Modélisation de la Parole avec Tacotron2 : Analyse acoustique et phonétique des plongements de caractère

Martin Lenglet, Olivier Perrotin, Gérard Bailly

Modélisation de la Parole avec Tacotron2 : Analyse acoustique et phonétique des plongements de caractère

Martin Lenglet Olivier Perrotin Gérard Bailly
Univ. Grenoble Alpes, CNRS, Grenoble-INP, GIPSA-lab, Grenoble, France
martin.lenglet,olivier.perrotin,gerard.bailly@grenoble-inp.fr

RÉSUMÉ

Les réseaux de neurones bouleversent depuis plusieurs années les applications de traitement automatique de la parole. Cependant, le bon de performances rendu possible par ces technologies se fait généralement au détriment de la compréhensibilité et de l'interprétabilité de ces nouveaux modèles. Pourtant, l'apprentissage statistique, au coeur de ces nouveaux usages, constitue une source potentielle d'informations importante sur le langage, à condition de réussir à identifier et localiser ces paramètres dans des réseaux de plusieurs millions de neurones. Ce papier propose une étude des plongements internes d'un modèle de synthèse vocale de type Tacotron2 entraîné sur le Français. Cette analyse suggère que le réseau apprend à représenter sa séquence d'entrée en une suite de cibles acoustiques et phonétiques, dépendantes de leur contexte. La mise en évidence de l'encodage de ces paramètres permet d'imaginer leur contrôle de manière plus naturelle.

ABSTRACT

Language Modeling with Tacotron2 : a Phonetic and Acoustic Analysis of Text Embeddings

In recent years, deep learning breakthroughs met a huge success in automatic speech processing. However, the leap forward in performances is accompanied by a lack of interpretability and understandability of these new models. Nevertheless, statistical learning constitutes a valuable source of information about language if analyzed with the right tools and methodology. This paper presents a study of text embeddings computed by a state of the art synthesis model Tacotron2 trained on French data. Our analysis shows that this network is able to compute an in-context sequence of acoustic and phonetic targets from the given input sequence. Identification and localization of these acoustic parameters would enable a more natural control over the synthesis.

MOTS-CLÉS : Synthèse de parole, réseau de neurones, plongement, réseau avec attention, transcription phonétique.

KEYWORDS: Speech synthesis, neural network, embeddings, attention network, phonetic transcription.

1 Introduction

Ces dernières années, les modèles d'apprentissage profond ont révolutionné l'approche du traitement automatique de la parole. En particulier, la synthèse de parole, portée par des modèles tels que Tacotron2 (Shen *et al.*, 2018) ou FastSpeech2 (Ren *et al.*, 2020), atteint une qualité proche de la voix naturelle. Tous ces modèles ont une architecture commune : un encodeur convertit la séquence d'entrée (orthographique et/ou phonétique) en une représentation abstraite, qu'un décodeur lit pour générer la séquence de trames de mel-spectrogramme correspondante.

Différentes stratégies de représentation des éléments de la séquence d'entrée sont proposées dans la littérature. Tacotron2 (Shen *et al.*, 2018) associe par exemple 3 couches de convolutions d'empan limité (2 caractères de chaque côté) à une unité Long-Short-Term Memory (LSTM) bidirectionnelle. Dans Fastspeech2 (Ren *et al.*, 2020), la couche récurrente LSTM est remplacée par plusieurs couches de self-attention. Dans les deux cas, la volonté est de mettre en contexte ces plongements afin de préparer la génération du signal audio associé.

Cependant, l'apprentissage automatique des milliers de paramètres qui constituent ces couches de neurones empilées, sans autres contraintes que la structure imposée au réseau, complique la lecture des paramètres encodés dans ces représentations latentes. Dans le même temps, l'apprentissage statistique est la force de ces modèles, qui parviennent à extraire les invariants dans les données d'apprentissage, ainsi que les covariations entre ces paramètres. Identifier et localiser ces paramètres permettrait tout d'abord de développer une méthodologie d'analyse linguistique de la parole d'un ou plusieurs locuteurs basée sur un apprentissage statistique sur de larges corpora. L'analyse des représentations latentes apprises par le modèle pourrait également permettre de gagner en contrôle sur la synthèse, en biaisant le modèle tout en respectant les covariations apprises entre les paramètres acoustiques d'intérêt. Cet article présente notre analyse linguistique de l'espace latent en sortie de l'encodeur d'un modèle Tacotron2. La section 2 présente le contexte de l'étude proposée. L'ensemble des méthodes nécessaires à l'analyse de l'espace latent sont décrites en section 3. L'analyse des représentations phonétiques des entrées orthographiques apprises par le modèle est détaillée en section 4.2. La section 4.3 développe l'analyse acoustique des paramètres encodés dans les représentations de la séquence d'entrée donnée au modèle.

