ETIQUETAGE AUTOMATIQUE DU SIGNAL DE PAROLE CONTINUE A L'AIDE DE LA VARIATION RELATIVE D'ENERGIE DES SEQUENCES DE PHONEMES

DESI M. RINGOT P. ANDREEWSKY A.

CNRS - LIMSI - ORSAY BP 30 91406 ORSAY CEDEX FRANCE

A threshold-free system for automatic labelling of speech signal is described. Mainly we transform the phonetic strings into energetic strings, using context-based rules or a square matrix which formalises the relative variation of energy between any two phonemes. For 700 sentences database, 95% of the labels are well matched. The adaptation, for other languages is easy.

L'étiquetage automatique, c'est à dire l'attribution de valeurs phonétiques aux spectres obtenus à partir d'un signal de parole, a pour but d'amorcer la phase d'apprentissage indispensable pour effectuer un décodage acoustico-phonétique performant pour la reconnaissance de la parole continue sur des vocabulaires étendus. Dans le présent travail on rappelle d'une part les idées générales de l'étiquetage automatique du système SHERPA et on expose une nouvelle version du module de calcul du profil théorique de la courbe d'énergie lorsque l'on connait la chaîne phonétique correspondante. Cette modification a comme intérêt d'une part de mieux exprimer la théorie sous jacente à l'étiquetage dans SHERPA et, d'autre part, de permetre une extension plus facile de cette approche à d'autres langues.

On donne les résultats de l'étiquetage automatique sur 700 phrases utilisant un vocabulaire de 2000 mots différents.

I. TRANSFORMATION DE LA CHAINE PHONETIQUE EN UNE SUITE D'ALTERNANCES PHONETIQUES, A L'AIDE DE REGLES CONTEXTUELLES.

Les opérations suivantes sont effectuées:

I.1. Définition des classes phonétiques.

Elle repose sur le principe suivant: deux phonèmes qui ont, dans un contexte phonétique identique, un comportement identique sur la courbe d'énergie, appartiennent à la même classe.

O (occlusives) : /p,t,k,b,d,g/
F (fricatives) : /f,v/
S (sifflantes) : /s,z/
X (chuintantes) : /ʃ,ʒ/
N (nasales) : /m, n, ñ/
L (liquides) : /l, r/
I (semi-voyelles) : /y, w, u/
V (voyelles) : /a, e, ɛ; i,ɔ̃ o, u, y, ø, oe, ɛ̃, ɑ̃, e muet/
DEB, FIN

I.2. La chaîne phonétique prononcée est transformée en une chaîne de classes.

I.3. On écrit des règles phonétiques qui, à la chaîne de classes, va faire correspondre une suite de minimums (min), de maximums (Max) et d'alternances secondaires (Alts), à partir de laquelle on va calculer les paramètres de lissage de la courbe d'énergie, afin de transformer cette dernière en une suite de min, Max, Alts, ayant une interprétation phonétique.

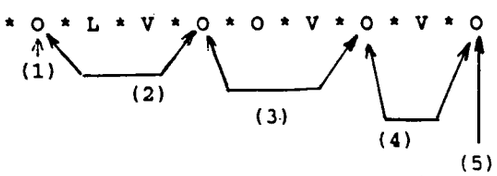
I.4. Les règles phonétiques opèrent sur les classes et ont par exemple la forme suivante: (le symbole "*" signifie: suivi de)

1) Une occlusive en début d'énoncé prend la valeur 0.
2) Si O1*L*Y*O2
   alors O1=0; L=½; V=1; O2=0
3) Si O1*O2*V1*O3
   alors O1,O2=0 ∂ 0; V1=1
4) Si O1*V1*O2 alors O1=0; V1=1
5) Une occlusive en fin d'énoncé prend la valeur 0.

Appliquées à la chaîne phonétique correspondant à "promptitude",

* p * r * ʃ * p * t * i * t * y * d

* O * L * V * O * O * V * O * V * O



ces règles nous donnent la suite de 0,1,1/2,à suivante:

0 1/2 1 0 ⓐ 0 1 0 1 0

que nous appelerons "chaîne des contrastes". Cette chaîne fournit le nombre théorique d'alternances qu'il faut conserver sur la courbe d'énergie et par conséquent le nombre de fluctuations qu'il va falloir lisser sur cette courbe.

Ces règles contextuelles, au nombre de 2000 environ, attribuent des valeurs du type min aux occlusives, Max aux voyelles, min, Max ou Alts aux autres classes de phonèmes en fonction du contexte phonétique immédiat de gauche ou de droite. Les étiquettes attribuées peuvent dans certains cas regrouper plusieurs étiquettes simples (/yʃ/).

Les règles transforment toute chaîne phonétique, quelqu'en soit la longueur, en une suite d'alternances du type (0,1,0) qui correspondent à des pseudo-syllabes (dont la définition ne se superpose pas avec celle de la syllabe classique).

