Classement automatique de phonèmes dans un cadre multilingue

C. Corredor-Ardoy, P. Boula de Mareüil, M. Adda-Decker, L. Lamel, J.L. Gauvain

LIMSI-CNRS, BP 133
91403 Orsay Cedex, FRANCE
Tél.: ++33 1 69 85 81 02 - Fax: ++33 1 69 85 80 88
e-mail: {corredor, mareuil, madda, lamel, gauvain}@limsi.fr
http://www.limsi.fr/TLP

ABSTRACT

In this article, we describe an approach for automatic multi-lingual phoneme classification. The classes were obtained by agglomerative hierarchical clustering. We used a similarity measure based on the likelihood between the acoustic frames and the Hidden Markov Models. The method was applied to French, English, German, Spanish (IDEAL corpus), as well as to Italian and Portuguese (SPEECHDAT corpus). The analysis of the clusters demonstrated that, despite the acoustic mismatch between these corpora, this approach remains robust. For 90 clusters, the obtained classes correspond, to a large extent, to well defined linguistic groups. A qualitative analysis of the results is given.

1. INTRODUCTION

La recherche d'une typologie des phonèmes constitue un des enjeux de l'étude de la parole [9, 17] dans un but d'enseignement [5] ou de traitement automatique de la langue [2, 3]. Dans cet article, nous décrivons un algorithme de classement visant à regrouper les phonèmes de six langues européennes (français, anglais, allemand, espagnol, italien et portugais), et nous donnons une interprétation linguistique des résultats.

Le classement automatique des phonèmes dans un cadre multilingue a été incité par le développement de corpus destinés à l'identification automatique de la langue (IAL) : OGI-TS [13], CALLHOME, CALLFRIEND, SPEECHDAT, IDEAL [4]... Il est possible de remplacer plusieurs systèmes de reconnaissance phonétique par un seul système commun et unique à toutes les langues [4].

Dans notre approche, l'analyse acoustique est fondée sur les coefficients cepstraux plutôt que sur l'extraction de formants. Cette dernière en effet nécessite une segmentation voisé/non voisé, la localisation des noyaux vocaliques et l'extraction de la valeur des formants, décisions qui sont sujettes à erreur.

Nous proposons un algorithme de classement hiérarchique. Berkling [2] a également appliqué cette méthode à l'IAL de 6 langues du corpus OGI, mais elle utilise une mesure de similitude fondée sur des distances entre vecteurs acoustiques. Köhler [11] utilise lui une mesure définie sur des vraisemblances acoustiques ; néanmoins, cette mesure a seulement été apliquée à 3 des 11 langues du corpus OGI. La normalisation introduite dans notre mesure de similitude, de même fondée sur des vraisemblances acoustiques, a permis d'appliquer le classement automatique dans un cadre plus étendu : celui des corpus IDEAL [4] et SPEECHDAT. La mesure utilisée s'est révélée robuste par rapport aux différentes conditions d'enregistrement et au contenu liguistique des corpus.

Dans un cadre monolingue, l'algorithme de classement hiérarchique peut aider à l'analyse des classes phonétiques, à la définition de l'ensemble des unités destiné à la reconnaissance de la parole, et au groupement d'un grand nombre de modèles markoviens en contexte.

La section 2 présente l'algorithme de classement automatique. La section 3 est consacrée au cadre expérimental : corpus et alphabets. La section 4 analyse les résultats obtenus en faisant diminuer le nombre de classes à partir des 235 phonèmes de départ, et détaille le regroupement en 90 classes, avant de conclure.

2. L'ALGORITHME DE CLASSEMENT AUTOMATIQUE

2.1. *Préliminaire*

Depuis l'apparition des premiers systèmes de reconnaissance vocale fondés sur une modélisation phonétique markovienne, il s'est avéré important de pouvoir mesurer la distance entre les modèles. Cette mesure est utilisée pour réduire le nombre des modèles en contexte, pour déterminer l'ensemble optimal des phonèmes dans une langue ou pour établir un ensemble phonétique commun à plusieurs langues. Parmi les algorithmes de classement automatique, l'algorithme *k*-moyennes [6] est très dépendant de l'initialisation des classes. Nous avons ici opté pour une arborescence hiérarchique, même si celle-ci n'autorise pas de retour en arrière (même mal calculée, une classe ne peut être reconsidérée).

