Transcription phonétique automatique pour la synthèse de la parole

Kévin Vythelingum[1, 2] Yannick Estève[2] Olivier Rosec[1]
(1) Voxygen, Pleumeur-Bodou, France
(2) LIUM, Le Mans Université, France
{kevin.vythelingum,yannick.esteve}@univ-lemans.fr,
{kevin.vythelingum,olivier.rosec}@voxygen.fr

RÉSUMÉ

La synthèse de la parole consiste à produire un signal de parole à partir d'une séquence de mots. Elle s'appuie sur un ensemble d'enregistrements de parole transcrits en mots et en chaînes phonétiques. La qualité de cette transcription influe directement sur la qualité globale des systèmes de synthèse. Or, les chaînes phonétiques sont généralement issues d'une phonétisation automatique du texte, qui ne varie donc pas d'un locuteur à l'autre. Dans ce travail, nous explorons différentes méthodes permettant d'obtenir des chaînes phonétiques dépendantes du signal de parole et du texte. Nous appliquons finalement nos résultats à la tâche de détection des erreurs de phonétisation. Autrement dit, nous cherchons à identifier des zones où les chaînes phonétiques initiales sont erronées. Sur des données en français, nous montrons que nous pouvons corriger de 76,6 à 90,7% des erreurs de phonétisation d'un système commercial en ne vérifiant que 3,6 à 18,5% des données.

ABSTRACT

Automatic phonemic transcription for text-to-speech synthesis

Text-to-speech synthesis (TTS) purpose is to produce a speech signal from an input text. This implies the annotation of speech recordings with word and phonemic transcriptions. The overall quality of TTS highly depends on the accuracy of phonemic transcriptions. However, they are generally automatically produced by grapheme-to-phoneme conversion systems, which don't deal with speaker variability. In this work, we explore ways to obtain signal-dependent and context-dependent phonemic transcriptions. We then apply our results on error detection of grapheme-to-phoneme conversion hypotheses in order to find where the phonemic transcriptions may be erroneous. On a French TTS dataset, we show that we can correct from 76.6 to 90.7 % of grapheme-to-phoneme conversion errors of a commercial system by checking only 3.6 to 18.5 % of phonemes.

MOTS-CLÉS : transcription phonétique, synthèse de la parole, détection automatique d'erreurs.

KEYWORDS: grapheme-to-phoneme conversion, text-to-speech synthesis, automatic error detection.

1 Introduction

La synthèse de la parole consiste à produire un signal de parole à partir d'une séquence de mots. La construction d'un tel système nécessite la création d'une base de données de signal segmenté (BDS), c'est-à-dire d'un corpus composé d'un ensemble d'enregistrements de parole transcrits en mots et

10.21437/JEP.2018-12

segmentés en unités acoustiques, chacune décrite par un phonème.

Plusieurs paradigmes peuvent être distingués en synthèse de parole, dont la synthèse par corpus, où le signal de parole résulte de la sélection et de la concaténation d'unités acoustiques (Hunt & Black, 1996), et la synthèse de parole paramétrique, où les paramètres acoustiques du signal de parole sont directement prédits à partir de la description linguistique de la séquence à prononcer (Zen *et al.*, 2009). Dans le premier cas, la BDS sert de corpus de sélection des unités acoustiques. Dans le second cas, elle permet l'apprentissage des modèles acoustiques. Récemment, des travaux ont montré que certains composants des systèmes de synthèse de la parole peuvent être remplacés par des modèles neuronaux, appris grâce à des descriptions phonétiques (Van den Oord *et al.*, 2016; Arık *et al.*, 2017a,b), ou non (Wang *et al.*, 2017; Shen *et al.*, 2017).

La transcription phonétique des BDS est généralement issue d'une phonétisation automatique du texte. De nombreuses approches ont été proposées dans la littérature. Parmi celles-ci, les plus populaires sont la recherche dans un dictionnaire, la phonétisation par règles (Béchet, 2001), les modèles de séquences jointes (Bisani & Ney, 2008; Galescu & Allen, 2002) et les systèmes inspirés de la traduction automatique qui considèrent des séquences de caractères à traduire en séquences de phonèmes (Laurent *et al.*, 2009; Rao *et al.*, 2015; Yao & Zweig, 2015).