Comme on l'a vu dans l'exemple ci-dessus, les valeurs ½ et ⓐ peuvent s'insérer dans l'alternance (0,1,0), ce qui correspond à des fluctuations intrasyllabiques possibles. Le nombre d'alternances (0,1,0) et le nombre total de ½ et de ⓐ sont déterminants pour la procédure de lissage.

II. TRANSFORMATION DE LA CHAINE PHONETIQUE A L'AIDE D'UNE MATRICE DES CONTRASTES D'ENERGIE DES PHONEMES.

L'approche par règles contextuelles suppose une étude exhaustive des contextes phonétiques, assez longue à mettre en place pour une langue donnée et son adaptation ensuite à d'autres langues reste complexe. Nous exposons une méthode différente qui repose sur une meilleure définition théorique du problème, et qui est plus facilement généralisable.

Le principe de base consiste à utiliser l'évolution de l'énergie d'un phonème à l'autre, au cours de l'émission d'une chaîne phonétique continue.

Dans cette approche, les voyelles se situent toujours aux maximas d'énergie. Deux voyelles ou plus, qui se suivent sans hiatus, sont regroupées en un seul maximum. Tant que, à partir d'une voyelle ou d'un groupe de voyelle, l'énergie théorique des phonèmes successifs décroit, on est toujours dans le maximum. Par exemple, dans /artist/, le premier maximum est constitué par /ar/. Dès que l'énergie se met à croître, c'est que l'on est passé par un minimum (dans /artist/, on décroit à partir de /a/ en passant par /r/ pour aller jusqu'à /t/ et on remonte à partir du /t/ qui est un minimum énergétique). A partir du minimum d'énergie, et jusqu'au maximum vocalique suivant il peut s'insérer des alternances secondaires (½ ou ⓐ): occlusive suivie de liquide, occlusive suivie d'occlusive, occlusive suivie de n'importe quelle consonne.

Pour l'application de ces principes, on a utilisé une matrice carrée de dimension 10: 8 classes phonétiques et 2 symboles de début et fin d'énoncé.

Nous allons nous contenter de donner la partie de la matrice indispensable pour traiter la chaîne phonétique "promptitude" /prʃptityd/

	O	L	V	...	DEB	FIN
O	ⓐ	½	+	...	ø	ø
L	-	-	+	...	ø	ø
V	-	-	=	...	ø	-
...				...		
DEB	ø	ø	+	...	ø	ø
FIN	ø	ø	ø	ø	ø	ø

La signification de ces symboles est simple:

- Sur la première colonne (par laquelle on entre dans la matrice), on trouve la classe du phonème de gauche.
- Sur la première ligne, on trouve la classe du phonème de droite.
- "+" signifie que le phonème de droite a une énergie supèrieure à celle du phonème de gauche.
- "-" signifie que le phonème de droite a une énergie infèrieure à celle du phonème de gauche.
- "½" signifie que l'on peut avoir une alternance secondaire du type ½ entre phonème de gauche et phonème de droite.
- "=" signifie que les phonèmes de gauche et de droite ont une énergie similaire au sens de l'algorithme.
- "ø" correspond à une séquence impossible.

Si on utilise cette matrice pour transformer la chaîne phonétique /prʃptityd/, ou plus exactement sur la chaîne de classes phonétiques correspondante (O L V O O V O V O), on commence par attribuer la valeur 0 aux occlusives et la valeur 1 aux voyelles; ensuite on attribue à L dans O L V la valeur ½ et on écrit un à entre les deux occlusives de la séquence O O; on obtient finalement la séquence :

0 ½ 1 0 ⓐ 0 1 0 1 0

III. LE LISSAGE.

L'objectif du lissage est de localiser les spectres correspondants aux étiquettes des minimums (consonnes ou groupes consonnantiques) et aux étiquettes des maximums (voyelles ou groupes vocaliques). Les spectres correspondants aux fluctuations intrasyllabiques sont délibérément lissés mais ensuite, éventuellement, étiquetés après un réexamen de l'aspect de la courbe à l'intèrieur de la pseudo-syllabe.

Pour cela on effectue un double lissage, dont l'intensité est guidée par la structure de la chaîne des contrastes, d'abord sur l'axe des énergies puis ensuite sur celui des temps. L'importance relative des lissages énergétique et temporel peut être paramètrée.

Remarque: il est indispensable d'effectuer un lissage sur les deux axes. En effet un lissage sur l'énergie seule risquerait de gommer les fluctuations liées à des pseudo-syllabes peu contrastées sur l'énergie (par exemple: /si/, /my/). Un lissage sur l'axe des temps, outre le fait qu'il pallie à l'inconvénient précédent, permet d'éliminer des fluctuations petites sur l'axe des temps mais parfois relativement importantes sur l'axe des énergies (par exemple l'explosion du /k/).

