Développement d'un système de reconnaissance automatique de la parole en coréen avec peu de ressources annotées

Antoine Laurent Lori Lamel
Spoken Language Processing Group
CNRS-LIMSI, BP 133
91403 Orsay cedex, France
laurent@limsi.fr, lamel@limsi.fr

RÉSUMÉ

Ce papier décrit le développement d'un système de reconnaissance automatique de la parole pour le coréen. Le coréen est une langue alpha-syllabique, parlée par environ 78 millions de personnes dans le monde. Le développement de ce système a été mené en utilisant très peu de données annotées manuellement. Les modèles acoustiques ont été adaptés de manière non supervisée en utilisant des données provenant de différents sites d'actualités coréens. Le corpus de développement contient des transcriptions approximatives des documents audio : il s'agit d'un corpus transcrit automatiquement et aligné avec des données provenant des mêmes sites Internet. Nous comparons différentes approches dans ce travail, à savoir, des modèles de langue utilisant des unités différentes pour l'apprentissage non supervisé et pour le décodage (des caractères et des mots avec des vocabulaires de différentes tailles), l'utilisation de phonèmes et d'unités "demi-syllabiques" et deux approches différentes d'apprentissage non supervisé.

ABSTRACT

Development of a speech recognition system for Korean with little annotated training data

This paper describes the development of a speech-to-text transcription system for the Korean language. Korean is an alpha-syllabary language spoken by about 78 million people worldwide. System development was carried out with only a small amount of manually transcribed audio data from LDC. Additional audio data downloaded from several Korean were included in the training corpus and used in an unsupervised manner. Korean web texts were also downloaded and used for language modeling. The development corpus is associated with approximative transcripts obtained by aligning automatic transcripts with texts taken from the same web site. We compare several systems : the use of different LM (in terms of vocabulary size and using chars instead of words) for the unsupervised training and decoding, the use of phonemes and "half-syllable" acoustic units and two different approaches for unsupervised training.

MOTS-CLÉS : Reconnaissance automatique de la parole, apprentissage non supervisé, langue sous dotée.

KEYWORDS: Speech recognition system, unsupervised training, under-resourced language.

1 Introduction

Dans le cadre du projet RAPMAT[1], il était nécessaire de réaliser un système de reconnaissance automatique de la parole pour le coréen. Hors, ne disposant pas de données pour l'apprentissage d'un tel système, nous avons tout d'abord développé un système de base en utilisant les ressources (en quantité très limitée) disponible via le LDC[2], puis étudié une solution d'apprentissage non supervisé que nous présentons dans cet article.

Traditionnellement, les systèmes de reconnaissance automatique de la parole sont entrainés sur de grandes quantités de données audio transcrites et sur une importante quantité de textes écrits (Lamel et Vieru, 2010). Cependant, l'obtention d'un tel corpus est couteux et requiert des locuteurs natifs dans la langue pour laquelle le système est développé. L'un des coûts de développement les plus cités est celui du temps nécessaire à l'obtention des données audio transcrites, puisque cela nécessite à la fois de la main d'oeuvre et du temps. De nombreuses sources audio (radio, television, web, ...) peuvent être trouvées, cependant, dans la majorité des cas, les données audio ne sont pas associées à des transcriptions précises (Lamel *et al.*, 2002).

Plusieurs projets de recherche se sont intéressés à la réduction de ces coûts (Kimball *et al.*, 2004) et quelques données d'entrainement audio, comme dans le programme DARPA Gale sont associés à des transcriptions rapides (Cieri *et al.*, 2004). Quelques sources audio sont associées avec du texte, comme des sous-titres, des résumés ou des transcriptions plus éloignées (provenant par exemple de page Internet). Une variété d'approches a été étudiée, la plupart s'appuyant sur la supervision d'un modèle de langue. Les différentes approches varient en terme d'utilisation ou non des mesures de confiance (Collan *et al.*, 2005; Bisani et Ney, 2008; Wessel et Ney, 2005) et selon le degré de supervision (Lamel *et al.*, 2002). Les auteurs de (Ma et Schwartz, 2008) proposent un apprentissage itératif.

