Identification non-supervisée de pseudo-phones à l'aide de k-means et de réseaux convolutifs

Céline MANENTI, Thomas PELLEGRINI, Julien PINQUIER

IRIT, Université de Toulouse, UPS, Toulouse, France
{celine.manenti, thomas.pellegrini, julien.pinquier}@irit.fr

Résumé – Retrouver de manière non supervisée des unités sous-lexicales et identifier des pseudo-mots dans la parole est un problème qui intéresse actuellement les chercheurs en parole. Dans ce travail, nous trouvons des pseudo-unités à l'aide des k-means et d'un réseau de neurones convolutif, en s'aidant d'une segmentation phonétique manuelle ou automatique (multi-langue). Pour vérifier la portabilité de notre approche, nous comparons les résultats pour trois langues différentes : l'anglais, le français et le xitsonga. L'originalité de notre travail réside dans l'utilisation non supervisée d'un réseau de neurones qui d'ordinaire s'utilise dans un cadre supervisé. Comme exemple de résultats sur le corpus de Xitsonga, nous avons obtenu une pureté au niveau phonétique de 46% et de 42%, avec respectivement des segmentations manuelles et automatiques, en utilisant 30 regroupements. Sur la base de ces pseudo-phones, nous avons pu identifier environ 200 pseudo-mots.

Abstract –
Unsupervised discovery of sub-lexical units and pseudo-words in speech is a problem that currently interests speech researchers. In this paper, we retrieve phone-like pseudo-units using k-means and convolutive neural networks (CNNs), using either a manual or an automatic (cross-language) segmentation at phone-level. To assess the portability of our approach to any language, we compare the results for three different languages: English, French and Xitsonga. The originality of our work lies in the unsupervised use of a supervised type of neural networks: CNNs. As an example of our results, on the Xitsonga corpus, we obtained a phone purity of 46% and 42% with manual and automatic pre-segmentations respectively, when considering an *a priori* number of 30 clusters. On the basis of these pseudo-phones, we were able to indetify about 200 pseudo-words.

1 Introduction

Alors que les corpus annotés abondent dans les langues les plus parlées, la grande majorité des langues ou dialectes est peu dotée en annotations manuelles. Pour pallier à ce problème, la découverte non supervisée de pseudo-unités linguistiques dans un flux continu de parole gagne du terrain depuis quelques années, encouragée par exemple par des initiatives telles que le *Zero Resource Speech Challenge* [1] organisé en 2015 et 2017. Pour trouver les unités de la parole, il est possible d'utiliser des matrices de similarité et de la programmation dynamique (*Segmental Dynamic Time Warping*, S-DTW) [2]. La similarité utilisée peut être la distance cosinus entre les probabilités *a posteriori* ou « posteriorgrammes », données par un modèle acoustique phonétique entraîné sur un corpus annoté manuellement [3]. Dans [4], les modèles DTW et chaînes de Markov cachées sont également utilisés sur des posteriorgrammes pour trouver des pseudo-mots. Nous avons donc cherché à obtenir ces probabilités *a posteriori* d'une manière non supervisée.

Pour obtenir ces probabilités, un regroupement des fenêtres peut être utilisé. Dans [5], les k-means sont utilisés sur les paramètres générés par un auto-encodeur (AE), également appelés *Bottleneck Features* (BnF), après binarisation. Les k-means sont utilisés de manière similaire dans [6], avec AE et *graph clustering*. De plus en plus utilisés dans les technologies liées à la parole, les réseaux de neurones se déclinent en plusieurs versions non supervisées. Les AE apprennent à reconstruire les données d'entrée, après plusieurs transformations effectuées par des couches de neurones. Les paramètres intéressants se trouvent dans les couches cachées du réseau, où l'information contenue dans les données y est reformulée.

Dans le contexte de la découverte non supervisée d'unités de parole, des variantes des AE ont émergé, tels que le *correspondance AE* (cAE) [7]. Le cAE ne cherche plus à reconstruire les données d'entrée mais d'autres données, préalablement mises en correspondance. Ils nécessitent donc une première étape de regroupement de segments de parole en paires similaires (pseudo-mots, etc.) trouvées par une DTW. Il existe un autre type d'AE qui évite l'étape de la DTW en forçant à reconstruire les données voisines, en utilisant les propriétés de stabilité de la parole : les *segmental AEs* [5].

