Détection des erreurs de phonétisation pour la synthèse de parole

Kévin Vythelingum

COMUE L'UNAM
Ecole Doctorale STIM
Sciences et Technologies de l'Information et Mathématiques

Spécialité : Informatique
Laboratoire : LIUM
Equipe : Language and Speech Technology

Détection des erreurs de phonétisation pour la synthèse de parole

Vythelingum, Kévin
Mél : kevin.vythelingum@univ-lemans.fr

Résumé : En synthèse de parole par corpus, la création d'une voix passe par la transcription d'enregistrements de parole lue en séquences phonétiques. Celle-ci est produite par une phonétisation automatique du texte lu, suivie d'une validation manuelle de la transcription. Cette étape de correction est potentiellement longue alors que peu d'erreurs de phonétisation sont présentes. Nous proposons une méthode de détection des erreurs de phonétisation fondée sur l'utilisation du signal de parole. Cette méthode met en oeuvre un modèle acoustique issu de la reconnaissance de la parole pour aligner un lexique phonétisé sur le signal transcrit en mots. Nous montrons que nous détectons de 50.98% à 79.99% des erreurs de phonétisation selon le lexique utilisé, pour un phonétiseur évalué à 1.8% de taux d'erreur. De plus, les phonèmes annotés comme erronés représentent entre 2.9% et 3.6% du corpus, ce qui permet de réduire de façon importante la quantité de données à valider manuellement.

Mots clés : *Phonétisation, Synthèse de parole, Reconnaissance de la parole, Alignement forcé*
Collaborations : Voxygen

1 Introduction

La parole est un langage articulé naturel utilisé universellement. Chercher à comprendre les mécanismes concernés dans sa production et tenter de reproduire ce phénomène est sans conteste un enjeu scientifique majeur. La synthèse de parole est une activité consistant à proposer des modèles permettant de produire artificiellement un signal de parole à partir d'une description linguistique. Cette dernière est le plus souvent une séquence de mots : on parle alors de synthèse de parole à partir du texte.

Un signal de parole peut être décrit en associant des symboles à certaines classes d'éléments sonores : ce sont les phonèmes. Pour générer un signal à partir d'un texte, celui-ci est d'abord transformé en une séquence phonétique prononçable sans ambiguïté. Ensuite, deux paradigmes peuvent être distingués. Le premier est de proposer des modèles de description du spectre sonore de la parole pour chaque phonème, permettant de générer un signal artificiel. Le deuxième est de sélectionner et de concaténer des morceaux de parole réelle issus de l'enregistrement de la lecture d'un texte. Nous nous intéressons ici à cette dernière approche, appelée synthèse par corpus, qui non seulement produit de la parole souvent perçue comme plus naturelle, mais aussi permet de reproduire la voix de locuteurs humains existants.

Pour construire une voix de synthèse par corpus, un texte est lu par un locuteur dont on souhaite reproduire la voix. Ce texte est d'abord phonétisé, puis la séquence phonétique est alignée temporellement sur le signal de parole enregistré. Ainsi, des unités acoustiques pouvant être concaténées les unes aux autres sont repérées par des phonèmes alignés sur le signal. Pour garantir une synthèse de parole de bonne qualité, il est donc nécessaire que la phonétisation soit juste, c'est-à-dire qu'elle décrive au mieux ce qu'a prononcé le locuteur lors de sa lecture du texte. Dans ce travail, nous nous appuyons sur des modèles acoustiques issus de la reconnaissance de la parole afin d'obtenir une séquence phonétique dépendante du signal. Nous vérifions alors que cette dernière, confrontée à la phonétisation du texte, permet de repérer d'éventuelles erreurs d'annotation.

2 Phonétisation

La phonétisation est l'action de transcrire un texte en une séquence de phonèmes. Cette séquence de phonèmes permet ensuite de sélectionner une séquence d'unités acoustique pour former un signal de parole. Plusieurs méthodes permettent de passer d'une représentation orthographique d'un texte à une représentation phonétique.