Dans cette étude, le développement du système est très légèrement supervisé. Nous avons utilisé un petit corpus annoté pour construire un premier modèle de langue et un modèle acoustique et ensuite seulement des données audio pour améliorer nos modèles. Les modèles de langue ont été construits en utilisant plusieurs sources textuelles différentes. Nous proposons de construire plusieurs systèmes en utilisant différents modèles de langue pour l'apprentissage non supervisé et pour le décodage (en terme de taille du vocabulaire, en utilisant des caractères à la place des mots), en utilisant les phonèmes et les "demi-syllables" comme unités acoustiques et en utilisant deux approches différentes pour l'apprentissage non supervisé des modèles acoustiques.

La section suivante donne un aperçu des caractéristiques de la langue coréenne, suivie par une description de l'approche et du corpus utilisé dans cette étude. Ensuite, les modèles de langue, la liste des phonèmes et les modèles acoustiques sont décrits, avant de présenter le contexte expérimental et les différents résultats obtenus.

2 La langue coréenne

Pendant plus d'un millénaire, le coréen était écrit avec des caractères Chinois adaptés, appelés Hanja. Au XVème siècle, un système d'écriture national appelé Hangul a été proposé et complète-

[1]Ce travail a été en partie financé par le projet DGA RAPID RAPMAT : http ://www.limsi.fr/tlp/rapmat.html
[2]https ://www.ldc.upenn.edu/

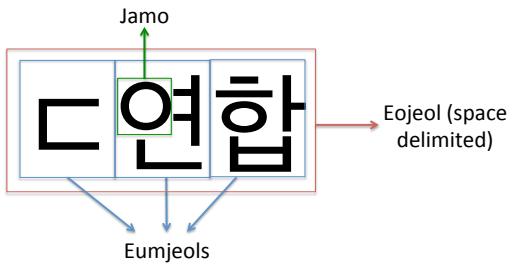

FIG. 1 – Symbole coréen

ment adopté au XXème siècle. Comme présenté dans (Kwon et Park, 2003), en coréen, un espace est placé entre deux “mots-phrases” adjacents, qui représentent deux ou trois mots Anglais au sens sémantique. C’est un système alpha-syllabique (Taylor, 1979). Comme décrit dans (Lee et Lee, 2009), les ensembles de Jamo (représentant des phonèmes) sont groupés en Eumjeols (syllables), et les séquences d’Eumjeols sont regroupées en Eojeol (mots séparés par des espaces), voir figure 1. Chaque Eumjeols est composé de deux ou trois éléments : le Choseong (consonne initiale), le Jungseong (la voyelle) et le Jongseong (optionnel, la consonne finale).

La plupart des résultats des systèmes de reconnaissance automatique de la parole développés pour le coréen sont substantiellement moins bon que les résultats obtenus avec des langues plus dotées telles que le Français ou l’Anglais. Comme les Eojeol représentent généralement plus d’un mot, le vocabulaire du système de reconnaissance basé sur les mots (éléments séparés par des espaces) doit contenir un nombre élevé d’entrées (K. *et al.*, 1999). Par exemple, un corpus Anglais contenant 40 millions de mots contient environ 190000 mots distincts (Lamel et Vieru, 2010), alors que notre corpus de 95 millions de mots en coréen contient environ 2 millions de mots différents. Une solution proposée pour ce problème est la décomposition des mots en morphèmes (comme présenté dans (Kwon et Park, 2003; Choi *et al.*, 2004)). Une autre explication des mauvais résultats des différents systèmes est le manque de données de parole et de texte pour l’apprentissage des systèmes. Seulement un nombre très restreint de corpus sont disponibles pour le coréen, et la plupart des travaux précédent se basent sur des corpus non redistribués.

3 Approche et corpus

L’approche général retenue pour ce travail est comparable à celle présentée dans (Ma et Schwartz, 2008; Lamel *et al.*, 2000, 2002; Lamel et Vieru, 2010) dans le sens ou un système de reconnaissance de la parole est utilisé pour produire des transcriptions approximatives pour l’apprentissage des modèles acoustiques. Dans cette étude, les données audio sont transcrites en lots et en procèdent à des itérations successives, les modèles sont estimés sur une quantité plus importante de données. Nous proposons également une autre approche dans laquelle à chaque itération, l’ensemble du corpus audio est transcrit pour l’adaptation des modèles acoustiques. Notre approche a été développée et testée sur un corpus pour lequel nous n’avons pas de référence exacte. Comme personne au laboratoire n’a de connaissances sur le coréen, nous ne savons pas à quel point notre corpus, extrait de sites web d’actualités coréen, est distant de l’audio provenant de

ces mêmes sites web.