Les phonèmes des six langues ont ainsi été groupés par classement automatique hiérarchique [6]. Dans la phase d'initialisation de l'algorithme, chaque phonème est assigné à une classe. Après chaque itération, les deux classes de mesure de similitude maximale sont regroupées. La procédure se répète jusqu'à l'obtention du nombre de classes désiré. À noter que certaines classes peuvent contenir plusieurs phonèmes d'une même langue, car aucune contrainte sur l'origine linguistique des phonèmes n'a été prise en compte[1]

[1] Notre approche entre donc dans le type de classement maximal (inter-langue + intra-langue) proposé par Berkling [3].

2.2. *La mesure de similitude*

La définition d'une mesure de similitude (ou dissimilitude) entre phonèmes ou allophones est un sujet qui a été largement traité par la communauté scientifique. Young [18] a défini la dissimilitude entre allophones en exprimant la divergence de deux gaussiennes en fonction de leur moyenne et de leur variance. Cependant, cette approche est seulement applicable à des modèles markoviens à un seul état et une seule gaussienne. En remplaçant le concept de divergence par la distance de Bhattacharyya, Mak [12] a proposé une expression alternative de la dissimilitude en fonction des paramètres des modèles. Cette approche est applicable à des modèles markoviens avec mélange de gaussiennes, mais à un seul état.
La comparaison de deux modèles markoviens à plusieurs états et mélange de gaussiennes requiert une mesure de similitude (ou dissimilitude) fondée sur la vraisemblance des données acoustiques (coefficients cepstraux) par rapport aux modèles markoviens. Dans cette ligne de travail, Juang [10] et Köhler [11] ont proposé une mesure de dissimilitude définie comme la différence des vraisemblances acoustiques. La mesure ainsi définie est dérivée du concept de divergence entre modèles markoviens.
Dans le cadre du travail présenté dans cet article, nous avons utilisé une mesure de similitude fondée sur le concept d'information mutuelle [14]. Cette mesure constitue une approximation de la probabilité a posteriori $\Pr(\lambda_i|\vec{\varphi}_j)$:

$$S(\varphi_i,\varphi_j) = f(\vec{\varphi}_j|\lambda_i)^\gamma / \sum_{k=1}^{n} f(\vec{\varphi}_j|\lambda_k)^\gamma \qquad (1)$$

où $\vec{\varphi}_j$ correspond aux données acoustiques du phonème φ_j ; λ_i est le modèle markovien du phonème φ_i ; λ_k est le modèle markovien de chacun des phonèmes à classifier ; f est la fonction de densité de probabilité des observations, connaissant les données d'apprentissage ; et n est le nombre d'unités phonétiques. Le coefficient γ a été introduit pour compenser l'hypothèse d'indépendance entre modèles (il a été fixé à 0,5 de façon empirique). Cette mesure étant asymétrique, nous avons utilisé une version symétrique, en prenant la moyenne des mesures de similitude :

$$S_s(\varphi_i,\varphi_j) = \frac{S(\varphi_i,\varphi_j) + S(\varphi_j,\varphi_i)}{2} \qquad (2)$$

Dans le cas où les classes contiennent plus d'un phonème, le concept de mesure de similitude phonétique s'étend à celui de mesure de similitude interclasse. Cette mesure est définie comme la moyenne des similitudes entre phonèmes[2] :

$$S(C_i,C_j) = \frac{1}{n_i n_j} \sum_{\varphi \in C_i} \sum_{\varphi' \in C_j} S_s(\varphi,\varphi') \qquad (3)$$

où n_i et n_j sont les nombres de phonèmes dans les classes C_i et C_j respectivement, et $S_s(\varphi,\varphi')$ est la mesure de similitude phonétique interclasse.

