-ID / IST, Universidade de Lisboa	Examinatrice
M. Nicolas BECKER, Designer sonore & artiste	Membre invité

REMERCIEMENTS

À mon père Norbert OBIN (1950-2019) et à mon frère Marc OBIN (1978-2021), tous deux disparus entre ma thèse de doctorat et ce manuscrit. Vous me manquez.

À ma mère Nicole URREA-OBIN (1952-), ♥

À ma famille,

À mes amis, à mes amours, à mes emmerdes...

À toutes les personnes qui m'ont inspiré,

À toutes les vagues que je n'ai pu surfer pendant que je rédigeais ce manuscrit,

À mes rapporteurs, et à tous ceux et celles qui prendront plaisir à me lire.

TABLE DES MATIÈRES

Humain, évolution et langage	vii
Langage et art	vii
Communication verbale et non verbale	ix
Simuler l'humain : une machine peut-elle rêver de moutons électriques?	x
Organisation de ce manuscrit	1
1 L'ART DE LA VOIX : SÉMILOGIE ET MODÈLES DE SIGNAUX POUR LA CARACTÉRISATION DE LA VOIX ACTÉE	4
1.1 Introduction	4
1.2 Caractérisation de la qualité vocale de voix actées : signal, modèle, information	7
Modèle de source glottique	7
Mesure Multi-Résolution du Rapport Harmonique sur Bruit (HNR)	8
Mesure d'Entropie Spectrale Multi-Résolution	9
Mesure d'entropie spectrale	11
Entropie de Rényi	11
Entropie Spectrale Multi-Résolution	12
Évaluation expérimentale	13
1.3 Synthèse et discussion 1	14
1.4 Sémiologie et apprentissage	15
Vers une taxonomie de la voix	17
Apprentissage de la similarité vocale : acoustique vs. perception	18
Modélisation de l'espace acoustique : modèle du Monde et supervecteur GMM	18
Analyse factorielle : espace de variabilité totale et i-vecteur	18
Mesures de la similarité acoustique	19
Mesure de la similarité perceptive	20
Évaluations expérimentales	21
Évaluation objective	22
Évaluation subjective	22
1.5 Synthèse et discussion 2	25
2 MODÉLISATION STATISTIQUE STRUCTURÉE DE SIGNAUX DE PAROLE POUR LA PERCEPTION SONORE	27
2.1 Introduction	27
2.2 Modèle NMF source/filtre avec contraintes inspirées de la physique	28
NMF et modèle source/filtre	29
Principe de la NMF	29
NMF avec modèle source/filtre	30
Modèle source/filtre avec contraintes inspirées de la physique	31
Contraintes usuelles	31
Contrainte de cohérence source/filtre	31
Contrainte adaptative	32
Expérience : séparation de sources sonores en environnement bruité	33
Méthodologie	33
Résultats et discussion	34
2.3 Factorisation en tenseurs non-négatifs pour la localisation binaurale de sources sonores avec a priori de HRTF	35
Principe de l'audition binaurale	36
Factorisation en tenseurs non-négatifs de signaux binauraux	38
Localisation de sources sonores par NTF	40
Localisation à partir de la matrice de mélange binaural estimée	40
Localisation à partir de l'image de la source	41

Expérience : localisation de sources sonores	41
Méthodologie	41
Expérience 1 : localisation d'une source de parole	43
Expérience 2 : vers la localisation de locuteurs multiples	44
2.4 Synthèse et discussion	45
3 MODÉLISATION GÉNÉRATIVE DE SIGNAUX DE PAROLE PAR APPRENTISSAGE NEURONAL DE REPRÉSENTATIONS STRUCTURÉES	48
3.1 Modélisation générative de la F0 pour la conversion vocale	53
Conversion neuronale de la F0 par modélisation séquence-à-séquence	54
Positionnement du problème et formulation neuronale	55
Détails d'implémentation	57
Expérience : conversion de neutre vers expressif	57
Méthodologie	57
Résultats et discussion	57
Conversion neuronale de la F0 à partir de représentation multi-échelle en ondelettes	59
Positionnement du problème et formulation neuronale	60
Détails d'implémentation	62
Expérience : conversion de l'expressivité	62
Méthodologie	64
Synthèse et discussion	65
3.2 Conversion neuronale de l'expressivité à partir de mel-spectrogrammes	67
Positionnement et formulation neuronale	67
Détails d'implémentation	70
Expérience : conversion de l'expressivité	71
Expérience objective : mesure de l'intelligibilité de la parole convertie	71
Expériences perceptives	72
Conclusion et transition	74
3.3 Apprentissage de représentations démêlées pour la manipulation des attributs de la voix	74
Conversion de l'identité vocale : histoire, problèmes, méthodes	74
Positionnement du problème et formulation neuronale	77
Pré-traitement et post-réseau	77
Encodeur du locuteur	78
Encodeur du contenu	78
Dissocier les informations sur le contenu et l'identité	79
Décodeur	79
VC neuronale avec préservation du temps et de la F0	80
Synchronisation temporelle	80
Préservation de la F0	81
Détails d'implémentation	83
Expérience : conversion de l'identité	83
Méthodologie	83
Résultats et discussion	84
Démêlage des attributs vocaux par apprentissage adversarial	85
Expérience : conversion de genre	86
Illustration préliminaire	86
Expérience objective : information mutuelle et visualisation des codes	87
Expérience subjective	88
Résultats et discussion	89
3.4 Synthèse et discussion	90
Considérations générales sur la génération de la parole	94
Les défis de la génération de la parole ... et au-delà	95
Problèmes d'apprentissage	95

Génération multimodale de comportements humains	96
Problèmes méthodologiques	96
Problèmes de sécurité : simuler des attaques pour apprendre à se défendre	97
Un chercheur engagé dans les défis de la société	97
Bibliographie	99

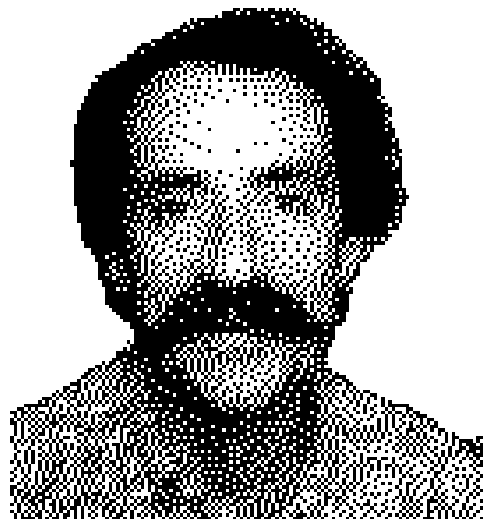
Svery aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it.

J. MCCARTHY, DARTMOUTH COLLEGE
M. L. MINSKY, HARVARD UNIVERSITY
N. ROCHESTER, I.B.M. CORPORATION
C. E. SHANNON, BELL TELEPHONE LABS

Swant to be a machine

ANDY WARHOL

XAVIER RODET,
FENÊTRE D'OUVERTURE DU LOGICIEL XSPECT



PRÉLIMINAIRES

Je m'intéresse à la communication chez l'humain, à ses comportements, et à leur simulation par ordinateur. Plus particulièrement, je m'intéresse aux productions et aux formes d'expression qui distinguent l'être humain — cet animal social et culturel — dans le monde vivant et animal : le langage et l'art. Chercheur à l'IRCAM, un institut dédié à la recherche sur le son et la musique, je me suis naturellement spécialisé sur la modalité sonore de la communication et de l'expression humaine, la parole et la musique. Je cherche à créer des "machines parlantes"¹ (Bailly, 1989; Black and Taylor, 1994; Fukada et al., 1994; Tokuda et al., 1995, 2002; Yoshimura et al., 1999; Black et al., 2007; Wang et al., 2017b) — et même chantantes — pour donner aux artistes la possibilité de créer de nouvelles voix humaines inouïes, sensibles, et sensationnelles. Pour ce faire, il est absolument nécessaire de craquer les codes de la compréhension du langage et de la communication humaine, et en particulier de sa forme orale : la parole. Cette introduction contextualise mon activité de recherche dans une perspective historique scientifique, artistique, et humaine.

¹ C'est en quelque sorte rendre au Golem la seule chose qui le séparait d'être humain, son souffle divin d'après la légende.

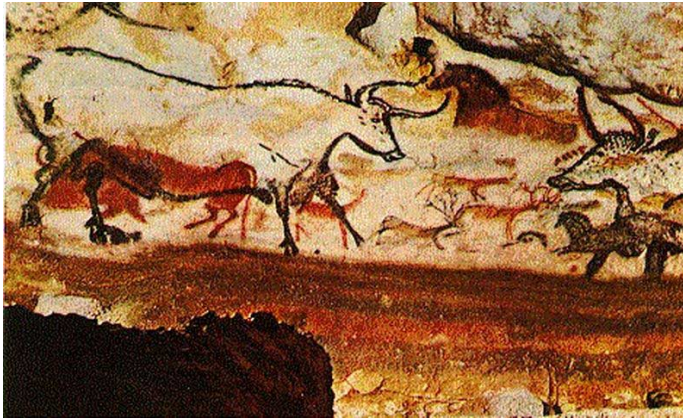
HUMAIN, ÉVOLUTION ET LANGAGE

Dans la longue évolution de la vie sur Terre depuis les premiers organismes unicellulaires jusqu'à aujourd'hui, le passage de l'*Homo Erectus* à l'*Homo Sapiens* a marqué un tournant déterminant dans l'histoire de l'humanité : l'émergence du langage. Le phénomène le plus important de ce passage est un développement important du cerveau, en taille (Weizenbaum, 1993) comme en nombre de connections synaptiques, ce qui a eu pour conséquence le développement de ses fonctions cognitives et sensori-motrices (Levinson and Holler, 2014). En outre, l'asymétrisation du cerveau a entraîné une spécialisation hémisphérique c'est-à-dire la localisation préférentielle de fonctions cognitives spécialisées dans l'une ou l'autre hémisphère, en particulier pour le langage et le raisonnement. L'*Homo Sapiens* est un être de raison : c'est l'avènement du symbolique et de ses représentations dans l'Humanité, du traitement symbolique de l'information. En conséquence, notre cognition filtre notre perception du monde physique à partir de nos sens mais aussi de nos échanges symboliques médiatisés par les mots du langage. Si l'*Homo Habilis* possédait une coordination motrice lui permettant de produire des sons, ils se limitaient vraisemblablement à des sons brefs exprimés comme réaction à son environnement, fondant les bases biologiques et physiologiques de la sensation et de l'expression des émotions et autres signaux d'alertes utiles à sa survie (Darwin, 1890). L'apparition de la parole coïnciderait avec l'apparition du langage chez l'*Homo Sapiens* (Lieberman, 1984). La thèse privilégiée suppose un abaissement progressif du larynx entraînant un agrandissement du pharynx du conduit vocal pour tendre vers la configuration actuelle de notre appareil vocal : l'être humain est devenu capable de produire des voyelles et de rendre fonctionnel son appareil vocal pour composer des sons pour communiquer. L'*Homo Sapiens* est également un animal social, le langage facilite la communication entre les individus (Weizenbaum, 1993) : communiquer, c'est partager sa représentation du monde avec autrui, c'est se confronter à l'altérité. Le développement du langage et de fonctions cognitives complexes coïncide alors avec l'apparition de la société et de la culture, de l'Homme cultivé. C'est l'origine des mythes et le fondement des sociétés dont l'Homme moderne est le descendant.

LANGAGE ET ART

Le langage, c'est l'abstraction du temps et de l'espace : par le langage, nous pouvons nous abstraire de l'immédiateté du présent pour figurer un autre temps et un autre lieu, et le partager avec autrui. L'art — du grec *habileté* — consiste à maîtriser les artifices et, à

ce titre, constitue sans doute la forme d'expression symbolique la plus aboutie chez l'être humain. Si l'art n'est pas un langage — au sens de la linguistique moderne (De Saussure, 1967) défini comme un système de signes fondés sur la dualité d'un signifiant et d'un signifié, il partage avec lui cette faculté d'abstraction et de manipulation des symboles, pour représenter le monde, le transfigurer, ou le sublimer². L'art — et la musique (Nattiez, 1987) — s'adresse à nos sens et à notre imagination sans nécessairement communiquer du sens. Cette manipulation des représentations symboliques est la source de la fiction et des mythes de l'Humanité (Witzel, 2013).



Fresque murale de la grotte de Lascaux.

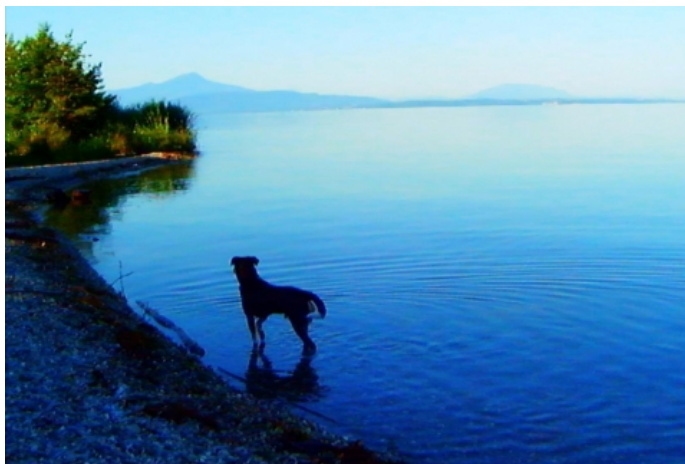
Par l'art, nous avons la possibilité de «sortir de nous», d'explorer «d'autres mondes» :

«Par l'art seulement nous pouvons sortir de nous, savoir ce que voit un autre de cet univers qui n'est pas le même que le nôtre et dont les paysages nous seraient restés aussi inconnus que ceux qu'il peut y avoir dans la lune. Grâce à l'art, au lieu de voir un seul monde, le nôtre, nous le voyons se multiplier, et autant qu'il y a d'artistes originaux, autant nous avons de mondes à notre disposition, plus différents les uns des autres que ceux qui roulent dans l'infini.» (Proust, 1927)

À ce titre, l'art partage avec la science et la philosophie l'appréhension et la compréhension d'un monde complexe et multiple. Le concept de l'"Umwelt" ou du "Monde propre" introduit par le biologiste et fondateur de l'éthologie Jakob von Uexküll postule qu'il existe autant de mondes que d'êtres sensibles, déterminés par un environnement sensoriel propre à chaque espèce ou chaque individu (von Uexküll, 1921). Plus encore, les neurosciences et la psychologie moderne ont montré que le monde est moins la résultante d'une perception, que d'une projection (Seth, 2021) — nous projetons littéralement le monde sur l'écran de notre conscience, en "mondovision". Par ailleurs, ce qui caractérise à la fois le langage et l'art comme représentations symboliques, c'est l'ambiguïté de ces symboles — la polysémie en est un exemple évidemment. Cette ambiguïté n'est pas une limite de la communication humaine (Browning and Le Cun, 2022), mais une nécessité et une force : c'est par ailleurs ce qui distingue l'humain de la machine. Dans chaque mot, il existe un interstice, une marge de liberté à son interprétation qui donne la liberté pour l'investir, s'approprier son sens, et construire des représentations singulières, une base neuro-psychologique du langage (Cyrulnik, 2023). Et c'est précisément cette incertitude et cette malléabilité du langage qui déterminent la possibilité de l'art et de la création artistique : dans l'histoire de l'humanité, l'art s'est graduellement substitué à Dieu pour sublimer le monde (Stiegler, 2013), c'est la figure de l'artiste moderne apparue au 19^{ème} siècle et que nous connaissons aujourd'hui.

Le langage et l'art constituent les productions les plus élaborées de l'être humain, et notamment comme organisation des sons par la parole et par la musique, la voix et le corps humain constituant les premiers instruments de musique connus. La musique, cette "sculpture dans le temps" (Sakamoto, 2018), représente à ce titre une forme d'art particulièrement élaborée. La musique occidentale procède d'une formalisation poussée

depuis les bases arithmétiques des intervalles musicaux chez Pythagore, la quantification du monde sonore des modes et des échelles tempérées — aujourd’hui codifiées dans le format d’échange numérique MIDI, les règles de leur organisation dans le traité de l’harmonie de Jean-Philippe Rameau (Rameau, 1722) jusqu’au traité d’instrumentation et d’orchestration d’Hector Berlioz (Berlioz, 1844), à l’atonalité, le sérialisme, ou la musique spectrale. L’évolution des techniques a permis d’automatiser la production et la diffusion des œuvres, de l’imprimerie de Guttenberg au cinéma, de l’industrialisation des moyens de production, la diffusion, la réception des œuvres au 20ème siècle jusqu’au numérique et à l’intelligence artificielle. Le film de Jean-Luc Godard "Adieu au langage" marque en quelque sorte le paroxysme d’une histoire du cinéma (Godard, 2006), et un retour sur les origines de l’art. "Ah Dieux", "Oh Langage" : de la caverne de Lascaux (et de Platon) à Godard, la boucle est bouclée.



Adieu au langage, Jean-Luc Godard (2014).

COMMUNICATION VERBALE ET NON VERBALE

Le langage n’est pas seulement écrit — ou verbal, pour le verbe —, il s’exprime également oralement et corporellement. Par notre voix, nous sommes capables d’émettre de grandes diversités de sons, c’est cette capacité que nous explorons dans notre première enfance avant d’acquérir le langage. La parole est la forme orale de notre langage, c’est-à-dire l’ensemble des sons que nous utilisons à des fins de communication. La communication orale ne se limite ainsi pas au sens des mots : "*miroir de l’âme*" (Aristote, 0 BC), la voix dévoile une part de notre intimité, de nos émotions, de nos intentions ; "*extension du corps dans l’espace*" (Despret, 2019), la voix est la projection de notre corps dans le monde extérieur, symboliquement et physiquement. La voix est la modalité qui articule la dualité du corps et de l’esprit de René Descartes (Descartes, 1647), l’intérieur et l’extérieur, l’acoustique et le symbolique. Si la linguistique moderne a posé les fondements des fonctions linguistiques du langage (De Saussure, 1967), ses fonctions stylistique (Léon, 1993), expressive, impressive, et poétique (Jakobson, 1960) demeurent encore largement inconnues et inexplorées. La parole est l’organisation des sons pour la communication verbale : d’une part, en articulant les sons et les symboles (Biber, 1988) — au figuré comme au propre avec les articulateurs du conduit vocal — ; et d’autre part, en les modulant par la prosodie vocale (Di Cristo, 1985; Nootboom, 1997; Campbell, 1992; Lacheret-Dujour and Beaugendre, 1998; Hirst et al., 2000), cette "*musique de la parole*" et ses dimensions sonores que sont la hauteur, l’intensité, le rythme, et la qualité vocale (Campbell and Mokhtari, 2003). Par ses fonctions linguistiques d’accentuation et de démarcation, la prosodie organise le sens et clarifie la structure d’un énoncé (Martin, 1987; Abney, 1992; Delais, 1994; Beckman et al., 2005), en particulier sa syntaxe (Selrik, 1981; Price et al., 1991; Beliã, 2016). La prosodie possède également des fonctions expressive (Wichmann, 2000; Beller et al., 2008; Shochi et al., 2009)

et stylistique (Hirschberg, 2000; Belião, 2013). En l'occurrence, la voix nous trahit de bien des manières c'est-à-dire est affectée par un grand nombre de facteurs humains, sociaux, et environnementaux : la voix véhicule des informations liées à notre physique et à notre physiologie (taille/poids, genre/âge, identité), notre psychologie (émotion ressentie ou exprimée (Ekman, 1992; Barrett, 2017), personnalité (Leary, 1957)), nos dispositions sociales (attitudes sociales) (Ajzen and Fishbein, 1980), nos origines socio-culturelles (accent, style de parole), nos habitudes et notre santé, etc. (Schuller and Batliner, 2013). La voix exprime nos émotions (Scherer et al., 1991; Bachorowski, 1999), en particulier par la prosodie vocale (Bänziger and Scherer, 2005), avec des universaux acoustiques (Ohala, 1996). Notre voix suggère notre disposition à l'égard d'autrui (Wichmann, 2000; Ponsot et al., 2018), avec comme dimension principale des premières impressions sociales l'amicalité (ami ou ennemi) et la dominance (supérieur ou inférieur) (Fiske et al., 2007). Nous adaptons notre style de parole en fonction de notre audience et de la situation de communication (Bell, 1984; Léon, 1993), renouvelant par la linguistique moderne le principe de l'éloquence introduit dans la rhétorique d'Aristote (Aristote, 3 BC).

Si l'être humain se distingue des autres primates par ses facultés de langage et de parole (Levinson and Holler, 2014), ils partagent avec eux au cours de l'évolution une communication non-verbale. En effet la communication ne se réduit pas au langage écrit : elle est par nature multimodale et implique le langage oral et le langage corporel. Les expressions faciales, le mouvement de la tête, des mains, et du corps composent un "système de systèmes" (McNeill et al., 2005; Argyle, 2013; Knapp et al., 2014) où chaque modalité ou canal de communication véhicule une information en propre. La théorie de la co-évolution (McNeill et al., 2005) postule que le langage et le geste font partie d'un seul système de pensée et de communication, où les deux modalités se renforcent ou se complètent dans la construction d'un sens et de sa communication à autrui. En particulier, l'être humain emploie continuellement sa "*prosodie visuelle*" (Graf et al., 2002), c'est-à-dire un ensemble de gestes corporels en conjonction avec sa communication verbale et sa prosodie vocale (thèse Mireille Fares). Dans sa théorie de l'évolution (Darwin, 1890), Charles Darwin considérait l'*expressivité* comportementale comme un facteur clef de survie d'une espèce. En outre, la complexité du répertoire de comportements de signalisation d'une espèce est liée à la complexité de son organisation sociale. Les bases biologiques et sociales communes aux primates humains et non-humains expliquent alors la similarité de comportement, et les formes particulièrement élaborées de la communication humaine et de son langage écrit, oral et corporel.

SIMULER L'HUMAIN : UNE MACHINE PEUT-ELLE RÊVER DE MOUTONS ÉLECTRIQUES ?

Depuis le mythe du Golem jusqu'à l'intelligence artificielle ChatGPT, l'humain n'a eu de cesse au cours de son histoire de chercher à créer des machines à son image, c'est-à-dire de simuler l'être humain et ses comportements. Cette histoire s'est incarnée par l'évolution des sciences et techniques à travers les âges de l'Humanité : depuis le Golem à l'âge de l'argile, les automates à l'âge du cuivre, les machines parlantes à l'âge de l'électricité, les robots humanoïdes à l'âge de l'informatique et du silicium, jusqu'aux agents virtuels et aux *deep fakes* à l'âge de l'intelligence artificielle. Si cette simulation s'est longtemps limitée à la reproduction de capacités motrices d'un être humain et de son fonctionnement bio-mécanique, l'intelligence artificielle telle que définie initialement par (Minsky et al., 1955) a rendu possible la simulation des capacités cognitives de raisonnement et de perception humaine, en premier lieu desquels le langage et la communication (Weizenbaum, 1966). Les capacités de simulation actuelles rendent non seulement possible la simulation d'apparences physiques mais également de comportements humains extrêmement réalistes. Les machines s'approprient les capacités qui jusqu'à alors distinguaient l'être humain, au premier rang desquelles le langage. Nous vivons aujourd'hui dans un monde cyber-physique, entourés et interagissant avec des machines capables de percevoir, d'interpréter,



Selfie.

de communiquer, et de simuler des comportements humains depuis les assistants vocaux comme "The Assistant" (Julia et al., 2001), les robots humanoïdes (Ishiguro et al., 2001), les agents virtuels (Hartholt et al., 2013), les *deep fakes* (Paris and Donovan, 2019), jusqu'à ChatGPT (OpenAI, 2022).

L'apparence extrêmement réaliste de ces comportements est troublante, à la limite de la vallée de l'étrange (Mori, 1970). Il n'en demeure pas moins que ces simulations, aussi réalistes soient-elles, ne le sont qu'en apparence : il s'agit de simulations de comportement humain en surface, qui ne relèvent en rien des processus biologiques et psychologiques qui les sous-tendent chez l'être humain (Bender et al., 2021). L'apparente humanité des machines intelligentes est révélatrice des biais de l'humain et de sa tendance à l'humano-centrisme, c'est-à-dire une tendance à humaniser tout ce qui l'entoure — objets inertes comme êtres vivants —, et en particulier de supposer des comportements similaires ou identiques aux siens à tout ce qui lui ressemble. Simuler par la machine, c'est quantifier les comportements, c'est-à-dire les simplifier en un nombre fini de particules, et les rendre programmables, c'est-à-dire prévisibles. Ceci n'est pas possible pour deux raisons principales : d'une part, parce que les bases organiques et psychologiques — y compris les pulsions et les désirs, conscients ou inconscients — qui sous-tendent en profondeur les comportements humains ne sont pas accessibles à la machine : un comportement est la résultante en surface d'un processus complexe de genèse et de choix arbitraires. D'autre part, l'humain est organiquement et spirituellement imparfait : par ses limites motrices et psycho-motrices, il ne peut réaliser identiquement un même geste. Ces imperfections font de l'être humain un animal stochastique, et c'est d'ailleurs la condition de la possibilité de son évolution, par hasard et sélection (Darwin, 1890). Mais dans la co-évolution des humains et des machines, il n'y a aucune raison d'attendre des machines qu'elles fonctionnent et se comportent comme un humain (Picq, 2019). Un humain n'est pas une machine, mais une machine n'est pas non plus un humain.

Les grandes avancées scientifiques réalisées en intelligence artificielle se sont opérées dans une démarche bio-inspirée et interdisciplinaire à l'interface de la biologie — le neurone formel formalisé dans (McCulloch and Pitts, 1943) —, de la cognition, de l'informatique et des "machines intelligentes" (Turing, 1950). L'intelligence artificielle (Minsky et al., 1955), la cybernétique (Wiener, 1961), et la robotique partagent la formalisation du fonctionnement

d'organismes vivants, en particulier l'humain, de l'ensemble de ses fonctions de perception, de décision, et d'action. Les réseaux de neurones artificiels sont exemplaires à cet effet : bio-inspiré, un réseau de neurones est une formalisation mathématique du fonctionnement réel du cerveau humain. En tant que modèle du cerveau, ils en constituent une approximation biologique mais une approximation fondée mathématiquement, à partir de laquelle il est possible d'apprendre par rétro-propagation de l'erreur de modélisation à partir de données. En parallèle, la communication a été théorisée par Claude Shannon (Shannon, 1948), fondant la théorie de l'information à l'origine des télécommunications modernes. Cette théorie propose un modèle fondateur de la communication numérique entre machines, et introduit l'unité de mesure de la quantité d'information de signaux fondée sur l'entropie. Par extension, cette théorie a été étendue à l'étude de systèmes organiques et à la communication humaine et animale, en opérant une réduction des réactions et des comportements à un ensemble discret et fini, et en complétant le schéma original de la communication avec la multimodalité et des boucles de rétroaction (Kopp et al., 2008). Nous pouvons ainsi mesurer la quantité d'information des systèmes de communication des espèces animales. La communication parlée constituerait *"le comportement le plus sophistiqué"* (Moore, 2007) connu à ce jour parmi les espèces animales — un canal de communication avec une capacité variant entre 50 b/s pour la forme écrite (Pierce and Karlin, 1957) et 20 kb/s pour la forme orale (Flanagan, 1972) notamment pour encoder acoustiquement les informations para- et extra-linguistiques. La mesure de la complexité du répertoire vocal d'une espèce animale est l'un des trois indicateurs de la complexité sociale de cette espèce (thèse Killian Martin encadrée avec l'éthologue Valérie Dufour, (Martin, 2023)), ouvrant une relation entre les télécommunications modernes et la théorie de l'évolution, une passerelle entre le monde numérique et le monde organique.

INTRODUCTION

ORGANISATION DE CE MANUSCRIT

Mon activité de recherche principale touche à la voix humaine, cette interface fascinante de l'humain entre son monde intérieur et son monde extérieur, entre le monde sensible et le monde symbolique. Ma démarche scientifique s'inscrit dans la lignée de la perspective historique présentée précédemment, et mes recherches se sont effectuées à l'IRCAM — un lieu de recherche sonore et musicale, de rencontres et d'expérimentations. Méthodologiquement, la simulation de comportements humains relève du paradigme de l'analyse par la synthèse : la synthèse permet de vérifier des hypothèses formulées par l'analyse, pour en retour les affiner et tendre progressivement vers une compréhension et une modélisation de plus en plus complète de cet objet d'étude complexe qu'est la voix humaine. En d'autres termes, il est nécessaire de bien formaliser pour bien simuler, en particulier l'ensemble des facteurs de variabilité de la communication humaine et de la parole. Nous verrons au cours de ce manuscrit que l'intelligence artificielle et l'apprentissage par réseaux de neurones ouvrent de nouvelles voies de modélisation et donc de compréhension de la voix humaine. Je soutiens que la connaissance humaine et l'apprentissage machine nécessitent une collaboration : d'une part, la spécification de connaissances humaines (linguistique, physiologique, psychologique) aide à l'apprentissage par la machine ; d'autre part, la connaissance humaine doit rester la finalité de la technologie et de l'intelligence artificielle.

Ce manuscrit aborde le sujet de la modélisation de signaux de parole par apprentissage à partir de données. Dans cette perspective, il pose et tente de répondre aux quatre questions fondamentales de l'apprentissage machine pour la simulation de la parole humaine, à savoir les représentations et modèles de signaux, les données, les modèles d'apprentissage, et leur évaluation.

1. **Q1. Quelles représentations ?** La spécification de modèles paramétriques de signaux spécifiques à la parole humaine présente l'avantage de formuler un modèle dont les paramètres sont intuitifs à interpréter et à manipuler (typiquement : fréquence fondamentale, énergie, bruit, enveloppe spectrale). Mais chaque paramètre seul est insuffisant pour expliquer de manière satisfaisante la variabilité des signaux de parole, et leur inter-relation est complexe. En outre, les vocodeurs généralement utilisés pour la synthèse ont des plages limitées de manipulation, et les signaux manipulés deviennent rapidement dégradés et non-réalistes
2. **Q2. Quelles données ?** Par rapport aux modalités textuelles ou visuelles, les bases de données de parole humaine sont extrêmement limitées. Pendant ma thèse, les bases de données se comptaient au mieux en dizaine d'heures, mais peuvent désormais atteindre 50,000 heures. Les facteurs de variabilité au premier rang desquels le langage, la langue, mais aussi le para-linguistique et la variété intra- et inter-individuelles rendent extrêmement difficile de collecter des données exhaustives et représentatives.
3. **Q3. Quels modèles d'apprentissage ?** Les modèles d'apprentissage doivent être conçus spécifiquement pour les signaux de parole, que ce soit d'un point de vue de la production, de la perception, ou plus largement de la cognition. La spécification de connaissances humaines sur la nature et la forme des signaux doit aider à apprendre de manière différenciée la variabilité associée à chacun des facteurs de variabilité.
4. **Q4. Comment évaluer ?** La simulation de comportements humains pose des problèmes évident d'évaluation. Les métriques objectives issues des protocoles de l'apprentissage machine ou du traitement du signal ne sont pas adaptées pour rendre

compte du naturel ou de l'expressivité d'une voix. En particulier, et comme pointé dans et (Obin et al., 2011c) et (Obin et al., 2012), la parole possède une grande variabilité phonétique et prosodique avec équivalence ou similarité de sens : définir une seule cible ne fait pas beaucoup de sens dans ce contexte. Il faut donc toujours de l'humain pour évaluer des productions humaines. Si des protocoles d'évaluation sont aujourd'hui couramment employés, comme les échelles MOS ou MUSHRA, ils se limitent encore à évaluer l'intelligibilité ou le naturel des voix et donc essentiellement à ses fonctions linguistiques. Il est nécessaire d'imaginer de nouveaux protocoles pour évaluer les fonctions stylistiques et expressives de la parole humaine. Je n'aborderai cependant que marginalement cette question dans le présent manuscrit, car elle n'a pas été centrale dans mes recherches. J'y reviendrai cependant dans la conclusion, à la lumière de mes travaux actuels.

Les réseaux de neurones rendent possible de lier ces trois questions fondamentales presque en une seule : contrairement au paradigme historique de l'apprentissage ou la représentation du signal et l'apprentissage pour une tâche étaient séparés, l'apprentissage par réseaux de neurones permet de s'affranchir de cette séparation en incluant l'apprentissage des représentations au sein de l'apprentissage lui-même. En d'autres mots, le réseau adapte sa perception en fonction de la tâche à accomplir.

Dans la suite de ce manuscrit, je vais présenter les travaux que j'ai réalisés à l'IRCAM dans l'équipe Analyse et Synthèse des sons (A/S) au sein du laboratoire des Sciences et Technologies de la Musique et du Son (STMS) selon trois axes présentés chronologiquement et thématiquement : une perspective de cognition (2011-2015) dans le Chapitre 1, une perspective de perception (2015-2018) pour l'élaboration de modèles de séparation et de localisation de sources sonores dans le Chapitre 2, et enfin dans une perspective de production (2018-2023) pour l'élaboration de modèles génératifs de la parole humaine dans le Chapitre 3. Je ne prétends pas dans ce manuscrit présenter exhaustivement mes activités de recherche : au contraire, j'ai choisi de les présenter sous un angle et une focale que j'estime le plus représentatif de ma démarche scientifique et de mes engagements scientifiques, artistiques, et sociétaux. La description détaillée de l'intégralité de mes activités de recherche, d'enseignement, de responsabilité, de diffusion, et de création est présentée en Annexe de ce manuscrit.

Durant les dix années qui séparent la soutenance de ma thèse (Obin, 2011) de ce manuscrit, j'ai publié 33 articles de conférences, 7 articles de revues, supervisé 12 stages de Master 2, encadré 7 thèses (3 soutenues, 4 en cours), piloté un projet ANR TheVoice (2017-2021), et deux projets Sorbonne Université ROUTE (Robot à l'écoute, 2014-2016) et ReVOLT (2022-2023), et été le coordinateur scientifique pour l'IRCAM de 3 projets ANR (EXOVOICES, 2022-2026, BRUEL, 2022-2026, et DeTOX, 2023-2025). En particulier, j'ai co-encadré les thèses de Clément Le Moine Veillon (2019-2023) sur la *Conversion neuronale des attitudes sociales dans les signaux de parole* sous la direction d'Axel Roebel (Le Moine, 2023), Mireille Farès (2019-2023) sur la *Génération de Comportements Humains Multimodaux avec Style* sous la direction de Catherine Pelachaud (Fares, 2023), et Killian Martin (2019-2022) sur la *Complexité vocale et contrôle cognitif chez le corbeau freux (Corvus frugilegus)* sous la direction de Valérie Dufour (Martin, 2023). J'encadre actuellement les thèses en cours de Léane Salais (2021-) sur la *Manipulation des attributs de la voix par apprentissage de représentations neuronales démêlées* sous la direction d'Axel Roebel (Salais, 2021), et de Théodor Lemerle (2023-) sur la *synthèse de la parole expressive pour la lecture d'histoires* sous la direction d'Axel Roebel (Lemerle, 2023). En musicologie, j'ai accompagné officiellement ou officieusement une nouvelle de musicologues de la voix parlée, scandée, ou chantée, comme la thèse d'Olivier Migliore intitulée *Analyser la prosodie musicale du punk, du rap et du ragga français (1977-1992) à l'aide de l'outil informatique* sous la direction d'Yvan Nommick (Migliore, 2016) et la thèse de Lisa La Pietra sur *L'indépendance de la voix du XXIème siècle. Ethique et Esthétique de l'interprétation vocale entre le Belcanto et l'intelligence artificielle* sous la direction de Giordano Ferrari (La Pietra, 2023). Les travaux d'Olivier Migliore ont été récompensés par le prix

Jean-Jacques Nattiez de la Société Française d'Analyse Musicale (SFAM) pour son article (Migliore, 2023).

Je suis actuellement le responsable du Master en Ingénierie des Systèmes Intelligents (ISI) à la Faculté des Sciences de Sorbonne Université, en formation initiale et en apprentissage. J'ai créé et je suis l'organisateur du premier colloque international sur les agents virtuels socialement intelligents (SIVA — *Socially Intelligent human-like Virtual Agents*, 2023)¹. Je suis membre du comité d'organisation de l'édition 2023 du colloque international *Speech Synthesis Workshop* (SSW — Grenoble, 2023) et du challenge Blizzard (Grenoble, 2023) sur la synthèse de la parole à partir du texte, et ai été membre du comité d'organisation de la 3ème édition du colloque international sur l'interaction vocale dans et entre les humains, les animaux, et les robots (VIHAR — Paris, France). Je suis le co-fondateur de SOPhIA l'association des étudiants de Sorbonne Université en Intelligence Artificielle en collaboration avec le Centre de Sorbonne Université pour l'Intelligence Artificielle (SCAI, avec Xavier Fresquet et Gérard Biau), et j'ai été chercheur invité sur l'axe "Working Living in a Cyber Physical AI world : which interactions between human being and machine?" au Forum Franco-Japonais sur l'Intelligence Augmentation and Amplification + Society (CNRS, EHESS, JST, 2022). Je suis également le fondateur et l'animateur d'évènements trans-disciplinaires et trans-communautaires comme les Deep Voice, Paris depuis 2020 un évènement de 2-3 jours visant à rassembler les acteurs de la science et de l'innovation dans les technologies vocales ou encore les *Fast-Forward* le lieu de rencontres expérimentales et informelles du design sonore pour le cinéma et des technologies du son à l'IRCAM. J'ai rejoint en 2022 le *Voice Lab* pour la promotion de la culture et la défense de la souveraineté numérique française, en particulier sur la création de ressources matérielles, logistiques, et algorithmiques pour la création de voix artificielles librement accessibles et exploitables, et en 2023 l'Association Française pour le Son à l'Image (AFSI). Enfin, je me suis impliqué dans des projets artistiques impliquant la voix humaine réelle et artificielle, comme la pièce *Luna Park* de Georges Aperghis (2011), le film *Marilyn* de Philippe Parreno (2012) et avec Nicolas Becker, le film *Annette* de Leos Carax (2021) avec Erwan Kerzanet, ou encore récemment l'oeuvre *Anima* (2022) d'Alexander Schubert, la recreation de la voix de Dalida pour l'émission *L'Hôtel du Temps* de Thierry Ardisson qui redonne vie à des personnalités disparues par l'entremise des Deep Fakes, la recreation de la voix d'un des pères fondateurs de la science-fiction Isaac Asimov pour le documentaire *Isaac Asimov, l'étrange testament du père des robots* de Mathias Théry (2022), ou la reconstitution artificielle du discours de *l'Appel du 18 Juin* du Général De Gaulle en collaboration avec Le Monde (2023).

La science partage avec l'art cette faculté de l'humain à interroger et à représenter le monde qui nous entoure — avec des objectifs et par des moyens différents —, à commencer par nous-mêmes, l'humain : cette "machine" surprenante! Je vous invite maintenant à parcourir avec moi l'épopée de ma recherche réalisée au cours de ces dix dernières années. Comme Ulysse, ce voyage est peuplé de stations, de rencontres, de découvertes, et de leçons — parfois dans la réussite, et souvent dans l'échec. Les trois stations principales de ce parcours sont marquées par les chapitres 1, 2, et 3 de ce manuscrit. La comparaison s'arrête là car, contrairement à Ulysse, le chercheur-arpenteur ne retourne jamais à ses origines. Dans la suite de ce manuscrit, je vous propose donc un cheminement à travers une sélection d'articles repensés dans la perspective générale de mes directions de recherche. J'ai profité de l'occasion pour reformuler les problématiques, repenser les notations, et corriger les éventuelles erreurs, afin de vous présenter ces recherches dans un cadre homogène et cohérent. Tout au long de ce manuscrit, je me suis essayé au difficile exercice de rassembler la rigueur formelle et l'interprétation générale dans un flot continu. Un ensemble d'illustrations sonores est disponible à l'adresse <https://nubo.ircam.fr/index.php/s/zGtyAFs6Fdp9F4S> pour vous accompagner dans votre parcours des chapitres de ce manuscrit. Le mot de passe nécessaire pour y accéder vous sera envoyé prochainement.

¹ Pour les amateurs de science-fiction, SIVA est un acronyme en référence au premier tome de la Trilogie Divine de l'auteur Philipp K. Dick dont le titre français est : Système Intelligent Vivant et Agissant

L'ART DE LA VOIX : SÉMIOLOGIE ET MODÈLES DE SIGNAUX POUR LA CARACTÉRISATION DE LA VOIX ACTÉE

1.1 INTRODUCTION

La communication orale ne se limite pas à la communication d'un message linguistique avec autrui (De Saussure, 1967). Notre voix encode et transmet de nombreuses informations non-verbales, certaines études tendent même à supporter l'hypothèse que la communication est essentiellement non-verbale (Mehrabian, 1972) : par la prosodie et le langage corporel, nous sommes capables de transmettre — sciemment ou indirectement — des informations sur notre identité, notre genre et notre âge, nos origines socio-géographiques, notre environnement professionnel, notre état de santé, notre état émotionnel, notre attitude vis-à-vis de notre énoncé ou de notre interlocuteur, nos valeurs morales. Ces informations sont dites d'ordre para-linguistiques lorsqu'elles participent à la compréhension de la communication (par exemple pour les émotions), ou extra-linguistiques lorsqu'il s'agit d'informations lorsqu'elles y sont extérieures (comme par exemple pour l'identité, le genre, ou l'âge). Notre voix, comme une modalité particulière de notre comportement, porte la trace de notre physiologie (notre corps), et de notre psycho-physiologie (la réaction biologique à notre environnement extérieur et le développement à travers lequel nous construisons une personnalité), généralement sous forme d'états temporaires (comme pour une émotion) ou de traits plus ou moins stables dans le temps (par exemple, notre personnalité). L'évolution de l'être humain et de ses facultés de sociabilisation ont progressivement fait basculer l'expression d'une réaction à l'environnement à une forme d'expression conventionnelles reproduites dans un environnement social, par détachement de la forme d'expression et de son origine biologique et environnementale (Ohala, 1996). Comme admirablement formulé dans (Morlec, 1997) : «on passe [...] de manière continue du "savoir" au "faire-savoir" puis au "faire-croire"». Par les conventions et les artifices, le faire-croire constitue la réalité de l'expression de l'acteur et du comédien, et de la construction vocale de la fiction d'un personnage, d'un rôle, d'une personnalité.

La personnalité — du latin *persona* — désigne le masque de théâtre antique grec ¹. En psychologie, la personnalité est un ensemble de traits émotionnels, d'attitudes, et de comportement qui constituent l'individualité d'une personne (McCrae and Costa, 1990). Par extension, la personnalité vocale regroupe l'ensemble des traits pouvant entrer dans l'expression et la perception d'une personnalité par la modalité vocale. Elle peut être simulée par un comédien ou un acteur pour incarner un personnage ou un rôle fictionnel. La définition de la personnalité vocale s'inscrit dans la lignée des recherches initiées en psychologie et en para-linguistique sur les états et les traits d'un locuteur — notamment par (Schuller and Batliner, 2013). Outre la question de savoir si une personnalité peut se transcrire et se manifester intégralement dans une voix, l'analyse de la personnalité vocale pose en premier lieu la question de la sémiologie de la voix, c'est-à-dire la capacité à définir un vocabulaire permettant de décrire les qualités et les traits perçus dans une voix. En effet, nous ne percevons pas une voix en tant qu'un signal brut, nous l'interprétons symboliquement à partir de nos représentations mentales, non seulement pour décoder le message communiqué mais pour tenter d'inférer les dispositions sociales, l'état émotionnel, l'identité, les origines sociales, professionnelles, ou encore de traits associés à la personnalité supposée de notre interlocuteur ². Une sémiologie de la voix nécessite alors de définir un vocabulaire permettant de décrire les qualités et les traits perçus dans une voix, puis de caractériser les corrélats acoustiques et leur intrication dans la réalisation de chacun de ces qualités ou de ces traits. Si il existe peut-être des modèles de personnalité vocale



¹ "All the world's a stage, And all the men and women merely players". As You Like It (1623). William Shakespeare.

² A ce sujet, le film Her nous rappelle à quel point une voix même dépourvue de corporalité peut être suggestive : on peut en déduire aussi bien des caractéristiques physiologiques que des émotions, des attitudes, des comportements à travers ses inflexions et sa prosodie.

similaires à ceux plus généraux proposés par la psychologie, nos recherches dans le domaine des voix actées tendent à suggérer que la configuration et l'importance relative de ces traits dépendent d'un domaine d'application et de ses finalités. En particulier, le vocabulaire généralement utilisé pour qualifier des voix de personnages de fiction ³ couvre tout un spectre allant de stéréotypes sociaux, moraux, ou fonctionnels définissant non pas nécessairement un *personnage* mais un rôle aux caractéristiques phonatoires et articulatoires (timbre voilée, voix rauque, etc...).

Il peut sembler contre-intuitif de s'intéresser à la voix actée, alors même que la psychologie s'en émancipe (Bänziger et al., 2012) pour mieux étudier les émotions dans la parole ordinaire, de tous les jours, ou spontanée — c'est-à-dire de l'expression d'émotions authentiquement ressenties (Cowie et al., 2011). Toutefois, la voix actée constitue un objet d'étude à part entière : celui des artifices et des conventions tout d'abord, et à terme peut-être celui de l'interprétation. En outre, la collection de données "ressenties" est une question de recherche en soi, et la création de bases de données de bonne qualité est plus facilement réalisable avec des voix actées. La voix actée offre un terrain privilégié pour l'étude de ces traits de personnalité dans la mesure où au théâtre comme au cinéma les comédiens ont tendance à grossir les traits pour ne pas laisser d'ambiguïté pour le spectateur sur les contours de la personnalité et des intentions d'un personnage et de la fiction - son rôle. En outre, l'étude de la voix actée nous permet d'observer des formes d'expressions comme le "chuchotement de scène" des comédiens de théâtre qui sont de purs artifices de convention que nous n'observons pas dans la parole de tous les jours. Il est d'autant plus intéressant que souvent un comédien est capable d'incarner par la voix des personnalités variées - réelles (doublage) ou fictionnelles (animation), ce que nous avons été amené à nommer sa "palette vocale". Pour ne citer qu'un exemple célèbre en France, on pense évidemment à la voix de Roger Carrel qui a incarné une quantité phénoménale de voix au rang desquelles les voix françaises de Peter Sellers et Peter Ustinov, la voix de Charlie Chaplin dans *Le Dictateur*, C-3PO dans la saga *Star Wars*, *Hercule Poirot*, Kermit dans la série *Le Muppet Show*, *Alf*, *Astérix*, *Winnie l'Ourson* : tout autant de personnages avec des personnalités et des voix extrêmement variées. La contrepartie des voix actées est qu'elles sont particulièrement expressives — en fait bien au-delà des limites du registre habituel de la parole conversationnelle, lue, ou spontanée — ce qui rend leur analyse résiste à la capacité des algorithmes traditionnels de traitement du signal ⁴. Cette limitation a nécessité de développer des algorithmes permettant d'extraire des descripteurs robustes à partir des signaux de parole - c'est-à-dire non sujets à des biais d'estimation (comme c'est le cas par exemple pour l'analyse de la F0 qui peut être fausse ou multiple dans le cas de phénomène de distorsions chaotiques ou alors l'estimation des paramètres d'un modèle de source glottique, comme nous le verrons au cours de ce chapitre.

Ce premier chapitre est consacré à la cognition de voix expressives, depuis la caractérisation acoustique jusqu'aux représentations abstraites de voix actées expressives, c'est-à-dire la catégorisation des voix selon des catégories reflétant des propriétés physiologiques et psychologiques d'un individu - sa personnalité vocale. Dans une première partie, je présente mes travaux réalisés sur la caractérisation de voix expressives, qui par la nature même de ces voix, souvent dans des registres extrêmes de la production vocale humaine, échappe à l'analyse. Les recherches résumées dans ce chapitre présentent une évolution des algorithmes de caractérisation de la qualité vocale de la voix actée, depuis les algorithmes basés sur des modèles de signaux comme l'estimation des caractéristiques de la source glottique et des rapports harmoniques à bruit (Obin, 2012) ou de la distorsion de phase (Degottex and Obin, 2014) à l'implémentation d'algorithmes robustes de mesure de niveaux de bruit multi-résolution basés sur les mesures d'entropies spectrales de Rényi (Obin and Liuni, 2012). Le résultat principal de cette première partie est que les modèles de signaux formalisés pour l'analyse de voix "standards" ne sont pas adaptés pour l'analyse de voix expressives. La proposition de descripteurs robustes — non soumis à l'estimation des paramètres d'un modèle de signal — notamment basés sur de l'analyse multi-résolution

³ Ce vocabulaire n'est pas partagé entre les acteurs du métier, entre les comédiens, les chercheurs de voix, ou les directeurs artistiques - ce qu'ont permis de mettre en lumière les enquêtes sociologiques réalisées pendant le projet ANR TheVoice.

⁴ Il est révélateur que la plupart des publications issues de ces recherches ont été présentées dans des sessions sur la parole "pathologique" ou "anormale"



Le chemin à parcourir dans l'analyse et la modélisation de voix expressives n'en est encore qu'à ses débuts - que ce soit dans ses corrélats acoustiques, dans ses fonctions linguistiques, mais aussi et possiblement esthétiques.

de la qualité vocale a permis d'améliorer substantiellement la caractérisation de ces voix. Dans un second temps, je présente mes travaux réalisés sur la thématique originale de la recommandation de voix assistée par ordinateur et de la mesure de la similarité perceptive entre des voix (Obin et al., 2014b). La problématique principale de cette recherche a été de mesurer la similarité perceptive entre des voix actées, et de la formaliser de manière à être capable de modéliser les critères utilisés par un opérateur humain professionnel dans son choix de voix similaires (par exemple pour le doublage ou plus généralement pour la post-production de voix). L'hypothèse formulée est que ce choix n'est pas opéré sur la base de similarités acoustiques mais sur une représentation mentale de la personnalité incarnée par une voix, et donc d'un ensemble de traits de personnalité dont l'un des objectifs a été de tenter de les définir de concert avec des professionnels de ce secteur d'activité. Le résultat principal de ces recherches est double : d'une part, les mesures de similarités acoustiques telles qu'utilisées pour la reconnaissance ou la vérification de locuteur sont inadaptées à rendre compte d'échelle de similarité entre de nombreuses voix ; et d'autre part, le passage par des représentations symboliques intermédiaires de la voix permet de mieux modéliser les processus cognitifs à l'œuvre dans la perception de la similarité vocale (Obin and Roebel, 2016).

Projets associés à ce chapitre (par ordre chronologique)

Projet FEDER Voice4Games (2011-2014)

Encadrements associés à ce chapitre (par ordre chronologique de fin)

Grégoire Bachman (2013-2014), post-doc, projet FEDER Voice4Games

François Lamare (2012, M2 ATIAM), Xavier Favory (2013), Rong Gong (2014, M2 ATIAM), Stéphane Rivaud (2015, M2 ATIAM)

Publications associées à ce chapitre (par ordre chronologique)

Obin, N. (2012). Cries and Whispers - Classification of Vocal Effort in Expressive Speech. In Interspeech, Portland, USA.

Obin, N. and Liuni, M. (2012). On the Generalization of Shannon Entropy for Speech Recognition. In IEEE workshop on Spoken Language Technology, Miami, USA.

Obin, N., Lamare, F., and Roebel, A. (2013). Syll-O-Matic : an Adaptive Time- Frequency Representation for the Automatic Segmentation of Speech into Syllables. In International Conference on Acoustics, Speech, and Signal Processing, Vancouver, Canada.

Obin, N., Roebel, A., and Bachman, G. (2014). On Automatic Voice Casting for Expressive Speech : Speaker Recognition vs. Speech Classification. In International Conference on Acoustics, Speech, and Signal Processing, Florence, Italy.

N.Obin, Belião, J., Veaux, C., and Lacheret, A. (2014). SLAM : Automatic Stylization and Labelling of Speech Melody. In Speech Prosody, page 246–250.

Degottex, G. and N.Obin (2014). Phase Distortion Statistics as a Representation of the Glottal Source : Application to the Classification of Voice Qualities. In Interspeech, Singapore, Singapore.

Obin, N. and Roebel, A. (2016). Similarity Search of Acted Voices for Automatic Voice Casting. IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP), 24(9) :1638–1647.

1.2 CARACTÉRISATION DE LA QUALITÉ VOCALE DE VOIX ACTÉES : SIGNAL, MODÈLE, INFORMATION

Dans une première partie, je présente l’historique de mes recherches liées à la caractérisation de la qualité vocale de voix expressives, avec comme application la classification automatique de l’effort vocal en trois catégories : chuchotée / douce (incluant le chuchotement de théâtre, c’est-à-dire le *stage whisper*), normale, et forte/criée réalisée à partir de bases de données de voix actées correspondant aux enregistrements de jeux-vidéos multilingues (français et anglais). Je commence par présenter les descripteurs issus de l’analyse par des algorithmes traditionnels de traitement du signal comme le modèle de glotte de LF, ou la décomposition sinusoïde / bruit, pour aller vers des représentations robustes issues de la théorie de l’information comme les mesures d’entropie de Shannon, Wiener, ou Rényi.

L’effort vocal correspond à l’ajustement de l’intensité vocale d’un locuteur en fonction de la distance de communication avec l’auditeur. Un changement de l’effort vocal provoque un changement de l’intensité vocale qui conduit à des modifications substantielles dans la configuration des mécanismes de production vocale, particulièrement phonatoires. Les études sur l’effort vocal supposent généralement 5 configurations phonatoires : chuchoté, doux, normal, fort et crié ⁵. Un grand nombre d’études ont été consacrées à la description des mécanismes impliqués dans la production des chuchotements (Monoson and Zemlin, 1984; Solomon et al., 1989; Sundberg et al., 2010), et des cris (Rostolland, 1982), et les différences de configuration entre la parole douce, normale et forte (Holmberg et al., 1988). Le vrai chuchotement ⁶ fait référence à l’excitation du conduit vocal par des plis vocaux à moitiés fermés. Le chuchotement de scène ⁷ est la simulation de la parole chuchotée par des acteurs professionnels afin que la voix soit suffisamment forte pour qu’elle soit audible et compréhensible par le public. Le chuchotement de scène diffère du vrai chuchotement par son caractère soufflé et partiellement voisé (Solomon et al., 1989). La voix criée est liée à une augmentation de la F0 due au augmentation de la pression sous-glottale utilisée pour augmenter l’intensité vocale (Harwardt, 2002), et des non-linéarités en raison de l’interaction non linéaire du flux d’air et des tourbillons d’air près des plis vocaux produisant des signaux d’excitation supplémentaires. Une augmentation de l’effort vocal affecte également les caractéristiques du conduit vocal, de la tension musculaire du conduit vocal et de ses résonances.

Modèle de source glottique

La source glottique tient une place particulièrement importante dans l’expression de voix expressives et en particulier de l’effort vocal. Basé sur le modèle paramétrique de source glottique de Liljencrants-Fant (Fant, 1995), (Degottex et al., 2009, 2011) a proposé une implémentation de ce modèle dans un algorithme de séparation de la source glottique et du conduit vocal. Dans ce modèle, la source glottique est représentée par le coefficient de relaxation glottique LF-RD qui caractérise la forme de l’impulsion glottique (Degottex et al., 2011). A partir de l’estimation de la forme des impulsions glottique, un algorithme détermine avec précision les instants de fermeture glottique (GCI) dans le signal de parole (Degottex et al., 2009). Ce modèle de source glottique permet en outre de décrire les changements de qualité vocale induits par des gradations de mode de phonation, depuis une voix tendue jusqu’à une voix relâchée comme illustré sur la **Figure 1.1**.

Nous définissons la régularité des GCIs de la manière suivante :

$$\Delta GCI(n) = |\Delta^2 \log(GCI(n) - GCI(n - 1))| \quad (1.1)$$

⁵ La parole Lombard et sous stress pourraient constituer deux autres configurations liés à des modes de phonations spécifiques.

⁶ En anglais, “true whisper” ou chuchotement à faible effort.

⁷ En anglais, “stage whisper” ou chuchotement à fort effort.

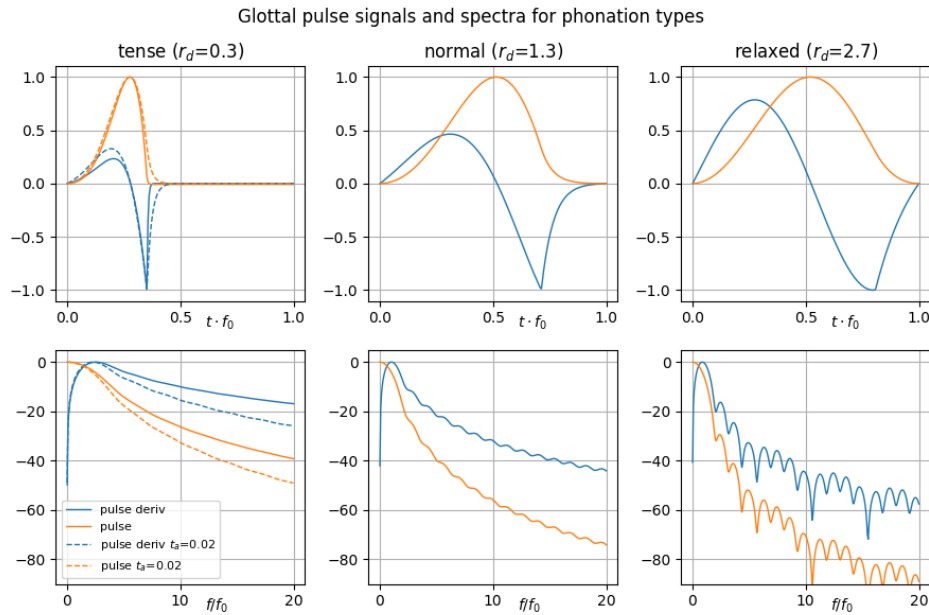


FIGURE 1.1 – Caractérisation de l’impulsion glottique et de sa dérivée (en haut) et des spectres correspondants (en bas) pour les modes de phonations : voix tendue (à gauche), voix normale (au milieu), voix relâchées (à droite). D’après (Degottex et al., 2011).

où : $\Delta^2(\cdot)$ est l’opérateur de dérivée numérique du second ordre et $GCI(n)$ la n -ème position temporelle des GCIs.

Dans la suite, nous référons à ces deux caractéristiques comme les caractéristiques de qualité vocale (VQ) dans la mesure où elles rendent compte explicitement de mécanismes phonatoires liés à la qualité vocale : le coefficient R_D donne une mesure du degré de *tension* dans la voix, et la régularité des impulsions glottiques GCIs reflète les phénomènes de *craquement* dans la voix et leur absence le *soufflement* dans la voix.

La **Figure 1.2** présente la distribution des caractéristiques VQ (R_D , Δ GCI) pour des voix chuchotées/douces, normales, ou fortes/crées. Cette figure illustre que les changements de l’effort vocal induisent des différences importantes dans la configuration de la source glottique. En particulier, ces distributions confirment qu’un accroissement de l’effort vocal est lié à une augmentation de la tension du muscle du conduit vocal. Par ailleurs, on observe une dispersion importante de la régularité des impulsions glottiques pour les voix chuchotées/douces, ce qui reflète vraisemblablement la présence de bruit, de souffle, et de craquements pour ces voix.

