Journée commune AFIA-TLH / AFCP – "Extraction de connaissances interprétables pour l'étude de la communication parlée"

Corinne Fredouille, Maëva Garnier, Olivier Perrotin, Marie Tahon

Journée commune AFIA-TLH – AFCP

Extraction de connaissances interprétables pour l'étude de la communication parlée

Avignon, 11 Décembre 2023

Organisée par :

Corinne Fredouille[1,2,3], Maëva Garnier[2,5], Olivier Perrotin[2,5], Marie Tahon[1,2,4]

[1]AFIA, [2]AFCP

[3]Avignon Université, LIA,

[4]Le Mans Université, LIUM

[5]Université Grenoble Alpes, CNRS, GIPSA-lab, France

AfIA

Association française pour l'Intelligence Artificielle

Préambule

L'Association Française pour l'Intelligence Artificielle (AFIA), au travers de son collège Technologies du Langage Humain (TLH), organise avec l'Association Francophone de la Communication Parlée (AFCP), une première journée commune sur le thème "Extraction de connaissances interprétables pour l'étude de la communication parlée" le lundi 11 décembre 2023 dans les locaux d'Avignon Université.

L'objectif de cette journée est de réunir chercheur.euse.s dont l'objet d'étude est la communication parlée, que ce soit du point de vue des Sciences Humaines et Sociales (SHS) ou du Traitement Automatique des Langues et de l'Intelligence Artificielle. Il s'agira au cours de cette journée d'aborder la question de l'extraction de connaissances interprétables dans le signal de parole par le biais d'approches automatiques, en particulier basées sur des apprentissages profonds, pour l'étude de la communication parlée au sens large. Ces études pourront porter sur des thématiques comme l'analyse de la parole dans le domaine de la phonétique ou de la linguistique, la caractérisation du locuteur pour des tâches de reconnaissance, de segmentation et regroupement en locuteurs, de comparaison de voix (criminalistique), l'analyse de la voix/parole pathologique, l'analyse des informations para-linguistiques (autre que le locuteur) comme la parole expressive, les émotions, les accents régionaux, etc., l'étude de comportements cognitifs autour de l'acquisition de la parole, ... Côté Traitement Automatique des Langues et de l'Intelligence Artificielle, les thèmes autour des modèles auto-supervisés de représentation de la parole, de l'explicabilité des modèles, de l'évaluation de l'interprétabilité et de la pertinence des explications, des boucles interactives avec l'utilisateur, pourront également être abordés.

Cette journée est ainsi l'occasion de montrer des approches automatiques déjà existantes d'extractions de connaissances interprétables, pour répondre aux besoins des chercheur.euse.s en SHS mais également d'exprimer de la part de ces derniers, de nouveaux besoins.

Elle s'adresse aussi bien aux jeunes chercheur.euse.s qu'aux chercheur.euse.s plus avancé.e.s du domaine. Elle est ouverte à la présentation de travaux à différents stades d'avancement voire à la présentation de projets de recherche en voie d'être lancés.

La journée débute par l'intervention d'un conférencier invité puis est rythmée par des communications orales et des posters[1]. Une discussion animée en fin de journée dresse un bilan des applications abordées (génération et reconnaissance automatique, développement du langage, apprentissage L2, production et pathologies vocales) ainsi que des méthodes d'exploration. Deux explorations principales émergent des présentations : l'observation d'espaces de représentation appris par les modèles, et l'interprétation du comportement du modèle par rapport à une tâche incluant des experts humains. Les échanges ont également permis de mettre en avant l'importance de l'interdisciplinarité pour l'exploration et la validation des explications et interprétations obtenues empiriquement.

[1] Le programme complet est disponible sur le lien : https://www.afcp-parole.org/journee-commune-afia-tlh-afcp-11-12-23/

Mot de Martine ADDA-DECKER, présidente de l'AFCP

Cher·e·s collègues,

En tant que présidente de l'AFCP, Association Francophone de la Communication Parlée, je vous souhaite la bienvenue à cette journée qui s'intéresse à l'étude de la communication parlée à travers le prisme du traitement automatique des langues, de l'intelligence artificielle et l'apprentissage profond.

En effet, comme dans beaucoup d'autres domaines, l'IA, l'approche neuronale et l'apprentissage profond ont permis des avancées majeures dans les technologies concernant la parole comme la transcription automatique de la parole, la synthèse, la traduction, la reconnaissance du locuteur... Si les performances permettent aujourd'hui d'innombrables applications, on peut en déduire que la modélisation est devenue plus précise, plus complète qu'avec les approches précédentes. Ceci devrait pouvoir bénéficier également à nos connaissances sur la parole et la communication parlée et cet objectif est un des buts de cette journée. Dans quelle mesure ces modèles sont comparables avec des représentations et processus cognitifs humains? quels parallèles, quelles différences? Ce sont des questions qu'on peut se poser et que vous allez certainement évoquer pendant cette journée.

Ainsi, les approches d'apprentissage automatique apportent de nouveaux moyens pour étudier et caractériser la voix et la parole. La matière première pour les modèles si performants et précis estimés via des méthodes mathématiques sophistiquées et des architectures complexes, sont les données, les data, les corpus en tout genre : oral, écrit, video... Les modèles exploitent au mieux ces données pour optimiser la tâche pour laquelle ils ont été conçus. Au-delà des méthodes, les données (ou le choix de données) jouent donc aussi un rôle crucial, car elles représentent la "vérité de terrain". Des données partielles risquent d'aboutir à des vérités partielles ou biaisés. Nous devons garder cela à l'esprit, afin ne pas abandonner nos propres jugements et réflexions d'humains face aux décisions de machines très puissantes et devenues très "savantes".
Je profite aussi de cette journée pour dire quelques mots sur l'AFCP: il s'agit d'une association à but non-lucratif (loi 1901) consacrée au soutien, au développement, à la diffusion et à la promotion des différentes spécialités des sciences de la communication parlée, dans la communauté francophone. Elle a été créée il y a plus de 20 ans, comprend environ 150 membres et est animé par un CA de 18 membres élus. N'hésitez pas à devenir membre et à candidater au CA (https://www.afcp-parole.org/). Les adhésions se font souvent au moment des JEP qui sont notre conférence biennale, et qui se joignent une fois sur deux à la conférence TALN/RECITAL. La prochaine édition JEP/TALN/RECITAL aura lieu à Toulouse au mois de juillet 2024.

Je vous souhaite des échanges fructueux et j'espère vous revoir toutes et tous en juillet 2024 à Toulouse pour les JEP/TALN/Récital.

Martine Adda-Decker

Co-organisation et Comité Scientifique

La journée est co-organisée par Marie Tahon et Corinne Fredouille du collège TLH de l'AFIA et Maëva Garnier et Olivier Perrotin de l'AFCP. Elle est soutenue par le comité scientifique suivant :

Au nom de l'AFCP

Nicolas Audibert (LPP, Paris)
Jean-François Bonastre (INRIA, LIA, Avignon)
Philippe Boula de Mareuil (LISN, Paris Saclay)
Olivier Crouzet (LLING, Nantes)
Maëva Garnier (GIPSA-Lab, Grenoble)
Damien Lolive (ENSSAT, IRISA, Rennes)
Julie Mauclair (IRIT, Toulouse)
Slim Ouni (LORIA, Nancy)
Olivier Perrotin (GIPSA-Lab, Grenoble)

Au nom de l'AFIA

Florian Boudin (L2SN, Nantes)
Davide Buscaldi (LIPN, Paris Nord)
Gaël Dias (GREYC, Caen)
Emmanuelle Esperança-Rodier (LIG, Grenoble)
Corinne Fredouille (LIA, Avignon)
José Moreno (IRIT, Toulouse)
Aurélie Névéol (LISN, Paris Saclay)
Yannick Parmentier (LORIA, Nancy)
Mathieu Roche (TETIS, CIRAD)
Marie Tahon (LIUM, Le Mans)

Programme

Conférence invitée

Session orale 1

Session orale 2

Session posters

Représentations de la parole issues de modèles neuronaux : une étude empirique

Yannick ESTEVE

Laboratoire Informatique d'Avignon

Une des raisons du succès des réseaux de neurones profonds tient à leur capacité à apprendre des représentations pertinentes des données qu'ils ont à traiter.

Historiquement, pour le traitement automatique de la parole comme pour d'autres domaines, la préparation des données, ou plus précisément le choix des caractéristiques alimentant les algorithmes d'apprentissage automatique, s'avérait être une tâche déterminante à réaliser en amont de ces apprentissages.

À l'ère de l'apprentissage profond et auto-supervisé, où d'énormes quantités de données peuvent être exploitées par des capacités de calcul toujours plus importantes, nous laissons aux modèles neuronaux le soin d'apprendre par eux-mêmes ces représentations de la parole, sous forme de représentations vectorielles dans des espaces continus, parfois transformées en unités discrètes.

Dans le cadre de cet exposé, je reviendrai sur différents travaux auxquels j'ai participé ces dernières années, dont l'un des points communs est l'exploitation ou l'analyse de ces représentations de la parole : représentation vectorielle de l'apparence acoustique des mots, représentation de l'expressivité, du locuteur, du contenu linguistique, ou encore de la sémantique. Nécessairement, ces travaux seront mis en relation avec l'état de l'art, évolutif, des domaines concernés.

Exploring the multidimensional representation of unidimensional speech acoustic parameters extracted by deep unsupervised models

Maxime JACQUELIN[1,2], Maëva GARNIER[1], Laurent GIRIN[1], Rémy VINCENT[2], Olivier PERROTIN[1]

[1]Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, F-38000 Grenoble, France
[2]Vogo, F-38190 Bernin, France

Understanding latent representations of speech encoded by deep unsupervised models is fundamental for unlocking the full potential of neural approaches for signal analysis, transformation, and generation. While prior studies have identified the directions of variation of individual acoustic parameters such as fundamental frequency or formant frequencies within deep latent spaces, one also demonstrated that those variations are often explained by multiple latent dimensions. This thus calls for. the following question: Why are multiple dimensions needed for the encoding of one-dimensional parameters within these latent spaces? Among the possible multiple interactions between acoustic parameters, our hypothesis, explored in this study, is that the different dimensions may reflect the different sources of inter- and intra-individual variability of each acoustic parameter.

In the framework of a variational autoencoder (VAE) trained on a multi-speaker database, this work proposes a novel methodology to identify the role of these intricate dimensions within the latent space. Specifically, for interpreting the multi-dimensional aspect of the representation of individual acoustic parameters, we: 1) tailored two test datasets with either controlled variation of single acoustic parameters (synthetic speech) or uncontrolled co-variations of all acoustic parameters (natural speech) ; 2) analyzed the direction of variation of those parameters in the latent space of the VAE with linear analysis, including principal component analysis, linear discriminant analysis and linear regression.