Dans cette étude, nous explorons plusieurs approches, l'objectif étant de donner un aperçu des techniques qui semblent fonctionner et de celles qui semblent ne pas donner de bons résultats.

Nous avons utilisé un ensemble de documents audio et textuels pour l'apprentissage des différents modèles. Nous avons utilisé un corpus provenant de LDC qui contient des données d'émissions de journaux transcrites. Ce corpus ne contient qu'une petite quantité de données (9 heures). Nous avons également utilisé, comme données textuelles, les corpus LDC Korean newswire, newswire 2 et les transcriptions du corpus LDC de conversations téléphoniques pour l'apprentissage des modèles de langue (25M + 55M + 230k mots). Nous avons également utilisé des données provenant de trois différents sites d'actualités coréen : VOA[3], RFA[4] et NHK[5]. Pour NHK, 406 heures de données ont été téléchargées depuis novembre 2007 avec des transcriptions approximatives (5,5M mots). Ces transcriptions approximatives correspondent au contenu HTML des pages web associées à des données audio. Les deux autres sources de données ont été téléchargées depuis octobre 2013 : nous avons récupéré 5 heures de données de RFA et 4 heures de VOA. Les transcriptions pour RFA, VOA et NHK extraites des pages web, ne couvrent qu'une partie de l'audio et ne sont pas alignées. Une partie de RFA, VOA et NHK a été utilisé comme corpus de développement / test approximatif. Le corpus a été automatiquement transcrit en utilisant notre système d'amorçage utilisant uniquement les données de LDC, puis, en utilisant un algorithme basé sur la programmation dynamique (DTW), nous avons aligné les sorties automatiques du système avec les contenu HTML des pages et nous avons écarté les parties pour lesquels aucun alignement n'a pu être réalisé. Le même corpus a été utilisé pour le développement et le test de notre système.

4 Modèles de langue

Les données textuelles de LDC et les transcriptions approximatives de NHK, VOA et RFA ont été utilisées pour construire les différents modèles de langue. Nous avons construits des modèles 2-gram, 3-gram et 4-gram en utilisant l'ensemble des 2 millions de mots distincts de notre corpus d'apprentissage. Ces modèles de langue ngram ont été obtenus en interpolants les différents modèles entrainés sur les sous ensembles de données en utilisant la technique du lissage Kneser-Ney modifié. Les poids d'interpolations ont été automatiquement estimés en utilisant l'algorithme EM de façon à minimiser la perplexité du modèle final sur le corpus de développement/test. Ce corpus de développement est composé d'environ 10k mots. Nous avons également construit des modèles de langue contenant les 200k unigram les plus probables dans le modèle de langue 1-gram interpolé, et nous avons construit un modèle de langue à base de caractères.

La normalisation des textes coréens a été une grosse partie de ce travail. Comme nous n'avons pas de connaissance de la langue, nous avons utilisé la littérature disponible pour nettoyer les textes. Nous avons défini une liste de caractères obsolètes et ne respectant pas l'encodage EUC-KR (en utilisant les informations disponibles sur http ://en.wikipedia.org/wiki/Hangul), et retiré l'ensemble des phrases contenant des éléments de cette liste de notre corpus d'apprentissage. Nous avons également retiré les phrases contenant des séquences illégales de symboles, ou

[3]http ://www.voakorea.com/
[4]http ://www.rfa.org/korean/
[5]http ://www.nhk.or.jp/korean/

Représentation phonétique : C w a k
Représentation "demi-syllabique" : Cw ak

FIG. 2 – Exemple de découpage en “demi-syllabe”

contenant des symboles composés uniquement d’un seul Jamo. Les phrases contenant des mots anglais ont également été retirées. 122 caractères ont été identifiés comme étant des variantes du caractère espace, et donc substitués. Un dictionnaire contenant 25251 mots est livré avec le corpus de conversations téléphonique du LDC (ldc2003l02), ainsi qu’un outil pour phonétiser automatiquement les mots nouveaux. Nous avons également utilisé cet outil pour trouver les séquences illégales de symboles.