Mesure Multi-Résolution du Rapport Harmonique sur Bruit (HNR)

Au-delà d’une description précise des caractéristiques de la source glottique qui nécessite l’emploi de modèles de source glottique avancés mais complexes et encore insuffisamment fiables, nous avons fait l’hypothèse simple que le mélange entre des impulsions périodiques et du bruit résiduel est fortement lié aux mécanismes de modes de phonation associées à l’effort vocal et aux qualités vocales. Une mesure du rapport harmonique-à-bruit (Griffin and Lim, 1985) (HNR) est introduite pour caractériser l’ensemble de ces phénomènes liés à l’effort vocal ou plus généralement à la qualité vocale. Une version multi-résolution, c’est-à-dire la mesure du HNR par bande de fréquences, permet de rendre compte de ces phénomènes de manière différenciée en fonction des fréquences. Par opposition aux mesures de décision de voisement (voiced/unvoiced), de coefficient de voisement, ou

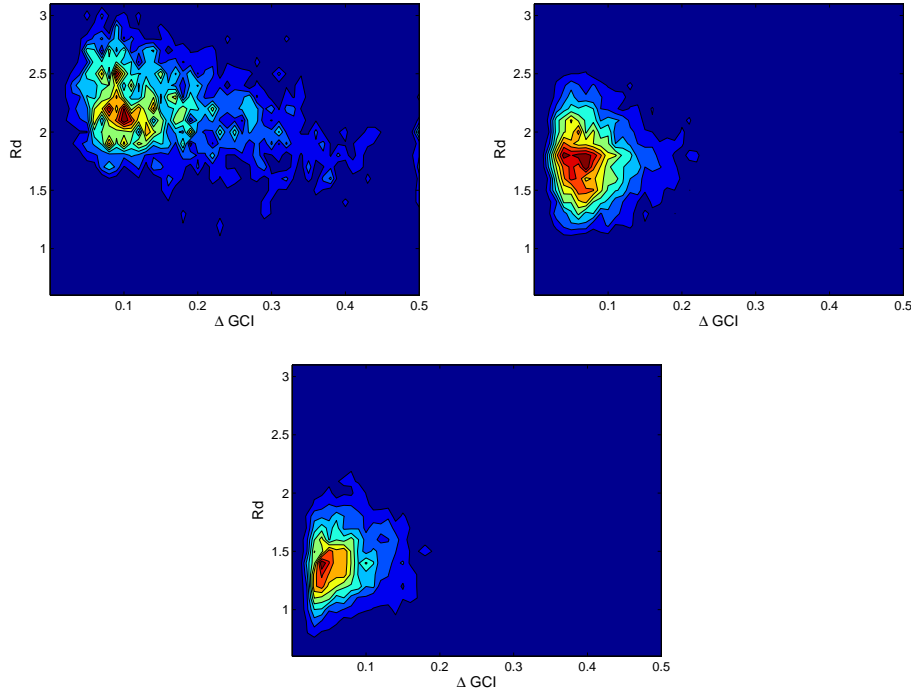


FIGURE 1.2 – Distribution des caractéristiques de qualité vocale (Δ GCI, R_d) pour les voix chuchotées/douces (en haut à gauche), normales (en haut à droite), et fortes/crées (en bas).

de fréquence maximale de voisement (Voiced/Unvoiced Frequency ou VUF), la mesure proposée permet de caractériser de manière différenciée le rapport de voisement et de bruit en fonction des régions fréquentielles. Dans chaque bande de fréquence i , le HNR est mesuré comme :

$$\text{HNR}^{(i)} = \frac{\sum_{k=1}^{K^{(i)}} |A_H(k)|^2}{\sum_{n=1}^{N^{(i)}} |A(n)|^2} \quad (1.2)$$

où : i représente la i -ème bande de fréquence, $|A_H(k)|$ l'amplitude de la k -ème harmonique, et $|A(n)|$ l'amplitude du n -ème point fréquentiel du spectre dans la bande fréquentielle considérée, $K^{(i)}$ le nombre totale d'harmoniques dans la i -ème bande de fréquence, et $N^{(i)}$ le nombre total de points fréquentiel de la i -ème bande de fréquence. Suivant cette définition, le HNR est égal à 0 lorsqu'il n'existe pas de contenu harmonique dans la bande de fréquence, et égal à 1 si la bande de fréquence ne contient que du contenu harmonique. Ici, la décomposition harmonique et bruit du spectre à court-terme est obtenu en utilisant l'algorithme de classification sinusoïde et bruit présenté dans (Zivanovic et al., 2008). La mesure définie a été calculée sur une échelle de 25 filtres en échelle mel.

Mesure d'Entropie Spectrale Multi-Résolution

Dans (Obin and Liuni, 2012), nous avons formalisé avec Marco Liuni une mesure de niveau de bruit basée sur l'entropie de Rényi et appliqué ces mesures pour la tâche de classification de qualités vocales. Cette recherche prolonge et renforce les recherches précédemment décrites, le niveau de bruit constitue une mesure simple et fiable permettant de capturer la plupart des modifications de la source glottique et des modes de phonation associés à un grand nombre de qualités vocales (Laver, 1980). Les mesures existantes utilisées pour mesurer le niveau de bruit dans un signal audio reposent soit sur une décomposition harmonique et bruit telle que décrite précédemment ou bien alors sur des mesures d'entropie spectrales, comme l'entropie de Shannon (Shannon, 1948; Misra et al.,

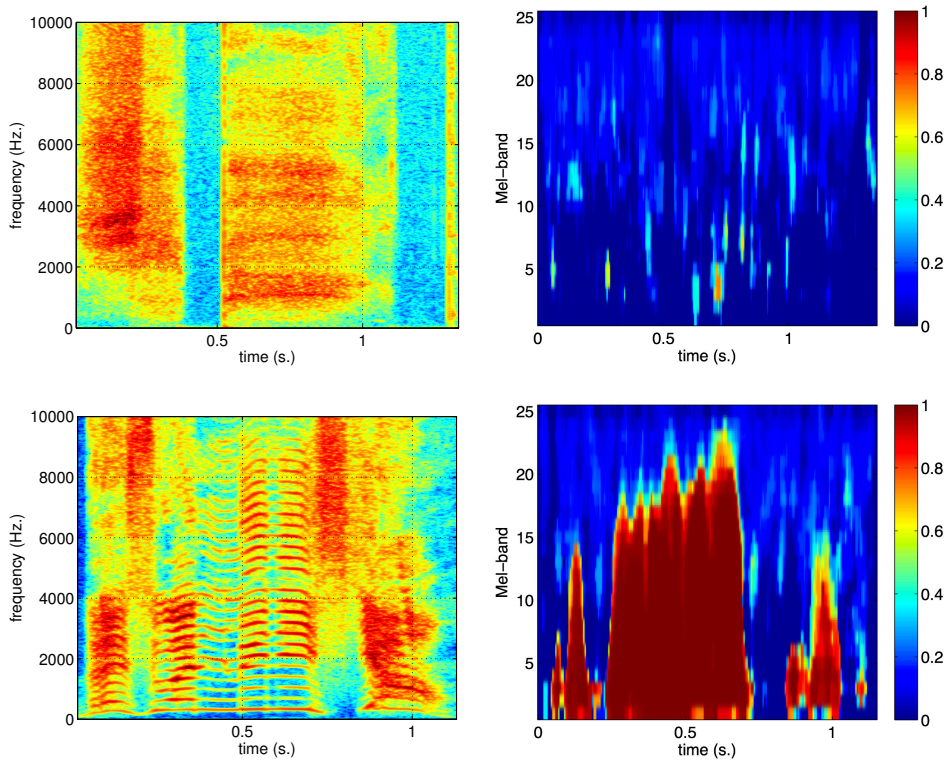


FIGURE 1.3 – Illustration des mesures de HNR obtenues pour un enregistrement de voix chuchotée (en haut) et criée (en bas).

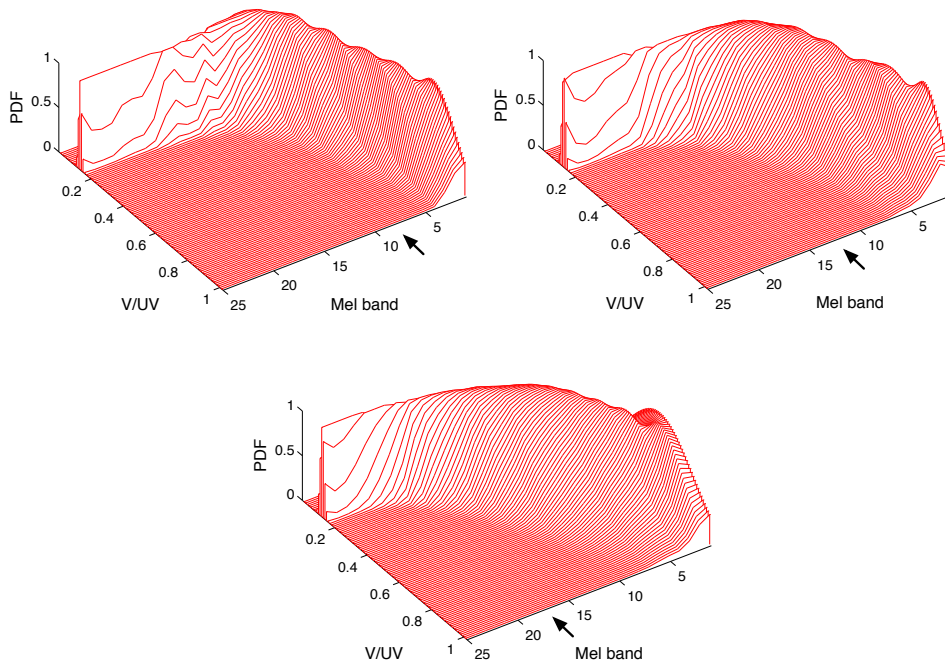


FIGURE 1.4 – Distribution des mesures HNR en échelle mel pour les voix chuchotées/douces (en haut à gauche), normales (en haut à droite), et fortes/criées (en bas). Les flèches noires indiquent la fréquence au-delà de laquelle aucun contenu harmonique est observé. Par exemple : $f_W = 9$, $f_N = 14$, et $f_S = 18$ pour $HNR \leq 0.2$

2004) ou l'entropie de Wiener (Wiener, 1954; Johnston, 1988; Dubnov, 2004). L'avantage de ces dernières mesures est qu'elles reposent exclusivement sur une valeur calculée directement sur le signal et donc ne sont pas sujettes aux erreurs d'estimation des algorithmes reposant sur des modèles de signaux (comme le modèle de source glottique ou le modèle sinusoïde et bruit). En particulier, l'entropie de Rényi présente l'avantage sur les autres mesures d'entropie de se focaliser sur le contenu harmonique (présence d'une amplitude prééminente à l'intérieur de la distribution des amplitudes spectrales) ou sur le contenu bruité (distribution uniforme des amplitudes spectrales) - et sans nécessiter de décomposition sinusoïde et bruit du signal.

Mesure d'entropie spectrale

Avec une normalisation appropriée, le spectre de puissance d'un signal audio peut être interprété comme une densité de probabilité. À partir de cette interprétation, le formalisme du domaine des probabilités et de la théorie de l'information peut être appliqué aux signaux audio. En particulier, l'entropie peut être utilisée pour déterminer une mesure de concentration d'une densité temps-fréquence. Cette mesure peut alors être interprétée comme une mesure du degré de périodicité (alternativement de bruit) d'un signal audio.

Entropie de Rényi

Definition 1.2.1 Étant donné une densité de probabilité discrète et finie $P = (P_1, \dots, P_N)$ et un nombre réel $\alpha \geq 0$, $\alpha \neq 0$, l'entropie de Rényi de P est définie comme,

$$H_\alpha[P] = \frac{1}{1-\alpha} \log_2 \sum_{n=1}^N P_n^\alpha, \quad (1.3)$$

où : P est entre crochet pour indiquer qu'une densité discrète est considérée.

Parmi les propriétés générales de l'entropie de Rényi (Rényi, 1961; Beck and Schögl, 1993; Baraniuk et al., 2001), les principales sont rappelées ci-après.

- 1) pour toute densité de probabilité discrète P , l'entropie de Rényi $H_\alpha[P]$ tend vers l'entropie de Shannon de P lorsque l'ordre α tend vers un.
- 2) $H_\alpha[P]$ est une fonction non croissante de α , telle que :

$$\alpha_1 < \alpha_2 \Rightarrow H_{\alpha_1}[P] \geq H_{\alpha_2}[P]. \quad (1.4)$$

Pour des densités discrètes et finies, le cas $\alpha = 0$ peut aussi être considéré, ce qui donne simplement le logarithme du nombre d'éléments de P ; en conséquence, $H_0[P] \geq H_\alpha[P]$ pour toute valeur possible de α .

- 3) pour tout ordre α , l'entropie de Rényi H_α est maximum lorsque P est distribué uniformément, et est minimum et égal à zéro lorsque P a une seule valeur non nulle.

L'avantage principal des entropies de Rényi est la dépendance sur l'ordre α , ce qui donne accès à différentes représentations de la concentration en fonction de chaque valeur de α . En particulier, les petites valeurs de α tendent à souligner le contenu bruité du signal, tandis que les grandes valeurs de α tendent à souligner le contenu harmonique du signal.

Entropie Spectrale Multi-Résolution

La mesure de l'entropie spectrale multi-résolution, c'est-à-dire des mesures d'entropie spectrale par bande de fréquences, est employée pour mesurer l'entropie des distributions spectrales de manière différenciée et locale en fréquence. Pour chaque bande de fréquence i , l'ENTROPIE DE RÉNYI est mesurée comme :

$$H_{\alpha}^{(i)} = \frac{1}{1-\alpha} \log_2 \sum_{n=1}^{N^{(i)}} \left(\frac{|A(n)|^2}{\sum_{n=1}^{N^{(i)}} |A(n)|^2} \right)^{\alpha} \quad (1.5)$$

où : $|A(n)|$ est l'amplitude du n -ème point fréquentiel dans la bande considérée, et le terme $\sum_{n=1}^{N^{(i)}} |A(n)|^2$ au dénominateur correspond au facteur de normalisation du spectre de puissance dans la bande de fréquence considérée pour définir une densité de probabilité. La mesure définie a été calculée sur une échelle de 25 filtres en échelle mel. La **Figure 1.5** illustre l'entropie de Rényi multi-résolution sur des enregistrements de voix chuchotée et crlée, pour différentes valeurs de l'ordre α .

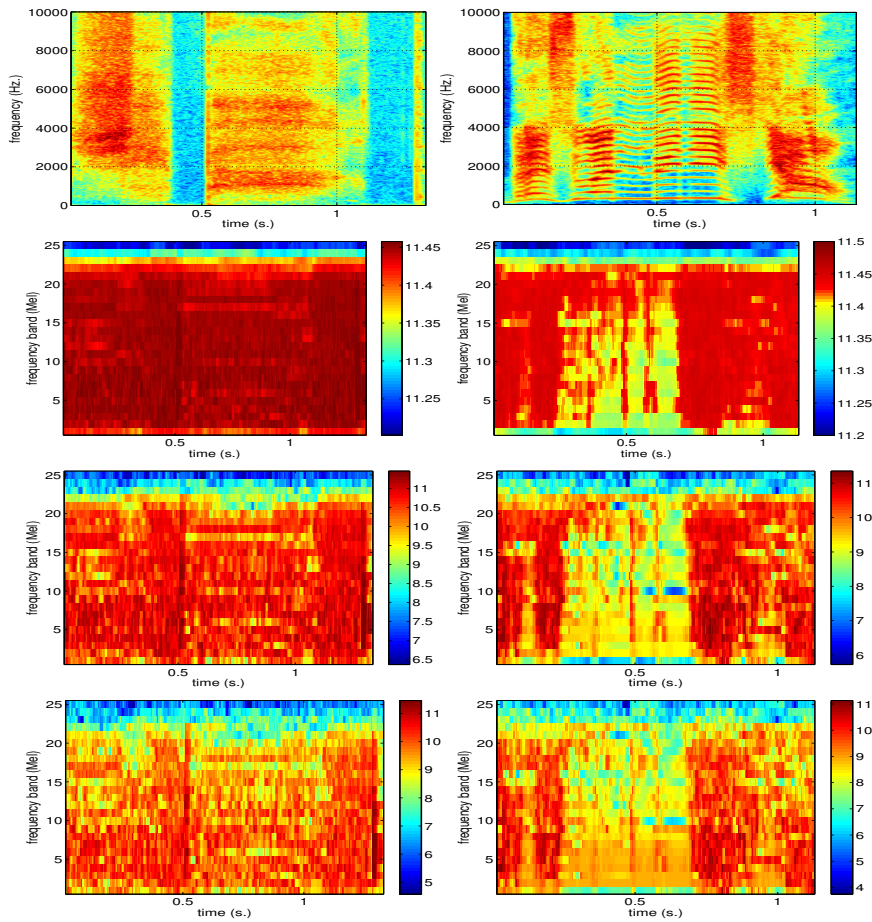


FIGURE 1.5 – Mesures d'entropie de Rényi sur des enregistrements de voix chuchotée (à gauche) et crlée (à droite). De haut en bas : spectrogramme, et entropie de Rényi pour les ordres $\alpha = 0.001$, $\alpha = 1$ (entropie de Shannon), et $\alpha = 3$.

L'ensemble des mesures introduites précédemment ont été évaluées expérimentalement sur des tâches de reconnaissance de l'effort vocal (chuchoté/doux, normal, fort/cré) ou de qualités vocales (normal, soufflé, craqué) à partir des bases de données de voix actées naturellement expressives. Les bases utilisées correspondent à l'intégralité des sessions d'enregistrement des voix de jeux de rôle massivement multi-joueurs (MM-RPG), et possiblement dans plusieurs langues : anglais-américain, français, allemand. Chaque MM-RPG comprend environ 10 à 20 heures d'enregistrements de parole dans des conditions de studio professionnel, réparties en des dizaines de milliers de fichiers correspondants à des phrases isolées du script. L'architecture de classification est basée sur un système GMM-UBM/SVM qui était alors un standard en reconnaissance de locuteur (Reynolds et al., 2000). Un modèle du monde universel (GMM-UBM) modélise la distribution acoustique des voix par un mélange de Gaussiennes (GMM) à partir de vecteurs représentant les caractéristiques à court-terme du signal audio (MFCC et autres mesures). Chaque enregistrement est alors représenté comme un super-vecteur par adaptation MAP des vecteurs de moyennes du GMM-UBM sur l'enregistrement. Les supervecteurs sont alors utilisés comme représentation en entrée d'un algorithme de classification par MACHINES À VECTEURS DE SUPPORT (SVM) qui cherche à déterminer par projection dans un espace de haute dimension l'hyperplan qui maximise la marge de séparation entre les classes pour opérer la classification (Campbell et al., 2006). Les détails d'implémentation et de méthodologie d'expérimentation sont décrits dans les articles mentionnés. Je me contente dans la suite de présenter et de commenter les principaux résultats obtenus.

Le **Tableau 1.1** présente les scores de reconnaissances de l'effort vocal pour les modèles appris à partir de représentations MFCC, la combinaison des représentations proposées avec les MFCC, et l'ensemble optimal de représentations proposées.

MASS EFFECT	WHISPERED	NORMAL	SHOUTED	TOTAL
MFCC	69.4	82.5	91.0	81.1
MFCC + TEO	72.5	83.6	91.6	82.5
MFCC + HNR	75.0	84.3	91.2	83.5
MFCC + VQ	76.9	84.7	92.0	84.7
⋮	⋮	⋮	⋮	⋮
MFCC + TEO + HNR + VQ	79.1	88.4	93.5	87.0

TABLEAU 1.1 – F-mesure obtenue avec les MFCC et les combinaisons des mesures proposées avec les MFCC sur le MM-RPG MASS EFFECT.

Le **Tableau 1.2** présente les performances obtenues pour les mesures de représentations de bruits proposés, et en particulier pour des valeurs du coefficient α de l'entropie de Rényi. Les résultats s'interprètent en deux temps. Dans un premier temps, les mesures de source glottique (VQ) et de rapport harmonique-à-bruit multi-résolution (HNR) apportent bien de l'information supplémentaire pour la reconnaissance de l'effort vocal par rapport aux caractéristiques traditionnelles comme les MFCCs. Une analyse post-hoc des résultats indiquent que, par exemple, pour les mesures de source glottique, il semblerait que les erreurs du modèle et leur distribution soient informatives au même titre que les valeurs de ces paramètres pour la caractérisation et la reconnaissance des modes d'effort vocal. Dans un second temps, une comparaison des mesures de bruits dans la voix montre que l'entropie de Rényi multi-résolution est un indicateur robuste de l'effort vocal. L'entropie de Rényi se compare favorablement pour la caractérisation du bruit à la mesure de rapport harmonique-à-bruit (HNR) présentée précédemment et à la mesure de platitude spectrale ou entropie de Wiener (SFM). L'entropie de Rényi obtient de meilleures performances en particulier pour les petites valeurs du coefficient α qui ont tendance à réhausser le

MASS EFFECT	WHISPERED	NORMAL	SHOUTED	TOTAL	DRAGON AGE	WHISPERED	NORMAL	SHOUTED	TOTAL
FRENCH					GERMAN				
MFCC	69.4	82.5	91.0	81.1	MFCC	73.1	77.2	87.1	79.0
MFCC + SFM	73.1	84.4	92.6	83.3	MFCC + SFM	73.4	77.9	87.5	79.6
MFCC + HNR	75.0	84.3	91.2	83.5	MFCC + VUV	74.3	78.7	87.2	80.1
entropy					entropy				
MFCC + α					MFCC + α				
$\alpha = 0.001$	75.0	84.0	91.3	83.4	$\alpha = 0.001$	74.6	78.0	87.2	79.9
$\alpha = 0.01$	75.2	84.2	92.0	83.8	$\alpha = 0.01$	75.1	78.5	87.5	80.4
$\alpha = 0.1$	75.0	84.3	91.3	83.5	$\alpha = 0.1$	74.5	78.0	87.5	80.1
$\alpha = 0.2$	74.6	84.0	91.6	83.4	$\alpha = 0.2$	74.5	78.2	87.3	80.0
$\alpha = 0.4$	74.5	84.0	91.5	83.4	$\alpha = 0.4$	74.2	78.0	87.3	79.8
$\alpha = 0.6$	74.2	84.0	91.3	83.2	$\alpha = 0.6$	74.0	77.9	87.0	79.6
$\alpha = 0.8$	74.2	84.0	91.2	83.0	$\alpha = 0.8$	74.0	77.7	87.0	79.6
$\alpha = 1$ (SHANNON)	74.2	83.9	91.0	83.0	$\alpha = 1$ (SHANNON)	73.8	77.5	87.1	79.5
$\alpha = 2$	75.0	84.3	91.3	83.5	$\alpha = 2$	73.6	77.8	87.4	79.6
$\alpha = 3$	74.5	84.0	91.5	83.3	$\alpha = 3$	72.8	77.0	87.4	79.1
$\alpha = 4$	74.6	83.8	91.5	83.2	$\alpha = 4$	72.5	77.1	86.7	78.8
$\alpha = 5$	74.1	83.8	91.2	83.0	$\alpha = 5$	72.4	77.0	86.5	78.5

TABLEAU 1.2 – F-mesure obtenue avec de la représentation MFCC, les MFCC + les mesures signal de rapport harmonique-à-bruit : (HNR), et les MFCC + les mesures D'ENTROPIE (SFM ou entropie de Wiener, entropie de SHANNON, et entropie de RÉNYI) pour les MM-RPG MASS EFFECT (en français) et DRAGON AGE (en allemand).

contenu bruité dans le signal de parole, ce qui semble une information pertinente pour la reconnaissance des modes d'effort vocal pour des voix expressives.

1.3 SYNTHÈSE ET DISCUSSION 1

Pour conclure la première partie de ce chapitre, je dresse une brève synthèse des principales contributions réalisées au cours de ces recherches, et ouvre des pistes de réflexion à partir des résultats obtenus.

Contributions

C1. Définition de mesures pour la caractérisation de la qualité vocale de voix expressives, basées sur des **modèles de signaux source glottique / conduit vocal** (par exemple, le modèle LF-Rd), le **rapport harmonique-à-bruit** (HNR) mesuré à partir d'une décomposition du spectre d'amplitude en sinusoides et bruit, ou sur des mesures d'entropie spectrale comme l'**entropie de Rényi**.

Ces recherches ont apporté des premiers éléments de réponses aux 3 questions de recherche formulées dans l'introduction.

Q1. et Q3. Quels modèles de signaux pour quels modèles d'apprentissage? Les voix expressives échappent à l'analyse par les modèles paramétriques de signaux (comme le vocodeur ou le modèle LF-Rd) dans la mesure où ces voix constituent des cas limites des hypothèses d'application de ces modèles. En termes de modélisation pour la caractérisation, les expérimentations menées au cours de ces recherches ont en particulier démontré le manque de robustesse de l'estimation de ces paramètres pour des voix expressives. Le passage graduel de modèles paramétriques fortement contraints (c'est-à-dire reposant

sur des hypothèses fortes sur la nature du signal, comme par exemple le modèle LF-Rd), à des modèles paramétriques avec des contraintes relâchées (par exemple, le modèle sinusoïde et bruit utilisé dans les mesures de rapport harmonique-à-bruit HNR), et enfin à des caractérisations non-paramétriques (par exemple fondées sur des mesures d'information directement à partir du signal) ont permis d'implémenter des algorithmes robustes pour la caractérisation de voix expressives. Non seulement ces mesures ne sont pas sujettes à des erreurs d'estimation particulièrement fréquentes dans l'analyse de voix expressives mais elles se sont révélées efficaces pour la caractérisation de la qualité vocale et notamment la description du bruit dans la voix et plus généralement de l'aspect organisé ou chaotique du signal de voix (comme à travers les mesures d'entropie spectrale). En outre, les expérimentations ont également montré la nécessité de caractériser le signal de voix dans son intégralité plutôt que par des paramètres séparés et identifiés préalablement. L'hypothèse sous-jacente est que les paramètres de la voix (hauteur, intensité, bruit, etc...) ne sont pas à même de rendre compte de la diversité des voix observées, mais surtout que ces paramètres sont couplés de manière complexe. Dans la mesure où les modèles de signaux existants ne rendent pas compte de cette complexité, la solution envisagée déjà à l'époque a été de modéliser le signal de voix dans son intégralité et déplacer les problèmes de la complexité à l'apprentissage machine que ce soit ici la modélisation statistique de la distribution acoustique ou l'identification de séparation non-linéaire pour la classification. Cette hypothèse marque une première étape de glissement entre les approches signal (connaissances humaines formulées explicitement) et les approches apprentissage (émergence de l'information à partir des données), qui va être largement reprise et exploitée quelques années plus tard avec les réseaux de neurones profonds, comme nous le verrons dans le Chapitre 3.

1.4 SÉMILOGIE ET APPRENTISSAGE

La question de la sémiologie de la voix traverse comme une question de fond l'ensemble de mes projets réalisés depuis 2012, et plus particulièrement les projets Voice4Games, TheVoice, et MoVE liés à la mesure de la similarité perceptive entre des voix actées. Menés en collaboration avec des utilisateurs experts (comédiens, chercheurs de voix, et directeurs artistiques), des chercheurs en sociologie (projet ANR TheVoice) ou en psychologie (projet IDF MoVE), ces projets ont eu le mérite d'étendre mes thématiques de recherche en signal-information-apprentissage en traitement automatique de la parole pour inclure les représentations abstraites de la voix, issues de la physiologie, de la linguistique, de la psychologie, ou de la sociologie.

Dans une seconde partie, je présente mes travaux préliminaires pour tenter de formuler une proposition d'une sémiologie de la voix — en particulier expressives, c'est-à-dire la qualification de la voix par des représentations symboliques, et dont la finalité serait de proposer une représentation standardisée à la fois minimale (un minimum de termes nécessaires à la description) et exhaustive (les termes utilisés permettent de rendre compte de la plupart des voix et de leur spécificité). La question de la sémiologie de la voix traverse en fond l'ensemble de mes projets réalisés sur la voix depuis 2012. Si elle s'est d'abord présentée par intuition et par nécessité pratique⁸, elle s'est graduellement affirmée comme une problématique centrale dans mes recherches, menées en collaboration avec des sociologues pour la qualification de voix actées¹. Aujourd'hui, l'expressivité et la sémiologie constituent des lignes de force de ma recherche, comme la manipulation ou la génération réaliste de la voix avec pour application l'étude des biais cognitifs chez l'être humain dans le projet EMERGENCE ReVOLT (REvealing human bias with real Time VOlcal deep fakes, 2022-2023) en collaboration avec les neuro-scientifiques Pablo Arias et

⁸ La définition d'une taxonomie vocale a été un aller-retour entre descriptions scientifiques existantes et spécifications liées à des besoins particuliers.

¹. Dans le cadre du projet ANR TheVoice avec les sociologues du Centre Norbert Elias à Avignon en partenariat avec le Festival de Théâtre d'Avignon.

Jean-Julien Aucouturier, l'étude de l'effet de l'expressivité de voix réelles ou artificielles sur l'imagination dans le projet ANR EXOVOICES (2023-2027) en cours avec Jerome Sackur du Laboratoire de Sciences Cognitives et Psycholinguistique (LSCP) à l'École Normale Supérieure avec la thèse de Théodor Lemerle (Lemerle, 2023), ou encore la manipulation des attributs de la voix dans la thèse de Léane Salais (Salais, 2021).

Cet objectif s'est heurté de fait à deux problèmes principaux. D'une part, si la littérature est riche en représentations symboliques de la voix, ces représentations sont hétérogènes et relèvent de disciplines différentes, selon l'angle et la focale avec lesquels on s'intéresse à la voix. Ces disciplines incluent : la physiologie de l'appareil phonatoire et les mécanismes de production vocale souvent issus des sciences médicales (anatomie), les sciences du langage pour la description des formes et des fonctions du langage oral (phonétique (Przedlacka, 2012), prosodie (Pfitzinger, 2006), et phonologie (Pierrehumbert, 1980; Gussenhoven, 2004) mais aussi style et expressivité (Léon, 1993)), de la psychologie (émotions (Darwin, 1890; Ekman and Friesen, 1971), personnalité et attitudes sociales) (Schuller and Batliner, 2013), voir de la sociologie (Demichel-Basnier, 2019) et de l'anthropologie (Le Breton, 2019; Smits, 2021). L'étude des principaux traits perçus dans une voix ou chez une personne a permis de formuler une première proposition consensuelle : l'âge et le genre (Kido and Kasuya, 2001) pour ce qui est des caractéristiques physiologiques dépendant de facteurs biologiques externes, la description phonétique des modes de phonation et d'articulation (Laver, 1980), les cinq émotions primaires pour ce qui est des états émotionnels ressentis ou exprimés (Ekman and Friesen, 1971), ou encore l'amicalité et la dominance pour ce qui est des principaux traits de personnalité en psychologie sociale (Fiske et al., 2007). Néanmoins, certaines descriptions manquent encore clairement de consensus : la nature catégorielle ou dimensionnelle des émotions, ou alors sont complexes et très spécialisées : par exemple, les traits phonologiques d'une langue. Par ailleurs, les représentations existantes sont souvent limitées à la description de voix de laboratoire ou de tous les jours et inadaptées à la description des voix actées et naturellement expressives. Pour ne donner qu'un exemple, la qualité vocale - encore largement décrite comme des modes de phonations de nature pathologique - constitue une dimension prosodique essentielle dans la communication orale (Campbell and Mokhtari, 2003). Ses fonctions linguistiques sont encore largement inexplorées, en particulier ses fonctions esthétiques (Jakobson, 1960) dans le contexte de l'interprétation et de voix actées. Par ailleurs, les descriptions scientifiques fonctionnelles ne sont pas non plus nécessairement adaptées aux besoins de domaines d'application impliquant des axes de perception et de description fortement guidés par le contexte, la situation, ou l'objectif ⁹

Cette section présente les premières tentatives de formalisation d'une taxonomie de la voix actée qui ont été réalisées dans le cadre du projet FEDER Voice4Games (2011-2015), et présentées initialement dans (Obin et al., 2014b) et approfondies dans (Obin and Roebel, 2016). Cette formalisation a découlé de l'hypothèse selon laquelle la perception de la similarité entre des voix n'est pas purement acoustique mais intègre de nombreux facteurs associés à des représentations symboliques ou mentales utilisées pour qualifier et caractériser les attributs d'une voix. L'association de cette représentation structurée de la voix avec les mesures de signal et d'information et l'apprentissage automatique présentés dans la section précédente démontrent l'hypothèse de recherche suivante : la perception de la similarité entre des voix se manifeste principalement à une similarité d'un ensemble de représentations abstraites, c'est-à-dire l'actualisation d'une voix à transmettre un ensemble de traits prédominants, qu'à une pure similarité acoustique. **Le résultat principal de ces recherches est que la mesure de la similarité entre des voix à partir de représentations symboliques intermédiaires rend mieux compte de la similarité perceptive entre des voix actées qu'une mesure fondée exclusivement sur une similarité acoustique mesurée directement à partir du signal audio.**

⁹ Ainsi dans le doublage, une voix n'est jugée que par sa capacité à incarner de manière crédible un rôle (un rôle est défini par l'ensemble des traits physiques, psychologiques, et moraux d'un personnage et de sa fonction au sein d'un récit. Dans les fictions, ces traits sont généralement stéréotypés pour ne pas introduire d'ambiguïté dans les traits d'un personnage), ses caractéristiques doivent optimiser cet objectif.

La description proposée pour qualifier une voix s'intéresse exclusivement aux facteurs para- et extra-linguistiques et a fait l'objet d'une étude approfondie de la littérature sur la description des états et des traits d'un individu et en particulier à partir de sa voix. Elle couvre une grande variété d'informations : les traits physiologiques (Kido and Kasuya, 2001), comme l'âge et le genre (Schuller et al., 2010); les traits acoustiques associés à des modes de phonation ou d'articulation, de qualité vocale ou du timbre, les traits correspondent à l'ensemble des caractéristiques externes persistantes d'un individu, par exemple ses traits de personnalité (Schuller et al., 2012a); et les états (ressentis ou exprimés) correspondent à l'ensemble des caractéristiques internes et temporaires d'un individu, par exemple ses états émotionnels (Scherer et al., 1991; Schuller et al., 2009, 2012b). Un élément important qui est ressorti de mes diverses collaborations relatives à des problématiques de qualification de voix actées est la notion prédominante de stéréotypes de rôle. Finalement, une voix se définit principalement par la manière d'incarner un rôle, c'est-à-dire les caractéristiques physiques et les valeurs morales d'un personnage, sa relation aux autres personnages, et sa fonction dans le déroulement d'une trame.

TAXONOMIE VOCALE		
TRAITS BIOLOGIQUES	SEXE	homme, femme
	ÂGE	enfant, adolescent, jeune adulte, adulte, vieux, très vieux
ÉTATS ÉMOTIONNELS	ÉMOTIONS	neutre, tristesse, joie, colère, peur, surprise,
		excitation, tendresse, stress, autre
INTERPRÉTATION	ATTITUDE/MODALITÉ	affirmation, confirmation, exclamation, interrogation, ordre, autre
	SITUATION	action, conversation, information, monologue, autre
	ARCHÉTYPES	annonce, intelligence artificielle, soldat basique, brute, commandant, héros, neutre, vieux sage, soldat débutant, soldat vétéran, autre .
PHONATION	QUALITÉ VOCALE	soufflée, craquée, rauque
	TENSION	relâchée, normale, tendue, pressée
	VOCAL EFFORT	chuchotée/douce, normale, forte/criée
ARTICULATION	ARTICULATION	hypo-, normale, hyper- articulée
PROSODIE	TESSITURE	extrême-bas, basse, medium, haut, extrême-haut
	REGISTRE DE HAUTEUR	plate, normale, étendue
	DÉBIT	lent, normal, rapide
TIMBRE	TIMBRE	clair, voilé

TABLEAU 1.3 – Taxonomie utilisée pour la qualification des voix.

La taxonomie retenue est présentée dans le **Tableau 1.3** comprend 14 classes (par exemple : sexe, âge, émotion, qualité vocale) et 68 labels (par exemple, pour la qualité vocale : soufflée, craquée, rauque). Par soucis de clarté, les termes "classe" et "labels" sont utilisés par analogie à la classification multi-classe et multi-label : une classe réfère à un ensemble contenant de multiples instances possibles (par exemple : l'émotion inclue les instances : colère, joie, peur tristesse), et un label réfère à une instance particulière (par exemple : colère, joie, peur, tristesse sont des labels). La représentation multi-label d'un enregistrement, proposée et illustrée sur la **Figure 1.6**, est encodée sous la forme d'un vecteur binaire dont on fait l'hypothèse qu'elle représente sa *signature vocale*. Cette taxonomie a été utilisée pour annoter manuellement une sélection de 4 000 enregistrements à partir des 20 000 enregistrements de la version française du MM-RPG Mass Effect 3. Le jeu comprend 54 acteurs interprétant 500 rôles avec un maximum de 10 enregistrements utilisés pour chaque rôle. Un guide d'annotation a été rédigé pour décrire chacune des classes et labels utilisés, chaque label étant accompagné d'échantillons sonores jugés représentatifs et issus de locuteurs différents, et une interface en ligne a été implémentée sous la forme d'un

formulaire en ligne pour permettre de déployer une plate-forme d'annotation simple et rapide. L'annotation de la base a été réalisée par un annotateur non-expert, préalablement entraîné par deux annotateurs experts ¹⁰ à partir du guide d'annotation et des échantillons d'exemples. Des campagnes d'annotations pilotes ont été menées sur de petits ensembles d'enregistrements (environ 50 à 100) tirés au hasard jusqu'à ce que l'annotateur non-expert présente un accord satisfaisant avec les annotations produites par les annotateurs experts. L'accord inter-annotateur final obtenu présente un score alpha de Krippendorff's moyen de $\alpha = 0.52$ (Hayes and Krippendorff, 2007) ce qui représente un accord satisfaisant, particulièrement en regard de la diversité des labels considérés pour l'annotation. La base de 20 000 enregistrements a ensuite été complètement annotée par l'annotateur formé.

¹⁰ *Moi-même et un directeur artistique.*

Apprentissage de la similarité vocale : acoustique vs. perception

Cette section présente les détails des modèles acoustiques et des mesures de similarité utilisées pour mesurer la similarité de voix : la première mesure la similarité vocale directement dans l'espace acoustique, la deuxième mesure la similarité vocale dans un espace perceptif par l'intermédiaire des représentations abstraites définies dans la taxonomie décrite précédemment.

Modélisation de l'espace acoustique : modèle du Monde et supervecteur GMM

Le Modèle du Monde (UBM) est utilisé pour modéliser la distribution de l'intégralité de l'espace acoustique (Reynolds et al., 2000), généralement à partir de mélanges de Gaussiennes à matrice de covariance diagonale (GMM-UBM). La vraisemblance du vecteur \mathbf{o} décrivant les caractéristiques acoustiques du signal de parole sachant les paramètres λ du modèle GMM-UBM s'écrit,

$$p(\mathbf{o}|\lambda) = \sum_{i=1}^M \alpha_i p_i(\mathbf{o}) \quad (1.6)$$

où : M est le nombre de composantes du mélange, $\lambda = \{\alpha_i, \mu_i, \Sigma_i\}_{i \in [1, M]}$ représente les poids, moyennes, et variances de la i -ème Gaussienne, et $p_i(\mathbf{o}) = \mathcal{N}(\mathbf{o}|\mu_i, \Sigma_i)$ correspond à la vraisemblance de la i -ème composante du mélange où \mathcal{N} indique une distribution Gaussienne.

À partir de cette modélisation acoustique, les vecteurs de moyennes $\mu = \{\mu_1, \dots, \mu_M\}$ de l'UBM sont adaptés à chaque enregistrement de parole $\mathbf{o} = [\mathbf{o}_1, \dots, \mathbf{o}_T]$ par l'estimateur du Maximum A Posteriori (MAP) (Reynolds et al., 2000). Finalement, chaque enregistrement est représenté par son vecteur de moyennes adaptées du mélange :

$$\mu^{\text{adapt}} = [\mu_1^{\text{adapt}\top}, \dots, \mu_M^{\text{adapt}\top}]^\top \quad (1.7)$$

où : \top représente l'opérateur de transposition, et μ^{adapt} , défini comme un supervecteur-GMM, est la concaténation de tous les vecteurs de moyenne de moyennes adaptées de l'UBM.

Analyse factorielle : espace de variabilité totale et i -vecteur

Un i -vecteur est la projection linéaire d'une représentation de haute-dimension d'un enregistrement de parole (un supervecteur-GMM est de dimension $M \times D \gg 1$ dans un espace de faible dimension nommé l'espace de variabilité totale (Dehak et al., 2011) ¹¹ :

$$\mu' = \mu + \mathbf{T}\mathbf{x} \quad (1.8)$$

¹¹ *C'est une version simplifiée de l'analyse factorielle alors développée pour séparer la variabilité acoustique liée au locuteur de la variabilité liée au canal de captation/transmission, comme formulée par (Kenny et al., 2008).*

où : μ' est le supervecteur-GMM d'un enregistrement de parole, μ est le supervecteur-GMM correspondant aux vecteurs de moyennes du modèle UBM, \mathbf{T} est la matrice de variabilité totale de dimensions $(DM \times q)$, et \mathbf{x} est un vecteur de dimension q suivant une distribution Gaussienne, définie comme un *i-vecteur*. La matrice de variabilité totale \mathbf{T} est estimée par Espérance-Maximisation (EM), et le *i-vecteur* \mathbf{x} d'un enregistrement de parole est déterminé par l'estimateur du maximum a posteriori (MAP). Une propriété importante des *i-vecteurs* est que leur dimensionalité est très inférieure à celles des super-vecteurs.²

Un ensemble de transformations est appliqué sur le *i-vecteur* pour compenser a posteriori les variabilités entre sessions d'enregistrements et pour normaliser les distributions par locuteur (respectivement par classe), par exemple par Analyse Linéaire Discriminante (LDA, (Dehak et al., 2011)), la Normalisation de la Covariance Intra-Classe (WCCN, (Hatch et al., 2006)), la normalisation de la longueur (LN, (Garcia-Romero and Espy-Wilson, 2011)), la normalisation radiale de facteur propre (EFR, (Bousquet et al., 2012)), ou la Normalisation Sphérique de la Nuisance (SN, (Bousquet et al., 2012; Larcher et al., 2013)).

Mesures de la similarité acoustique

La première contribution a été d'établir des mesures de similarité *acoustique* entre deux voix et d'évaluer leur capacité à rendre compte de leur similarité perceptive. Pour ce faire, j'ai défini des mesures de similarité acoustique à partir des mesures classiquement utilisées en reconnaissance ou en authentification de locuteur. Si le tandem SVM / supervecteur-GMM (Campbell et al., 2006) a longtemps constitué une référence en reconnaissance de locuteur, les avancées réalisées sur la modélisation de l'espace acoustique et l'introduction de la représentation *i-vecteur* ont permis de définir des mesures plus efficaces, par exemple la distance cosinus (Dehak et al., 2009) ou à partir de modèles génératifs (Dehak, 2009) comme par exemple l'analyse discriminante linéaire probabiliste (PLDA) (Prince and Elder, 2007).

La distance cosinus entre deux *i-vecteurs* \mathbf{x}_{src} and \mathbf{x}_{tgt} est définie simplement comme :

$$s(\mathbf{x}_{src}, \mathbf{x}_{tgt}) = \frac{\langle \mathbf{x}_{src}, \mathbf{x}_{tgt} \rangle}{\|\mathbf{x}_{src}\| \|\mathbf{x}_{tgt}\|} \quad (1.9)$$

où $\langle \cdot, \cdot \rangle$ est l'opérateur de produit scalaire. Cette mesure popularisée pour l'authentification de locuteur présente l'avantage d'être une mesure simple et directe de la similarité et ne nécessite aucun apprentissage. Par ailleurs, la distance cosinus fait l'hypothèse que seulement l'angle entre les *i-vecteurs* est porteur d'information, et non leurs amplitudes respectives. La distance cosinus est aujourd'hui toujours pertinente puisqu'elle est utilisée pour mesurer la similarité entre des locuteurs à partir de l'approximation neuronale des *i-vecteurs* par des *x-vecteurs* (Li et al., 2017) ou récemment pour mesurer la similarité entre locuteurs pour le transfert d'identité dans le cadre de la synthèse de la parole à partir du texte (Chen and Garner, 2023).

Pour la PLDA, la mesure de similarité s'exprime sous la forme d'un rapport de vraisemblance (Dehak et al., 2010) :

$$s(\mathbf{x}_{src}, \mathbf{x}_{tgt}) = \frac{p(\mathbf{x}_{src}, \mathbf{x}_{tgt} | \mathcal{H}_1)}{p(\mathbf{x}_{src} | \mathcal{H}_0) p(\mathbf{x}_{tgt} | \mathcal{H}_0)} \quad (1.10)$$

où : \mathcal{H}_1 est l'hypothèse selon laquelle les deux vecteurs sont issus de la même distribution latente (identité du locuteur ou classe), et \mathcal{H}_0 est l'hypothèse selon laquelle les deux vecteurs

². De l'ordre de 40 à 100 pour les *i-vecteurs* contre $M = 1024$ ou 2048 (Gaussiennes) $\times D = 13$ (MFCC) pour les super-vecteurs.

sont issus de distributions latentes distinctes. Une solution analytique pour calculer ce rapport est détaillée dans (Prince, 2012).

Mesure de la similarité perceptive

La deuxième contribution a été d'établir une mesure de similarité *perceptive* par l'intermédiaire de la description sémantique proposée. L'hypothèse principale est que la perception de la similarité entre des voix repose principalement sur des similarités qualitatives, par exemple basée sur des représentations abstraites constituant des dimensions principales de perception de la similarité entre des voix (typiquement, l'âge ou le genre). La mesure perceptive de similarité proposée est calculée à partir d'une classification automatique des attributs de la voix et comparaison de ces attributs entre les enregistrements de voix. Tout d'abord, une couche de classification multi-label par SVM est ajoutée au sommet de la représentation acoustique non-supervisée par supervecteur-GMM ou i-vecteur. Les scores de classification sont convertis en probabilité a posteriori d'appartenance à chaque label. Ces probabilités sont enfin concaténées pour former un vecteur qui représente la *signature vocale* d'un enregistrement. Ce vecteur est utilisé pour définir la mesure de similarité entre deux enregistrements de voix. L'ensemble de l'architecture est illustré sur la Figure 1.6.

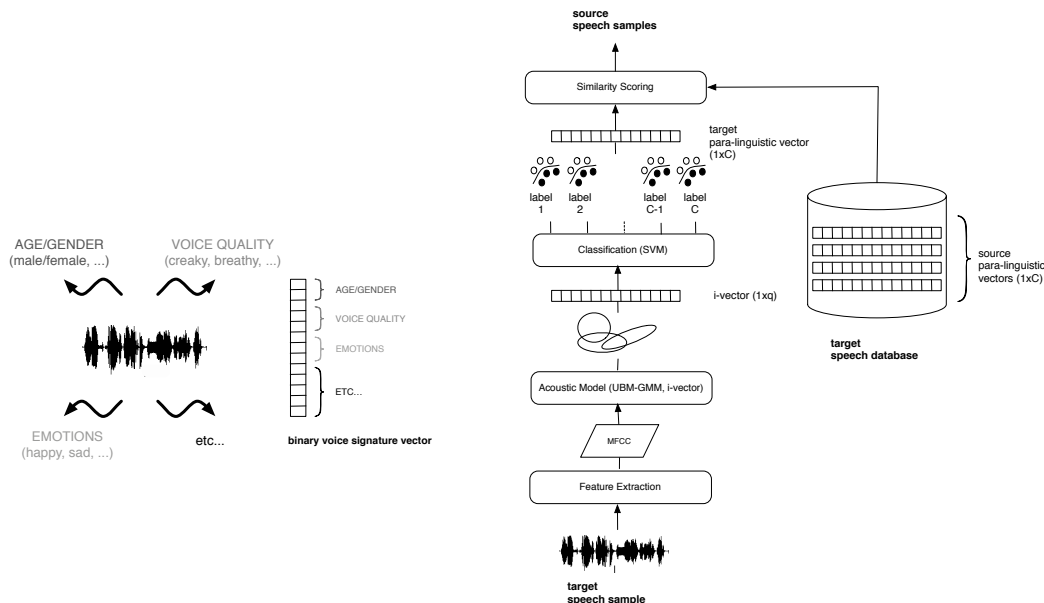


FIGURE 1.6 – Architecture du système de mesure cognitive de la similarité vocale. Figure de gauche : définition du vecteur de *signature vocale* à partir des représentations binaires des attributs de la voix. Figure de droite : en bas, la modélisation acoustique non-supervisée; en haut, la couche de classification multi-label et la mesure de similarité. Source : (Obin and Roebel, 2016)

Pour construire une architecture de classification multi-label, la tâche de classification de labels multiples est transformée systématiquement en multiples tâches de classification binaires (Zhang and Zhou, 2014). Pour ce faire, chaque label de la description symbolique ("la qualité vocale est : rauque", ou "l'émotion est : colère" est transformée en une représentation binaire équivalente (c'est-à-dire : oui/non). Au-delà des aspects algorithmiques, l'idée sous-jacente est que la représentation n'est pas totalitaire et exclusive mais est ouverte à la nuance et au mélange. Ainsi, une voix peut être à la fois triste et joyeuse¹² ou masculine et féminine dans une certaine proportion. Ensuite, un classifieur est entraîné sur chacune des tâches binaires séparément, ce qui résulte en C classifieurs indépendants de type un-contre-tous (Rifkin and Klautau, 2004). Une machine à vecteur

¹² Pour s'en convaincre, il suffit d'écouter n'importe quelle œuvre de Mozart, Beethoven, ou Schubert. Cette idée de mélange émotionnel a depuis fait son chemin en traitement automatique de la parole Zhou et al. (2022).

de support (SVM) est utilisée pour réaliser la classification binaire. Pour chaque label c , la classification d'un vecteur \mathbf{x} (respectivement, supervecteur ou i -vecteur) correspondant à la représentation acoustique d'un enregistrement est obtenue d'après la fonction de décision :

$$f_c(\mathbf{x}) = \sum_{i=1}^N \omega_c^i K(\mathbf{x}, \mathbf{x}_c^i) + b_c \quad (1.11)$$

où : $\Theta_c = \{\omega_c^i, \mathbf{x}_c^i, b_c\}_{i=1}^N$ sont les paramètres de l'hyper-plan de marge maximum déterminé pendant l'entraînement (respectivement les poids, vecteurs de supports, et biais) et $K(., .)$ le noyau du SVM (Cortes and Vapnik, 1995).

Dans un SVM classique, la décision binaire correspondant au vecteur d'observation \mathbf{x} est assignée en fonction de la fonction de décision :

$$\hat{y}_c = \text{sign}(f_c(\mathbf{x})) \quad (1.12)$$

où : $\text{sign}(x) = +1$ pour $x \geq 0$ et -1 sinon, tel quel y_c est égal à 1 quand la décision est positive, et à 0 quand la décision est négative.

Cette fonction de décision est convertie en l'estimation d'une probabilité a posteriori z_c (Wu et al., 2004) conditionnée sur le label c , telle que :

$$z_c = p(y = 1 | \mathbf{x}, \Theta_c), \quad c \in [1, \dots, C] \quad (1.13)$$

Enfin, les probabilités a posteriori de tous les labels sont concaténées pour former un vecteur \mathbf{z} qui représente la *signature vocale* d'un enregistrement de voix :

$$\mathbf{z} = [z_1, \dots, z_C]^\top \quad (1.14)$$

où : z_c est la probabilité a posteriori du c -ème label conditionnellement au vecteur d'observation acoustique \mathbf{x} .

Finalement, la similarité *perceptive* entre un enregistrement de voix source \mathbf{x}_{src} et cible \mathbf{x}_{tgt} est mesurée comme la distance d entre leur signature vocale :

$$s(\mathbf{x}_{\text{src}}, \mathbf{x}_{\text{tgt}}) = d(\mathbf{z}_{\text{src}}, \mathbf{z}_{\text{tgt}}) \quad (1.15)$$

où : \mathbf{z}_{src} et \mathbf{z}_{tgt} représentent la signature vocale des enregistrements source et cible, respectivement. Dans cette recherche, $d(., .)$ est définie comme la divergence de Kullback-Leibler symétrique $KL(., .)$ qui constitue une mesure de divergence naturelle entre des distributions de probabilités. (Cover and Thomas, 1991).

Évaluations expérimentales

Dans cette sous-section, je présente les principaux résultats obtenus à travers cette recherche. Les enregistrements de la version française du MM-RPG Mass Effect 3 ont été utilisés pour l'évaluation. L'expérimentation a été menée en deux temps : tout d'abord une optimisation et une évaluation objective de la modélisation acoustique, de son optimisation pour la reconnaissance de locuteur, et de la classification multi-label. Ensuite, une évaluation subjective relative à la perception de la similarité entre des voix. L'ensemble des détails d'implémentation et des résultats sont décrits dans (Obin and Roebel, 2016).

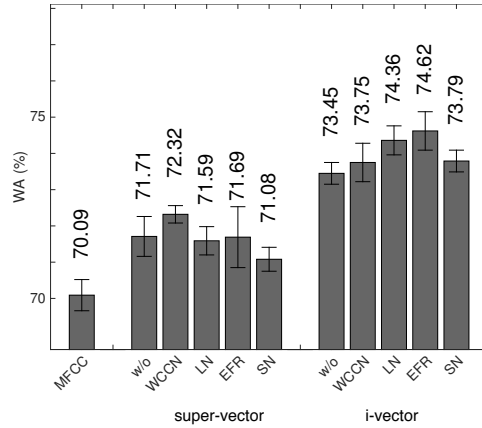


FIGURE 1.7 – Moyenne des scores de classification (BA%) et intervalle de confiance à 95% pour chaque configuration. Source : (Obin and Roebel, 2016)

Évaluation objective

Pour la partie objective de l'évaluation expérimentale, le résultat principal est la performance de classification des attributs vocaux, dans la mesure où elle constitue une condition *sine qua non* de l'architecture proposée pour mesurer une similarité perceptive qui repose sur ces classifications. La métrique de performance utilisée pour mesurer la performance est la précision équilibrée qui correspond à la moyenne non-pondérée de la moyenne arithmétique du rappel positif et négatif calculé pour chaque label binaire.

$$BA = \frac{R_P + R_N}{2} = \frac{1}{2} \left(\frac{T_P}{T_P + F_N} + \frac{T_N}{F_P + T_N} \right) \quad (1.16)$$

où : R_P and R_N sont les rappels positif et négatif, T_P , T_N , F_P , et F_N sont respectivement les nombres de vrais positifs, vrais négatifs, faux positifs, et faux négatifs. Comme moyenne non-pondérée, cette mesure présente l'avantage de ne pas être sensible aux déséquilibres de proportion des labels (Wanger, 1993; Velez et al., 2007). La synthèse des scores de classification est présentée sur la Figure 1.7. La principale observation est que les attributs sont en moyenne reconnus de manière fiable avec une performance de 74.62%. D'autre part, les représentations à partir de i-vecteurs conduisent à des performances significativement meilleures (au sens où l'écart entre les moyennes est statistiquement significatif) que celles obtenues par les représentations classiques par MFCC (70.09%) ou supervecteurs 72.32%.

Les résultats présentés dans le Tableau 1.4 montrent des performances variables selon les labels. En particulier, certains labels sont reconnus de manière fiable à très fiable comme le genre (environ 95%), l'âge (environ 80%), la qualité vocale (environ 77%), la tension vocale (environ 74%), l'effort vocal (environ 80%), et le timbre (environ 73%), la tessiture (environ 83%), la situation (environ 79%) ou les archétypes (environ 80%). En revanche, d'autres labels dont les représentations sont moins consensuelles (pour les émotions) ou moins facilement accessibles par les représentations utilisées (pour l'articulation ou le débit) sont reconnus de manière beaucoup plus mitigée : les émotions (environ 66%), l'articulation (environ 63%), le débit (environ 67%), ou encore le registre de hauteur (environ 66%)

Évaluation subjective

Pour évaluer la capacité des architectures proposées à rendre compte de la similarité perceptive entre des voix, nous avons élaboré une expérience subjective sur la perception de la similarité entre des voix dans deux langues différentes. Pour cela, nous avons utilisé les versions anglais-américain (langue source) et française (langue cible) du MM-

CLASSE	LABEL	MFCC	SUPER-VECTEUR					I-VECTEUR				
			W/O	WCCN	LN	EFR	SN	W/O	WCCN	LN	EFR	SN
Genre	MASCULIN	92.77	92.93	93.74	93.63	93.69	93.35	94.04	94.99	94.92	95.62	94.69
	FÉMININ	92.60	93.52	93.95	93.50	93.47	93.61	94.30	94.96	94.82	95.24	94.46
Âge	ENFANT	96.40	93.81	95.09	94.94	94.88	94.79	95.78	96.40	96.14	96.30	95.50
	ADOLESCENT	87.03	85.46	87.12	82.24	86.93	86.57	90.08	93.48	91.93	87.56	94.62
	JEUNE ADULTE	68.95	72.25	73.44	73.16	72.87	73.20	75.10	76.62	75.95	77.50	75.48
	ADULTE	60.63	64.54	65.61	64.13	63.97	64.84	68.16	69.05	68.49	69.82	68.03
	ÂGÉ	67.63	70.55	71.93	70.70	70.97	71.22	73.75	74.53	74.74	75.46	73.69
	TRÈS ÂGÉ	68.76	67.25	69.41	66.42	67.36	69.14	72.29	73.25	72.59	75.60	72.21
Qualité vocale	SOUFFLÉE	70.03	70.67	72.12	72.44	72.55	71.14	73.61	73.94	74.60	74.86	74.33
	CRAQUÉE	73.63	74.22	75.50	75.66	75.40	75.67	76.14	76.55	78.42	77.71	75.81
	RAUQUE	71.53	73.21	74.44	73.87	74.50	73.65	77.11	76.91	77.53	77.88	77.26
Tension	RELÂCHÉE	72.50	71.89	73.13	73.01	72.26	73.34	73.62	74.36	75.57	76.05	74.82
	NORMALE	64.92	67.65	68.33	67.78	67.58	67.54	68.56	68.91	69.23	69.48	68.60
	TENDUE	62.05	62.53	63.78	62.55	63.36	62.52	63.32	64.05	64.37	64.58	63.81
	PRESSÉE	80.57	82.16	83.91	83.43	83.33	83.59	83.60	83.95	84.19	84.44	84.03
Effort vocal	CHUCHOTÉ/DOUX	80.24	82.01	83.00	82.64	83.33	82.54	83.27	83.31	83.61	83.28	83.73
	NORMAL	68.34	72.50	74.23	72.12	72.08	72.07	73.04	74.47	75.21	76.24	73.87
	FORT/CRIÉ	78.44	77.82	78.35	77.85	77.89	78.63	79.39	80.02	80.95	81.37	79.29
Articulation	HYPO	58.76	58.02	60.26	56.51	56.50	58.41	59.92	61.09	58.78	59.71	59.82
	NORMALE	57.29	59.75	58.62	58.66	58.04	58.16	58.87	59.48	59.48	59.78	59.22
	HYPER	65.11	68.90	69.99	67.99	68.71	68.07	68.34	68.53	69.17	69.20	68.95
Timbre	CLAIR	69.07	70.58	71.91	70.31	70.41	71.00	72.87	73.16	73.45	73.36	73.62
	VOILÉ	69.07	70.46	72.13	70.56	70.45	71.14	72.73	73.18	73.85	73.22	73.35
Registre Fo	EXTRÊME-GRAVE	91.53	90.41	91.17	91.20	91.20	91.04	91.64	92.72	93.04	92.50	92.07
	GRAVE	83.72	85.59	86.25	85.59	86.00	86.14	86.39	86.15	87.10	86.50	86.22
	MOYEN	67.67	70.92	71.87	71.19	71.03	71.28	70.49	71.24	71.50	72.93	71.14
	AIGU	72.39	73.86	74.72	74.49	73.95	74.84	73.85	74.13	74.56	75.23	74.12
	EXTRÊME-AIGU	85.83	86.72	88.14	87.57	87.96	88.21	87.57	87.69	88.43	87.64	88.08
Attitude / Modalité	AFFIRMATION	66.89	69.57	69.40	68.93	69.46	69.66	69.64	69.97	70.05	70.22	69.78
	CONFIRMATION	60.51	59.27	61.42	61.42	61.60	62.99	61.81	64.45	63.58	65.49	61.05
	EXCLAMATION	67.28	68.04	68.99	67.95	68.27	67.95	68.63	68.69	68.83	69.35	68.90
	INTERROGATION	58.08	62.31	62.90	62.14	61.92	62.86	59.80	58.47	61.67	59.07	59.47
	ORDRE	64.95	66.77	68.26	66.66	67.20	65.98	68.71	68.45	69.60	68.70	68.87
Émotion	COLÈRE	61.98	62.38	63.36	62.85	62.69	62.58	63.40	64.99	64.56	65.11	64.44
	EXCITATION	66.18	66.74	67.25	66.79	68.10	67.67	67.04	68.18	68.20	68.22	67.41
	JOIE	53.89	55.81	56.45	55.82	56.71	55.42	58.40	59.24	58.96	60.72	59.61
	NEUTRE	61.42	62.99	64.65	64.05	63.45	64.24	65.75	65.23	65.84	65.69	66.01
	TRISTESSE	60.65	61.99	63.92	63.05	64.05	63.28	62.68	63.78	64.06	64.42	63.15
	PEUR	63.59	63.68	64.98	64.04	64.31	63.78	64.45	63.97	64.93	66.85	65.96
	STRESS	80.22	78.71	80.06	79.52	79.89	79.65	79.59	79.56	80.42	81.12	80.09
	SURPRISE	57.74	58.00	59.09	57.74	60.58	59.40	58.01	60.46	60.58	60.27	60.97
TENDRESSE	62.33	62.42	64.51	63.25	63.19	63.87	64.31	64.22	64.64	64.87	64.87	
Situation	ACTION	83.24	81.76	83.26	83.23	82.97	82.63	82.91	82.83	84.28	82.60	82.79
	DIALOGUE	73.69	75.79	76.00	75.77	75.60	75.81	77.25	77.03	78.45	77.30	77.80
	INFORMATION	69.53	76.22	78.94	77.80	76.90	77.02	78.48	80.22	78.89	81.73	80.07
	MONOLOGUE	64.25	65.45	66.06	65.22	66.34	67.99	69.13	66.57	72.54	73.76	68.20
Archétype	ANNONCEUR	82.63	74.67	81.29	83.72	82.74	79.00	87.99	90.20	89.65	89.60	87.99
	INTELLIGENCE ART.	87.78	89.34	87.62	85.19	88.52	87.29	90.86	91.56	89.96	93.24	91.10
	SOLDAT	68.71	70.37	71.02	69.22	69.23	70.56	71.73	72.31	73.29	72.80	71.53
	BRUTE	73.39	75.10	76.75	76.98	76.36	76.70	78.00	78.94	79.62	79.74	78.59
	COMMANDANT	64.96	63.99	66.47	66.00	65.31	66.30	68.49	69.58	70.26	70.58	70.00
	HÉROS	68.58	70.67	75.54	70.90	70.90	73.36	77.52	76.40	76.52	78.11	76.53
	SOLDAT DÉBUTANT	69.23	72.54	73.97	74.26	73.78	73.78	76.87	77.36	79.31	78.98	76.15
	SOLDAT VÉTÉRAN	70.76	70.89	73.15	71.32	71.56	72.08	72.59	73.81	74.14	74.51	73.40
Total		70.09	71.71	72.32	71.59	71.69	71.08	73.45	73.75	74.36	74.62	73.79

TABLEAU 1.4 – Scores de classification (BA%) en fonction des labels pour des représentations supervecteur (à gauche) et i-vecteur (à droite).