Ne dépendant que du texte, la phonétisation automatique ne varie pas selon les locuteurs et selon les situations. Lors de la création d'une BDS pour une nouvelle voix, il est donc nécessaire de procéder à des ajustements, soit au niveau du paramétrage du phonétiseur, soit en corrigeant manuellement les phonèmes qui divergent du signal constaté. Dans Brognaux *et al.* (2014), les auteurs montrent que la correction manuelle des transcriptions phonétiques de BDS en français peut améliorer la qualité de la synthèse de la parole. De plus, des améliorations de la synthèse ont été constatées dans Dall *et al.* (2016) lorsque la phonétisation est meilleure. Il est donc essentiel que la transcription phonétique des BDS soit la plus juste possible. Autrement dit, qu'elle s'accorde avec le signal de parole.

Précédemment, nous avons montré que nous pouvions détecter des erreurs de phonétisation en exploitant partiellement le signal de parole grâce à l'alignement forcé d'un lexique phonétisé par un modèle acoustique (Vythelingum *et al.*, 2017). Autrement dit, nous cherchions à identifier des zones où les chaînes phonétiques initiales étaient erronées. Dans ce travail, nous explorons différentes méthodes permettant d'obtenir des chaînes phonétiques dépendantes et indépendantes du signal de parole et du texte. Nous appliquons finalement nos résultats à la tâche de détection des erreurs de phonétisation.

L'article est organisé de la façon suivante : nous détaillons tout d'abord les différents systèmes de transcription phonétique, puis nous expliquons la tâche de détection d'erreurs et les métriques utilisées, avant de donner les résultats obtenus.

2 Systèmes de transcription phonétique

2.1 Transcription phonétique à partir du texte

2.1.1 Système de phonétisation par règles

Le système de phonétisation par règles (Fig. 1) est un système propriétaire utilisé pour la synthèse de la parole en français. Il prend en entrée une séquence de mots et donne en sortie la transcription

phonétique de ceux-ci.

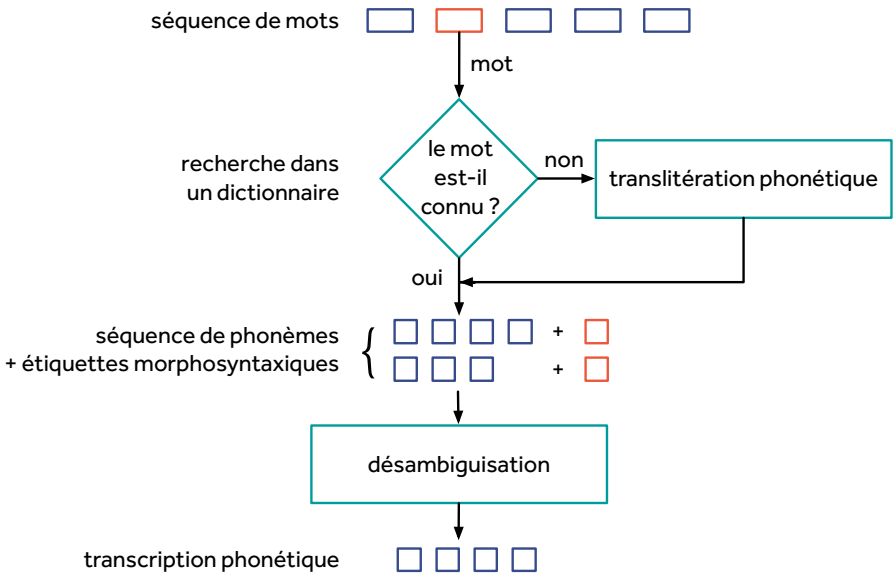

FIGURE 1 – Système de phonétisation par règles

Chaque mot est recherché individuellement dans un dictionnaire. S'il est présent, nous obtenons une ou plusieurs variantes phonétiques, associées à une étiquette morphosyntaxique. Sinon, des règles de translitération sont appliquées. Finalement, un module de désambiguisation permet de choisir une des variantes phonétiques proposées selon le contexte lexical.