Dans ce papier, notre but est de trouver les sous-unités lexicales (phones, mots) de manière non supervisée. Les systèmes non supervisés ne peuvent pas trouver exactement ces unités : nous appelons les unités trouvées les pseudo-phones et les pseudo-mots. Ces pseudos-unitgés sont définies par des ensembles de segments de parole similaires. Un pseudo-mot ne correspond pas forcément à un mot : il peut s'agir de seulement une partie d'un mot ou bien de plusieurs mots fréquemment trouvés ensemble, tels que "I think that". Nous avons testé l'utilisation

d'AE pour résoudre notre problème mais les résultats ont été décevants (<30% de pureté). Nous avons donc cherché une approche différente : nous utilisons un réseau supervisé et non pas un AE. Nous avons opté pour un CNN (Convolutional Neural Network), auquel nous donnons à apprendre des clusters trouvés par des k-means. Dans ce papier, nous décrivons tout d'abord notre modèle puis nos données avant de parler de nos expériences et résultats puis de conclure.

2 Description du système

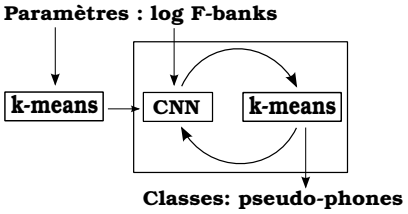

FIGURE 1 – Architecture générale de notre système itératif

La figure 1 montre le schéma général de notre approche, consistant principalement en un CNN qui apprend à prédire les classes attribuées par des k-means. Le premier k-means utilise en entrée des bancs de filtres (log F-Bank), puis les k-means des itérations suivantes utilisent les probabilités calculées par le CNN (posteriorgrammes). Le CNN, lui, prend toujours en entrée 40 log F-Bank, extraits sur des fenêtres de 16 ms, avec un recouvrement aux 3/4. Nous considérons cette entrée sur un voisinage de 6 fenêtres.. Ces deux étapes (CNN et k-means) sont répétées tant que le coût d'apprentissage du CNN décroît. Un des atouts de cette méthode est l'uniformisation par vote majoritaire au niveau des segments phonétiques des échanges entre les k-means et le CNN.

Actuellement, notre modèle a besoin de connaître les frontières des phonèmes ou phones afin d'uniformiser sur les segments les information échangées entre l'algorithme des k-means et le CNN. Nous utilisons et comparons deux segmentations différentes : une segmentation manuelle fournie avec le corpus et une segmentation automatique multi-langues (apprise sur des langues et testée sur d'autres), issue d'un travail précédent [8].

Lors de leur phase d'apprentissage, les réseaux de neurones supervisés ont besoin de connaître les classes associées aux exemples des données d'apprentissage. Ici, les vraies classes annotées manuellement ne sont pas disponibles. Nous utilisons à la place les pseudo-classes trouvées précédemment grâce aux k-means et à l'information de segmentation. Nous utilisons un CNN avec deux couches de convolution suivies d'une couche entièrement connectée et d'une couche finale de sortie. La première couche de convolution est composée de 30 filtres de taille 4×3 et la seconde de 60 filtres de taille 3×3. Un taux d'apprentissage de 0,007 a été utilisé avec des mises à jour de paramètres de type *momentum de Nesterov*.

3 Données

Nous avons utilisé 3 corpus, de langue, taille et conditions différentes : BUCKEYE, BREF80 et NCHLT.

BUCKEYE est un corpus d'anglais américain de parole spontanée, avec environ 30 minutes d'enregistrement par locuteur et 40 locuteurs différents. Il y a une soixantaine de phones différents, d'une durée médiane de 70 ms. Le nombre important de phones différents est dû au choix des concepteurs du corpus de séparer certaines prononciations différentes, principalement des prononciations nasales. Nous avons utilisé 13 heures provenant de 26 locuteurs différents, correspondant à la partie d'entraînement du découpage utilisé dans *Zero Resource Speech challenge* de 2015.