PRG Mass Effect 3 pour sélectionner 50 échantillons représentatifs et répartis de manière équilibrée entre les locuteurs et leurs genres (50% d'échantillons de voix masculines, 50% d'échantillons de voix féminines) et chacun d'une durée approximative de 5 secondes. Le protocole expérimental est le suivant : pour chaque échantillon *source* présenté dans un ordre aléatoire, il est présenté les 3 échantillons dans la langue *cible* les plus similaires recommandés par :

1. la similarité *acoustique* mesurée à partir du meilleur système de reconnaissance de locuteur obtenu au cours des expérimentations objectives (i-vector + sphNorm + PLDA);
2. la similarité *perceptive* mesurée à partir du meilleur système de classification multi-label obtenu au cours des expérimentations objectives (i-vector + EFR (noNorm) + SVM);

Le sujet doit alors juger la similarité des échantillons *cibles* proposés par rapport à l'échantillon *source* sur une échelle à 5 degrés : très dissimilaire (-2), plutôt dissimilaire (-1), peu similaire (0), plutôt similaire (+1), très similaire (+2). L'expérience a été réalisée dans une chambre d'écoute avec 30 participants natifs en français (20 hommes et 10 femmes, entre 20 et 35 ans.

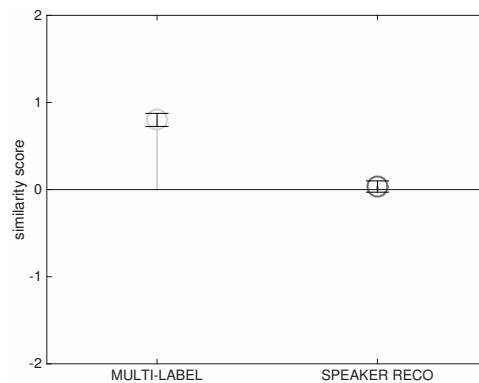


FIGURE 1.8 – Score de similarité moyen et intervalle de confiance à 95% pour les deux architectures *perceptive* (à gauche) et *acoustique* (à droite). L'échelle de similarité est : très dissimilaire (-2), plutôt dissimilaire (-1), peu similaire (0), plutôt similaire (+1), très similaire (+2). Source : (Obin and Roebel, 2016)

La synthèse des jugements de similarité obtenus lors de cette expérience est présentée sur la Figure 1.8. Les échantillons recommandés par l'architecture *perceptive* sont jugés en moyenne comme plutôt similaires à l'échantillon source, alors que les échantillons recommandés par l'architecture *acoustique* ne sont jugés en moyenne que faiblement similaires. Par ailleurs, les jugements entre les deux architectures présentent une différence statistiquement significative. Cette expérience subjective confirme l'hypothèse selon laquelle la mesure de la similarité fondée sur des représentations symboliques intermédiaires des attributs de la voix rend mieux compte de la perception de la similarité entre des voix. En outre, l'expérience objective démontre la faisabilité opérationnelle de la mesure de la similarité cognitive entre des voix. Le passage par des représentations symboliques intermédiaires permet de simplifier un problème complexe en un sous-problème plus simple, et de supprimer beaucoup de variabilité acoustique non pertinente, ce qui simplifie d'autant la tâche de similarité et la quantité de données nécessaire pour l'apprentissage de ces similarités. Cette stratégie actualise finalement le principe de diviser pour mieux régner.

Pour conclure ce chapitre, je dresse une brève synthèse des principales contributions réalisées au cours de ces recherches, et ouvre des pistes de réflexion à partir des résultats obtenus.

Contributions

C2. Définition d'une **taxonomie pour qualifier les attributs d'une voix**, sur la base des descriptions existantes en physiologie, psychologie, et linguistique.

C3. Implémentation d'une architecture de **mesure de la similarité perceptive entre des voix par apprentissage machine**, fondée sur la similarité entre les attributs d'une voix (issus de la taxonomie proposée en C2.)

Ces années de recherche ont été riches en contributions et en enseignements. Elles ont été pour moi l'occasion de fonder et d'explorer un champ de recherche sur un objet nouveau en traitement automatique de la parole, à savoir l'analyse, la modélisation et la caractérisation de la voix actée naturellement expressive. La nature même de cet objet le rend difficilement accessible à l'analyse par les modèles de signaux de parole traditionnels. En l'occurrence, les voix actées n'ont pas grand chose à voir avec les voix rencontrées en parole ordinaire, de tous les jours, ou spontanées. Elles évoluent régulièrement dans des registres extrêmes, s'appuient sur un grand nombre de modes de phonation ou d'articulation, et font intervenir des mécanismes complexes (non-linéarité, fonctionnement chaotique, etc...). Un facteur clef de ces travaux a été la possibilité d'accéder à des données généralement inaccessibles à la recherche et l'observation de voix mettant en jeu une diversité et une complexité de mécanismes qui ne sont généralement pas observables sur des bases de données de recherche, encore moins avec cette qualité et cette quantité.

Ces recherches ont apporté des premiers éléments de réponse aux 3 questions de recherche formulées dans l'introduction.

Q1. Quelles représentations ? Les représentations du signal de voix ne sont pas seulement de nature acoustique, mais également de nature symbolique Le premier élément de réponse porte sur la définition d'une ontologie de la voix permettant de qualifier verbalement une voix pour en décrire les attributs principaux. Une première tentative de taxonomie a été proposée, principalement sur la base des descriptions existantes dans les domaines de la physiologie, de la linguistique, et de la psychologie. Cette représentation symbolique de la voix s'est révélée un moyen efficace pour mesurer la similarité perceptive entre des voix — autrement dit, comme une mise en lumière des attributs principaux rentrant en compte dans la perception d'une voix. Nous avons également démontré qu'il est possible d'associer efficacement ces représentations symboliques avec de l'apprentissage machine pour caractériser des voix actées et naturellement expressives. Par ailleurs, la contribution à la définition d'une taxonomie partagée est importante aussi bien dans la communauté de la parole que pour les professionnels de la voix. Néanmoins, cette contribution reste largement préliminaire et soulève un certain nombre de limitations et nous n'en sommes qu'au début de l'étude des voix actées. En particulier, un nombre non négligeable des attributs utilisés par les professionnels pour qualifier une voix ne sont pas ou peu documentés dans la littérature scientifique, ce qui rend leur étude et leur standardisation difficile. Les travaux menés dans le projet ANR TheVoice (2017-2021) sur la question suggère que chaque application possède un vocabulaire propre

qui varie en fonction des considérations expressive et esthétique, des médias et des canaux de communication utilisés (Quillot et al., 2020). Cette première recherche a été fondatrice d'une démarche plus générale visant à intégrer les représentations symboliques de la voix avec des méthodes de traitement du signal et d'apprentissage machine selon le paradigme : description (symbolique), représentation (acoustique), apprentissage (modélisation générative), comme nous le verrons de manière approfondie dans le Chapitre 3 sur la manipulation des attributs de la voix. Cette direction nécessite par ailleurs de créer des passerelles avec des disciplines des sciences humaines et expérimentales, comme la psycho-acoustique, la sociologie, ou la psychologie, pour à terme réussir à mieux formaliser ces descriptions. En retour, les modèles signaux et l'apprentissage machine peuvent permettre de définir de nouvelles expérimentations dans ces disciplines et d'accéder à de nouvelles connaissances notamment sur la psychologie et le comportement humain.

Les travaux présentés dans ce premier chapitre ont constitué les premières tentatives d'analyse, de caractérisation, de représentation, et de modélisation de voix expressives. Ils présentent l'avantage d'ouvrir la boîte de Pandore de la parole expressive actée, en particulier en considérant comme objet d'étude les fonctions expressives, impressives, et esthétiques de la parole humaine — comme de leur cognition et leur réception ancrées tant biologiquement que culturellement. Le fruit de ces recherches ne sont que très préliminaires, car les voix expressives et actées, par leur diversité comme par leurs registres extrêmes, demeurent encore largement inexplorées et réfractaires aux modèles de signaux de parole actuels. Depuis, cette boîte de Pandore a continué de s'ouvrir, en particulier à travers les thèses menées au Laboratoire d'Informatique d'Avignon par Adrien Gresse (Gresse, 2020) et Matthias Quillot (Quillot, 2020). Le doublage de parole automatique d'une langue à une autre est désormais ¹³ un axe de recherche rendu possible par l'apprentissage par réseaux de neurones profonds (Federico et al., 2020) possiblement avec un redessinnage des mouvements des lèvres pour assurer la synchronisation labiale (Prajwal et al., 2020). La caractérisation et la modélisation de la palette expressive de la voix actée a été l'objet principal du projet ANR TheVoice (2017-2021) que j'ai coordonné, et la quête d'ontologies de la voix humaine se poursuit à travers le projet ANR EVA (Explicit Voice Attributes, soumis en 2023 et porté par Olivier Le Blouch d'Orange), poursuit cette recherche sous l'angle des réseaux de neurones et de l'apprentissage de représentations. La voix humaine et son expressivité présente de multiples facettes, comme autant de masques ajustables et combinables, à l'interface de la cognition, de la sociologie, de la linguistique. Le traitement du signal et de l'information, et l'apprentissage machine fournissent des moyens pour tenter d'analyser, de modéliser, et de simuler l'expressivité humaines dans toute sa diversité et sa complexité.

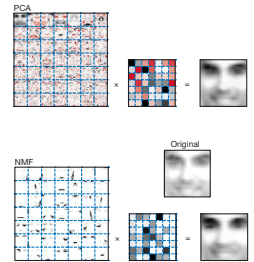
¹³ Le doublage n'est cependant pas une simple traduction littérale d'une langue à une autre : comme pour la traduction littéraire, c'est une adaptation stylistique à une langue et pour une culture. La forme des masques et leur fonction varient d'une langue à une autre, en fonction des cultures.

2.1 INTRODUCTION

Ce deuxième chapitre est consacré à la formalisation de modèles structurés pour la modélisation statistique de signaux de parole pour des tâches de perception de la parole. La perception humaine consiste à localiser, séparer, et interpréter les sources de son environnement pour pouvoir l’appréhender et interagir avec. Les capacités de perception auditive de l’être humain sont avant tout déterminées par sa physiologie : un humain écoute le monde avec deux oreilles — et un corps —, son écoute est donc binaurale. Cette capacité est essentiellement psycho-acoustique, c’est-à-dire liée à la physique du système auditif de l’être humain. Nous percevons notre environnement sonore en trois dimensions (azimuth, élévation, et distance en coordonnées sphériques) (McKinley and Ericson, 2005; Makous and Middlebrooks, 1990). La séparation de sources sonores ressort d’un processus cognitif : l’effet “cocktail party” (Cherry, 1953). Nous sommes en effet capables, dans un environnement complexe, bruyant, et réverbéré, de focaliser notre attention pour filtrer sélectivement les sources sonores d’intérêt — par exemple, une conversation particulière. L’analyse de scène auditive assistée par ordinateur — ou CASA, en anglais — consiste à simuler artificiellement par la machine les facultés de capacité de perception auditive de l’être humain, une machine d’écoute. Dans ce contexte, la factorisation en matrices non-négatives (Paatero, 1997; Lee and Seung, 1999) a longtemps constitué un paradigme de recherche pour la modélisation de signaux sonores sur des tâches de perception, et en particulier de séparation de sources sonores (Benaroya et al., 2003; Vincent et al., 2007; Smaragdis, 2007; Dessein et al., 2010; Ozerov and Fevotte, 2010). Dans ce chapitre, je présente mes travaux réalisés entre 2015 et 2018 sur la factorisation en matrices (respectivement, en tenseurs) non-négatives (NMF ou NTF) comme modèle d’écoute appliquée sur des tâches de perception de signaux de parole, comme la séparation et la localisation de sources sonores. En particulier, nous avons essayé d’introduire et de formaliser explicitement à l’intérieur de modèles d’apprentissage des contraintes et des conditions explicitant des connaissances a priori sur la structure de ces signaux. En d’autres mots, il s’agit d’intégrer formellement des modèles de signaux ou des hypothèses physiques directement dans les algorithmes d’apprentissage pour restreindre l’espace des solutions à un sous-ensemble de solutions physiquement réalistes.

La factorisation en matrices non-négatives (NMF) est un algorithme de réduction de dimensionalité (Paatero, 1997; Lee and Seung, 1999) qui a été popularisé pour ses applications en traitement de l’image, de texte, et au son. Par définition, la factorisation en matrices non-négatives d’un signal consiste en sa projection linéaire sur un hyper-plan défini par les bases de la NMF avec comme contrainte spécifique que les termes des bases et de la combinaison linéaire sont non-négatifs, c’est-à-dire à valeurs positives ou nulles. Cette factorisation se distingue des autres algorithmes de réduction de dimensionalité par la non-négativité de la décomposition, interprétable physiquement comme une combinaison de parties définies non-négativement¹. La NMF propose un modèle de perception fondé sur des bases neuro-physiologiques (Lee and Seung, 1999) : d’une part, les connexions synaptiques sont nécessairement excitatrices ou inhibitrices, et l’activation des neurones positive ou nulle; d’autre part, l’asymétrisation du cerveau et sa spécialisation au cours de l’évolution humaine ont entraîné une parcimonie de l’activité neuronale.²

¹ Par opposition à l’analyse en composantes principales (PCA), la contrainte de non-négativité empêche l’apparition d’“énergie noire”.



Source : (Lee and Seung, 1999)

² Cet argument se retrouve également dans les réseaux de neurones, et notamment sur la formalisation de la parcimonie de l’activité des neurones ou des connexions synaptiques (Srivastava et al., 2014).

L'application aux signaux sonores, sous la forme d'un spectrogramme d'amplitude ou de puissance, a été naturelle sous l'hypothèse d'additivité des signaux sonores ³, avec des applications comme modèle d'écoute pour la perception de signaux sonores comme la séparation de sources sonores (Vincent et al., 2007) ou la transcription de musique (Dessein et al., 2010, 2013). Tout d'abord, la NMF présente un certain nombre de bonnes propriétés en termes de convergence (Badeau et al., 2011). Des formulations spécifiques ont permis de modéliser des mélanges de signaux sonores linéaires instantanés (Benaroya et al., 2003) aux mélanges convolutionnels (Ozerov and Fevotte, 2010; Mysore and Smaragdis, 2012). La formulation de variantes fondées sur des fonctions de coût dérivées des familles de fonctions des Beta-divergences (qui incluent la divergence de Kullback-Leibler et la divergence d'Itakura-Saito) présentent des propriétés avantageuses pour les signaux sonores de stabilité ou d'invariance à la mise à l'échelle (Févotte and Idier, 2011). Enfin, l'intégration de modèles signaux spécifiques, comme le modèle source/filtre de signaux sonores (Durrieu et al., 2009). Dans un premier temps, je présente l'intégration d'un modèle signal source/filtre (SF) pour la factorisation d'un signal de parole (SF-NMF) avec formalisation de contraintes physiques (Bouvier et al., 2016); et dans un second temps, la formalisation d'un modèle de signal binaural et son intégration à une factorisation NMF multi-canal (Benaroya et al., 2018). Ces implémentations sont appliquées sur des tâches variées de traitement automatique de la parole, comme la séparation parole et bruit (Sun and Mysore, 2013; Virtanen et al., 2013) ou la séparation de sources multi-canal ou la localisation de sources sonores (Ozerov and Fevotte, 2010; Benaroya et al., 2018).

³ Le sonore supporte mieux cette hypothèse que la vision, en négligeant cependant les interactions de phases constructives ou destructives entre les sources sonores.

Projets associés à ce chapitre (par ordre chronologique)
Projet EMERGENCE ROUTE (Sorbonne Université, 2015-2016)
Encadrements associés à ce chapitre (par ordre chronologique de fin)
Laurent Benaroya (2015-2016), post-doc, projet EMERGENCE ROUTE Damien Bouvier (M2 ATIAM, 2015), Guillaume Doras (M2 ATIAM, 2016), Wilson Raugel (M2 ISI, 2016), Félix Rohrlach (M1 Acoustique, 2017)
Publications associées à ce chapitre (par ordre chronologique)
W. Phokhinnan, N. Obin, S. Argentieri (2023). Binaural Sound Localization in Noisy Environments Using Frequency-Based Audio Vision Transformer (FAViT), Interspeech, Dublin, Ireland, 2023. L. Benaroya, N. Obin, M. Liuni, A. Roebel, W. Raugel, S. Argentieri (2018). Binaural Localization of Multiple Sound Sources by Non-Negative Tensor Factorization. IEEE/ACM Transactions on Audio, Speech and Language Processing, vol. 26, no 6, p. 1072-1082, February 2018. D. Bouvier, N.Obin, M. Liuni, A. Roebel (2016). A Source/Filter Model with Adaptive Constraints for NMF-based Speech Separation International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Shanghai, China, 2016.

2.2 MODÈLE NMF SOURCE/FILTRE AVEC CONTRAINTES INSPIRÉES DE LA PHYSIQUE

Dans cette première partie, je présente les recherches menées avec Damien Bouvier, Laurent Benaroya et Axel Roebel sur l'implémentation d'algorithmes de NMF avec des contraintes physiques et leur application sur une tâche de séparation de la parole (Bouvier et al., 2016). Dans cette tâche, nous faisons l'hypothèse que le signal audio observé est la somme d'un signal de parole d'intérêt et d'un environnement sonore considéré comme du bruit (par exemple, des sons environnementaux ou alors de la parole de fond). Dans la suite de cette section, nous faisons l'hypothèse de mélanges linéaires instantanés. La tâche

de séparation de parole est au moins partiellement supervisée, c'est-à-dire que nous faisons également l'hypothèse qu'il existe des exemples de la source de parole d'intérêt à partir desquels les bases d'une NMF peuvent être apprises (Virtanen and Klapuri, 2006; Durrieu et al., 2009; Mysore and Smaragdis, 2012; Le Magoarou et al., 2014). Dans certains cas plus limités, on peut faire la même hypothèse pour les bruits de l'environnement, comme dans (Sun and Mysore, 2013; Virtanen et al., 2013). Plus récemment, des implémentations hybrides nommées deep-NMF (Le Roux et al., 2015) ont enfin été proposées pour apprendre les bases d'une NMF multi-couches par rétro-propagation de manière similaire à un réseau de neurones profond.

NMF et modèle source/filtre

Principe de la NMF

On considère une matrice d'observation $\mathbf{X} \in \mathbb{R}^{+F \times T}$, par exemple la matrice correspondant au module du spectrogramme d'un signal audio $x(t)$, où F est le nombre de points fréquentiels de la transformée de Fourier et T le nombre de trames. La factorisation en matrices non-négatives (NMF) de \mathbf{X} consiste à approximer cette matrice comme un produit de deux matrices \mathbf{W} et \mathbf{H} :

$$\mathbf{X} \simeq \mathbf{V} = \mathbf{W}\mathbf{H} \quad (2.1)$$

où : $\mathbf{W} \in \mathbb{R}^{+F \times S}$ et $\mathbf{H} \in \mathbb{R}^{+S \times T}$ sont également des matrices à coefficients non-négatifs et S représente le nombre de bases utilisées pour la factorisation. \mathbf{W} représente la matrice de bases du dictionnaire et \mathbf{H} la matrice d'activation des bases du dictionnaire. En d'autres mots, $\mathbf{W}\mathbf{H}$ est la projection linéaire de \mathbf{V} sur l'hyper-plan défini par les bases \mathbf{W} avec comme poids \mathbf{H} . Suite à la décomposition, les signaux des sources séparées sont obtenus par filtrage de Wiener à partir des spectrogrammes estimés pour chacune des sources (Benaroya et al., 2006).

Le problème de la NMF consiste à déterminer les paramètres (\mathbf{W}, \mathbf{H}) qui minimisent une certaine fonction de coût $C(\mathbf{X}|\mathbf{V})$. La Beta-divergence est une famille de divergence qui constitue une fonction de coût usuelle en NMF — définie pour $\beta \in \mathbb{R} \setminus \{0, 1\}$:

$$c = d_\beta(X_{f,n}|V_{f,n}) = \frac{1}{\beta(\beta-1)} \left(X_{f,n}^\beta + (\beta-1)V_{f,n}^\beta - \beta X_{f,n} V_{f,n}^{\beta-1} \right) \quad (2.2)$$

La Beta-divergence présente de bonnes propriétés pour son application à des signaux audio, en particulier la stabilité à la mise à l'échelle (Févotte et al., 2009) :

$$d_\beta(\lambda x|\lambda y) = \lambda^\beta d_\beta(x|y) \quad (2.3)$$

Les divergences de Kullback-Leibler (Virtanen, 2007) et d'Itakura-Saito (Févotte et al., 2009) constituent des cas particuliers limites de la Beta-divergence, respectivement pour $\beta = 1$ et $\beta = 0$.

La solution du problème de la NMF est obtenue efficacement par l'algorithme de la mise à jour multiplicative (MU) dérivé de la descente de gradient (Févotte and Idier, 2011). A la i -ème itération, la mise-à-jour de \mathbf{W} et \mathbf{H} est obtenue par application de la mise-à-jour multiplicative suivante :

$$\Theta^{(i+1)} \longleftarrow \Theta^{(i)} \otimes \frac{\nabla_{\Theta^{(i)}}^- c}{\nabla_{\Theta^{(i)}}^+ c} \quad (2.4)$$

où : Θ représente respectivement \mathbf{W} ou \mathbf{H} , $\nabla_{\Theta^{(i)}}^+$ et $\nabla_{\Theta^{(i)}}^-$ représentent les parties positive et négative du gradient de la fonction de coût \mathcal{C} relativement à $\Theta^{(i)}$, \otimes représente le produit d’Hadamard, et la division est terme-à-terme.

NMF avec modèle source/filtre

La NMF avec modèle source/filtre (Durrieu et al., 2009) (SF-NMF) permet de représenter explicitement la structure des signaux de parole au sein de la décomposition NMF. La décomposition SF-NMF d’un signal de parole \mathbf{V}^S s’écrit :

$$\begin{aligned} \mathbf{V}^S &= \mathbf{V}_{\text{ex}} \otimes \mathbf{V}_{\Phi} \\ &\simeq \underbrace{(\mathbf{W}_{\text{ex}} \mathbf{H}_{\text{ex}})}_{\text{excitation}} \otimes \underbrace{(\widehat{\mathbf{W}}^{\Phi} \mathbf{H}_{\Phi})}_{\text{filter}} \end{aligned} \quad (2.5)$$

où : \mathbf{V}_{ex} et \mathbf{V}_{Φ} sont respectivement le spectrogramme d’amplitude de la source d’excitation et du filtre, \mathbf{W}_{ex} et \mathbf{H}_{ex} sont les termes de la décomposition NMF de la matrice d’excitation \mathbf{V}_{ex} (où \mathbf{W}_{ex} est un dictionnaire fixé à l’avance, comprenant un ensemble de bases périodiques et bruitées), et $\widehat{\mathbf{W}}^{\Phi}$ et \mathbf{H}_{Φ} sont les termes de la décomposition de la matrice de filtre \mathbf{V}_{Φ} . Pour garantir la lissitude des filtres $\widehat{\mathbf{W}}^{\Phi}$, ce terme est à son tour décomposé comme un produit $\mathbf{W}_{\Phi} \mathbf{U}_{\Phi}$ où \mathbf{W}_{Φ} est un dictionnaire fixé de filtres élémentaires lisses (ici, des fenêtres de Hanning) et \mathbf{U}_{Φ} est la matrice de combinaison linéaire des filtres élémentaires pour approximer le filtre de parole $\widehat{\mathbf{W}}^{\Phi}$. L’architecture du modèle SF-NMF est illustrée en Figure 2.1.

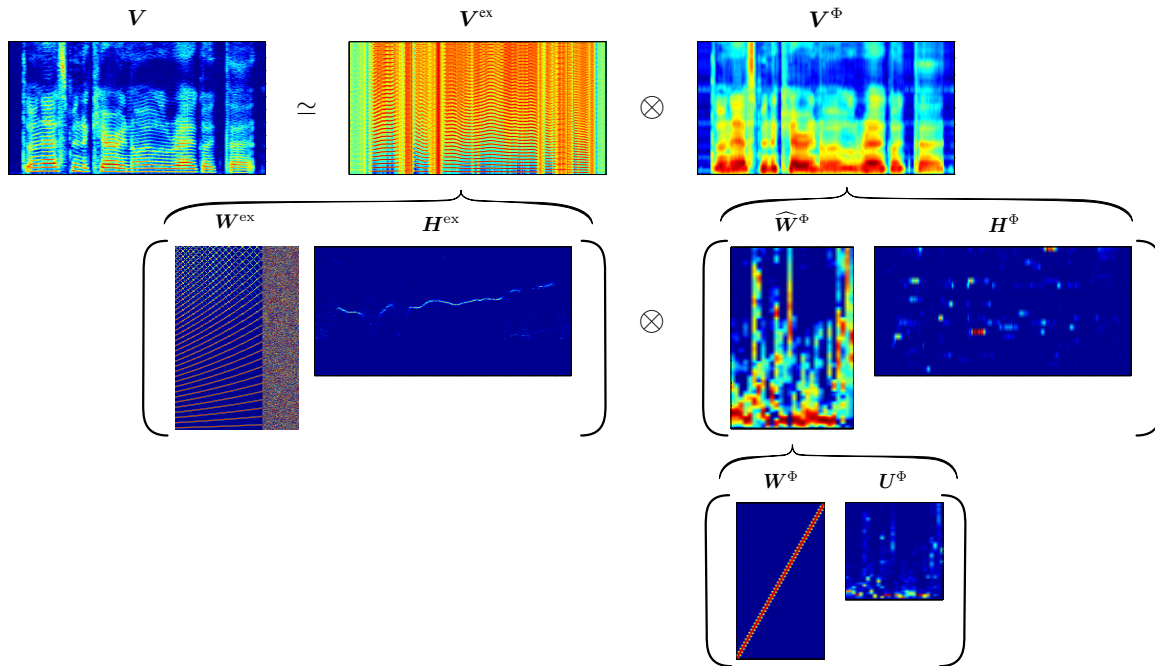


FIGURE 2.1 – Illustration de la décomposition source/filtre d’un spectrogramme d’amplitude telle qu’exprimée par l’Équation (2.5). Source : (Bouvier et al., 2016)

Dans le cadre de la séparation de la parole, nous faisons l’hypothèse que le signal observé \mathbf{V} est un mélange linéaire instantané d’un signal de parole \mathbf{V}^S et d’un signal de bruit de fond \mathbf{V}^N . \mathbf{V} est alors approximé par décomposition NMF en $\widetilde{\mathbf{V}}$ de la manière suivante :

$$\mathbf{V} \simeq \widetilde{\mathbf{V}} = (\mathbf{W}_{\text{ex}} \mathbf{H}_{\text{ex}}) \otimes (\mathbf{W}_{\Phi} \widehat{\mathbf{U}}^{\Phi} \mathbf{H}_{\Phi}) + \mathbf{W}^N \mathbf{H}^N \quad (2.6)$$

Le terme correspondant au signal de bruit de fond \mathbf{V}^N est exprimé comme une décomposition NMF de termes \mathbf{W}^N et \mathbf{H}^N . En supplément de (Durrieu et al., 2009), nous avons commencé par rendre l'apprentissage des termes de la SF-NMF semi-supervisé. En particulier, nous avons proposé d'apprendre les matrices de filtre à partir de signaux propres de parole. Pour ce faire, la matrice de filtre \mathbf{V}_Φ est d'abord estimée à partir d'un algorithme d'estimation de l'enveloppe spectrale (Villavicencio et al., 2006) à partir des signaux propres de parole. Cette matrice est alors approximée suivant la décomposition NMF de la matrice de filtre :

$$\mathbf{V}_\Phi \simeq \mathbf{W}_\Phi \mathbf{U}_\Phi \mathbf{H}_\Phi \quad (2.7)$$

De plus, nous avons par ailleurs utilisé un alignement phonétique (Lanchantin et al., 2008) pour apprendre des matrices de filtre spécifiques pour chaque phonème séparément, par exemple une base pour chaque phonème.

Modèle source/filtre avec contraintes inspirées de la physique

L'idée principale a été d'introduire des contraintes physiques pour la séparation de parole par SF-NMF pour spécifier les solutions à des solutions physiquement acceptables. Pour ce faire, nous avons ajouté des termes de régularisation pour pénaliser les solutions qui ne respectent pas un modèle de signal de parole. Ce faisant, la fonction de coût \mathcal{C} est modifiée par l'addition d'une contrainte de régularisation \mathcal{P} :

$$\mathcal{C} = D_{\text{IS}} \left(\mathbf{V} | \tilde{\mathbf{V}} \right) + \lambda \mathcal{P}(\tilde{\mathbf{V}}) \quad (2.8)$$

où : λ est une valeur positive déterminant le poids de la contrainte. La règle de mise-à-jour pour la i -ème itération s'écrit alors avec le terme de régularisation :

$$\Theta^{(i+1)} \longleftarrow \Theta^{(i)} \otimes \frac{\nabla_{\Theta^{(i)}}^- D_{\text{IS}} + \lambda \nabla_{\Theta^{(i)}}^- \mathcal{P}}{\nabla_{\Theta^{(i)}}^+ D_{\text{IS}} + \lambda \nabla_{\Theta^{(i)}}^+ \mathcal{P}} \quad (2.9)$$

Pour nous permettre d'exploiter les contraintes de régularisation dans le cadre du modèle SF-NMF, nous avons spécifié les contraintes usuelles pour le modèle SF-NMF, proposé une contrainte de cohérence source/filtre, et formulé un algorithme d'adaptation dynamique des pondérations de ces contraintes au cours de la séparation.

Contraintes usuelles

Tout d'abord, nous avons utilisé les contraintes existantes en NMF (Bertin, 2009) que nous avons formulées spécifiquement pour le modèle SF-NMF. La contrainte de parcimonie (Joder et al., 2013) favorise l'activation d'une seule base d'excitation de la source glottique et d'une seule base de filtre résonateur à chaque instant; la contrainte de décorrélation normalisée (Li et al., 2001) pénalise l'activation simultanée des bases d'activation, afin de spécialiser les bases et de réduire la redondance entre ces bases; et la contrainte de lissitude temporelle (Virtanen, 2007) favorise la continuité des activations de sources et de filtres au cours du temps. La formulation de ces contraintes dans le cadre du modèle SF-NMF vise principalement à spécifier ce modèle par des contraintes inspirées de la physique, comme le modèle source/filtre de la parole ou la contrainte de continuité temporelle des paramètres de ce modèle.

Contrainte de cohérence source/filtre

Dans un second temps, nous avons proposé la formulation d'une contrainte de cohérence entre les termes de source et filtre du modèle source/filtre de la parole. Cette contrainte repose sur l'idée que la réalisation de chaque phonème correspond à l'actualisation d'une

source d'excitation et d'un filtre particulier. Ainsi, les filtres vocaux correspondants à des phonèmes voisés seront systématiquement associés à une source d'excitation périodique; et les filtres vocaux correspondant à des phonèmes non-voisés seront systématiquement associés à une source d'excitation bruitées. La contrainte proposée est définie de manière à empêcher des associations non-réalistes entre des excitations et des filtres comme illustré sur la [Figure 2.2](#). Cette contrainte, inspirée de la contrainte de décorrélation normalisée, s'écrit :

$$\mathcal{P}_\phi = \sum_{\substack{k \in \text{periodic} \\ l \in \text{unvoiced}}} \frac{[\mathbf{H}_{\text{ex}k} \mathbf{H}_\phi^T]_{kl}}{\|\mathbf{H}_{\text{ex}k}\|_{\ell_2} \|\mathbf{H}_\phi l\|_{\ell_2}} + \sum_{\substack{k \in \text{noise} \\ l \in \text{voiced}}} \frac{[\mathbf{H}_{\text{ex}k} \mathbf{H}_\phi^T]_{kl}}{\|\mathbf{H}_{\text{ex}k}\|_{\ell_2} \|\mathbf{H}_\phi l\|_{\ell_2}} \quad (2.10)$$

Le terme de gauche dans la somme est une mesure de la corrélation entre les bases d'excitation périodique et les bases de filtre correspondant à des phonèmes non-voisés, normalisée par leur puissance; le terme de droite dans la somme est une mesure de la corrélation entre les bases d'excitation bruitée et les filtres correspondant à des phonèmes voisés, normalisée par leur puissance.

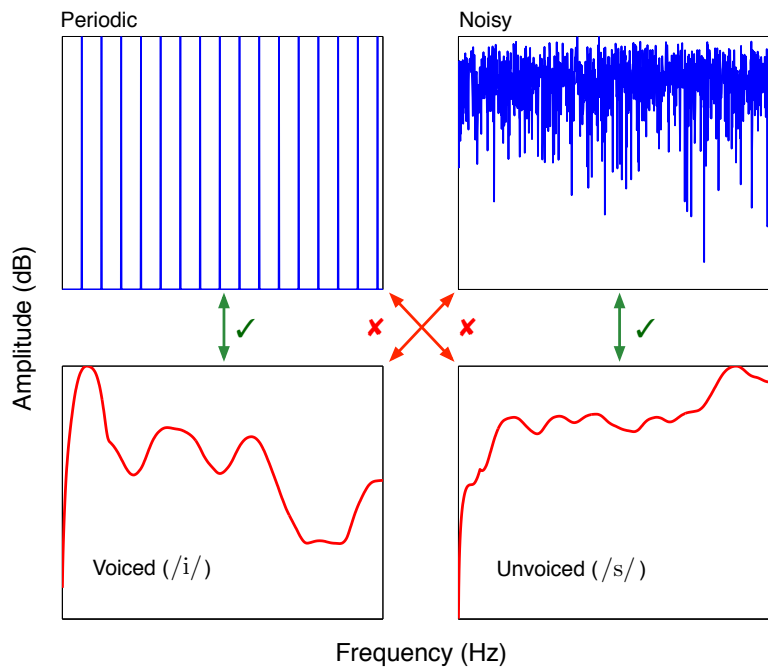


FIGURE 2.2 – Illustration schématique d'une combinaison réaliste (croix vertes) et non-réalistes (croix rouges) d'une source d'excitation (en haut) et d'un filtre (en bas). Source : (Bouvier et al., 2016)

Contrainte adaptative

Finalement, nous avons proposé un algorithme d'adaptation dynamique des pondérations des contraintes au cours de la séparation des sources sonores. L'une des limitations principales de l'application des pondérations des contraintes dans l'algorithme de la NMF réside dans le choix de la valeur de ces pondérations. Une petite valeur peut conduire à un effet négligeable ou nul de la contrainte, tandis qu'une grande valeur peut conduire à une trop forte prise en compte de la contrainte au détriment de la reconstruction. Dans les deux cas, ces choix peuvent conduire à de mauvaises solutions, en fonction du point d'initialisation. La seconde contribution consiste donc en l'adaptation du poids de

la contrainte à chaque itération durant la décomposition. La valeur du poids évolue de faible à forte en fonction de l'évolution de la reconstruction (plus précisément, du taux d'évolution de la valeur de la β -divergence). A la i -ème itération, le poids de contrainte λ est mis-à-jour comme :

$$\lambda^{(i)} = \lambda_{\max} \frac{D_{\text{IS}}(\mathbf{V}|\tilde{\mathbf{V}}^{(i-1)})}{D_{\text{IS}}(\mathbf{V}|\tilde{\mathbf{V}}^{(i-2)})} \quad (2.11)$$

où : λ est initialisé à 0 pour les deux premières itérations, puis varie dans l'intervalle $[0, \lambda_{\max}]$, avec λ_{\max} la valeur maximale de λ fixée à l'avance. Plus la valeur de la β -divergence est petite, plus la contrainte sera faible; plus la valeur de la contrainte est grande, plus la contrainte sera forte. L'idée sous-jacente est de laisser plus de liberté à la reconstruction lorsque celle-ci évolue fortement, puis d'augmenter graduellement le poids de la contrainte lorsque la reconstruction se stabilise. La Figure 2.3 présente une illustration de l'effet de l'adaptation proposée sur la fonction de coût. La différence principale réside dans la courbe de pondération des contraintes : sans adaptation, les contraintes sont fortes au début de la séparation; avec adaptation, les contraintes sont faibles au début de la séparation. Le coût de reconstruction diminue alors plus rapidement avec ce relâchement de contrainte initial. L'apparition des pondérations liées aux contraintes permet alors de réguler les solutions vers des solutions physiquement réalistes.

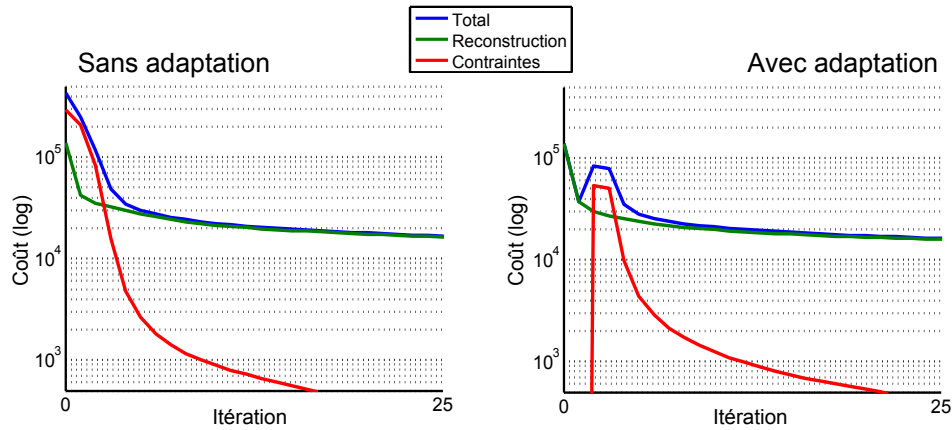


FIGURE 2.3 – Évolution du coût de reconstruction (en vert), du coût de contrainte (en rouge) et du coût total (en bleu) avec (à droite) et sans (à gauche) l'adaptation de la pondération des contraintes. Source : (Bouvier et al., 2016)

Expérience : séparation de sources sonores en environnement bruité

Methodologie

Le modèle SF-NMF semi-supervisé et les contraintes proposées ont été évalués sur une tâche de séparation de parole, et comparés avec des algorithmes de l'état-de-l'art, comme l'algorithme SF-NMF (Durrieu et al., 2009) non-supervisé (originellement proposé pour la séparation de la voix chantée et de la musique) et l'algorithme supervisé ASNA (Virtanen et al., 2013). Les algorithmes SF-NMF et ASNA nous donnent des bornes inférieures et supérieures utilisées comme indicateurs du positionnement de l'algorithme proposé : SF-NMF constitue la borne inférieure dans la mesure où il constitue la version non-supervisée de l'algorithme proposé, ASNA constitue la borne supérieure dans la mesure où cet algorithme est totalement supervisé, c'est-à-dire qu'il suppose les sources de bruits connues à l'avance et des bases de bruits pré-entraînées. Les algorithmes ont été évalués sur des mélanges instantanés réalisés à partir de signaux de parole propres issus de la base TIMIT d'anglais-américain (Zue et al., 1990) et de la base de sons environnementaux

(Dean et al., 2010). Les détails d’implémentation et le protocole expérimental sont décrits dans (Bouvier et al., 2016). Les métriques utilisées pour l’évaluation sont les métriques standards en séparation de sources audio (Vincent et al., 2006) : le rapport-signal-à-bruit (SNR, en dB), la distorsion-signal-à-bruit (SDR, en dB). Nous avons également évalué la qualité perceptive des signaux de parole séparés (PESQ) avec l’algorithme décrit dans (Rix et al., 2001).

		Algorithmes								
		Références		Proposé						
		ASNA (Virtanen et al., 2013)	SF-NMF (Durrieu et al., 2009)	# 1	# 2	# 3	# 4	# 5	# 6	# 7
Apprentissage	Parole	✓		✓	✓	✓	✓	✓	✓	✓
	Bruit	✓								
SF-NMF			✓	✓	✓	✓	✓	✓	✓	✓
Contraintes				SoA	cohérence	all	SoA	cohérence	all	
Adaptation				Sans			Avec			
-6	SDR	5.8	4.4	4.0	4.1	5.0	5.2	4.1	5.2	5.4
	PESQ	2.00	1.22	1.91	1.91	1.94	1.92	1.91	2.01	2.01
+0	SDR	10.7	7.8	9.1	9.2	9.0	8.9	9.2	9.8	9.8
	PESQ	2.44	1.54	2.30	2.30	2.24	2.23	2.30	2.34	2.35
+6	SDR	15.0	9.7	13.0	12.8	11.1	10.9	13.0	12.8	12.9
	PESQ	2.85	1.82	2.62	2.61	2.46	2.44	2.62	2.59	2.62
Mean	SDR	10.5	7.3	8.7	8.7	8.4	8.3	8.7	9.3	9.4
	PESQ	2.43	1.52	2.28	2.27	2.21	2.20	2.28	2.31	2.33

TABLEAU 2.1 – Mesures de SDR et PESQ obtenues pour les algorithmes comparés, en fonction du SNR (-6 dB, 0 dB, +6 dB). *SoA* correspond aux contraintes de l’état-de-l’art (parcimonie, décorrélation, et lissitude), *cohérence* à la contrainte de cohérence source/filtre proposée, et *adaptation* à l’adaptation de la pondération des contraintes. Dans les colonnes, nous présentons les algorithmes comparés : ASNA (Virtanen et al., 2013), SF-NMF (Durrieu et al., 2009), et les algorithmes proposés. Dans les premières lignes, nous présentons les spécificités des algorithmes comparés : la ligne apprentissage indique si les bases de parole et de bruit sont pré-apprises ; la ligne contraintes indique si les algorithmes utilisent des contraintes et lesquelles ; enfin, la ligne adaptation indique pour l’algorithme proposé si l’algorithme d’adaptation de la pondération des contraintes est activé.

Résultats et discussion

Le **Tableau 2.1** présente les scores de séparation pour les algorithmes comparés. Premièrement, l’algorithme source/filtre semi-supervisé (#1) améliore la séparation par rapport à l’algorithme source/filtre non-supervisé SF-NMF, ce qui démontre l’intérêt de pré-apprendre les bases de parole humaine préalablement pour guider la séparation. Deuxièmement, l’algorithme semi-supervisé proposé présente des performances supérieures à l’algorithme non-supervisé (V-IMM) et légèrement inférieures à l’algorithme supervisé (ASNA). Par rapport aux bornes mentionnées précédemment, l’algorithme proposé se situe proche de la borne supérieure de l’algorithme ASNA, mais sans nécessiter de connaissance ou d’apprentissage du bruit environnant. Deuxièmement, les contraintes source/filtre sans adaptation n’améliorent pas la séparation (#1 vs. #2, #3, et #4), mais l’adaptation de ces contraintes améliorent substantiellement la séparation (#5, #6, et #7), et tout particulièrement lorsque l’ensemble des contraintes sont considérées. Ceci

démontre notamment l'importance d'un apprentissage en plusieurs phases, en introduisant progressivement les contraintes : une première phase pour encoder les bases nécessaires à la reconstruction, et une deuxième phase pour moduler ces bases en fonction des contraintes formulées. Un examen approfondi des contraintes montre que les contraintes existantes (decorrélation, parcimonie, et lissitude) ont un effet faible sur la séparation (#5), alors que la contrainte de cohérence source/filtre proposée (# 6) a un effet important sur la séparation, en particulier pour des niveaux de bruit élevés.

2.3 FACTORISATION EN TENSEURS NON-NÉGATIFS POUR LA LOCALISATION BINAURALE DE SOURCES SONORES AVEC A PRIORI DE HRTF

Dans cette seconde section de ce chapitre, je présente les recherches menées avec Laurent Benaroya pour structurer une architecture NMF pour contraindre les solutions sur des tâches spécifiques, comme la localisation binaurale de sources sonores et présentées dans (Benaroya et al., 2018). L'écoute binaurale correspond à une branche de la robotique visant à créer des robots anthropomorphiques, en particulier capables des mêmes capacités perceptives qu'un être humain. En conséquence, un robot humanoïde a deux oreilles pour écouter son environnement sonore et pour être capable de localiser, séparer, et interpréter des sources sonores multiples possiblement mobiles dans un environnement bruité et réverbéré. L'écoute binaurale repose généralement sur deux indicateurs interauraux : la différence interaurale de temps ITD (respectivement, de phase IPD), et la différence interaurale de niveau (ILD). Ces indicateurs reflètent l'effet de la présence de la tête (et du corps) d'un être humain entre les deux capteurs que constituent les oreilles gauche et droite sur la perception de son environnement sonore. Ceci joue le rôle de filtres en gain et en phase pour chaque oreille en fonction de la localisation spatiale de la source sonore, caractérisée par ses fonctions de transfert liées à la tête (HRTF). La localisation binaurale (avec ou sans HRTF) a généralement été traitée par des algorithmes de traitement du signal, comme les algorithmes PHAT (Knapp and Carter, 1976), DUET (Jourjine et al., 2000), ou encore MESSL (Mandel and Ellis, 2007; Mandel et al., 2010). Par ailleurs, les algorithmes de localisation binaurale utilisant les HRTFs sont limités à la localisation d'une seule source sonore (Viste and Evangelista, 2004; Raspaud et al., 2010) dans la mesure où la réponse à des sources sonores multiples n'est pas linéaire, comme indiqué par la suite dans les [Équations \(2.13\) et \(2.14\)](#) qui ne s'appliquent que pour des sources considérées isolément.

L'idée de ces recherches était d'intégrer les connaissances sur les HRTFs dans un algorithme d'apprentissage pour associer les avantages de l'apprentissage machine et des connaissances humaines a priori. Cette idée a résulté dans l'implémentation d'un algorithme de factorisation en tenseurs non-négatifs (alternativement, factorisation multi-canaux en matrices non-négatives). En audio, la NMF a originellement été introduite en audio comme une représentation parcimonieuse de l'information temps-fréquence contenue dans un canal audio — par l'intermédiaire de bases spectrales et de leurs activations au cours du temps. Son extension à la représentation de signaux multi-canaux (FitzGerald et al., 2005; Ozerov and Fevotte, 2010) a permis d'intégrer implicitement l'information spatiale sous la forme de facteurs de mélange des sources sonores à travers les canaux. Mais ces facteurs de mélanges ne peuvent généralement pas être convertis en une position spatiale explicite (comme l'azimut d'une source sonore), et se limitent à des valeurs scalaires utilisées seulement à des fins d'illustration (Mitchell and Essa, 2006). La contribution principale est dans cette partie de formaliser une architecture NTF à deux canaux dans laquelle les facteurs de mélanges sont des vecteurs représentant explicitement le gain en amplitude du module des filtres HRTFs, en encodant explicitement l'information spatiale de sources sonores. Par ailleurs, l'algorithme proposé permet à la fois de réaliser la localisation et la séparation de sources sonores.

On considère une source sonore émettant, à un temps t en échantillons, un signal $s(t)$ dont la position dans l'espace est définie par sa distance d (en m.), son azimuth θ (en rad.) et son élévation ψ (en rad.) par rapport au centre de la tête d'un robot, comme illustré sur la **Figure 2.4**. La représentation temps-fréquence $S_{k,n}$ correspondante est obtenue par la transformée de Fourier à court-terme, avec les indexes fréquentiels $k \in [1, \dots, F]$ (en points fréquentiels) et les indexes temporels $n \in [1, \dots, T]$ (en trames). On définit $y_l(t)$ et $y_r(t)$ les signaux gauche et droite perçus par le robot, avec comme représentation fréquentielle le tenseur $\mathcal{Y} \in \mathbb{C}^{2 \times F \times T}$ et les représentations temps-fréquence correspondante gauche et droite $\mathcal{Y}_{l,k,n}$ et $\mathcal{Y}_{r,k,n}$. Les signaux binauraux perçus sont obtenus par les modifications

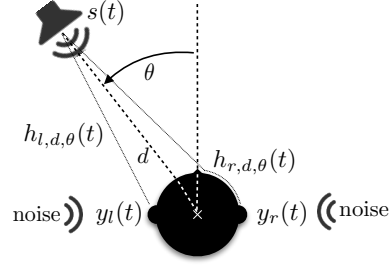


FIGURE 2.4 – Illustration de la localisation binaurale d'une source sonore. Une source sonore émet un signal $s(t)$ à la position $(d, \theta, \psi = 0^\circ)$, où l'azimuth $\theta = 0^\circ$ correspond à une source placée en face de la tête. Le signal source se propage aux oreilles gauche et droite de la tête, définissant les signaux binauraux gauche et droit $y_l(t)$ et $y_r(t)$. Ces deux signaux sont reliés au signal source $s(t)$ par l'intermédiaire de la réponse impulsionnelle liée à la tête $h_{l,d,\theta}(t)$ et $h_{r,d,\theta}(t)$ respectivement. Additionnellement, un bruit diffus est simulé produisant un bruit additif supplémentaire sur les signaux gauche et droit. Source : (Benaroya et al., 2018)

engendrées le corps du robot sur l'onde sonore incidente, incluant son torse, sa tête, et les pavillons auditifs. Ces effets sont captés à travers les réponses impulsionnelles liées à la tête (HRIRs) $h_{l,\cdot}(t)$ et $h_{r,\cdot}(t)$ des canaux gauche l et droit r , dont les transformées de Fourier définissent les fonctions de transfert liées à la tête (HRTFs) $H_{l,\cdot}$ et $H_{r,\cdot}$, fonctions de la position spatiale de la source sonore paramétrée en coordonnées sphériques par son azimuth θ , son élévation φ , et sa distance d . Dans le reste de cette partie, nous nous intéresserons uniquement à l'azimuth θ de la source, et ignorerons les termes d'élévation et de distance. Dans l'hypothèse anéchoïque, la relation entre la source émettrice et les signaux binauraux perçus s'écrit alors :

$$\begin{cases} \mathcal{Y}_{l,k,n} = H_{l,\theta,k} S_{k,n} \\ \mathcal{Y}_{r,k,n} = H_{r,\theta,k} S_{k,n} \end{cases} \quad (2.12)$$

Dérivée de la théorie Duplex introduite dans (Rayleigh, 1907), la localisation binaurale de sources sonores repose sur deux indices interauraux. La différence inter-aurale de niveau (ILD) est définie comme la différence des niveaux perçus par l'oreille gauche et droite,

$$ILD(k, n) = 20 \log \frac{|\mathcal{Y}_{l,k,n}|}{|\mathcal{Y}_{r,k,n}|}. \quad (2.13)$$

La différence interaurale de phase (IPD) est définie par la différence de chemin parcourue par l'onde sonore pour arriver à l'oreille gauche et droite,

$$IPD(k, n) = \angle \frac{\mathcal{Y}_{l,k,n}}{\mathcal{Y}_{r,k,n}}, \quad (2.14)$$

où : \angle représente la phase en radians d'un nombre complexe.

En injectant les HRTFs telles que définis par l'Équation (2.12) dans l'Équation (2.13) et l'Équation (2.14), on peut alors écrire les indices binauraux comme une fonction de l'azimuth,

$$ILD_{\theta}^{\text{hrtf}}(k) = 20 \log \frac{|H_{l,\theta,k}|}{|H_{r,\theta,k}|}, \quad (2.15)$$

$$IPD_{\theta}^{\text{hrtf}}(k) = \angle \frac{H_{l,\theta,k}}{H_{r,\theta,k}}. \quad (2.16)$$

La carte des indices binauraux théoriques calculés à partir de mesures de HRTFs d'un buste Kemar en fonction de l'azimuth sont représentés sur la Figure 2.5.

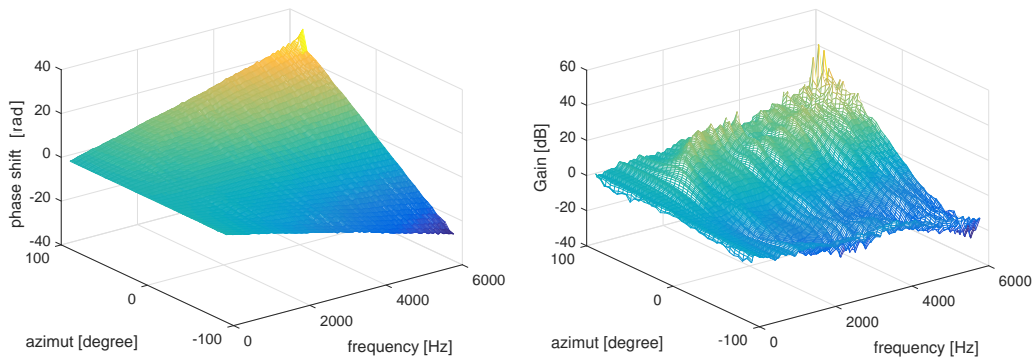


FIGURE 2.5 – Illustration des indices IPD (à gauche) et ILD (à droite) en fonction de l'azimuth de la source et de la fréquence à partir des mesures de HRTFs réalisées dans (Gardner and Martin, 1995). Source : (Benaroya et al., 2018)

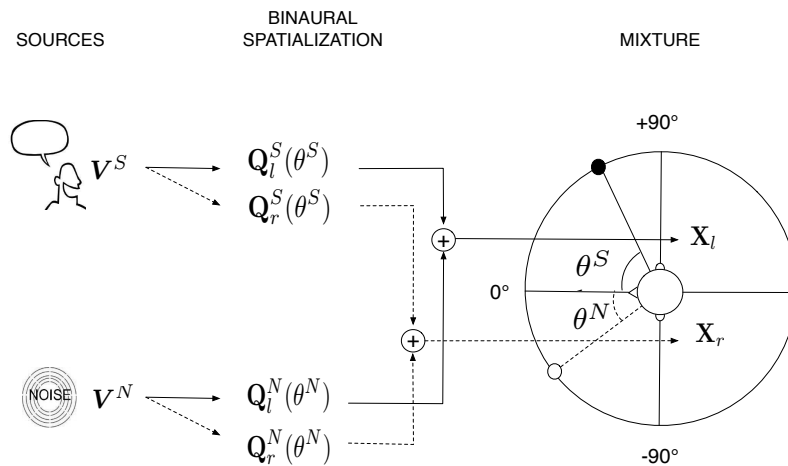


FIGURE 2.6 – Illustration de la localisation binaurale de sources sonores par NTF dans le contexte de la séparation d'une source de parole et d'une source de bruit. Pour la source de parole : V^S est le spectrogramme d'amplitude de la source de parole, $Q_l^S(\theta^S)$ et $Q_r^S(\theta^S)$ sont les HRTFs associés à l'azimuth θ^S . Pour la source de bruit, qui est considérée comme une source ponctuelle dans l'illustration : V^N est le spectrogramme d'amplitude de la source bruitée, $Q_l^N(\theta^N)$ et $Q_r^N(\theta^N)$ sont les HRTFs associés à l'azimuth θ^N . Source : (Benaroya et al., 2018)

La factorisation en tenseurs non-négatifs (NTF) est une généralisation de la NMF à plusieurs canaux, comme par exemple issus de signaux captés par des antennes de microphones (FitzGerald et al., 2005; Ozerov and Fevotte, 2010). Cette généralisation repose sur l'hypothèse selon laquelle les sources sont distribuées de manière différenciée en fonction des canaux, ce qui est traduit par l'intermédiaire de matrices de mélanges sur les canaux. La NTF introduit implicitement la notion de mélange spatial dans le cadre de la NMF et permet ainsi de passer d'une représentation temps-fréquence à une représentation temps-fréquence-espace. Cette section présente la formulation de la NTF dans le cadre de la localisation binaurale de sources sonores multiples. La NTF binaurale est obtenue en fixant le nombre de canaux $c \in \{l, r\}$, où $c = l$ pour l'oreille gauche et $c = r$ pour l'oreille droite. On obtient ainsi un spectrogramme binaural $\mathcal{X} = \mathbf{X}_l \times \mathbf{X}_r \in \mathbb{R}^{+2 \times F \times T}$ et $\mathbf{X}_c \in \mathbb{R}^{+F \times T}$, qui est approximé par un tenseur \mathcal{V} par factorisation NTF. La contribution de chaque source u à l'approximation du spectrogramme binaural \mathcal{V} est représentée par le produit terme-à-terme d'une matrice temps-fréquence $\mathbf{V} \in \mathbb{R}^{+F \times T}$ et d'une matrice de mélange spatial $\mathbf{Q} \in \mathbb{R}^{+2 \times F}$ comme présenté dans (Ozerov and Fevotte, 2010) :

$$u_{c,k,n} = \mathbf{Q}_{c,k} \cdot \mathbf{V}_{k,n} \quad (2.17)$$

où : c , k et n représentent respectivement les indexes de canal, de fréquence, et de trame.

Le spectrogramme binaural d'amplitude \mathcal{X} s'écrit alors :

$$\mathcal{V} = \sum_{m=1}^M \mathcal{U}^S[m] + \sum_{p=1}^P \mathcal{U}^N[p]. \quad (2.18)$$

où : les approximations spectrales binaurales de la $m^{\text{ème}}$ source de parole $\mathcal{U}^S[m]$ et de la $p^{\text{ème}}$ source de bruit $\mathcal{U}^N[p]$, sont obtenues par multiplication des approximations spectrales $\mathbf{V}^{S,(m)}$ et $\mathbf{V}^{N,(p)}$ par les matrices de mélange spatial $\mathbf{Q}^{S,(m)}$ et $\mathbf{Q}^{N,(p)}$ fonctions du canal c et de la fréquence k , comme indiqué par l'Équation (2.17).