2.1.2 Système de phonétisation par modèles de séquences jointes

Un modèle de séquences jointes est un modèle statistique permettant d'associer des séquences de lettres à des séquences de phonèmes (Bisani & Ney, 2008; Galescu & Allen, 2002). Le principe est d'effectuer un alignement des séquences de lettres, appelés graphèmes, et des phonèmes. Les séquences de couples *(graphèmes, phonèmes)* sont ensuite modélisées par un modèle de langage. La phonétisation d'un mot se fait donc en deux étapes : la génération des différentes séquences de couples *(graphèmes, phonèmes)* possibles selon les graphèmes du mot, puis la désambiguisation des variantes phonétiques grâce au modèle de langage.

Nous utilisons l'outil Phonétisaurus (Novak *et al.*, 2012, 2013) pour le modèle d'alignement phonétique. Ce dernier est associé à un modèle de language 6-gramme construit avec SRILM (Stolcke, 2002; Stolcke *et al.*, 2011) sur l'alignement graphèmes-phonèmes du corpus d'apprentissage.

Contrairement au système de phonétisation par règles, le système de phonétisation par modèles de séquences jointes ne peut traiter que des mots isolés. Il nous est donc davantage utile pour enrichir un lexique phonétisé que pour directement annoter en phonèmes les BDS.

2.1.3 Système de phonétisation par réseau de neurones

Pour obtenir une hypothèse de transcription phonétique dépendante du contexte lexical des mots, nous avons développé un modèle neuronal fondé sur l'architecture encodeur-décodeur décrite dans

(Bahdanau *et al.*, 2015). L'encodeur est un réseau de neurones récurrent avec une couche de GRU (Gated Reccurent Unit) de taille 128 bidirectionnelle. Il prend en entrée des mots segmentés au niveau caractères projetés dans un espace à 64 dimensions. Le décodeur est quant à lui composé de deux couches de GRU avec un mécanisme d'attention.

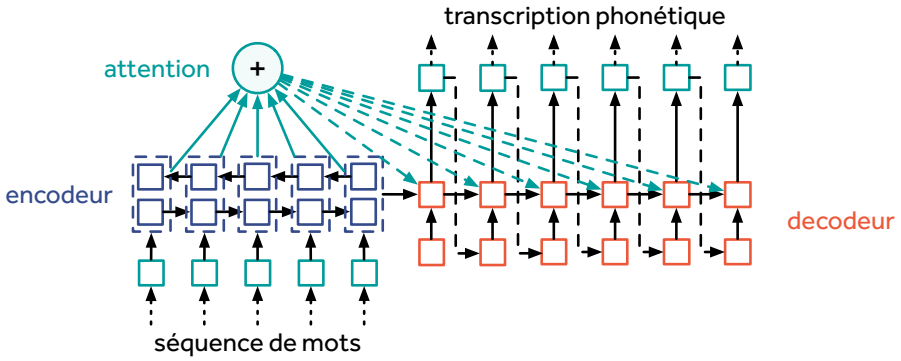

FIGURE 2 – Système de phonétisation par réseau de neurones

Nous utilisons l'outil Nmtpy (Caglayan *et al.*, 2017) dans sa configuration par défaut pour mettre en oeuvre ce modèle de traduction automatique neuronale. Ceci nous permet d'obtenir une hypothèse de phonétisation dépendante du texte sans réaliser d'alignement préalable entre caractères et phonèmes.

2.2 Transcription phonétique par alignement forcé

Pour prendre en compte le signal acoustique dans l'annotation en phonèmes des BDS, nous procédons à un alignement forcé du texte sur la parole. Ceci est fait grâce à un modèle acoustique d'une part, et à un lexique phonétisé d'autre part.

Tout d'abord, nous entraînons un modèle acoustique HMM-GMM (Hidden Markov Models - Gaussian Mixture Models) sur des paramètres PLP (Perceptual Linear Prediction) avec une adaptation au locuteur fMLLR (feature space Maximum Likelihood Linear Regression). Ensuite, nous entraînons un modèle HMM-DNN (Hidden Markov Models - Deep Neural Networks) sur les mêmes paramètres acoustiques en utilisant l'alignement produit par le modèle HMM-GMM. Notre modèle acoustique neuronal est composé d'une couche d'entrée de 360 dimensions, correspondant à des paramètres acoustiques de 40 dimensions concaténés aux paramètres voisins sur une fenêtre de 8 trames). De plus, 5 couches cachées comportant 3000 dimensions et une couche de sortie de 10553 dimensions composent le modèle. Les modèles acoustiques sont construits grâce à Kaldi (Povey *et al.*, 2011) sur des BDS existantes.