[2] La mesure de similitude entre deux classes peut aussi être calculée comme le maximum ou le minimum des similitudes entre phonèmes [6].

3. CADRE EXPÉRIMENTAL

3.1. *Description des corpus*

L'algorithme de classement hiérarchique a été appliqué aux enregistrements du corpus IDEAL [4] et du corpus SPEECHDAT. Le corpus IDEAL a été conçu pour le développement de systèmes d'identification automatique de la langue. Il contient plus de 300 appels téléphoniques en français, anglais, allemand et espagnol. Le contenu linguistique d'un appel est très varié : 18 phrases lues (phrases de journal, phrases phonétiquement équilibrées, dates, heures, adresses, etc.), 12 questions d'ordre général (code de l'appel, sexe, âge, code postal, etc.) et 6 questions destinées à recueillir de la parole spontanée. Dans le cadre de ce travail, nous avons sélectionné environ 7500 phrases par langue, dont 40% sont lues et 60% constituent des réponses aux questions d'ordre général. Le corpus téléphonique SPEECHDAT a lui été construit pour le développement de systèmes de reconnaissance de la parole dans plusieurs langues européennes. En ce qui concerne l'italien et le portugais, nous avons sélectionné environ 9500 phrases phonétiquement équilibrées par langue.

3.2. *Définition des alphabets phonétiques*

Les unités phonétiques définissant chaque langue ont été sélectionnées différemment. Pour le français, l'anglais et l'allemand, nous avons repris les jeux de phonèmes utilisés par le LIMSI dans le cadre de la reconnaissance de la parole continue de ces langues (34, 44 et 46 phonèmes pour le français, l'anglais et l'allemand respectivement) [7, 1]. Les 24 phonèmes de l'espagnol constituent l'ensemble de base traditionnellement retenu [16]. Enfin pour l'italien et le portugais les jeux phonétiques sont issus de SAMPA [8] (49 et 38 phonèmes pour l'italien et le portugais respectivement). Au total, le nombre d'unités phonétiques est de 235 ; le tableau 2 illustre le nombre de phonèmes par langue et par catégorie.

Table 1: nombre de phonèmes par langue (français, anglais, allemand, espagnol, italien, portugais) et par catégorie (V=voyelles, C=consonnes, syl.=syllabiques).

	Fr.	*An.*	*Al.*	*Es.*	*It.*	*Po.*
V simples	14	14	19	5	7	14
diphtongues	-	6	3	-	-	2
glides ou C syl.	3	2	4	-	2	2
C simples	17	20	20	17	17	20
affriquées	-	2	-	1	4	-
géminées	-	-	-	1	19	-
total	34	44	46	24	49	38

3.3. *Modélisation phonétique*

Chaque phonème est caractérisé au moyen d'un Modèle de Markov Caché à trois états indépendant du contexte (CI-CDHMM: Context Independent Continuous Density Hidden Markov Model). À chaque état, la loi d'émission d'observation est définie comme la somme pondérée de 32 gaussiennes. Pour chacune des langues, l'ensemble des

phrases a été phonétiquement aligné. Cette segmentation a été obtenue en utilisant la transcription orthographique des phrases, le dictionnaire des mots sous forme graphémique et phonémique, et les modèles phonétiques.
À partir des vecteurs cepstraux correspondant aux 235 phonèmes, et à partir des leurs modèles markoviens associés, nous avons calculé la matrice de vraisemblances acoustiques. Avant d'effectuer cette opération, nous avons dû limiter le nombre de vecteurs acoustiques par phonème à moins de 20 000 échantillons (7 000 vecteurs cepstraux en moyenne[3]), pour réduire le temps de calcul de la matrice de vraisemblance. Une fois cette matrice calculée, nous avons appliqué l'algorithme de classement hiérarchique en faisant varier le nombre de classes désirées.