La m -ième source image de parole est approximée par décomposition SF-NMF :

$$\mathbf{V}^{S,(m)} \simeq \left(\mathbf{W}_{\text{ex}} \mathbf{H}_{\text{ex}}^{S,(m)} \right) \otimes \left(\mathbf{W}_{\Phi}^{S,(m)} \hat{\mathbf{U}} \Phi \mathbf{H}_{\Phi}^{S,(m)} \right) \quad (2.19)$$

La p -ième source image de bruit est approximée par décomposition NMF :

$$\mathbf{V}^{N,(p)} \simeq \mathbf{W}^{N,(p)} \mathbf{H}^{N,(p)} \quad (2.20)$$

Le modèle NTF décrit par l'Équation (2.18) est semi-supervisé (Joder et al., 2012; Weninger et al., 2012) : les spectrogrammes d'amplitude des sources de parole $\mathbf{V}^{S,(m)}$ sont décomposés suivant un modèle SF-NMF dont les bases sont apprises de manière supervisée à partir d'enregistrements de parole propres puis fixées pendant la séparation, et les spectrogrammes d'amplitudes des sources de bruit $\mathbf{V}^{N,(p)}$ sont décomposées suivant un modèle NMF classique de manière non-supervisée pendant la séparation. Les matrices de mélanges $\mathbf{Q}^{S,(m)}$ et $\mathbf{Q}^{N,(p)}$ sont également estimées de manière non-supervisée pendant la séparation, et sont considérées comme une estimation du module des HRTFs correspondant à chacune des sources. Un résumé des matrices utilisées pour la décomposition NTF binaurale est présenté dans le Tableau 2.2.

TABLEAU 2.2 – Statuts des matrices et tenseurs NMF/NTF correspondant respectivement à la m -ième source de parole et à la p -ième source de bruit pendant les phases d'apprentissage et de séparation. Les paramètres des matrices sont soit **fixés** a priori et non-entraînaibles, soit **libres** (c'est-à-dire, entraînaibles), soit encore **figés** (c'est-à-dire, non-entraînaibles après pré-apprentissage).

	apprentissage	séparation
\mathbf{W}_{ex}	fixe	figé
$\mathbf{H}_{\text{ex}}^{S,(m)}$	-	libre
$\mathbf{W}_{\Phi}^{S,(m)}$	libre	figé
$\hat{\mathbf{U}}^{\Phi}$	fixe	figé
$\mathbf{H}_{\Phi}^{S,(m)}$	libre	libre
\mathbf{W}^N	-	libre
\mathbf{H}^N	-	libre
$\mathbf{Q}^{S,(m)}$	-	libre
$\mathbf{Q}^{N,(p)}$	-	libre

Les paramètres de la NMF sont estimés par minimisation de l'erreur de reconstruction définie comme la somme à travers les canaux des β -divergence entre les spectrogrammes d'amplitude observé et approximé :

$$\mathcal{D}_{\beta}(\mathcal{X}|\mathcal{V}) = \sum_{c \in \{l,r\}} \mathcal{D}_{\beta}(\mathbf{X}_c|\mathbf{V}_c) \quad (2.21)$$

où : $\mathcal{D}_{\beta}(\mathbf{X}_c|\mathbf{V}_c)$ est la divergence définie par l'Équation (2.2). Les contributions de chaque canal à la fonction de coût sont indépendantes, c'est-à-dire qu'il n'existe pas de termes inter-canaux. De manière similaire à la dérivation des MU pour le modèle NMF, nous pouvons dériver avec la β -divergence les MU des paramètres du modèle NTF binaural proposé à partir des Équations (2.18), (2.20) et (2.5), (2.21). Pour des raisons de clarté, nous formulons les MU à partir des notations matricielles \mathbf{X}_c , \mathbf{V}_c , $\mathbf{Q}_c^{S,(m)}$ et $\mathbf{Q}_c^{N,(p)}$ avec $c \in \{l, r\}$ plutôt qu'avec les notations tensorielles \mathcal{X} , \mathcal{V} et $\mathbf{Q}^{S,(m)}$ et $\mathbf{Q}^{N,(p)}$. Ainsi, les règles de mise-à-jour multiplicatives (MU) des matrices libres \mathbf{H}_{ex} , \mathbf{H}_{Φ} , \mathbf{W}^N , \mathbf{H}^N , \mathbf{Q}_c^S , et \mathbf{Q}_c^N s'écrivent de la manière suivante :

Matrices d'activation des sources de parole :

$$\mathbf{H}_{\text{ex}m} \leftarrow \mathbf{H}_{\text{ex}m} \odot \frac{(\mathbf{W}_{\text{ex}m})^{\top} [\sum_{c \in \{l,r\}} \text{diag}(\mathbf{Q}_c^{S,(m)}) (\mathbf{X}_c \odot \mathbf{V}_c^{\beta-2} \odot \mathbf{V}_{\Phi m})]}{(\mathbf{W}_{\text{ex}m})^{\top} [\sum_{c \in \{l,r\}} \text{diag}(\mathbf{Q}_c^{S,(m)}) (\mathbf{V}_c^{\beta-1} \odot \mathbf{V}_{\Phi m})]} \quad (2.22)$$

$$\mathbf{H}_{\Phi m} \leftarrow \mathbf{H}_{\Phi m} \odot \frac{(\mathbf{W}_{\Phi m})^{\top} [\sum_{c \in \{l,r\}} \text{diag}(\mathbf{Q}_c^{S,(m)}) (\mathbf{X}_c \odot \mathbf{V}_c^{\beta-2} \odot \mathbf{V}_{\text{ex}m})]}{(\mathbf{W}_{\Phi m})^{\top} [\sum_{c \in \{l,r\}} \text{diag}(\mathbf{Q}_c^{S,(m)}) (\mathbf{V}_c^{\beta-1} \odot \mathbf{V}_{\text{ex}m})]} \quad (2.23)$$

Matrice des sources de bruit :

$$\mathbf{W}_p^N \leftarrow \mathbf{W}_p^N \odot \frac{[\sum_{c \in \{l,r\}} \text{diag}(\mathbf{Q}_c^{N,(p)}) (\mathbf{X}_c \odot \mathbf{V}_c^{\beta-2})] (\mathbf{H}_p^N)^{\top}}{[\sum_{c \in \{l,r\}} \text{diag}(\mathbf{Q}_c^{N,(p)}) \mathbf{V}_c^{\beta-1}] (\mathbf{H}_p^N)^{\top}} \quad (2.24)$$

$$\mathbf{H}_p^N \leftarrow \mathbf{H}_p^N \odot \frac{(\mathbf{W}_p^N)^{\top} [\sum_{c \in \{l,r\}} \text{diag}(\mathbf{Q}_c^{N,(p)}) (\mathbf{X}_c \odot \mathbf{V}_c^{\beta-2})]}{(\mathbf{W}_p^N)^{\top} [\sum_{c \in \{l,r\}} \text{diag}(\mathbf{Q}_c^{N,(p)}) \mathbf{V}_c^{\beta-1}]} \quad (2.25)$$

Matrices de mélange binaural :

$$\mathbf{Q}_c^{S,(m)} \leftarrow \mathbf{Q}_c^{S,(m)} \odot \frac{[\mathbf{X}_c \odot \mathbf{V}_c^{\beta-2} \odot \mathbf{V}_m^S] \mathbf{1}_{T \times 1}}{[\mathbf{V}_c^{\beta-1} \odot \mathbf{V}_m^S] \mathbf{1}_{T \times 1}} \quad (2.26)$$

$$\mathbf{Q}_c^{N,(p)} \leftarrow \mathbf{Q}_c^{N,(p)} \odot \frac{[\mathbf{X}_c \odot \mathbf{V}_c^{\beta-2} \odot \mathbf{V}_p^N] \mathbf{1}_{T \times 1}}{[\mathbf{V}_c^{\beta-1} \odot \mathbf{V}_p^N] \mathbf{1}_{T \times 1}} \quad (2.27)$$

où : $\mathbf{1}_{T \times 1}$ est un vecteur de dimension T ne contenant que des 1. Considérant un vecteur $\mathbf{u} \in \mathbb{R}^F$, $\text{diag}(\mathbf{u})$ est la matrice diagonale dans $\mathbb{R}^{F \times F}$ où les coefficients de \mathbf{u} sont sur la diagonale.

Localisation de sources sonores par NTF

À partir de l'algorithme de factorisation de signaux binauraux proposé, nous avons proposé deux stratégies pour la localisation de sources sonores : la localisation est réalisée de manière interne ou externe à la factorisation.

Localisation à partir de la matrice de mélange binaural estimée

La première idée est de réaliser la localisation à partir de la matrice de mélange $\mathbf{Q}^{S,(m)}$. En effet, $\mathbf{Q}^{S,(m)}$ encode l'amplitude des HRTFs - noté par la suite, MHRTF - correspondant respectivement aux oreilles gauche et droite de la tête binaurale. En conséquence, elle peut être utilisée directement pour calculer la position spatiale de chaque source sonore, par comparaison des MHRTFs estimées à partir de la matrice de mélange et les vraies MHRTFs mesurées sur la tête binaurale. Pour simplifier les notations, nous ignorons dans la suite l'index m de la source de parole, sans perte de généralité. A partir de la matrice de mélange $\mathbf{Q}^{S,(m)}$, nous pouvons calculer directement l'ILD correspondant à la source de parole :

$$\text{ILD}^{\text{est}}(k) = 20 \log_{10} \frac{Q_{l,k}^S}{Q_{r,k}^S} \quad (2.28)$$

Comme la matrice de mélange \mathbf{Q}^S est supposée constante en temps, l'ILD estimé l'est par conséquent également. L'hypothèse de stationnarité sur \mathbf{Q}^S correspond à l'hypothèse que la source de parole est fixe dans le temps.

La localisation de la source de parole est alors réalisée par minimisation avec les ILDs de référence calculés à partir de la base de données des HRTFs de la tête binaurale. L'azimuth est ainsi estimé par minimisation de l'erreur quadratique :

$$\hat{\theta} = \underset{\theta}{\text{argmin}} \|\text{ILD}^{\text{est}} - \text{ILD}_{\theta}^{\text{hrtf}}\|_2 \quad (2.29)$$

Pour être en accord avec la théorie Duplex (Rayleigh, 1907) selon laquelle l'ILD n'est exploitable que dans les hautes-fréquences, nous avons limité la plage de fréquence pour l'estimation de l'ILD dans l'Équation (2.29) à la bande-passante en hautes-fréquences (1500 to 4500 Hz). La Figure 2.7 présente une comparaison entre l'ILD estimé par NTF et l'ILD théorique correspondant calculé à partir d'une base de données de HRTFs.

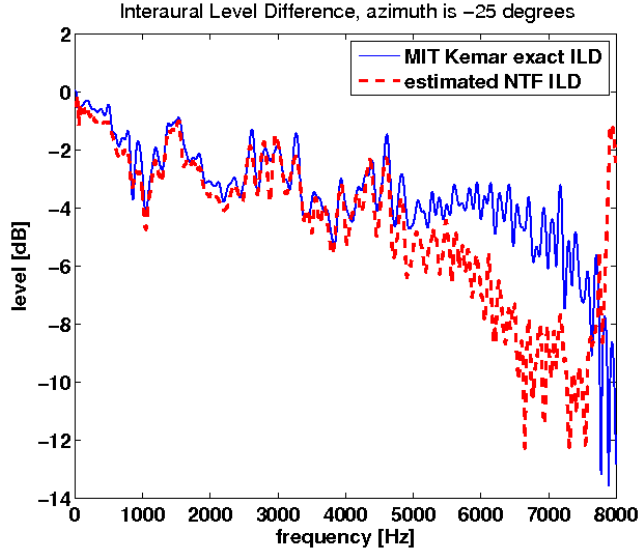


FIGURE 2.7 – Comparaison de l’ILD mesuré à partir des HRTFs de la base de données MIT pour une tête Kemar (trait bleu plein), et l’ILD estimé à partir de la matrice de mélange \mathbf{Q}^S par l’algorithme NTF proposé (trait rouge pointillé). Cet exemple a été réalisé avec une phrase du locuteur FCJFo de la base TIMIT spatialisée à un azimuth de -25° sans bruit dans une chambre anéchoïque. Source : (Benaroya et al., 2018)

Localisation à partir de l’image de la source

Une autre solution consiste à réaliser la séparation sources sonores, en calculant les images correspondant à chacune des sources de parole. L’image $\hat{\mathbf{S}}_m^S$ d’une source de parole se calcule par filtrage de Wiener généralisé (Benaroya et al., 2006) :

$$\hat{\mathbf{S}}_{c,m}^S = \mathbf{Y}_c \odot \frac{\mathbf{Q}_c^{S,(m)} \odot \mathbf{V}^{S,(m)}}{\mathbf{V}_c} \quad (2.30)$$

où : $\mathbf{y} \in \mathbb{C}^{2 \times F \times T}$ est la TFCT de l’image stéréo du mélange et \mathbf{V}_c est défini dans l’Équation (2.18).

La localisation de la source sonore peut alors être estimée à partir de l’ILD de l’image de chaque source :

$$\text{ILD}(k, n) = 20 \log_{10} \left(\frac{Q_{l,k}^S}{Q_{r,k}^S} \cdot \frac{|Y_{l,k,n}|}{|Y_{r,k,n}|} \cdot \frac{V_{r,k,n}}{V_{l,k,n}} \right) \quad (2.31)$$

Cette expression est égale à l’Équation (2.28) pour le cas où les spectrogrammes d’amplitude $|Y_{c,k,n}|$ et leurs approximations $V_{c,k,n}$ sont égales, par exemple lorsque la factorisation est de rang plein.

La Figure 2.8 illustre la localisation réalisées pour deux locuteurs en présence d’un bruit diffus à 0dB.

Expérience : localisation de sources sonores

Méthodologie

Pour réaliser l’évaluation expérimentale de l’algorithme de localisation proposé, nous avons créé une large base de scènes sonores binaurales simulées avec le logiciel Ircam Spat (Carpentier et al., 2015), par spatialisation et mélange de sources de parole issus de

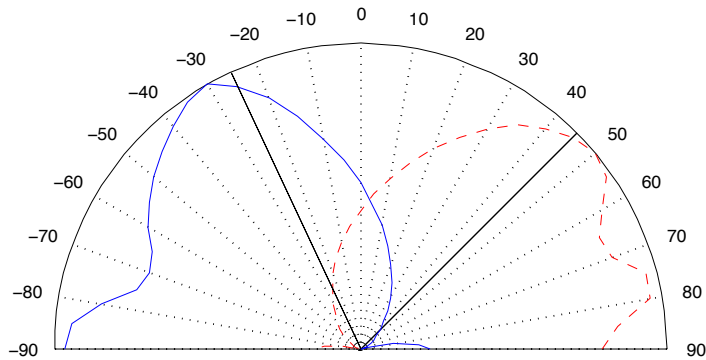


FIGURE 2.8 – Tracé radar en coordonnées polaires de l’inverse de l’erreur en fonction de l’azimuth pour la localisation de deux sources sonores dans un environnement bruité. Cet exemple est réalisé pour deux locuteurs de la base TIMIT localisés respectivement à -25° et $+40^\circ$ en présence d’un bruit diffus issu de la base QUT-NOISE avec un SNR de 0 dB. Le trait plein indique la position réelle de la source et les courbes l’inverse de la fonction d’erreur pour les deux locuteurs. Source : (Benaroya et al., 2018)

la base de données de locuteurs TIMIT (Zue et al., 1990) et la base de données de bruits environnementaux QUT-NOISE-TIMIT (Dean et al., 2010). Les sources sonores ont été utilisées uniquement pour simuler un bruit diffus, et le mélange des sources de parole avec les sources de bruit a été réalisé pour quatre niveaux de rapport signal sur bruit (SNR) : infini, +6, 0, et -6 dB (Viña et al., 2011). Nous avons évalué sur cette base les performances de localisation obtenues par l’algorithme de NTF binaural proposé selon les deux stratégies mentionnées : la localisation binaurale à partir de la matrice de mélange ($NTF + Q^S$), et la localisation binaurale réalisée à partir de l’image des sources estimée par séparation (NTF+sep). Ces deux versions de l’algorithme de localisation binaurale proposées ont été comparées avec des algorithmes existants pour la localisation binaurale de sources sonores : GCC-PHAT (Blandin et al., 2012), DUET (Jourjine et al., 2000), et une localisation binaurale basée sur l’Équation (2.29) (ILD). Pour métrique d’évaluation, nous nous sommes inspirés de la métrique de l’exactitude grossière (GA en %) proposée dans (Woodruff and Wang, 2012) définie comme le pourcentage de localisation correcte avec un seuil de tolérance fixe autour de la vraie valeur, et à partir de laquelle nous avons proposé un équivalent fondé sur la perception humaine en modulant de seuil en fonction de l’azimuth de la source. En effet, la littérature en perception auditive montre que la précision d’un être humain pour localiser une source sonore n’est pas constante et dépend de la position relative de la source sonore (Stevens and Newman, 1936; Butler, 1986). Ainsi, un être humain est plus précis pour localiser une source en face de lui (quelques degrés d’incertitude à 0 degré), et moins précis pour localiser des sources sur les bords (environ 10 à 20 degrés d’incertitude pour des sources à ± 90 degrés). La mesure proposée a en conséquence un seuil de tolérance plus étroit en face (0 degré), et plus étendu sur les côtés (± 90 degrés).

Nous avons procédé à deux séries d’expérience : la première pour la localisation d’un locuteur en présence d’un bruit diffus, et la seconde avec deux locuteurs dans les mêmes conditions. La seconde expérience était essentiellement pour s’assurer du fonctionnement de l’algorithme dans le cas de sources multiples. Pour le cas avec un seul locuteur, les expériences ont été réalisées pour un ensemble d’énoncés de 10 locuteurs de la base TIMIT spatialisés de -90 degrés à $+90$ degrés (pour référence : 0 degré est en face) avec un pas de 10 degrés, 4 bruits tirés de la base QUT-NOISE-TIMIT, les 4 niveaux SNRs. Pour le cas avec deux locuteurs, la procédure est la même à l’exception que nous avons tiré des paires de locuteurs au hasard parmi les locuteurs sélectionnés pour l’expérience.

Le **Tableau 2.3** présente la performance des algorithmes comparés pour la localisation d'une source de parole mélangée avec du bruit diffus, en fonction du SNR.

TABLEAU 2.3 – **Exactitude grossière (%), Bruit diffus, Un locuteur.** Comparaison des algorithmes de localisation binaurale de source sonore.

SNR	-6	0	+6	$+\infty$
<i>binLoc</i>				
GCC-PHAT	6.6	15.9	45.9	100
DUET	6.8	11.0	29.7	98.6
ILD	16.2	45.1	78.4	100
<i>NTF + sep</i>				
NTF + sep + GCC-PHAT	10.0	33.9	67.6	100
NTF + sep + DUET	10.4	37.7	68.1	98.6
NTF + sep + ILD	70.1	94.3	99.6	100
<i>NTF + Q^S</i>				
NMF + est Q ^S	79.9	96.2	99.7	100

Premièrement, nous pouvons observer que les algorithmes existants de localisation binaurale sont extrêmement sensibles à la présence de bruit, avec une performance qui chute radicalement. Deuxièmement, les algorithmes de localisation utilisant la NTF pour séparer les sources sonores en amont de la localisation (*NTF + sep*) montrent que la dégradation de la localisation liée à la présence de bruit se réduit considérablement. En particulier, l'algorithme basé sur l'estimation de l'ILD présente des performances relativement stables en fonction des niveaux de bruit. Enfin, l'algorithme proposé de localisation à partir de la matrice de mélange (*NTF est, NMF + est Q^S*) présente les meilleures performances, avec une complexité moindre, et pour tous les SNRs considérés. Un examen détaillé des erreurs de localisation révèle des erreurs plus prononcées sur les côtés (± 90 degrés) qu'en face (0 degré), ce qui est en conformité avec les connaissances sur la capacités des êtres humains à localiser des sources sonores ([Stevens and Newman, 1936](#); [Butler, 1986](#)).

Le **Tableau 2.4** présente les scores de l'algorithme proposé sur la tâche connexe de séparation. La performance de séparation est mesurée à partir des métriques recommandées dans le domaine ([Vincent et al., 2007](#)) : SDR (rapport signal à distorsion), ISR (rapport image source à distorsion spatiale), SIR (rapport source à interférence), et SAR (rapport source à artefacts).

TABLEAU 2.4 – **SDR, ISR, SIR, SAR, Bruit diffus, Un locuteur.** Comparaison des algorithmes proposés pour la séparation de source sonore.

SNR	SDR	ISR	SIR	SAR
MU				
-6	-1.0	8.3	-2.8	13.4
0	0.7	9.3	4.1	15.1
+6	0.5	3.3	11.1	17.3

Nous observons que l'algorithme proposé présente un SAR élevé. Ceci indique moindre une distorsion artefactuelle ou présence d'artefacts résiduels dans les sources séparées, c'est-à-dire une bonne séparation des sources dans le domaine spectro-temporel. Néanmoins, cette performance se dégrade avec l'augmentation du niveau de bruit. Par ailleurs, nous observons également un ISR élevé et relativement stable avec l'augmentation du niveau de

bruit. Ceci indique une moindre distorsion spatiale, c'est-à-dire une meilleure séparation des sources dans le domaine spatial, et par conséquent une meilleure estimation de l'information spatiale pour la localisation. Il existe en l'occurrence probablement une relation entre le SAR et le ISR dans le contexte de la localisation binaurale : si on suppose que le système formé de la factorisation temps-fréquence et de la factorisation spatiale est à énergie constante, l'énergie des deux factorisations se compensent. Et par conséquence, une erreur sur l'une factorisation provoque une erreur sur l'autre factorisation par compensation.

Le **Tableau 2.5** présente les scores de localisation obtenus pour des variantes de l'algorithme proposé $NTF + Q^S$. En particulier, nous avons examiné :

- l'effet de la nature du modèle pour la source de parole, soit une NMF standard (NMF), soit une NMF source/filtre (SF-NMF) telle que présentée dans la section précédente.
- l'effet de la nature du modèle pour la source de bruit : soit absente (w./o Q^N), soit présentée mais non localisée (cst $Q^N = 1$, i.e., la source de bruit est supposée en face), soit présente et localisée (est Q^N).
- la capacité de généralisation à des locuteurs inconnus, c'est-à-dire l'utilisation de paires de locuteurs possiblement différents à l'apprentissage et au test : soit le même locuteur (same speak.), soit deux locuteurs différents (diff. speak.)

TABLEAU 2.5 – **Exactitude grossière (%), Bruit diffus, Un locuteur.** Comparaison des variantes de l'algorithme de localisation binaurale proposé. Les variantes considérées comprennent 1) les modèles de source : NMF ou SF-NMF pour le modèle de source de locuteur, et le modèle de source de bruit est toujours NMF ; 2) entraîné et évalué sur le même locuteur ou sur un locuteur différent ; 3) et le modèle spatial de source de bruit : pas de source (w/o Q^N), une source fixée a priori et positionnée en face (w. cst Q^N), et une source dont la position est estimée w. est Q^N

SNR		-6	0	+6	$+\infty$
NMF	w. speak. w. est Q^N	79.9	96.2	99.7	100
SF-NMF	same speak. w/o Q^N	10.1	17.6	39.3	100
SF-NMF	same speak. w. cst Q^N	52.1	71.6	85.5	80.5
SF-NMF	same speak. w. est Q^N	86.3	96.6	99.5	100
SF-NMF	diff. speak. w. est Q^N	85.7	95.9	98.9	100

Premièrement, l'utilisation d'un modèle source/filtre pour les sources de parole améliore la localisation par rapport à un modèle NMF standard. Deuxièmement, la performance de localisation semble relativement insensible au locuteur (SF-NMF same speak. w. est Q^N vs. SF-NMF diff. speak. w. est Q^N). Naturellement, la connaissance a priori du locuteur améliore la localisation, mais la différence est relativement faible avec un locuteur inconnu (environ 1% pour tous les SNRs). Troisièmement, l'intégration d'une source de bruit améliore substantiellement la localisation, en particulier pour un modèle de source de bruit localisé. Un examen des résultats obtenus indique que plus le niveau de bruit est élevé plus le modèle de bruit localisé aide à la localisation (+14.0% GA at +6 dB, +25.0% GA at 0 dB, et +34.2% GA at -6 dB). Cette observation est probablement liée à la réduction de la distorsion spatiale mentionnée dans le **Tableau 2.4**.

Expérience 2 : vers la localisation de locuteurs multiples

Le **Tableau 2.6** présente les résultats obtenus pour la localisation de deux locuteurs dans du bruit diffus. Dans ce scénario expérimental, le nombre de sources présentes est spécifié comme un méta-paramètre supposé connu.

TABLEAU 2.6 – **Exactitude grossière (%), Bruit diffus, Deux locuteurs.** Comparaison de la localisation pour une et deux sources de parole. Pour les sources de parole : modélisation temps-fréquence par NMF ou NMF source/filtre SF-NMF. Pour les sources de bruit : modélisation temps-fréquence par NMF, avec une matrice de mélange ignorée (w/o Q^N), fixée à une valeur unitaire (w. cst Q^N), ou estimée (w. est Q^N)

SNR		-6	0	+6	$+\infty$
<i>un locuteur</i>					
NMF	w. est Q^N	79.9	96.2	99.7	100
SF-NMF	w. est Q^N	86.3	96.6	99.5	100
<i>deux locuteurs</i>					
NMF	w. est Q^N	63.1	74.9	76.1	75.2
SF-NMF	w/o Q^N	47.2	58.4	69.8	87.5
SF-NMF	w. cst Q^N	59.7	69.8	74.1	74.2
SF-NMF	w. est Q^N	64.7	72.0	75.3	79.4
SF-NMF	w. est Q^N w perm.	68.7	74.5	76.5	80.0

Tout d'abord, nous observons les mêmes tendances que celles présentées pour la localisation d'un locuteur : l'algorithme avec un modèle source/filtre NMF pour la source de parole et un modèle de source de bruit avec estimation de sa matrice de mélange spatial présente les meilleurs scores de localisation. Le score de localisation est d'environ 60-75% pour l'algorithme $NTF + Q^S$ (SF-NMF w. est Q^N) et augmente jusqu'à 70-80% si nous ignorons les erreurs de localisation dues à une simple inversion des locuteurs (SF-NMF w. est Q^N w. perm). Malheureusement, la localisation se dégrade de manière non négligeable avec l'augmentation du niveau de bruit : comme discuté précédemment, ce phénomène s'explique probablement par une moins bonne séparation des locuteurs, et est amplifié par le nombre de sources à localiser.

2.4 SYNTHÈSE ET DISCUSSION

Pour conclure ce chapitre, je propose une synthèse des principales contributions réalisées au cours des recherches présentées, et en discute les propriétés au regard des questions de recherche formulées. Les contributions apportées dans ce chapitre sont :

Contributions

C1. Un algorithme de factorisation en matrices non négatives inspiré du modèle source/filtre présenté dans (Durrieu et al., 2010). En particulier, la formulation de contraintes source/filtre inspirées par la physique et un algorithme de contraintes adaptatives pour réguler l'application de ces contraintes au cours de l'apprentissage. Cette stratégie est similaire aux stratégies d'amorçage utilisés pour l'apprentissage de réseaux de neurones, avec pour idée principale de contraindre l'espace des solutions à un sous-espace répondant à un ensemble de spécifications ou de contraintes implicites. Cet algorithme a été évalué expérimentalement sur une tâche de séparation de sources sonores en environnement bruité.

C2. Un algorithme de factorisation en matrices non-négatives multi-canal, avec la formulation explicite d'un modèle d'écoute binaural avec connaissance a priori des fonctions de transfert liées à la tête (HRTFs). Cet algorithme a été évalué expérimentalement sur une tâche de localisation binaurale de sources sonores en environnement bruité. La capacité de spécifier les modèles spectro-temporel et spatial pour chacune des sources sonores montre une grande expressivité de l'algorithme pour s'adapter à des problèmes de séparation et de localisation variés.

Ces contributions ont été évaluées dans le cadre restreint de mélanges linéaires instantanés de sources sonores de parole et de sources de bruit, ce qui constitue un premier pas pour l'évaluation d'algorithmes de séparation et de localisation dans des environnements réalistes. Méthodologiquement, nous avons créé par simulation une grande base de données d'enregistrements binauraux de sources sonores spatialisées. Néanmoins, il reste encore à confronter ces algorithmes pour des sources multiples (> 2), avec des mélanges convolutionnels correspondant à des environnements réverbérés, et à des environnements hybrides mélangeant des sources de bruit dans des environnements réverbérés. Enfin, il reste à généraliser la localisation spatiale à l'ensemble des coordonnées spatiales : l'azimut, l'élévation, et la distance. Pour aller plus loin dans la direction proposée, nous avons essayé de formuler la matrice de mélange binaurale à partir des a priori sur les dictionnaires de HTRFs supposés connus et une matrice d'activation de ces HTRFs au cours du temps. Malheureusement et de manière contre-intuitive, des évaluations expérimentales n'ont pas permis d'améliorer la localisation. Un examen des résultats semble indiquer que les HTRFs ne constituent de l'espace de perception binaurale les HTRFs ne seraient pas linéairement indépendants, et donc une position pourrait s'expliquer de manière équivalente par différentes combinaisons de HTRFs. Des formulations neuronales (Ma et al., 2018; Jiang et al., 2020; Liu et al., 2022) ont depuis été proposées pour répondre à ces problèmes, auxquels nous continuons de contribuer (Phokhinnan et al., 2023).

Dans la perspective de ce chapitre sur la perception de la parole humaine pour des tâches de séparation et de localisation de sources sonores, les contributions proposées permettent de proposer des éléments de réponse à la question de l'apprentissage **Q3**.

Q3. Quels modèles d'apprentissage? Les contributions proposées démontrent clairement l'intérêt de la spécification de modèles d'apprentissage, notamment à partir de la formalisation d'hypothèses physiques. Nous avons vu en particulier que le modèle source/filtre de signaux de parole et le modèle de HTRFs de l'écoute binaurale contribuent à rendre opérationnels des algorithmes d'apprentissage sur des tâches de séparation et de localisation de sources sonores, dans des environnements réalistes. A cet égard, la NMF présente un formalisme exemplaire dans le cadre de la règle de la mise-à-jour multiplicative qui rend possible l'intégration formelle explicite de modèles de signaux à l'intérieur de l'algorithme d'apprentissage, à condition de pouvoir les écrire sous forme matricielle ou tensorielle et de pouvoir en calculer les gradients sur les variables considérées.

La formalisation de ces contraintes permet de réduire l'espace des solutions à un sous-ensemble de solutions physiquement acceptables. Nous avons vu également que cette intégration pouvait bénéficier d'une intégration graduelle pour éviter de sur-spécifier dès l'initialisation ce sous-espace. Si des recherches ont contribué à rendre l'apprentissage des algorithmes NMF distribuables pour permettre le passage à l'échelle sur de grandes bases de données (Serizel et al., 2016), optimisables de manière similaire à un réseau de neurones (Mairal et al., 2012) et multi-couches (Le Roux et al., 2015), la linéarité de cet algorithme a montré ses limites en termes de capacité de généralisation pour des tâches de perception de signaux de sonores.

→ Nous arrivons à un moment déterminant dans mon parcours de recherche, qui marque entre 2015 et 2018 une transition du paradigme d'apprentissage vers les algorithmes d'apprentissage non-linéaire et en particulier les réseaux de neurones profonds. Ces algorithmes présentent des avantages importants, en particulier pour l'apprentissage de représentations et de relations complexes, et associent l'apprentissage de représentation et la prise de décision dans une seule optimisation. Par ailleurs, ils sont aussi bien utilisables pour des tâches de perception comme de production : la modélisation générative de signaux de parole humaine par réseaux de neurones est l'objet du dernier chapitre de ce manuscrit.

MODÉLISATION GÉNÉRATIVE DE SIGNAUX DE PAROLE PAR APPRENTISSAGE NEURONAL DE REPRÉSENTATIONS STRUCTURÉES

Le troisième et dernier chapitre de ce mémoire, qui couvre la période 2018 à 2022, est consacré à la production et en particulier de la génération de signaux de parole. Tout d’abord cette période marque un retour à mon activité de recherche principale et préférée : la synthèse (TTS, pour *Text-To-Speech synthesis*) et la transformation de la voix (VC, pour *Voice Conversion*), puis par extension aux comportements humains multimodaux comme la génération d’expressions faciales et de gestes en deux ou trois dimensions. Cet axe de recherche est un axe historique stratégique de l’équipe Analyse et Synthèse des sons (A/S) et de l’IRCAM depuis sa création en 1978, avec successivement et entre autre les travaux de Philippe Depalle, Xavier Rodet, Axel Roebel, Geoffroy Peeters, Thomas Hueber, Gregory Beller, Snorre Farnér, Pierre Lanchantin, Christophe Veaux, et Gilles Degottex. Ces travaux couvrent des contributions dans tous les domaines de la génération et transformation de la parole, incluant les modèles de signaux pour la synthèse et la transformation de la voix (Depalle and Poirot, 1991; Depalle et al., 1994; Peeters, 2001; Röbel and Rodet, 2005; Röbel, 2010; Degottex et al., 2013), la synthèse de la parole à partir du texte par corpus et modélisation statistique (Hueber, 2005; Lanchantin et al., 2010; Obin, 2011), la conversion statistique de l’identité de la voix (Villavicencio et al., 2009; Lanchantin and Rodet, 2010), ou de l’expressivité (Beller, 2009; Veaux and Rodet, 2011). La modélisation générative de comportements humains réalistes est une application particulièrement intéressante pour la modélisation de séquences. La parole humaine, comme les expressions faciales ou les gestes, représente une production humaine complexe utilisée dans le contexte de communication avec un ou plusieurs autres êtres humains. Les facteurs de variabilité sont multiples, couvrant les facteurs linguistique (le texte sémantique à communiquer), para-linguistique (toute information utile à la communication mais ne relevant pas de l’information sémantique, comme par exemple les émotions exprimées ou le style), et extra-linguistique (le reste des informations véhiculées mais a priori non utile à la communication, comme l’identité d’un locuteur, son âge ou son genre). Le problème principal réside dans le fait que l’ensemble de ces facteurs ne sont pas directement accessibles à l’observation, mais sont emmêlés dans le signal de parole. Qui plus est, l’influence de ces facteurs sur le signal de parole est complexe : entres autres, multi-paramétrique, non-linéaire, non-stationnaire, et multi-échelle. La modélisation générative nécessite de modéliser explicitement ou implicitement l’information associée à chacun de ces facteurs pour pouvoir ensuite les manipuler pendant la génération, à partir de l’observation statistique sur un ensemble de données. Si la déclaration de Dartmouth de 1956 conjecture que la machine doit pouvoir simuler tous les aspects de l’intelligence humaine (Minsky et al., 1955), la modélisation générative de comportements humains (langage, parole, expressions faciales, etc...) ou de productions humaines (par exemple, poésie, peinture, ou musique) en constitue en quelque sorte un paroxysme. Néanmoins, les capacités des machines ont longtemps été particulièrement limitées dans ce domaine. En l’occurrence, si les algorithmes ont égalé ou dépassé les capacités humaines sur de nombreuses tâches de perception (He et al., 2016), ce n’est en revanche que très récemment que des résultats similaires ont été observés sur des tâches de production. Cette limitation est multi-factorielle : les algorithmes de génération laissent toujours des traces caractéristiques dans le signal généré, perceptibles ou non perceptibles. Ces traces sont désignées comme des *artefacts*, c’est-à-dire étymologiquement “effets de l’art” ou encore “effets de l’artifice”.



FIGURE 3.1 – Visages générés par le styleGAN de Nvidia (Karras et al., 2019)



FIGURE 3.2 – Edmond de Belamy. Premier portrait peint par une IA. Collectif Obvious.

Les humains semblent particulièrement sensibles à la présence d'artefacts dans des comportements ou productions humaines, ce qui résulte généralement en une impression de manque de *réalisme* ou de *naturel*. Par ailleurs, le déterminisme et le manque de contextualisation des algorithmes sont souvent des indices de l'origine machinique de la simulation, incluant sa monotonie et sa prévisibilité ¹. D'Eliza en 1965 (Weizenbaum, 1966) à Chat-GPT en 2022 (OpenAI, 2022), les algorithmes génératifs ont réalisé des avancées spectaculaires pour aujourd'hui permettre des générations réalistes (parfois même ultra ou hyper-réaliste!) de comportement ou de production humaines. Si de telles réalisations ont été rendues possibles pour le langage (les désormais célèbres LLMs et autres chat-GPT) et pour la vision (Dall-E, MidJourney, et Stable Diffusion), qu'en est-il pour le son et en particulier pour la parole? C'est tout l'objet de ce chapitre qui présente mon cheminement personnel dans les évolutions radicales de ces dix dernières années, notamment par l'apprentissage de réseaux de neurones profonds.

La modélisation générative de la parole humaine est une tâche complexe car :

- Le signal de parole varie en fonction de nombreux facteurs (linguistique, paralinguistique, et extra-linguistique). En particulier, la variabilité linguistique est particulièrement importante, et dépend de la langue ;

¹ Ce sont les fameuses voix "robotiques", comme le vocodeur pour VOice enCODER inventé par Homer Dudley en 1939 en est à la fois la source et l'archétype (Dudley et al., 1939)

- Les signaux de parole sont multi-paramétriques, non-linéaires, non-stationnaires, et multi-échelles ;
- Cette variabilité n'est accessible qu'à travers l'observation de réalisations particulières du signal de parole dans lequel l'ensemble de ces facteurs et de leur influence sur l'observation acoustique sont emmêlés.

L'histoire moderne de la génération de la parole par des machines démontre la complexité de la tâche et la variété des approches proposées, mais aussi une évolution progressive des modèles de signaux et de la connaissance humaine vers un apprentissage de bout-en-bout, en particulier avec les réseaux de neurones profonds — et en parallèle de l'accroissement exponentiel des données disponibles et des capacités matérielles et logistiques des machines. Ce basculement s'articule en 3 ou 4 phases : depuis la synthèse de parole par corpus (Bailly, 1989; Black and Taylor, 1994; Fukada et al., 1994; Hunt and Black, 1996), à la modélisation statistique paramétrique par chaînes de Markov cachées (HMM), d'abord de paramètres séparés (Tokuda et al., 1995, 1999) puis de la représentation multi-paramétrique du signal de parole (Yoshimura et al., 1999; Tokuda et al., 2000, 2002), et un ensemble d'améliorations successives (Zen et al., 2007) sur la modélisation de la source d'excitation glottique (Yoshimura et al., 2001; Lanchantin et al., 2010), la modélisation dynamique des paramètres (Tokuda et al., 2003; Toda and Young, 2009), l'adaptation de la variance (Toda and Tokuda, 2007), ou encore la modélisation du style ou de l'expressivité (Tachibana et al., 2005; Yamagishi, 2006; Obin et al., 2011b). Une autre période transitoire a ouvert la voie à l'hybridation des modèles de signaux et d'apprentissage, d'abord entre synthèse par sélection d'unités et modélisation statistique (Boidin and Boffard, 2008; Veaux et al., 2010; Obin, 2011; Obin et al., 2012, 2015), puis par intégration de modèles par réseaux de neurones (Zen et al., 2013) dans la synthèse de parole par HMM, pour finalement basculer dans la phase actuelle de la modélisation intégrale par réseaux de neurones (Van den Oord et al., 2016; Wang et al., 2017b; Shen et al., 2018).

Dans cette perspective, la modélisation générative de la parole humaine soulève les trois questions de recherche suivantes :

Q1. Quelles représentations ? Quels modèles de signaux utilisés pour représenter les signaux de parole ? Les représentations multi-paramétriques comme le modèle source/filtre ou similaire utilisés par des vocodeurs comme SuperVP (Röbel, 2010) ou STRAIGHT (Kawahara, 2006) présentent l'avantage de proposer un modèle des signaux de parole sous la forme de paramètres interprétables en termes acoustique et fonctionnel (e.g., F0, gain, bruit, enveloppe spectrale), et dont la manipulation est physiquement intuitive. Néanmoins, ces représentations historiques fondées sur un modèle de signaux présentent des limitations très importantes pour la modélisation générative. En particulier, les paramètres présentent entre eux des inter-corrélations complexes, la modification de l'un des paramètres nécessitant la modification de l'ensemble des autres paramètres de manière coordonnée et cohérente. En l'occurrence, la manipulation séparée ou non cohérente introduit rapidement une dégradation du signal de parole, jugé non réaliste. Pour donner un exemple concret, la manipulation de la F0 seule ne reste réaliste que dans une plage de valeur limitée autour de la F0 observée, se fondant sur une hypothèse d'invariance des autres paramètres, une hypothèse dont le domaine de validité est très limité. Non seulement, les résonances du conduit vocal changent avec la F0, mais la qualité vocale également comme par exemple dans l'apparition de souffle et de craquement dans la voix pour des accents bas généralement observés en fin de phrase. Quelles(s) autre(s) représentation(s) adopter ? Pourquoi ?

→ Nous verrons notamment qu'avec l'émergence des vocodeurs neuronaux et les capacités d'apprentissage des réseaux de neurones, il devient envisageable d'apprendre à partir de représentations compressées perceptivement comme les mel-spectrogrammes.

Q2. Quelles données ? Les bases de données appaillées ou parallèles sont extrêmement limitantes pour l'apprentissage. Pendant longtemps, le paradigme de la conversion de voix supposait l'existence de bases de données parallèles dans lesquelles les mêmes énoncés étaient soit prononcés par des locuteurs différents pour la conversion d'identité, soit prononcés par un même locuteur avec des émotions différentes pour la conversion des émotions. La construction de telles bases de données présente l'avantage de contrôler la variabilité linguistique des données et facilitent l'apprentissage de conversions. En revanche, elles sont extrêmement coûteuses en temps humains et par conséquent limitées en taille. Cette limitation empêche d'exploiter pleinement les capacités d'apprentissage et de généralisation des réseaux de neurones. En outre, les capacités de généralisation sont extrêmement limitées par la variabilité linguistique réduite observée dans ces bases. En outre, la recherche en traitement automatique de la parole souffre d'un manque de données exploitables pour l'apprentissage, et ce problème se démultiplie par le nombre de langues existantes dans le monde numérique. Quand l'apprentissage de grands modèles de langage ou de modèles génératifs pour l'image exploitent respectivement 300 milliards de mots (Chat-GPT 3.5 et 4.0, 2023) et des millions ou dizaines de millions d'images (DALL-E 2.0), les bases de données de parole se sont longtemps restreintes à une dizaine ou centaine d'heures, aujourd'hui 50,000 à 60,000 heures pour les plus grandes bases connues (Wang et al., 2023; Betker, 2023). Doit-on penser le formalisme neuronal pour apprendre à partir de peu de données ? Peut-on s'émanciper de cette contrainte ? Comment apprendre une conversion sans exemples appaillés ou parallèles ?

Q3. Quel paradigme d'apprentissage ? Dans un premier temps : comment intégrer explicitement des informations au cours de l'apprentissage ? Dans un second temps : comment apprendre des conversions à partir de données à la volée ? Les réseaux de neurones, par leur expressivité et par leur capacité à modéliser efficacement des distributions arbitrairement complexes (Hornik et al., 1989) ont permis de réaliser des avancées importantes dans les domaines de la génération, en particulier pour la synthèse de la parole à partir du texte (Shen et al., 2018) et pour la conversion de la voix (Kaneko and Kameoka, 2017; Kameoka et al., 2018; Kaneko et al., 2019; Kaneko et al., 2019), et ouvert de nouvelles problématiques de recherche jusqu'alors inimaginables, comme le transfert de style (Wang et al., 2018; Pan and He, 2021) ou la synthèse de parole multilingue (Zhang et al., 2019). Néanmoins, les paradigmes d'apprentissage ont été longtemps contraints par les données disponibles pour l'apprentissage. En particulier, les paradigmes d'apprentissage ont longtemps été contraints par l'accès à des bases de données parallèles. Ces bases présentaient l'avantage de contrôler la variabilité linguistique pour l'apprentissage (Stylianou et al., 1998) mais en limitaient grandement les capacités de généralisation — notamment à cause de la faible quantité de données parallèles disponibles. Les paradigmes récents d'apprentissage permettent désormais d'apprendre à partir de données non contraintes, mais nécessitent en contre-partie une plus grande spécification des modèles d'apprentissage — typiquement pour encoder l'information linguistique désormais non contrôlée dans les données. Il semble ainsi exister une relation entre le degré de spécification du modèle d'apprentissage et le degré de spécification des données disponibles pour l'apprentissage. Cette relation peut se formuler de la manière suivante : plus les données sont contraintes, plus le modèle est libre ; et moins les données sont contraintes, plus le modèle doit être spécifié. Que ce soit dans les données ou le modèle, l'information doit être spécifiée pour pouvoir être apprise efficacement.

→ Nous verrons notamment l'évolution des paradigmes d'apprentissage formulés comme des problèmes de traduction (Le Moine, 2023) ou d'adaptation de domaine (Le Moine et al., 2021b) pour la conversion de l'identité ou de l'expressivité à partir de bases de données parallèles à des paradigmes d'auto-encodeurs à partir de données à la volée. La capacité d'apprendre des modèles génératifs de conversion à partir de données à la volée ouvrirait de nouvelles possibilités pour la recherche, en s'affranchissant des limitations liées à la structure des données utilisées pour l'apprentissage. Le paradigme de l'auto-encodeur,

qui consiste en la formulation neuronale du codage de données, pose néanmoins deux problèmes principaux pour la conversion de la parole : comment démêler efficacement les informations du signal de parole (contenu linguistique, identité, émotions, etc...) dans des codes distincts ? Comment encoder efficacement ces informations pour que leur manipulation soit effective en conversion ?

Le reste de ce chapitre présente le cheminement de ma recherche dans le domaine de la modélisation générative de la parole humaine, depuis les représentations structurées du signal de parole à l'apprentissage de représentations structurées par réseaux de neurones. Dans un premier temps, je présente les travaux menés sur la modélisation de la F0 pour la conversion de l'expressivité. En particulier, je présenterai l'une des premières formalisations d'une architecture neuronale séquence-à-séquence (S2S) (Robinson et al., 2019) pour une tâche de conversion de la voix ; puis l'apprentissage de représentations multi-niveaux de la F0 par apprentissage neuronal de filtres d'ondelettes CWT (Le Moine et al., 2021b). Dans un second temps, je présente le changement d'une représentation signal paramétrique (par exemple, F0) à une représentation signal non-paramétrique (en l'occurrence, le spectrogramme en échelle mel) pour l'apprentissage de la conversion de la voix. En particulier, je présente une architecture neuronale avec spécification de l'information linguistique (Le Moine, 2023) pour préserver le contenu linguistique au cours de la conversion et améliorer les capacités de généralisation de la conversion à des énoncés non vus à l'apprentissage. Dans un dernier temps, j'introduis le principe de la conversion neuronale fondée sur une architecture d'encodeur-décodeur et l'apprentissage adversarial de représentations structurées (identité, contenu, genre/âge), avec application à la conversion de l'identité (Bous et al., 2022) et du genre (Benaroya et al., 2023). Ma recherche sur la modélisation générative de signaux de parole a commencé en 2018 avec le stage de Carl Robinson (Robinson et al., 2019). Elle a ensuite été portée principalement par le projet ANR TheVOICE (2017-2021) sur la conversion neuronale de l'identité vocale menée avec Laurent Benaroya (Bous et al., 2022; Bengio, 2023); et par la thèse de Clément Le Moine - Veillon (2019-2023) (Le Moine, 2023) dans le cadre du projet AI4IDF MoVE. Enfin, elle s'est récemment étendue à la génération multimodale de comportements humains à partir du signal de parole avec la thèse de Mireille Fares (2019-2023) (Fares, 2023). Si les contributions de ces recherches sont principalement algorithmiques, ces recherches ont par ailleurs contribué à créer des ressources matérielles librement accessibles comme la base de données Att-HACK (?), et méthodologiques, par exemple pour la validation écologique de bases de données (Salais et al., 2022; Le Moine, 2023).

Projets associés à ce chapitre (par ordre chronologique)
Projet ANR TheVoice (2017-2021) Projet IDF MoVE (2019-2022) Projet ANR ARS (2019-2023)
Encadrements associés à ce chapitre (par ordre chronologique de fin)
Laurent Benaroya (2020-2021), post-doc, projet ANR TheVoice Clément Le Moine (2019-2022), thèse de doctorat, financement projet MoVE, école doctorale EDITE Mireille Farès (2019-2022), thèse de doctorat, co-encadrement avec Catherine Pelachaud (ISIR, CNRS, Sorbonne Université, bourse SCAI, école doctorale EDITE Léane Salais (2021-2024), thèse de doctorat, bourse ministère de la recherche, école doctorale EDITE Pascal Pham (2017, M2 ATSI, Université Paris Sud), Carl Robinson (2018, M2 Télécom Paris Sud), Clément Lemoine (2019, M2 ATIAM), Yujia Yang (2020, M2 ATIAM), Léane Salais (2021, M2 DAC), Luc Sterkers (2022, M2 DAC), Théodor Lemerle (2022, M2 ATIAM)
Publications associées à ce chapitre (par ordre chronologique)
Robinson, C., Obin, N., and Roebel, A. (2019). Sequence-to-Sequence Modelling of Fo for Speech Emotion Conversion. In <i>IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)</i> , pages 6830–6834. Le Moine, C. and Obin, N. (2020). Att-HACK : an Expressive Speech Database with Social Attitudes. In <i>Speech Prosody</i> . Ferro, R., Obin, N., and Roebel, A. (2021). CycleGAN Voice Conversion of Spectral Envelopes using Adversarial Weights. In <i>European Signal Processing Conference (EUSIPCO)</i> . Fares, M., Pelachaud, C., and Obin, N. (2021). Multimodal Modeling of Expressiveness for Human-Machine Interaction. In <i>WACAI</i> . Le Moine, C., Obin, N., and Roebel, A. (2021). Towards End-to-End Fo Voice Conversion based on Dual-GAN with Convolutional Wavelet Kernels. In <i>European Signal Processing Conference (EUSIPCO)</i> , pages 36–40. Le Moine, C., Obin, N., and Roebel, A. (2021). Speaker Attentive Speech Emotion Recognition. In <i>Interspeech</i> . Bous, F., Benaroya, L., Obin, N., and Roebel, A. (2022). Voice Reenactment with Fo and Timing Constraints and Adversarial Learning of Conversions. In <i>European Signal Processing Conference (EUSIPCO)</i> . Fares, M., Pelachaud, C., and Obin, N. (2022). Transformer Network for Semantically-Aware and Speech-Driven Upper-Face Generation. In <i>European Signal Processing Conference (EUSIPCO)</i> . Salais, L., Arias, P., Le Moine, C., , Rosi, V., Teytaut, Y., Obin, N., and Roebel, A. (2022). Production Strategies of Vocal Attitudes. In <i>Interspeech</i> . Benaroya, L., Obin, N., and Roebel, A. (2023) Manipulating Voice Attributes by Adversarial Learning of Structured Disentangled Representations, in <i>Entropy</i> , 25(2), 375, February 2023 Fares, M., Grimaldi, M., Pelachaud, C., and Obin, N. (2023). Zero-Shot Style Transfer for Gesture Animation driven by Text and Speech using Adversarial Disentanglement of Multimodal Style Encoding. In <i>Frontiers in Artificial Intelligence</i> , en cours de publication, 2023. Fares, M., Obin, N. and Pelachaud, C. (2023) Behavior Generation Model for Socially Interactive Agents, In <i>Gesture and Speech in Interaction (GeSpin)</i> , Nijmegen, Netherlands.

3.1 MODÉLISATION GÉNÉRATIVE DE LA F0 POUR LA CONVERSION VOCALE

La fréquence fondamentale (F0) est sans aucune doute le paramètre de la prosodie le plus étudié, et dont l'importance des fonctions linguistiques (Delattre, 1966; Pierrehumbert, 1980; Pierrehumbert and Hirschberg, 1990; Hirst and Di Cristo, 1998; Lacheret-Dujour and Beaugendre, 1998) (pour l'anglais-américain et le français), para- et extra- linguistiques

(Scherer et al., 1991; Léon, 1993; Ohala, 1994, 1996; Campbell and Mokhtari, 2003) (en stylistique, en psychologie, ou en éthologie) a été largement mise en évidence dans la littérature à partir de la seconde moitié du 20^{ème} siècle. Pour cette raison, un grand nombre de recherches ont visé à proposer des modèles de signaux depuis la représentation des courbes de F0 (Maeda, 1974; Fujisaki, 1981) jusqu'à leur codage symbolique (Beckman and Ayers, 1997; Campione et al., 2000; Post et al., 2006; Obin et al., 2014a) à l'origine des théories de l'intonation et de la phonétique et phonologie en linguistique. Ces recherches ont soulevé des problèmes importants relatifs à la modélisation de la F0. D'une part, la F0 est une grandeur non-nécessairement définie à chaque trame d'un signal audio : elle ne prend une valeur que pour les trames dites voisées, c'est-à-dire correspondant à une variation quasi-périodique de la source d'excitation glottique ; elle n'est pas définie pour les trames dites non-voisées. Par ailleurs, la F0 instancie des fonctions linguistique et para-linguistiques sur différents domaines temporels ², ce qui a donné lieu à l'émergence de modèles paramétriques multi-linéaires. Dans la littérature générative, la modélisation de la F0 se base sur tout ou partie de la chaîne de traitement suivante : 1) La séquence observée de F0 est pré-traitée pour créer une séquence de valeur de F0 définie à chaque index temporel, par exemple par interpolation entre les segments voisés successifs ; 2) Une stylisation des courbes de F0, c'est-à-dire une réduction de sa variabilité sur des courbes "simplifiée" fonctionnellement équivalentes, c'est-à-dire sans modifier le sens du message véhiculé. Cette stylisation repose soit sur des équivalences perceptives comme la JND ou autres métriques issues de la psycho-acoustique ('t Hart et al., 1990; d'Alessandro and Mertens, 1995; Mertens, 2004, 2013), ou alors sur des modèles de signaux comme la décomposition de la F0 sur des bases de fonctions lentement variable en temps. Ces fonctions sont soit pré-définies formellement, par exemple à partir de la Transformée en Cosinus Discrète (DCT) (Teutenberg et al., 2008) ou apprises sur des bases de données (Obin and Beliã, 2018) ; 3) Une modélisation paramétriques linéaire (Hirst and Espesser, 1993; Fukada et al., 1994) ; multi-échelle, par exemple sous la forme de modèles multi-linéaires (Fujisaki, 1981; Taylor, 1998; Mishra et al., 2006; Branislav et al., 2018) ou des représentations adaptatives comme les ondelettes par exemple avec la Transformée Continue en Ondelettes (CWT) (Luo et al., 2017; Sisman and Li, 2018) ou des réseaux de neurones récurrents multi-couches (Wang et al., 2017a). L'ensemble de ces étapes, en partie ou en totalité, ont été largement utilisées sur des tâches de modélisation générative des signaux de parole, comme la synthèse de parole à partir du texte (TTS) (Fukada et al., 1994; Black et al., 2007; Latorre and Akamine, 2008; Obin et al., 2011a; Obin, 2011; Wang et al., 2017a, 2018) ou la conversion de la voix (VC) (Veaux and Rodet, 2011; Ming et al., 2016; Yin et al., 2016; Wang et al., 2017a; Sisman et al., 2017; Sisman and Li, 2018; Kaneko and Kameoka, 2018). Cette première section présente mes recherches principales sur la thématique de la modélisation de la F0 (Obin et al., 2014a; Obin and Beliã, 2018) et en particulier sur la conversion neuronale de la F0 (Robinson et al., 2019; Le Moine et al., 2021b).

² Syllabe, mot, phrase prosodique, ou des unités latentes ne correspondant pas nécessairement à des unités linguistiques bien définies... La définition et l'inventaire de ces domaines ne font pas consensus encore aujourd'hui dans la communauté linguistique. Les unités comme les entités symboliques sont l'objet de débats animés et toujours actuels. Par exemple, un article de référence sous forme de critique des représentations phonologiques de l'intonation et de sa transcription : ToBI or not ToBI? (Wightman, 2002)!

Conversion neuronale de la F0 par modélisation séquence-à-séquence

La première contribution en termes de modélisation générative de la F0 pour la conversion de la voix a été de proposer l'une des premières formulations du problème de la conversion sous la forme d'une architecture neuronale séquence-à-séquence (Robinson et al., 2019), c'est-à-dire comme un problème de traduction de la F0 d'une modalité expressive à une autre modalité expressive. Les modèles séquence-à-séquence (S2S) constituent un sous-ensemble des algorithmes génératifs de modélisation de séquence, comme les réseaux de neurones récurrents (RNN) ou les LSTM (LSTM) ou les transformeurs, sont aujourd'hui régulièrement utilisés pour la synthèse de parole à partir du texte (Fan et al., 2014; Zen, 2015; Zen and Sak, 2015; Ronanki et al., 2016; Wang et al., 2017a) ou la conversion de la voix (Ming et al., 2016; Li et al., 2016). Les modèles S2S en constituent une extension dans la mesure où les deux séquences considérées sont de longueur variables et généralement différentes : une séquence observée en entrée, et une séquence à prédire en sortie conditionnellement

à la séquence d'entrée. Le problème S2S est formulé sous la forme d'un encodeur de la séquence d'entrée, et d'un décodeur de la séquence de sortie conditionnellement aux états de l'encodeur. Dans le cas de réseaux de neurones, l'encodeur et le décodeur prennent la forme d'un réseau de neurones récurrent. Historiquement, le décodeur était conditionné uniquement au dernier état de l'encodeur récurrent supposé transmettre toute l'information de la séquence d'entrée (Sutskever et al., 2014a), avant d'être remplacé par un mécanisme d'attention linéaire permettant d'exploiter dynamiquement lors du décodage l'information transmise par l'ensemble des états de l'encodeur (Bahdanau et al., 2015). Ces modèles s'appliquent pour de nombreux problèmes comme historiquement en traitement automatique du langage naturel pour la traduction de texte (Sutskever et al., 2014a), avant de s'étendre vers d'autres domaines comme les signaux audio et la parole (Wang et al., 2017b; Wan et al., 2017), par exemple par transduction neuronale de la modalité texte dans la modalité son. Cette architecture S2S a permis de repenser la tâche de conversion de la voix comme un problème de traduction acoustique d'une modalité expressive vers une autre ³. En outre, la conversion neuronale permet non seulement de traduire la F0 d'une modalité expressive à une autre, mais également de dilater/compresser dynamiquement les contours de F0 et par conséquent les durées et le rythme associé, considéré comme la deuxième dimension acoustique de la prosodie.

³ Sous la contrainte de préserver certaines informations contenues dans la voix d'origine, comme le contenu linguistique. Nous reviendrons dessus plus tard dans ce manuscrit.

Positionnement du problème et formulation neuronale

Soit $\mathbf{x} = [x_1, \dots, x_{T_x}]$ la séquence de F0 correspondant au signal de parole dans une émotion source (par défaut, l'émotion "neutre" ⁴), et $\mathbf{y} = [y_1, \dots, y_{T_y}]$ la séquence de F0 correspondant au signal de parole dans une émotion cible, avec T_x la longueur de la séquence source et T_y la longueur de la séquence cible.