5 Modélisation acoustique

5.1 Liste des phonèmes et unités acoustiques

Les mots d’origine étrangère mis à part, le coréen est écrit en utilisant 14 consonnes simples et 5 consonnes doubles qui sont formées à partir des consonnes simples. Il y a 9 voyelles simples et 12 voyelles complexes. Ces voyelles complexes sont appelées diphtongues, elles sont la combinaison de 2 voyelles simples. Les mots coréens sont écrits de la gauche vers la droite. L’ensemble de phonèmes retenu pour ce travail comprend 25 phonèmes, 13 consonnes (les phonèmes K et k ont été regroupés en k, T et tt en t, P et pp en p, c et cc en C et S et ss en s – au format Sampa), 9 voyelles et 3 unités pour les évènements acoustiques (respiration, silence, filler).

Des modèles avec des “demi-syllabes” comme unité phonémique ont également été construits. Partant de la représentation phonétique de chaque symbole, les phonèmes ont été regroupés pour former ces “demi-syllabes” : chaque phonème jusqu’à rencontrer la première voyelle est considéré comme la première partie de la syllabe, et depuis cette voyelle jusqu’à la fin de la représentation phonétique du symbole est considéré comme la deuxième partie (voir figure 2).

5.2 Modèles acoustiques

Un modèle acoustique pour les phonèmes anglais a été utilisé pour initialiser notre modèle acoustique coréen. Une correspondance entre les phonèmes anglais et les phonèmes coréens a été faite et les modèles extraits ont été utilisés pour initialiser notre nouveau modèle. Les paramètres cepstraux standards (perceptual linear prediction - PLP) ont été utilisés. Les vecteurs de paramètres cepstraux sont composés de 39 paramètres : 12 coefficients cepstraux plus l’énergie sous forme logarithmique, associés à leur dérivées premières et secondes. Les modèles acoustiques sont des mélanges de gaussiennes à états partagés, gauche-droite dépendant du contexte. Les

modèles de phonèmes tri phones dépendants du contexte sont indépendants des mots, mais dépendants de leur position à l'intérieur des mots. Les états partagés sont obtenus au moyen d'un arbre de décision. Les modèles acoustiques sont indépendants du genre avec un apprentissage adaptatif (Speaker Adaptative Training - SAT). Le silence est modélisé par un seul état à 1024 gaussiennes.

6 Résultats expérimentaux

L'une des difficultés concernant le développement d'un système de reconnaissance de la parole basé sur les mots en coréen est le fait que le vocabulaire doit contenir un nombre élevé d'entrées. Comme présenté dans le tableau 1, en utilisant un modèle de langue contenant 200k mots (200k LM), le taux de mots hors vocabulaire (Out-Of-Vocabulary – OOV) sur notre corpus de développement est élevé (presque 10%), et est toujours de 3% avec un modèle de langue de 2 millions de mots (2M LM) contenant tous les mots du corpus d'apprentissage.

TAB. 1 – Perplexité et taux d'OOV des différents modèles de langue

Modèle	PPL	OOV (%)
2M LM (4g)	732	3.02
200k LM (4g) LM	1596	9.69

Le tableau 2 montre les résultats en termes de taux d'erreur mots (Word Error Rate – WER) et en termes de taux d'erreur sur les caractères (CER) en utilisant le modèle de langue 200k LM pour le décodage du corpus de développement.