Le lissage s'effectue donc en deux temps et utilise deux nombres fournis par la chaîne des contrastes; le premier correspond au nombre total de pseudo-syllabes et donne le nombre d'extremums qu'il convient de conserver après les deux lissages; le deuxième correspond au nombre de fluctuations intrasyllabiques possibles et est déterminant pour évaluer l'importance relative des deux lissages. Cette importance relative est également régulée par un paramètre d'ajustement, si les critères de contrôle (Cf ci-dessous) de la bonne qualité de l'étiquetage ne sont pas vérifiés.

Le double lissage est obtenu par suppression itérative des fluctuations énergétiques puis temporelles jusqu'à obtention du nombre théorique d'extremums.

Soulignons que dans cette procédure de lissage, il n'est fait appel à aucun seuil ni sur l'énergie, ni sur le temps, ce qui représente un facteur de portabilité intra et multilocuteurs très important.

IV. PROCEDURE D'ETIQUETAGE.

La procédure d'étiquetage est relativement simple.

Dans un premier temps, on attribue aux extremums sélectionnés par le lissage, des étiquettes phonétiques consonnantiques (simples ou multiples) aux minimas et vocaliques (simples ou multiples) aux maximas. La valeur et l'ordre des différentes étiquettes sont donnés par la procédure de transformation des chaînes phonétiques.

Dans un deuxième temps, on essaye, à l'intèrieur même de la pseudo-syllabe, de dissocier les étiquettes multiples chaque fois que la présence d'une fluctuation énergétique intrasyllabique rend cela possible. Par exemple, l'étiquette complexe /pr/ dans la pseudo-syllabe /prʃ/, pourra être dissociée en /p/ et /r/ si entre les extremums correspondant au min: /pr/ et max: /ʃ/, il existe une fluctuation de la courbe. Il existe ainsi, compte tenu de l'énergie relative des phonèmes constituant des étiquettes multiples, plusieurs cas de figure que nous ne détaillerons pas ici.

V. PROCEDURES DE CONTROLE.

Ces procédures détectent au sens de certains critères, les phrases qui présentent un risque d'étiquetage défectueux. On propose alors une solution de réétiquetage en faisant varier le paramètre d'ajustement qui contrôle l'importance relative des lissages énergétiques et temporels.

La qualité du nouvel étiquetage est à son tour vérifiée sur l'ensemble des critères et il est remis éventuellement en cause. On effectue ainsi au plus six tentatives (le paramètre d'ajustement varie six fois); si à la sixième tentative les critères ne sont toujours pas vérifiés, la phrase est automatiquement rejetée du corpus d'apprentissage.

Actuellement seuls deux critères sont opérationnels. L'un vérifie que les N occlusives d'un énoncé sont placées sur les N minimas les plus bas de la courbe d'énergie; il détecte les erreurs globales d'étiquetage. L'autre vérifie que deux étiquettes consécutives sont séparées par un minimum de deux spectres; ce critère détecte la plupart des erreurs purement locales.

VI. RESULTATS.

L'étiquetage automatique a été testé sur un corpus de 700 phrases de longueur variable de 5 à 10 mots. La qualité de l'étiquetage est évaluée par rapport aux performances de l'étiquetage manuel par un phonéticien; les occurences d'un même mot à des parties différentes du corpus ont la même courbe d'énergie et les étiquettes correspondantes sont placées aux mêmes endroits de la courbe.

Phrases rejetées (critères de contrôle non vérifiés): 15%.
Etiquettes bien placées: 95%.
Etiquettes complexes: 15%.
complexes dissociées: 50%

Les 5% d'erreurs d'étiquetage ont peu de répercussion sur l'ensemble du système car les autres modules de l'apprentissage comportent des contrôles internes qui permettent de les détecter.

L'adaptation de ce système d'étiquetage à une autre langue que le français est simple. Il suffit de modifier le contenu de la matrice de transformation des chaînes phonétiques, en fonction du système phonétique de la langue. L'adaptation est en cours de réalisation pour l'espagnol et l'italien.

BIBLIOGRAPHIE

ANDREEWSKY A. DESI M. FLUHR C. POIRIER F. "Une méthode de mise en correspondance d'une chaîne phonétique et de sa forme accoustique", 11ème ICA, Revue d'Accoustique, 1983, p. 245.

ANDREEWSKY A. DESI M. POIRIER F. "Le système SHERPA -de l'étiquetage phonétique automatique à la reconnaissance par analyse ternaire", 5ème Congrès RFIA, 1985, p.

DESI M. POIRIER F. "Le système SHERPA: étiquetage et classification automatique par apprentissage pour le décodage automatique et la parole continue", Thèse de Doctorat en Sciences, Paris-Sud Orsay, 1985.

LENNIG N. "Automatic alignment of natural speech whith a corresponding transcription", Speech communication, 1983, p.190-192.

MERCIER G. "Accoustic-phonetic decoding and adaptation in continuous speech recognition", Automatic Speech Analysis and Recognition, Reidel Publishing Co, 1982.

WAGNER M. "Automatic labeling of continuous speech whith a Given Phonetic Transcription using Dynamic Programming Algorithms", IEEE Acoustics Speech and Signal Processing, Catalog N°81CH1610-5, 1981, p.1156-1159.