Le lexique utilisé pour l'alignement forcé est issu de la transcription phonétique des mots du corpus de test par le système de phonétisation par règles. Comme certaines hypothèses de prononciation sont manquantes, nous enrichissons ce lexique avec des hypothèses du système de phonétisation par modèles de séquences jointes. Afin de déterminer le nombre optimal d'hypothèses de prononciation à ajouter par mot, nous réalisons plusieurs alignements forcés avec à chaque fois un lexique plus ou moins enrichi.

2.3 Transcription phonétique à partir du signal de parole

2.3.1 Système de reconnaissance de phonèmes HMM-DNN

L'alignement forcé dépendant des variantes de prononciations présentes dans le lexique, le modèle acoustique est limité pour choisir l'hypothèse de phonétisation la plus probable. Une autre manière de prendre en compte le signal de parole dans la transcription phonétique des BDS est de réaliser une reconnaissance de phonèmes sur la partie acoustique des données.

Le système de reconnaissance de phonèmes utilise le même modèle acoustique que pour l'alignement forcé, le lexique étant constitué de la liste des phonèmes. Pour le décodage, nous avons appris un modèle de langage trigramme au niveau phonèmes sur le corpus d'apprentissage du modèle acoustique. Ce dernier ayant été validé manuellement, il nous permet d'apprendre les contraintes phonotactiques sur des données fiables.

2.3.2 Système de reconnaissance de phonèmes de bout en bout

Un système de reconnaissance de phonèmes de bout en bout est également évalué. Il s'agit de l'architecture de Deepspeech 2 (Amodei *et al.*, 2015), composée d'un réseau de neurones avec 2 couches de convolution et 5 couches de GRU bidirectionnelles de taille 800. Le système est entraîné grâce à la fonction d'activation CTC (Connectionist Temporal Classification) (Graves *et al.*, 2006) afin d'éviter la phase d'alignement entre les séquences de phonèmes et le signal. En effet, comme l'alignement temporel n'est pas requis pour la tâche de transcription phonétique, nous pouvons alors éviter d'induire des erreurs inhérentes à cette tâche.

3 Détection des erreurs de phonétisation

Nous cherchons à détecter les erreurs de phonétisation du système à base de règles. Pour cela, nous comparons ses sorties à des transcriptions corrigées manuellement, afin d'annoter l'ensemble des phonèmes avec des étiquettes *correct* ou *erreur*. Nous obtenons donc la référence pour la détection d'erreurs. Ensuite, les hypothèses des différents systèmes de transcription phonétique sont alignées avec les sorties du phonétiseur à base de règles : des phonèmes différents donneront une étiquette *erreur*, tandis que des phonèmes identiques donneront l'étiquette *correct*. Nous obtenons ainsi les différentes hypothèses de détection d'erreurs. Finalement, nous comparons référence et hypothèses de détections d'erreurs pour évaluer cette tâche.

Plusieurs métriques sont utilisées pour l'évaluation. D'une part, nous utilisons *précision* et *rappel* pour déterminer respectivement la proportion de vraies alarmes et la proportion d'erreurs détectées. D'autre part, nous calculons le pourcentage de données à valider, c'est-à-dire la proportion de phonèmes qu'un annotateur doit vérifier manuellement pour corriger l'ensemble des erreurs détectées. Ce dernier correspond au rapport entre le nombre de phonèmes supposés erronés et le nombre de phonèmes de la référence :

$$\text{données à valider} = \frac{\text{nombre de phonèmes supposés erronés}}{\text{nombre de phonèmes de la référence}}$$

Nous cherchons à maximiser précision et rappel, tandis que nous cherchons à minimiser la quantité de données à valider.