Le problème de la modélisation S2S consiste à calculer la probabilité de la séquence de F0 de l'émotion cible \mathbf{y} conditionnellement à la séquence de F0 de l'émotion source \mathbf{x} ,

$$p(\mathbf{y}|\mathbf{x}) \quad (3.1)$$

Par applications successives du théorème de Bayes, cette probabilité peut s'exprimer sous la forme,

$$p(\mathbf{y}|\mathbf{x}) = \prod_{t=1}^{T_y} p(y_t | \mathbf{y}_{<t}, \mathbf{x}) \quad (3.2)$$

Dans (Cho et al., 2014), la résolution de ce problème passe par la décomposition de l'Équation (3.2) en deux sous-problèmes intermédiaires,

$$\mathbf{c} = q(\mathbf{x}) \quad (3.3)$$

$$p(\mathbf{y}|\mathbf{x}) = p(\mathbf{y}|\mathbf{c}) = \prod_{t'=1}^{T_y} p(y_{t'} | \mathbf{y}_{<t'}, \mathbf{c}) \quad (3.4)$$

où : \mathbf{c} est un vecteur de dimension fixe d_h et q une fonction de $\mathbb{R}^{d_x \times T_x}$ dans \mathbb{R}^{d_h} , d_x et T_x respectivement la dimension et la longueur de la séquence \mathbf{x} . L'Équation (3.3) correspond à l'encodage de la séquence d'entrée, et l'Équation (3.4) correspond au décodage de la séquence de sortie conditionnellement à l'encodage de la séquence d'entrée. Le tout prend la forme d'un auto-encodeur séquentiel.

La partie encodeur peut être exprimée sous la forme d'un réseau de neurones récurrent,

$$\mathbf{h}_t = f(\mathbf{x}_t, \mathbf{h}_{t-1}) \quad (3.5)$$

$$\mathbf{c} = \mathbf{h}_{T_x} \quad (3.6)$$

⁴ L'émotion "neutre" est une abstraction pratique analogue du chiffre zéro en arithmétique ou en algèbre, avec lequel il partage une définition négative. Largement remis en question dans les communautés de l'informatique affective, le neutre ne correspond à aucune réalité biologique, psychologique, ou acoustique

où : \mathbf{h}_t est l'état latent de l'encodeur récurrent à l'instant t de dimension d_h , f est la fonction d'activation de l'encodeur récurrent. L'Équation (3.6) transcrit l'hypothèse selon laquelle le dernier état de l'encodeur récurrent contient toute l'information pertinente de la séquence d'entrée \mathbf{x} .

La partie décodeur peut également être exprimée à l'aide d'un réseau de neurones récurrent,

$$p(\mathbf{y}'_t | \mathbf{y}_{<t'}, \mathbf{x}) = g(\mathbf{y}'_t, \mathbf{h}'_t, \mathbf{c}) \quad (3.7)$$

où : \mathbf{h}'_t , est l'état latent du décodeur récurrent à l'instant t' de dimension d'_h , et g est la fonction d'activation du décodeur récurrent.

L'introduction de mécanismes d'attention (Bahdanau et al., 2015) permet de calculer le contexte encodé \mathbf{c} tel qu'exprimé par l'Équation (3.6) de manière dynamique sous la forme d'une combinaison linéaire des états latents de l'encodeur. Dans cette formulation, le contexte \mathbf{c} est cette fois-ci fonction de l'instant t' du décodeur,

$$\mathbf{c}_{t'} = \sum_{t=1}^{T_x} \alpha_{t,t'} \mathbf{h}_t \quad (3.8)$$

où : $\alpha_{t,t'}$ sont les poids d'attention entre l'état latent de l'encodeur \mathbf{h}_t à l'instant t et l'état latent du décodeur $\mathbf{h}'_{t'}$ à l'instant t' .

Le mécanisme d'attention permet d'identifier les informations pertinentes pour la prédiction à l'instant t' de la séquence de sortie. Une illustration de l'architecture S2S appliquée au problème de la conversion de la F_0 est présentée sur la Figure 3.3, et un exemple de conversion sur la Figure 3.4.

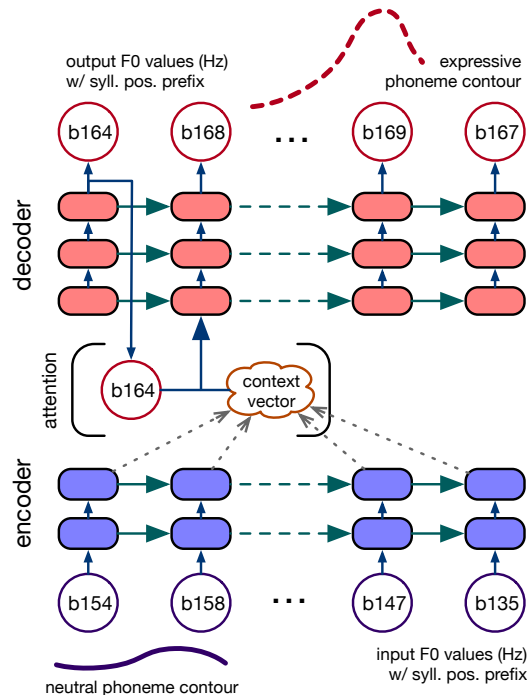


FIGURE 3.3 – Architecture S2S avec mécanisme d'attention pour la conversion de la F_0 . Source : (Robinson et al., 2019)

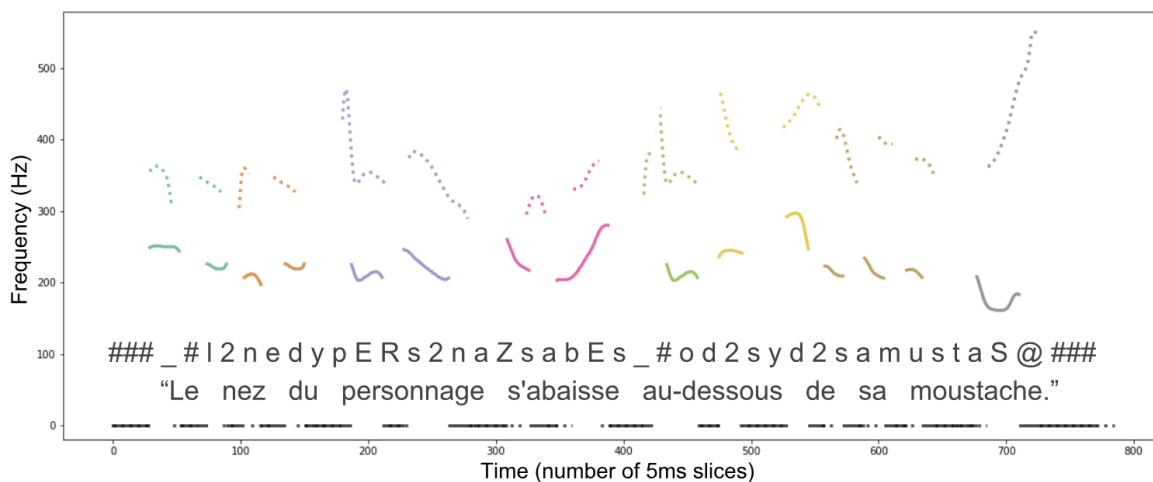


FIGURE 3.4 – Contours de F0 originaux en traits pleins (source=neutre) et convertis en traits pointillés (cible=joie) pour l'énoncé “Le nez du personnage s'abaisse au-dessous de sa moustache”.
Source : (Robinson et al., 2019)

Détails d'implémentation

Dans cette étude, l'architecture S2S a été employée pour convertir les contours de F0 sur les noyaux vocaliques de chaque syllabe, avec ajout d'informations contextuelles supplémentaires pour différencier les syllabes. La F0 est préalablement quantifiée entre 50 Hz et 550 Hz et encodée sous forme de vecteur un-chaud en entrée et en sortie du réseau. L'ensemble des détails d'implémentation de l'architecture proposée sont décrits dans (Robinson et al., 2019).

Expérience : conversion de neutre vers expressif

L'architecture proposée a été évaluée expérimentalement sur une base de données d'émotions en Français décrite dans (Veaux and Rodet, 2011). Cette base de données comporte un locuteur masculin et un locuteur féminin prononçant 10 phrases différentes dans les 4 émotions primaires (joie, peur, tristesse, colère) avec 5 niveaux d'intensité, et une version de contrôle “neutre”. Un modèle de conversion de F0 a été entraîné pour chaque pair de neutre à émotion.

Méthodologie

L'évaluation expérimentale a été réalisée sous la forme d'une expérience de perception. Un échantillon (réel ou transformé) était présenté au participant avec deux consignes : 1) Identifier l'émotion exprimée (choix forcé parmi les 4 émotions primaires), et 2) Juger du naturel de l'échantillon sur une échelle MOS à 5 degrés (mauvais, faible, passable, bon, excellent). Les participants devaient évaluer 20 échantillons, présentés dans un ordre aléatoire. L'expérience s'est déroulée sur une plate-forme en ligne, et les participants étaient encouragés à réaliser l'expérience dans un environnement silencieux et au casque. 87 participants ont contribué à cette étude.

Résultats et discussion

Le **Tableau 3.1** présente les taux de reconnaissance des émotions obtenus à partir des échantillons interprétés par les comédiens, et le **Tableau 3.2** ceux obtenus sur les phrases converties. Dans cette deuxième expérience, nous avons également étudié l'influence

de l'introduction d'un conditionnement linguistique de l'encodeur sur la position de la syllabe à l'intérieur de la phrase. Les taux de reconnaissance obtenus à partir des échantillons originaux montre (ou confirme) que la communication des émotions est loin d'être sans ambiguïté. Ainsi, la joie ou la tristesse sont des émotions clairement identifiées par les participants (respectivement, 91.3% et 85.7%), alors que la colère et la peur sont beaucoup plus ambiguës (respectivement, 70.3% et 58.9%). En particulier, ces deux émotions sont souvent confondues entre elles. Les sources possibles d'ambiguïté possibles sont nombreuses : de l'émission/expression à la réception/perception, en passant par des stratégies individuelles d'expression et des représentations mentales susceptibles de varier entre individus.

		Perceived				
		Joy	Sadness	Anger	Fear	
Original	Acted	Joy	91.3%	2.4%	6.3%	0.0%
		Sadness	4.4%	85.7%	0.0%	33.3%
		Anger	2.2%	2.4%	70.3%	7.8%
		Fear	2.1%	9.5%	23.4%	58.9%
		Total	100%	100%	100%	100%

TABLEAU 3.1 – Matrice de confusion obtenue à partir des réponses des participants pour les échantillons correspondant aux émotions actées par les comédiens.

Par comparaison, les taux de reconnaissance obtenus à partir des échantillons convertis sont plutôt satisfaisants. En particulier, la joie, la colère, et la tristesse sont reconnus de manière consistante (respectivement, 74.8%, 64.9%, et 50.1%), avec des taux comparables à ceux obtenus pour les émotions actées. La tristesse est moins bien reconnue et souvent confondue avec la peur.

		Perceived				
		Joy	Sadness	Anger	Fear	
Converted	with Cond.	Joy	74.8%	13.0%	9.2%	12.9%
		Sadness	17.6%	40.2%	18.5%	22.5%
		Anger	5.9%	13.6%	64.9%	14.5%
		Fear	1.7%	33.2%	7.4%	50.1%
		Total	100%	100%	100%	100%
Converted	w/o Cond.	Joy	67.8%	20.9%	30.7%	17.0%
		Sadness	19.4%	25.7%	7.7%	25.5%
		Anger	9.6%	25.5%	42.4%	21.3%
		Fear	3.2%	27.9%	19.2%	36.2%
		Total	100%	100%	100%	100%

TABLEAU 3.2 – Matrices de confusion obtenues à partir des réponses des participants pour les échantillons convertis. En haut : le modèle avec conditionnement linguistique ; en bas : le modèle sans conditionnement linguistique.

Pour ce qui est du naturel, les échantillons d'émotions actées originaux sont jugés d'un naturel entre bon et excellent (4.22 en moyenne, avec un écart-type de 0.84). Par comparaison, les échantillons convertis sont jugés entre passage et bon (3.40 en moyenne, avec un écart-type aux alentours de 1.0). Nous observons donc une dégradation non négligeable du naturel après conversion, qui peut être dû à des artefacts engendrés

soit par des erreurs de prédiction du modèle de conversion, soit par l'algorithme de synthèse utilisé (ici, le super vocodeur de phase (Röbel, 2010)). Il doit être noté que les deux métriques perceptives utilisées dans cette expérience ne sont clairement pas indépendantes : en particulier, la dégradation du signal engendre nécessairement une gêne pour la reconnaissance d'une émotion.

Pour le conditionnement linguistique, le modèle avec conditionnement améliore substantiellement les taux de reconnaissance. En particulier, l'exploitation du contexte linguistique apporte une amélioration de 7.0% pour la joie, de 13.9% pour la peur, de 14.5% pour la tristesse, et de 22.5% pour la colère en termes de taux de reconnaissance. L'interprétation de ces résultats est double : d'une part, la prise en compte du contexte souligne l'inter-dépendance des contours de F0 sur les syllabes mais surtout leur évolution dynamique sur des temporalités plus longues. Par ailleurs, ils montrent également l'efficacité de l'encodage du contexte proposé. Remis dans une perspective plus large des travaux en modélisation S2S, le conditionnement proposé est concomitant des conditionnements de position qui accompagneront l'émergence des architectures transformeurs présenté pour la première fois en 2018 (Vaswani et al., 2017). En conclusion, la première implémentation d'une tâche de conversion de voix à partir d'architecture S2S avec attention a été prometteuse. Néanmoins, l'hypothèse faite d'un modèle séquentiel linéaire montre aussi ses limites à modéliser les variations de F0 par essence multi-échelles.

Conversion neuronale de la F0 à partir de représentation multi-échelle en ondelettes

Pour palier les limitations de modèles séquentiels, nous nous sommes orientés vers des modélisations multi-échelles. Dans un premier temps, nous avons tenté de formuler une architecture neuronale hiérarchique de la F0 par superposition de couches récurrentes associées à des unités linguistiques bien déterminées (par exemple : syllabe, mot, phrase prosodique, etc...). Mais ces tentatives se sont révélées infructueuses, et n'ont pas donné suite. Nous avons alors décidé de redéfinir nos hypothèses de manière moins stricte, en laissant la liberté au réseau d'apprendre les unités temporelles porteuses d'information de F0 pertinente pour la modélisation et la conversion. Pour ce faire, nous nous sommes donc orientés vers la Transformée Continue en Ondelettes (CWT) qui a été introduite pour la représentation de la F0 sur des tâches de conversion (Ming et al., 2015, 2016; Luo et al., 2017; Sisman and Li, 2018; Luo et al., 2019; Zhou et al., 2020b). La CWT calcule une décomposition du signal de F0 sur des bases d'ondelettes, ce qui offre une représentation des variations de la F0 sur plus échelles temporelles (Ming et al., 2015). Une formulation de la CWT (Luo et al., 2017) a permis de calculer cette décomposition sur des échelles temporelles fixes (e.g., phonème, syllabe, mot, phrase prosodique, etc...). Enfin, un algorithme d'adaptation d'échelle (CWT-AS) (Luo et al., 2019) permet de sélectionner les échelles temporelles optimales pour la conversion, au sens de la maximisation de la distance moyenne entre les émotions dans l'espace de représentation de la CWT. Après sélection de ces échelles, un réseau dual-GAN est employé pour apprendre les conversions de F0 entre une émotion source et une émotion cible (Luo et al., 2019). Nous avons fondé notre recherche dans la lignée directe de ces travaux, avec deux contributions principales : 1) L'intégration d'un coût de reconstruction de la F0 pour la sélection des échelles ; et 2) L'optimisation conjointe des échelles de représentation et des conversions dans une seule architecture neuronale. Pour reprendre la terminologie utilisée en début de ce chapitre, la partie dédiée à la stylisation de la F0 par AS-CWT et la partie de conversion sont intégrées dans une seule architecture neuronale optimisée pour la conversion. La suite de cette section présente les détails de l'algorithme proposé et les résultats expérimentaux sur une tâche de conversion expressive de la voix.

Positionnement du problème et formulation neuronale

Soit \mathcal{X}^{src} et \mathcal{X}^{tgt} les ensembles de phrases respectivement associées à des expressivités src et tgt. À partir de cet ensemble, des paires de phrases source x^{src} et cible x^{tgt} sont sélectionnées, avec les séquences de F0 source et cible correspondantes. Chaque paire de phrase possède le même contenu linguistique, et est prononcée par le même locuteur : seule l'expressivité varie. La Figure 3.5 présente l'architecture neuronale proposée. Cette architecture est composée d'un pré-réseau dont la tâche est d'encoder la F0 des séquences source et cible, et d'un réseau principal dont la tâche est d'apprendre la conversion entre l'expressivité src et tgt.

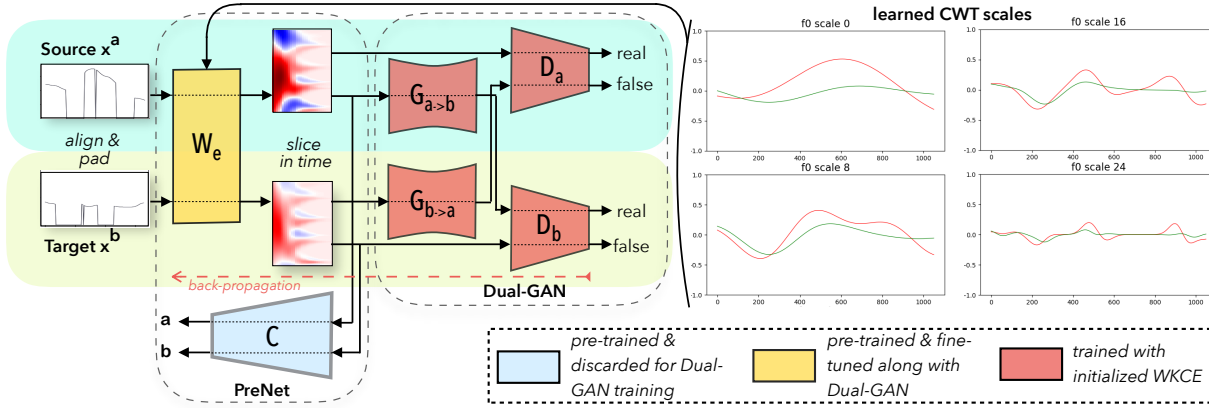


FIGURE 3.5 – Architecture neuronale pour l'encodage multi-échelle et la conversion de la F0. À droite : décomposition de la F0 sur 4 des bases d'ondelettes apprises pour les expressivité source (en rouge) et cible (en vert).

Préalablement à l'application de la CWT, les séquences de F0 sont linéairement interpolées en échelle logarithmique entre les segments voisins successifs pour obtenir une fonction x de F0 définie et continue à tout instant $t \in [1, T]$ d'une phrase. L'auto-encodeur opère des convolutions 1-D entre la séquence x de F0 tel que pour un noyau d'ondelette défini par l'ondelette mère $\psi_s \in \mathbb{R}^T$,

$$\psi_s = \frac{2\pi^{-\frac{1}{4}}}{\sqrt{3}} \left(1 - \left(\frac{t}{s}\right)^2\right) e^{-\frac{1}{2}\left(\frac{t}{s}\right)^2} \quad (3.9)$$

où : s est un paramètre d'échelle de l'ondelette.

La contribution \mathbf{h}_x^s de l'ondelette d'échelle temporelle s au signal de F0 x s'écrit alors comme la convolution entre x et ψ_s . La transformée WCKE résultante $W_e(x) \in \mathbb{R}^{N \times T}$ s'écrit alors,

$$W_e(x) = [\mathbf{h}_x^{s_0}, \dots, \mathbf{h}_x^{s_N}] \quad (3.10)$$

En notant W_r l'opération de transformée inverse, le signal reconstruit \hat{x} s'écrit alors,

$$\hat{x} = W_r(W_e(x)) = \frac{d_j \sqrt{d_t}}{C_d Y_0} \sum_{i=0}^{N-1} \mathbf{h}_x^{s_i} + \bar{x} \quad (3.11)$$

avec : \bar{x} le vecteur moyenne de x dans le temps, $d_t = 1.2$, $d_j = 0.125$, $C_d = 3.541$ et $Y_0 = 0.867$ (cf. (Torrence and Compo, 1998) pour une description détaillée du choix de ces valeurs).

Cette représentation multi-échelle de la F0 est utilisée en entrée d'une architecture neuronale générative fondée sur le transfert entre la distribution de l'expressivité source et la distribution de l'expressivité cible. Comme illustrée sur la Figure 3.5, l'architecture proposée est composée : d'une part, d'un pré-réseau dont l'objectif est d'apprendre

des représentations latentes pour l'encodage de la F0; et d'autre part, un réseau génératif antagoniste Dual-GAN dont l'objectif est d'apprendre la correspondance entre la distribution des expressivité source et cible, comme présenté dans (Xia et al., 2017). L'une des particularités de l'architecture proposée est que les bases d'ondelettes de la CWT sont explicitement optimisées de bout-en-bout dans l'architecture proposée, depuis l'encodage jusqu'au transcodage.

Le pré-réseau est formulé sous la forme d'un auto-encodeur des distributions des expressivités source et cible dont l'objectif est d'apprendre un encodage compact de la F0; et d'un classifieur dont l'objectif est d'apprendre un encodage différencié selon les expressivités. Pour ce faire, on définit une première fonction de perte de reconstruction \mathcal{L}_{rec} entre les signaux de F0 \mathbf{x}^a et \mathbf{x}^b échantillonnés à partir des distributions source et cible $P(\mathbf{x}^a)$ and $P(\mathbf{x}^b)$,

$$\begin{aligned} \mathcal{L}_{rec}(W_e) = & \mathbb{E}_{\mathbf{x}^a \sim P(\mathbf{x}^a)} (\|W_r(W_e(\mathbf{x}^a)) - \mathbf{x}^a\|_1) + \\ & \mathbb{E}_{\mathbf{x}^b \sim P(\mathbf{x}^b)} (\|W_r(W_e(\mathbf{x}^b)) - \mathbf{x}^b\|_1) \end{aligned} \quad (3.12)$$

où : \mathbb{E}_X représente l'espérance mathématique sur la variable aléatoire X , et $\|x\|_1$ la norme L_1 de x .

Une seconde fonction de perte de classification L_{cl} est définie dans l'espace latent W_e comme l'entropie croisée entre les expressivités source et cible prédites $\hat{a} = C(W_e(\mathbf{x}^a))$ et $\hat{b} = C(W_e(\mathbf{x}^b))$, et les vraies valeurs a et b des expressivités,

$$\begin{aligned} \mathcal{L}_{cl}(W_e) = & \mathbb{E}_{\mathbf{x}^a \sim P(\mathbf{x}^a)} [a * C(W_e(\mathbf{x}^a))] + \mathbb{E}_{\mathbf{x}^b \sim P(\mathbf{x}^b)} [b * C(W_e(\mathbf{x}^b))] \\ & + \mathbb{E}_{\mathbf{x}^b \sim P(\mathbf{x}^b)} (1 - a)[1 - \log(C(W_e(\mathbf{x}^a)))] \\ & + \mathbb{E}_{\mathbf{x}^a \sim P(\mathbf{x}^a)} (1 - b)[1 - \log(C(W_e(\mathbf{x}^b)))] \end{aligned} \quad (3.13)$$

La fonction de perte totale du pré-réseau \mathcal{L}_{pN} s'écrit alors,

$$\mathcal{L}_{pN}(W_e) = \alpha \mathcal{L}_{rec}(W_e) + \beta \mathcal{L}_{cl}(W_e) \quad (3.14)$$

où : α et β sont les poids de reconstruction et classification du pré-réseau. Le pré-réseau peut être pré-entraîné séparément pour servir d'initialisation à l'optimisation de l'ensemble de l'architecture proposée.

Le transfert de la F0 entre l'expressivité source et l'expressivité cible est appris par une architecture générative antagoniste Dual-GAN (Xia et al., 2017) à partir des représentations latentes encodées précédemment. Un réseau génératif antagoniste GAN (Goodfellow et al., 2014a) vise à approximer la distribution d'une variable aléatoire par le couplage d'un réseau générateur G dont l'objectif est d'échantillonner des données issues de la distribution, et d'un réseau discriminateur D dont l'objectif est de distinguer entre les données issues de la vraie distribution et les données générées par le générateur. Les deux réseaux sont optimisés alternativement avec des objectifs antagonistes dans un jeu à somme nulle.

L'architecture Dual-GAN est l'extension du principe du GAN pour le transfert entre les distributions d'un domaine source et d'un domaine cible. L'apprentissage de ce transfert est réalisé par l'optimisation de deux tâches duales : le transfert de la distribution du domaine source vers la distribution du domaine cible, et vice-versa. On note $G_{a \rightarrow b}$ et $G_{b \rightarrow a}$ les générateurs de la distribution du domaine source vers la distribution du domaine cible (respectivement, cible vers source), et D_a et D_b les discriminateurs des domaines source et cible.

Une première fonction de perte de transfert $\mathcal{L}_{a \leftrightarrow b}$ est définie comme l'erreur de transcodage entre la F0 prédite du domaine cible $G_{a \rightarrow b}(W_e(\mathbf{x}^a))$ à partir de la F0 du

domaine source \mathbf{x}^a et la vraie valeur de la F0 du domaine cible \mathbf{x}^b (respectivement, du domaine cible vers le domaine source),

$$\begin{aligned} \mathcal{L}_{a \leftrightarrow b}(G_{a \rightarrow b}, G_{b \rightarrow a}, W_e) = & \mathbb{E}_{(\mathbf{x}^a, \mathbf{x}^b) \sim \mathcal{P}(\mathbf{x}^a, \mathbf{x}^b)} (\|W_\tau(G_{a \rightarrow b}(W_e(\mathbf{x}^a))) - \mathbf{x}^b\|_1) \\ & + \mathbb{E}_{(\mathbf{x}^a, \mathbf{x}^b) \sim \mathcal{P}(\mathbf{x}^a, \mathbf{x}^b)} (\|W_\tau(G_{b \rightarrow a}(W_e(\mathbf{x}^b))) - \mathbf{x}^a\|_1) \end{aligned} \quad (3.15)$$

Une deuxième fonction de perte \mathcal{L}_{adv} est définie comme la perte antagoniste entre le générateur du domaine source vers le domaine cible $G_{a \rightarrow b}$ et le discriminateur du domaine cible D_b (et vice-versa),

$$\begin{aligned} \mathcal{L}_{adv}(G_{a \rightarrow b}, G_{b \rightarrow a}, D_a, D_b, W_e) = & \mathbb{E}_{\mathbf{x}^a \sim \mathcal{P}(\mathbf{x}^a)} [D_a(W_e(\mathbf{x}^a))] + \mathbb{E}_{\mathbf{x}^b \sim \mathcal{P}(\mathbf{x}^b)} [D_b(W_e(\mathbf{x}^b))] \\ & + \mathbb{E}_{\mathbf{x}^b \sim \mathcal{P}(\mathbf{x}^b)} [1 - \log(D_a(G_{b \rightarrow a}(W_e(\mathbf{x}^b))))] \\ & + \mathbb{E}_{\mathbf{x}^a \sim \mathcal{P}(\mathbf{x}^a)} [1 - \log(D_b(G_{a \rightarrow b}(W_e(\mathbf{x}^a))))] \end{aligned} \quad (3.16)$$

Une troisième fonction de perte \mathcal{L}_{dual} est définie comme,

$$\begin{aligned} \mathcal{L}_{dual}(G_{a \rightarrow b}, G_{b \rightarrow a}, W_e) = & \mathbb{E}_{(\mathbf{x}^a, \mathbf{x}^b) \sim \mathcal{P}(\mathbf{x}^a, \mathbf{x}^b)} (\|W_e(\mathbf{x}^a) * G_{a \rightarrow b}(W_e(\mathbf{x}^a)) \\ & - W_e(\mathbf{x}^b) * G_{b \rightarrow a}(W_e(\mathbf{x}^b))\|_1) \end{aligned} \quad (3.17)$$

La fonction de perte totale de l'architecture Dual-GAN s'écrit alors,

$$\begin{aligned} \mathcal{L}_{DG}(G_{a \rightarrow b}, G_{b \rightarrow a}, D_a, D_b, W_e) = & \lambda \mathcal{L}_{a \leftrightarrow b}(G_{a \rightarrow b}, G_{b \rightarrow a}, W_e) \\ & + \mathcal{L}_{adv}(G_{a \rightarrow b}, G_{b \rightarrow a}, D_a, D_b, W_e) \\ & + \gamma \mathcal{L}_{dual}(G_{a \rightarrow b}, G_{b \rightarrow a}, W_e) \end{aligned} \quad (3.18)$$

où : λ et γ sont les poids de transformation et dual du réseau Dual-GAN.

Détails d'implémentation

L'encodeur W_e est composé d'une couche convolutionnelle avec des filtres en ondelettes couvrant $N = 32$ échelles temporelles échantillonnées linéairement de d_{\min} ms. à d_{\max} ms. Le classifieur C est composé de couches convolutionnelles comprenant de 32 à 128 filtres de taille (3×3) avec des activations linéaires rectifiées, une couche de désactivation aléatoire avec un facteur de 0.2, une opération d'agrégation en 2-dimension utilisant un pas de dimension (2×2) et une agrégation maximale de dimension (4×4) . Ces couches sont suivies d'une couche d'aplatissement, d'une couche entièrement connectée de dimension 1,000 avec activation linéaire rectifiée, et d'une couche entièrement connectée de dimension 2 avec une activation exponentielle normalisée. L'architecture du Dual-GAN comprenant les générateurs $G_{a \rightarrow b}$ et $G_{b \rightarrow a}$, et les discriminateurs D_a et D_b reprennent l'implémentation décrite dans (Luo et al., 2019). L'ensemble des détails d'implémentation et d'apprentissage sont décrits dans (Le Moine et al., 2021a).

Expérience : conversion de l'expressivité

L'architecture proposée a été évaluée expérimentalement sur la base de données AttHACK (Le Moine and Obin, 2020) créée spécifiquement pendant la thèse de Clément Le Moine Veillon sur l'expression vocale et la conversion neuronale des attitudes sociales. Cette base de données comprend 25 locuteurs interprétant 100 phrases avec 4 attitudes sociales : amicale, distante, dominante, et séductrice, 3 à 5 répétitions pour chaque phrase, pour un total d'environ 30 heures de parole. L'expérience consiste en la comparaison de trois configurations : une architecture *baseline* et deux versions *config_A* et *config_B* de l'architecture proposée.

1. *baseline* L'architecture AS-CWT + Dual-GAN telle que présentée dans (Luo et al., 2019). Les échelles de CWT sont pré-sélectionnées par un algorithme d'échelles adaptatives (AS) en fonction des paires d'expressivité considérées ;
2. *config_A* Le pré-réseau CWT *We* est optimisé uniquement pour l'objectif de reconstruction ($\alpha = 1, \beta = 0$);
3. *config_B* Le pré-réseau CWT *We* est optimisé à la fois pour l'objectif de reconstruction de la F0 et de classification des attitudes ($\alpha = 1, \beta = 1$)

Dans les trois configurations, le même algorithme Dual-GAN est utilisé pour l'apprentissage de la conversion (Luo et al., 2019). Pour cette étude, des modèles dépendant du locuteur ont été utilisés pour l'expérimentation : une femme (Fo8) et un homme (Mo7).

En guise d'illustration préliminaire, la **Figure 3.6** présente les distributions des échelles temporelles sélectionnées par l'algorithme CWT-AS de base et pour les deux versions de l'algorithme proposé. L'émergence des échelles temporelles prosodiques à partir des données est en filiation directe avec les travaux initiés par (Holm, 2003) et prolongés par (Gerazov et al., 2018a,b). Nous pouvons observer deux tendances : tout d'abord, la configuration *config_B* présente une couverture temporelle plus étendue que les autres configurations que les configurations *baseline* et *config_A*. Cette observation est conforme avec l'hypothèse multi-échelle de la prosodie, couvrant aussi bien les micro-variations de l'échelle du phonème aux macro-variations des contours de phrase. Par ailleurs, les échelles temporelles associées à la configuration *config_B* apparaissent également plus variées en fonction des paires d'attitudes. Cette observation semble suggérer que ces échelles se différencient en fonction des attitudes considérées, chaque attitude se caractérisant par un ensemble d'échelles temporelles privilégiées.

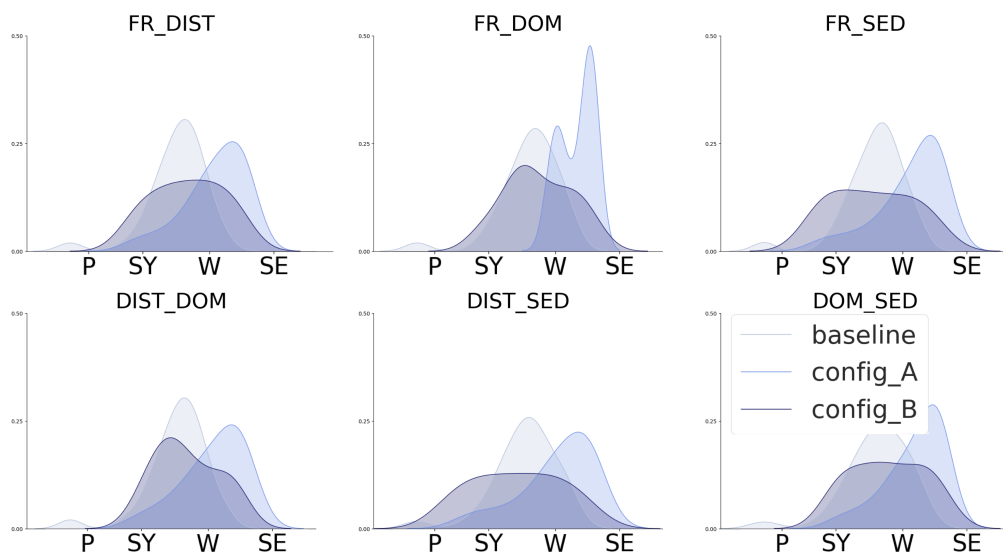


FIGURE 3.6 – Distribution des échelles temporelles CWT scale obtenues pour les trois algorithmes comparés : *baseline*, *config_A* et *config_B*. Ces distributions sont calculées pour les six paires possible de conversion d'attitude pour le locuteur Fo8, indiquée en haut de chaque sous-figure : FR pour amical, DOM pour dominant, SED pour séducteur, et DIS pour distant. Chaque sous-figure présente les distributions des échelles temporelles apprises pour les trois algorithmes comparés. Pour donner un repère temporel, les marqueurs P, SY, W et SE indiqués en abscisse correspondent respectivement aux durées moyennes des phonèmes, syllabes, mots et phrases.

Dans un premier temps, une évaluation objective a été conduite pour comparer le modèle *baseline* et les configurations *config_A* et *config_B* proposées. Les évaluations ont été menées pour l'ensemble des 4 paires d'attitudes, ce qui a résulté en 12 conversions (6 en sens direct, 6 en sens inverse). Un exemple de conversion est illustré sur la Figure 3.7, sur laquelle nous pouvons observer que la courbe de F0 générée par l'architecture proposée *config_B* est la plus proche de la F0 cible. Cette observation est confirmée quantitativement par les résultats de l'évaluation objective présentée ci-dessous. Le Tableau 3.3 présente les mesures de la racine carrée de l'erreur quadratique moyenne (RMSE, en Hz.) entre : 1) la F0 originale et la F0 reconstruite ; 2) la F0 convertie et la F0 correspondant à l'attitude cible. En termes de reconstruction, nous observons que les architectures proposées permettent d'encoder des représentations de la F0 plus précises par rapport à l'architecture *baseline* qui ne comprend pas de fonction de perte de reconstruction. Cette contrainte est essentielle dans la mesure où la précision de la reconstruction de la F0 est essentielle pour assurer un rendu fidèle et naturel ⁵. D'autre part, nous observons que l'erreur de reconstruction est plus grande dans le cas de la *config_B* que de la *config_A* : c'est un résultat attendu puisque seule la reconstruction est optimisée dans la *config_A* tandis que la *config_B* optimise les deux objectifs de reconstruction et de classification. Les mêmes tendances sont globalement observées pour la tâche de transformation, avec une augmentation de l'erreur pour l'ensemble des configurations. Cette fois-ci, la *config_B* présente la meilleure performance générale, ce qui indique que l'intégration de la classification dans l'encodage de la F0 permet d'obtenir des codes mieux structurés et plus efficaces pour la conversion.

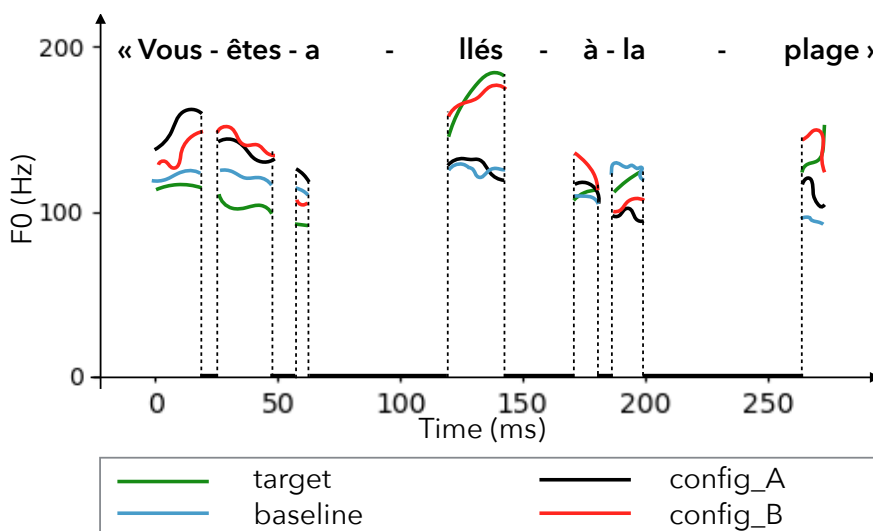


FIGURE 3.7 – Exemple de conversion de la F₀ depuis une attitude distante à dominante pour le locuteur Mo7 de la base de données Att-HACK. Les courbes de couleurs différentes indiquent la courbe de F₀ dans l'attitude cible, et les autres couleurs les conversion de F₀ obtenues avec l'algorithme *baseline* et les algorithmes proposés *config_A* et *config_B*.

Dans un second temps, nous avons conduit une évaluation perceptive pour comparer l'architecture *baseline* avec les deux configurations proposées. Le protocole utilisé pour cette évaluation propose une évolution par rapport à l'expérience réalisée dans la section précédente, en substituant à une expérience de reconnaissance par une expérience de préférence à une cible (test XAB). Ce choix méthodologique permet de construire une mesure de similarité entre les expressivités des conversions et de la cible, sans nécessiter de verbaliser la perception du sujet, typiquement par l'intermédiaire d'une catégorie d'émotion ou d'attitude. Le protocole consiste à présenter au participant deux enregistrements correspondant à une conversion réalisée par l'architecture *baseline* et une conversion réalisée

⁵ La perception humaine de la hauteur et de ses variations est complexe. En particulier pour la perception de la hauteur dans la parole, on distingue généralement une écoute "linguistique" utilisée pour distinguer des unités phonologiques du langage ('t Hart, 1981), comme des accents ou des modalités ; et une écoute purement "musicale" c'est-à-dire uniquement concernée par les variations mélodiques. Pour l'écoute musicale, le seuil de discrimination est de 4 cents pour des tons purs et statiques ('t Hart et al., 1990), et il existe des seuils équivalents pour les glissandi (D'Alessandro et al., 1998). Pour une conversion transparente mélodiquement, les algorithmes de conversion de la F0 doivent donc viser une précision de l'ordre de ou inférieure ou égale à ces seuils. Je remercie Christophe d'Alessandro pour cette communication.

Models	RMSE (Hz)	
	Reconstruction	Transformation
<i>baseline</i>	17.32	21.71
<i>config_A</i>	9.16	19.15
<i>config_B</i>	13.4	18.83

TABLEAU 3.3 – Comparaison de la RMSE obtenue pour les architectures *baseline*, *config_A* and *config_B* sur les tâches de reconstruction et de conversion.

par l'une des deux configurations *config_A* ou *config_B*. Le participant doit alors juger lequel des deux enregistrements est le plus similaire avec un enregistrement réel de l'attitude cible. La [Figure 3.8](#) présente les résultats de cette expérience. Les deux configurations présentent une préférence significative par rapport à l'architecture *baseline*. Cette tendance est particulièrement marquée pour la *config_B* avec une préférence moyenne d'environ 40% pour l'architecture *config_B* contre environ 15% pour l'architecture *baseline*.

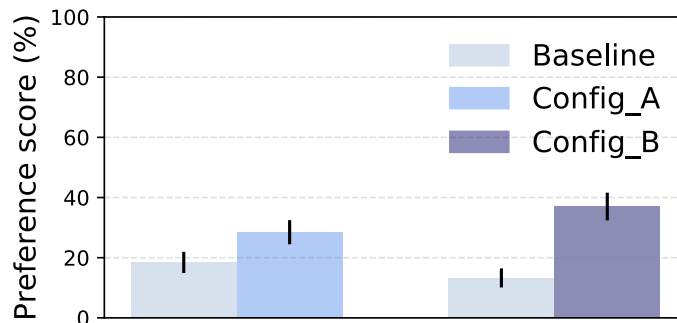


FIGURE 3.8 – Distribution des scores de préférence XAB : moyenne et intervalle de confiance à 95% entre l'architecture *baseline* et les architectures *config_A* (à gauche) et *config_B* (à droite).

Synthèse et discussion

Pour conclure partiellement ce chapitre, je dresse une brève synthèse des principales contributions réalisées au cours des recherches présentées, et en discute les propriétés au regard des questions de recherche formulées.

Les contributions apportées dans le première partie de ce chapitre sont :

Contributions

- C1.** Conversion neuronale de la F0 de la parole neutre à la parole expressive, par modélisation séquence-à-séquence ([Robinson et al., 2019](#)) et par modélisation multi-échelle ([Le Moine et al., 2021b](#)).
 - C2.** Construction de bases de données expressives en français, avec des émotions ([Veaux and Rodet, 2011](#)) ou des attitudes sociales (?) actées.
 - C3.** Contributions méthodologiques sur l'évaluation et la perception de la conversion de la parole expressive)
-

Les réponses partielles aux questions de recherche formulées et les solutions envisagées pour la suite sont :

Q1. Quelles représentations? La conversion de la F0 seule est insuffisante pour convertir efficacement et de manière réaliste l'expressivité de la parole.

Il y a plusieurs raisons à cela : D'une part, la F0 n'est pas le seul paramètre qui encode l'expressivité dans la communication parlée. L'ensemble des paramètres de la parole, aussi bien prosodique que phonétique, interviennent et doivent être modélisés. D'autre part, la transformation de la F0 seule — c'est-à-dire en préservant le reste des paramètres de la parole — conduit à une dégradation du signal de parole non-réaliste. L'ensemble des paramètres du signal de parole sont inter-dépendants et doivent en conséquence être modifiés de manière cohérente lors de la conversion pour obtenir un rendu réaliste.

→ Une solution envisagée est d'apprendre directement les conversions soit à partir du signal brut soit à partir de représentations intermédiaires contenant toutes l'information du signal de parole. L'introduction des vocodeurs neuronaux (Morise et al., 2016; Shen et al., 2018; Prenger et al., 2019; Kong et al., 2020; Roebel and Bous, 2022) a rendu réalisable cette solution. La représentation utilisée pour l'apprentissage de la conversion est le spectrogramme d'amplitude en échelle mel. Cette représentation est une représentation compressée du spectrogramme suivant une compression fréquentielle en échelle logarithmique. Elle contient toute l'information du signal de parole, puisqu'un vocodeur neuronal peut reconstruire de manière quasi transparente le signal de parole à partir de cette représentation.

Q2. Quelles données? Les bases de données appareillées ou parallèles sont extrêmement limitantes pour l'apprentissage. La construction de bases de données appareillées ou parallèles présente l'avantage de contrôler la variabilité linguistique des données et facilite l'apprentissage de conversions. En revanche, elles sont extrêmement coûteuses en temps humains et par conséquent limitées en taille. Cette limitation empêche d'exploiter pleinement les capacités d'apprentissage et de généralisation des réseaux de neurones.

→ La solution envisagée est d'augmenter les degrés de liberté des données, en réduisant les degrés de libertés des modèles d'apprentissage. En d'autres termes, le modèle d'apprentissage doit être structuré pour encoder explicitement les facteurs de variabilités considérés.

Q3. Quel modèle d'apprentissage? Les modèles présentés apprennent directement les conversions. Ils sont en l'état sous-spécifiés et incapables d'apprendre à partir de données à la volée. Dans un premier temps : comment intégrer explicitement des informations au cours de l'apprentissage? Dans un second temps : comment apprendre des conversions à partir de donnée à la volée?

→ Les architectures d'auto-encodeurs présentent l'avantage de pouvoir être apprises directement à partir des données à la volée, en encodant des représentations compressées du signal de parole à partir desquelles une reconstruction quasi transparente est possible. En contrepartie, elles présentent également deux problèmes : comment démêler efficacement les informations du signal de parole (contenu linguistique, identité, émotions, etc...) dans des codes distincts? Comment encoder efficacement ces informations pour que leur manipulation soit effective en conversion?

Le reste de ce chapitre présente mes travaux initiés dans les directions de recherche envisagées, à savoir à terme de tendre vers : un apprentissage de la conversion à partir représentations compactes et complètes, à partir de bases de données à la volée, et à partir d'architecture neuronale structurée permettant d'encoder de manière efficace les informations contenues dans le signal de parole pour pouvoir les manipuler à la génération.

Cette évolution est présentée en deux temps : dans un premier temps, je présente la suite des travaux sur la conversion de l’expressivité en intégrant l’encodage de l’information linguistique pour pallier aux limitations des bases de données parallèles ; dans un deuxième temps, je présente mes derniers travaux sur la conversion de l’identité vocale à partir d’architectures auto-encodeurs et m’intéresse à la problématique de l’apprentissage de représentations démêlées.

3.2 CONVERSION NEURONALE DE L’EXPRESSIVITÉ À PARTIR DE MEL-SPECTROGRAMMES

Avec l’apparition des vocodeurs neuronaux, la conversion neuronale de la parole ne s’est plus limitée à la conversion d’un seul paramètre comme la F0 mais cherche désormais à apprendre la conversion de l’ensemble des paramètres de ces vocodeurs (Zhou et al., 2020a,b; Kameoka et al., 2020), qu’ils soient paramétriques (Morise et al., 2016) ou non (Shen et al., 2018; Prenger et al., 2019; Kong et al., 2020; Roebel and Bous, 2022). En particulier, les mel-spectrogrammes se sont imposés comme une représentation compacte et complète à partir de laquelle il est possible de reconstruire le signal de parole de manière quasi-transparente. Je présente maintenant une première évolution du paradigme de conversion neuronale : la fonction de conversion est apprise directement à partir des mel-spectrogrammes mais toujours à partir de bases de données parallèles ; pour pallier au manque de généralisation associée à ces bases, un encodage explicite de l’information liée au contenu linguistique est intégrée à l’algorithme de conversion neuronal (Le Moine, 2023).

Positionnement et formulation neuronale

Le positionnement du problème est le même que dans la section précédente. La formulation neuronale proposée reprend la formulation du problème de la conversion présentée comme d’un problème de traduction de séquence à séquence, et implémentée cette fois-ci au moyen d’une architecture transformeur. Ces architectures se sont imposées comme la norme pour la modélisation de séquences, en particulier par la généralisation du mécanisme d’attention supposé reproduire artificiellement le mécanisme d’attention cognitive. Ce mécanisme d’attention linéaire présente en particulier de bonnes propriétés de mémoire temporelle de l’information, de complexité algorithmique, et de distribution du calcul (Vaswani et al., 2017). La formulation neuronale proposée pour la conversion de l’expressivité reprend l’architecture présentée dans (Kameoka et al., 2021) pour la conversion de l’identité, en l’adaptant aux représentations mel-spectrogrammes et à la tâche de conversion de l’expressivité, comme illustré sur la Figure 3.9.

Le problème consiste à convertir le mel-spectrogramme d’un énoncé c prononcé par un locuteur s d’une expressivité source $\mathbf{X}_c^{s,src}$ vers le mel-spectrogramme correspondant dans une expressivité cible $\mathbf{X}_c^{s,tgt}$. Pour simplifier, nous les noterons respectivement \mathbf{X}^{src} et \mathbf{X}^{tgt} dans la suite. La formulation de la solution neuronale proposée consiste à apprendre la traduction entre les expressivités source et cible, à partir d’une architecture transformeur. Cette architecture principale est augmentée de deux réseaux pré-net et post-net permettant respectivement en amont et aval du transformeur, utilisés respectivement pour changer la représentation utilisée en entrée et sortie du transformeur.

En amont, un pré-net est utilisé pour encoder les mel-spectrogrammes source et cible :

$$\tilde{\mathbf{X}}^{src} = \mathbb{E}_{src}^{pre}(\mathbf{X}^{src}) \quad (3.19)$$

$$\tilde{\mathbf{X}}^{tgt} = \mathbb{E}_{tgt}^{pre}(\mathbf{X}^{tgt}) \quad (3.20)$$

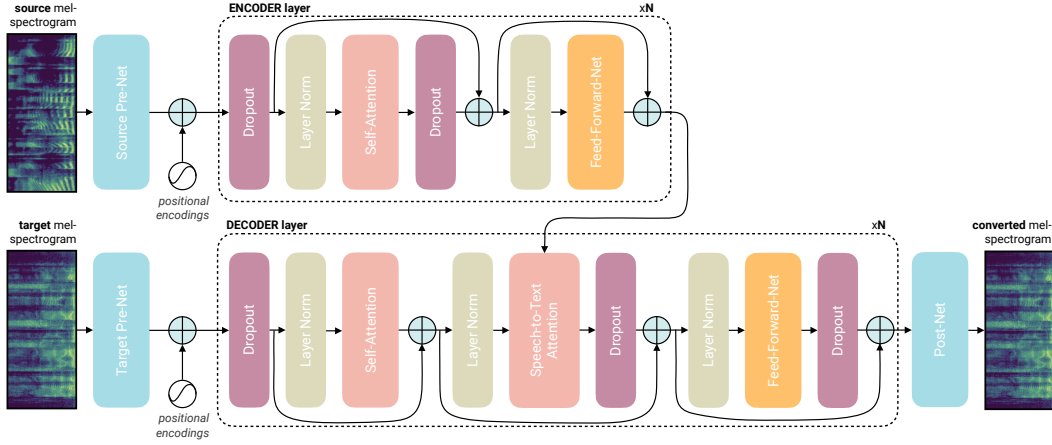


FIGURE 3.9 – Architecture neuronale proposée pour la conversion de l’expressivité d’un énoncé avec une expressivité source vers le même énoncé avec une expressivité cible. Le détail de cette architecture est donné dans le texte de cette section.

En aval, un post-net est utilisé pour reconstruire le mel-spectrogramme cible à partir de la sortie du transformer :

$$\hat{\mathbf{X}}^{\text{src} \rightarrow \text{tgt}} = \text{D}^{\text{post}}(\tilde{\mathbf{X}}^{\text{src} \rightarrow \text{tgt}}) \quad (3.21)$$

L’encodeur transformer E encode le mel-spectrogramme de l’énoncé avec l’expressivité source $\{\mathbf{X}_c^{\text{s,src}}\}$ en une séquence de vecteurs de contexte $\mathbf{Z}_c^{\text{s,src}}$ caractérisant l’énoncé source :

$$\mathbf{Z}^{\text{src}} = E(\tilde{\mathbf{X}}^{\text{src}}) \quad (3.22)$$

Nous notons les sorties des couches successives l de l’encodeur $\tilde{\mathbf{X}}_{(l)}^{\text{src}}$.

Le décodeur transformer décode séquentiellement l’énoncé avec l’expressivité cible à partir du vecteur de contexte de l’énoncé source \mathbf{Z}^{src} et des valeurs précédemment décodées de l’énoncé cible :

$$\tilde{\mathbf{x}}_{j+1}^{\text{src} \rightarrow \text{tgt}} = \text{D}(\tilde{\mathbf{x}}_j^{\text{tgt}}, \mathbf{Z}^{\text{src}}) \quad (3.23)$$

Nous notons les sorties des couches successives l du décodeur $\tilde{\mathbf{X}}_{(l)}^{\text{src} \rightarrow \text{tgt}}$.

Ce décodeur fonctionne de manière auto-régressive : à partir d’une initialisation par un vecteur nul, le décodeur décode itérativement la valeur à l’instant $j + 1$ conditionnellement à la valeur précédemment décodée à l’instant j et à l’encodage de l’énoncé source \mathbf{Z}^{src} . Avec les notations utilisées pour cette architecture, nous avons alors l’égalité :

$$\tilde{\mathbf{x}}_{j+1}^{\text{tgt}} = \text{E}^{\text{pre}}(\tilde{\mathbf{x}}_{j+1}^{\text{src} \rightarrow \text{tgt}}) \quad (3.24)$$

Par ailleurs, un mécanisme d’attention est utilisé pour sélectionner dynamiquement dans la séquence source \mathbf{Z}^{src} le vecteur de contexte représentant l’information utile pour le décodage à l’instant j de la séquence cible. Ce mécanisme d’attention s’écrit sous la forme d’une matrice d’attention $\mathbf{A} \in \mathbb{R}_+^{\text{T}_{\text{src}} \times \text{T}_{\text{tgt}}}$ qui mesure la similarité entre chaque code de la séquence décodée et tous les codes de la séquence encodée. Dans l’architecture proposée, cette attention est multi-tête mais nous ne présentons pas les notations correspondantes pour préserver la clarté de la lecture.

La version plusieurs-à-plusieurs de cet algorithme consiste à rajouter explicitement le conditionnement sur l'expressivité k à chaque couche de l'encodage de l'énoncé source et du décodage de l'énoncé cible :

$$\tilde{\mathbf{X}}_{(l)}^k \leftarrow [\tilde{\mathbf{X}}_{(l)}^k, k] \quad (3.25)$$

$$\tilde{\mathbf{X}}_{(l)}^{k \rightarrow k'} \leftarrow [\tilde{\mathbf{X}}_{(l)}^{k \rightarrow k'}, k'] \quad (3.26)$$

où : les expressivités k et k' sont représentées sous la forme d'un vecteur un-chaud.

Une fonction de perte de reconstruction \mathcal{L}_{rec} est définie comme la norme L_1 entre la séquence des valeurs prédites pour l'expressivité $k \rightarrow k'$ et la séquence observée de l'expressivité k' décalée d'une trame vers la droite :

$$\mathcal{L}_{rec}^{(k,k')}(E, D) = \mathbb{E}_{\mathbf{X}^k, \mathbf{X}^{k'}} ([\hat{\mathbf{X}}^{t, k \rightarrow k'}]_{1:T_t-1, :} - [\mathbf{X}^{t, k'}]_{2:T_t, :1}) \quad (3.27)$$

Cette fonction de perte est également fonction des paramètres des pré-encodeur E_{src}^{pre} et E_{tgt}^{pre} , et du post-décodeur D^{post} . Pour des raisons de lisibilité, je ne les fais pas apparaître dans les formulations proposées.

En langage naturel, il est courant d'observer des dépendances à long terme dans le passé comme dans le futur, notamment pour la traduction où l'ordonnement des mots peut être modifié très significativement d'une langue à une autre. En parole pour la traduction d'une expressivité à une autre dans une même langue, il n'existe pas de tels phénomènes. Les dépendances sont alors locales dans le temps et nous faisons l'hypothèse que l'attention est monotone dans le temps et qu'en l'occurrence que l'espace d'attention se limite à un corridor autour du temps présent. Pour implémenter cette contrainte sur la matrice d'attention \mathbf{A} , une fonction de perte d'attention diagonale est définie, comme :

$$\mathcal{L}_{da}^{(k,k')}(D) = \mathbb{E}_{\mathbf{X}^k, \mathbf{X}^{k'}} \frac{1}{T_s T_t L_{dec} H} \sum_{l=1}^{L_{dec}} \|\mathbf{G}_{T_s \times T_t} \odot \mathbf{A}_{(l)}^{(k,k')}\|_1 \quad (3.28)$$

où $\mathbf{A}_l^{k,k'}$ représente la matrice d'attention source-cible associée à la couche l du décodeur, et $\mathbf{G}_{T_s \times T_t}$ est une matrice de poids exponentiellement décroissants autour de la diagonale. L'ajout de cette fonction de perte permet de restreindre l'espace des solutions et d'éviter une dispersion de l'attention qui résulte en une dégradation de la conversion.

La fonction de perte totale de l'architecture s'écrit alors :

$$\mathcal{L}_{vtn}^{(k,k')}(E, D) = \mathcal{L}_{rec}^{(k,k')}(E, D) + \lambda_{da} \mathcal{L}_{da}^{(k,k')}(D) \quad (3.29)$$

où : λ_{da} est le facteur de pondération de la contrainte sur l'attention diagonale.

Discussion L'architecture présentée est une formulation de l'architecture présentée dans (Kameoka et al., 2021) adaptée de la conversion de l'identité à la conversion de l'expressivité. Si l'attention diagonale permet d'améliorer la consistance de l'attention en la contraignant localement, nous avons observé régulièrement une dégradation de la conversion en fin d'énoncé, avec la conséquence de soit dégrader ou rendre incompréhensible le contenu linguistique, soit plus grave de le modifier. Ce phénomène a été particulièrement observé pour des énoncés qui n'ont pas été vus lors de l'apprentissage. En l'état, cette architecture est linguistiquement agnostique la variabilité linguistique étant contrôlée à l'apprentissage par l'intermédiaire des bases de données appareillées. Une conséquence directe à l'inférence est la limitation liée à la faible variabilité linguistique de ce genre de base de données.

Pour répondre à cette limitation, nous avons proposé d'introduire un module de reconnaissance de parole pour réaliser la transcription parole à texte en sortie de la conversion, comme illustré sur la [Figure 3.10](#).

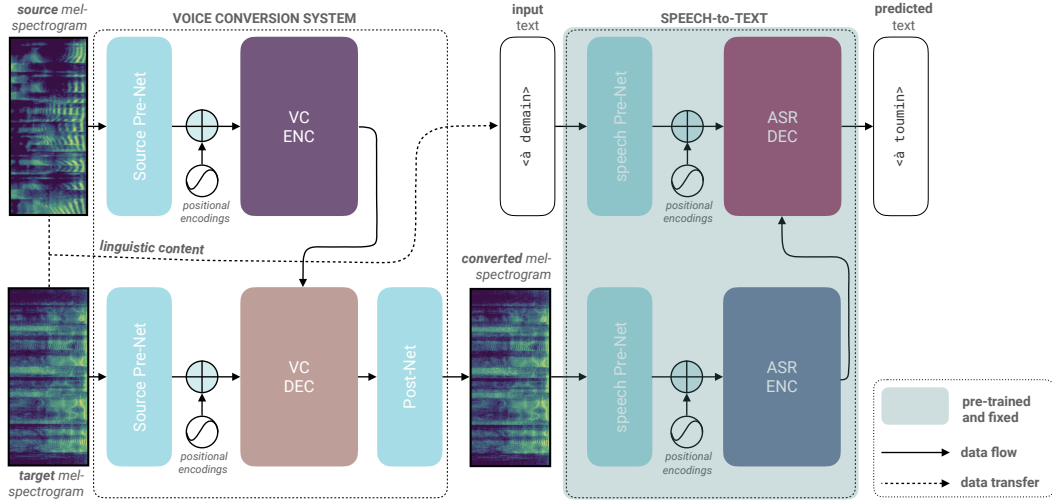


FIGURE 3.10 – Architecture proposée pour la conversion de l’expressivité avec préservation explicite du contenu linguistique. Partie gauche : architecture neuronale de conversion décrite précédemment. Partie droite : module neuronal de reconnaissance de la parole proposé en sortie de la conversion, pour assurer la préservation du contenu linguistique avec l’énoncé source.