Our investigation first confirmed that each acoustic parameter mentioned above, essential in characterizing speech, is encoded within the VAE's latent space on multiple directions. Among those multiple dimensions, we have demonstrated that one of them directly encodes the global shape of the parameter distribution seen in the training set, pointing out the impact of the training dataset on the performance of our model. Then, we proved that parameter values belonging to each mode of the distribution are encoded on additional distinct dimensions. In the particular case of the fundamental frequency, the parameter distribution is bimodal (corresponding to the two genders) and values belonging to different modes are encoded on two additional and distinct dimensions. Given those findings, we aimed to identify latent directions of variation of acoustic parameters within and between modes of multi-modal distributions, and found disentangled directions in the VAE latent space that explain the between- and within-gender variations.

In summary, our research underscores the pivotal role of latent spaces in deep unsupervised models for speech representation learning. While several studies have used latent space dimension reduction, addressed the orthogonality of the different directions that explain a given parameter, or identified the variation of acoustic parameters in the latent space, this work is one of the few to interpret the multidimensional representation of each unidimensional acoustic parameter, by introducing a systematic methodology that combines the use of specifically designed test sets and linear analysis methods. We believe that our research illuminates the need for more interpretable representations, and that our findings on the unsupervised representation of the inter-and intra-individual variability of each acoustic parameter are a first

step towards finely controllable speech encoding-decoding models, crucial for speech analysis, transformation and synthesis.

Keywords: representation learning, speech encoding, variational autoencoder, source-filter mode

Vers une représentation automatique du rythme de la parole

Jérôme FARINAS [1], Corine ASTESANO [2]

[1] IRIT, Université de Toulouse, CNRS, Toulouse INP, UT3, Toulouse, France
[2] Laboratoire de Neuro-Psycho-Linguistique LNPL, Université Toulouse II, France

Une bonne structuration prosodique en groupements cohésifs hiérarchisés optimise la segmentation de la parole et l'accès au sens pour l'auditeur. Parmi les trois principes organisateurs de la parole (intonation, accentuation et rythme (Di Cristo, 2011)), le rythme constitue le socle de la prosodie puisqu'il permet l'intégration temporelle des proéminences accentuelles et des unités intonatives en lien avec les règles métriques sous-jacentes d'une langue donnée (Arvaniti, 2012). Robin Vaysse, dans son doctorat, a proposé une représentation spectrale du rythme générée automatiquement à partir du signal de parole (Vaysse, 2023) : le spectre de modulation d'amplitude (Envelope Modulation Spectrum – EMS). Cette méthode permet de visualiser les répartitions d'énergie du rythme de la parole. L'EMS est obtenu en calculant l'enveloppe du signal auquel est appliqué un filtre 300-1000 Hz afin de capturer l'énergie des voyelles (Tilsen & Johnson, 2008 ; Vaysse et al. 2021) ; voir courbe noire, figure 1). Un spectre de puissance est ensuite appliqué pour représenter les fréquences de 0 à 10 Hz (courbe orange). Un lissage (courbe rouge pointillée) est également appliqué pour englober les constituants prosodiques de niveau similaire. Sur la figure 1, les niveaux prosodiques sont représentés en couleur et sont issus d'une annotation manuelle : syllabes, mot prosodique (pw; (Astesano, 2019)), syntagme accentuel (AP), syntagme intermédiaire (ip), syntagme intonatif (IP) ; (Di Cristo, 2011). Nous souhaitons nous baser sur cette représentation, et proposer une détection automatique des zones des niveaux prosodiques. Nous pourrions également combiner à cette représentation, pour les fréquences de 0 à 4 Hz, une représentation issue de la courbe de l'intonation, ce qui permettrait d'améliorer la précision dans cette zone. L'évolution temporelle de cette représentation pourrait également être étudiée, et par exemple matérialisée par un spectrogramme du rythme (cf. figure 2).

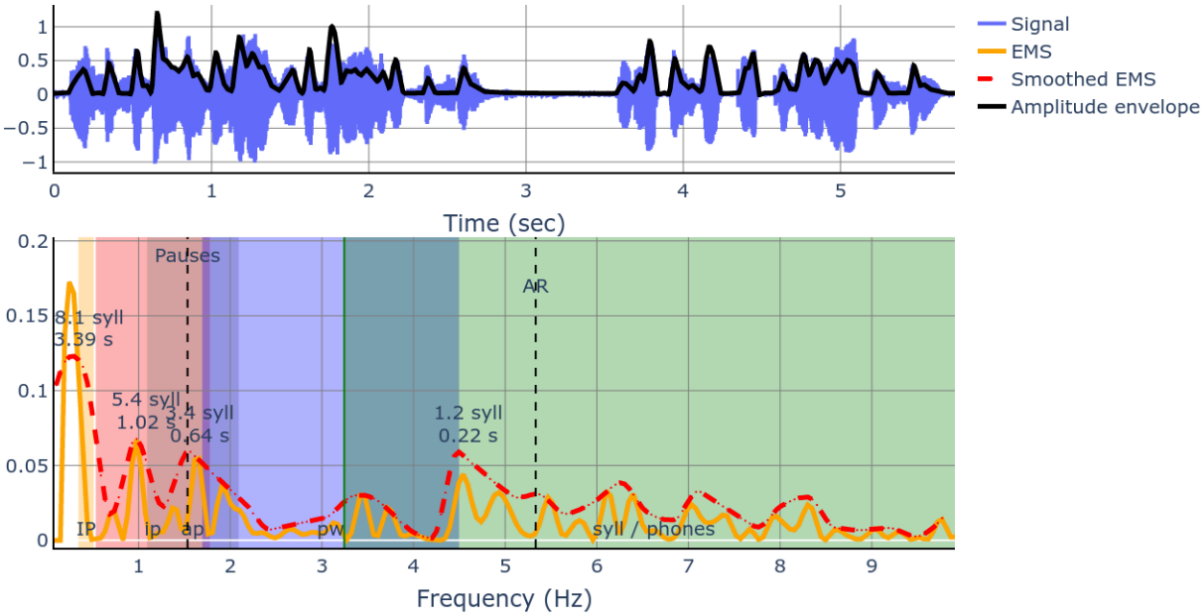


Figure 1 : EMS sur l'extrait "Monsieur Seguin n'avait jamais eu de bonheur avec ses chèvres. Il les perdait toutes de la même façon". Les intervalles correspondants aux niveaux prosodiques annotés manuellement sont indiqués en couleur : orange pour l'IP, rouge pour l'ip, gris pour l'ap, bleu pour le pw et vert pour la syllabe.

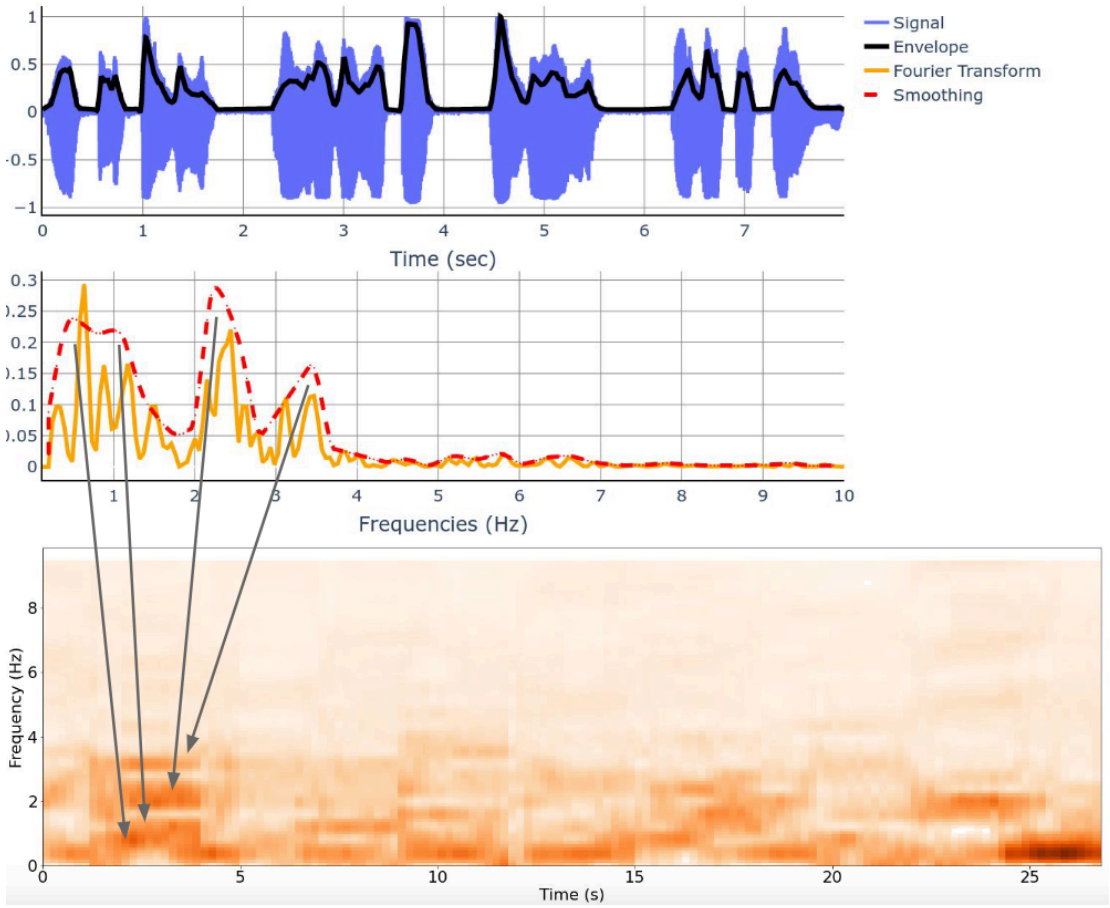


Figure 2 : Exemple de spectrogramme du rythme. La partie haute représente le signal et son enveloppe d'amplitude sur la phrase "Monsieur Seguin n'avait jamais eu de bonheur avec ses chèvres". En dessous, on retrouve le spectre de modulation d'amplitude et tout en bas le spectrogramme sur l'ensemble de la lecture. Les flèches montrent comment les pics du spectre apparaissent dans le spectrogramme.

Références bibliographiques

Arvaniti, A. 2012. The Usefulness of Metrics in the Quantification of Speech Rhythm. Journal of Phonetics, 3, 351–373.

Astésano, C. (2019) The prosodic word as the domain of French accentuation - Empirical evidence. Phonetics and Phonology in Europe, PaPE 2019, Lecce : 170-171.

Di Cristo, A. (2011). Une approche intégrative des relations de l'accentuation au phrasé prosodique du français. Journal of French Language Studies, 21(1), 73-95.

Tilsen, S. & Johnson, K. (2008). Low-frequency fourier analysis of speech rhythm. The Journal of the Acoustical Society of America, 124(2): 34–39.