4 Résultats

Les données que nous avons utilisées pour l'évaluation de nos systèmes sont issues des bases de données de synthèse de Voxygen. Les modèles sont appris sur environ 50 heures de parole énoncées par 9 locuteurs, soit 90 135 séquences de mots. Les modèles sont testés sur environ 10 heures de parole énoncées par 3 locuteurs, soit 16 328 séquences de mots. Les locuteurs du corpus de test ne sont pas présents dans le corpus d'apprentissage. L'un d'eux, noté *locuteur 1* dans la suite, a la particularité d'avoir un accent africain sénégalais Bien que son accent diffère des autres locuteurs au niveau acoustique, il n'induit pas l'application de règles de phonétisation différentes. Ainsi, cela nous permettra d'évaluer la robustesse de nos modèles acoustiques.

4.1 Évaluation des systèmes de transcription phonétique

Tout d'abord, nous évaluons la performance des systèmes de transcription phonétique sur la tâche de phonétisation. Celle-ci est donnée en terme de taux d'erreur de phonétisation, pour chaque locuteur et pour l'ensemble des données (Tab. 1). Il s'agit de comptabiliser les substitutions, insertions et omissions.

#	Système	Locuteur 1	Locuteur 2	Locuteur 3	Total
(0)	phonétisation par règles	0,8	2,3	1,3	1,8
(1)	phonétisation par modèles de séquences jointes	8,4	9,0	7,4	8,5
(2)	phonétisation par réseau de neurones	**0,1**	**2,1**	**0,5**	**1,4**
(3)	alignement forcé lexique de base	5,0	2,9	2,7	3,1
(4)	alignement forcé lexique de base + 1 variante	**4,9**	**2,4**	**1,8**	**2,5**
(5)	alignement forcé lexique de base + 2 variantes	7,5	3,0	2,2	3,3
(6)	alignement forcé lexique de base + 3 variantes	10,1	3,5	3,0	4,2
(7)	reconnaissance de phonèmes HMM-DNN	54,6	13,3	12,4	18,3
(8)	reconnaissance de phonèmes CNN-RNN	**49,3**	**11,9**	**10,9**	**16,4**

TABLE 1 – Taux d'erreur de transcription phonétique (%) des différents systèmes étudiés

La transcription phonétique à partir du texte donne les meilleurs résultats, excepté pour la phonétisation par modèles de séquences jointes, qui ne prend pas en compte le contexte lexical des mots. La phonétisation par réseau de neurones améliore même la transcription phonétique par rapport à celle obtenue à base de règles. Les erreurs corrigées lors de l'étape de validation manuelle des données composant le corpus d'apprentissage du modèle ont donc été capturées.

Pour l'alignement forcé, nous faisons varier de 0 à 3 le nombre de variantes de prononciation que nous ajoutons à un lexique de base. Le lexique de base est constitué avec le phonétiseur par règles tandis que les variantes additionnelles sont produites par le phonétiseur par modèles de séquences jointes. Nous observons que la meilleure transcription est obtenue en ajoutant une seule variante. Cela montre que certaines variantes utiles n'ont pas été produites par le phonétiseur par règles mais qu'une trop grande variabilité dans le lexique nuit à l'alignement forcé.

Concernant la reconnaissance de phonèmes, le modèle neuronal de bout en bout permet de gagner deux points de taux d'erreur sur le modèle HMM-DNN, et ce pour chaque locuteur. Bien que le taux d'erreur soit beaucoup plus élevé que pour les systèmes dépendant du texte, il est possible que les erreurs soient différentes des autres systèmes, et permettent de détecter certaines erreurs de phonétisation.

Nous observons finalement que le taux d'erreur est plus élevé pour le locuteur 1 pour les systèmes dépendants du signal de parole. Les modèles acoustiques utilisés ne sont donc pas robustes à la particularité de l'accent de ce locuteur.

4.2 Application à la détection des erreurs de phonétisation

Nous évaluons les systèmes de transcription phonétique sur la tâche de détection des erreurs de phonétisation du système à base de règles. Les résultats sont donnés en termes de précision, rappel, et de quantité de données à valider pour corriger les erreurs détectées. Ceci permet d'estimer quel effort la validation nécessitera et quelle quantité d'erreurs pourra-t-on espérer corriger.