La première consiste à appliquer des règles de réécriture des mots reposant sur une analyse lexicale et morpho-syntaxique du texte. Ces règles sont complétées par un dictionnaire d'exceptions pour couvrir certains mots qui suivent des conventions différentes, comme souvent les noms propres ou les mots d'origine étrangère.

Lorsqu'un lexique phonétisé existe, des techniques de modélisation statistique permettent de l'étendre à d'autres mots, comme les modèles de séquences [1] ou plus récemment les modèles fondés sur des réseaux de neurones

[2, 3]. Ces modèles se fondent sur un alignement entre les lettres et les phonèmes pour déterminer les associations fréquentes.

Quelle que soit la méthode employée, la phonétisation à partir du texte est source d'erreur dans la mesure où un locuteur humain est susceptible d'utiliser des variantes de prononciation imprévues ou difficilement gérées par le phonétiseur. Pour la synthèse de parole, il est alors indispensable de procéder à une vérification manuelle de l'annotation phonétique des données enregistrées. Une méthode de détection des erreurs de phonétisation permettrait de faciliter cette tâche.

3 Détection des erreurs

3.1 Motivations

La phonétisation n'utilise que le texte pour produire une chaîne phonétique. Or, dans le cadre de la construction de bases de données de synthèse, l'enregistrement audio de la parole est disponible. Exploiter le signal de parole pour produire une chaîne phonétique permettrait de détecter certaines zones où l'annotation automatique à partir du texte est erronée.

En reconnaissance de la parole, on cherche à modéliser la réalisation acoustique de chaque phonème. Pour cela, il est nécessaire d'annoter en phonèmes un corpus de parole. Cependant, cette tâche étant difficile à effectuer manuellement, on préfère en pratique transcrire la parole en mots puis utiliser un lexique de prononciation afin d'obtenir la phonétique. Le processus consistant à aligner les phonèmes sur le signal de parole s'appelle l'alignement forcé. Cet alignement est réalisé conjointement à la modélisation acoustique, selon une procédure itérative.

Une fois le modèle acoustique construit à l'aide d'un corpus d'apprentissage, il est possible de l'utiliser pour aligner les phonétisations d'un lexique sur des données supplémentaires. Lorsque plusieurs variantes de prononciation sont proposées par le lexique, l'alignement forcé sélectionne la plus problable au regard de l'observation acoustique. C'est cette sortie qui nous intéresse : nous obtenons de cette façon une phonétisation automatique dépendante du signal.

Etant donné que la phonétisation automatique à partir du texte et par alignement forcé est réalisée à l'aide de sources différentes, nous émettons l'hypothèse que les zones de divergence entre les deux sources correspondent à des erreurs de phonétisation, ou au moins à des zones douteuses nécessitant validation manuelle. La comparaison des deux phonétisations permettrait alors de détecter des erreurs.

3.2 Architecture

Nous considérons un corpus d'enregistrement de parole lue transcrit en mots et en phonèmes. Celui-ci est séparé en deux, une partie étant réservée pour l'apprentissage des modèles et l'autre pour l'évaluation. Un modèle acoustique est alors réalisé avec le *toolkit* de reconnaissance de la parole Kaldi [4]. Ce modèle, fondé sur des modèles de Markov cachés et des réseaux de neurones profonds, est construit selon une approche considérée comme standard [5].

Ensuite, le texte du corpus d'évaluation est d'une part phonétisé par alignement forcé (phn_1), d'autre part phonétisé avec un outil interne à Voxygen à l'aide de règles de réécriture des mots par analyse lexicale et morpho-syntaxique (phn_2). Cette dernière phonétisation est corrigée manuellement pour obtenir la référence d'évaluation (phn_3).