Vaysse, R., Farinas, J., Astésano, C., André-Obrecht, R. (2021) Automatic Extraction of Speech Rhythm Descriptors for Speech Intelligibility Assessment in the Context of Head and Neck Cancers. Interspeech 2021, 1912-1916, https://doi.org/10.21437/Interspeech.2021-1736

Vaysse, R. (2023) Caractérisation automatique du rythme de la parole : application aux cancers des voies aéro-digestives supérieures et à la maladie de Parkinson. Sciences de l'information et de la communication. Université Paul Sabatier - Toulouse III, 21 mars 2023. https://theses.hal.science/tel-04198849.

Explication de la segmentation audio à l'aide d'un proxy et de la factorisation matricielle non négative

Théo MARIOTTE[1], Antonio ALMUDEVAR[2], Alfonso ORTEGA[2], Marie TAHON[1]

[1]LIUM, Le Mans Université, France
[2]ViVoLab, University of Zaragoza, Spain

La segmentation du signal audio est une tâche clef pour de nombreuses applications de traitement automatique de la parole (reconnaissance de la parole, segmentation et regroupement en locuteurs...). Elle consiste à détecter des segments homogènes contenant un ou plusieurs événements (Gimeno et al., 2020 ; Lebourdais et al. 2023). Dans ces travaux, nous proposons un modèle unique, que nous appellerons *teacher*, pour la détection jointe d'activité vocale (VAD), de parole superposée (OSD), de musique (MD) et de bruit (ND). L'approche proposée vise à expliquer les décisions de ce *teacher*, vu comme un modèle "boîte noire" et préalablement entraîné. Ce système est composé du modèle pré-entraîné WavLM, permettant d'extraire une séquence de caractéristiques à partir du signal audio, suivit d'un réseau convolutif temporel (TCN) permettant la modélisation de séquence. Bien que cette approche permette des performances remarquables, les décisions prises par ce système sont difficilement explicables.

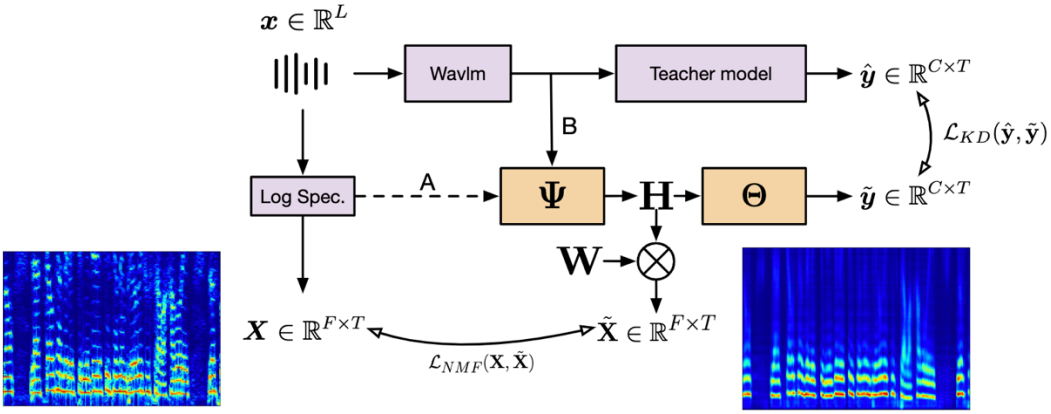


Figure1. Principe d'apprentissage du *proxy* à partir des prédictions du *teacher*.

Un second modèle, appelé *proxy*, est développé afin d'expliquer les décisions prises par le *teacher*. Ce dernier est entraîné à segmenter le signal audio à partir des distributions de sorties du *teacher* et utilise la factorisation matricielle non négative (NMF) (Lee & Seung, 2000) pour expliquer les décisions. L'architecture est inspirée des travaux (Parekh et al. 2023) et illustrée en Figure 1. Plus précisément, ce modèle est constitué de deux modules :

- Ψ: ce modèle prend en entrée une séquence de caractéristiques (WavLM ou spectrogramme) et prédit un embedding non-négatif $\mathbf{H} \in \mathbb{R}_+^{K \times T}$. T représente le nombre de trames temporelles et K un rang choisi.
- Θ: ce modèle, composé d'une couche neuronale linéaire sans biais, prédit la pseudo-probabilité de chaque classe à partir de **H**.

D'autre part, l'embedding **H** est exploité pour reconstruire le spectrogramme de puissance du signal d'entrée $X \in \mathbb{R}_+^{K \times T}$ à l'aide d'un dictionnaire NMF pré-entraîné $\mathbf{W} \in \mathbb{R}_+^{K \times T}$ tel que $\tilde{\mathbf{X}} = \mathbf{WH}$ où F représente le nombre de fréquences. L'embedding **H** est commun à la décision et à la reconstruction. Le dictionnaire **W** agit comme un décodeur entre l'espace des embeddings et le domaine des fréquences. Ces propriétés permettent de représenter les décisions prises par

le modèle dans l'espace des fréquences, permettant ainsi d'identifier les composantes fréquentielles représentatives de chaque classe.

Les résultats obtenus avec cette approche montrent d'abord que le proxy obtient des performances similaires, voire supérieures au *teacher*. De plus, le proxy permet d'expliquer les décisions prises par le modèle à deux échelles. Dans un premier temps, les explications sont extraites à l'échelle d'un segment, permettant une explication locale de la segmentation. Dans un second temps, une analyse empirique montre que l'approche proposée permet d'extraire des prototypes de classes. Ces représentations permettent d'identifier à l'échelle globale les composantes représentatives de chaque classe

Mots-clés : Segmentation audio, factorisation matricielle non négative, explication des décisions.

Références bibliographiques

Gimeno, P., Viñals, I., Ortega, A., Miguel, A. and Lleida, E. (2020) Multiclass audio segmentation based on recurrent neural networks for broadcast domain data. EURASIP Journal on Audio, Speech, and Music Processing:1–19, 2020.

Lebourdais, M., Mariotte, T., Tahon, M., Larcher, A. et al. (2023) Joint speech and overlap detection: a benchmark over multiple audio setup and speech domains. arXiv preprint arXiv:2307.13012.

Lee, D. and Seung, H. S. (2000) Algorithms for non-negative matrix factorization. Advances in neural information processing systems, 13.

Parekh, J., Parekh, S., Mozharovskyi, P. et al. (2023) Tackling interpretability in audio classification networks with non-negative matrix factorization. arXiv preprint arXiv:2305.07132.

Comment l'oreille humaine détecte-elle la somnolence ?

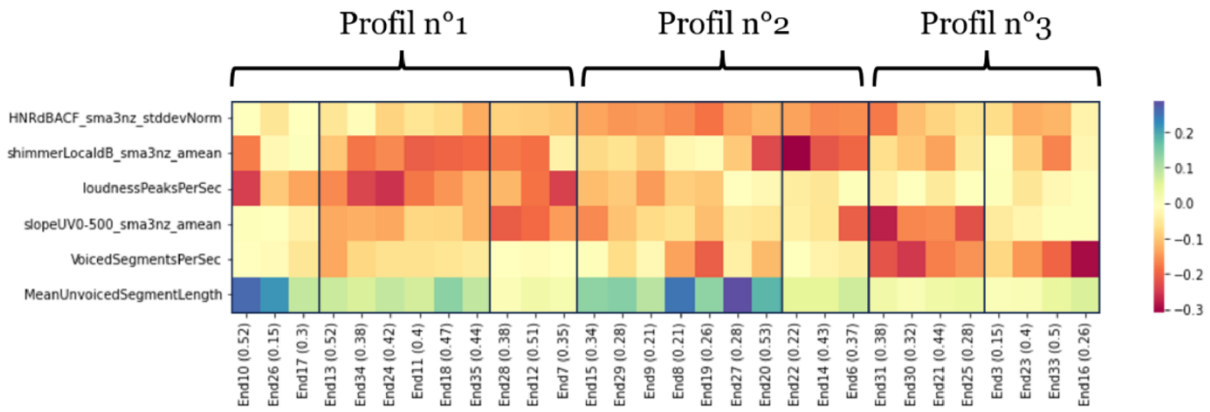
Vincent P. MARTIN[1], Nathan SALIN[2], Colleen BEAUMARD[2,3], Jean-Luc ROUAS[2]

[1] Deep Digital Phenotyping Research Unit, Department of Precision Health, Luxembourg Institute of Health, L-1415 Strassen, Luxembourg.
[2] Univ. Bordeaux, CNRS, Bordeaux INP, LaBRI, UMR 5800, F-33400 Talence, France
[3] Univ. Bordeaux, CNRS, SANPSY, UMR 6033, F-33000 Bordeaux, France

Lors d'une précédente étude perceptuelle1, nous avons demandé à 30 annotateurs naïfs d'estimer la somnolence des locuteurs du corpus SLEEP, afin d'étudier la faisabilité pour l'oreille humaine de détecter la somnolence telle qu'annotée dans le corpus (somnolence subjective, annotée avec l'échelle de somnolence de Karolinska2). Cependant, afin de minimiser le temps de passation de cette étude et maximiser le nombre de participants, nous n'avons pas collecté le retour des annotateurs sur les indices déterminants dans leur perception de la somnolence. L'objectif de ce travail est de déterminer, a posteriori, les indices sur lesquels se sont appuyés les annotateurs pour évaluer le niveau de somnolence dans l'étude perceptuelle. Pour cela, nous en étudions les poids des descripteurs audio de 30 systèmes d'estimation automatique (machine learning), chacun entraîné à reproduire les annotations d'un participant.

Nous avons utilisé les 90 échantillons de test du corpus SLEEP3 ayant été annotés dans les études perceptuelles. Les annotations sont normalisées par locuteur (z-score). Nous avons extrait les descripteurs eGEMAP des enregistrements audio grâce à la boîte à outils openSMILE4 (moyennes et écart-types des descripteurs de bas niveau seulement, n=46). Concernant l'estimation automatique, l'étude des descripteurs utilisés nécessite que notre système soit complètement explicable et transparent. Nous avons donc combiné une Analyse en Composantes Principales (ACP) et une régression linéaire, pour lesquels le produit matriciel des coefficients donne la contribution de chaque descripteur à la classification. Conformément aux précédents travaux sur le corpus SLEEP, nous avons calculé la performance des modèles avec la corrélation de Spearman entre les scores estimés par le pipeline et le label qu'il était entraîné à reproduire, au sein d'une procédure validation croisée 5-fold. Enfin, nous avons utilisé un algorithme de regroupement hiérarchique (clustering) afin d'identifier des profils d'annotateurs.

Les 30 modèles entraînés ont une performance comprise entre $\rho = 0.12$ et $\rho = 0.53$ (moyenne $\rho = 0.36$). La classification hiérarchique sur les descripteurs les plus importants tous annotateurs confondus (médiane sur les annotateurs > 5%) conduit à 3 profils, qui sont détaillés ci-dessous:



En entraînant des algorithmes de machine learning à reproduire les évaluations des annotateurs d'une étude perceptuelle, nous avons réussi à identifier les caractéristiques sur lesquelles ils se sont basés pour produire cette évaluation

Mots-clés : Somnolence, Étude perceptuelle, Production vocale, Analyse de système d'apprentissage automatique.