2 Visualisation des plongements dans la littérature

La question de l'information encodée dans les plongements appris par un réseau d'apprentissage profond et la visualisation de ces paramètres est au coeur de la volonté naissante d'explicabilité de ces nouvelles méthodes de traitement de l'information (Burkart & Huber, 2021). L'exploration manuelle de l'espace latent en sortie d'un encodeur de type Tacotron (Wang *et al.*, 2017) entraîné sur des entrée orthographiques a montré une tendance à la représentation phonétique de ces entrées (Perquin *et al.*, 2020). Les auteurs montrent par visualisation t-SNE que les plongements se rassemblent par groupes de phonèmes, chaque groupe rassemblant des plongements d'un ou plusieurs graphèmes, qui dans leur contexte correspondent à ce phonème commun.

A l'échelle de la phrase complète, l'analyse des plongements de style ou de locuteur, massivement utilisés en synthèse expressive multi-locuteurs, s'inscrit dans cette même démarche. La visualisation par UMAP des plongements de style appris par encodeur de référence (Skerry-Ryan *et al.*, 2018) combiné à un modèle DCTTS (Tachibana *et al.*, 2018) a mis en évidence les corrélations entre les dimensions de l'espace réduit et certains paramètres acoustiques d'intérêt pour le contrôle expressif (Tits *et al.*, 2019). De même, l'observation des plongements de locuteur (Hsu *et al.*, 2019) révèle une organisation de l'espace latent par proximité vocale entre les locuteurs, séparant notamment les voix masculines des voix féminines.

Bien que ces travaux mettent en lumière l'information acoustique et phonétique potentiellement encodée dans l'espace latent des plongements, les méthodes non-linéaires de transformation de l'espace utilisées ne permettent pas de localiser les paramètres acoustiques dans l'espace latent initialement construit par le modèle. Nous souhaitons ainsi prolonger cette analyse pour transformer ces espaces latents en espaces de contrôle opérationnels pour la synthèse.

3 Méthodes proposées

Cette section décrit les méthodes d'analyse de l'espace latent en sortie de l'encodeur Tacotron2 utilisées dans ce papier. Dans l'objectif de vérifier que les entrées orthographiques sont interprétées de façon phonétique par l'encodeur, l'ensemble des méthodes décrites dans cette section sont appliquées sur des séquences d'entrées orthographiques.

3.1 Lecture de la carte d'attention

Dans Tacotron2 (Shen *et al.*, 2018), un réseau avec attention (Bahdanau *et al.*, 2014) réalise l'interface entre l'encodeur et le décodeur. Pour chaque trame de mel-spectrogramme générée de façon auto-regressive par le décodeur, cette couche d'attention calcule le poids assigné à chaque élément de la séquence d'entrée. Ces poids, appelés poids d'attention, indiquent l'importance relative des éléments de la séquence d'entrée pour la trame à générer. Nous faisons donc l'hypothèse que la lecture de ces poids permet de réaliser la segmentation automatique du spectrogramme en sortie en fonction des graphèmes donnés en consigne. Notre méthode de lecture de la carte d'attention s'applique à tous les caractères de la séquence d'entrée selon la procédure ci-dessous :

1. Vérification que le poids d'activation maximum de ce caractère sur la séquence dépasse un seuil fixé à 0.35. Ce seuil permet d'exclure les caractères qui ne sont pas suffisamment représentés dans le vecteur de contexte calculé par la couche d'attention. Ces caractères sont qualifiés de "muets" dans la suite de cet article.
2. Sur la trame pendant laquelle ce maximum est atteint, on vérifie que le poids d'attention maximum est sur le caractère en question. Sinon, on considère également ce caractère comme muet.
3. Si le poids est maximum sur le caractère, on considère que l'attention de la trame se porte sur ce caractère, de même pour toutes les trames adjacentes pour lesquelles le poids d'attention est maximum sur ce caractère.

Cette méthode permet d'estimer la durée d'attention de chaque caractère, ainsi que les trames pendant lesquelles ce caractère a été généré. Les performances de cette méthode sont évaluées en section 4.3.1.