TAB. 2 – WER et CER en utilisant 200k LM pour le décodage

Sources		200k LM		2M LM	
Audio	heures	WER	CER	WER	CER
LDC	9	50.6	32.1	50.6	32.1
Web	10	50.2	32.8	50.1	32.6
LDC+Web	19	49.1	31.5	48.8	31.0
LDC+Web	34	48.7	30.4	47.7	30.4
LDC+Web	79	48.0	29.9	47.6	29.7

Dans ce tableau, 200k LM et 2M LM correspondent au modèle de langue utilisé pour l'apprentissage non supervisé. Nous pouvons remarquer que le WER est relativement élevé. Cela peut être expliqué en partie par le fait que notre corpus de développement est un corpus approximatif et par le fait que, comme dit précédemment, chaque mot correspond à deux ou trois mots anglais au sens sémantique (Kwon et Park, 2003). En utilisant seulement un modèle acoustique estimé de façon non supervisé sur 10 heures de données audio provenant d'Internet (appelé "Web"), le WER du système est meilleur qu'en utilisant le corpus de LDC estimé sur un corpus annoté de 9 heures. Cela peut être expliqué par le fait que notre corpus de développement sur lequel le système est testé provient lui aussi de différents sites Web. L'ajout de nouvelles données permet d'observer une diminution en termes de WER et de CER.

Le tableau 3 montre les résultats obtenus en utilisant le modèle de langue de 2 millions de mots pour le décodage du corpus de développement. Nous pouvons remarquer que les résultats obtenus avec ce modèle sont meilleurs en terme de WER et de CER que ceux obtenus avec le modèle 200k et que la même tendance est visible lors de l'adaptation non supervisée du modèle acoustique.

TAB. 3 – WER et CER en utilisant 2M LM pour le décodage

Sources Audio	heures	200k LM WER	200k LM CER	2M LM WER	2M LM CER
LDC	9	47.5	29.4	47.5	29.4
Web	10	48.1	30.5	46.6	29.5
LDC+Web	19	45.2	28.2	42.3	27.0
LDC+Web	34	44.7	27.7	41.1	25.7
LDC+Web	79	44.0	27.0	40.2	25.4
LDC+Web	150	43.1	26.4	39.9	25.2

D'autres expériences, dans lesquelles un modèle de langue à base de caractères (char LM) pour l'apprentissage non supervisé, ont également été menées. Les modèles acoustiques résultant de ces expériences ont été utilisés pour décoder le corpus de développement avec le modèle 2M LM et avec le modèle char LM. Le décodage avec le modèle 2M LM ne donne pas de bons résultats en terme de WER (44,5% avec notre meilleur système), mais les résultats sont comparables, en terme de CER, en utilisant les modèles char LM ou 2M LM. Les résultats sont présentés dans le tableau 4.

TAB. 4 – WER et CER en utilisant char LM pour l'apprentissage non supervisé

Sources Audio	Heures	2M LM WER	2M LM CER	Char LM CER
LDC+Web	79	44.5	26.7	27.0
LDC+Web	150	44.0	26.4	26.5

Le tableau 5 montre les résultats obtenus en utilisant la deuxième technique d'apprentissage non supervisée : l'ensemble des données audio est transcrit, le modèle adapté, puis l'opération de transcription recommence. La diminution en termes de WER et de CER est plus rapide qu'en utilisant l'autre méthode : dès la première itération les résultats sont meilleurs que lorsqu'on ajoute les données au fur et à mesure.

TAB. 5 – WER et CER en utilisant toutes les données à chaque itération

Sources Audio	Iteration	2M LM WER	2M LM CER
LDC+Web 150h	1	39.7	25.2
LDC+Web 150h	2	39.7	25.3
LDC+Web 150h	3	39.6	25.3
LDC+Web 150h	4	39.5	25.1

Les expériences dans d'autres langues montrent une diminution plus rapide et plus importante des WER et CER en utilisant les deux méthodes non supervisées présentées. Nous pensons

que cette plus petite amélioration (WER de 47,5 à 39,5% et CER de 29,5 à 25,1%) observée pour le coréen est due à l'utilisation de notre corpus approximatif. Un corpus annoté devrait être disponible prochainement (un coréen natif a été recruté pour réaliser les transcriptions manuellement).

Nous pouvons remarquer dans le tableau 6 que les résultats en utilisant des "demi-syllabes" ne sont pas encourageants. Nous espérons également que l'expertise du transcripteur coréen va nous permettre d'améliorer la représentation utilisée.