BREF80 est un corpus de français lu. Le jeu de phones en comporte 35, d'une durée médiane de 70 ms. Nous avons utilisés une heure enregistrée par 8 locuteurs différents.

NCHLT est un corpus de xitsonga, une langue d'Afrique du Sud. Nous avons utilisé une demi-heure d'enregistrements de 4 locuteurs différents, également issue des données du *Zero Resource Speech challenge*. Sur cette demi-heure, il y a 49 phones différents, d'une durée médiane de 90 ms.

4 Expériences et résultats

Dans cette section, nous rapportons d'abord les résultats obtenus en utilisant la segmentation en phones manuelle, nous permettant ainsi d'évaluer uniquement la tâche de regroupement de segments de paroles en pseudo-phones. Nous évaluons ensuite notre modèle à l'échelle des pseudo-mots, en utilisant ou non la segmentation manuelle.

4.1 Résultats au niveau des phones

Pour évaluer nos résultats, nous calculons la pureté des pseudos-phones, avec N le nombre de segments phonétiques, K le nombre de pseudo-phones, C le nombre de phones et n_j^i le nombre de segments du phone j automatiquement classé en tant que pseudo-phone i :

$$\frac{1}{N}\sum_{i=1}^{K} \arg\max_{j\in[1,C]}(n_j^i)$$

Dans un premier temps, nous avons cherché à optimiser les résultats du premier k-means (utilisé pour initialiser le processus en attribuant des numéros de classe aux segments phonétiques). Les paramètres qui influent ces résultats sont l'entrée (log F-bank), le nombre de fenêtres de contexte et le nombre de moyennes utilisées. Nous avons testé différentes tailles de voisinage et pu constater que l'influence de ce paramètre était d'au plus 1% sur les résultats. La meilleure valeur est autour de six fenêtres. Le choix du nombre de groupes se situe idéalement aux environs du nombre de phones recherchés, c'est-à-dire en général une trentaine.

L'utilisation du CNN nous apporte une plus grande amélioration, pour un faible nombre de classes. Ainsi, il fait gagner environ 10% pour 15 classes, 5% pour une trentaine et un très faible pourcentage au-delà de la centaine. Il est intéressant de comparer les résultats avec ceux obtenus de manière supervisée. La table 1 montre les valeurs de pureté obtenues avec un CNN supervisé, qui sont comme attendues supérieures à celles obtenues avec l'approche non supervisée.

TABLE 1 – Pureté par segment obtenue pour chaque corpus : BUCKEYE (An), BREF80 (Fr) et NCHLT (Xi)

Language	An	Fr	Xi
Pureté (%) avec l'apprentissage supervisé	60	62	66
Pureté (%) pour 30 groupes	29	43	46

L'une des applications possibles est d'aider à l'annotation manuelle de corpus. Dans le cas où un humain étiquette chacune des classes attribuées par le modèle avec des étiquettes phonétiques réelles, regroupant ainsi les doublons, nous pouvons envisager d'utiliser plus de moyennes que le nombre de phonèmes présents dans la langue considérée. Nous avons examiné l'évolution de la pureté en fonction du nombre de classes dans la figure 2. Nous voyons que, pour peu de classes, les résultats s'améliorent rapidement avec l'augmentation du nombres de classes. Mais, à partir d'une centaine de classes, la pureté commence à évoluer plus lentement : pour gagner environ 4%, le nombre de moyennes doit être multiplié par un facteur 10.

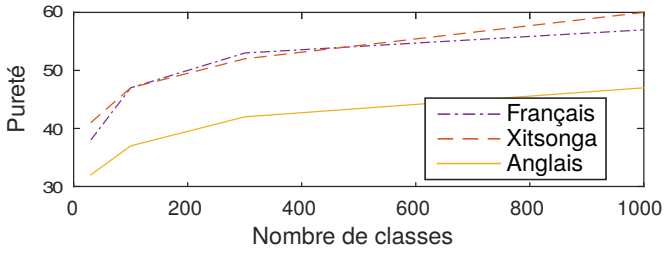

FIGURE 2 – Influence du nombre de classes sur la pureté

La table 1 donne des informations pour évaluer la qualité des résultats :

— le pourcentage de pureté obtenu en classification supervisée, par le même type de CNN que celui que nous utilisons dans le cadre non supervisé. En comparaison, l'état de l'art est d'environ 80% de précision sur le corpus TIMIT [9].