Références bibliographiques

Martin, V. P., Ferron, A., Rouas, J.-L. & Philip, P. "Prediction of Sleepiness Ratings from Voice by Man and Machine": a perceptual experiment replication study. in ICASSP 2023 (2023). doi:10.1109/ICASSP49357.2023.10096193.

Åkerstedt, T. & Gillberg, M. Subjective and objective sleepiness in the active individual. Int J Neurosci 52, 29–37 (1990).

Schuller, B. et al. The INTERSPEECH 2019 Computational Paralinguistics Challenge: Styrian Dialects, Continuous Sleepiness, Baby Sounds & Orca Activity. in Interspeech 2019 (2019). doi:10.21437/Interspeech.2019-1122.

Eyben, F. & Schuller, B. Opensmile. ACM SIGMultimedia Rec. 6, 4–13 (2015).

Interprétabilité pour l'identification de locuteurs. Retour sur le projet JSALT 2023

Marie TAHON[1], Imen BEN AMOR[2] , Nicolas DUGUE[1] , Jean-François BONASTRE[2,3]
[1]LIUM, [2] LIA, [3] INRIA

Lors de l'édition 2023 du workshop JSALT, nous nous sommes attaqués au sujet de l'explicabilité dans le cas de systèmes de diarization. C'est une tâche clé pour la plupart des technologies vocales telles que la transcription automatique, l'identification du locuteur, et la prédiction de dialogue. Celles-ci sont régulièrement utilisées dans des scénarios multi-locuteurs, incluant la TV/radio, les réunions ou des conversations médicales. Dans la plupart de ces domaines, la tendance actuelle pour l'IA explicable est un processus fondamental pour améliorer la transparence des décisions prises par des modèles apprentis automatiquement : l'utilisateur final, qu'il soit médecin, juge, ou data scientist, doit justifier les choix qu'il prend à partir des sorties du systèmes.

Dans ce contexte, nous avons proposé plusieurs approches pour expliquer le comportement de différents modèles de segmentation, de diarization, d'identification du locuteur ou de prédiction d'état émotionnel grâce à différentes techniques utilisées pour l'analyse d'image (integrated gradient (Suraj & Fleuret, 2019), prototypes (Li et al. 2018)), ou d'informations textuelles (projection de l'espace d'embeddings (Boluukbasi et al. 2016)). Cependant expliquer le fonctionnement d'un modèle, ou la structuration de son *espace de représentation*, ne permet pas de fournir des attributs interprétables à un utilisateur non expert car cet espace n'est pas interprétable per se. La projection dans un espace binaire a été montré efficace en terme d'explicabilité (Ben-Amor & Bonastre, 2022 ; Bonastre & Ben-Amor, 2022, Bonastre et al. 2011). Il faut donc aller plus loin et développer un alignement entre l'espace de représentation des modèles et un *espace informatif* constitué de variables explicites directement extraites de notre *espace perceptif*.

Dans ce résumé, nous présentons les travaux réalisés autour de la question de l'interprétabilité dans le cas particulier de l'identification des locuteurs. L'*espace de représentation*, ici des embeddings de locuteurs (x-vecteurs (Snyder et al., 2018)), sont projetés dans un espace positif, parcimonieux et de grande dimension, où chaque dimension est supposée interpréter la présence d'un attribut (Subramanian et al., 2018 ; Prouteau et al. 2022). Chacune de ces dimensions est rendue binaire sans que cela n'affecte de façon drastique les performances en vérification du locuteur, reconnaissance du genre, des émotions et diarization. Pour aller plus loin vers l'interprétabilité, nous avons défini un *espace informatif*, constitué de descripteurs prosodiques et acoustiques maitrisés par les experts (Eyben et al., 2016). Plusieurs méthodes statistiques ont été mises en place pour lier chacune des dimensions binaires avec ces descripteurs. Nous avons montré que les dimensions les plus importantes pour le genre, respectivement les émotions, pouvaient être associées à des familles de descripteurs liées aux formants et la fréquence fondamentale, respectivement à la prosodie. Ainsi l'approche proposée pose un premier jalon vers l'interprétabilité pour le traitement automatique de la parole en proposant une méthodologie pour lier un *espace de représentation* à un *espace informatif*.

Références bibliographiques

Ben-Amor, I. and Bonastre, J. F.. BA-LR: Binary-Attribute-based Likelihood Ratio estimation for forensic voice comparison. In nternational workshop on biometrics and forensics (IWBF) (2022)

Bolukbasi, T., et al. "Man is to computer programmer as woman is to homemaker? debiasing word embeddings." Advances in neural information processing systems 29 (2016)

Bonastre, J.F., Ben Amor, I.. BA-LR : une approche transparente de comparaison de voix en criminalistique. XXXIVe Journées d'Études sur la Parole-- JEP 2022, Nantes, France. pp.646-654 (2022)

Bonastre, J. F., Bousquet, P. M., Matrouf, D., & Anguera, X. (2011, May). Discriminant binary data representation for speaker recognition. ICASSP, Prague Czech Republic (2011). Snyder, D., Garcia-Romero, D., Sell, G., Povey, D. and Khudanpur, S. "X-Vectors: Robust DNN Embeddings for Speaker Recognition," ICASSP, Calgary, Canada (2018).

Eyben, F., Scherer, K., Schuller, B., Sundberg, J., André, E., et al.: The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. IEEE Trans. on Affect. Computing 7(2), 190–202 (2016).

Li, O., Liu, H., Chen, C., and Rudin, C.. Deep learning for case-based reasoning through prototypes: a neural network that explains its predictions. In Proceedings of the Thirty-Second AAAI, Article 432, 3530–3537 (2018)

Prouteau , T., Dugué, N., Camelin, N. and Meignier. S.. « Are Embedding Spaces Interpretable? » Results of an Intrusion Detection Evaluation on a Large French Corpus. LREC 2022, Marseille, France (2022).

Subramanian, A., et al. "Spine: Sparse interpretable neural embeddings." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 32. No. 1. (2018).

Suraj, S. and Fleuret, F.. "Full-gradient representation for neural network visualization." Advances in neural information processing systems 32(2019).

A multimodal dynamical variational autoencoder for audiovisual speech representation learning

Samir SADOK[1], Simon LEGLAIVE[1], Laurent GIRIN[2], Xavier ALAMEDA-PINEDA[3], Renaud SEGUIER[1]

[1]CentraleSupélec, IETR UMR CNRS 6164, France
[2]Univ. Grenoble Alpes, CNRS, Grenoble-INP, GIPSA-lab, France
[3]Inria, Univ. Grenoble Alpes, CNRS, LJK, France

Les données de haute dimension telles que les images naturelles ou les signaux vocaux présentent une forme de régularité qui empêche leurs dimensions de varier de manière indépendante. Cela suggère qu'il existe une représentation latente de dimension plus petite à partir de laquelle les données observées de haute dimension ont été générées. La découverte des caractéristiques explicatives cachées des données complexes est l'objectif de l'apprentissage de représentation, et les modèles génératifs de variables latentes profondes se sont révélés prometteurs en tant qu'approches non supervisées. En particulier, l'autoencodeur variationnel (VAE) (Kingma & Welling, 2014), qui est doté à la fois d'un modèle génératif et d'un modèle d'inférence, permet l'analyse, la transformation et la génération de divers types de données. Au cours des dernières années, le VAE a été étendu de nombreuses manières, notamment pour traiter des données qui sont multimodales (Wu & Goodman, 2018 ; Stutter et al. 2020) ou dynamiques (c'est-à-dire séquentielles) (Girin et al. 2020). Dans cette présentation, nous présenterons un VAE multimodal et dynamique (MDVAE) appliqué à l'apprentissage de la représentation non supervisée de la parole audiovisuelle. L'espace latent est structuré pour dissocier les facteurs dynamiques latents partagés entre les modalités (par exemple, les mouvements des lèvres de l'orateur) de ceux qui sont spécifiques à chaque modalité (par exemple, les variations de hauteur de l'orateur ou les mouvements des yeux). Une variable latente statique est également introduite pour coder l'information qui reste constante dans le temps au sein d'une séquence de parole audiovisuelle (par exemple, l'identité de l'orateur ou son état émotionnel global). Le modèle est entraîné de manière non supervisée sur un ensemble de données de parole émotionnelle audiovisuelle, en deux étapes. Dans la première étape, un VAE à vecteurs quantifiés (VQ-VAE) (Van Den Oord et al. 2017) est appris indépendamment pour chaque modalité, sans modélisation temporelle. La deuxième étape consiste à apprendre le MDVAE, dont les entrées sont les représentations intermédiaires du VQ-VAE avant la quantification. La disjonction entre l'information statique et dynamique, ainsi que l'information spécifique à la modalité par rapport à celle partagée, se produit au cours de cette deuxième étape de formation. Des résultats expérimentaux seront présentés, mettant en évidence les caractéristiques des données de parole audiovisuelle encodées dans les différents espaces latents, comment le modèle multimodal proposé peut être bénéfique par rapport à un modèle unimodal, et comment la représentation apprise peut être exploitée pour effectuer des tâches ultérieures.

Pour plus de détails, veuillez consulter le document sur MDVAE disponible à cette adresse: https://arxiv.org/abs/2305.03582

Mots-clés : Modélisation générative profonde, Apprentissage de représentations désentrelacées, Autoencodeur variationnel, Données multimodales et dynamiques, Traitement de la parole audiovisuelle.

Références bibliographiques

Girin, L., Leglaive, S., Bie, X., Diard, J., Hueber, T. and Alameda-Pineda, X. (2020). Dynamical variational autoencoders: A comprehensive review. arXiv preprint arXiv:2008.12595.

Kingma, DP. and Welling, M. (2014). Auto-encoding variational bayes. in international conference on learning representations (iclr).

Sutter, T., Daunhawer, I. and Vogt, J. (2020). Multimodal generative learning utilizing jensen-shannon-divergence. Advances in Neural Information Processing Systems, 33:6100–6110.

Van Den Oord, A., Vinyals, O. et al. (2017) Neural discrete representation learning. Advances in neural information processing systems, 30.

Wu, M. and Goodman, N. (2018) Multimodal generative models for scalable weakly-supervised learning. Advances in Neural Information Processing Systems, 31.