4. ANALYSE DES RÉSULTATS

4.1. *Réduction du nombre de classes*

On peut regrouper de diverses façons les phonèmes d'une langue, a fortiori de plusieurs. Il est classique, en phonétique, d'adopter une représentation sous la forme d'une arborescence plus ou moins hiérarchisée, reflétant un degré d'analyse plus ou moins profond. Ceci est possible avec notre algorithme, où l'on choisit à l'avance le nombre de classes, en faisant décroître ce nombre. Avec 2 classes, on retrouve en gros une bipartition voyelle-consonne ; avec une dizaine, on retombe à peu près sur les groupes suivants, qui s'excluent mutuellement : voyelles ouvertes, voyelles fermées, fricatives, occlusives, nasales, liquides, glides. Même si certains phonèmes sont réfractaires au regroupement, et même si d'autres regroupements ne sont pas ceux que nous souhaiterions, il est intéressant d'observer la pertinence linguistique des classes produites par un calcul sur des traits purement acoustiques (les coefficients cepstraux).
Toutefois, notre but ici se situe à un niveau d'abstraction moindre : nous souhaitons perdre le moins possible d'information phonématique. Nous ne prétendons pas définir formellement ce concept, qu'il est difficile d'évaluer de manière objective. Le phonème et ses allophones sont traditionnellement définis pour une langue donnée, par l'analyse en paires minimales de commutation, liée à la fonction distinctive. Des recommandations et des conventions sont publiées chaque année par des spécialistes, pour décider d'inclure officiellement ou non des symboles et des diacritiques dans l'Alphabet Phonétique International (API). En même temps, ce n'est pas un hasard, intuitivement, si des sons voisins, dans plusieurs langues, sont représentés avec le même symbole API [15]. C'est cette proximité que l'algorithme décrit ci-dessus permet d'explorer objectivement.
Partant des 235 phonèmes de notre ensemble initial (pour le français, l'anglais, l'allemand, l'espagnol, l'italien et le portugais), on a très rapidement des géminées de l'italien qui se regroupent avec les consonnes simples correspondantes, mais ce n'est que vers 90 classes qu'émerge une classe /t/, commune aux 6 langues. Avec 80 classes, les /z/ se retrouvent regroupés, mais le /o/ est rattaché au /u/. Avec la part d'arbitraire que cela comporte (mais il faut trouver un compromis), c'est le regroupement en 90 classes que nous allons maintenant décrire plus en détails. De surcroît, 90 est grosso modo le nombre de symboles API différents qui sont utilisés pour décrire nos 6 langues.

4.2. *Interprétation des 90 classes de phonèmes*

On compte 35 singletons (correspondant à presque autant de voyelles que de consonnes) : des exemples en sont le /ɥ/ français, le /ð/ et le /ʌ/ anglais, le /ç/ allemand, qui sont propres à leurs langues. Parmi les 55 classes restantes, reportées dans le tableau 2, certaines rassemblent uniquement 6 phonèmes de symbole API identique à travers les 6 langues (plus un 7^{e} pour les consonnes géminées de l'italien) : /i/, /s/, /m/, /t/, /p/, /k/. Dans ces chiffres, il faut bien sûr compter avec le bruit introduit initialement par des jeux de phonèmes hétérogènes (notamment en ce qui concerne les traits de durée, les diphtongues, les glides, les affriquées, les géminées), plus ou moins fins en allophones (vocaliques). Les affriquées et les diphtongues se voient souvent isolées - même si les /aɪ/ allemand et anglais sont regroupés. Notre modèle ne permet pas de déterminer s'ils s'apparieraient avec 2 segments (par exemple /t/ et /ʃ/ pour /ʧ/).
Les classes dérivées automatiquement correspondent en général à des unités linguistiques identiques ou similaires. Il convient néanmoins de mentionner que 12 d'entre elles sont des paires de phonème d'une même langue : par exemple, le /z/ et le /v/ anglais. Ce cas peut être expliqué par le fait que l'essentiel de l'énergie du [z] est concentrée au dessus de 4 kHz. Par conséquent, le filtrage effectué par la bande téléphonique nous prive d'une grande part de l'information. Inversement, les /u/, /w/, /j/, /l/, /g/, ainsi que les fricatives sonores (notamment /ʒ/) ne sont pas tous rassemblés.