3.2 Identification des schémas d'activation des graphèmes

La lecture des cartes d'attention de Tacotron2 a montré que tous les graphèmes n'étaient pas ciblés par l'attention. En effet, la correspondance graphème-phonème n'est généralement pas de 1 pour 1. Dans les langues à orthographe opaque comme le français, la prononciation correcte d'une lettre nécessite la connaissance d'un large empan de lettres (Bosse & Valdois, 2009). Un phonème est donc souvent associé à plusieurs graphèmes ; ce phonème est alors appelé phonème complexe. Dans ce cas, on note que l'attention du modèle se porte sur un seul graphème, selon un schéma qui dépend du contexte.

Les schémas les plus courants sont résumés en figure 1, et les règles de correspondance en table 1. Aux règles de la table 1 s'ajoute la représentation des diphtongues par l'association des 2 ou 3 phonèmes qui la constitue, séparés par '&' ("x" => /k&s/ par exemple). Ces règles permettent d'identifier les graphèmes porteurs de l'information phonétique (/_/ dans le cas des graphèmes muets), et sont utilisées pour explorer l'espace latent en sortie de l'encodeur comme décrit en section 4.2.

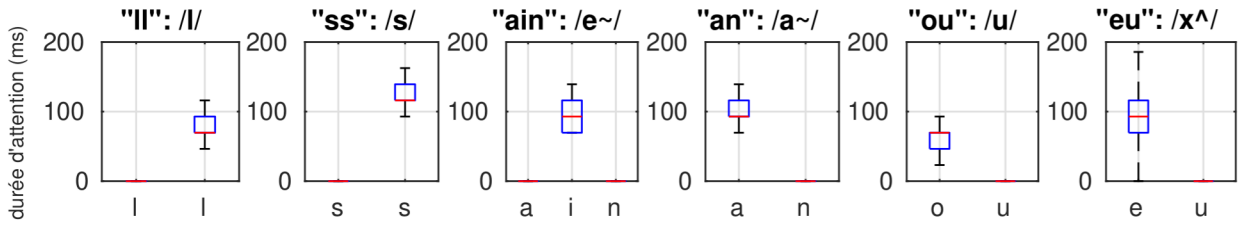


FIGURE 1 – Schémas des durées d'activation (en ms) de 6 phonèmes complexes.

Schémas	Activation	Exemples
C C	_ C	"nn", "ll", "ss"
V V	V _	"an", "ou", "au"
V V V	_ V _	"eau", "ain"

TABLE 1 – Règles d'activation des graphèmes. C et V représentent un caractère dans un phonème consonne et voyelle respectivement. _ représente un caractère muet.

4 Expériences et Résultats

4.1 Modèle et données

Le modèle Tacotron2 utilisé dans cette étude diffère légèrement de l'implémentation partagée par NVIDIA [1]. Comme (Lenglet *et al.*, 2021), notre modèle implémente la correction de Gate Loss, ainsi que la possibilité de présenter des entrées orthographiques et/ou phonétiques. Le décodeur génère 2 trames de spectrogramme à la fois. Nous avons observé que générer 2 trames à la fois permettait d'accélérer l'inférence, sans détériorer les performances du modèle. De plus, une tâche parallèle de prédiction phonétique à partir de la sortie de l'encodeur est mise en place pour les entrées orthographiques. Après mise en contexte par l'encodeur, une couche de projection linéaire associée à une fonction softmax réalise la classification des plongements orthographiques parmi l'ensemble des phones décrit par (Bailly *et al.*, 2021) [2], auquel s'ajoutent les diphtongues, pour un total de 63 phones. Pour établir une classification par graphème, le corpus a été annoté phonétiquement en suivant les règles d'activation observées en section 3.2. L'erreur d'entropie croisée de cette couche de classification est ajouté à l'erreur globale du modèle avant la rétro-propagation du gradient. Cette tâche parallèle permet d'aider à structurer l'espace latent lors de l'apprentissage et est évaluée en section 4.2.2.