Finalement, la phonétisation à partir du texte est comparée à la phonétisation de référence pour identifier les zones d'erreur. La phonétisation phn_1 est également comparée à phn_2, ce qui constitue les hypothèses d'erreurs. L'évaluation de la détection des erreurs de phonétisation consiste alors à comparer les hypothèses d'erreurs à celles identifiées par la référence. La figure 1 montre une vue d'ensemble de l'architecture du système de détection des erreurs de phonétisation.

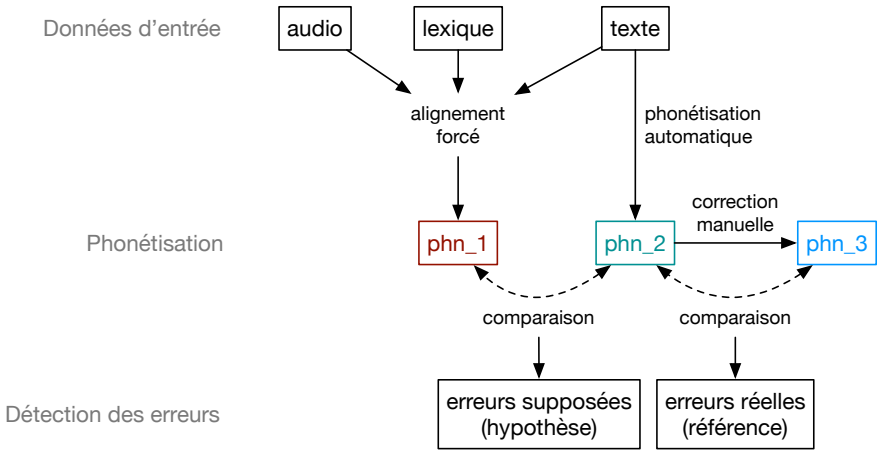

Fig. 1 – architecture du système de détection des erreurs de phonétisation

4 Expériences

4.1 Données

Les différentes expérimentations seront menées sur des enregistrements de parole lue en français, issus des bases de données de synthèse de Voxygen. Nous réservons 50h de parole pour l'apprentissage du modèle acoustique, et 10h pour l'évaluation de la détection des erreurs [Tab. 1].

	Apprentissage	Évaluation
audio	50h	10.8h
segments	90 135	16 328
mots (uniques / occurrences)	33 191 / 618 155	15 607 / 125 433

Tab. 1 – répartition des données pour l'apprentissage et l'évaluation

Plusieurs lexiques seront expérimentés pour l'alignement forcé. Un premier lexique est construit par phonétisation des mots du corpus d'évaluation à l'aide d'un phonétiseur automatique (lex_1). Cependant, certaines variantes de prononciation peuvent être absentes ou erronées. Or, si la phonétisation attendue n'est pas proposée par le lexique, il n'y a aucune possibilité pour que l'alignement propose une solution adéquate. Pour estimer le biais induit par cette limitation, nous utilisons un second lexique qui reprend le précédent en incluant la phonétisation manuelle de référence (lex_2).

4.2 Évaluation

La phonétisation automatique est d'abord évaluée en taux d'erreur de phonétisation *(PER, Phone Error Rate)* par comparaison à la phonétisation de référence. Il s'agit de comptabiliser les substitutions, insertions et omissions. Par ailleurs, la détection des erreurs est évaluée en *précision* et *rappel* en comparant les annotations en erreurs hypothèse et référence. La *précision* mesure la capacité du système à proposer des erreurs qui en sont réellement, tandis que le *rappel* mesure la capacité du système à trouver l'ensemble des phonèmes erronés.

4.3 Résultats

Nous commençons par évaluer la phonétisation automatique du corpus d'évaluation [Tab. 2]. Ce dernier comporte 427 768 phonèmes selon la transcription manuelle de référence.