Dans un premier temps, nous considérons les différents systèmes de manière isolée (Tab. 2).

#	Système	Précision	Rappel	Données à valider
(1)	phonétisation par modèles de séquences jointes	13,4	**64,7**	8,6
(2)	phonétisation par réseau de neurones	**68,4**	58,5	**1,5**
(3)	alignement forcé lexique de base	30,9	51,0	3,0
(4)	alignement forcé lexique de base + 1 variante	**39,8**	64,8	**2,9**
(5)	alignement forcé lexique de base + 2 variantes	32,5	72,7	4,0
(6)	alignement forcé lexique de base + 3 variantes	27,0	**75,4**	5,0
(7)	reconnaissance de phonèmes HMM-DNN	7,8	**85,2**	19,3
(8)	reconnaissance de phonèmes CNN-RNN	**8,5**	83,0	**17,3**

TABLE 2 – Évaluation des systèmes étudiés sur la tâche de détection des erreurs de phonétisation (%)

Nous observons que les systèmes permettant la meilleure précision et offrant la plus petite quantité de données à valider sont ceux qui obtenaient les taux d'erreurs de phonétisation les plus faibles. Cependant, le meilleur rappel est obtenu à l'inverse par ceux dont la transcription divergeaient le plus de la référence. Nous comprenons qu'il est nécessaire de trouver un compromis entre temps passé et nombre d'erreurs corrigées lors de la validation manuelle. Les différents systèmes permettent donc de cibler différents niveaux de qualité de transcription. Nous remarquons ainsi que valider 1,5% des données permet de corriger 58,5% des erreurs, valider 2,9% des données permet de corriger 64,8% des erreurs, valider 5,0% des données permet de corriger 75,4% des erreurs et valider 17,3% des données permet de corriger 83,0% des erreurs.

Dans un second temps, nous considérons la combinaison des hypothèses de phonétisation des meilleurs systèmes étudiés (Tab. 3). Nous avons trois systèmes, soit celui qui obtenait le plus faible taux d'erreur de transcription phonétique pour chaque source utilisée. Nous profitons ainsi de la complémentarité des transcriptions phonétiques issues du signal de parole et du texte.

Nous observons que la combinaison de l'alignement forcé avec la phonétisation à partir du texte permet un gain important en rappel pour la même précision par rapport à l'alignement forcé seul. Nous passons en effet de 64,8% à 76,6% de rappel. De plus, la combinaison de la reconnaissance

Systèmes combinés	Précision	Rappel	Données à valider
(2) + (4)	**38,4**	76,6	**3,6**
(2) + (8)	8,8	87,2	17,5
(4) + (8)	8,5	87,7	18,1
(2) + (4) + (8)	8,6	**90,7**	18,5

TABLE 3 – Évaluation de la combinaison de systèmes sur la tâche de détection des erreurs de phonétisation (%)

de phonèmes avec les autres systèmes permet les rappels les plus importants. En combinant les trois systèmes retenus, nous pouvons corriger 90,7% des erreurs en validant 18,5% des données.

5 Conclusion

Nous comparons plusieurs méthodes pour obtenir une transcription phonétique à partir de signal de parole et de texte. Les systèmes étudiés, bien qu'ayant des taux d'erreurs de transcription différents, permettent d'obtenir des hypothèses complémentaires. Nous proposons d'appliquer ces résultats dans le cadre de la détection des erreurs de phonétisation, une tâche très utile en synthèse de parole pour accélérer le processus de développement de nouvelles voix. En combinant les hypothèses de phonétisation complémentaires des différents systèmes étudiés, nous parvenons à identifier des zones dans les données annotées où la transcription phonétique semble erronée. Dans le but d'augmenter la qualité des bases de données de synthèse, un annotateur humain peut donc en un temps minimum améliorer globalement les transcriptions. En effet, sur des données en français, nous montrons que nous pouvons corriger de 76,6 à 90,7% des erreurs de phonétisation d'un système commercial en ne vérifiant que 3,6 à 18,5% des données. Dans un prochain travail, nous chercherons à valider notre approche sur d'autres langues, notamment celles disposant de moins de ressources en termes de données et de connaissances linguistiques.