5. CONCLUSION

Nous avons appliqué une approche de classement automatique aux phonèmes de six langues (français, anglais, allemand, espagnol, italien et portugais). Cette méthode est fondée sur un algorithme de classement hiérarchique, qui seul permet un suivi de l'évolution d'un groupement agglomératif. À partir d'une initialisation avec 235 phonèmes, nous avons itéré l'algorithme tout en analysant la pertinence linguistique des groupes obtenus à chaque étape. Nous avons retenu le nombre de 90 classes, afin de ne pas perdre trop d'information phonématique. Ces classes correspondent en majeure partie à une réalité linguistique.
Nous avons aussi analysé les résultats du classement dans les niveaux les plus bas de l'agglomération hiérarchique - où l'on retrouve la dichotomie voyelle-consonne. Ceci est un autre facteur qui démontre l'efficacité de cette méthode. Nous devons aussi noter la capacité de cette approche à traiter des données de différentes origines. De l'analyse des résultats obtenus à chaque itération, nous pouvons observer qu'il n'y a pas eu un classement des phonèmes par corpus, ni un bon classement pour les phonèmes de certaines langues et mauvais pour les phonèmes des autres. Ceci semble démontrer la robustesse de la mesure de

[3] Certains phonèmes en effet, comme les géminées de l'italien, ont peu de représentants. Pour ces phonèmes rares, la validité des modèles statistiques peut être remise en question.

similitude, face aux variations acoustiques des conditions d'enregistrement, et au contenu linguistique des différents corpus d'apprentissage. De fait, la fonction de densité de probabilité correspondant aux vecteurs acoustiques du phonème φ_j, dans l'expression de la mesure de similitude, effectue une normalisation sur toutes les données acoustiques. Cette approche a été appliquée dans le cadre de l'identification automatique de 4 langues (français, anglais, allemand, espagnol). Les résultats obtenus pour le calcul des classes de phonèmes de six langues permettent l'extension du décodeur, de quatre à six langues. On peut même envisager d'utiliser les modèles acoustiques correspondant aux classes de phonèmes des langues connues, pour la reconnaissance phonétique d'une nouvelle langue, dont on dispose uniquement de corpus de parole sans transcription phonétique.

Quantifier la proximité entre les phonèmes et les allophones d'une même langue peut contribuer à la définition d'un ensemble économique de phonèmes. Regrouper les phonèmes de plusieurs langues trouve enfin des applications en synthèse vocale multilingue, et en phonétique descriptive, didactique ou corrective.

Table 2: regroupements d'au moins deux phonèmes pour 90 classes, à travers les 6 langues étudiées (français, anglais, allemand, espagnol, italien, portugais) - une classe par ligne.

Fr.	*An.*	*Al.*	*Es.*	*It.*	*Po.*
i	i	i	i	i	i
e	-	e: ɪ eɪ	-	-	-
ɛ	-	ɛ: ɛ	e	e	ɛ
y	-	y	-	-	-
œ ø	-	-	-	-	-
ə	-	ə	-	-	-
a	æ ɑ: aʊ	a ɑ	a	a	a
ɔ	ɒ	ɔ	o	ɔ o	ɔ
o	ɔ:	o	-	-	o
u	-	-	u	u	u w
-	-	-	-	-	e ẽ
-	-	-	-	-	ĩ j̃
ɑ	-	-	-	-	ɐ
ɔ	-	-	-	-	ũ õ
-	ɜˑ ə	-	-	-	ɐ
-	ɪ e	-	-	-	-
-	ɛ ɛə	-	-	-	-
-	aɪ	aɪ	-	-	-
-	-	ɐ̯ əʁ	-	-	-
-	-	u ʊ	-	-	-
w	-	-	-	w	-
j	-	-	-	j	j
-	-	aʊ	-	-	i̯
s	s	s	s	s ss	s
z	-	-	-	z	z
ʃ ʒ	ʃ	ʃ	-	ʃ ʃʃ	ʒ
f	f	f	f θ	f ff	f
v	-	-	-	v	v
n	n	n ən	n	n	-
-	-	-	-	nn	n
m	m	m	m	m mm	m
ɲ	-	-	ɲ	ɲ ɲ	ɲ
l	-	l	l	l ll	l
ʁ	-	ʁ	-	-	-
p	p	p	p	p pp	p
t	t	t	t	t tt	t
k	k	k	k	k kk	k
g	g	g	-	-	-
-	b	b	-	-	-
-	d	d	-	-	-
-	-	-	b	b bb	b
-	-	-	d	d dd	d
b d	-	-	-	-	-
-	-	-	g	g	g
-	ʤ ʧ	-	-	-	-
-	-	-	-	ʤ ʤʤ	-
-	-	-	-	ʧ ʧʧ	-
-	-	-	-	ts tsts	-
-	z v	-	-	-	-
-	ŋ	ŋ	-	-	-
-	l	-	-	-	l
-	h	h	-	-	-
-	-	-	x	-	R
-	-	-	ʃ ʎ	λλ	ʎ
-	-	-	r r̄	r rr	r