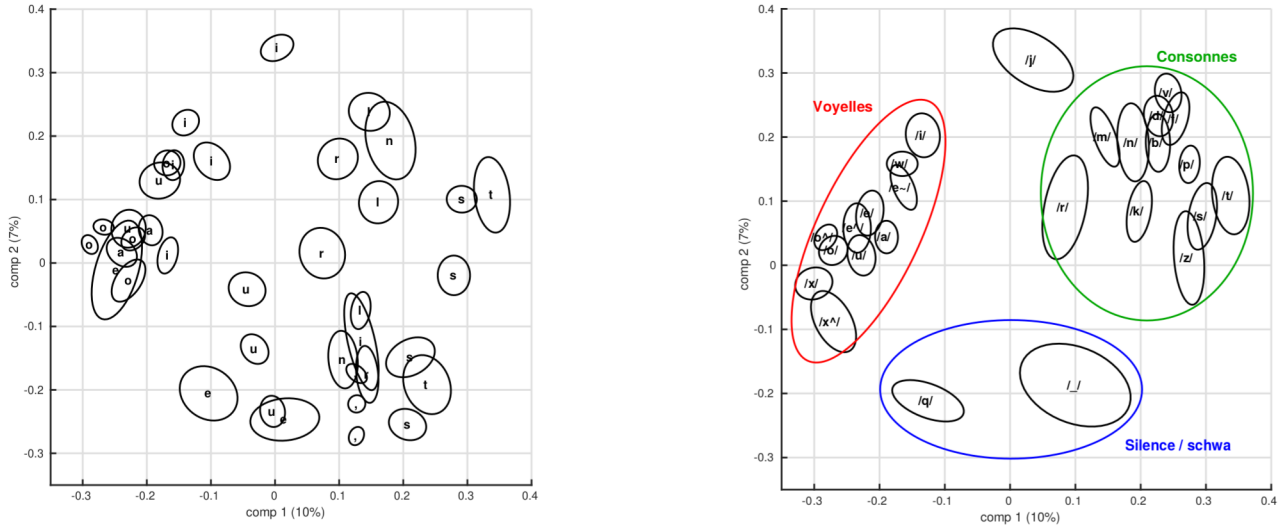
Ce modèle est entraîné sur la nouvelle segmentation d'une partie du corpus français M-AILABS proposée par (Bailly *et al.*, 2021). Nous avons sélectionné un ensemble de 29557 phrases tirées de 4 romans lus par Nadine Eckert-Boulet. Toutes les phrases sélectionnées sont présentes sous forme orthographique et phonétique. 5% de ce corpus est mis de côté pour le test, soit 1477 phrases. Le modèle est entraîné sur les 2 types d'entrées jusqu'à stabilisation de l'erreur sur la base de test, ce qui représente environ 100 époques. Pour les sections 4.2 et 4.3, les 1477 phrases du corpus de test sont ensuite synthétisées en utilisant l'entrée orthographique. Les plongements des graphèmes d'entrée, mis en contexte par l'encodeur sont enregistrés en parallèle de la synthèse, ainsi que la prédiction phonétique associée. Le vocodeur utilisé est Waveglow (Prenger *et al.*, 2019).

1. https ://github.com/NVIDIA/tacotron2
2. https ://zenodo.org/record/4580406#%23.YI_qIyaxXmE

4.2 Représentation phonétique des plongements de caractère

4.2.1 Visualisation de l'espace phonétique

Afin de visualiser l'espace latent, la distance cosinus est calculée entre toutes les paires de plongements du corpus de test. La matrice de distances ainsi obtenue est projetée en dimensions réduites par mise à l'échelle multidimensionnelle (MDS) (Kruskal, 1978). Les 2 premières composantes de cette projection sont données en figure 2, en considérant les plongements par leur graphème d'origine (2a) ou leur correspondance phonétique (2b), obtenue par la méthode décrite en section 3.2.



(a) Par graphème (15 graphèmes plus fréquents + ",") (b) Par phone (25 phones plus fréquents + /_/)

FIGURE 2 – MDS des plongements du corpus de test. Les ellipses montrent l'étalement de chaque groupe selon les 2 demi-axes principaux, avec une amplitude d'un écart-type.

Plusieurs groupes apparaissent pour chaque graphème, confirmant les résultats de (Perquin *et al.*, 2020). En revanche, la représentation par phonème fait apparaître des groupes uniques et distincts. De plus, on retrouve une proximité entre les plongements correspondant à des phonèmes de même type : les phonèmes de voyelles se distinguent des consonnes et des silences. Au sein même des consonnes, une séparation apparaît entre les consonnes plosives, nasales et fricatives. Le groupe des phonèmes silencieux /_/ regroupe la plus grande variété de graphèmes différents : tous les graphèmes muets tels que décrits en section 3.2, ainsi que les ponctuations et les espaces.