— le pourcentage de pureté obtenu par le modèle non supervisé. Les scores sont donnés pour 30 pseudo-classes : il y a près de 20% d'écart entre le petit corpus xitsonga lu et le grand corpus d'anglais spontané.

C'est avec les corpus français et xitsonga que nous avons obtenu les meilleurs résultats, en supervisé ou non. Les raisons probables sont qu'ils sont composés de textes lus, sont les plus petits corpus et ont peu de locuteurs différents.

En étudiant en détail la composition des regroupements, nous avons constaté qu'avoir des scores de pureté de 40% (ou de 30% pour l'anglais) ne signifie pas avoir 60% de phones totalement différents du label phonétique attribué. Les groupes sont généralement constitués de deux ou trois lots d'exemples appartenant à des classes phonétiques proches. Ainsi, les trois phones les plus fréquents dans chaque groupe représentent en moyenne 70% des échantillons du groupe pour le français ou le xitsonga et 57% pour l'anglais.

4.2 Résultats au niveau des mots

Pour trouver les pseudo-mots, nous regroupons entre elles les séquences égales de plus de 5 pseudo-phones. Considérer des séquences plus petites génère trop de pseudo-mots faux. Pour évaluer les pseudo-mots trouvés, nous comparons les transcriptions phonétiques correspondant aux segments de parole regroupés. si les séquences phonétiques sont égales, le pseudo-mot est correct. Sinon, nous comptons le nombre de différences entre les segments regroupés et tolérons jusqu'à 2 différences.

Dans le tableau 2, nous examinons trois caractéristiques des pseudo-mots identifiés pour évaluer nos résultats :

— nombre de pseudo-mots trouvés,

— nombre de pseudo-mots dont la transcription phonétique manuelle entre les exemples regroupés a au plus deux phones différents,

— nombre de pseudo-mots dont toutes les séquences les représentant ont exactement la même transcription phonétique.

TABLE 2 – Statistiques des pseudo-mots

Langue	An	Fr	Xi
# heures	13	1	1/2
# exemples de phones	586k	36k	10k
Segmentation manuelle			
# pseudo-mots	3304	671	231
# pseudo-mots $\leq$ 2 différences	1171	415	172
# pseudo-mots identiques	334	188	76
Segmentation automatique			
# pseudo-mots	3966	540	200
# pseudo-mots $\leq$ 2 différences	843	269	120
# pseudo-mots identiques	40	32	25

La transcription manuelle du corpus français contient 3201 suites de phones égales. Avec notre approche non supervisée, nous obtenons 671 pseudo-mots, dont 188 sont corrects et 227 avec une ou deux différences dans leurs transcriptions phonétiques. Nous nous trouvons ainsi avec 415 pseudo-mots avec au plus deux différences entre la transcription phonétique manuelle des exemples les définissant. Les résultats obtenus sur les deux autres corpus sont moins bons. Proportionnellement,

nous avons environ dix fois moins de pseudo-mots que pour le corpus français. En comparaison, dans un travail réalisé sur quatre heures du corpus ESTER, 1560 pseudo-mots ont été trouvés, dont 672 étaient suffisamment précis selon leurs critères, avec un algorithme de recherche pseudo-mot optimisé basé sur la DTW et les matrices d'auto-similarité [10].

En utilisant une **segmentation automatique** multilingue, la pureté des groupes de phonèmes diminue légèrement. Nous obtenons 42% pour le xitsonga (une diminution de 4%), 40% pour le français (-3%) et les mêmes résultats pour l'anglais. Concernant la découverte de pseudo-mots, nous obtenons les résultats affichés dans la table 2. Le nombre de pseudo-mots trouvés est similaire à celui trouvé avec la segmentation manuelle, mais le score de pureté est inférieur, comme prévu. Pour le français et le xitsonga, la moitié des pseudo-mots trouvés ont moins de deux erreurs, pour l'anglais il est inférieur à un quart.