La fonction de perte de reconnaissance \mathcal{L}_{asr} s’écrit comme l’entropie croisée entre la séquence de caractères observée $\mathbf{c} = [c_1, \dots, c_{N_c}]$ et la séquence $\hat{\mathbf{c}}$ transcrite à partir de la parole convertie :

$$\mathcal{L}_{\text{asr}}(E_{\text{asr}}, D_{\text{asr}}) = \mathbb{E}_{\mathbf{x}^k, \mathbf{x}^{k'}} - \sum_{i=1}^{N_c} c_i \log(\hat{c}_{i,j}) \quad (3.30)$$

La fonction de perte totale incluant la fonction de perte de reconnaissance s’écrit alors finalement :

$$\mathcal{L}_{\text{vtN} \times \text{asr}}(E, D) = \mathcal{L}_{\text{vtN}}(E, D) + \lambda_{\text{asr}} \mathcal{L}_{\text{asr}}(E, D) \quad (3.31)$$

Comme indiqué par l’équation précédente, le module de reconnaissance est pré-entraîné séparément de l’architecture principale, et n’est utilisé que pour fournir une fonction de perte de reconnaissance pour l’apprentissage des paramètres du transformeur. Par l’introduction d’une spécification explicite de l’information linguistique dans l’architecture de conversion neuronale, nous faisons l’hypothèse que cette information doit permettre la préservation de l’information linguistique au cours de la conversion. Par ailleurs, cet encodage linguistique doit également permettre une meilleure généralisation à des contextes linguistiques et des énoncés non vus lors de l’apprentissage.

Détails d’implémentation

Comme illustré sur la Figure 3.9, l’encodeur est composé de L_E blocs identiques de transformer, chaque bloc étant formé de la succession d’une couche d’auto-attention et d’une couche entièrement connectée. Des connexions résiduelles et des normalisations sont appliquées à chacune des couches pour prévenir la dispersion de l’information transmise par le gradient. Le décodeur est composé similairement de L_D blocs identiques de transformer, formé de la succession d’une couche d’auto-attention, d’une couche dense, de connexions résiduelles et de normalisation par lot. En supplément, une couche d’auto-attention multi-tête (MHSA) est utilisée pour estimer dynamiquement quelles trames de l’énoncé source correspondent contextuellement à la trame en cours de décodage de l’énoncé cible. À l’apprentissage, une stratégie d’enseignement forcé est employée : la vérité terrain de

l'énoncé cible est utilisée en entrée à chaque trame du décodeur, en substitution de la prédiction auto-régressive mentionnée précédemment. Cette stratégie permet d'éviter le phénomène de propagation temporelle de l'erreur et ainsi d'accélérer et de faciliter la convergence pendant l'apprentissage. À l'inférence, l'énoncé cible n'est pas disponible : l'inférence commence en initialisant la cible par un vecteur nul, puis en itérant l'inférence auto-régressive. L'ensemble des détails d'architecture et d'implémentation sont présentés dans (Le Moine, 2023)

Expérience : conversion de l'expressivité

L'architecture proposée a été évaluée expérimentalement sur la base de données ATT-HACK (Le Moine and Obin, 2020) présentée précédemment. Dans une première étude générale, nous avons comparé deux configurations de l'architecture de base inspirée de (Kameoka et al., 2021) et adaptée à la conversion de l'expressivité : **VTN-s** et **VTN-l**, respectivement une petite et une grande version de l'architecture transformeur. Pour une étude détaillée de l'impact du module de reconnaissance proposé, nous avons comparé des configurations avec trois facteurs de pondérations λ_{asr} de ce module. Cette seconde étude a été menée uniquement avec le grand modèle **VTN-l** qui présentait de meilleures performances que le modèle petit **VTN-s**.

- **VTN-s** : un petit modèle avec $L_{\text{enc}} = 1$ et $L_{\text{dec}} = 1$.
- **VTN-l** : un grand modèle avec $L_{\text{enc}} = 2$ et $L_{\text{dec}} = 2$.
- **VTN-lxASR-li** : le grand large modèle avec module ASR \mathcal{C}^{asr} et une pondération faible ($\lambda_{\text{asr}} = 0.1$).
- **VTN-lxASR-me** : le grand modèle avec module ASR \mathcal{C}^{asr} et une pondération moyenne ($\lambda_{\text{asr}} = 0.5$).
- **VTN-lxASR-st** : le grand modèle avec module ASR \mathcal{C}^{asr} et une pondération importante ($\lambda_{\text{asr}} = 1.0$).

L'ensemble des algorithmes ont été entraînés dans un mode plusieurs-à-plusieurs, et évalués sur un ensemble de test de phrases non vues lors de l'apprentissage.

La [Figure 3.11](#) illustre les conversions obtenues avec les algorithmes **VTN-l** et **VTN-lxASR-st** et les matrices d'alignement correspondantes pour deux phrases. Pour les conversions réalisées avec l'algorithme **VTN-l** (ligne du haut), nous observons une dégradation du contenu linguistique systématiquement vers la fin des énoncés. Ce phénomène peut s'expliquer comme une erreur cumulée lors de l'inférence auto-régressive, susceptible d'entraîner une divergence de la conversion. Ce phénomène disparaît avec l'intégration du module de reconnaissance **VTN-lxASR-st** (ligne du milieu), et présentant un alignement plus consistant avec la référence de l'attitude cible (ligne du bas).

Expérience objective : mesure de l'intelligibilité de la parole convertie

Dans un premier temps, nous voulions nous assurer de l'efficacité du module de reconnaissance de parole à préserver le contenu linguistique au cours de la conversion. La [Figure 3.12](#) présente les taux d'erreur de caractère (CER) et de mot (WER) calculés à partir des transcriptions de référence. Tout d'abord, le large modèle **VTN-l** apparaît comme substantiellement plus performant que le petit modèle **VTN-s** en termes d'erreurs CER et WER. Néanmoins, le taux d'erreur de ce modèle reste non négligeable, ce qui entraîne des dégradations de la conversion non seulement en termes d'expressivité mais également en termes de contenu linguistique soit dégradé soit modifié. L'ajout du module de reconnaissance de parole diminue considérablement et significativement la perte de contenu linguistique. Pour donner un exemple, l'algorithme de base **VTN-l** présente un CER de 17.3% alors que sa version avec module de reconnaissance **VTN-lxASR-st** seulement

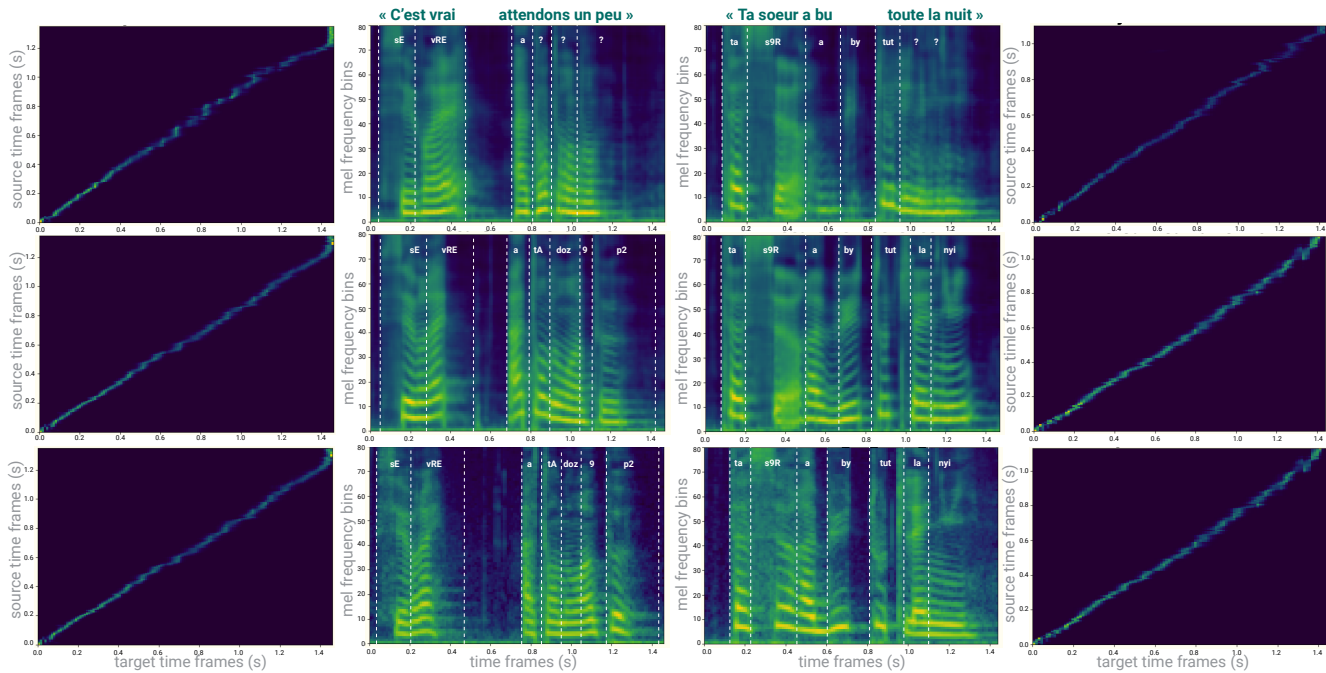


FIGURE 3.11 – Illustrations des conversions réalisées de distant à séductif et les attentions obtenues pour les phrases "C'est vrai attendons un peu" (à gauche) et "Ta soeur a bu toute la nuit" (à droite). La ligne du haut présente les conversions réalisées avec l'algorithme VTN-I; la ligne du milieu avec l'algorithme VTN-lxASR-st; et la ligne du bas la référence de la phrase pour l'attitude cible.

un CER de 3,8%. Ceci démontre clairement l'efficacité du module de reconnaissance dans la préservation du contenu linguistique au cours de la conversion.

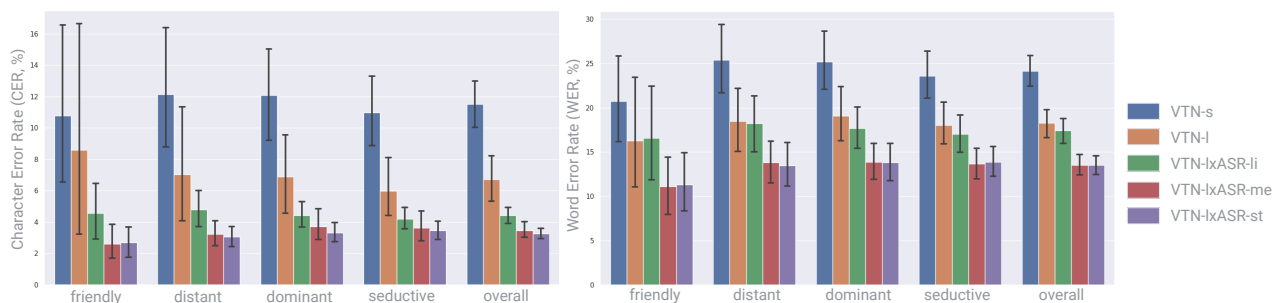


FIGURE 3.12 – Taux d'erreur de reconnaissance pour les algorithmes de conversions comparés VTN-s, VTN-l, VTN-lxASR-li, VTN-lxASR-me et VTN-lxASR-st en fonction des 4 attitudes. À gauche : taux d'erreur de caractère (CER); à droite : taux d'erreur de mot (WER). Les barres présentent le taux d'erreur moyen et les moustaches l'intervalle de confiance à 95%.

Nous nous limitons ici à la présentation de l'efficacité de l'algorithme proposé dans la préservation du contenu linguistique lors de la conversion. Une description détaillée des résultats obtenus, notamment en termes d'erreur de prédiction de F0 et de distorsion spectrale sont présentés dans (Le Moine, 2023).

Expériences perceptives

Dans un second temps, nous avons évalué les algorithmes proposés VTN-s, VTN-l, VTN-lxASR-li, et VTN-lxASR-st dans le cadre d'une expérience perceptive. Pour ce faire, les

conversions ont été réalisées avec les quatre configurations pour dix phrases sélectionnées aléatoirement. La référence de la cible a été ajoutée comme stimulus de contrôle. Chaque participant devait évaluer 20 paires d'échantillons sélectionnées aléatoirement parmi l'ensemble des paires d'échantillons générées. Le participant devait juger ces paires selon les deux indications suivantes : 1) *lequel des deux échantillons est le plus intelligible?*, et 2) *lequel des deux échantillons véhicule le plus l'attitude cible considérée?*. Les réponses étaient formulées selon l'échelle MOS présentée précédemment. Si nous notons A et B les deux échantillons, les réponses possibles étaient alors *Surtout A, Plutôt A, Entre les deux, Plutôt B* and *Surtout B*. Nous pouvons observer une nouvelle évolution méthodologique pour évaluer la perception d'une attitude véhiculée, en substituant à un échantillon cible de référence seulement l'attitude désignée. Cette évolution se base sur l'hypothèse que la présentation d'un exemple comme une référence contraint trop fortement les représentations possibles par rapport à cette instance. L'adéquation à l'attitude désignée laisse ainsi ouvert les possibles variations associées aux représentations de chaque participant. L'expérience a permis de récolter les réponses de 140 participants.

Pour l'intelligibilité, la **Figure 3.13** présente les scores de préférence obtenus pour les algorithmes comparés et la référence. Les résultats de l'expérience perceptive confirment ceux obtenus lors de l'évaluation objective précédente. L'algorithme avec forte contrainte de préservation **VTN-IxASR-st** obtient en moyenne les meilleurs scores de préférence par rapport à l'ensemble des algorithmes comparés. En particulier, il est considéré en moyenne comme plutôt plus intelligible que les algorithmes de base, et passablement préféré à l'algorithme avec une contrainte de préservation faible **VTN-IxASR-li**. Par ailleurs, l'intégration de la contrainte de préservation améliore toujours l'intelligibilité de la parole convertie.

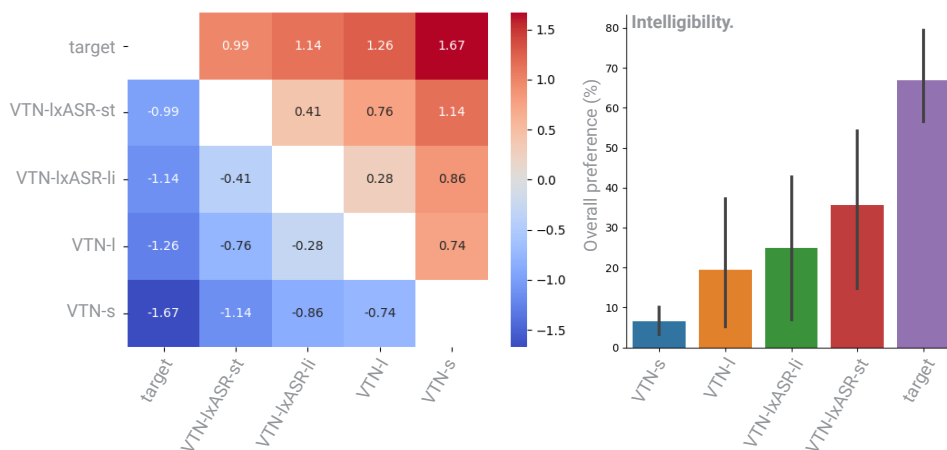


FIGURE 3.13 – Perception de l'intelligibilité. À gauche : matrice de confusion des scores d'intelligibilité moyens pour les 4 algorithmes et la référence. Un score positif indique une préférence de l'algorithme y (ligne) par rapport à l'algorithme x (colonne). À droite : proportion de paires préférées pour chaque configuration.

Pour la perception de l'attitude véhiculée, la **Figure 3.14** présente les scores de préférence obtenus pour les 4 algorithmes comparés et la référence de l'attitude cible. Les résultats s'interprètent de la même manière que pour l'intelligibilité. Les résultats présentent une tendance similaire à celle observée précédemment, cependant moins marquée : l'algorithme avec forte contrainte de préservation du contenu linguistique **VTN-IxASR-st** est préféré à l'ensemble des autres algorithmes en terme de similarité à l'attitude cible. En particulier, il est passablement préféré aux algorithmes de base et présenté une très faible préférence par rapport à l'algorithme avec une contrainte de préservation faible **VTN-IxASR-li**.

À partir de ces deux observations, nous pouvons conclure que l'intégration de la contrainte de préservation du contenu linguistique lors de la conversion non seulement

améliore l'intelligibilité de la parole convertie mais également la conversion vers l'attitude désirée, en particulier sur des nouvelles phrases avec des contenus linguistiques non observés pendant l'apprentissage. Pour bien interpréter ces résultats, il faut observer que ces deux critères ne sont clairement pas indépendants : la dégradation de l'intelligibilité gêne certainement la perception de l'attitude véhiculée. Inversement, l'amélioration de l'intelligibilité facilite la perception de l'attitude véhiculée.

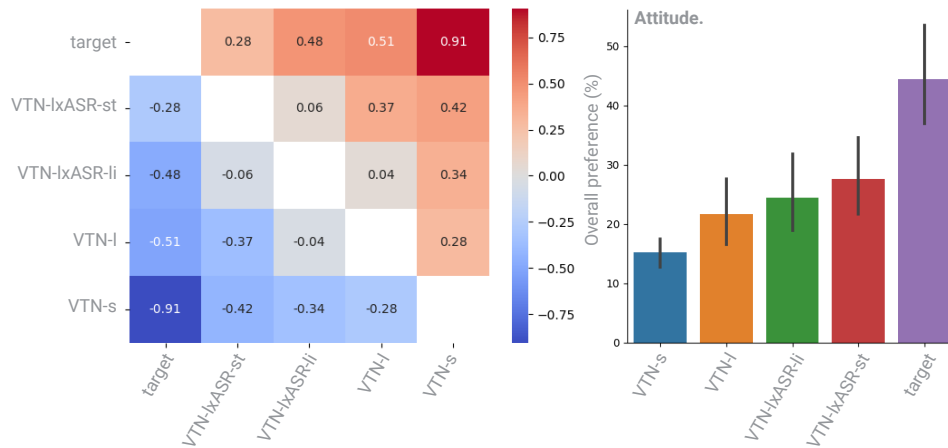


FIGURE 3.14 – Perception de la similarité à l'attitude cible. À gauche : matrice de confusion des scores de préférence moyens pour les 4 algorithmes et la référence. Un score positif indique une préférence de l'algorithme y (ligne) par rapport à l'algorithme x (colonne). À droite : proportion de paires préférées pour chaque configuration.

Conclusion et transition

Pour proposer un positionnement dans la perspective des questions de recherche formulées Q2 et Q3, ces travaux constituent une première étape de structuration de l'apprentissage, en spécifiant les informations encodées dans le réseau. Cette spécification démontre son efficacité, même dans le contexte d'un apprentissage à partir de données parallèles, pour améliorer la généralisation du modèle en particulier à des contextes linguistiques non observés pendant l'apprentissage. Pour pouvoir passer à un apprentissage à partir de données à la volée, cette spécification va s'imposer comme une nécessité, et provoquer un changement important de paradigme d'apprentissage dans la dualité des problématiques associées aux questions de recherche Q2 et Q3. C'est l'objet de la section suivante, dernière contribution de ce manuscrit.

3.3 APPRENTISSAGE DE REPRÉSENTATIONS DÉMÊLÉES POUR LA MANIPULATION DES ATTRIBUTS DE LA VOIX

Conversion de l'identité vocale : histoire, problèmes, méthodes

L'identité constitue depuis environ 30 ans la branche principale de la recherche en conversion vocale (VC), et partage une position privilégiée avec la synthèse de parole à partir du texte (TTS) dans le domaine de la modélisation générative de la parole. La tâche consiste à partir de l'enregistrement d'un énoncé prononcé par un locuteur *source* d'en modifier les caractéristiques acoustiques pour qu'il soit perçu comme ayant été prononcé par un locuteur *cible*. La notion d'identité vocale est éminemment complexe et débattue, notamment dans le domaine de la biométrie. L'identité vocale se manifeste en premier lieu par des déterminants physiologiques : la géométrie de l'appareil vocal et en particulier celle du conduit vocal et de ses résonances est un facteur d'identification.

Mais l'identité est aussi culturelle : nous construisons en partie notre manière de parler, par imitation de notre environnement familial, géographique et culturel, ce qui contribue également à notre identité. Par ailleurs, cette identité est dynamique ; contrairement au visage ou à l'iris, la voix n'est pas un instantané, son intégration temporelle est nécessaire préalablement à son interprétation, en particulier en terme d'identité. Cette identité est par ailleurs mouvante : elle se modifie selon de nombreux facteurs, en particulier avec l'âge. La conversion d'identité vocale a historiquement été formulée comme un problème d'apprentissage de la correspondance statistique un-à-un entre la distribution acoustique d'un locuteur *source* et d'un locuteur *cible*. L'apprentissage de cette correspondance a été simplifié, en limitant les facteurs de variabilité acoustique et notamment la variabilité linguistique associée au texte prononcé. Pour ce faire, des bases de données dites parallèles étaient conçues spécifiquement, c'est-à-dire constituées d'un ensemble d'énoncés communs aux deux locuteurs, et préalablement alignés temporellement, permettant ainsi une mise en correspondance trame à trame des locuteurs source et cible. Pendant l'apprentissage, la distribution acoustique conjointe entre les locuteurs source et cible est modélisée avec un modèle de mélange de Gaussiennes (GMM) (Stylianou et al., 1998). Lors de la conversion, une régression linéaire est effectuée sur cette distribution conjointe afin de déterminer les caractéristiques vocales du locuteur cible conditionnellement à celles du locuteur source. Les GMMs ont été le paradigme principal la recherche en VC pendant plus de 10 ans, avec de nombreuses améliorations apportées par rapport à la formulation d'origine notamment pour l'apprentissage à partir de bases multi-locuteurs (Toda et al., 2007).

La VC neuronale, c'est-à-dire la VC basée sur des réseaux de neurones, a été introduite pour la première fois par (Desai et al., 2009). Cette première proposition consistait néanmoins seulement à remplacer dans la formulation historique la modélisation de la distribution acoustique par un réseau de neurones. Néanmoins les nombreuses avancées réalisées dans la théorie comme dans l'application des réseaux de neurones (Goodfellow et al., 2014b; Sutskever et al., 2014b; Bahdanau et al., 2015; Hsu et al., 2016; Zhu et al., 2017) ont amené à reformuler le problème de la conversion d'identité vocale dans une perspective neuronale. Ainsi, nous pouvons observer un net changement de paradigme depuis la VC un-à-un apprise à partir de bases de données parallèles de paires de locuteurs à la VC plusieurs-à-plusieurs apprises à partir de bases multi-locuteurs et non parallèles. Dans la lignée historique de la VC biunivoque, les premiers algorithmes de VC neuronaux ont été proposés pour apprendre la correspondance acoustique à partir de paires de phrases provenant de locuteurs source et cible non nécessairement pré-alignées, puis non nécessairement parallèles. La VC séquence-à-séquence (S2S) (Tanaka et al., 2019; Kameoka et al., 2020) formule la VC comme un problème de traduction entre une identité source et une identité cible. Cette traduction est opérée sous la forme d'un encodeur et d'un décodeur séquentiels, à l'interface desquels un mécanisme d'attention (Bahdanau et al., 2015) réalise l'alignement entre l'encodage des séquences des locuteurs source et cible, optimisant ainsi l'apprentissage séquentiel de la conversion. La VC par cycle-GAN (Kaneko and Kameoka, 2017; Kaneko et al., 2019; Fang et al., 2018) présente la VC comme un problème d'adaptation de domaine (Zhu et al., 2017) dans lequel l'identité du locuteur constitue le domaine à adapter. Un auto-encodeur couplé à un réseau génératif adversarial (GAN) est utilisé pour apprendre la fonction de conversion de la distribution acoustique de l'identité source vers celle de l'identité cible. L'introduction d'un cycle vise à apprendre l'adaptation de domaine à partir de données non-appareillées, par exemple des images présentant des structures et des textures variées et non alignées ou des phrases présentant des contenus linguistiques variés. L'introduction du cycle permet alors de réduire l'espace des solutions possibles et de favoriser la préservation de la structure pendant la conversion - c'est-à-dire dans notre cas du contenu linguistique. Les discriminateurs présentent l'avantage de pouvoir apprendre explicitement des conversions à partir de bases non appareillées, c'est-à-dire pour lesquelles il n'existe pas de référence objective pour le signal converti, cette lacune étant palliée partiellement par des mesures probabilistes sur les distributions des locuteurs

source et cible. En contrepartie, l'apprentissage adversarial présente les désavantages liés à la complexité et la stabilité des apprentissages par GANs. Malheureusement, le paradigme de la VC un-à-un, à partir de bases de données parallèles ou non, demeure limité par nature. La quantité de données disponibles est trop faible pour pleinement bénéficier des avantages de l'apprentissage de réseaux de neurones profonds et en l'occurrence pour apprendre efficacement la conversion.

Pour pallier ces insuffisances, les efforts de recherche se sont progressivement concentrés pour formuler le problème de la conversion neuronale selon un paradigme de conversion plusieurs-à-plusieurs et entraînable à partir de bases de données non-parallèles, idéalement à partir de données à la volée récupérées sur internet. Cette évolution a permis un passage à l'échelle de la conversion neuronale et d'une amélioration substantielle de la qualité de la conversion. Par ailleurs, la décomposition de l'apprentissage en une phase de pré-apprentissage à partir de très grandes quantités de données, et d'une adaptation (fine-tuning) à partir de peu de données permet aujourd'hui de combiner haute-qualité de la conversion et faible quantité de données nécessaires pour le transfert d'identité. En particulier, le starGAN VC (Kameoka et al., 2018; Kaneko et al., 2019) généralise le principe du cycle-GAN VC à la VC neuronale plusieurs-à-plusieurs en introduisant un encodage explicite de l'identité du locuteur. Par extension du cycleGAN, le décodeur est explicitement conditionné sur l'identité du locuteur. Ce conditionnement présente un double avantage : d'une part, il ouvre la voie de l'apprentissage à partir de bases de données multi-locuteurs, et d'autre part simplifie l'apprentissage du cycleGAN puisqu'il n'est plus nécessaire d'apprendre les poids d'un encodeur et d'un décodeur dont les points sont partagés pour la conversion dans les deux sens. En plus des fonctions de perte du cycleGAN, une perte de classification est ajoutée pour déterminer l'identité du locuteur à partir du signal de parole converti.

Dans le même temps, une autre voie de recherche formule le problème de la conversion simplement sous la forme d'un auto-encodeur conditionné sur l'identité du locuteur (Zhou et al., 2018; Lu et al., 2019; Qian et al., 2019). Ce paradigme présente l'avantage de grandement simplifier l'apprentissage : une seule perte de reconstruction est utilisée mesurable à partir du signal de parole observé. L'architecture est similaire à celle d'un starGAN VC sauf que les locuteurs source et cible sont les mêmes pendant l'apprentissage, permettant donc d'exploiter la comparaison à une référence observée par une simple mesure d'erreur de reconstruction. Néanmoins, ce paradigme présente la contrepartie que la conversion n'est pas explicitement apprise pendant l'apprentissage, et que par conséquent il n'y a pas de garantie que le conditionnement sur l'identité soit opérationnel pendant la conversion. Pendant la conversion, il suffit de manipuler l'attribut de l'identité du locuteur en entrée du décodeur pour convertir l'énoncé du locuteur source avec l'identité souhaitée du locuteur cible. Cette architecture décompose le problème de la conversion en deux parties distinctes : une phase de pré-entraînement pendant laquelle l'architecture est entraînée sur une grande quantité de données multi-locuteurs, ce qui permet en outre d'apprendre efficacement l'espace latent dans lequel les identités des locuteurs sont projetées à partir d'un énoncé ; une phase d'ajustement fin qui permet d'optimiser la conversion pour un locuteur cible. Cette phase vise principalement à déterminer la position du locuteur cible dans l'espace latent des locuteurs. Cette architecture constitue parmi les premières tentatives de VC neuronale à partir de peu d'exemples (Lu et al., 2019; Qian et al., 2019) — à la limite à partir d'un seul énoncé.

En complément de la VC neuronale par auto-encodeur et par GAN, une piste de recherche actuelle consiste à apprendre explicitement des représentations structurées de l'information contenue dans un signal de parole — en particulier, en explicitant l'encodage de l'information liée au contenu linguistique de l'énoncé (c'est-à-dire sa transcription orthographique ou phonémique) et l'encodage de l'information liée à l'identité du locuteur. En particulier, les algorithmes de VC neuronale ont commencé à intégrer des informations

sur le contenu linguistique et l'identité du locuteur (Zhang et al., 2019), par exemple en proposant un encodage explicite de l'information linguistique à partir d'algorithmes de transcription externes (Phonetic Posterior-Grams, PPG (Sun et al., 2016; Mohammadi and Kim, 2019)). L'émergence de l'apprentissage de représentations démêlées (Higgins et al., 2018), c'est-à-dire la capacité à encoder de manière différenciée les informations contenues dans un signal, a permis encore une fois de franchir une étape supplémentaire pour l'apprentissage de la conversion, en donnant la possibilité d'apprendre l'ensemble des encodages de manière cohérente au sein d'une seule architecture — et par conséquent d'optimiser l'apprentissage de la conversion. Ce problème peut s'écrire sous une forme neuronale et être formulé en adoptant une stratégie de démêlage fondée sur la théorie de l'information (Belghazi et al., 2018) : typiquement, par limitation de l'information (Tishby and Zaslavsky, 2015) (dimensionner les codes pour qu'ils n'encodent que l'information souhaitée) ou par apprentissage adversarial (Goodfellow et al., 2014b) (apprendre des codes séparés pour chaque source d'information et statiquement indépendants les uns des autres). Dans (Qian et al., 2020b), trois limitations d'information sont utilisés pour encoder séparément les paramètres vocaux de la hauteur, du timbre et du rythme, tandis que (Zhang et al., 2020; Yuan et al., 2021) utilisent l'apprentissage adversarial pour apprendre de manière différenciée les informations linguistique et de locuteur.

Pour résumer l'évolution de la VC et en particulier de la VC neuronale : on observe un lien direct entre les degrés de liberté des données utilisées pour l'apprentissage, et les degrés de liberté des algorithmes d'apprentissage. En d'autres mots : plus les données sont contraintes (c'est-à-dire, dont on contrôle les facteurs de variabilité, comme c'est le cas pour les bases parallèles), moins l'algorithme doit être spécifié ; plus les données sont libres, plus l'algorithme doit être spécifié. Cette spécification est introduite sous la forme de connaissances a priori utilisées pour construire une architecture neuronale, typiquement pour intégrer les informations sur le contenu linguistique et l'identité du locuteur. Je présente dans la suite de cette section les travaux qui ont occupé mes dernières années de recherche et mes contributions à cette histoire : principalement, le projet ANR TheVoice (2017-2022) dont j'ai été le coordinateur principal et qui a abouti aux publications (Bous et al., 2022) et (Benaroya et al., 2023), et les thèses de Léane Salais et de Théodor Lemerle (en cours).

Positionnement du problème et formulation neuronale

L'architecture neuronale de la VC présentée est directement inspirée de (Zhang et al., 2020). L'idée principale de cette architecture VC est que les représentations linguistique et locuteur sont encodées de manière démêlée au cours de l'apprentissage. Deux encodeurs sont utilisés pour apprendre les représentations linguistique et locuteur séparément, et une stratégie adversariale est introduite pour rendre l'information de l'encodage linguistique indépendante de l'identité du locuteur. Cette architecture est illustrée sur la [Figure 3.15](#). Les entrées de l'architecture VC sont la matrice du signal de parole \mathbf{A} , représentée par le mel-spectrogramme calculé sur T trames temporelles, et la séquence de T phonèmes \mathbf{p} correspondant à la transcription phonétique du texte d'entrée aligné sur le signal de parole correspondant. Deux encodeurs, E^c et E^s , sont utilisés pour coder le contenu linguistique et l'identité du locuteur.

Pré-traitement et post-réseau

La VC neuronale proposée fonctionne sur une représentation en spectrogramme mel du signal de parole. Pour l'analyse du signal, nous suivons le paramétrage proposé dans (Liu et al., 2018), c'est-à-dire que le signal d'entrée est sous-échantillonné à 16kHz, converti en STFT en utilisant une fenêtre de Hanning de 50ms avec un pas d'avancement de 12,5ms et nombre de points fréquentiels de 2048, puis compressé perceptivement sur une

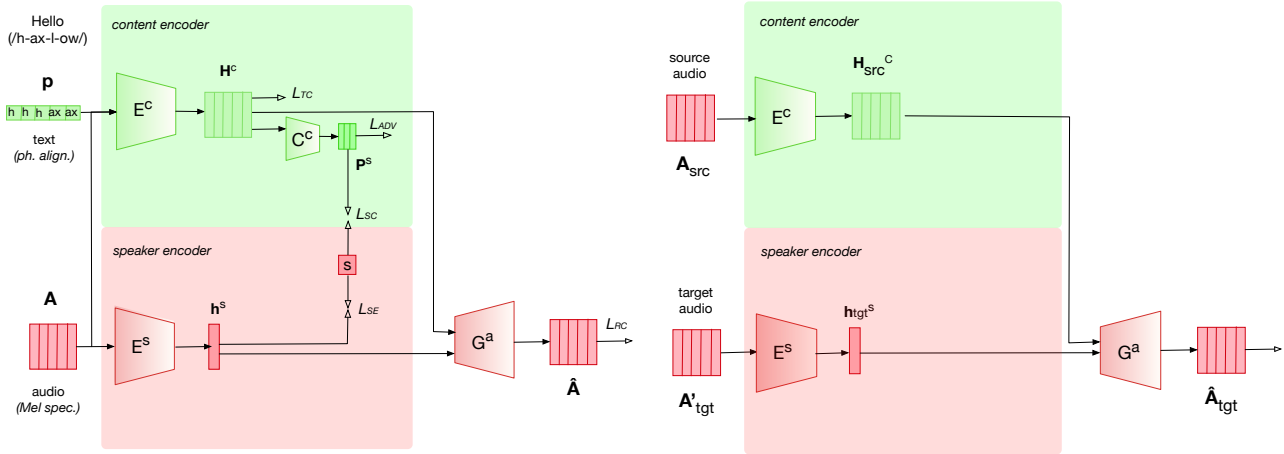


FIGURE 3.15 – Architecture de la VC neuronale avec apprentissage adversarial des représentations contenu et locuteur démixées. À gauche : phase d’apprentissage. À droite : phase de conversion. Source : (Benaroya et al., 2023)

échelle fréquentielle mel avec 80 points fréquentiels en échelle mel. Nous normalisons les logarithmes des mel-spectrogrammes en normalisant chaque point de fréquence séparément par rapport à la moyenne et à l’écart type de l’ensemble des données. Pour restituer le signal de parole à partir d’un mel-spectrogramme généré, nous utilisons l’algorithme d’inversion Mel présenté dans (Roebel and Bous, 2022).

Encodeur du locuteur

L’encodeur du locuteur E^s convertit le signal de parole A en un vecteur indépendant du temps h^s , à partir de l’hypothèse que l’identité d’un locuteur ne varie pas au cours d’un énoncé :

$$h^s = E^s(A) \quad (3.32)$$

La fonction de perte associée à la reconnaissance du locuteur \mathcal{L}_{SE} est définie comme l’entropie croisée entre l’identité du locuteur prédite à partir de h^s par un classifieur C_s^s , et la véritable identité du locuteur s codée sous la forme d’un vecteur un-chaud.

$$\mathcal{L}_{SE}(C_s^s|E^s) = \mathbb{E}_A CE(C_s^s(h^s), s) \quad (3.33)$$

où $CE(\dots)$ représente l’entropie croisée entre deux variables aléatoires ⁶.

Encodeur du contenu

L’encodeur de contenu E^c convertit la séquence de phonèmes p ou le signal de parole A en une représentation linguistique partagée H^c par le biais d’une perte contrastive (voir (Zhang et al., 2020) pour plus de détails) :

$$H^c = E^c(A) \quad (3.34)$$

L’apprentissage d’un encodage partagé entre les modalités audio et textuelle peut être lié à l’adaptation de domaine pour des données multimodales, dans laquelle on souhaite apprendre un code indépendant de la modalité d’entrée, dans la mesure où elles sont supposées encoder la même information. En l’occurrence, l’encodeur du contenu est entraîné pour transcrire le contenu phonétique à partir du signal vocal. Dans l’architecture présentée, l’encodage linguistique a la même longueur T que la séquence de phonèmes

⁶ On notera que l’entropie croisée peut être interprétée en termes de divergence de Kullback-Leibler entre les distributions des deux variables considérées, c’est-à-dire la quantité supplémentaire d’information nécessaire pour coder la vraie distribution mais avec l’utilisation des a priori issus de la distribution estimée. Par ailleurs, l’entropie croisée avec l’activation exponentielle normalisée peut être interprétée directement en termes d’information mutuelle entre les vraies étiquettes et les étiquettes prédites dans le cas d’une tâche de classification (Qin et al., 2015). Cette formulation présente l’interprétation possible de l’apprentissage de réseaux de neurones à la lumière de la théorie de l’information.

alignée (ainsi que le mel-spectrogramme), ce qui signifie que l'information temporelle est entièrement préservée tout au long de l'encodage.

La fonction de perte associée à la reconnaissance du contenu \mathcal{L}_{TC} est définie comme l'entropie croisée entre le phonème prédit à partir de \mathbf{h}_n^c par le classifieur C^c et l'étiquette du phonème réel correspondant \mathbf{p}_n pour la n -ème trame de temps :

$$\mathcal{L}_{TC}(C^c|E^c) = \mathbb{E}_{\mathbf{p}} CE(C^c(\mathbf{h}_n^c), \mathbf{p}_n). \quad (3.35)$$

Dissocier les informations sur le contenu et l'identité

Afin de démêler les informations associées au contenu et à l'identité du locuteur, une stratégie adversariale est introduite pour rendre l'encodage du contenu \mathbf{H}^c indépendant de l'identité du locuteur. L'idée du démêlage est la suivante : la séparation des informations évite de possibles phénomènes de redondance dans les encodages, ou de fuite de l'information d'un encodage à l'autre. Cette séparation est absolument essentielle pour les rendre opérationnelles lors de la génération, c'est-à-dire que la modification de ces codes entraîne une modification correspondante dans le signal de parole généré. On peut interpréter cette architecture comme une compression maximale de l'information véhiculée par le signal de parole pour permettre de le reconstruire de la manière la plus transparente possible. La fonction de perte associée à la classification du locuteur est définie comme l'entropie croisée entre l'identité du locuteur prédite à partir de \mathbf{h}_n^c par le classifieur C_s^c et l'identité réelle du locuteur \mathbf{s} . Une fonction de perte adversariale $\mathcal{L}_{ADV}(E^c)$ est additionnellement définie avec pour objectif opposé d'apprendre une représentation du contenu \mathbf{H}^c , à partir de laquelle l'identité du locuteur ne peut pas être reconnue par le classifieur du locuteur :

$$\mathcal{L}_{ADV}(E^c|C_s^c) = \mathbb{E}_{\mathbf{A}} \|\mathbf{u} - C_s^c(\mathbf{h}_n^c)\|_2^2 \quad (3.36)$$

où \mathbf{u} représente une valeur tirée sur une distribution uniforme dans laquelle tous les locuteurs ont la même probabilité $1/S$, S étant le nombre total de locuteurs dans la base de données.

Décodeur

Un décodeur G^a conditionné par les codes de contenu \mathbf{H}^c et de locuteur \mathbf{h}^s est utilisé pour reconstruire une approximation $\hat{\mathbf{A}}$ du signal de parole original \mathbf{A} :

$$\hat{\mathbf{A}} = G^a(\mathbf{h}^s = E^s(\mathbf{A}), \mathbf{H}^c = E^c(\mathbf{A})) \quad (3.37)$$

La fonction de perte de reconstruction \mathcal{L}_{RC} est définie entre le mel-spectrogramme du signal de parole reconstruit $\hat{\mathbf{A}}$ et le mel-spectrogramme du signal de parole original \mathbf{A} .

$$\mathcal{L}_{RC}(E^s, E^c, G^a) = \mathbb{E}_{\mathbf{A}} \|G^a(E^s(\mathbf{A}), E^c(\mathbf{A})) - \mathbf{A}\|_1 \quad (3.38)$$

Pendant l'apprentissage, le réseau de VC neuronale est pré-entraîné sur un ensemble de données composées de multiples locuteurs. Comme l'architecture VC repose principalement sur un auto-encodeur, il n'y a pas de manipulation ou de conversion d'attributs pendant l'apprentissage. Pendant la conversion, l'encodeur de contenu E^c calcule le code de contenu \mathbf{H}_{src}^c à partir du mel-spectrogramme \mathbf{A}_{src} d'un énoncé du locuteur source :

$$\mathbf{H}_{src}^c = E^c(\mathbf{A}_{src}) \quad (3.39)$$

En parallèle, l'encodeur de locuteur E^s calcule le code de locuteur \mathbf{h}_{tgt}^s à partir du mel-spectrogramme \mathbf{A}'_{tgt} d'un énoncé du locuteur cible :

$$\mathbf{h}_{tgt}^s = E^s(\mathbf{A}'_{tgt}) \quad (3.40)$$

Enfin, le décodeur G_a est conditionné sur les codes contenus \mathbf{H}_{src}^c et locuteur \mathbf{h}_{tgt}^s pour générer l'énoncé correspondant à $\hat{\mathbf{A}}_{tgt}$ avec l'identité du locuteur cible,

$$\hat{\mathbf{A}}_{tgt} = G_a(\mathbf{h}_{tgt}^s = E^s(\mathbf{A}'_{tgt}), \mathbf{H}_{src}^c = E^c(\mathbf{A}_{src})) \quad (3.41)$$

De cette manière, un énoncé ayant le contenu linguistique de l'énoncé source est prononcé avec l'identité du locuteur cible.

VC neuronale avec préservation du temps et de la F0

Dans la formulation du problème de la VC dans la littérature actuelle, la seule information qui est conservée de l'énoncé du locuteur source pendant la conversion est l'énoncé. Autrement dit il ne reste aucune caractéristique acoustique du locuteur source après conversion, le problème de la VC est formulé de telle sorte que la phase d'encodage consiste en une conversion parole-à-texte de compression du temps et la phase de décodage en une conversion texte-à-parole de décompression du temps. C'est un changement de positionnement important par rapport au problème d'origine (Stylianou et al., 1998) qui consistait à réaliser la transformation par modification directe sur le signal, ce qui présentait l'avantage de préserver une partie des caractéristiques du locuteur source. Ceci pose de sérieuses limitations pour des applications comme par exemple la post-synchronisation ou le doublage. D'une part, l'interprétation portée par la voix d'un acteur disparaît complètement lors de la conversion ou alors est générée de manière incontrôlée pendant la phase de synthèse. D'autre part, la synchronisation temporelle entre la conversion et l'énoncé source n'est pas assurée ce qui rend cette formulation impropre pour des exploitations audiovisuelles dans lesquelles la synchronisation temporelle du son et de l'image est essentielle en particulier en ce qui concerne la synchronisation labiale. Pour pallier cette limitation, nous avons proposé une stratégie de VC neuronale par transfert de timbre c'est-à-dire en préservant explicitement le rythme et les variations de F0 du locuteur source pendant la conversion. Cette formulation rend beaucoup plus flexible les applications en associant les compétences de l'acteur et de la machine : dans ce processus, l'acteur conserve l'intégralité de son interprétation. Cette formulation fonctionne de manière analogue à la capture de mouvement mais pour la voix : dans la capture de mouvement (MOCAP), un acteur réalise physiquement une interprétation dont la gestuelle est captée et utilisée comme un squelette sur lequel est appliqué une texture de synthèse pour réaliser un rendu 3D. Dans notre cas, le squelette gestuel est remplacé par le squelette vocal de l'acteur (structure textuelle et interprétation) sur lequel est appliqué le timbre de la voix cible. Pour réaliser cet objectif, nous avons donc proposé deux modifications sur l'architecture initiale : la préservation des temps et la préservation de la F0 pendant la conversion.

Synchronisation temporelle

Dans (Zhang et al., 2020), l'encodeur de contenu E^r effectue une opération de compression temporelle depuis les T trames temporelles du signal audio \mathbf{A} vers les N indices de phonèmes du code linguistique \mathbf{H}^r à l'aide d'un modèle S2S auto-régressif. De la même manière, le décodeur G^a effectue l'opération de décompression opposée depuis les N indices de phonèmes de \mathbf{H}^r (resp. \mathbf{H}^t) vers les T' trames du signal audio reconstruit $\hat{\mathbf{A}}$. Le décodeur a une structure similaire à celle du Tacotron, utilisé pour la synthèse de la parole (Wang et al., 2017b; Shen et al., 2018). Afin de préserver la synchronisation temporelle entre les signaux vocaux originaux et reconstruits, la dimension temporelle

de longueur T est préservée tout au long du réseau, du signal de parole original \mathbf{A} au code linguistique \mathbf{H}^t , et au signal de parole reconstruit $\hat{\mathbf{A}}$, comme illustré sur la [Figure 3.16](#). Un opérateur de compression temporelle D est utilisé pour calculer la fonction de perte de reconnaissance phonétique \mathcal{L}_{TC} . Pendant la conversion, l’encodage du contenu est réalisé uniquement à partir du signal de parole \mathbf{A} — ne nécessitant donc pas d’alignement préalable. Par ailleurs, la partie S2S auto-régressive de l’encodeur de contenu E^t et du décodeur G^a est modifiée en conséquence par l’utilisation d’architectures récurrentes simples. L’encodeur de contenu E^t est composé de deux couches LSTM bi-directionnelles de dimension 128 suivies d’une couche entièrement connectée (FC) de dimension 128, résultant en un code linguistique de dimension $(128 \times T)$. Le décodeur G^a utilise deux LSTM bi-directionnelles de dimension 128 chacune et une couche entièrement connectée de dimension 80 qui produit un mel-spectrogramme approximatif ayant les mêmes dimensions que le mel-spectrogramme d’entrée, c’est-à-dire $(80 \times T)$. Ces simplifications permettent des conversions synchrones en temps et un gain de temps de calcul conséquent : environ 33 % du temps de calcul pour l’apprentissage sur un serveur avec un seul GPU. Contrairement à ([Zhang et al., 2020](#)) où l’entrée linguistique est simplement la transcription phonétique du texte de l’énoncé prononcé, la VC neuronale synchronisée en temps nécessite de connaître en entrée à l’apprentissage l’alignement temporel des phonèmes avec le signal de parole. Pour ce faire, nous utilisons un algorithme d’alignement audio et texte pour produire un alignement forcé entre la séquence de phonèmes et le signal de parole représenté par le mel-spectrogramme \mathbf{A} .

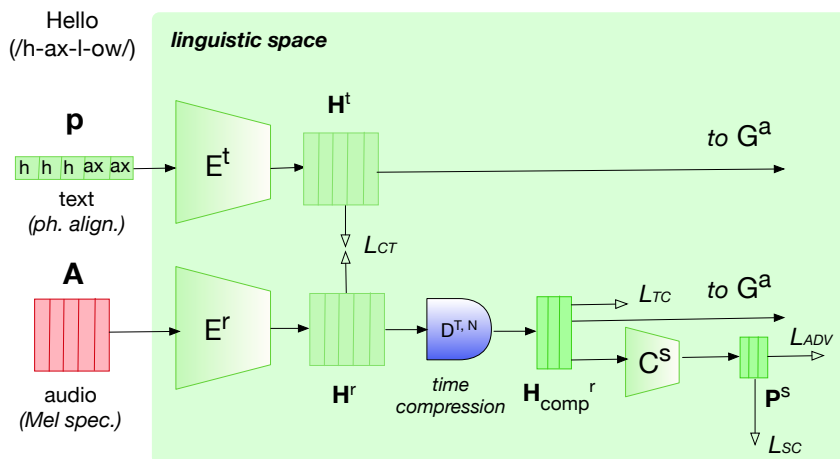


FIGURE 3.16 – Architecture du VC neuronale synchronisée en temps proposé. La synchronisation temporelle signifie que la parole convertie est alignée temporellement avec le signal de parole original. Par souci de simplicité, seule la partie linguistique de l’architecture est présentée.

Préservation de la F0

Afin de préserver la F0 du signal de parole source pendant la conversion, une fonction de perte sur la F0 est explicitement formulée. Pour ce faire, nous définissons un encodeur de F0 E^{F0} à partir du mel-spectrogramme :

$$\mathbf{h}^{F0} = E^{F0}(\mathbf{A}) \quad (3.42)$$

L’encodeur E^{F0} a la même architecture que celle décrite dans ([Roebel and Bous, 2022](#)). Il est pré-entraîné sur l’ensemble d’entraînement de la base de données vocales VCTK, puis fixé pendant l’entraînement VC, c’est-à-dire qu’il n’est utilisé que pour calculer la fonction de perte F0 \mathcal{L}_{F0} . La fonction de perte associée à la F0 est calculée comme l’erreur quadratique

moyenne entre les valeurs F0 du mel-spectrogramme reconstruit $\hat{\mathbf{h}}^{F0}$ et les valeurs F0 du mel-spectrogramme original \mathbf{h}^{F0} .

$$\mathcal{L}_{F0}(\mathbf{h}^{F0}, \hat{\mathbf{h}}^{F0}) = \mathbb{E}_{\mathbf{A}} \|\mathbf{E}^{F0}(\mathbf{A}) - \mathbf{E}^{F0}(\hat{\mathbf{A}})\|_2^2 \quad (3.43)$$

$$= \frac{1}{T} \sum_{t=1}^T (\mathbf{h}^{F0}(t) - \hat{\mathbf{h}}^{F0}(t))^2 \quad (3.44)$$

Pendant l'apprentissage, cette perte n'est rétro-propagée qu'à travers le décodeur G^a pour apprendre à générer un mel-spectrogramme correspondant à la séquence de F0 désirée. Cette proposition a l'avantage de formuler explicitement une contrainte sur la F0, contrairement à (Qian et al., 2020a) basée uniquement sur une stratégie de goulot d'étranglement de l'information dans lequel la préservation de la F0 est observée expérimentalement mais sans garantie. Cette perte est ajoutée à la perte de reconstruction :

$$\mathcal{L}_{GEN}(G_a) = \mathcal{L}_{RC}(G_a) + \lambda_{F0} \mathcal{L}_{F0}(G_a) \quad (3.45)$$

où λ_{F0} est une pondération variant linéairement de 10^{-6} à 10^{-2} , ce qui a pour effet d'augmenter progressivement l'importance de la préservation de la F0 par rapport à la perte de reconstruction pendant l'apprentissage.

L'architecture proposée à l'apprentissage et à l'inférence est présentée sur la Figure 3.17.

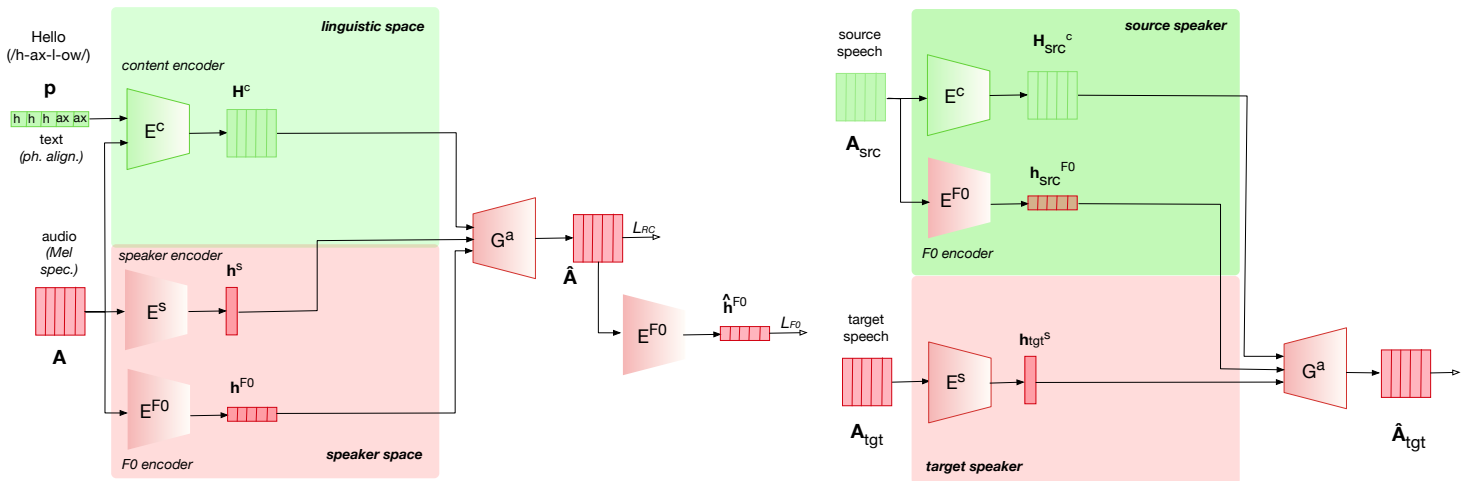


FIGURE 3.17 – Architecture de la VC neuronale proposée. À gauche : La VC à l'apprentissage comprend les encodeurs de contenu et de locuteur E^c et E^s , et un décodeur G^a qui reconstruit le signal de parole $\hat{\mathbf{A}}$ en fonction des codes de contenu et de locuteur \mathbf{H}^c et \mathbf{h}^s . Un encodeur de F0 pré-entraîné E^{F0} est utilisé pour calculer une perte de F0 entre la F0 du mel-spectrogramme de la parole originale et la F0 du mel-spectrogramme de la parole reconstruite. À droite : Lors de la conversion, le contenu \mathbf{H}_{src}^c et la F0 \mathbf{h}_{src}^{F0} sont encodés à partir du l'énoncé du locuteur source, et le code locuteur \mathbf{h}_{tgt}^s à partir d'enregistrements du locuteur cible. Source : (Bous et al., 2022)

Pendant la conversion, la F0 peut être transférée à partir d'un énoncé d'un locuteur source ou fixé arbitrairement (par exemple, en appliquant une transposition ou en définissant des valeurs arbitraires). Dans cet article, la F0 utilisée pour le conditionnement a été adaptée à la tessiture du locuteur cible afin d'éviter une conversion non naturelle qui serait due à une différence importante entre les tessitures respectives du locuteur source et du locuteur cible (typiquement lors de la conversion d'un homme en femme ou inversement). Pour ce faire, les valeurs de F0 correspondant à la phrase du locuteur source ont été normalisées par rapport à la moyenne et à l'écart-type du $\log(F0)$ du locuteur cible.

Détails d'implémentation

L'ensemble des détails d'implémentation de l'architecture proposée sont décrits dans (Bous et al., 2021). Le [Tableau 3.4](#) présente les différences d'implémentation de l'architecture VC par rapport à (Zhang et al., 2020).

TABLEAU 3.4 – Détails de configuration du modèle. FC désigne à une couche entièrement connectée, BLSTM une couche LSTM bi-directionnelle, et Tanh à la fonction d'activation tangente hyperbolique. La flèche droite \rightarrow indique l'ordre des couches successives du réseau.

E^r	2 couches BLSTM-Dropout(0.2), 256 cellules chaque direction \rightarrow FC-512-Tanh
E^s	2 couches BLSTM-Dropout(0.2), 128 cellules chaque direction \rightarrow agrégation par moyenne \rightarrow FC-128-Tanh
G^a	2 couches BLSTM, 64 cellules chaque direction \rightarrow FC-80

Expérience : conversion de l'identité

L'architecture proposée a été évaluée expérimentalement sur la base de données VCTK (Yamagishi et al., 2019) en anglais. La base de données VCTK contient les enregistrements de 110 locuteurs et les transcriptions textuelles correspondant aux énoncés prononcés. Chaque locuteur a lu environ 400 phrases sélectionnées dans des journaux anglais, ce qui représente un total d'environ 44 heures de parole. Tous les locuteurs sont inclus dans les ensembles d'apprentissage et de validation. Pour chaque locuteur, nous divisons la base de données en un ensemble d'apprentissage avec 90% des phrases et un ensemble de test avec 10% d'entre elles. La durée totale de la base de données est d'environ 27 heures après avoir supprimé les silences au début et à la fin de chaque enregistrement.

Méthodologie

L'évaluation expérimentale a été réalisée par des mesures objective et subjective. Tout d'abord, la préservation de la F0 a été évaluée en calculant la racine carrée de l'erreur quadratique moyenne entre la F0 du signal de parole original et la F0 du signal reconstruit pour les données issues de la base de test. L'expérience subjective suit les protocoles d'évaluation utilisés dans les tâches de conversion de l'identité (comme par exemple dans les challenges VCC (Lorenzo-Trueba et al., 2018; Zhao et al., 2020)). À partir d'un enregistrement présenté comme stimulus, le sujet doit juger de : 1) la similarité du signal de parole avec le locuteur cible, et 2) la qualité du signal de parole. Chacun de ces critères est évalué en utilisant une échelle MOS à 5 degrés. Pour la similarité : très similaire, plutôt similaire, faiblement similaire, plutôt dissimilaire, très dissimilaire, pour la qualité : excellente, bonne, passable, faible, mauvaise. Chaque participant devait juger 15 échantillons de parole sélectionnés au hasard parmi le nombre total d'échantillons de parole produits pour les expériences subjectives. L'expérience a été menée en ligne. Les participants ont été encouragés à réaliser l'expérience dans un environnement calme et au casque ou avec des écouteurs. Quatre locuteurs ont été utilisés pour l'expérience : deux hommes (p232 et p274) et deux femmes (p253 et p300) avec huit phrases choisies au hasard par locuteur dans l'ensemble de validation. Les conversions ont été calculées entre les locuteurs masculins (H \rightarrow H) et entre les locuteurs féminins (F \rightarrow F), ce qui a donné lieu à deux configurations de conversion masculines et deux configurations de conversion

féminines. Quatre configurations ont été comparées : 1) le signal audio original, et le signal de parole converti avec : 2) la VC synchrone en temps avec préservation de la F0, (désigné par F0 *cond.*), 3) la VC synchrone en temps avec préservation de la F0 et avec un discriminateur entraîné uniquement sans conversion d'identité (désigné par F0 *cond. w/adv same id*), et 4) la VC synchrone en temps avec préservation de la F0 et avec un discriminateur entraîné avec conversion d'identité (désigné par F0 *cond. w/adv diff id*).

Résultats et discussion

Le **Tableau 3.5** présente la racine carré de l'erreur quadratique moyenne de la F0 du signal de parole converti avec les configurations proposées pour les 4 locuteurs utilisés dans l'expérience. L'erreur est d'environ 5 Hz en moyenne, ce qui n'est pas audible dans la plupart des cas ('t Hart, 1981). Cela démontre l'efficacité de la stratégie de préservation de la F0 proposée, même en combinaison avec un apprentissage adversarial.

TABLEAU 3.5 – Racine carrée de l'erreur quadratique moyenne sur la F0 du signal de parole converti (en Hz.).

VC system	M-to-M	F-to-F	M-to-F	F-to-M
F0 cond w/ adv. same id	2.712	5.686	2.970	5.090
F0 cond w/ adv. diff id	2.574	6.246	4.171	5.252

Le **Tableau 3.6** présente les scores MOS moyens obtenus auprès de 25 participants pour les configurations comparées et les enregistrements de parole de référence.

TABLEAU 3.6 – Scores MOS moyens obtenus pour les configurations de VC comparées.

VC system	Male-to-Male		Female-to-Female		TOTAL	
	Similarity	Naturalness	Similarity	Naturalness	Similarity	Naturalness
Orig : target speaker	4.92	4.94	4.98	4.97	4.98	4.96
F0 cond.	3.90	3.38	3.93	2.85	3.92	3.09
F0 cond w/ adv. same id	3.90	3.15	3.94	2.91	3.96	3.14
F0 cond w/ adv. diff id	3.91	3.16	4.23	3.21	4.06	3.18

Premièrement, l'algorithme VC de base utilisant la synchronisation temporelle et la préservation de la F0 présente une similarité avec le locuteur cible de passable à bonne (MOS=3.92 pour la similarité), même si les durées et la F0 sont hérités du locuteur source. Le naturel est également jugé passable (MOS=3.09). On peut observer que le naturel est moins bon pour les locuteurs féminins (MOS=2.85) que pour les locuteurs masculins (MOS=3.38). Cela indique que l'application des durées et de la F0 provenant d'un locuteur différent ne dégrade pas de manière substantielle la similarité et la qualité de la parole convertie.