Ces observations suggèrent que l'encodeur de Tacotron2 apprend à représenter dans un espace phonétique les plongements de la séquence d'entrée, même si cette dernière est entièrement orthographique. Les ellipses de dispersion des groupes phonétiques suggèrent que le modèle est capable de générer des variations pour chaque phonème, basées sur le contexte dans lequel celui-ci se trouve. Ces ellipses sont plus restreintes pour les consonnes que les voyelles, ce qui est cohérent avec le phénomène de coarticulation qui impacte davantage les voyelles que les consonnes (Modarresi *et al.*, 2004). La structuration de cet espace latent relève donc d'un compromis entre 1) la distinction à faire entre les phonèmes en vue de leur prononciation et 2) laisser suffisamment de marge à chaque phonème pour permettre une coarticulation naturelle. L'évaluation du rapport entre les distances intra-classes et inter-classes, menée en amont de cette étude, montre que l'apprentissage conjoint des entrées orthographiques et phonétiques, ainsi que la tâche de prédiction phonétique à partir des entrées orthographiques, favorisent la désintrication des ellipses phonétiques. La proximité entre les phonèmes dont la prononciation est proche suggère que cet espace encode des informations acoustiques qui pourront être utilisées par le décodeur lors de l'inférence ; cette hypothèse sera explorée en section 4.3.

4.2.2 Évaluation de la prédiction phonétique

Les prédictions phonétiques obtenues par le classifieur à partir de la sortie de l'encodeur sont comparées aux transcriptions du corpus original par les règles de la section 3.2. La matrice de confusion du modèle est donnée en figure 3a.

Le modèle atteint une précision de 99% sur l'ensemble des phones. Ces performances sont supérieures à celles de (Perquin *et al.*, 2020) qui obtenait un taux d'erreur de 12.8% sur les phonèmes en entraînant un classifieur sur la sortie de l'encodeur d'un modèle Tacotron (Wang *et al.*, 2017). Ce gain peut s'expliquer par l'utilisation d'un modèle de synthèse vocale plus récent, entraîné sur des entrées à la fois orthographiques et phonétiques, et avec cette tâche supplémentaire de transcription 1 pour 1 entre graphèmes et phonèmes dès le début de son entraînement.

L'analyse plus détaillée de la matrice de confusion fait apparaître quelques erreurs fréquentes, résumées dans le tableau 2. Ces "erreurs" relèvent davantage d'un choix de style adopté par le locuteur que d'erreurs phonétiques. La majorité des distinctions entre la prédiction et le corpus d'origine apparaissent en fin de mot, dans le cadre de liaisons facultatives. L'écoute des synthèses concernées révèle que la prédiction du modèle est en accord avec la prononciation produite, mais que cette prononciation diffère de la prononciation adoptée dans le corpus original. Cette méthode de transcription automatique Graphème à Phonème par l'ajout d'une tâche supplémentaire lors de l'apprentissage d'un modèle de synthèse vocale permet donc d'associer deux avantages majeurs : 1) une excellente précision, et 2) une transcription adaptée à la synthèse correspondante, pertinente vis-à-vis des habitudes de style de parole du locuteur original (liaisons facultatives, schwas, etc..).

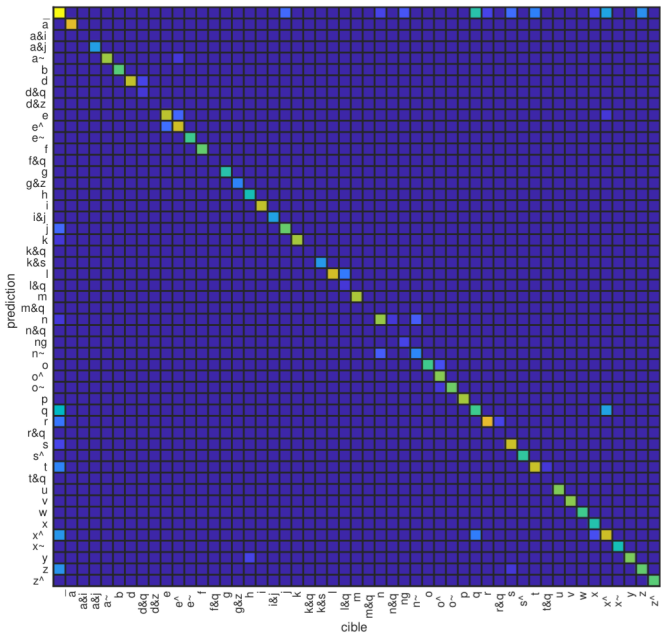
Phonèmes confondus		**Explications**	**Exemples**
/q/	/x^/	Schwas ou fin de mot appuyée	"quelqu**e**s rares fenêtres"
/o/	/o^/	Choix d'harmonisation vocalique	"**O**tons nos souliers"
/r/, /s/, /z/, /t/	/_/	Liaisons facultatives	"si tu n'es pas_heureux"

TABLE 2 – Confusions courantes révélées par la matrice de la figure 3a.