	PER
Phonétisation par règles	1.8%
Alignement forcé lex_1	3.1%
Alignement forcé lex_2	2.5%

Tab. 2 – évaluation du taux d'erreur de phonétisation

La phonétisation par règles fait globalement peu d'erreurs, avec seulement 7 677 phonèmes erronés. Ce sont ces erreurs que nous cherchons à détecter. Par ailleurs, l'alignement forcé donne un meilleur score lorsque les entrées correspondant à la phonétisation de référence sont présentes dans le lexique. Cela montre que les alternatives de prononciation proposées uniquement par phonétisation automatique limitent effectivement l'alignement forcé.

Ensuite, nous évaluons la détection des erreurs de phonétisation [Tab. 3].

	Erreurs supposées	Erreurs détectées	Précision	Rappel
Alignement forcé avec lex_1	12 677	3 914	30.87%	50.98%
Alignement forcé avec lex_2	15 210	6 141	40.37%	79.99%

Tab. 3 – évaluation de la détection des erreurs

La détection d'erreur avec lex_2 donne de bons résultats, avec un rappel de presque 80% en moyenne. Malheureusement, nous descendons à un peu plus de 50% de rappel seulement lorsque nous considérons uniquement la phonétisation automatique. Cela veut dire que la phonétisation par alignement forcé a tendance à reproduire les erreurs de la phonétisation par règles.

Nous pouvons remarquer que dans le cas de l'alignement forcé avec lex_1, seulement 2,9% des phonèmes sont annotés comme incorrects. Ainsi, d'après la mesure de rappel, la validation manuelle de ces seuls phonèmes permettrait de corriger plus de 50% des erreurs de phonétisation. Avec un meilleur lexique, d'après les résultats de l'alignement forcé avec lex_2, nous pouvons corriger près de 80% des erreurs en vérifiant uniquement 3.6% du corpus.

L'alignement forcé d'un lexique phonétisé sur le signal de parole permet donc de réduire de façon importante la quantité de données à valider manuellement dans le processus de création d'une voix de synthèse de parole. De plus, améliorer les variantes de prononciation proposées par le lexique promet des gains encore plus significatifs.

5 Conclusion

La phonétisation automatique du texte à l'aide de règles de réécriture des mots par analyse lexicale et morpho-syntaxique fait relativement peu d'erreurs. Cependant, dans le cadre de la création de voix pour la syntèse de parole par corpus, la correction de ces erreurs de phonétisation nécessite de valider une grande quantité de données manuellement. L'exploitation d'un modèle acoustique, habituellement utilisé pour la reconnaissance de la parole, permet d'obtenir une phonétisation dépendante du signal de parole grâce à l'alignement forcé d'un lexique phonétisé. La confrontation des deux phonétisations obtenues permettent de repérer des zones d'erreur, d'autant plus si le lexique employé propose des variantes de prononciation pertinentes. Dans un travail futur, nous chercherons à améliorer les variantes de prononciations proposées par les lexiques, notamment en explorant l'usage de méthodes de phonétisation statistiques. De plus, nous appliquerons notre méthode de détection d'erreurs à d'autres langues, où il n'existe pas toujours de phonétiseur par règles aussi abouti que pour le français.

Références

[1] Maximilian Bisani and Hermann Ney. Joint-sequence models for grapheme-to-phoneme conversion. *Speech Communication*, 50(5) :434–451, 2008.

[2] Kanishka Rao, Fuchun Peng, Hasim Sak, and Françoise Beaufays. Grapheme-to-phoneme conversion using long short-term memory recurrent neural networks. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 4225–4229. IEEE, 2015. 00009.

[3] Kaisheng Yao and Geoffrey Zweig. Sequence-to-sequence neural net models for grapheme-to-phoneme conversion. *arXiv preprint arXiv :1506.00196*, 2015. 00007.

[4] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukáš Burget, Ondřej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlíček, Yanmin Qian, Petr Schwarz, and others. The Kaldi speech recognition toolkit. 2011.

[5] Geoffrey Hinton, Li Deng, Dong Yu, George E. Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath, and others. Deep neural networks for acoustic modeling in speech recognition : The shared views of four research groups. *Signal Processing Magazine, IEEE*, 29(6) :82–97, 2012.