Deuxièmement, l'algorithme VC proposé avec perte adversariale sur les identités variables améliore les scores dans quasiment tous les cas par rapport à l'algorithme VC de base. La similarité globale avec le locuteur cible est bonne (MOS=4.06) et la qualité de la conversion est moyenne (MOS=3.18). L'amélioration de la similarité est particulièrement importante pour les locuteurs féminins (MOS=4.23) alors que, dans le même temps, la différence de qualité entre les conversions masculines et féminines est beaucoup moins prononcée (MOS=3.16 pour les locuteurs masculins et MOS=3.21 pour les locuteurs féminins). Cela indique que l'ajout du discriminateur permet non seulement d'améliorer la qualité de la parole convertie (en supprimant les artefacts perceptibles) mais aussi d'augmenter la similarité avec le locuteur cible. En outre, l'utilisation de différentes identités de locuteur avec la perte adversariale améliore les scores dans tous les cas par rapport à l'utilisation de la seule identité réelle du locuteur. Ceci est probablement dû au

fait que le discriminateur est plus efficace lorsqu'il est soumis à une plus grande variété de phrases et d'identités de locuteurs.

Démêlage des attributs vocaux par apprentissage adversarial

Dans la section précédente, nous avons présenté une architecture VC basée sur l'encodage du contenu et de l'identité du locuteur. Dans la présente section, nous présentons une recherche préliminaire pour tendre vers l'apprentissage de représentations démêlées à partir d'un signal de parole permettant l'encodage et la manipulation d'attributs vocaux. La contribution principale, appliquée à la manipulation d'un attribut a priori simple — le genre — repose sur l'intégration d'un encodage en cascade du code de l'identité du locuteur en un code non-genré et un code de genre. Cette décomposition est réalisée par apprentissage adversarial à partir du code du locuteur avec une architecture à curseurs (Lample et al., 2017), comme illustré sur la Figure 3.18, et comme proposée en parallèle pour la représentation des attributs d'un locuteur dans le domaine de la préservation de la vie privée (Noé et al., 2021). L'architecture Fader est un réseau génératif qui a été proposé pour la manipulation des attributs dans une image. Cette architecture prend la forme générale d'un auto-encodeur dans lequel le décodeur est conditionné en plus sur la variable d'attribut. Pour rendre ce conditionnement opérationnel, une stratégie adversariale est introduite pour faire en sorte que le code résultant de l'encodeur devienne indépendant sur la variable correspondant à l'attribut. Cette architecture a montré des résultats particulièrement prometteurs pour la manipulation d'attributs binaires d'un visage humain — mais également et de manière plus surprenante de leur interpolation continue. Nous en reprenons ici l'idée pour l'appliquer à la manipulation des attributs de la voix humaine.

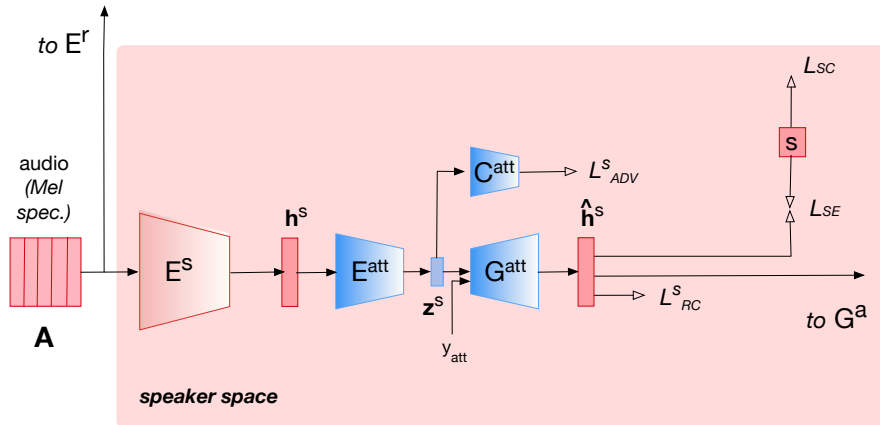


FIGURE 3.18 – Architecture proposée pour l'apprentissage démêlé de l'identité du locuteur. Le code du locuteur \mathbf{h}^s est démêlé en un code d'attribut att et un code de locuteur \mathbf{z}^s indépendants sur l'attribut att. Par souci de simplicité, seul l'espace locuteur de l'architecture est présenté dans cette figure. Source : (Benaroya et al., 2023)

Le réseau à curseurs est un auto-encodeur dans lequel le code locuteur \mathbf{h}^s est encodé par E^{att} en un code latent de faible dimension \mathbf{z}^s ,

$$\mathbf{z}^s = E^{\text{att}}(\mathbf{h}^s) \quad (3.46)$$

Inversement, le décodeur G^{att} tente de reconstruire le code locuteur $\hat{\mathbf{h}}^s$ à partir du code latent \mathbf{z}^s et de la variable d'attribut y_{att} ,

$$\hat{\mathbf{h}}^s = G^{\text{att}}(\mathbf{z}^s, y_{\text{att}}) \quad (3.47)$$

L'objectif du réseau Fader est de pouvoir reconstruire la variable d'entrée $\hat{\mathbf{h}}^s$ à partir du code latent \mathbf{z}^s et de la variable de conditionnement y_{att} . La question principale de ce problème est que l'information de genre est déjà encodée dans la variable \mathbf{h}^s , ce qui fait que le code d'attribut y_{att} est redondant avec cette information et donc probablement inopérant. La solution proposée par le réseau Fader est d'ajouter un module adversarial qui va permettre de rendre la variable de code latent \mathbf{h}^s indépendante de la variable d'attribut y_{att} , les deux informations étant par ailleurs complémentaires.

Tout d'abord, une perte de reconstruction de l'auto-encodeur $\mathcal{L}_{\text{RC}}^{\text{S}}$ est définie comme l'erreur absolue moyenne entre le code locuteur \mathbf{h}^s et le code locuteur reconstruit $\hat{\mathbf{h}}^s$:

$$\mathcal{L}_{\text{RC}}^{\text{S}}(\text{E}^{\text{att}}, \text{G}^{\text{att}}) = \mathbb{E}_{\mathbf{h}^s} \|\mathbf{h}^s - \text{G}^{\text{att}}(\text{E}^{\text{att}}(\mathbf{h}^s), y_{\text{att}})\|_1 \quad (3.48)$$

L'objectif de cette première fonction de perte est que l'encodeur E^{att} code l'information \mathbf{z}^s de telle sorte que le décodeur G^{att} soit capable de reconstruire l'entrée originale à partir du code latent \mathbf{z}^s et de l'attribut de conditionnement y_{att} .

Deuxièmement, une perte du discriminateur $\mathcal{L}_{\text{D}}^{\text{S}}$ est définie comme l'entropie croisée entre l'attribut prédit par le classifieur C^{att} et l'attribut réel y_{att} , représenté sous la forme d'un vecteur un-chaud :

$$\mathcal{L}_{\text{D}}^{\text{S}}(\text{C}^{\text{att}}|\text{E}^{\text{att}}) = \mathbb{E}_{\mathbf{h}^s} \text{CE}(y_{\text{att}}, \text{C}^{\text{att}}(\text{E}^{\text{att}}(\mathbf{h}^s))) \quad (3.49)$$

L'objectif de cette deuxième fonction de perte est que le classifieur C^{att} soit capable de prédire l'attribut correct y_{att} à partir du code latent \mathbf{z}^s .

Troisièmement, une fonction de perte adversariale $\mathcal{L}_{\text{ADV}}^{\text{S}}$ est définie comme l'entropie croisée entre l'attribut prédit par le classifieur C^{att} et l'attribut erroné $1 - y_{\text{att}}$:

$$\mathcal{L}_{\text{ADV}}^{\text{S}}(\text{E}^{\text{att}}|\text{C}^{\text{att}}) = \mathbb{E}_{\mathbf{h}^s} \text{CE}(1 - y_{\text{att}}, \text{C}^{\text{att}}(\text{E}^{\text{att}}(\mathbf{h}^s))). \quad (3.50)$$

L'objectif de cette fonction de perte est que le classifieur C^{att} ne soit pas capable de prédire l'attribut correct y_{att} à partir du code latent \mathbf{z}^s . Cette perte est définie de manière à rendre le code latent \mathbf{z}^s indépendant sur la variable y_{att} .

Finalement, la fonction de perte adversariale totale du réseau Fader peut s'écrire comme,

$$\mathcal{L}_{\text{RC}}^{\text{S}}(\text{E}^{\text{att}}, \text{G}^{\text{att}}|\text{C}^{\text{att}}) = \mathcal{L}_{\text{RC}}^{\text{S}}(\text{E}^{\text{att}}, \text{G}^{\text{att}}) + \lambda \mathcal{L}_{\text{ADV}}^{\text{S}}(\text{E}^{\text{att}}|\text{C}^{\text{att}}). \quad (3.51)$$

Expérience : conversion de genre

L'architecture proposée a été évaluée expérimentalement sur la base de données VCTK précédemment décrite, et en utilisant les méta-données sur le genre des locuteurs de la base (en l'occurrence : homme ou femme).

Illustration préliminaire

La [Figure 3.19](#) présente quatre spectrogrammes superposés avec les contours de F0 (en traits pleins rouges). La phrase "Ask her to bring these things with her from the store" est prononcée par un locuteur (à gauche) et par une locutrice (à droite). Les figures du haut montrent les signaux originaux et les figures du bas correspondent à la conversion conditionnée sur genre opposé. L'algorithme de conversion du genre transpose clairement la F0 moyenne conformément à ce que nous aurions fait pour convertir entre des locuteurs masculins et féminins à l'aide d'un vocodeur traditionnel (± 1 octave) ([Farner et al., 2009](#)). Cependant, contrairement à ce que nous aurions fait en utilisant des vocodeurs traditionnels,

la transposition est ici dynamique, modifiant également les contours de l'intonation. En outre, l'algorithme crée une voix craquée sur les derniers mots de l'énoncé lors de la conversion d'une voix masculine à une voix féminine, alors qu'il fait l'inverse lors de la conversion d'une voix féminine à une voix masculine. Nous faisons l'hypothèse que la présence ou l'absence de craquement vocal reflète une tendance générale des voix masculines et féminines présentes dans la base de données.

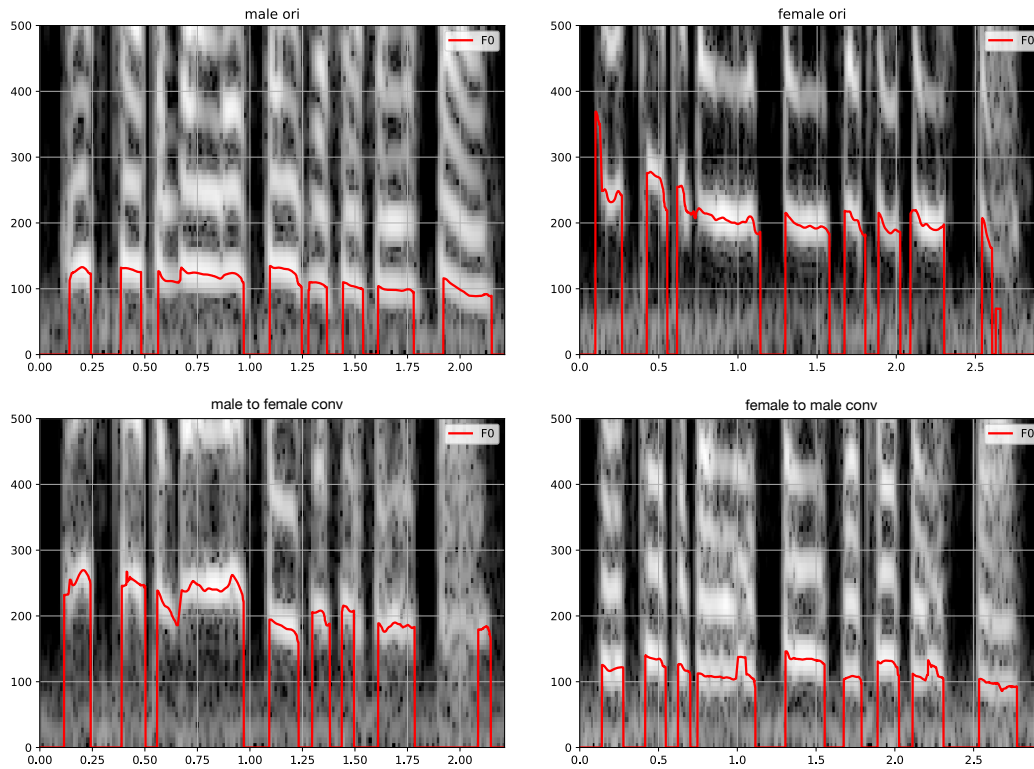


FIGURE 3.19 – Visualisation des spectrogrammes et des courbes de F0 pour l'énoncé "Ask her to bring these things with her from the store.". En haut : deux spectrogrammes des signaux de parole originaux d'un locuteur masculin (gauche) et d'une locutrice féminine (droite). En bas : Spectrogrammes des deux signaux après conversion du genre à l'aide de l'architecture proposée. Le trait rouge continu correspond à la F0. L'axe des y indique la fréquence en Hertz, tandis que l'axe des x indique le temps en secondes. Source : (Benaroya et al., 2023)

Expérience objective : information mutuelle et visualisation des codes

Le **Tableau 3.7** présente l'estimation de l'information mutuelle entre le vrai genre avec le code locuteur d'origine et le code locuteur reconstruits avec différents conditionnements de l'attribut de genre. Ce score est calculé à l'aide d'un estimateur de l'information mutuelle entre les variables discrètes et continues, comme décrit dans (Gao et al., 2017). La dimension de la variable continue est réduite de 128 à 8 à l'aide d'une analyse en composantes principales (ACP), préalablement au calcul de l'information mutuelle. Les coordonnées de l'ACP utilisées pour tracer les visualisations sur la **Figure 3.20** sont sélectionnées comme celles maximisant l'information mutuelle avec l'information de genre. L'information mutuelle correspondant au code latent \mathbf{z}^s est très faible, notamment par comparaison à celle obtenue pour le code locuteur \mathbf{h}^s . Ceci indique que le code latent \mathbf{z}^s devient essentiellement indépendant du genre, comme l'illustre en complément la **Figure 3.20**. Par ailleurs, la mesure de l'information mutuelle nous permet également de vérifier que le conditionnement sur la variable de genre est effective pour reconstruire un code

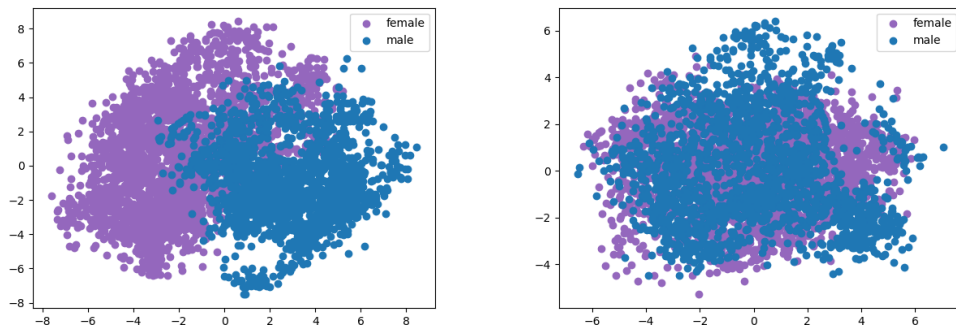


FIGURE 3.20 – À gauche : visualisation par ACP des encodages du locuteur \mathbf{h}^s sur l'ensemble d'évaluation; les composantes sélectionnées sont 1 et 2. À droite : Visualisation par ACP des codes latents \mathbf{z}^s sur l'ensemble d'évaluation; les composantes sélectionnées sont 3 et 7. Source : (Benaroya et al., 2023)

TABLEAU 3.7 – Estimation de l'information mutuelle entre le vrai genre et le code locuteur \mathbf{h}^s , le code latent \mathbf{z}^s , et les codes locuteurs reconstruits $\hat{\mathbf{h}}^s$ avec différents conditionnements w .

information mutuelle	
Code original \mathbf{h}^s	0.47
Genre est. ($w = \tilde{w}$)	0.44
Genre inv. ($w = 1 - \tilde{w}$)	0.38
Dé-genre ($w = 1/2$)	0.16
Code latent \mathbf{z}^s	0.11

locuteur genré $\hat{\mathbf{h}}^s$: l'information mutuelle avec le conditionnement sur le genre réel ou son opposé est élevée et comparable à celle calculée avec le code locuteur d'origine. Par ailleurs, le conditionnement avec un poids intermédiaire $w = 1/2$ entre les genres possède également une information mutuelle faible, ce qui suggère qu'il est possible d'interpoler continûment dans l'espace de genre.

Dans (Benaroya et al., 2023), nous présentons une série d'évaluations objectives supplémentaires dont je fais l'économie dans ce manuscrit sur la reconnaissance du genre et de l'identité du locuteur sur les codes locuteurs, le code latent, et les codes locuteurs reconstruits avec différents conditionnements sur le genre. Les résultats principaux sont que : le conditionnement sur le genre est effectif sur la capacité du classifieur de genre à identifier le genre à partir du code, et un classifieur d'identité du locuteur est capable de déterminer l'identité du locuteur à partir des codes reconstruits avec n'importe quel conditionnement de genre. De manière extrêmement intéressante, ce dernier résultat indique qu'il est possible d'encoder l'identité d'un locuteur indépendamment de son genre. En d'autres mots, il existe une identité non-genrée du point de vue de la machine.

Expérience subjective

Une évaluation subjective a été réalisée pour déterminer si l'architecture proposée permet de convertir efficacement le genre dans la voix.

Algorithme de référence Il n'existe pas à notre connaissance d'algorithme neuronal de conversion du genre dans la littérature. Nous avons donc utilisé un algorithme de

traitement du signal comme algorithme de référence pour l'expérience perceptive. Les algorithmes classiques de transformation de la voix manipulent le genre en modifiant la moyenne de la fréquence fondamentale (F0) et les positions des résonances du conduit vocal (appelées formants). En raison des différences physiologiques entre les organes vocaux féminins et masculins, notamment la taille des plis vocaux et du conduit vocal, ces deux paramètres ont des valeurs moyennes qui diffèrent statistiquement pour les voix masculines et féminines. Ces différences ont été mesurées et documentées dans la littérature (Peterson and Barney, 1952; Iseli et al., 2007). Considérant que ces paramètres font partie des configurations physiologiques des organes vocaux, ils font partie de l'identité du locuteur; il a été montré dans (Farner et al., 2009) qu'une transposition constante et indépendante de la F0 et des formants peut être utilisée pour modifier avec succès le genre et l'âge perçus d'une voix. À la suite de ces conclusions, nous utilisons les paramètres suivants pour la conversion du genre : la F0 est transposée de \pm une octave (± 1200 cents) et l'enveloppe spectrale est transposée de ± 3 demi-tons (c'est-à-dire ± 300 cents). Une transposition positive est utilisée pour la conversion homme-femme, tandis qu'une transposition négative est utilisée pour la conversion femme-homme. Un vocodeur de phase à forme d'onde invariante (Röbel, 2010) est utilisé pour la manipulation du signal en utilisant l'estimateur d'enveloppe spectrale présenté dans (Röbel and Rodet, 2005) pour la représentation des filtres résonateurs.

Protocole expérimental La tâche consistait à écouter un échantillon de parole (convertie ou non) et à juger si la voix est typiquement perçue comme : *féminine*, *plutôt féminine*, *incertain*, *plutôt masculine*, ou *masculine*. Chaque sujet a dû juger vingt échantillons de parole choisis au hasard parmi tous les échantillons de parole produits pour les expériences subjectives. Quatre locuteurs de la base VCTK ont été utilisés pour l'expérience, deux hommes (p232 et p274) et deux femmes (p253 et p300), avec cinq phrases choisies au hasard par locuteur dans l'ensemble de validation. Six configurations ont été comparées :

- le signal de parole original (*True*);
- le signal de parole converti avec un vocodeur de phase de référence (*phase voc.*);
- le signal de parole converti avec l'algorithme VC original (*VC*);
- le signal de parole converti avec l'algorithme VC proposé (*base*) pour cinq valeurs de conditionnement du paramètre de genre $\tilde{w} \in \{0, 1/4, 1/2, 3/4, 1\}$;
- le signal de parole converti avec l'algorithme VC proposé mais sans la perte adversariale de l'encodeur de genre (*nofader*), pour cinq valeurs du paramètre \tilde{w} ;
- le signal de parole converti avec l'algorithme VC proposé, avec le décodeur ré-entraîné avec l'encodeur de genre (*adapt*), pour cinq valeurs du paramètre \tilde{w} .

Résultats et discussion

La Figure 3.21 présente les scores obtenus sur le genre et la qualité perçus pour les configurations comparées (moyenne et intervalle de confiance à 95%). En ce qui concerne la qualité perçue, les échantillons du signal de parole original ont le score le plus élevé (4.6), les échantillons générés avec l'algorithme VC neuronale ont des scores similaires à ceux rapportés dans (Zhang et al., 2020) (2.90), et les échantillons convertis avec le vocodeur de phase ont des scores assez bas (1.6). Ce dernier score est vraisemblablement dû à l'utilisation de paramètres par défaut, mais indique aussi clairement la limitation de la conversion vocale basée sur des algorithmes de traitement du signal. Les trois versions de l'algorithme VC proposé ont des scores similaires et comparables à ceux de l'algorithme original (entre 3.0 et 4.0) : 2.9 pour la VC de *base*, 3.11 pour la VC *nofader*, et 2.97 pour la VC *adapt*. VC. Ce résultat montre que l'intégration de l'encodeur de genre ne dégrade pas la qualité de la conversion. On observe par ailleurs une dégradation de la qualité pour les conversions de femme vers homme. Néanmoins, cette tendance tend à disparaître avec

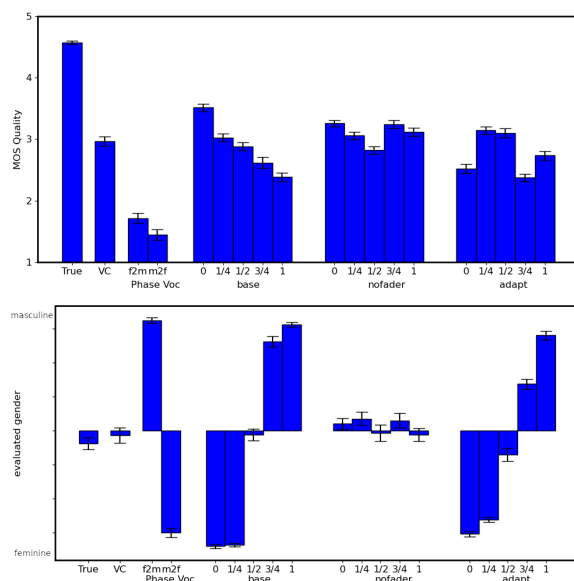


FIGURE 3.21 – En haut : scores MOS obtenus pour les six configurations (moyenne et intervalle de confiance à 95%). En bas : genre de voix perçu pour les six configurations (moyenne et intervalle de confiance à 95%). Source : (Benaroya et al., 2023)

la configuration de la VC *adapt* c'est-à-dire lorsque le décodeur est ré-entraîné en même temps que l'encodeur de genre, ce qui démontre l'intérêt d'adapter le décodeur à partir des nouveaux codes locuteurs.

Pour le genre perçu, le genre réel des locuteurs est facilement reconnu pour les échantillons de parole originaux, le signal de parole converti avec l'algorithme VC original, et le signal de parole converti avec le vocodeur de phase. Comme attendu, l'algorithme VC avec l'encodeur de genre sans perte adversariale est totalement inefficace pour convertir le genre. Pour l'algorithme VC avec l'encodeur de genre incluant la perte adversariale, le conditionnement sur le genre est efficace pour manipuler le genre perçu pendant la conversion : le genre perçu varie de manière consistante avec le genre indiqué par le conditionnement. On observe néanmoins une discontinuité pour les valeurs intermédiaires, en particulier pour les valeurs autour de $w = 1/2$ correspondant donc à un genre intermédiaire. Cette discontinuité est particulièrement marquée pour l'algorithme VC de *base* : dans la mesure où l'algorithme ne voit pas d'exemples intermédiaires au cours de l'apprentissage, il saute brusquement d'une voix féminine à une voix masculine et échoue à générer des voix avec un genre intermédiaire. Cette transition semble en revanche moins marquée pour l'algorithme VC *adapt*, ce qui indique encore une fois que le ré-entraînement du décodeur peut améliorer la conversion y compris en terme de genre généré.

3.4 SYNTHÈSE ET DISCUSSION

Les contributions principales de la deuxième moitié de ce chapitre sont :

Contributions

C1. La conversion neuronale de l'identité et l'expressivité à partir de représentations mel-spectrogramme (Benaroya et al., 2023; Le Moine, 2023) utilisant un vocodeur neuronal universel (Roebel and Bous, 2022).

C2. L'intégration explicite d'information linguistique dans l'apprentissage de la conversion de l'émotion (Le Moine, 2023) et l'apprentissage de représentations structurées de la voix pour la conversion de l'identité (Bous et al., 2022). En particulier, l'apprentissage de représentations démêlées notamment par stratégie adversariale permet d'étendre la conversion de la voix à de nouveaux attributs, comme le genre (Benaroya et al., 2023).

C3. La progression graduelle de l'apprentissage à partir de bases appareillées ou parallèles (Le Moine, 2023) à l'apprentissage à partir de bases non contraintes soit par architecture auto-encodeur (Benaroya et al., 2023) soit par architecture adversariale (Bous et al., 2022).

Je présente enfin les conclusions finales aux questions de recherche pour ce dernier chapitre. Cette conclusion démontre finalement que les trois questions finissent par être entrelacées en une seule grande question d'apprentissage.

Q1. Quelles représentations ? La représentation mel-spectrogramme permet d'encoder l'intégralité des caractéristiques acoustiques de la voix tout en proposant un encodage compressé du signal de parole.

La représentation mel-spectrogramme est une représentation compressée avec perte basée sur une échelle Mel logarithmique en fréquence reproduisant la sélectivité fréquentielle de la perception auditive humaine. Cette représentation est incomplète : elle n'est pas inversible avec reconstruction parfaite. Néanmoins, des algorithmes de reconstruction ont été proposés pour reconstruire la forme d'onde à partir du mel-spectrogramme, soit à partir de modèles de signaux (Griffin and Lim, 1985) soit plus récemment avec des vocodeurs neuronaux (Van den Oord et al., 2016; Roebel and Bous, 2022) avec une reconstruction quasiment transparente. La représentation mel-spectrogramme est une représentation non-paramétrique qui contient toute l'information du signal de parole et offre donc la possibilité d'apprendre directement les interrelations complexes entre les caractéristiques acoustiques à partir d'une représentation complète du signal de parole. Des alternatives paramétriques existent pour un contrôle explicite des paramètres vocaux, comme par exemple le vocodeur neuronal *WORLD* (Morise et al., 2016). Mais cette représentation — mel-spectrogramme ou représentation paramétrique — n'est qu'un point d'entrée : la représentation effective étant déterminée par l'apprentissage en même temps et en fonction de la tâche de conversion.

→ La représentation mel-spectrogramme semble être une étape intermédiaire vers de nouvelles formes d'encodage des signaux de parole. En particulier, les stratégies d'apprentissage auto-supervisé comme celles utilisées pour les modèles de langage (Devlin et al., 2019) et adapté aux signaux audio (Chung and Glass, 2018; Tagliasacchi et al., 2020) permettent d'apprendre des représentations universelles et contextualisées en fonction des modalités considérées. En particulier, les représentations audio multi-échelles (Zeghidour et al., 2022; Défossez et al., 2022) peuvent être particulièrement intéressantes pour modéliser la multi-temporalité de la parole — en particulier sur des tâches de génération et de manipulation de signaux de parole.

Q2. Quelles données ? L'évolution des paradigmes d'apprentissage a permis de s'affranchir de la contrainte des bases appareillées et parallèles. Les nouveaux paradigmes d'apprentissage ne nécessitent plus de bases parallèles ou appareillées, ce qui ouvre potentiellement la capacité d'apprendre à partir de grandes masses de données — idéalement à partir de tous les enregistrements de parole accessibles en ligne. Les quantités de données utilisées pour l'apprentissage sont en quelques années passées de quelques heures d'un locuteur à des milliers d'heures en multi-locuteurs (Yamagishi et al., 2019).

→ La contrainte pour la génération reste encore forte : l'apprentissage de réseaux génératifs de parole nécessite encore à ce jour des données propres : locuteur isolé, parole monologuée, conditions d'enregistrements bonnes et homogènes. Pour s'affranchir de cette contrainte, il est nécessaire de formaliser des algorithmes capables de générer de la parole artificielle propre — ou avec des conditions spécifiées et contrôlées — à partir de données hétérogènes et dégradées. Par ailleurs, le problème de l'annotation — généralement produite par des humains, lents et imprécis — constitue toujours un facteur limitant et une question ouverte, par exemple pour modéliser les facteurs de variabilité para-linguistiques, comme les émotions ou les attitudes sociales. Enfin, la présence de biais dans les données (Karkkainen and Joo, 2021) n'a pas encore été posée à ce jour dans la communauté de la parole. Leur considération est essentielle pour que les données reflètent la diversité langagière et individuelles de la parole humaine — que ce soit pour la préservation de langues minoritaires ou la manipulation continue du concept de genre plus fluide et représentatif de la diversité observée. Les initiatives Common Voice de Mozilla ou le Voice Lab en France constituent de premières tentatives dans le sens de données ouvertes et variées, mais qui reste largement à développer et à réguler — y compris juridiquement.

Q3. Quels modèles d'apprentissage? La spécification de modèles d'apprentissage structurés rend possible l'apprentissage efficace des informations véhiculées dans un signal de parole à partir des données à la volée — c'est-à-dire sans contrainte. Le couplage avec un vocodeur neuronal permet de décomposer un problème complexe en deux problèmes plus simples ciblés sur deux tâches différentes. Les paradigmes d'apprentissage d'auto-encodeur et adversarial ont permis d'apprendre à partir de bases de données à la volée. L'encodeur-décodeur permet de définir une fonction de perte de reconstruction, la sortie étant l'entrée connue; l'architecture adversariale permet d'apprendre des générations pour des exemples dont on ne possède pas d'observation réelle et par conséquent d'aggrandir artificiellement la quantité de données exploitables pour l'apprentissage. La spécification d'une représentation structurée de la parole humaine intégrée dans l'architecture d'apprentissage couplée à des stratégies d'apprentissage de représentations démêlées permet d'encoder de manière différenciée les informations véhiculées par la parole humaine comme le contenu linguistique, l'identité du locuteur, et son genre. Cette architecture peut être directement interprétée à la lumière de la théorie de l'information, comme des systèmes de codage et de décodage de l'information ([Dudley, 1939](#); [Huffman, 1952](#); [ISO/IEC 13818-3, 1998](#)) mais dont les codes neuronaux sont appris à partir des données. Les recherches présentées dans ce chapitre constituent une preuve de concept de cette voie de recherche.

→ Les trois questions de recherche initialement présentées peuvent en réalité se formuler en une seule question de recherche : les données, les représentations et les modèles d'apprentissage ne sont que les trois faces d'un seul et même problème général de l'apprentissage et de l'encodage de l'information par la machine à partir de données. La spécification du problème d'apprentissage permet de relâcher les contraintes sur les données et de rendre possible l'apprentissage à partir de données à la volée, voir même à partir de données issues d'autres modalités ([Hawthorne et al., 2022](#)). Autrement dit : les spécifications du modèle permettent d'augmenter le nombre de degrés de liberté sur les données. La spécification des modèles d'apprentissage permet également d'apprendre directement les représentations à partir des données à la volée. Les modèles de signaux paramétriques ne sont pas totalement rejetés, leur hybridation à des architectures neuronales peut faciliter l'apprentissage et diminuer la complexité des modèles. Autrement dit : la spécification de connaissances humaines dans les modèles d'apprentissage facilite leur apprentissage et leur interprétabilité — par des humains.

CONCLUSION ET PERSPECTIVES

La conclusion de ce manuscrit prendra la forme d'une ouverture — la recherche scientifique est par essence une "œuvre ouverte", un "*work-in-progress*" — et d'une tentative d'identifier les défis actuels et futurs de la modélisation et la génération de la parole à l'ère de l'intelligence artificielle. J'initierai cette conclusion par un bilan et une ouverture sur les défis scientifique et technique, avant de conclure par une réflexion plus générale sur mon statut de chercheur dans le monde et les sociétés d'aujourd'hui et de demain.

CONSIDÉRATIONS GÉNÉRALES SUR LA GÉNÉRATION DE LA PAROLE

Depuis 2018, il est possible de créer des voix artificielles jugées aussi réalistes que des voix humaines (Shen et al., 2018) à partir d'environ 20 heures d'enregistrements de cette voix ou de transférer le style de parole d'une voix sur une autre (Wang et al., 2018; Pan and He, 2021). Depuis lors, les recherches sur la modélisation générative de la parole se sont multipliées et accélérées : en l'espace de 5 ans, il est maintenant possible de transférer l'identité vocale d'une personne à partir d'un échantillon de 3 à 5 secondes de sa voix (Wang et al., 2023; Betker, 2023) avec des bases de données de 50,000 à 60,000 heures d'enregistrements de parole. Également, la synthèse multilingue n'est aujourd'hui plus quelque chose d'inimaginable — et même plutôt bien avancée. Il semble ainsi possible aujourd'hui de générer et éditer de la parole artificielle avec n'importe quelle identité et dans n'importe quelle langue (Le et al., 2023). La recherche sur la génération de la parole est-elle pour autant close? La réponse est non, et j'en expose les raisons principales. La parole accuse toujours un retard historique par rapport aux modalités du texte et de la vision comme les deep fakes (Paris and Donovan, 2019), Chat-GPT (OpenAI, 2022), Dall-E et MidJourney (Ramesh et al., 2022), ou Drag your GAN (Pan et al., 2023). Mais ce n'est pas la seule raison, ni la raison la plus importante.

L'apprentissage de larges modèles (Betker, 2023) à partir de bases à la volée trouve des limitations importantes à la fois scientifique et technique (Ettinger, 2020). La taille actuelle des modèles de génération pose des problèmes importants d'intégration et de portabilité, et leur puissance est sans commune mesure avec la puissance d'un cerveau humain — de l'ordre de 40 Watts. Les architectures proposées telles que (Wang et al., 2023; Betker, 2023) sont largement sous-spécifiées et ne bénéficient principalement que de la quantité de données utilisées pour l'apprentissage. Au-delà de la course à la performance, l'apprentissage neuronal doit être repensé et des alternatives légères, soutenables — notamment énergétiquement et écologiquement —, et explicables, doivent être proposées. La spécification de connaissances humaines expertes dans de telles architectures neuronales contribuera à tendre vers ces objectifs tout en participant à l'accroissement de la connaissance humaine sur le langage, la parole, et la communication parlée. Par ailleurs, la génération de parole à partir de bases à la volée introduit une quantité de biais incontrôlés et non quantifiés actuellement. Le plus important de ces biais est la proportion des langues représentées dans le monde numérique. L'intelligence artificielle et les grands modèles de langage ont une tendance à fortement amplifier cette distorsion et par conséquent à accélérer le phénomène de disparition de langues minoritaires, peu ou pas dotées, et des cultures associées. Des propositions ont été faites dans le domaine de la vision pour limiter ces biais (Karkkainen and Joo, 2021) à l'origine dans les données, en revanche il n'existe pas vraiment à ce jour de formalisation satisfaisante pour leur prise en compte durant l'apprentissage et l'effet de ces biais demeure largement une inconnue. La question principale n'est pas tant : doit-on éliminer les biais? Ils sont probablement inhérents à l'humain, cognitivement et statistiquement. Il est en revanche important de

les identifier et de mieux en comprendre les causes et les effets sur la génération pour en améliorer la compréhension et la maîtrise. Plus spécifiquement, le phénomène de “*fuite d’information*” (en anglais, “*leakage*”) est caractéristique des méthodes de transfert neuronal : il se caractérise par le fait que lors du transfert d’identité ou de style, non seulement l’identité ou le style cible ne sont pas totalement transférés mais ces informations se mélangent avec l’identité ou le style source. En outre, il n’existe que peu d’études à ce jour pour essayer de comprendre ce qui est en réalité transféré lors d’un transfert de style (Sigurgeirsson and King, 2023). La proposition d’une méthodologie expérimentale pour objectiver ces biais dans les données, à l’apprentissage, ou pendant l’évaluation est nécessaire à la fois pour mieux comprendre, modéliser, et à la fin générer artificiellement de la parole. Enfin, les contraintes sur la qualité des données exploitables pour l’apprentissage sont encore très fortes : les bases pour la génération doivent être isolées, monologuées, et de qualité sonore bonne et homogène pour la génération de voix artificielle. Par comparaison avec les ressources disponibles pour le texte ou l’image, les contraintes exposées font de la parole humaine une ressource rare qui ne peut bénéficier de l’effet des grands modèles que marginalement. Vu différemment, ces contraintes constituent une opportunité pour élaborer des modèles d’apprentissage spécifiques à la modalité orale pour apprendre les informations liées aux différents facteurs de variabilité de la voix, de la parole et des canaux de communication analogique ou numérique — entre autres : microphones, compression, bruit, réverbération.

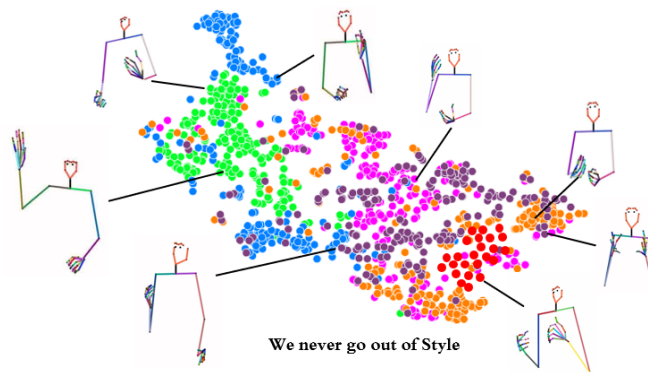
LES DÉFIS DE LA GÉNÉRATION DE LA PAROLE ... ET AU-DELÀ

Problèmes d’apprentissage

Si les problèmes de l’intelligibilité et de la naturalité de la parole artificielle semblent désormais en grande partie résolus, nous sommes encore loin d’avoir appréhendé et solutionné l’ensemble des fonctions de parole telles que postulées par (Jakobson, 1960) et en particulier l’ensemble des fonctions expressives, stylistiques, et esthétiques. En particulier, la génération de la parole expressive, située ou contextualisée, en interaction, et dans des langues différentes constitue les grandes lignes d’aujourd’hui et les défis de demain pour la recherche sur la parole. La multiplication des facteurs ou des attributs modélisés simultanément pose un problème évident pour l’apprentissage : la quantité de données diminue exponentiellement avec le nombre de facteurs, ce qui rend le problème de la généralisation, de l’interpolation, ou de l’extrapolation complexe. Par ailleurs, les avancées réalisées en apprentissage de réseaux de neurones offrent de nouvelles possibilités et voies de recherche inimaginables auparavant. La première de ces possibilités est la génération de voix artificielles non-humaines mais crédibles — soit parce que ces voix n’existent pas en réalité, soit parce qu’elles possèdent des capacités qui outrepassent les capacités vocales humaines notamment physiques. La recreation neuronale de la voix du chanteur Farinelli pour l’artiste Judith Deschamps (Deschamps, 2022) par Frederick Bous et Axel Roebel (Bous and Roebel, 2022) offre un aperçu de ces possibilités, en créant une voix artificielle capable de chanter avec une tessiture et une couleur impossible à réaliser par un être humain. La génération de voix inexistantes constitue également un axe de recherche émergent, notamment pour créer des voix de synthèse à la demande et avec des personnalités sur-mesure ou par hybridation avec d’autres sons en fonction des besoins et des applications. La capacité de transférer un style ou un certain nombre de propriétés d’une voix directement par l’exemple ouvre également de nouvelles voies pour la recherche comme pour les applications. La possibilité de s’émanciper a priori de toute ontologie et de catégories complexes, non-consensuelles, et toujours à définir constitue un changement de paradigme important de la démarche scientifique, qui reste largement à formaliser et à explorer.

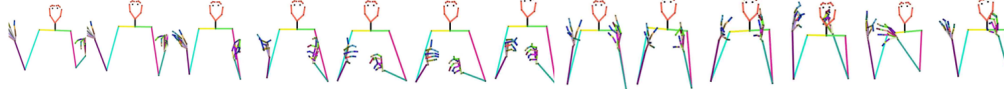
Génération multimodale de comportements humains

La voix n'existe pas sans corps : elle est incarnée par essence. Les avancées réalisées par l'intelligence artificielle permettent désormais déjà d'envisager des problèmes complexes comme la génération multimodale de comportements humains, notamment la coordination spatiale et temporelle de l'ensemble des modalités de la communication humaine, en particulier du langage oral et du langage corporel — les expressions faciales, les poses corporelles, les gestes des mains, etc. Cette problématique a été l'objet de la thèse réalisée avec Mireille Fares (Fares, 2023) (2019-2023) dans laquelle nous avons contribué à proposer des modèles de générations de gestes corporels à partir du signal de parole et du transfert de style multimodal incluant les modalités textuelles, orales, et corporelles. Cette thèse a été également l'occasion de proposer de nouveaux protocoles expérimentaux pour évaluer le transfert du style, en proposant une méthodologie permettant d'évaluer séparément la gestuelle associée au contenu linguistique d'un énoncé et la gestuelle spécifique associée au style d'un individu ou d'un groupe d'individus, comme illustré sur la **Figure 3.22**.

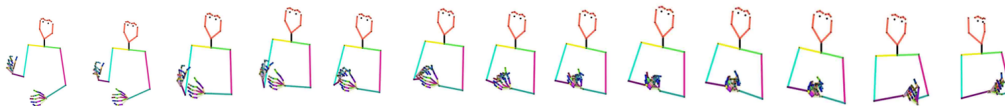


Source: Lec_cosmic

Spilled into the surroundings, and then you form a next generation star



Target: Corden



Lec_cosmic → Corden



FIGURE 3.22 – En haut : visualisation de l'espace de gestes corporels. En bas : transfert de style entre deux individus pour l'énoncé : "Spilled into the surroundings, and then you form a next generation star". Source : (Fares et al., 2023)

Problèmes méthodologiques

L'évaluation de la génération de la parole a historiquement été pensée pour mesurer l'intelligibilité et la naturalité de la parole humaine comme artificielle. Les protocoles MOS

ou MUSHRA actuellement en vigueur en sont des exemples. A nouveaux temps, nouvelles possibilités : de nouvelles méthodologies sont à définir pour évaluer la voix et la parole dans de nouvelles perspectives. C'est entre autres exemples l'évaluation du transfert de style ou d'identité ou la coordination des gestes et de la parole pour la génération multimodale. L'évaluation ne doit d'ailleurs plus se limiter à une évaluation de la parole sous l'angle de la production mais de la réception et de l'impression, c'est-à-dire l'effet d'une voix humaine ou artificielle sur un auditeur humain. Nous devons pour cela nous inspirer des mesures psychométriques proposées en psychologie expérimentale pour évaluer l'attention, l'engagement et l'imagination d'une personne face à une voix artificielle et de manière générale pour mieux comprendre la cognition de la voix humaine. L'imagination stimulée par une lecture d'histoire avec une voix expressive — réelle ou artificielle — est par exemple l'objet principal du projet ANR EXOVOICES (2023-2027) en cours avec Jérôme Sackur du Laboratoire de Sciences Cognitives et Psycholinguistique (LSCP) à l'Ecole Normale Supérieure et la thèse de Théodor Lemerle ([Lemerle, 2023](#)) (2023-2026) que je co-encadre avec Axel Roebel. C'est également l'enjeu principal du projet EMERGENCE ReVOLT (REvealing human bias with real Time VOcal deep fakes, 2022-2023) en collaboration avec Pablo Arias et Clément Le Moine qui vise à proposer de nouvelles méthodologies d'évaluation des biais humains par l'intermédiaire de la manipulation de l'apparence vocale d'une personne — et notamment de son genre.

Problèmes de sécurité : simuler des attaques pour apprendre à se défendre


La capacité à générer des voix ou à manipuler des voix de manière ultra-réaliste pose le problème général de l'authentification et de la certification des données numériques — plus particulièrement dans la mesure ces manipulations tendent à devenir imperceptibles par un être humain. Des certifications doivent être proposées pour maîtriser l'intégralité du cycle de vie d'une donnée numérique vocale depuis son acquisition jusqu'à sa diffusion pour en certifier l'authenticité ou pour détecter des données manipulées. C'est particulièrement vrai pour l'usurpation d'identité, mais plus largement pour l'ensemble des manipulations envisageables, comme les émotions, etc... C'est l'objet des recherches en cours avec les projets ANR BRUEL (ElaBoRation d'Une méthodologie d'EvaLuation des systèmes d'identification par la voix, 2022-2026) pour la certification de deep fake audio et la thèse de Mathilde Abrassart (2023-2026), ou le projet ASTRID DeTOX (Lutte contre les vidéos hyper-truquées de personnalités françaises, 2023-2025) pour l'élaboration d'algorithmes de détection de deep fakes audiovisuels spécialisés pour des personnalités publiques pour qui il existe de nombreuses données accessibles en ligne pour construire des générateurs de deep fakes.

UN CHERCHEUR ENGAGÉ DANS LES DÉFIS DE LA SOCIÉTÉ

Deux mille vingt-trois marque un tournant dans l'histoire de l'intelligence artificielle avec l'émergence des grands modèles de langage et leur démocratisation — c'est-à-dire leur libre mise à disposition à tous publics. Dans les débats actuels et les prises de position sur l'intelligence artificielle ([Bengio, 2023](#); [Le Cun, 2023](#); [Hinton, 2023](#)) qui se cristallisent autour de ces récents déploiements, la voix ne fait pas exception ([Wang et al., 2023](#); [Betker, 2023](#)). Le chercheur en intelligence artificielle n'a dans les temps modernes jamais été aussi exposé publiquement, il doit aujourd'hui exprimer sa voix et s'engager au-delà de ses recherches scientifiques sur les effets sociétaux et politiques du fruit de ces recherches — en particulier pour sensibiliser aux mauvais usages et lutter contre toutes les formes d'obscurantisme qui menacent les sociétés humaines. La question principale est aujourd'hui : non pas quand ? mais pour qui et pour quoi ? Dans ce contexte, la question de l'éthique devient centrale. Une régulation et une meilleure législation est absolument nécessaire pour réguler le "far-west" ([Bengio, 2023](#)) que constitue une


utilisation non-contrôlée des données personnelles, du droit à l'image (et à la voix), et du droit d'auteur. La récente rédaction d'un manifeste sur l'intelligence artificielle par l'union des artistes de la voix¹ dénote la montée des préoccupations autour de l'intelligence artificielle. Les deux premières éditions des Deep Voices, Paris ont initié des réflexions dans ce sens, sur les effets psychologiques et cognitifs d'une immersion affective dans un monde cyber-physique ou virtuel (2021) ou sur la diversité et l'inclusion des identités, des genres, des dialectes, et des langues dans un monde fortement numérisé (2022). Ces préoccupations sont également prégnantes dans le projet REVoLT (2022-2023) sur la cognition des deep fakes et leur utilisation en cognition pour mesurer les biais humains, ou encore le colloque international SIVA (dont la première édition a eu lieu en janvier 2023) sur la génération et la psychologie des deep fakes et des agents virtuels socialement intelligents.

Si la première révolution industrielle était la machination des corps par l'externalisation des capacités physiques humaines dans des machines-outils, la révolution de l'intelligence artificielle opère la machination des esprits et donc l'externalisation des capacités cognitives de l'être humaine dans les prothèses que sont les ordinateurs, les smartphones, et les objets connectés. Il doit aussi se positionner contre l'industrialisation des esprits (Stiegler, 2013). Pour rendre calculable et prédictible, la machine opère un certain nombre de réduction de l'individu et de sa pensée. Elle réduit, uniformise, et détermine les comportements. Pour que cette révolution soit fructueuse, il est nécessaire que l'intelligence artificielle augmente et non réduise les capacités humaines, sensori-motrices, cognitives, et créatives. En aucun cas, l'intelligence artificielle ne doit se substituer à l'humain, elle doit au contraire l'étendre et le stimuler dans une co-évolution de l'humain et de la machine. Par ailleurs, les imperfections du corps et de l'esprit de l'humain sont précisément ce qui le distingue et rend possible sa diversité, son expressivité, et sa créativité : une condition nécessaire à son émergence et à sa persistance dans l'évolution (Darwin, 1890). Cette faculté de l'arbitraire à créer de la surprise et à effectuer des comportements non-motivés en apparence comme l'art constitue une force pour l'Humanité. Pour clôturer ce manuscrit, les pionniers de l'intelligence artificielle avaient pressenti cette capacité comme la conjecture qui clôt la proposition de Dartmouth relative à l'aléatoire et à la créativité (Minsky et al., 1955) :

 fairly attractive and yet clearly incomplete conjecture is that the difference between creative thinking and unimaginative competent thinking lies in the injection of a some randomness. The randomness must be guided by intuition to be efficient. In other words, the educated guess or the hunch include controlled randomness in otherwise orderly thinking.

J. MCCARTHY, DARTMOUTH COLLEGE
M. L. MINSKY, HARVARD UNIVERSITY
N. ROCHESTER, I.B.M. CORPORATION
C. E. SHANNON, BELL TELEPHONE LABS

Ce à quoi répond l'artiste — et que confirme aujourd'hui les connaissances en neuro-psychologie et neuro-linguistique :

 here is a crack in everything, that's how the light gets in

LEONARD COHEN

1. <https://www.unitedvoiceartists.com>

BIBLIOGRAPHIE

- Abney, S. (1992). Prosodic structure, performance structure and phrase structure. In *Human Language Technology : Proceedings of the workshop on Speech and Natural Language*, pages 425–428, Morristown, NJ, USA. Association for Computational Linguistics.
- Ajzen, I. and Fishbein, M. (1980). *Understanding attitudes and predicting social behaviour*. Prentice-Hall, Englewood Cliffs, New Jersey.
- Argyle, M. (2013). *Bodily communication*. Routledge.
- Aristote (329 BC - -323 BC). *Rhétorique*.
- Aristote (350 BC). *Poétique*.
- Bachorowski, J. A. (1999). Vocal expression and perception of emotion. *Current Directions in Psychological Science*, 8 :53–57.
- Badeau, R., Bertin, N., and Vincent, E. (2011). Stability analysis of multiplicative update algorithms for non-negative matrix factorization. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic.
- Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural Machine Translation by Jointly Learning to Align and Translate. In *International Conference on Learning Representations (ICLR)*.
- Bailly, G. (1989). Integration of rhythmic and syntactic constraints in a model of generation of french prosody. *Speech Communication*, 8(2) :137–146.
- Bänziger, T., Mortillaro, M., and Scherer, K. R. (2012). Introducing the geneva multimodal expression corpus for experimental research on emotion perception. *Emotion*, 12(5) :1161–79.
- Baraniuk, R. G., Flandrin, P., Janssen, A. J., and Michel, O. (2001). Measuring Time-Frequency Information Content Using the Rényi Entropies. *IEEE Transactions on Information Theory*, 47(4) :1391–1409.
- Barrett, L. (2017). The theory of constructed emotion : an active inference account of interoception and categorization. *Social Cognitive and Affective Neuroscience*, 12.
- Beck, C. and Schögl, F. (1993). *Thermodynamics of Chaotic Systems*. Cambridge University Press, Cambridge, Massachusetts, USA.
- Beckman, M. and Ayers, G. (1997). Guidelines for ToBI labelling. Technical report, Linguistics Department, Ohio State University.
- Beckman, M. E., Hirschberg, J., and Shattuck-Hufnagel, S. (2005). *Prosodic Typology - The Phonology of Intonation and Phrasing*, chapter The Original ToBI System and the Evolution of the ToBI Framework, pages 9–54. Oxford University Press.
- Belghazi, I., Rajeswar, S., Baratin, A., Hjelm, R. D., and Courville, A. C. (2018). MINE : Mutual Information Neural Estimation. In *International Conference on Machine Learning (PMLR)*.
- Belião, J. (2013). Characterizing Genres through Syntax and Prosody. In *European Summer School in Logic, Language and Information*, pages 1–12, Düsseldorf , Germany.

- Belião, J. (2016). *How syntax and prosody are mapping? A study of synchronization and congruence*. Thèse de doctorat, Université Paris Ouest - Nanterre La Défense.
- Bell, A. (1984). Language style as audience design. *Language in society*, 13(2) :145–204.
- Beller, G. (2009). *Analyse et modèle génératif de l'expressivité : application à la parole et à l'interprétation musicale*. Thèse de doctorat, Université Pierre et Marie Curie.
- Beller, G., Obin, N., and Rodet, X. (2008). Articulation Degree As a Prosodic Dimension of Expressive Speech. In *Speech Prosody*, Campinas, Brazil.
- Benaroya, E., Donagh, L. M., Gribonval, R., and Bimbot, F. (2003). Non negative sparse representation for wiener based source separation with a single sensor. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 6, pages 613–616, Hong-Kong, China.
- Benaroya, L., Bimbot, F., and Gribonval, R. (2006). Audio source separation with a single sensor. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(1) :191–199.
- Benaroya, L., Obin, N., Liuni, M., Roebel, A., Rauml, W., and Argentieri, S. (2018). Binaural localization of multiple sound sources by non-negative tensor factorization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(6) :1072–1082.
- Benaroya, L., Obin, N., and Roebel, A. (2023). Manipulating voice attributes by adversarial learning of structured disentangled representations. *Entropy*, 25(2).
- Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the dangers of stochastic parrots : Can language models be too big? . In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.
- Bengio, Y. (28 avril 2023). Aujourd'hui, l'intelligence artificielle, c'est le Far West! Nous devons ralentir et réguler. https://www.lemonde.fr/idees/article/2023/04/28/yoshua-bengio-chercheur-aujourd-hui-l-intelligence-artificielle-c-est-le-far-west-nous-devons-ralentir-et-reguler_6171336_232.html, consulté le 31 juillet 2023.
- Berlioz, H. (1844). *Le Traité d'instrumentation et d'orchestration* . Henry Lemoine : Paris, France.
- Bertin, N. (2009). *Les factorisations en matrices non-négatives. Approches contraintes et probabilistes, application à la transcription automatique de musique polyphonique*. PhD thesis, Télécom ParisTech.
- Betker, J. (2023). Better Speech Synthesis through Scaling. *arXiv :2305.07243 [cs.SD]*.
- Biber, D. (1988). *Variation Across Speech and Writing*. Cambridge University Press, Cambridge.
- Black, A. and Taylor, P. (1994). Assigning intonation elements and prosodic phrasing for english speech synthesis from high level linguistic input. In *International Conference on Spoken Language Processing*, pages 715–718, Yokohama, Japan.
- Black, A., Zen, H., and K.Tokuda (2007). Statistical parametric speech synthesis. In *International Conference on Audio, Speech, and Signal Processing (ICASSP)*, pages 1229–1232.
- Blandin, C., Ozerov, A., and Vincent, E. (2012). Multi-source TDOA Estimation in Reverberant Audio using Angular Spectra and Clustering. *Elsevier Signal Processing*, 92 :1950–1960.

- Boidin, C. and Boffard, O. (2008). Generating Intonation from a Mixed CART-HMM Model for Speech Synthesis. In *Interspeech*, pages 2130–2133, Brisbane, Australia.
- Bous, F., Benaroya, L., Obin, N., and Roebel, A. (2021). Voice Reenactment with Fo and timing constraints and adversarial learning of conversions. In *European conference on signal processing (EUSIPCO)*, Belgrade, Serbia.
- Bous, F., Benaroya, L., Obin, N., and Roebel, A. (2022). Voice Reenactment with Fo and timing constraints and adversarial learning of conversions. In *European Signal Processing Conference (EUSIPCO)*.
- Bous, F. and Roebel, A. (2022). A Bottleneck Auto-Encoder for Fo Transformations on Speech and Singing Voice. *Information*, 13(3) :102.
- Bousquet, P.-M., Larcher, A., Matrouf, D., Bonastre, J.-F., and Plhot, O. (2012). Variance-Spectra based Normalization for I-vector Standard and Probabilistic Linear Discriminant Analysis. In *Odyssey : The Speaker and Language Recognition Workshop*, pages 157–164, Singapore, Singapore.
- Bouvier, D., Obin, N., Liuni, M., and Roebel, A. (2016). A Source/Filter Model with Adaptive Constraints for NMF-based Speech Separation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 131–135, Shanghai, China.
- Branislav, G., Bailly, G., Mohammed, O., Xu, Y., and Garner, P. N. (2018). A variational prosody model for the decomposition and synthesis of speech prosody. In *ArXiv e-prints*.
- Browning, J. and Le Cun, Y. (2022). AI And The Limits Of Language. <https://www.noemamag.com/ai-and-the-limits-of-language/>, consulté le 31 juillet 2023.
- Butler, R. A. (1986). The Bandwidth Effect on Monaural and Binaural Localization. *Journal of Hearing Research*, 21 :67–73.
- Bänziger, T. and Scherer, K. R. (2005). The role of intonation in emotional expressions. *Speech Communication*, 46(3) :252 – 267. Quantitative Prosody Modelling for Natural Speech Description and Generation.
- Campbell, N. (1992). Prosodic Encoding of English Speech. In *International Conference on Spoken Language Processing*, pages 663–666, Edmonton, Canada.
- Campbell, N. and Mokhtari, P. (2003). Voice quality : the 4th prosodic dimension. In *International Congress of Phonetic Sciences*, pages 2417–2420, Barcelona, Spain.
- Campbell, W. M., Sturim, D. E., and Reynolds, D. A. (2006). Support Vector Machines using GMM Supervectors for Speaker Verification. *IEEE Signal Processing Letters*, 13(5) :308–311.
- Campione, E., Hirst, D., and Véronis, J. (2000). *Automatic Stylisation and Symbolic Coding of Fo : Implementations of the INTSINT Model*, chapter Intonation. Research and Applications. Kluwer, Dordrecht.
- Carpentier, T., Noisternig, M., and Warusfel, O. (2015). Twenty Years of Ircam Spat : Looking Back, Looking Forward. In *International Computer Music Conference (ICMC)*, pages 270–277, Denton, USA.
- Chen, H. and Garner, P. N. (2023). Diffusion Transformer for Adaptive Text-to-speech. In *ISCA Speech Synthesis Workshop (SSW)*, Grenoble, France.
- Cherry, E. C. (1953). Some Experiments on the Recognition of Speech, with One and with Two Ears. *The Journal of the Acoustical Society of America*, 25(5) :975–979.

- Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In *Empirical Methods in Natural Language Processing (EMNLP)*.
- Chung, Y.-A. and Glass, J. (2018). Speech2Vec : A Sequence-to-Sequence Framework for Learning Word Embeddings from Speech. In *ISCA Interspeech*, pages 811–815.
- Cortes, C. and Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, 20(3) :273–297.
- Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory*. John Wiley.
- Cowie, R., Douglas-Cowie, E., McRorie, M., Sneddon, I., Devillers, L., and Amir, N. (2011). *The HUMAINE database*, chapter Issues in Data Collection. Springer.
- Cyrulnik, B. (2023). *La nuit j'écirai des soleils*. Odile Jacob.
- d'Alessandro, C. and Mertens, P. (1995). Automatic pitch contour stylization using a model of tonal perception. *Computer Speech and Language*, 9(3) :257–288.
- Darwin, C. (1890). *The Expression of the Emotion in Man and Animals*. John Murray, London.
- De Saussure, F. (1967). *Cours de linguistique générale*. Payot : Paris.
- Dean, D. B., Sridharan, S., Vogt, R. J., and Mason, M. W. (2010). The QUT-NOISE-TIMIT corpus for the evaluation of voice activity detection algorithms. In *Interspeech*, pages 3110–3113.
- Degottex, G., Lanchantin, P., Roebel, A., and Rodet, X. (2013). Mixed source model and its adapted vocal tract filter estimate for voice transformation and synthesis. *Speech Communication*, 55(2) :278–294.
- Degottex, G. and Obin, N. (2014). Phase Distortion Statistics as a Representation of the Glottal Source : Application to the Classification of Voice Qualities. In *Interspeech*, Singapore, Singapore.
- Degottex, G., Roebel, A., and Rodet, X. (2009). Glottal Closure Instant Detection from a Glottal Shape Estimate. In *13th International Conference on Speech and Computer*, pages 226–231, St-Petersburg, Russia.
- Degottex, G., Roebel, A., and Rodet, X. (2011). Phase Minimization for Glottal Model Estimation. *IEEE Transactions on Acoustics, Speech and Language Processing*, 19(5) :1080–1090.
- Dehak, N. (2009). *Discriminative and Generative Approaches for Long- and Short-Term Speaker Characteristics Modeling : Application to Speaker Verification*. PhD. thesis, Ecole de Technologie Supérieure.
- Dehak, N., Dehak, R., Glass, J., Reynolds, D., and Kenny, P. (2010). Bayesian Speaker Verification with Heavy-Tailed Priors. In *Odyssey : The Speaker and Language Recognition Workshop*, Brno, Czech Republic.
- Dehak, N., Dehak, R., Kenny, P., Brummer, N., Ouellet, P., and Dumouchel, P. (2009). Support Vector Machines versus Fast Scoring in the Low-Dimensional Total Variability Space for Speaker Verification. In *Interspeech*, pages 4237–4240.
- Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., , and Ouellet, P. (2011). Front-End Factor Analysis for Speaker Verification. *IEEE transactions on Audio, Speech, and Language Processing*, 19(4) :788–798.