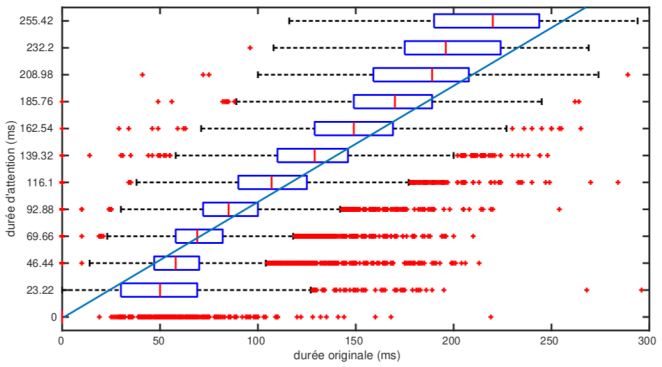
4.3 Analyse acoustique de l'espace des plongements

4.3.1 Segmentation automatique par l'attention

Afin d'identifier les paramètres acoustiques qui seraient encodés dans les plongements, le première étape consiste à isoler les portions du signal de sortie pendant lesquelles un plongement détermine la sortie audio. On utilise pour cela la procédure décrite en section 3.1. Afin de vérifier les performances de cette méthode, les durées d'activation ainsi calculées sont comparées aux durées des phonèmes dans le corpus initial obtenues par alignement semi-automatique. On adopte les règles définies en section 3.2 pour déterminer à quel plongement orthographique doit correspondre la durée du phonème : les caractères muets sont considérés comme ayant une durée de 0 secondes. Pour adopter le même rythme que la voix originale, les synthèses sont générées en prédiction, c'est-à-dire que les trames originales sont données au réseau en remplacement des trames calculées à chaque pas de temps du décodeur. Les résultats de cette comparaison sont donnés en figure 3b. La lecture de la carte d'attention permet de prédire les durées des phonèmes avec une corrélation de 0.88. La méthode proposée limite l'estimation des durées à un nombre pair de trames d'une longueur fixée par le modèle acoustique à 11.61ms (pour rappel : le modèle prédit 2 trames à la fois). Cette discrétisation de l'estimation peut expliquer une partie des erreurs de prédiction observées. Les performances de



(a) Matrice de confusion de la prédiction phonétique à partir de la sortie de l'encodeur. Tous les 80907 plongements du corpus de test sont pris en compte. Précision : 99%



(b) Comparaison entre les durées des phonèmes et la durée d'attention calculée suivant la méthode décrite en section 3.1.

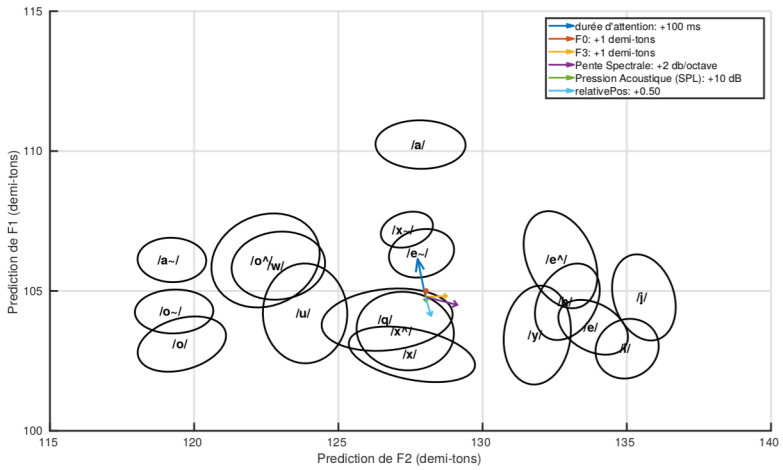
FIGURE 3 – Évaluation des méthodes proposées en section 3

cette méthode permettent d'envisager la segmentation automatique du corpus de test en vue d'une analyse acoustique.