Interprétation d'un score d'intelligibilité dans le cadre de l'évaluation de troubles de la parole au travers d'une représentation « profonde » de la parole

Sondes ABDERRAZEK[1], Corinne FREDOUILLE[1], Alain GHIO[2], Muriel LALAIN[2], Christine MEUNIER[2], Virginie WOISARD[3]

[1]LIA, Avignon Université
[2]Aix-Marseille Univ, LPL, CNRS, Aix-en-Provence
[3]UT2J, LNPL, Toulouse Université & Toulouse Hospital

Récemment, nous avons proposé un cadre analytique général, appelé Neuro-based Concept Detector (NCD), pour interpréter les représentations profondes d'un réseau de neurones (DNN). Basé sur les schémas d'activation des neurones cachés, ce cadre met en évidence la capacité des neurones à détecter un concept spécifique lié à la tâche finale de classification dévolue au DNN. Son principal atout est de fournir un outil d'interprétabilité générique pour tout type de DNN quel que soit le domaine d'application. Dans le domaine qui nous intéresse ici – la phonétique clinique - nous avons démontré l'émergence de traits phonétiques (« concept » évoqué plus haut) dans les couches de classification d'un modèle basé sur une architecture de type CNN (Convolutional Neural Network) pour une tâche de classification des phonèmes du français sur de la parole lue et saine. Appliquée à de la parole dégradée produite par des patients atteints d'un cancer de la tête ou du cou, nous avons montré que cette structure reflète automatiquement le niveau d'altération des traits phonétiques des productions des patients.

Le travail décrit ci-dessus fait partie d'un projet à long terme2 qui visait à déterminer les unités linguistiques qui contribuent le plus au maintien ou à la perte d'intelligibilité dans les troubles de la parole (Woisard et al., 2021). Trois étapes ont été identifiées pour atteindre cet objectif : (Step 1) Modélisation des caractéristiques des unités phonémiques de la parole "normale" grâce à un CNN dédié à une tâche de classification des phonèmes du français (Abderrazek et al., 2020), (Step 2) étude des propriétés de représentation du modèle profond en termes de contenu phonétique (Abderrazek et al., 2022a, Abderrazek et al., 2022b), (Step 3) transfert de cette modélisation et de ses propriétés de représentation dans une tâche de prédiction de l'intelligibilité, typiquement dans le contexte de parole normale et dégradée, en vue d'étudier sa capacité à fournir une interprétation de la contribution des unités phonémiques dans l'intelligibilité et sa variation (amélioration ou altération). Il s'agit ici de présenter nos avancées sur la réalisation de l'étape 3 à savoir l'intégration de l'approche NCD décrite ci-dessus dans un système automatique de prédiction de l'intelligibilité. Les premiers résultats montrent une corrélation autour de 0.85 entre les scores d'intelligibilité prédits automatiquement par le système développé en étape 3 et ceux évalués perceptivement par les experts. L'utilisation conjointe des approches développées dans les étapes 1 et 2 avec le système de prédiction de l'étape 3 permet de mettre en relation un score d'intelligibilité prédit avec un niveau d'altération des traits phonétiques du français, issu de l'approche NCD, pour interprétation par des experts cliniciens.

Références bibliographiques

[2] Ce travail a été réalisé dans le cadre du projet RUGBI ("Looking for Relevant linguistic Units to improve the intelliGiBIlity measurement of speech production disorders") financé par l'Agence Nationale de la Recherche (Grant N°ANR-18CE45-0008-04).

Abderrazek, S., Fredouille, C., Ghio, A., Lalain, M., Meunier, C., Woisard, V. (2020). Towards Interpreting Deep Learning Models to Understand Loss of Speech Intelligibility in Speech Disorders Step 1 : CNN model-based phone classification. Interspeech'20, Shanghai, China.

Abderrazek, S., Fredouille, C., Ghio, A., Lalain, M., Meunier, C., Woisard, V. (2022a). Towards interpreting deep learning models to understand loss of speech intelligibility in speech disorders Step 2 : contribution of the emergence of phonetic traits. ICASSP, Singapore.

Abderrazek, S., Fredouille, C., Ghio, A., Lalain, M., Meunier, C., Woisard, V. (2022b). Validation of the Neuro-Concept Detector framework for the characterization of speech disorders: A comparative study including Dysarthria and Dysphonia. Interspeech'22, Corée.

Woisard, V. and et al. (2021). C2SI corpus : a database of speech disorder productions to assess intelligibility and quality of life in head and neck cancers. Language Resources and Evaluation, vol. 55(1), 2021.

Self-supervised learning of the relationships between speech sounds, articulatory gestures and phonetic units

Marc-Antoine GEORGES, Jean-Luc SCHWARTZ, Thomas HUEBER

Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, 38000 Grenoble, France

Learning to speak is a hard task. It involves the control of a complex motor system for uttering speech sounds from articulatory gestures which is challenging, considering the nonlinear and many-to-many nature of the relationships between articulatory configurations and speech sounds. In parallel, the child has to learn the different levels of speech and language, and in particular the phonetic level, which provide the entry to the linguistic system. Importantly, children seem to learn the relationships between speech sounds, the corresponding articulatory gestures and the phonetic units in a weakly-supervised manner. In fact, children are almost never provided with an explicit segmentation and labeling of an auditory input at the phonetic level. Similarly, apart from visual information on the facial movements of the interlocutor (if available), children have almost no access to the articulatory gestures he/she should produce to reach an acoustic target (acoustic-to-articulatory mapping). Computer-based modeling and simulation can be used to better understand the underlying mechanisms of speech learning. To study the acoustic-to-articulatory inverse mapping, a common approach is to use an articulatory synthesizer based on a physically plausible model of the vocal tract and to learn "by exploration" the relationships between the sensory space and the motor commands (e.g. Kröger et al., 2014; Philippsen, 2021; Rasilo et Räsänen, 2017). However, models are often tuned and tested on simple data, such as vowels, isolated syllables and sometimes synthetic data generated by the articulatory model itself. Another recent line of research focus on self-supervised (deep) learning (SSL) models of speech trained on massive unlabeled audio datasets (Dupoux, 2018). As an example, (Lavechin et al., 2023) partly reproduced in vitro the perceptual narrowing effect documented in infants (i.e. the progressive loss of the ability to discriminate sounds that are not relevant in the infant's language) using a SSL model trained on a set of multilingual audiobooks using contrastive predictive coding. However, most studies based on SSL speech models do not incorporate knowledge about the speech production process and therefore cannot provide much light on its implication in speech learning.

Our work is at the crossroads between these two lines of research. In (Georges et al., 2021), we proposed a first computational model of speech learning trained from raw speech in a self-supervised manner and integrating explicit knowledge on speech production. We propose here an extension of this model combining i) a pre-trained neural articulatory synthesizer able to reproduce complex speech stimuli from a limited set of interpretable articulatory parameters, ii) a RNN-based acoustic-to-articulatory inverse model, and iii) a (discrete) speech unit discovery module based on vector-quantized variational autoencoders (VQ-VAE). First, using the ABX methodology (Schatz et al., 2013) to assess which linguistic invariants are encoded by the latent dimensions of the VQ-VAE, we show the complementarity of acoustic and articulatory levels for learning discriminative representations at the phonetic level (Georges et al., 2022).

Next, we evaluate the learning abilities of the proposed computational model when it is trained to "repeat" an auditory speech input, by first identifying a set of target discrete units, then estimating from these units the target articulatory gestures, and finally generating an audio speech signal using its articulatory synthesizer. We will discuss the limitations of the proposed model and present future perspectives.

Keywords: speech production, speech acquisition, acoustic-articulatory modeling, self-supervised learning

Références bibliographiques

Dupoux, E. (2018). Cognitive science in the era of artificial intelligence: a roadmap for reverse-engineering the infant language-learner. Cognition, 173, 43–59.

Georges M-A, Diard, J., Girin, L., Schwartz J-L, Hueber, T., "Repeat after me: self-supervised learning of acoustic-to-articulatory mapping by vocal imitation", Proc. of ICASSP, pp. 8252-8256, 2022.

Georges M-A, Schwartz J-L, Hueber, T., "Self-supervised speech unit discovery from articulatory and acoustic features using VQ-VAE", Proc. of Interspeech, 2022.

Kröger, B. J., Kannampuzha, J. & Kaufmann, E. (2014). Associative learning and self-organization as basic principles for simulating speech acquisition, speech production, and speech perception. EPJ Nonlinear Biomedical Physics, 2 (1), 1-28.

Lavechin, M., (2023) Artificial neural networks to analyze and simulate language acquisition in children, PhD Thesis, Univ. PSL.

Philippsen, A. (2021). Goal-directed exploration for learning vowels and syllables: a computational model of speech acquisition. KI-Künstliche Intelligenz, 35 (1), 53-70.

Rasilo, H. & Räsänen, O. (2017). An online model for vowel imitation learning. Speech Communication, 86, 1-23.

Schatz, T. et al., “Evaluating speech features with the minimal-pair ABX task: Analysis of the classical MFC/PLP pipeline”, in Proc of. Interspeech 2013, pp. 1–5 (2013).

Le nombre de schwas détecté automatiquement est-il un indicateur de l'état de somnolence chez des patients hypersomniaques ?

Colleen BEAUMARD [1,2], Vincent P. MARTIN [1,2,3], Yaru WU [4], Jean-Luc ROUAS [1], Pierre PHILIP [2]

[1] Université de Bordeaux, CNRS, Bordeaux INP, LaBRI, UMR 5800, F-33400 Talence
[2] Université de Bordeaux, CNRS, SANPSY, UMR 6033, F-33000 Bordeaux, France
[3] Deep Digital Phenotyping Research Unit, Department of Precision Health, Luxembourg Institute of Health, Strassen, Luxembourg
[4] Université de Caen, CRISCO/UAE4255

La Somnolence Diurne Excessive impacte négativement la vie des patients. Le suivi de cette pathologie étant à la fois chronophage et coûteux pour les patients et les cliniciens, nous souhaitons développer une solution écologique du suivi des symptômes de la SDE. Pour ce faire, nous nous sommes intéressés à l'analyse de la voix car sa collecte est non-invasive et simplifiée par l'utilisation des smartphones. Dans cette étude, nous utilisons le corpus TILE1 qui contient 660 enregistrements de 132 patients hypersomniaques de la clinique du sommeil du CHU de Bordeaux. Ces derniers ont effectué un Test Itératif de Latence d'Endormissement qui consiste en 5 opportunités de sieste au cours de la journée. Ils ont été enregistrés en train de lire 5 textes différents à voix haute. Leur latence d'endormissement (le temps entre le début du test et l'endormissement) a été mesurée à chaque opportunité de sieste grâce à des mesures polysomnographiques et correspond à la somnolence physiologique. Les patients ont également rapporté leur somnolence subjective à court-terme avec le remplissage du questionnaire de somnolence de Karolinska (KSS – Karolinska Sleepiness Scale) avant chaque lecture.

Des précédentes analyses sur ce corpus ont montré des liens entre plusieurs mécanismes impliqués dans la production de la parole (acoustiques, pauses de lecture et erreurs de lecture) et la somnolence subjective et physiologique. Cependant, aucune analyse n'a porté sur le lien entre phonétique et somnolence. Nous nous sommes intéressés au schwa, voyelle instable en français, et avons émis l'hypothèse que plus les patients sont somnolents, plus ils prononcent de schwas afin de planifier la suite de leur lecture. Un lien entre le nombre de schwas annotés manuellement dans un sous-corpus de 20 locuteurs (100 échantillons) du corpus TILE et la somnolence subjective et physiologique a été observé dans une précédente étude (Beaumard et al. 2023). Nous avons étendu les phonèmes considérés à /ø/, /œ/ et à leur combinaison (notée e), et validé une procédure de détection automatique de ces phonèmes robuste à la somnolence. Celle-ci est composée d'un système de Reconnaissance Automatique de la Parole de type TDNN-HMM produisant une sortie en mots, qui sont ensuite phonétisés grâce au Lexique 3.834 qui contient une prononciation standard du français.