- Delais, E. (1994). Rythme et structure prosodique en français. *French Generative Phonology : Retrospectives and Perspectives*, pages 131–150.
- Delattre, P. (1966). Les Dix Intonations de Base du Français. *The French Review*, 40(1) :1–14.
- Demichel-Basnier, S. (2019). *Sociologie des voix artificielles*. Presses universitaires de Grenoble, Grenoble, France.
- Depalle, P., García, G., and Rodet, X. (1994). A Virtual Castrato (!?). In *International Computer Music Conference (ICMC)*, pages 1720–1723, Aarhus, Denmark.
- Depalle, P. and Poirot, G. (1991). SVP : A Modular System for Analysis, Processing and Synthesis of Sound Signals. In *International Computer Music Conference (ICMC)*, Montréal, Canada.
- Desai, S., Raghavendra, E. V., Yegnanarayana, B., Black, A. W., and Prahallad, K. (2009). Voice conversion using Artificial Neural Networks. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3893–3896.
- Descartes, R. (1647). *Les Méditations métaphysiques*. Paris.
- Deschamps, J. (08 juin 2022). Avec Farinelli, j’ai fait le lien entre la castration et le transhumanisme. https://www.lemonde.fr/culture/article/2022/06/08/judith-deschamps-avec-farinelli-j-ai-fait-le-lien-entre-la-castration-et-le-transhumanisme_6129390_3246.html.
- Despret, V. (2019). *Habiter en oiseau*. Actes Sud.
- Dessein, A., Cont, A., and Lemaitre, G. (2010). Real-Time Polyphonic Music Transcription with Non-negative Matrix Factorization and Beta-Divergence. In *International Society for Music Information Retrieval Conference (ISMIR)*, pages 489–494, Utrecht, Netherlands.
- Dessein, A., Cont, A., and Lemaitre, G. (2013). *Matrix Information Geometry*, chapter Real-time detection of overlapping sound events with non-negative matrix factorization, pages 341–371. Springer.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Di Cristo, A. (1985). *De la Microprosodie à l’Intonosyntaxe*. Publications de l’Université d’Aix-en-Provence, Aix-en-Provence.
- Dubnov, S. (2004). Generalization of Spectral Flatness Measure for Non-Gaussian Linear Processes. *IEEE Signal Processing Letters*, 11(8) :698–701.
- Dudley, H., Riesz, R., and Watkins, S. (1939). A Synthetic Speaker. *Journal of the Franklin Institute*, 227(6) :739–764.
- Dudley, H. W. (1939). The Vocoder. *Bell Labs Rec.*, 18 :122–126.
- Durrieu, J.-L., Ozerov, A., Févotte, C., Richard, G., and David, B. (2009). Main Instrument Separation from Stereophonic Audio Signals using a Source/Filter Model. In *European Signal Processing Conference (EUSIPCO)*, pages 15–19.
- Durrieu, J.-L., Richard, G., David, B., and Févotte, C. (2010). Source/Filter Model for Unsupervised Main Melody Extraction from Polyphonic Audio Signals. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3) :564–575.

- Défosse, A., Copet, J., Synnaeve, G., and Adi, Y. (2022). High fidelity neural audio compression. *arXiv preprint arXiv :2210.13438*.
- D'Alessandro, C., Rosset, S., and Rossi, J.-P. (1998). The Pitch of Short-duration Fundamental Frequency Glissandos. *The Journal of the Acoustical Society of America*, 104(4) :2339–2348.
- Ekman, P. (1992). An Argument for Basic Emotions. *Cognition and Emotion*, 6(3-4) :169–200.
- Ekman, P. and Friesen, W. V. (1971). Constants Across Cultures in the Face and Emotion. *Journal of Personality and Social Psychology*, 17(2) :124–129.
- Ettinger, A. (2020). What BERT is not : Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8 :34–48.
- Fan, Y., Qian, Y., Xie, F.-L., and Soong, F. K. (2014). TTS Synthesis with Bidirectional LSTM based Recurrent Neural Networks. *Interspeech*.
- Fang, F., Yamagishi, J., Echizen, I., and Lorenzo-Trueba, J. (2018). High-Quality Nonparallel Voice Conversion Based on Cycle-Consistent Adversarial Network. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5279–5283.
- Fant, G. (1995). The LF-Model Revisited. Transformations and Frequency Domain Analysis. Technical Report 2-3, K.T.H. Quarterly Progress Report and Status Progress. Departement for Speech, Music and Hearing.
- Fares, M. (2023). *Multimodal Expressive Gesturing with Style*. Thèse de doctorat, Sorbonne Université.
- Fares, M., Obin, N., and Pelachaud, C. (2023). TranSTYler : Multimodal Behavioral Style Transfer for Facial and Body Gestures Generation. In *ACM International Conference on Multimodal Interaction (ICMI)*, Paris, France.
- Farner, S., Roebel, A., and Rodet, X. (2009). Natural Transformation of Type and Nature of the Voice for Extending Vocal Repertoire in High-Fidelity Applications. In *Audio Engineering Society Conference : 35th International Conference : Audio for Games*.
- Federico, M., Enyedi, R., Barra-Chicote, R., Giri, R., Isik, U., Krishnaswamy, A., and Sawaf, H. (2020). From Speech-to-Speech Translation to Automatic Dubbing. In *International Conference on Spoken Language Translation (IWSLT)*, pages 257–264, Online. Association for Computational Linguistics.
- Févotte, C. and Idier, J. (2011). Algorithms for nonnegative matrix factorization with the β -divergence. *Neural Computation*, 23(9) :2421–2456.
- Fiske, S. T., Cuddy, A. J. C., and Glick, P. (2007). Universal Dimensions of Social Cognition : Warmth and Competence. *Trends in Cognitive Sciences*, 11(2) :77–83.
- FitzGerald, D., Cranitch, M., and Coyle, E. (2005). Non-negative Tensor Factorisation for Sound Source Separation. In *Irish Signals and Systems Conference (ISSC)*, Dublin, Ireland.
- Flanagan, J. L. (1972). *Speech analysis synthesis and perception*. Springer.
- Fujisaki, H. (1981). Dynamic characteristics of voice fundamental frequency in speech and singing. acoustical analysis and physiological interpretations. Technical report, K.T.H. Quarterly Progress Report and Status Progress. Departement for Speech, Music and Hearing.
- Fukada, T., Komori, Y., Aso, T., and Ohora, Y. (1994). A study of pitch pattern generation using HMM-based statistical information. In *International Conference on Spoken Language Processing*, pages 723–726, Genove, Italy.

- Févotte, C., Bertin, N., , and Durrieu, J.-L. (2009). Nonnegative matrix factorization with the itakura-saito divergence with application to music analysis. *Neural Computing*, 21(3) :793–830.
- Gao, W., Kannan, S., Oh, S., and Viswanath, P. (2017). Estimating Mutual Information for Discrete-Continuous Mixtures. In *Advances in Neural Information Processing Systems (NIPS)*, pages 5986–5997.
- Garcia-Romero, D. and Espy-Wilson, C. (2011). Analysis of i-vector Length Normalization in Speaker Recognition Systems. In *Interspeech*, page 249–252.
- Gardner, W. G. and Martin, K. D. (1995). HRTF measurements of a KEMAR. *Journal of the Acoustic Society of America*, 97(6) :3907–3908.
- Gerazov, B., Bailly, G., Mohammed, O., Xu, Y., and Garner, P. (2018a). A Variational Prosody Model for the Decomposition and Synthesis of Speech Prosody. In *Speech Prosody*, Poznan, Poland.
- Gerazov, B., Bailly, G., and Xu, Y. (2018b). A Weighted Superposition of Functional Contours Model for Modelling Contextual Prominence of Elementary Prosodic Contours. In *Interspeech*, pages 2524–2528, Hyderabad, India.
- Godard, J.-L. (2006). *Histoire(s) du cinéma*. Gallimard.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014a). Generative adversarial nets. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc.
- Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014b). Generative Adversarial Networks. In *Advances in Neural Information Processing Systems (NIPS)*.
- Graf, H. P., Cosatto, E., Strom, V., and Huang, F. J. (2002). Visual prosody : Facial movements accompanying speech. In *IEEE International Conference on Automatic Face Gesture Recognition*, pages 396–401. IEEE.
- Gresse, A. (2020). *L'Art de la Voix : Caractériser l'information vocale dans un choix artistique*. PhD. Thesis, Laboratoire d'Informatique d'Avignon, Université d'Avignon.
- Griffin, D. W. and Lim, J. S. (1985). A New Model-Based Analysis/Synthesis System. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 513–516, Tampa, Florida.
- Gussenhoven, C. (2004). *The Phonology of Tone and Intonation*. Cambridge University Press, Cambridge.
- Hartholt, A., Traum, D., Marsella, S. C., Shapiro, A., Stratou, G., Leuski, A., Morency, L.-P., and Gratch, J. (2013). All Together Now : Introducing the Virtual Human Toolkit. In *International Conference on Intelligent Virtual Humans*.
- Harwardt, C. (2002). Comparing the Impact of Raised Vocal Effort on Various Spectral Parameters. In *Interspeech*, pages 2941–2944, Florence, Italy.
- Hatch, A., Kajarekar, S., and Stolcke, A. (2006). Within-Class Covariance Normalization for SVM-based Speaker Recognition. In *International Conference on Spoken Language Processing*, pages 1471–1474, Pittsburgh, USA.

- Hawthorne, C., Jaegle, A., Cangea, C., Borgeaud, S., Nash, C., Malinowski, M., Dieleman, S., Vinyals, O., Botvinick, M., Simon, I., Sheahan, H., Zeghidour, N., Alayrac, J.-B., Carreira, J., and Engel, J. (2022). General-purpose, long-context autoregressive modeling with perceiver AR. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S., editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 8535–8558. PMLR.
- Hayes, A. F. and Krippendorff, K. (2007). Answering the Call for a Standard Reliability Measure for Coding Data. *Communication Methods and Measures*, 1 :77–89.
- He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.
- Higgins, I., Amos, D., Pfau, D., Racaniere, S., Matthey, L., Rezende, D., and Lerchner, A. (2018). Towards a Definition of Disentangled Representations. *arXiv preprint arXiv :1812.02230*.
- Hinton, G. (01 mai 2023). ‘The Godfather of A.I.’ Leaves Google and Warns of Danger Ahead. <https://www.nytimes.com/2023/05/01/technology/ai-google-chatbot-engineer-quits-hinton.html>, consulté le 31 juillet 2023.
- Hirschberg, J. (2000). *Prosody, Theory and Experiment : Studies*. Merle Horne.
- Hirst, D. and Di Cristo, A. (1998). *Intonation Systems : a survey of twenty languages*. Cambridge University Press, Cambridge.
- Hirst, D., Di Cristo, A., and Espresser, R. (2000). *Prosody : Theory and Experiments*, chapter Levels of representation and levels of analysis for the description of intonation systems. M. Horne.
- Hirst, D. and Espresser, R. (1993). Automatic Modelling of Fundamental Frequency using a Quadratic Spline Function. In *Travaux de l’Institut de Phonétique d’Aix*, volume 15, pages 71–85.
- Holm, B. (2003). *SFC : un modèle de superposition de contours multiparamétriques pour la génération automatique de la prosodie - Apprentissage automatique et application l’énunciation de formules mathématiques*. PhD. thesis, Institut de la Communication Parlée, Grenoble.
- Holmberg, E. B., Hillman, R. E., and Perkell, J. S. (1988). Glottal Airflow and Transglottal Air Pressure Measurements for Male and Female Speakers in Soft, Normal, and Loud Voice. *Journal of the Acoustical Society of America*, 84(2) :511–529.
- Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer Feedforward Networks are Universal Approximators. *Neural Networks*, 2(5) :359–366.
- Hsu, C.-C., Hwang, H.-T., Wu, Y.-C., Tsao, Y., and Wang, H.-M. (2016). Voice Conversion from Non-parallel Corpora Using Variational Auto-encoder. In *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pages 1–6.
- Hueber, T. (2005). *Talkapillar : Système d’analyse, de synthèse et de transformation de la parole à partir du texte*. Mémoire de stage, École d’Ingénieurs en Chimie et Sciences du Numérique de Lyon.
- Huffman, D. (1952). A Method for the Construction of Minimum-Redundancy Codes. *Proceedings of the Institute of Radio Engineers*, 40(9) :1098–1101.
- Hunt, A. J. and Black, A. W. (1996). Unit selection in a concatenative speech synthesis system using a large speech database. In *International Conference on Acoustics, Speech, and Signal Processing*, page 373–376, Atlanta, USA.

- Iseli, M., Shue, Y.-L., and Alwan, A. (2007). Age, Sex, and vowel dependencies of Acoustic Measures related to the Voice source. *The Journal of the Acoustical Society of America*, 121(4) :2283–2295. Publisher : Acoustical Society of America.
- Ishiguro, H., Ono, T., Imai, M., Maeda, T., Kanda, T., and Nakatsu, R. (2001). Robovie : an interactive Humanoid Robot. *Industrial Robot*, 28(6) :498–504.
- ISO/IEC 13818-3 (1998). Information technology — Generic coding of moving pictures and associated audio information. Standard, International Organization for Standardization.
- Jakobson, R. (1960). *Style and Langage*, chapter Linguistics and Poetics. Cambridge Massachusetts Institute of Technology Press, New-York.
- Jiang, S., Wu, L., Yuan, P., Sun, Y., and Liu, H. (2020). Deep and CNN fusion Method for Binaural Sound Source Localisation. *The Journal of Engineering*, 2020(13) :511–516.
- Joder, C., Weninger, F., Eyben, F., Virette, D., and Schuller, B. (2012). Real-Time Speech Separation by Semi-supervised Nonnegative Matrix Factorization. In *International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, pages 322–329, Tel Aviv, Israel.
- Joder, C., Weninger, F., Virette, D., and Schuller, B. (2013). A Comparative Study on Sparsity Penalties for NMF-based Speech Separation : Beyond Lp-norms. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 858–862.
- Johnston, J. D. (1988). Transform Coding of Audio Signals using Perceptual Noise Criteria. *IEEE Journal on Selected Areas in Communications*, 6(2) :314–332.
- Jourjine, A., Rickard, S., , and Yilmaz, O. (2000). Blind Separation of Disjoint Orthogonal Signals : Demixing n Sources From 2 Mixtures. In *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, page 2985–2988, Istanbul, Turkey.
- Julia, L. E., Bing, J. G., and Dubreuil, J. (2001). System and method for speech activated navigation. *Patent : 20030078781*.
- Kameoka, H., Huang, W.-C., Tanaka, K., Kaneko, T., Hojo, N., and Toda, T. (2021). Many-to-Many Voice Transformer Network. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29 :656–670.
- Kameoka, H., Kaneko, T., Tanaka, K., and Hojo, N. (2018). StarGAN-VC : Non-Parallel Many-to-Many Voice Conversion Using Star Generative Adversarial Networks. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 266–273.
- Kameoka, H., Tanaka, K., Kwaśny, D., Kaneko, T., and Hojo, N. (2020). ConvS2S-VC : Fully Convolutional Sequence-to-Sequence Voice Conversion. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28 :1849–1863.
- Kaneko, T. and Kameoka, H. (2017). Parallel-Data-Free Voice Conversion Using Cycle-Consistent Adversarial Networks. *arXiv :1711.11293*. arXiv : 1711.11293.
- Kaneko, T. and Kameoka, H. (2018). CycleGAN-VC : Non-parallel Voice Conversion Using Cycle-Consistent Adversarial Networks. In *European Signal Processing Conference (EUSIPCO)*, pages 2100–2104.
- Kaneko, T., Kameoka, H., Tanaka, K., and Hojo, N. (2019). CycleGAN-VC2 : Improved CycleGAN-based Non-parallel Voice Conversion. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6820–6824.
- Kaneko, T., Kameoka, H., Tanaka, K., and Hojo, N. (2019). StarGAN-VC2 : Rethinking Conditional Methods for StarGAN-Based Voice Conversion. In *ISCA Interspeech*, pages 679–683.

- Karkkainen, K. and Joo, J. (2021). FairFace : Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1548–1558.
- Karras, T., Laine, S., and Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4396–4405.
- Kawahara, H. (2006). STRAIGHT, exploitation of the other aspect of VOCODER : Perceptually isomorphic decomposition of speech sounds. *Acoustical Science and Technology*, 27(6) :349–353.
- Kenny, P., Ouellet, P., Dehak, N., Gupta, V., and Dumouchel, P. (2008). A Study of Inter-Speaker Variability in Speaker Verification. *IEEE transactions on Audio, Speech, and Language Processing*, 16(5) :980–988.
- Kido, H. and Kasuya, H. (2001). Everyday Expressions associated with Voice Quality of Normal Utterance — Extraction by Perceptual Evaluation. *Journal of the Acoustic Society of Japan*, 57(5) :337–344.
- Knapp, C. and Carter, G. (1976). The Generalized Cross-Correlation Method for Estimation of Time Delay. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 24(4) :320–327.
- Knapp, M. L., Hall, J. A., and Horgan, T. G. (2014). *Nonverbal Communication in Human Interaction*. Wadsworth/Cengage learning.
- Kong, J., Kim, J., and Bae, J. (2020). HiFi-GAN : Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 17022–17033. Curran Associates, Inc.
- Kopp, S., Allwood, J., Grammer, K., Ahlsen, E., and Stocksmeier, T. (2008). Modeling Embodied Feedback with Virtual Humans. In Wachsmuth, I. and Knoblich, G., editors, *Modeling Communication with Robots and Virtual Humans*, pages 18–37, Berlin, Heidelberg. Springer Berlin Heidelberg.
- La Pietra, L. (2023). *L'indépendance de la voix du XXIème siècle. Ethique et Esthétique de l'interprétation vocale entre le Belcanto et l'intelligence artificielle*. Thèse de doctorat, Université Vincennes — Saint-Denis.
- Lacheret-Dujour, A. and Beaugendre, F. (1998). *La prosodie du français*. CNRS, Paris.
- Lample, G., Zeghidour, N., Usunier, N., Bordes, A., Denoyer, L., and Ranzato, M. (2017). Fader networks : Manipulating images by sliding attributes. In *Advances in Neural Information Processing Systems(NIPS)*, pages 5967–5976.
- Lanchantin, P., Degottex, G., and X.Rodet (2010). A HMM-based Speech Synthesis System using a new Glottal Source and Vocal-tract Separation Method. In *International Conference on Audio, Speech, and Signal Processing (ICASSP)*, pages 4630–4633, Dallas, USA.
- Lanchantin, P., Morris, A., Rodet, X., and Veaux, C. (2008). Automatic Phoneme Segmentation with Relaxed Textual Constraints. In *International Conference on Language Resources and Evaluation*, pages 2403–2407, Marrakech, Morocco.
- Lanchantin, P. and Rodet, X. (2010). Dynamic Model Selection for Spectral Voice Conversion. In *Interspeech*, pages 1720–1723, Makuhari, Japan.
- Larcher, A., Bonastre, J.-F., Fauve, B., Lee, K. A., Levy, C., Li, H., Mason, J. S., and Parfait, J.-Y. (2013). ALIZE 3.0 - Open Source Toolkit for State-of-the-Art Speaker Recognition. In *Interspeech*, Lyon, France.

- Latorre, J. and Akamine, M. (2008). Multilevel Parametric-base Fo Model for Speech Synthesis. In *Interspeech*, pages 2274–2277, Brisbane, Australia.
- Laver, J. (1980). *The Phonetic Description of Voice Quality*. Cambridge : Cambridge University Press.
- Le, M., Vyas, A., Shi, B., Karrer, B., Sari, L., Moritz, R., Williamson, M., Manohar, V., Adi, Y., Mahadeokar, J., and Hsu, W.-N. (2023). Voicebox : Text-Guided Multilingual Universal Speech Generation at Scale. *arXiv :2306.15687 [eess.AS]*.
- Le Breton, D. (2019). *Éclats de voix. Une anthropologie des voix*. Métailie, Paris.
- Le Cun, Y. (28 avril 2023). L'idée même de vouloir ralentir la recherche sur l'IA s'apparente à un nouvel obscurantisme. https://www.lemonde.fr/idees/article/2023/04/28/yann-le-cun-directeur-a-meta-l-idee-meme-de-vouloir-ralentir-la-recherche-sur-l-ia-s-apparente-a-un-nouvel-obscurantisme_6171338_3232.html, consulté le 31 juillet 2023.
- Le Magoarou, L., Ozerov, A., and Duong, N. Q. (2014). Text-informed Audio Source Separation. Example-based Approach using Non-negative Matrix Partial co-Factorization. *Journal of Signal Processing Systems*, pages 1–15.
- Le Moine, C. (2023). *Neural Conversion of Social Attitudes in Speech Signals*. Thèse de doctorat, Sorbonne Université.
- Le Moine, C. and Obin, N. (2020). Att-HACK : An Expressive Speech Database with Social Attitudes. In *Speech Prosody*, pages 744–748, Tokyo, Japan.
- Le Moine, C., Obin, N., and Roebel, A. (2021a). Speaker Attentive Speech Emotion Recognition. In *Interspeech*, pages 2866–2870, Brno, Czech Republic.
- Le Moine, C., Obin, N., and Roebel, A. (2021b). Towards End-to-End Fo Voice Conversion based on Dual-GAN with Convolutional Wavelet Kernels. In *European Signal Processing Conference (EUSIPCO)*, pages 36–40, Dublin, Ireland.
- Le Roux, J., Hershey, J. R., and Wenginger, F. (2015). Deep NMF for Speech Separation. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 66–70.
- Leary, T. (1957). *Interpersonal diagnosis of personality. A functional theory and methodology for personality evaluation*. New-York : Ronald Press.
- Lee, D. D. and Seung, H. S. (1999). Learning the Parts of Objects by Non-negative Matrix Factorization. *Nature*, 401 :788–791.
- Lemerle, T. (2023). *Expressive Text-to-Speech Synthesis for Storytelling Performance*. Thèse de doctorat, Sorbonne Université.
- Levinson, S. C. and Holler, J. (2014). The Origin of Human Multi-Modal Communication. *Philosophical Transactions of the Royal Society B : Biological Sciences*, 369(1651).
- Li, C., Ma, X., Jiang, B., Li, X., Zhang, X., Liu, X., Cao, Y., Kannan, A., and Zhu, Z. (2017). Deep Speaker : an End-to-End Neural Speaker Embedding System. *arXiv :1705.02304 [cs]*. *arXiv : 1705.02304*.
- Li, R., Wu, Z., Meng, H., and Cai, L. (2016). DBLSTM-based Multi-Task Learning for Pitch Transformation in Voice Conversion. In *2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP)*.

- Li, S. Z., Hou, X. W., Zhang, H., and Cheng, Q. (2001). Learning Spatially Localized, Parts-based Representation. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages I–207.
- Lieberman, P. (1984). *The Biology and Evolution of Language*. Harvard University Press.
- Liu, H., Yuan, P., Yang, B., Yang, G., and Chen, Y. (2022). Head-related Transfer Function–reserved Time-Frequency Masking for Robust Binaural Sound Source Localization. *CAAI Transactions on Intelligence Technology*, 7(1) :26–33.
- Liu, L.-J., Ling, Z.-H., Jiang, Y., Zhou, M., and Dai, L.-R. (2018). WaveNet Vocoder with Limited Training Data for Voice Conversion. In *Interspeech 2018*, pages 1983–1987.
- Lorenzo-Trueba, J., Yamagishi, J., Toda, T., Saito, D., Villavicencio, F., Kinnunen, T., and Ling, Z. (2018). The Voice Conversion Challenge 2018 : Promoting Development of Parallel and Nonparallel Methods. In *Speaker Odyssey : The Speaker and Language Recognition Workshop*, pages 195–202.
- Lu, H., Wu, Z., Dai, D., Li, R., Kang, S., Jia, J., and Meng, H. (2019). One-Shot Voice Conversion with Global Speaker Embeddings. In *ISCA Interspeech*, pages 669–673.
- Luo, Z., Chen, J., Takiguchi, T., and Ariki, Y. (2017). Emotional Voice Conversion using Neural Networks with Arbitrary Scales Fo based on Wavelet Transform. *EURASIP Journal on Audio, Speech, and Music Processing*, 2017(1).
- Luo, Z., Chen, J., Takiguchi, T., and Ariki, Y. (2019). Emotional Voice Conversion Using Dual Supervised Adversarial Networks With Continuous Wavelet Transform Fo Features. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(10) :1535–1548.
- Léon, P. (1993). *Précis de Phonostylistique - Parole et Expressivité*. Nathan, Paris.
- Ma, N., Gonzalez, J. A., and Brown, G. J. (2018). Robust Binaural Localization of a Target Sound Source by Combining Spectral Source Models and Deep Neural Networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(11) :2122–2131.
- Maeda, S. (1974). A characterization of fundamental frequency contours of speech. Technical report, MIT Research Laboratory of Electronics. Quarterly Progress Report.
- Mairal, J., Bach, F., and Ponce, J. (2012). Task-driven dictionary learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(4) :791–804.
- Makous, J. and Middlebrooks, J. (1990). Two-dimensional sound localization by human listeners. *The Journal of the Acoustical Society of America*, 87(5) :2188–2200.
- Mandel, M. I. and Ellis, D. P. W. (2007). EM Localization and Separation Using Interaural Level and Phase Cues. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, page 275–278, New York, USA.
- Mandel, M. I., Weiss, R. J., and Ellis, D. P. W. (2010). Model-based Expectation-Maximization Source Separation and Localization. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(2) :382–394.
- Martin, K. (2023). *Complexité vocale et contrôle cognitif chez le corbeau freux (Corvus frugilegus)*. Thèse de doctorat, Universtié de Tours.
- Martin, P. (1987). Prosodic and rhythmic structures in french. *Linguistics*, 25(5) :925–950.
- McCrae, R. R. and Costa, P. T. (1990). *Personality in Adulthood*. New York : The Guilford Press.

- McCulloch, W. S. and Pitts, W. (1943). Calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5 :115–133.
- McKinley, R. L. and Ericson, M. A. (2005). Human Auditory Localization Performance in Azimuth. *The Journal of the Acoustical Society of America*, 85(S1) :S38–S39.
- McNeill, D., Bertenthal, B., Cole, J., and Gallagher, S. (2005). Gesture-first, but no gestures? *Behavioral and Brain Sciences*, 28(2) :138–139.
- Mehrabian, A. (1972). *Nonverbal Communication*. Walter De Gruyter, New-York.
- Mertens, P. (2004). The Prosogram : Semi-Automatic Transcription of Prosody based on a Tonal Perception Model. In *Speech Prosody*, pages 549–552, Nara, Japan.
- Mertens, P. (2013). Automatic Labelling of Pitch Levels and Pitch Movements in Speech Corpora. In *Tools and Resources for the Analysis of Speech Prosody (TRASP)*, Aix-en-Provence.
- Migliore, O. (2016). *Analyser la prosodie musicale du punk, du rap et du ragga français (1977-1992) à l'aide de l'outil informatique*. Thèse de doctorat, Université Montpellier 3 - Paul Valéry.
- Migliore, O. (2023). Analyser l'interprétation vocale des musiques populaires avec le logiciel Audiosculpt : chronique d'une pratique de l'analyse musicale assistée par ordinateur. *Musurgia*.
- Ming, H., Huang, D., Dong, M., , Li, H., Xie, L., and Zhang, S. (2015). Fundamental frequency modeling using wavelets for emotional voice conversion. In *Affective Computing Intell. Interact.*, page 804, Åi809.
- Ming, H., Huang, D., Xie, L., Wu, J., Dong, M., and Li, H. (2016). Deep Bidirectional LSTM Modeling of Timbre and Prosody for Emotional Voice Conversion. In *Interspeech 2016*.
- Minsky, M. L., Rochester, N., Shannon, C. E., and McCarthy, J. (1955).
- Mishra, T., Van Santen, J., , and Klabbers, E. (2006). Decomposition of Pitch Curves in the General Superpositional Intonation Model. In *Speech Prosody*, Dresden, Germany.
- Misra, H., Ikbal, S., Boulard, H., and Hermansky, H. (2004). Spectral Entropy based Feature for Robust ASR. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 193–196.
- Mitchell, P. R. and Essa, I. A. (2006). Estimating the Spatial Position of Spectral Components in Audio. In *Independent Component Analysis and Blind Signal Separation (ICA)*, pages 666–673, Dublin, Ireland.
- Mohammadi, S. H. and Kim, T. (2019). One-Shot Voice Conversion with Disentangled Representations by Leveraging Phonetic Posteriorgrams. In *ISCA Interspeech*, pages 704–708.
- Monoson, P. and Zemlin, W. R. (1984). Quantitative Study of a Whisper. *Folia Phoniat*, 36(2) :53–65.
- Moore, R. K. (2007). Spoken Language Processing : Piecing Together the Puzzle. *Speech Communication*, 49(5) :418–435.
- Mori, M. (1970). The Uncanny Valley. *Energy*, 7 :33–35.
- Morise, M., Yokomori, F., and Ozawa, K. (2016). WORLD : A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications. *IEICE Transactions on Information and Systems*, E99.D(7) :1877–1884.

- Morlec, Y. (1997). *Génération multiparamétrique de la prosodie du français par apprentissage automatique*. PhD. thesis, Institut de la Communication Parlée, Grenoble.
- Mysore, G. J. and Smaragdis, P. (2012). A Non-negative Approach to Language informed Speech Separation. In *Latent Variable Analysis and Signal Separation*, pages 356–363. Springer.
- Nattiez, J.-J. (1987). *Musicologie générale et sémiologie*. Christian Bourgois : Paris.
- Nooteboom, S. (1997). The prosody of speech : melody and rhythm. *The handbook of phonetic sciences*, 5 :640–673.
- Noé, P.-G., Mohammadamini, M., Matrouf, D., Parcollet, T., Nautsch, A., and Bonastre, J.-F. (2021). Adversarial Disentanglement of Speaker Representation for Attribute-Driven Privacy Preservation. In *Interspeech*, pages 1902–1906.
- Obin, N. (2011). *MeLos : Analysis and Modelling of Speech Prosody and Speaking Style*. PhD. Thesis, Ircam - UPMC.
- Obin, N. (2012). Cries and Whispers - Classification of Vocal Effort in Expressive Speech. In *Interspeech*, pages 2234–2237, Portland, USA.
- Obin, N. and Belião, J. (2018). Sparse Coding of Pitch Contours with Deep Auto-Encoders. In *International Conference on Speech Prosody*, pages 799–803.
- Obin, N., Belião, J., Veaux, C., and Lacheret, A. (2014a). SLAM : Automatic Stylization and Labelling of Speech Melody. In *Speech Prosody*, pages 246–250.
- Obin, N., Lacheret, A., and Rodet, X. (2011a). Stylization and Trajectory Modelling of Short and Long Term Speech Prosody Variations. In *Interspeech*, pages 2029–2032, Florence, Italy.
- Obin, N., Lanchantin, P., Lacheret, A., and Rodet, X. (2011b). Discrete/Continuous Modelling of Speaking Style in HMM-based Speech Synthesis : Design and Evaluation. In *Interspeech*, page Submitted, Florence, Italy.
- Obin, N., Lanchantin, P., Lacheret, A., and Rodet, X. (2011c). Reformulating Prosodic Break Model into Segmental HMMs and Information Fusion. In *Interspeech*, pages 1829–1832, Florence, Italy.
- Obin, N. and Liuni, M. (2012). On the Generalization of Shannon Entropy for Speech Recognition. In *IEEE workshop on Spoken Language Technology*, Miami, USA.
- Obin, N. and Roebel, A. (2016). Similarity Search of Acted Voices for Automatic Voice Casting. *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, 24(9) :1638–1647.
- Obin, N., Roebel, A., and Bachman, G. (2014b). On Automatic Voice Casting for Expressive Speech : Speaker Recognition vs. Speech Classification. In *International Conference on Acoustics, Speech, and Signal Processing*, Florence, Italy.
- Obin, N., Veaux, C., and Lanchantin, P. (2012). Making Sense of Variations : Introducing Alternatives in Speech Synthesis. In *Speech Prosody*, pages 179–182, Shanghai, China.
- Obin, N., Veaux, C., and Lanchantin, P. (2015). *Exploiting Alternatives for Text-To-Speech Synthesis : From Machine to Human*, pages 189–202. Springer Berlin Heidelberg, Berlin, Heidelberg.
- Ohala, J. (1994). *The Frequency Codes underlies the Sound Symbolic Use of Voice Pitch*, chapter Sound symbolism, pages 325–347. Cambridge University Press.

- Ohala, J. (1996). Ethological theory and the expression of emotion in the voice. In *International Conference on Spoken Language Processing*, pages 1812–1815, Philadelphia, USA.
- OpenAI (2022). ChatGPT : Optimizing Language Models for Dialogue. <https://openai.com/blog/chatgpt/>.
- Ozerov, A. and Fevotte, C. (2010). Multichannel Nonnegative Matrix Factorization in Convolutional Mixtures for Audio Source Separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3) :550 – 563.
- Paatero, P. (1997). Least squares formulation of robust non-negative factor analysis. *Chemometrics and Intelligent Laboratory Systems*, 37(1) :23–35.
- Pan, S. and He, L. (2021). Cross-speaker Style Transfer with Prosody Bottleneck in Neural Speech Synthesis. In *ISCA Interspeech*.
- Pan, X., Tewari, A., Leimkühler, T., Liu, L., Meka, A., and Theobalt, C. (2023). Drag Your GAN : Interactive Point-based Manipulation on the Generative Image Manifold. In *ACM SIGGRAPH 2023 Conference Proceedings*.
- Paris, B. and Donovan, J. (2019). Deepfakes and cheapfakes. *United States of America : Data Society*.
- Peeters, G. (2001). *Modèles et modification du signal sonore adaptés aux caractéristiques locales*. Thèse de doctorat, Université Pierre et Marie Curie, Paris, France.
- Peterson, G. E. and Barney, H. L. (1952). Control methods used in a study of the vowels. *The Journal of the acoustical society of America*, 24(2) :175–184. Publisher : Acoustical Society of America.
- Pfützinger, H. R. (2006). Five Dimensions of Prosody : Intensity, Intonation, Timing, Voice Quality, and Degree of Reduction. In *Speech Prosody*, Dresden, Germany. Keynote.
- Phokhinnan, W., Argentieri, S., and Obin, N. (2023). Binaural Sound Localization in Noisy Environments Using Frequency-Based Audio Vision Transformer (FAViT). In *Interspeech*, Dublin, Ireland.
- Picq, P. (2019). *L'Intelligence artificielle et les chimpanzés du futur* . Odile Jacob : Paris.
- Pierce, J. R. and Karlin, J. E. (1957). Reading Rates and the Information Rate of a Human Channel. *Bell System Technical Journal*, 36(2) :497–516.
- Pierrehumbert, J. (1980). *The phonology and phonetics of English Intonation*. Phd. thesis, Massachusetts Institute of Technology.
- Pierrehumbert, J. and Hirschberg, J. (1990). *Intention in communication*, chapter The meaning of intonation in the interpretation of discourse, page 271–311. MIT Press.
- Ponsot, E., Burred, J. J., Belin, P., and Aucouturier, J.-J. (2018). Cracking the Social Code of Speech Prosody using Reverse Correlation. *Proceedings of the National Academy of Sciences*, 115(15) :3972–3977.
- Post, B., Delais-Roussarie, E., and Simon, A.-C. (2006). IVTS, un système de transcription pour la variation prosodique. *Bulletin de la Phonologie du Français Contemporain*, 6 :51–68.
- Prajwal, K. R., Mukhopadhyay, R., Namboodiri, V. P., and Jawa-har, C. V. (2020). A Lip Sync Expert Is All You Need for Speech to Lip Generation In The Wild. In *ACM International Conference on Multimedia*, pages 257–264, Seattle, USA.

- Prenger, R., Valle, R., and Catanzaro, B. (2019). Waveglow : A Flow-based Generative Network for Speech Synthesis. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3617–3621.
- Price, P. J., Ostendorf, M., Shattuck-Hufnagel, S., and Fong, G. (1991). The use of prosody in syntactic disambiguation. *Journal of the Acoustical Society of America*, 90 :2956–2970.
- Prince, S. (2012). *Computer Vision : Models Learning and Inference*. Cambridge University Press.
- Prince, S. J. D. and Elder, J. H. (2007). Probabilistic Linear Discriminant Analysis for Inferences about Identity. In *IEEE International Conference on Computer Vision*, pages 1751–1758, Rio de Janeiro, Brazil.
- Proust, M. (1927). *Le Temps Retrouvé*. Gallimard : Paris.
- Przedlacka, J. (2012). An introduction to English phonetics. *Journal of the International Phonetic Association*, 42(1) :113–115.
- Qian, K., Jin, Z., Hasegawa-Johnson, M., and Mysore, G. J. (2020a). Fo-Consistent Many-To-Many Non-Parallel Voice Conversion Via Conditional Autoencoder. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6284–6288.
- Qian, K., Zhang, Y., Chang, S., Hasegawa-Johnson, M., and Cox, D. (2020b). Unsupervised Speech Decomposition via Triple Information Bottleneck. In *International Conference on Machine Learning (ICML)*, pages 7836–7846.
- Qian, K., Zhang, Y., Chang, S., Yang, X., and Hasegawa-Johnson, M. (2019). AutoVC : Zero-Shot Voice Style Transfer with Only Autoencoder Loss. In *International Conference on Machine Learning (ICML)*, pages 5210–5219.
- Qin, Z., Kim, D., and Gedeon, T. (2015). Rethinking Softmax with Cross-Entropy : Neural Network Classifier as Mutual Information Estimator. In *International Conference on Machine Learning (ICML)*.
- Quillot, M. (2020). *Un premier pas vers la caractérisation de l'information véhiculée par les voix actées : dualité des informations personnage et locuteur*. PhD. Thesis, Laboratoire d'Informatique d'Avignon, Université d'Avignon.
- Quillot, M., Guillou, L., Gresse, A., Ferro, R., Röth, R., Malinas, D., Dufour, R., Roebel, A., Obin, N., Bonastre, J.-F., and Ethis, E. (2020). La voix actée : pratiques, enjeux, applications (acted voice : practices, challenges, applications). In *Journées d'Études sur la Parole (JEP)*, pages 525–533, Nancy, France.
- Rameau, J.-P. (1722). *Le Traité de l'harmonie réduite à ses principes naturels*. Paris, France.
- Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. (2022). Hierarchical Text-Conditional Image Generation with CLIP Latents. *arXiv :2204.06125 [cs.CV]*.
- Raspaud, M., Viste, H., and Evangelista, G. (2010). Binaural Source Localization by Joint Estimation of ILD and ITD. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(1) :68–77.
- Rayleigh, L. (1907). On our Perception of Sound Direction. *Philosophical Magazine Series 6*, 13(74) :214–232.
- Rényi, A. (1961). On Measures of Entropy and Information. In *Berkeley Symposium on Mathematical Statistics and Probability*, pages 547–561, Berkeley, California.
- Reynolds, D. A., Quatieri, T. F., and Dunn, R. B. (2000). Speaker Verification Using Adapted Gaussian Mixture Models. *Digital Signal Processing*, 10(1-3) :19–41.

- Rifkin, R. and Klautau, A. (2004). In Defense of One-Vs-All Classification. *Journal of Machine Learning Research*, 5 :101–141.
- Rix, A. W., Beerends, J. G., Hollier, M. P., and Hekstra, A. P. (2001). Perceptual evaluation of speech quality (PESQ) - a new method for speech quality assessment of telephone networks and codecs. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, pages 749–752.
- Robinson, C., Obin, N., and Roebel, A. (2019). Sequence-to-Sequence Modelling of Fo for Speech Emotion Conversion. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6830–6834.
- Roebel, A. and Bous, F. (2022). Neural Vocoding for Singing and Speaking Voices with the Multi-band Excited WaveNet. *Information*, 13(3) :103.
- Ronanki, S., Henter, G. E., Wu, Z., and King, S. (2016). A Template-Based approach for speech synthesis intonation generation using LSTMs. In *Interspeech 2016*.
- Rostolland, D. (1982). Acoustic Features of Shouted Voice. *Acustica*, 50 :118–125.
- Röbel, A. (2010). Shape-invariant Speech Transformation with the Phase Vocoder. In *International Conference on Spoken Language Processing (InterSpeech)*, pages 2146–2149.
- Röbel, A. and Rodet, X. (2005). Efficient Spectral Envelope Estimation and its application to Pitch Shifting and Envelope Preservation. In *International Conference on Digital Audio Effects (DAFx)*, pages 30–35.
- Sakamoto, R. (2018). *Le son pur de Ryuichi Sakamoto*. France Culture : Paris.
- Salais, L. (2021). *Manipulation des attributs de la voix par apprentissage de représentations neuronales demelees*. Thèse de doctorat, Sorbonne Université.
- Salais, L., Arias, P., Moine, C. L., , Rosi, V., Teytaut, Y., Obin, N., and Roebel, A. (2022). Production Strategies of Vocal Attitudes. In *Interspeech*, Incheon, Korea.
- Scherer, K. R., Banse, R., Wallbott, H. G., and Goldbeck, T. (1991). Vocal Cues in Emotion Encoding and Decoding. *Motivation and Emotion*, 15 :123–148.
- Schuller, B. and Batliner, A. (2013). *Computational Paralinguistics : Emotion, Affect and Personality in Speech and Language Processing*. Wiley.
- Schuller, B., Steidl, S., and Batliner, A. (2009). The Interspeech 2009 Emotion Challenge. In *Interspeech*, Brighton, UK.
- Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., and Narayanan, S. (2010). The Interspeech 2010 Paralinguistic Challenge. In *Interspeech*, Makuhari, Japan.
- Schuller, B., Steidl, S., Batliner, A., Nöth, E., Vinciarelli, A., Burkhardt, F., Van Son, R., Weninger, F., Eyben, F., Bocklet, T., Mohammadi, G., and Weiss, B. (2012a). The Interspeech 2012 Speaker Trait Challenge. In *Interspeech*, Portland, USA.
- Schuller, B., Steidl, S., Batliner, A., Schiel, F., and Krajewski, J. (2012b). The Interspeech 2011 Speaker State Challenge. In *Interspeech*, Florence, Italy.
- Selrik, E. (1981). On prosodic structure and its relation to syntactic structure. In *Nordic Prosody II*, pages 111–140, Trondheim, Norway.
- Serizel, R., Essid, S., and Richard, G. (2016). Mini-batch Stochastic approaches for Accelerated Multiplicative Updates in Nonnegative Matrix Factorisation with Beta-divergence. In *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, Vietri sul Mare, Italy.

- Seth, A. (2021). *Being you : A new science of consciousness*. Penguin, New-York.
- Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(3) :379–423.
- Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R., Saurous, R. A., Agiomvrgiannakis, Y., and Wu, Y. (2018). Natural TTS Synthesis by Conditioning Wavenet on Mel Spectrogram Predictions. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783.
- Shochi, T., Rilliard, A., Aubergé, V., and Erickson, D. (2009). *The role of prosody in Affective Speech*, chapter Intercultural Perception of English, French and Japanese Social Affective Prosody, pages 31–59. Peter Lang.
- Sigurgeirsson, A. T. and King, S. (2023). Do Prosody Transfer Models Transfer Prosody? In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.
- Sisman, B. and Li, H. (2018). Wavelet Analysis of Speaker Dependent and Independent Prosody for Voice Conversion. In *ISCA Interspeech*, pages 52–56, Hyderabad, India.
- Sisman, B., Li, H., and Tan, K. C. (2017). Transformation of Prosody in Voice Conversion. In *IEEE Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pages 1537–1546.
- Smaragdis, P. (2007). Convolutional speech bases and their application to supervised speech separation. *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, 15(1) :1–12.
- Smits, L. (2021). *Master of Voice*. Sternberg Press : New York.
- Solomon, N. P., McCall, G. N., Trosset, M. W., and Gray, W. C. (1989). Laryngeal Configuration and Constriction during Two Types of Whispering. *Journal of Speech and Hearing Research*, 32 :161–174.
- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout : A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research (JMLR)*, 15 :1929–1958.
- Stevens, S. S. and Newman, E. B. (1936). The Localization of Actual Sources of Sound. *American Journal of Psychology*, 21 :297–306.
- Stiegler, B. (2013). *De la misère symbolique*. Flammarion.
- Stylianou, Y., Cappé, O., and Moulines, E. (1998). Continuous Probabilistic Transform for Voice Conversion. *IEEE Transactions on Speech and Audio Processing*, 6(2) :131–142.
- Sun, D. L. and Mysore, G. J. (2013). Universal speech models for speaker independent single channel source separation. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 141–145.
- Sun, L., Li, K., Wang, H., Kang, S., and Meng, H. (2016). Phonetic Posteriorgrams for Many-to-One Voice Conversion without Parallel Data Training. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6.
- Sundberg, J., Scherer, R., Hess, M., and Müller, F. (2010). Whispering - A Single-Subject Study of Glottal Configuration and Aerodynamics. *Journal of Voice*, 24(5) :574–584.
- Sutskever, I., Vinyals, O., and Le, Q. V. (2014a). Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems (NIPS)*.

- Sutskever, I., Vinyals, O., and Le, Q. V. (2014b). Sequence to Sequence Learning with Neural Networks. In *International Conference on Neural Information Processing Systems (NIPS)*, page 3104–3112.
- 't Hart, J. (1981). Differential sensitivity to pitch distance, particularly in speech. *Journal of the Acoustical Society of America*, 69(3) :811–821.
- 't Hart, J., Collier, R., and Cohen, A. (1990). *A perceptual study of intonation : an experimental-phonetic approach to speech melody*. Cambridge Studies in Speech Science and Communication. Cambridge University Press.
- Tachibana, M., Yamagishi, J., Masuko, T., and Kobayashi, T. (2005). Speech Synthesis with Various Emotional Expressions and Speaking Styles by Style Interpolation and Morphing. *IEICE Transaction on Information. and Systems*, E88-D(11) :2484–2491.
- Tagliasacchi, M., Gfeller, B., Quitry, F. d. C., and Roblek, D. (2020). Pre-Training Audio Representations With Self-Supervision. *IEEE Signal Processing Letters*, 27 :600–604.
- Tanaka, K., Kameoka, H., Kaneko, T., and Hojo, N. (2019). AttS2S-VC : Sequence-to-Sequence Voice Conversion with Attention and Context Preservation Mechanisms. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6805–6809.
- Taylor, P. (1998). The TILT intonation model. In *International Conference on Spoken Language Processing*, pages 1383–1386, Sydney, Australia.
- Teutenberg, J., Watson, C., and Riddle, P. (2008). Modelling and Synthesising Fo contours with the Discrete Cosine Transform. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 3973–3976, Las Vegas, U.S.A.
- Tishby, N. and Zaslavsky, N. (2015). Deep Learning and the Information Bottleneck Principle. In *IEEE Information Theory Workshop (ITW)*.
- Toda, T., Ohtani, Y., and Shikano, K. (2007). One-to-Many and Many-to-One Voice Conversion Based on Eigenvoices. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 4, pages 1249–1252.
- Toda, T. and Tokuda, K. (2007). A Speech Parameter Generation Algorithm considering Global Variance for HMM-based Speech Synthesis. *IEICE Transactions on Information and Systems*, 90(5) :816–824.
- Toda, T. and Young, S. (2009). Trajectory training considering Global Variance for HMM-based Speech Synthesis. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 4025–4028, Taipei, Taiwan.
- Tokuda, K., Masuko, T., Miyazaki, N., and Kobayashi, T. (1999). Hidden Markov models based on multi-space probability distribution for pitch pattern modeling. In *International Conference on Audio, Speech, and Signal Processing*, pages 229–232, Phoenix, Arizona.
- Tokuda, K., Masuko, T., Yamada, T., Kobayashi, T., and Imai, S. (1995). An algorithm for speech parameter generation from continuous mixture HMMs with dynamic features. In *European Conference on Speech Communication and Technology*, pages 757–760, Madrid, Spain.
- Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., and Kitamura, T. (2000). Speech parameter generation algorithms for HMM-based speech synthesis. In *International Conference on Spoken Language Processing*, pages 2347–2350, Beijing, China.
- Tokuda, K., Zen, H., and Black, A. (2002). An HMM-based speech synthesis system applied to English. In *Workshop on Speech synthesis*, pages 227–230.

- Tokuda, K., Zen, H., and Kitamura, T. (2003). Trajectory modeling based on HMMs with the explicit relationship between static and dynamic features. In *European Conference on Speech Communication and Technology*, pages 865–868, Geneva, Switzerland.
- Torrence, C. and Compo, G. P. (1998). A practical guide to wavelet analysis. *Bulletin of the American Meteorological Society*, 79 :61–78.
- Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, 59(236) :433–460.
- Van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K. (2016). WaveNet : A Generative Model for Raw Audio.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is All you Need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30.
- Veaux, C., Lanchantin, P., , and Rodet, X. (2010). Joint Prosodic and Segmental Unit Selection for Expressive Speech Synthesis. In *Speech Synthesis Workshop (SSW7)*, pages 323–327, Kyoto, Japan.
- Veaux, C. and Rodet, X. (2011). Intonation Conversion from Neutral to Expressive Speech. In *Interspeech*, pages 2765–2768, Florence, Italy.
- Velez, D. R., White, B. C., Motsinger, A. A., Bush, W. S., Ritchie, M. D., Williams, S. M., and Moore, J. H. (2007). A Balanced Accuracy Function for Epistasis Modeling in Imbalanced Datasets using Multifactor Dimensionality Reduction. *Genetic Epidemiology*, 31(4) :306–315.
- Villavicencio, F., Röbel, A., and Rodet, X. (2006). Improving LPC Spectral Envelope Extraction of Voiced Speech by True-Envelope Estimation. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 869–872, Toulouse, France.
- Villavicencio, F., Röbel, A., and Rodet, X. (2009). Applying improved Spectral Modeling for High Quality Voice Conversion. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2009, 19-24 April 2009, Taipei, Taiwan*, pages 4285–4288.
- Vincent, E., Gribonval, R., and Févotte, C. (2006). Performance Measurement in Blind Audio Source Separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4) :1462–1469.
- Vincent, E., Sawada, H., Bofill, P., Makino, S., and Rosca, J. (2007). First Stereo Audio Source Separation Evaluation Campaign : Data, Algorithms and Results. In *International Conference on Independent Component Analysis and Blind Source Separation (ICA)*, pages 552–559, London, United Kingdom.
- Virtanen, T. (2007). Monaural Sound Source Separation by Nonnegative Matrix Factorization With Temporal Continuity and Sparseness Criteria. *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, 15(3) :1066–1074.
- Virtanen, T., Gemmeke, J. F., and Raj, B. (2013). Active-Det Newton Algorithm for Overcomplete Non-Negative Representations of Audio. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(11) :2277–2289.
- Virtanen, T. and Klapuri, A. (2006). Analysis of Polyphonic Audio using Source-Filter Model and Non-Negative Matrix Factorization. In *Advances in models for acoustic processing, neural information processing systems workshop*.

- Viste, H. and Evangelista, G. (2004). Binaural Source Localization. In *Digital Audio Effects (DAFx) Conference*, page 145–150, Naples, Italy.
- Viña, C., Argentieri, S., and Rébillat, M. (2011). A Spherical Cross-channel Algorithm for Binaural Sound Localization . In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2921–2926, Tokyo, Japan.
- von Uexküll, J. (1921). *Umwelt und Innenwelt der Tiere*. Springer.
- Wan, V., Agiomyrgiannakis, Y., Silen, H., and Vít, J. (2017). Google’s Next-Generation Real-Time Unit-Selection Synthesizer Using Sequence-to-Sequence LSTM-Based Autoencoders. In *Interspeech*, Stockholm, Sweden.
- Wang, C., Chen, S., Wu, Y., Zhang, Z., Zhou, L., Liu, S., Chen, Z., Liu, Y., Wang, H., Li, J., He, L., Zhao, S., and Wei, F. (2023). Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers. *arXiv :2301.02111 [cs.CL]*.
- Wang, X., Takaki, S., and Yamagishi, J. (2017a). An RNN-Based Quantized Fo Model with Multi-Tier Feedback Links for Text-to-Speech Synthesis. In *Interspeech*, Stockholm, Sweden.
- Wang, Y., Skerry-Ryan, R. J., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., Le, Q. V., Agiomyrgiannakis, Y., Clark, R., and Saurous, R. A. (2017b). Tacotron : Towards End-to-End Text-To-Speech Synthesis. In *Interspeech*, pages 4006–4010, Stockholm, Sweden.
- Wang, Y., Stanton, D., Zhang, Y., Ryan, R.-S., Battenberg, E., Shor, J., Xiao, Y., Jia, Y., Ren, F., and Saurous, R. A. (2018). Style Tokens : Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis. In *International Conference on Machine Learning (ICML)*, volume 80, pages 5180–5189.
- Wanger, H. (1993). Measuring Performance in Category Judgment Studies on Nonverbal Behavior. *Journal of Nonverbal Behavior*, 17(1) :3–28.
- Weizenbaum, J. (1966). Eliza—a computer program for the study of natural language communication between man and machine. *Commun. ACM*, 9(1) :36–45.
- Weizenbaum, J. (1993). Coevolution of neocortical size, group size and language in humans. *Behavioral and Brain Sciences*, 16(4) :681–694.
- Weninger, F., Feliu, J., and Schuller, B. (2012). Supervised and Semi-Supervised Suppression of Background Music in Monaural Speech Recordings. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 61–64.
- Wichmann, A. (2000). The attitudinal effects of prosody, and how they relate to emotion. In *ITRW on Speech and Prosody*, Newcastle, UK.
- Wiener, N. (1954). In *The human use of human beings : Cybernetics and society*, chapter Cybernetics in History, pages 15–27. Houghton Mifflin, Boston.
- Wiener, N. (1961). *Cybernetics Or Control and Communication in the Animal and the Machine*. MIT Press.
- Wightman, C. (2002). ToBI or not ToBI? In *Speech Prosody*, pages 25–29, Aix-en-Provence, France.
- Witzel, E. M. (2013). *The Origins of the World’s Mythologies*. Oxford University Press.
- Woodruff, J. and Wang, D. (2012). Binaural Localization of Multiple Sources in Reverberant and Noisy Environments. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(5) :1503 – 1512.

- Wu, T.-F., Lin, C.-J., and Weng, R. C. (2004). Probability Estimates for Multi-Class Classification by Pairwise Coupling. *Journal of Machine Learning Research*, 5 :975–1005.
- Xia, Y., Qin, T., Chen, W., Bian, J., Yu, N., and Liu, T.-Y. (2017). Dual Supervised Learning. In *International Conference on Machine Learning (ICML)*, pages 3789–3798.
- Yamagishi, J. (2006). *Average-Voice-Based Speech Synthesis*. PhD. thesis, Tokyo Institute of Technology.
- Yamagishi, J., Veaux, C., and Macdonald, K. (2019). CSTR VCTK Corpus : English Multi-speaker Corpus for CSTR Voice Cloning Toolkit (version 0.92). University of Edinburgh. The Centre for Speech Technology Research (CSTR).
- Yin, X., Lei, M., Qian, Y., Soong, F. K., He, L., Ling, Z.-H., and Dai, L.-R. (2016). Modeling Fo trajectories in hierarchically structured deep neural networks. *Speech Communication*, 76 :82–92.
- Yoshimura, T., Tokuda, K., Masuko, T., and Kitamura, T. (2001). Mixed-excitation for HMM-based speech synthesis. In *European Conference on Speech Communication and Technology*, pages 2259–2262.
- Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T. (1999). Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. In *European Conference on Speech Communication and Technology*, pages 2347–2350, Budapest, Hungary.
- Yuan, S., Cheng, P., Zhang, R., Hao, W., Gan, Z., and Carin, L. (2021). Improving Zero-Shot Voice Style Transfer via Disentangled Representation Learning. In *International Conference on Learning Representations (ICLR)*.
- Zeghidour, N., Luebs, A., Omran, A., Skoglund, J., and Tagliasacchi, M. (2022). SoundStream : An End-to-End Neural Audio Codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, 30 :495–507.
- Zen, H. (2015). Statistical Parametric Speech Synthesis : from HMM to LSTM-RNN. In *RTTH Summer School on Speech Technology*, Barcelona, Spain.
- Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A., and Tokuda, K. (2007). The HMM-based Speech Synthesis System Version 2.0. In *Speech Synthesis Workshop*, pages 294–299, Bonn, Germany.
- Zen, H. and Sak, H. (2015). Unidirectional Long Short-Term Memory Recurrent Neural Network with Recurrent Output Layer for Low-Latency Speech Synthesis. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4470–4474, Brisbane, Australia.
- Zen, H., Senior, A., and Schuster, M. (2013). Statistical Parametric Speech Synthesis using Deep Neural Networks. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7962–7966.
- Zhang, J.-X., Ling, Z.-H., and Dai, L.-R. (2020). Non-Parallel Sequence-to-Sequence Voice Conversion With Disentangled Linguistic and Speaker Representations. *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, 28 :540–552.
- Zhang, M.-L. and Zhou, Z.-H. (2014). A Review on Multi-Label Learning Algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 26(8) :1819 – 1837.
- Zhang, Y., Weiss, R. J., Zen, H., Wu, Y., Chen, Z., Skerry-Ryan, R. J., Jia, Y., Rosenberg, A., and Ramabhadran, B. (2019). Learning to Speak Fluently in a Foreign Language : Multilingual Speech Synthesis and Cross-Language Voice Cloning. In *Interspeech*, pages 2080–2084.

- Zhao, Y., Huang, W.-C., Tian, X., Yamagishi, J., Das, R. K., Kinnunen, T., Ling, Z., and Toda, T. (2020). Voice Conversion Challenge 2020 : Intra-Lingual Semi-Parallel and Cross-Lingual Voice Conversion. In *Interspeech*, pages 80–98.
- Zhou, C., Horgan, M., Kumar, V., Vasco, C., and Darcy, D. (2018). Voice Conversion with Conditional SampleRNN. In *ISCA Interspeech*, pages 1973–1977.
- Zhou, K., Sisman, B., and Li, H. (2020a). Transforming Spectrum and Prosody for Emotional Voice Conversion with Non-Parallel Training Data. In *The Speaker and Language Recognition Workshop (Odyssey 2020)*, pages 230–237.
- Zhou, K., Sisman, B., Rana, R., Schuller, B. W., and Li, H. (2022). Speech Synthesis with Mixed Emotions. *IEEE Transactions on Affective Computing*, pages 1–16.
- Zhou, K., Sisman, B., Zhang, M., and Li, H. (2020b). Converting Anyone’s Emotion : Towards Speaker-Independent Emotional Voice Conversion. In *Interspeech*, pages 3416–3420.
- Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017). Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2242–2251.
- Zivanovic, M., Röbel, A., and Rodet, X. (2008). Adaptive Threshold Determination for Spectral Peak Classification. *Computer Music Journal*, 32(2) :57–67.
- Zue, V., Seneff, S., and Glass, J. (1990). Speech database development at MIT : TIMIT and beyond. *Speech Communication*, 9(4) :351–356.