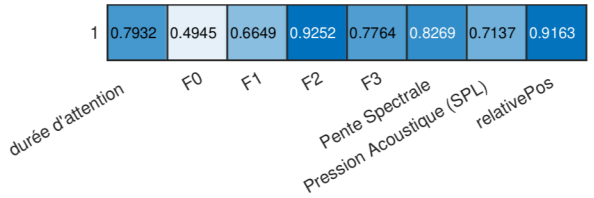
4.3.2 Corrélations acoustiques et contrôle

Suite aux observations de la section 4.3.1, la segmentation audio du corpus de test est réalisé en suivant la procédure décrite en section 3.1. On ne considère dans cette section que les plongements associés à des phonèmes voyelles, afin d'intégrer des mesures acoustiques dépendantes du voisement. 8 mesures sont considérées : la durée d'attention, F0, F1, F2, F3 la pente spectrale, le niveau de pression sonore, ainsi que la position relative du plongement dans la séquence d'entrée. Les mesures acoustiques sont effectuées automatiquement avec Praat (Boersma, 2001). Les plongements sont projetés dans un espace de dimensions réduites par MDS comme détaillé en section 4.2.1. Une régression multi-linéaire par paramètre est réalisée afin d'obtenir une estimation de ces paramètres en fonction de la position du plongement dans la MDS. Les coefficients de corrélation de ces régressions sont donnés en figure 4b. Pour représenter l'espace latent, on affiche en figure 4a les plongements de voyelles en fonction de l'estimation de leurs premier et deuxième formants. On retrouve en partie le triangle vocalique formé par le trinôme /a/, /i/ et /u/. Les voyelles ouvertes /o^/ et /e^/ sont bien placées entre les versions mi-fermées /o/ et /e/ et la voyelle la plus ouverte /a/. La prédiction des formants semble moins précise sur les phonèmes les moins fréquents dans le corpus : les nasales /a~/, /o~/, /x~/ et /e~/, ainsi que /o/.

Les régressions par paramètre permettent d'envisager l'espace des plongements comme un espace de contrôle effectif : chaque paramètre peut être modifié en déplaçant les plongements dans la direction entraînant le maximum de variation de ce paramètre. A titre d'exemple, le déplacement induit par une modification arbitraire des paramètres à l'étude, est affiché en figure 4a. Le contrôle de la synthèse par un biais appliqué dans cet espace profite des covariations apprises par le modèle. On note par exemple qu'augmenter la durée portée par l'attention aux plongements de 100ms est



(a) Gradients de variation des paramètres acoustiques



(b) Coefficients de corrélation entre les paramètres acoustiques mesurés et leur prédiction par régression multi-linéaire sur leurs coordonnées dans la MDS.

FIGURE 4 – Visualisation des paramètres acoustiques dans l'espace latent réduit par MDS

accompagné d'une augmentation du premier formant de 1.2 demi-tons, cohérente avec une durée plus importante des voyelles ouvertes (O'Shaughnessy, 1981). De même, une légère augmentation de F1 en augmentant F0 rejoint la simulation d'un effort vocal plus important (Liénard & Di Benedetto, 1999). À l'inverse, l'augmentation de la position relative du phonème est associée à une réduction de F1 qui mimique la réduction de l'ouverture des voyelles en fin de phrases. L'amplitude relativement faible des déplacements dans le plan de la figure 4a est rassurant vis-à-vis des possibilités de contrôle dans cet espace ; le spectre des phonèmes ne doit pas être dénaturé au point d'être confondu avec une autre voyelle lors de la modification d'un paramètre de durée par exemple.

La MDS étant un transformation linéaire de l'espace, l'inverse de la matrice de passage permet de projeter cette translation dans l'espace des plongements. Le vecteur résultant peut ensuite être ajouté à tous les plongements en sortie de l'encodeur avant le passage dans le décodeur, afin de modifier une ou plusieurs caractéristiques de la voix. L'évaluation du contrôle de la synthèse par cette méthode est encore à l'étude.

5 Conclusions

Cette article présente une méthode d'analyse acoustique et phonétique des représentations internes d'un modèle de synthèse vocale à l'état de l'art. Cette étude a montré que la séquence de graphèmes donnée en consigne au modèle de synthèse était mise en contexte par l'encodeur pour en calculer une représentation phonétique. De plus, cette analyse montre que les représentations phonétiques calculées par l'encodeur contiennent non seulement les cibles acoustiques à atteindre, mais également des informations de rythme et de positionnement dans la phrase. La localisation des paramètres acoustiques dans les plongements permet d'imaginer le contrôle de ces derniers, par exemple par l'ajout d'un biais global sur la phrase à synthétiser qui déplacerait tous les plongements dans une direction correspondant à la variation du paramètre choisi. Ce type de contrôle, ainsi qu'une analyse approfondie des indices supra-segmentaux de rythme et de position fera l'objet de futures recherches.