Étant en possession d'un système de détection de ces phonèmes fiable, nous pouvons maintenant vérifier s'il existe un lien entre la somnolence et le comportement phonologique des patients hypersomniaques dans l'entièreté du corpus TILE. Pour cela, nous avons réalisé quatre ANOVA multivariées à mesures répétées (une pour le nombre total de chaque phonème /ə/, /ø/, /œ/ ou e, détecté automatiquement par notre procédure) ayant comme facteurs explicatifs la somnolence subjective à court-terme (KSS) et la somnolence physiologique (latence d'endormissement).

Nous n'avons trouvé aucun effet significatif expliquant les variations inter-locuteurs du nombre de phonèmes automatiquement détectés. Seul un effet significatif des variations intra-locuteur du score à la KSS sur le phonème /ø/ (p=5,0e-3) et tous les phonèmes confondus e (p=4,1e-2) ont été mesurés. Il est donc possible de proposer une aide au diagnostic de la SDE avec la détection automatique des phonèmes. Des tendances statistiques ont été observées entre les variations intra- locuteurs du score à la KSS et le phonème /ə/ d'une part ; et entre la latence d'endormissement et tous les phonèmes confondus. Ces résultats exploratoires restent cependant à confirmer par des études de plus grande puissance statistique

Phonème	**/ə/**	**/ø/**	**/œ/**	**e**
KSS	.	**		*
Latence d'endormissement				.

. : $p<.1$ * : $p<.05$ ** : $p<.01$

Ces résultats confirment notre précédente hypothèse d'un lien entre somnolence et changement de comportement dans la production de phonème lors de la parole lue. Après cette étude centrée sur le nombre de chaque phonème, nous prévoyons d'analyser également leur durée ainsi que leur qualité acoustique afin de déterminer si ces informations supplémentaires vont dans le même sens que les résultats obtenus. De plus, un corpus similaire contenant des tâches de parole spontanée est en cours d'enregistrement au CHU de Bordeaux. Celui-ci nous permettra d'étendre et de comparer nos analyses avec la parole spontanée, afin de se rapprocher des conditions écologiques souhaitées pour la solution du suivi des symptômes de la SDE.

Références bibliographiques

Beaumard, C., Martin, V. P., Wu, Y., Rouas, J-L., & Philip, P. 2023. « Automatic Detection of Schwa in French Hypersomniac Patients ». Journée IA et Santé (PFIA).

Martin, V. P., Rouas, J-L., Micoulaud-Franchi, J-A., & Philip, P. 2020. « The Objective and Subjective Sleepiness Voice Corpora ». Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020).

Martin, V. P., Arnaud, B., Rouas, J-L., & Philip P. 2022. « Does Sleepiness Influence Reading Pauses in Hypersomniac Patients? » Speech Prosody 2022: 62-66

New, B., Pallier, C., Brysbaert, M., & Ferrand, L. 2004. « Lexique 2 : A New French Lexical Database ». Behavior Research Methods, Instruments, & Computers 36 (3) : 516-24

Détection et classification automatiques d'erreurs de prononciation en L2 : approche basée sur les connaissances didactiques.

Romain CONTRAIN, Julien PINQUIER, Lionel FONTAN, Isabelle FERRANE[1]

[1]IRIT, Université de Toulouse, CNRS, Toulouse INP, UT3, Toulouse, France
[2]Archean Labs, Montauban, France

Les outils d'Entraînement à la Prononciation Assisté par Ordinateur (EPAO) sont intéressants pour l'apprentissage des langues dans la mesure où la majorité des apprenants n'a pas accès à un professeur particulier pour travailler ces aspects. Ces outils doivent être capables d'effectuer la détection et le diagnostic des erreurs de prononciation avec suffisamment de fiabilité et de précision pour pouvoir fournir à l'apprenant des retours pertinents vis-à-vis des difficultés qu'il rencontre.

Dans ce domaine, nombre de travaux s'appuient sur un alignement forcé du signal de parole avec la prononciation canonique, ce qui permet d'évaluer les sons produits en connaissant les phones canoniques auxquels ils correspondent. La méthode la plus citée est le Goodness of Pronunciation (GOP) (Witt & Young, 2000), mais des méthodes plus récentes emploient des classifieurs basés sur des réseaux de neurones profonds ou des représentations issues de réseaux de neurones comme wav2vec 2.0 (Baevski et al. 2020). D'autres méthodes se basent sur une transcription phonétique suivie d'une comparaison avec la prononciation canonique, ce qui permet de fournir un diagnostic d'erreur. Dans Wu et al. (2021), les auteurs effectuent la phase de reconnaissance de phones selon deux architectures basées sur des Transformers. Le meilleur système obtient 81,3% de précision et 80,7% de rappel pour la détection et 10,0% d'erreur de diagnostic sur le corpus CU-CHLOE d'apprenants chinois de l'anglais.

Dans le cadre de nos travaux, nous nous plaçons dans le contexte d'une tâche de répétition de mots ou de phrases simples, sur la base de stimuli audio présentés aux apprenants. La prononciation correcte est connue et contient une difficulté à leur faire travailler. Les réalisations probables du phonème cible sont classées dans des catégories didactiques correspondant aux retours qu'un enseignant fournirait selon la présence et le type d'erreurs. Nous explorons deux approches qui se basent sur les deux tendances observées dans la littérature : l'une repose sur un alignement entre la prononciation cible et la production, et l'autre se base sur une transcription phonétique de la production.

La première approche réalise d'abord un alignement entre signaux avec l'algorithme Dynamic Time Warping (DTW), en utilisant des MFCC. Le segment de la production correspondant au phonème cible est ensuite classé dans l'une des catégories didactiques identifiées, en utilisant diverses mesures tirées de la littérature et un classifieur hiérarchique binaire basé sur la méthode des random forest. L'intérêt de cette approche est qu'elle est indépendante de la langue.

La seconde approche réalise une transcription phonétique de la production à l'aide d'un réseau récurrent bidirectionnel à mémoire courte et long terme (BiLSTM) (Li et al., 2020) pré-entraîné sur de la parole native de la L1 et de la L2 puis adapté à la parole d'apprenants. La transcription est ensuite alignée avec la prononciation cible via un alignement de Needleman-Wunsch (1970). La réalisation correspondant au phonème cible est alors classée dans la catégorie didactique avec laquelle elle est la plus similaire (selon des critères de similarité entre phones inspirés de Ghio et al. (2018)). Avec cette approche, l'utilisation de transfer learning permet de gérer le manque de données d'entraînement.

Notre étude a été réalisée sur un corpus de 7112 énoncés produits par 67 apprenants japonais du français et annotés au niveau phonétique par deux experts. Notre étude se limite pour l'instant aux phonèmes /ʒ/ et /y/, jugés parmi les plus importants pour l'apprentissage du français par des japonophones. Nous disposons ainsi de 1540 réalisations de /ʒ/ et de 1183 réalisations de /y/.

Pour le phonème /ʒ/, les résultats sont encourageants, avec des précisions assez élevées sur les deux catégories les plus fréquentes. Le système basé sur la transcription par un BiLSTM donne les meilleurs résultats. Nous obtenons ainsi une précision moyenne de 83,8% sur ces catégories. Pour le phonème /y/, les résultats sont moins bons, avec des précisions moins élevées sur les catégories d'erreur les plus fréquentes. Le système basé sur un alignement entre signaux et un classifieur hiérarchique binaire donne les meilleurs résultats. On détecte certes les prononciations correctes avec une précision de 96,7% mais la précision moyenne sur les catégories d'erreurs les plus fréquentes est de seulement 57,7%. Ces différences peuvent s'expliquer en partie par les différences de distribution des catégories entre les phonèmes : pour /ʒ/ le corpus compte 647 prononciations correctes et 862 représentants de la principale catégorie d'erreur, ce qui est assez équilibré. Par contre, pour /y/, il y a 869 prononciations correctes, mais seulement 151 et 87 représentants pour les deux catégories d'erreurs les plus fréquentes.

Cette première étude, son application à une paire de langues donnée (Japonais/Français) et aux phonèmes cibles faisant l'objet de difficultés typiques de cet apprentissage, a permis d'explorer deux approches potentiellement complémentaires pour la détection et de diagnostic d'erreurs de prononciation. Les résultats prometteurs vont se poursuivre par la prise en compte d'autres phonèmes cible. La méthodologie proposée peut également être généralisée à d'autres L1 ou d'autres paires de langues.

Références bibliographiques

A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2 : A framework for self-supervised learning of speech representations," in Advances in Neural Information Processing Systems, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 12 449–12 460.

A. Ghio, M. Lalain, L. Giusti, G. Pouchoulin, D. Robert, M. Rebourg, C. Fredouille, I. Laaridh, and V. Woisard, "Une mesure d'intelligibilité par décodage acoustico-phonétique de pseudo-mots dans le cas de parole atypique," in XXXIIe Journées d'Etudes sur la Parole. ISCA, 2018, pp. 285–293.

X. Li, S. Dalmia, J. Li, M. Lee, P. Littell, J. Yao, A. Anastasopoulos, D. R. Mortensen, G. Neubig, A. W. Black et al., "Universal phone recognition with a multilingual allophone system," in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020, pp. 8249–8253.

S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," Journal of Molecular Biology, vol. 48, no. 3, pp. 443–453, 1970.

S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," Speech communication, vol. 30, no. 2-3, pp. 95–108, 2000.

M. Wu, K. Li, W.-K. Leung, and H. Meng, "Transformer Based End-to-End Mispronunciation Detection and Diagnosis," in Proc. Interspeech 2021, 2021, pp. 3954–3958.