TAB. 6 – WER et CER en utilisant 200k LM pour le décodage en unités "demi-syllabiques" et pour l'apprentissage

Sources Audio	Heures	200k LM WER	200k LM CER
LDC+Web	19	59.3	41.8
LDC+Web	79	53.8	34.5
LDC+Web	150	52.4	32

7 Conclusion

La description du développement d'un système de reconnaissance automatique de la parole pour le coréen, avec peu de données annotées à été présentée dans ce papier. Uniquement des données audio provenant de sites Web ont été utilisées pour l'adaptation non supervisées des modèles acoustiques. Pour cette étude, un corpus de développement/test approximatif a été utilisé. Ce corpus a été construit à partir de données (audio et textes) provenant de sites d'actualités coréens.

Les résultats montrent une diminution d'environ 17% en terme de WER et de 15% en terme de CER (relatif). Deux méthodes d'apprentissage non supervisées ont été proposées avec différents types de modèles de langue pour l'apprentissage et le décodage.

Un coréen natif va rejoindre l'équipe, il sera en mesure de transcrire notre corpus de développement approximatif et pourra nous aider pour la normalisation du texte et sa représentation acoustique.

La normalisation des données textuelles a été améliorée depuis le début de ce travail, nous aurons donc à recommencer l'apprentissage non supervisé du début pour construire de nouveaux modèles. Enfin, nous mettrons en place des modèles MLP (Multi-Layer Perceptron) pour améliorer nos modèles.

En utilisant peu de données annotées, nous avons développé un système de reconnaissance automatique de la parole en coréen. La méthode, nécessitant uniquement des données audio pour adapter le modèle acoustique, pourrait être appliquée à d'autres langues peu dotées, ou pour adapter des systèmes pour lesquels des ressources sont disponibles, à des conditions audio spécifiques.

Références

Bisani, M. et Ney, H. (2008). Joint-sequence models for grapheme-to-phoneme conversion. *In Speech Communication*, volume 50, pages 434–451.

Choi, I.-J., Kim, N.-H. et Yoon, S. Y. (2004). Large vocabulary continuous speech recognition based on cross-morpheme phonetic information. *In Interspeech 2004*, pages 453–456.

Cieri, C., Miller, D. et K., W. (2004). The fisher corpus : a resource for the next generations of speech-to-text. *In LREC 2004*, pages 69–71.

Collan, C., Bisani, M., Kanthak, S., Schlüter, R. et Ney, H. (2005). Cross domain automatic transcription on the tc-star epps corpus. *In ICASSP 2005*, volume 1, pages 825–828.

K., D., Schultz, T. et Waibel, A. (1999). Data-Driven Determination of Appropriate Dictionary Units for Korean LVCSR. *In Proceedings of International Conference on Speech Processing (ICSP'99)*, pages 323–327, Seoul, Korea.

Kimball, O., Kao, C., Iyer, R., Arvizo, T. et Makhoul, J. (2004). Using quick transcriptions to improve conversational speech models. *In Interspeech 2004*, pages 2265–2268.

Kwon, O.-W. et Park, J. (2003). Korean large vocabulary continous speech recognition with morpheme-based recognition units. *Speech Communication*, 39:287–300.

Lamel, L., Gauvain, J.-L. et Adda, G. (2000). Lightly supervised acoustic model training. *ITRW ASR*, 1:150–154.

Lamel, L., Gauvain, J.-L. et Adda, G. (2002). Lightly supervised and unsupervised acoustic model trainings. *Computer Speech and Language*, 16:115–129.

Lamel, L. et Vieru, B. (2010). Development of a speech-to-text transcription system for finnish. *In Workshop on Spoken Languages Technologies for Under Resourced Languages (SLTU 2010)*, pages 62–67, Penang, Malaysia.

Lee, J. et Lee, G. G. (2009). A data-driven grapheme-to-phoneme conversion method using dynamic contextual converting rules for korean tts systems. *Computer Speech and Language*, 23:423–434.

Ma, J. et Schwartz, R. (2008). Unsupervised versus supervised training of acoustic models. *In Interspeech 2008*, pages 2374–2377.

Taylor, I. (1979). The korean writing system : An alphabet ? a syllabary ? a logography ? *In Proceeding of Visible Language*, volume 2, pages 67–82, New York, USA.

Wessel, F. et Ney, H. (2005). Unsupervised training of acoustic models for large vocabulary continuous speech recognition. *In IEEE Transactions on Speech and Audio Processing*, pages 23–31.