Remerciements

Ces recherches sont financées par la BPI dans le cadre du projet THERADIA et par MIAI@Grenoble-Alpes (ANR-19-P3IA-0003). Ces travaux ont l'accès à HPC/IDRIS sous l'attribution 2021-AD011011542R1 faite par GENCI.

Références

BAHDANAU D., CHO K. & BENGIO Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv :1409.0473*.

BAILLY G., PERROTIN O. & LENGLET M. (2021). Ressources for End-to-End French Text-to-Speech Blizzard challenge.

BOERSMA P. (2001). Praat, a system for doing phonetics by computer. *Glot. Int.*, **5**(9), 341–345.

BOSSE M.-L. & VALDOIS S. (2009). Influence of the visual attention span on child reading performance : a cross-sectional study. *Journal of research in reading*, **32**(2), 230–253.

BURKART N. & HUBER M. F. (2021). A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research*, **70**, 245–317.

HSU W.-N., ZHANG Y., WEISS R. J., CHUNG Y.-A., WANG Y., WU Y. & GLASS J. (2019). Disentangling correlated speaker and noise for speech synthesis via data augmentation and adversarial factorization. In *ICASSP*, p. 5901–5905 : IEEE.

KRUSKAL J. B. (1978). *Multidimensional scaling*. Number 11. Sage.

LENGLET M., PERROTIN O. & BAILLY G. (2021). Impact of segmentation and annotation in french end-to-end synthesis. In *11th ISCA Speech Synthesis Workshop*, p. 13–18 : ISCA.

LIÉNARD J.-S. & DI BENEDETTO M.-G. (1999). Effect of vocal effort on spectral properties of vowels. *The Journal of the Acoustical Society of America*, **106**(1), 411–422.

MODARRESI G., SUSSMAN H., LINDBLOM B. & BURLINGAME E. (2004). An acoustic analysis of the bidirectionality of coarticulation in vcv utterances. *Journal of Phonetics*, **32**(3), 291–312.

O'SHAUGHNESSY D. (1981). A study of french vowel and consonant durations. *Journal of Phonetics*, **9**(4), 385–406.

PERQUIN A., COOPER E. & YAMAGISHI J. (2020). An investigation of the relation between grapheme embeddings and pronunciation for tacotron-based systems. *arXiv preprint arXiv :2010.10694*.

PRENGER R., VALLE R. & CATANZARO B. (2019). Waveglow : A flow-based generative network for speech synthesis. In *ICASSP*, p. 3617–3621 : IEEE.

REN Y., HU C., TAN X., QIN T., ZHAO S., ZHAO Z. & LIU T.-Y. (2020). Fastspeech 2 : Fast and high-quality end-to-end text to speech. *arXiv preprint arXiv :2006.04558*.

SHEN J., PANG R., WEISS R. J., SCHUSTER M., JAITLY N., YANG Z., CHEN Z., ZHANG Y., WANG Y., SKERRV-RYAN R. *et al.* (2018). Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *ICASSP*, p. 4779–4783 : IEEE.

SKERRY-RYAN R., BATTENBERG E., XIAO Y., WANG Y., STANTON D., SHOR J., WEISS R. J., CLARK R. & SAUROUS R. A. (2018). Towards end-to-end prosody transfer for expressive speech synthesis with tacotron. *arXiv preprint arXiv :1803.09047*.

TACHIBANA H., UENOYAMA K. & AIHARA S. (2018). Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention. In *ICASSP*, p. 4784–4788.

TITS N., WANG F., HADDAD K. E., PAGEL V. & DUTOIT T. (2019). Visualization and interpretation of latent spaces for controlling expressive speech synthesis through audio analysis. *arXiv preprint arXiv :1903.11570*.

WANG Y., SKERRY-RYAN R., STANTON D., WU Y., WEISS R. J., JAITLY N., YANG Z., XIAO Y., CHEN Z., BENGIO S., LE Q., AGIOMYRGIANNAKIS Y., CLARK R. & SAUROUS R. A. (2017). Tacotron : Towards end-to-end speech synthesis. In *Proceedings of Interspeech*, p. 4006–4010, Stockholm, Sweden.