REFERENCES

[1] M. Adda-Decker, G. Adda, L. Lamel, J.L. Gauvain, "Developments in large Vocabulary Continuous Speech Recognition of German", *ICASSP-96*.

[2] K.M. Berkling, E. Barnard, "Language Identification of six Languages Based on a Common Set of Broad Phonemes", *ICSLP-94*.

[3] K.M. Berkling, *Automatic Language Identification with Sequences of Language-Independent Phoneme Clusters*, PhD thesis, OGI, 1996.

[4] C. Corredor-Ardoy, J.-L. Gauvain, M. Adda-Decker, L. Lamel, "Language Identification With Language-Independent Acoustic Models", *Eurospeech-97*.

[5] P. Delattre, *Comparing the phonetic features of English, German, Spanish and French*, Julius Gross Verlag, Heidelberg, 1965.

[6] R.O. Duda, P.E. Hart, *Pattern Classification and Scene Analysis*, Wiley-Interscience, 1973.

[7] J.L. Gauvain, L. Lamel, G. Adda, M. Adda-Decker, "Speaker-independent Continuous Speech Dictation", *Speech Communication 15*, pp. 21-37.

[8] D. Gibbon, R. Moore, R. Winski (eds.), *Handbook of Standards and Resources for Spoken Language Systems*, Mouton de Gruyter, Berlin, 1997.

[9] A.G. Haudricourt, J.M.C. Thomas, *La notation des langues. Phonétique et phonologie*, Imprimerie de l'Institut Géographique National, Paris, 1967.

[10] B.H. Juang, L.R. Rabiner, "A Probabilistic Measure for Hidden Markov Models", *AT&T Technical Journal* 64(2), 1985.

[11] J. Köhler, "Multi-lingual Phoneme Recognition Exploiting Acoustic-Phonetic Similarities of Sounds", *ICSLP-96*.

[12] B. Mak, E. Barnard, "Phone Clustering using the Bhattacharyya Distance", *ICSLP-96*.

[13] Y.K. Muthusamy, R.A. Cole, B.T. Oshika, "The OGI multi-language telephone speech corpus", *ICSLP-92*.

[14] J.G. Proakis, *Digital Communications*, McGraw-Hill Series in Electrical Engineering, 1989.

[15] G.K. Pullum & W.A. Ladusaw, *Phonetic Symbol Guide*, The University of Chicago Press, Chicago and London, 1996.

[16] A. Quilis et J.A. Fernández, 1992, *Curso de fonética y fonología españolas*, Consejo Superior de Investigaciones Científicas, Madrid, 1992.

[17] N. Vallée, L.J. Boë, C. Abry, J.L. Schwartz, A. Berrah, "La matérialité des structures sonores du langage", *JEP-96*.

[18] S.J. Young, P.C. Woodland, "The use of state tying in continuous speech recognition", *Eurospeech-93*.