Références

AMODEI D., ANUBHAI R., BATTENBERG E. *et al.* (2015). Deep speech 2 : End-to-end speech recognition in english and mandarin. *CoRR*, **abs/1512.02595**.

ARIK S., CHRZANOWSKI M., COATES A. *et al.* (2017a). Deep voice : Real-time neural text-to-speech. In *arXiv :1702.07825v2*.

ARIK S., DIAMOS G., GIBIANSKY A. *et al.* (2017b). Deep voice 2 : Multi-speaker neural text-to-speech. In *arXiv :1705.08947v1*.

BAHDANAU D., CHO K. & BENGIO Y. (2015). Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations (ICLR)*.

BÉCHET F. (2001). Lia phon : un système complet de phonétisation de textes. In *Traitement Automatique des Langues (TAL)*, p. 47–67.

BISANI M. & NEY H. (2008). Joint-sequence models for grapheme-to-phoneme conversion. In *Speech Communication*, p. 434–451.

BROGNAUX S., B. P., DRUGMAN T. & D. L. (2014). Speech synthesis in various communicative situations : Impact of pronunciation variations. In *Proceedings of InterSpeech*.

CAGLAYAN O., GARCÍA-MARTÍNEZ M., BARDET A. *et al.* (2017). Nmtpy : A flexible toolkit for advanced neural machine translation systems. *Prague Bull. Math. Linguistics*, **109**, 15–28.

DALL R., BROGNAUX S., RICHMOND K. *et al.* (2016). Testing the consistency assumption : Pronunciation variant forced alignment in read and spontaneous speech synthesis. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 5155–5159.

GALESCU L. & ALLEN J. F. (2002). Pronunciation of proper names with a joint n-gram model for bi-directional grapheme-to-phoneme conversion. In *Proceedings of InterSpeech*.

GRAVES A., FERNÁNDEZ S., GOMEZ F. & SCHMIDHUBER J. (2006). Connectionist temporal classification : labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine learning*, p. 369–376.

HUNT A. J. & BLACK A. W. (1996). Unit selection in a concatenative speech synthesis system using a large speech database. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, p. 373–376.

LAURENT A., DELÉGLISE P. & MEIGNIER S. (2009). Grapheme to phoneme conversion using an smt system. In *Proceedings of InterSpeech*.

NOVAK J. R., DIXON P. R., MINEMATSU N. *et al.* (2012). Improving wfst-based g2p conversion with alignment constraints and rnnlm n-best rescoring. In *Proceedings of InterSpeech*.

NOVAK J. R., MINEMATU N. & HIROSE K. (2013). Failure transitions for joint n-gram models and g2p conversion. In *Proceedings of InterSpeech*.

POVEY D., GHOSHAL A., BOULIANNE G. *et al.* (2011). The kaldi speech recognition toolkit. In *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*.

RAO K., PENG F., SAK H. & BEAUFAYS F. (2015). Grapheme-to-phoneme conversion using long short-term memory recurrent neural networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 4225–4229.

SHEN J., PANG R., WEISS R. J. *et al.* (2017). Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *arXiv :1712.05884*.

STOLCKE A. (2002). Srilm – an extensible language modeling toolkit. In *Proceedings of InterSpeech*.

STOLCKE A., ZHENG J. & WANG W. (2011). Srilm at sixteen : Update and outlook. In *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding*.

VAN DEN OORD A., DIELEMAN S., ZEN H. *et al.* (2016). Wavenet : A generative model for raw audio. In *arXiv :1609.03499v2*.

VYTHELINGUM K., ESTÈVE Y. & ROSEC O. (2017). Error detection of grapheme-to-phoneme conversion in text-to-speech synthesis using speech signal and lexical context. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop*.

WANG Y., SKERRY-RYAN R., STANTON D. *et al.* (2017). Tacotron : Towards end-to-end speech synthesis. In *arXiv :1703.10135*.

YAO K. & ZWEIG G. (2015). Sequence-to-sequence neural net models for grapheme-to-phoneme conversion. In *Proceedings of InterSpeech*.

ZEN H., TOKUDA K. & BLACK A. W. (2009). Statistical parametric speech synthesis. In *Speech Communication*, p. 1039–1064.