5 Conclusions

Dans cet article, nous avons décrit nos expériences sur la découverte d'unités de parole fondée d'abord sur une approche simple utilisant les k-means sur des caractéristiques acoustiques usuelles. À partir de ces résultats, nous avons ensuite développée une version améliorée, dans laquelle un CNN est entraîné en apprenant les pseudo-phones trouvés par l'algorithme des k-means. Cette solution diffère de l'approche standard fondée sur les AE rapportés dans la littérature.

Notre modèle n'est pas encore totalement non supervisé : il a besoin d'une pré-segmentation au niveau des phones et, comme attendu, de meilleurs résultats ont été obtenus avec une segmentation manuelle. Néanmoins, la perte de performance due à l'utilisation de la segmentation automatique est faible. De plus, nous avons montré dans un travail précédent que cette segmentation peut être faite en utilisant un modèle de segmentation appris sur d'autres langues, pour lesquelles nous avons de grands corpus annotés manuellement. Cela nous permet d'appliquer notre approche à des langues sans aucune donnée annotée manuellement.

Nous avons testé notre approche sur trois langues : l'anglais américain, le français et une langue peu représentée appelée xitsonga. Avec le corpus de xitsonga, par exemple, et des segmentations manuelles et automatiques, nous avons pu obtenir des scores de pureté respectivement de 46% et 42%, avec 30 pseudo-phones. En utilisant ces 30 pseudo-phones, nous avons découvert environ 200 pseudo-mots. Dans toutes nos expériences, les résultats sur le corpus BUCKEYE, qui est composé de parole conversationnelle, sont moins bons que pour les deux autres corpus, qui sont constitués de parole lue. L'augmentation du nombre de locuteurs est également un facteur de diminution de la performance.

Notre prochain travail portera sur la segmentation non supervisée pour avoir un modèle entièrement non supervisé. De plus, nous avons présenté les premiers résultats sur la découverte de pseudo-mots fondée sur des séquences similaires de pseudo-phones. L'étape suivante consiste à appliquer des algorithmes de découverte de pseudo-mots, tels que des matrices de similarité.

Références

[1] M. Versteegh, R. Thiollière, T. Schatz, X.-N. Cao, X. Anguera, A. Jansen, and E. Dupoux, "The zero resource speech challenge 2015," in *Proc. INTERSPEECH*, pp. 3169–3173.

[2] A. S. Park and J. R. Glass, "Unsupervised pattern discovery in speech," *IEEE Trans. Audio, Speech & Language Processing*, vol. 16, no. 1, pp. 186–197, 2008.

[3] A. Jansen, K. Church, and H. Hermansky, "Towards spoken term discovery at scale with zero resources," in *INTERSPEECH 2010*. International Speech Communication Association, 2010, pp. 1676–1679.

[4] H. Wang, T. Lee, and C.-C. Leung, "Unsupervised spoken term detection with acoustic segment model," in *Speech Database and Assessments (Oriental COCOSDA), 2011 International Conference on*. IEEE, 2011, pp. 106–111.

[5] L. Badino, C. Canevari, L. Fadiga, and G. Metta, "An auto-encoder based approach to unsupervised learning of subword units," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 7634–7638.

[6] F. tian, bin Gao, Q. Cui, E. Chen, and T.-Y. Liu, "Learning deep representation for graph clustering," 2014, pp. 1293–1299.

[7] D. Renshaw, H. Kamper, A. Jansen, and S. Goldwater, "A comparison of neural network methods for unsupervised representation learning on the zero resource speech challenge," in *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015*, 2015, pp. 3199–3203.

[8] C. Manenti, T. Pellegrini, and J. Pinquier, "CNN-based phone segmentation experiments in a less-represented language (regular paper)," in *INTERSPEECH, San Francisco*. International Speech Communication Association (ISCA), septembre 2016, p. 3549.

[9] L. Badino, "Phonetic context embeddings for DNN-HMM phone recognition," in *Interspeech 2016, 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, September 8-12, 2016*, 2016, pp. 405–409.

[10] A. Muscariello, F. Bimbot, and G. Gravier, "Unsupervised Motif Acquisition in Speech via Seeded Discovery and Template Matching Combination," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 7, pp. 2031 – 2044, Sep. 2012.