Prédiction de la compréhensibilité de la parole d'apprenants de français

Verdiana DE FINO[1,2], Isabelle FERRANE[1], Lionel FONTAN[2], Julien PINQUIER[1]

[1]IRIT, Université de Toulouse, CNRS, Toulouse INP, UT3, Toulouse, France
[2]Archean Labs, Montauban, France

Dans le domaine de l'apprentissage des langues, se faire comprendre est un objectif primordial pour un apprenant, plus important peut-être que le caractère natif de la parole (Munro & Derwing, 1995). De plus en plus d'attention s'est portée sur l'idée d'utiliser des outils automatiques pour évaluer la compréhensibilité de ce type de productions. Ces motivations viennent principalement du fait de l'aspect chronophage de l'évaluation de la compréhensibilité par les enseignants (O'Brien et al., 2018) et de l'aspect subjectif de la compréhensibilité, qui peut par exemple paraître plus importante pour un auditeur familier avec l'accent L1 (langue maternelle) Kennedy & Trofimovich, 2008). Dans leur récente étude, Saito et al. (2022) sont parvenus à prédire la compréhensibilité de la parole d'apprenants japonais d'anglais et ont atteint une corrélation $r = 0,82$ ($p < 0,001$) entre les scores prédits et les scores terrains. Les paramètres utilisés pour atteindre ces performances sont principalement liés à des paramètres segmentaux et suprasegmentaux (phonétique-phonologie, fluence, mélodie). Or, étant donné que la compréhensibilité peut être impactée par plusieurs niveaux linguistiques, tels que le niveau lexical et le niveau syntaxique (Isaacs & Trofimovich, 2012), il serait intéressant de les inclure dans le processus de prédiction.

Nous proposons ici une approche automatique visant à prédire la compréhensibilité de la parole d'apprenants du français. Notre méthode est basée sur des paramètres issus de différents niveaux linguistiques, tels que la phonétique-phonologie, mais aussi le lexique, la syntaxe et le discours. Nous définissons la compréhensibilité comme « la capacité d'un auditeur à interpréter le sens du message oral produit par un locuteur, sans tenir compte de la précision ou de la justesse phonétique ou lexicale » (Woisard et al. 2013). Nous avons réalisé un corpus contenant au total 1960 fichiers audio produits par 40 apprenants japonais et neuf apprenants allemands de français. Ces productions correspondent à des traductions, en français et à l'oral, d'énoncés écrits en L1. Elles ont été évaluées de manière subjective par 80 annotateurs en termes de compréhensibilité, sur une échelle entre 1 (compréhensibilité nulle) et 5 (compréhensibilité totale). En utilisant l'algorithme de régression Random Forest, nous avons d'excellentes performances de prédiction : une corrélation $r = 0,97$ ($p < 0,001$) et une MAE (erreur absolue moyenne) de 0,15 (± 0,12) par rapport aux annotations manuelles (vérité terrain) pour les apprenants japonais, et une corrélation $r = 0,98$ ($p < 0,001$) et une MAE de 0,18 (± 0,17) pour les apprenants allemands. Ces performances sont obtenues suite à une sélection de trois paramètres, dont un de fluence phonétique et deux lexico-grammaticaux. De plus, en entraînant notre modèle avec les données des apprenants japonais et en prédisant la compréhensibilité des apprenants allemands, la qualité des résultats reste excellente : corrélation $r = 0,97$ ($p < 0,001$) et MAE de 0,20 (± 0,21). Notre système pourrait ainsi permettre de prédire la compréhensibilité de la parole d'apprenants issus d'une autre L1, et même être généralisable à d'autres paires de langues.

Références bibliographiques

Isaacs, T. and Trofimovich, P. (2012). Deconstructing comprehensibility: Identifying the Linguistic Influences on Listeners' L2 Comprehensibility Ratings. Studies in Second Language Acquisition, 34(3).

Kennedy, S. and Trofimovich, P. (2008). Intelligibility, Comprehensibility, and Accentedness of L2 Speech: The Role of Listener Experience and Semantic Context. Canadian Modern Language Review-revue Canadienne Des Langues Vivantes, 64(3).

Munro, M. J. and Derwing, T. M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. Language learning, 45(1) :73–97.

O'Brien, M., Derwing, T., Cucchiarini, C., Hardison, D., Mixdorff, H., Thomson, R., Strik, H., Levis, J., Munro, M., Foote, J. and Levis, G. (2018). Directions for the future of technology in pronunciation research and teaching. Journal of Second Language Pronunciation.

Saito, K., Macmillan, K., Kachlicka, M., Kunihara, T. and Minematsu, N. (2022). Automated Assessment of Second Language Comprehensibility: Review, Training, Validation, and Generalization Studies. Studies in Second Language Acquisition. 45(1).

Woisard, V., Espesser, R., Ghio, A. and Duez, D. (2013). De l'intelligibilité à la compréhensibilité de la parole, quelles mesures en pratique clinique ? Revue de Laryngologie Otologie Rhinologie, 134(1) :27–33.

Utilisation d'un modèle d'apprentissage auto-supervisé wav2vec 2.0 pour automatiser la détection de la nasalité en vue de caractériser les locuteurs

Lila KIM, Cédric GENDROT

Laboratoire de Phonétique et Phonologie (CNRS & U. Sorbonne Nouvelle)

La nasalité est omniprésente dans les langues du monde, effectuée par l'abaissement du voile du palais, et crée des effets acoustiques sur les sons nasals (Maeda, 1982). Elle est utilisée pendant la production de parole pour la distinction phonologique entre nasales et orales, que ce soit pour les voyelles (un /a/ et un /ã/ par exemple) ou bien les consonnes (un /b/ et un /m/ par exemple). Si elle ne permet pas l 'opposition phonologique dans une langue, elle peut être présente grâce au phénomène de la coarticulation nasale (par exemple, "ban" en anglais).

La qualité de voix a de grandes implications dans la caractérisation du locuteur (Gold & French, 2019). Elle peut être un élément permanent de la voix d'un locuteur due à des facteurs physiologiques, mais aussi sujette à la variabilité intra-locuteur, notamment dans le style de discours ou l'émotion (Nolan, 2005). Parmi les exemples connus de la qualité de voix, les nasales offrent une caractéristique fiable pour la reconnaissance des locuteurs (Kahn, 2011) en raison de la morphologie de la cavité nasale stable et variable entre locuteurs (Dang et al., 1994 ; Serrurier, 2006).

Cependant, l'analyse acoustique de la nasalité est complexe car le couplage de deux cavités provoquent des modifications acoustiques en engendrant des pôles et zéros nasales. Bien que les méthodes d'analyse aient été entreprises pour la nasalité (Chen, 1997 ; Styler, 2017), elles sont très influencées par les caractéristiques articulatoires propres à chaque son, et à chaque locuteur.

Notre recherche se concentre sur le développement d'un système de reconnaissance automatique de la nasalité sur toutes les productions de parole en utilisant des réseaux de neurones profonds. Notre travail vise à évaluer la capacité des réseaux de neurones à détecter la nasalité (tous phonèmes confondus) en nous appuyant sur un modèle de parole auto-supervisé "wav2vec 2.0" (Baevski et al., 2020), et en validant physiologiquement. Parmi les différentes variantes du modèle disponibles, notre choix s'est porté sur "LeBenchmark" qui a été pré-entraîné sur des données en français (Parcollet, 2013). Nous avons opté pour une approche de probing dans laquelle les caractéristiques des couches intermédiaires du modèle auto-supervisé sont utilisées pour entraîner un modèle à accomplir une tâche spécifique sans nécessiter de fine-tuning (Adi et al., 2016 ; Conneau et al., 2018 ; Ma et al., 2020 ; Shah et al., 2021 ; Triantafyllopoulos et al., 2022 ; Yang et al., 2023).

Avec des représentations vectorielles extraites, notre classifieur a été entraîné, et ensuite testé sur des données acoustiques pour lesquelles une mesure physiologique a été effectuée en guise de référence (Elmerich et al., 2023 ; Kim et al., 2023). Les résultats ont été encourageants, avec un taux d'exactitude global atteignant 89,92 % et des variations intra- et inter-locuteurs observées. Les données aérodynamiques ont été utilisées comme des indications sur la réalisation de la nasalité, et l'analyse de ces données a permis de fournir des explications approfondies concernant les classifications réalisées par le modèle.

Mots-Clés: nasalité, wav2vec 2.0, deep learning, caractérisation du locuteur, variabilité inter-locuteur.

Références bibliographiques

Adi, Y., Kermany, E., Belinkov, Y., Lavi, O., & Goldberg, Y. (2016). Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. arXiv preprint arXiv:1608.04207.

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in neural information processing systems, 33, 12449-12460.

Chen, M. Y. (1997). Acoustic correlates of English and French nasalized vowels. The Journal of the Acoustical Society of America, 102(4), 2360-2370.

Conneau, A., Kruszewski, G., Lample, G., Barrault, L., & Baroni, M. (2018). What you can cram into a single vector: Probing sentence embeddings for linguistic properties. arXiv preprint arXiv:1805.01070

Dang, J., Honda, K., & Suzuki, H. (1994). Morphological and acoustical analysis of the nasal and the paranasal cavities. The Journal of the Acoustical Society of America, 96(4), 2088-2100.

Gold, E., & French, P. (2019). International practices in forensic speaker comparisons: second survey. International Journal of Speech, Language and the Law, 26(1), 1-20.

Elmerich A., Kim L., Gendrot C., Amelot A., Crevier-Buchman L. & Maeda S. Nasality detection from acoustic data with a convolutional neural network and comparison with aerodynamic data. In Proceedings of the 20th International Congress of Phonetic Sciences, 1356-1360

Kahn, J. (2011, December). Parole de locuteur: performance et confiance en identification biométrique vocale. Avignon.

Kim L., Gendrot C., Elmerich A., Amelot A. & Maeda S. Détection de la nasalité du locuteur à partir de réseaux de neurones convolutifs et validation par des données aérodynamiques. In 18e Conférence en Recherche d'Information et Applications \\ 16e Rencontres Jeunes Chercheurs en RI \\ 30e Conférence sur le Traitement Automatique des Langues Naturelles \\ 25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues, pages 101–108. ATALA, 2023

Ma, D., Ryant, N., & Liberman, M. (2021, June). Probing acoustic representations for phonetic properties. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 311-315). IEEE.

Maeda, S. (1982). Acoustiic cues of vowel nasalization: a simulation study. Journal of Acoustical Society of America, 7(S1), S102.

Nolan, F. (2007). Voice quality and forensic speaker identification. Govor, 24(2), 111-128.

Parcollet, T., Nguyen, H., Evain, S., Boito, M. Z., Pupier, A., Mdhaffar, S., ... & Besacier, L. (2023). LeBenchmark 2.0: a Standardized, Replicable and Enhanced Framework for Self-supervised Representations of French Speech. arXiv preprint arXiv:2309.05472.

Serrurier, A. (2006). Modélisation tridimensionnelle des organes de la parole à partir d'images IRM pour la production de nasales-Caractérisation articulatori-acoustique des mouvements du voile du palais (Doctoral dissertation, Institut National Polytechnique de Grenoble-INPG).

Shah, J., Singla, Y. K., Chen, C., & Shah, R. R. (2021). What all do audio transformer models hear? probing acoustic representations for language delivery and its structure. arXiv preprint arXiv:2101.00387.

Styler, W. (2017). On the acoustical features of vowel nasality in English and French. The Journal of the Acoustical Society of America, 142(4), 2469-2482.

Triantafyllopoulos, A., Wagner, J., Wierstorf, H., Schmitt, M., Reichel, U., Eyben, F., ... & Schuller, B. W. (2022). Probing speech emotion recognition transformers for linguistic knowledge. arXiv preprint arXiv:2204.00400.

Yang, M., Shekar, R. C., Kang, O., & Hansen, J. H. (2023). What Can an Accent Identifier Learn? Probing Phonetic and Prosodic Information in a Wav2vec2-based Accent Identification Model. arXiv preprint arXiv:2306.06524.

A closer look at latent representations of end-to-end TTS models

Martin LENGLET, Olivier PERROTIN Gérard BAILLY

Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, F-38000 Grenoble, France

In recent years, deep neural architectures have displayed groundbreaking performances in various speech processing areas, including Text-To-Speech (TTS), but at the expense of interpretability of computed intermediate representations. However, statistical learning performed by these neural models constitutes a valuable source of information about speech and language.

The present study aims at developing statistical tools to narrow the gap between these new processing techniques and speech sciences. By identifying phonetic and acoustic features in model representations, the proposed methods help understanding how neural TTS are able to organise speech information on an unsupervised manner and provide new insights on phonetic regularities captured by statistical learning on massive data.

We introduce a methodology for the analysis of any phonetic or acoustic feature in any intermediate representations of state-of-the-art sequence-to-sequence TTS models: Tacotron2 and FastSpeech2, without the need for additional data and training process. In particular, we show that acoustic features measured on the output synthetic speech can be approximated by multi-linear predictors from the output of any layer of these models. The direction of variation of each acoustic feature in an intermediate representation is given by the regression coefficients. Analysis of the goodness of fit of the multi-linear regression (R2) for each model, each intermediate layer and each acoustic feature first demonstrated that segmental acoustic features (formant frequencies, spectral tilt, centre of gravity) are gradually encoded throughout both models, with the highest fit at the end of the decoder. This shows that segmental features are not completely encoded in the text encoder and that the decoder is needed to complement this information, likely modelling the co-articulations factors. The gradual encoding of segmental features also highlights the early computation of phonetic representations by the models. This hypothesis was confirmed by the adaptation of the proposed method to a linear phoneme classification task from the output of each layer. Supra-segmental features (fundamental frequency, duration, energy) on the other hand are mostly encoded at the output of the text encoder. The fundamental frequency and energy predictors natively implemented in FastSpeech2 constrain this behaviour, whereas Tacotron2 linearly encodes these features by default.

The identification of intermediate layers that display the best linear representation of acoustic features opens the route toward designing more careful control architectures for neural TTS. As an example, we showed how explicit biases can be inferred from the direction of variation of each acoustic feature calculated in intermediate representations, and added with a controllable gain to those representations to vary the corresponding acoustic feature value. This control mechanism was evaluated for various levels of internal representations, and we reached highly accurate control of acoustic features on the intermediate layers that displayed the highest regression goodness of fit. The localisation of phonetic representations in the model also allows for discrete control of phonological processes such as French liaisons and pauses. The combined control of continuous prosodic features and discrete representations was evaluated through listening test, which showed the benefits of the proposed embedding bias method to manipulate the speaking rate.

Overall, the proposed analysis highlighted how acoustic and phonetic features were linearly encoded into intermediate latent representations. The proposed methodology can be applied to any encoder-decoder architectures, as well as any acoustic parameters, either continuous or categorical, without the need for additional data and training process, and paves the way towards more controllable speech generation systems.

Keywords: Speech Synthesis, Representation learning, Language Modelling, Phonetic Analysis, Prosodic Control

Investigating the dynamics of hand and lips in French Cued Speech using attention mechanisms and CTC-based decoding

Sanjana SANKAR, Denis BEAUTEMPS, Frederic ELISEI, Olivier PERROTIN, Thomas HUEBER

Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, 38000 Grenoble, France

Hard of hearing or profoundly deaf people make use of cued speech (CS) as a communication tool to understand spoken language. By delivering cues that are relevant to the phonetic information, CS offers a way to enhance lipreading. In literature, there have been several studies on the dynamics between the hand and the lips in the context of human production. This article proposes a way to investigate how a neural network learns this relation for a single speaker while performing a recognition task using attention mechanisms. Further, an analysis of the learnt dynamics is utilized to establish the relationship between the two modalities and extract automatic segments. For the purpose of this study, a new dataset has been recorded for French CS. Along with the release of this dataset, a benchmark will be reported for word-level recognition, a novelty in the automatic recognition of French CS.

Keywords: cued speech, hearing impaired, assistive technology, explainability, corpus, multimodality

Comprendre les phénomènes permettant la gestion des tours de parole dans les contenus de médias audiovisuels

Rémi URO[1,2,3], Marie TAHON[3], David DOUKHAN[1], Albert RILLIARD[2,4]

[1] Institut National de l'Audiovisuel
[2] Laboratoire Interdisciplinaire des Sciences du Numérique
[3] Laboratoire d'Informatique de l'Université du Mans
[4] Universidade Federal do Rio de Janeiro

La gestion des interactions soulève des questions liées aux incivilités. Étudier ces phénomènes est utile pour mieux comprendre le rôle de différentes caractéristiques sociales (genre, position sociale, etc.) lors des interactions, particulièrement pour le cas des médias audiovisuels qui jouent un rôle descriptif comme prescriptif (Coulomb-Gully, 2011). Les interactions parlées constituent des objets complexes ; les décrire de manière quantitative nécessite une formalisation des objets d'étude et des outils de traitement automatique adaptés. Le projet ANR « Gender Equality Monitor » vise à fournir de tels outils, permettant par exemple l'analyse du temps de parole et d'apparition à l'écran des femmes et des hommes dans les médias (Doukhan et al., 2018, 2019 ; Uro & Doukhan, 2020). Dans ce cadre, je m'intéresse à la question des interruptions (voir le concept de manterrupting (Bennett, 2015)). À cette fin il est nécessaire d'analyser la dynamique des tours de parole, notamment les aspects liés à la terminalité d'un énoncé – c'est-à-dire déterminer si la personne produit un tour complet d'un point de vue sémantique et pragmatique, et pas seulement une fin de phrase (Levinson, 1983). L'approche choisie est de considérer que si des interlocuteur·ice·s humain·es sont capables de planifier leur prise de parole avant même que le tour courant soit terminé (Grosjean, 1996 ; Magyari & De Ruuiter, 2012), alors il est possible pour un système automatique d'effectuer de telles prédictions. Une expérience perceptive visant à déterminer les indices permettant aux humain·es de prévoir la terminalité d'un tour de parole a été menée et met en avant l'importance des informations prosodiques dans la gestion des tours de parole.

L'une des approches mises en œuvre pour aborder cette question de façon automatique est l'utilisation du modèle de Voice Activity Projection développé par (Skantze, 2017), permettant une prédiction de l'activité vocale d'un·e locuteur·ice dans le futur. Une adaptation de ce modèle basée sur des réseaux de neurones récurrents (LSTM), utilisant en entrée des paramètres prosodiques extraits par eGemaps (Eyben et al. 2015) afin de prédire la probabilité que le·a locuteur·ice courant·e continue son tour de parole dans les secondes suivantes, a été réalisée et entraînée sur le corpus Switch-board (Calhoun et al. 2010). Si les paramètres prosodiques d'entrée peuvent être interprétés, il n'est pas évident de savoir lesquels de ces paramètres sont les plus discriminants pour effectuer la prise de décision. Expliquer un tel modèle fournirait des informations précieuses pour mieux comprendre la façon dont les humain·es prévoient la gestion des tours de parole.

La méthode envisagée pour expliquer ce modèle s'inspire de (Wanying et al. 2021) et vise à interpréter les valeurs de Shapley issues d'une analyse DeepSHAP (Lundberg & Lee, 2017) afin de mettre en avant les paramètres prosodiques les plus influents sur la décision du modèle. Cela donnera un aperçu des phénomènes qui jouent un rôle dans la prédiction de la durée d'un tour de parole.

De plus, un corpus de parole issue d'émissions de télévision et de radio, annoté en terminalité, a été constitué ; il permettra d'entraîner ce modèle sur des données contenant de la parole spontanée correspondant à notre objet d'étude, qui présente des dynamiques différentes des interactions présentes dans les échanges téléphoniques disponibles dans le corpus Switchboard. Effectuer cette analyse SHAP sur un modèle entraîné avec ces données

mettra en lumière les différences entre les informations pertinentes pour effectuer cette tâche en comparant les modèles entraînés sur ces deux corpus.

Mots-clés : réseaux de neurones, Turn-taking, Incivilités, Interruptions, TRP, Médias, SHAP, Prosodie

Références bibliographiques

Jessica Bennett. How Not to Be 'Manterrupted' in Meetings — time.com. https ://time.com/3666135/sheryl-sandberg-talking-while-female-manterruptions/, 2015. [Accessed 26-Sep-2022].

Sasha Calhoun, Jean Carletta, Jason M. Brenier, Neil Mayo, Dan Jurafsky, Mark Steedman, and David Beaver. The nxt-format switchboard corpus : a rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue. Language Resources and Evaluation, 44(4) :387–419, Dec 2010.

Marlène Coulomb-Gully. Genre et médias : vers un état des lieux. Sciences de la société, (8383) :3–13, Nov 2011.

David Doukhan, Géraldine Poels, Zohra Rezgui, and Jean Carrive. Describing gender equality in french audiovisual streams with a deep learning approach. VIEW Journal of European Television History and Culture, 7(14) :103–122, 2018.

David Doukhan, Zohra Rezgui, Géraldine Poels, and Jean Carrive. Estimer automatiquement les différences de représentation existant entre les femmes et les hommes dans les médias. page 3, 2019.

Florian Eyben, Klaus Scherer, Björn Schuller, Johan Sundberg, Elisabeth Andre, Carlos Busso, Laurence Devillers, Julien Epps, Petri Laukka, Shrikanth Narayanan, and Khiet Truong. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. IEEE Transactions on Affective Computing, 7 :1–1, Jan 2015.

Wanying Ge, Jose Patino, Massimiliano Todisco, and Nicholas Evans. Explaining deep learning models for spoofing and deepfake detection with shapley additive explanations. (arXiv :2110.03309), Oct 2021.

Francois Grosjean. Using prosody to predict the end of sentences in english and french : Normal and brain-damaged subjects. Language and Cognitive Processes, 11(1–2) :107–134, Apr 1996.

Stephen C. Levinson. Pragmatics. Cambridge [England] ; New York : Cambridge University Press, 1983.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. CoRR, abs/1705.07874, 2017.

Lilla Magyari and Jan de Ruiter. Prediction of turn-ends based on anticipation of upcoming words. Frontiers in Psychology, 3, 2012.

Gabriel Skantze. Towards a general, continuous model of turn-taking in spoken dialogue using lstm recurrent neural networks. In Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue, page 220–230, Saarbrücken, Germany, 2017. Association for Computational Linguistics.

Rémi Uro and David Doukhan. Étude ina. Pendant le confinement, le temps de parole des femmes a baissé à la télévision et à la radio, 2020.