Découpage prosodique sur différents types de segmentations phonémiques

Natalia Segal, Katarina Bartkova

France Télécom R&D/TECH/SSTP
2 avenue de Pierre Marzin, 22307 Lannion, France
{natalia.segal; katarina.bartkova}@orange-ftgroup.com

ABSTRACT

This paper presents the assessment of prosodic boundary detection algorithms using the prosodic structure representation in the form of trees, with different types of phonemic segmentations. Two types of prosodic boundary detection algorithms were studied, first using linguistic and prosodic information and second using only prosodic information [5, 6]. The algorithms were applied to different kinds of phonemic segmentations in order to find out the limits of their applicability to various automatic speech processing tasks. We analyzed the degradation of performance according to phonetic segmentation quality, including manually verified segmentation, automatic alignment and phonemic decoding using a phoneme trigram. We also evaluated and compared the two algorithms on a larger spontaneous speech data base with automatic alignment.

Keywords: Prosody, intonation, spontaneous speech, prosodic segmentation.

1. INTRODUCTION

La détection automatique des frontières prosodiques à partir du signal s'avère importante pour différents domaines du traitement automatique de la parole. En reconnaissance de la parole un découpage du signal en unités prosodiques permet d'éliminer les candidats incorrects ou de restreindre l'espace de recherche. En synthèse de la parole à partir du texte, les bases de données de tailles importantes utilisées lors de la sélection d'unités doivent être étiquetées prosodiquement car la valeur des ces paramètres est prise en compte lors du choix des unités.

Les études concernant le découpage prosodique du signal de parole visent soit la transcription prosodique complète (codage), soit la modélisation automatique de certains évènements prosodiques.

En ce qui concerne le codage, il s'agit soit d'une annotation formelle des évènements prosodiques de surface [1] sans liaison directe avec le découpage et la modélisation des unités prosodiques, soit d'un codage des évènements prosodiques linguistiques (accent, frontière des unités prosodiques), mais qui est difficile à automatiser et très dépendant de l'organisation prosodique de la langue [7]. Pour le français, les systèmes de ce genre sont peu développés.

La modélisation statistique des évènements prosodiques concerne généralement des applications très limitées, telles que la modélisation des erreurs de prononciation ou le découpage du flux de parole en sujets de conversation [2]. Ce type d'études est également très peu développé pour le français.

Nous avons cherché à développer un algorithme de découpage en unités prosodiques qui pourrait en même temps représenter de façon adéquate la prosodie du français et être complètement automatique.

Nous avons développé, à partir d'une représentation de la structure prosodique de la langue française, élaborée pour la parole préparée [3], deux algorithmes de découpage automatique de la parole spontanée en mots prosodiques fonctionnant avec ou sans connaissances lexicales. L'applicabilité de ces deux algorithmes à la parole spontanée et aux données en sortie de la reconnaissance a été testé dans [5, 6]. L'approche avec connaissances lexicales s'est avérée avantageuse pour le découpage prosodique des données ayant une bonne qualité de segmentation phonémique car elle permet d'utiliser les contraintes lexicales. Mais elle est également plus sensible à la dégradation de la qualité de segmentation (en cas d'erreurs multiples dans les hypothèses des mots et des phonèmes), tandis que l'approche sans connaissances lexicales s'est avérée plus robuste.

Par la suite, nous avons voulu étudier la dégradation des performances en fonction de différents types de segmentation phonémique dont la qualité est plus ou moins éloignée de la segmentation manuelle pouvant être considérée comme "parfaite". Pour l'étude présentée ici, nous avons testé les données provenant d'une segmentation phonémique manuelle et d'une segmentation issue d'un modèle trigramme de phonèmes. Nous avons étudié la dégradation de performance et effectué une analyse manuelle des erreurs typiques. Ensuite nous avons testé les performances des deux algorithmes de découpage prosodique sur un corpus de données plus conséquent ne comportant pas de segmentation phonémique manuelle.

2. BASES DES DONNEES

Les 2 bases de données utilisées pour tester les approches de découpages prosodiques contiennent des messages en français de longueur variable, enregistrés à travers le réseau téléphonique. Dans les deux cas les

messages ont un caractère spontané. Nous avons supposé que chaque message a été prononcé par un utilisateur différent (homme ou femme).

Le premier corpus (**corpus MC**) contient 83 messages courts (environs 32 mots par message) laissés sur un répondeur par des utilisateurs avertis que leur message allait être enregistré. Il s'agit de messages de communications internes dans une entreprise.

Pour évaluer les performances des algorithmes de découpage prosodique en fonction de la qualité de segmentation phonémique, nous avons utilisé sur le corpus MC trois segmentations phonémiques différentes :

- segmentation phonémique vérifiée manuellement (ce qui explique la taille limitée du corpus) ;
- alignement forcé du signal avec la transcription provenant d'un phonétiseur automatique de texte;
- segmentation automatique par un modèle trigramme de phonèmes.

Le corpus a été également découpé manuellement en mots prosodiques, pour comparer les frontières manuelles et les frontières trouvées automatiquement par notre algorithme de découpage prosodique.

Le deuxième corpus (**corpus ES**) contient les résultats d'une enquête de satisfaction (plus de 1000 messages, environs 54 mots par message).

Le corpus ES a été utilisé pour comparer les performances des deux approches de segmentation prosodique (avec et sans connaissances lexicales) sur un grand corpus de données. L'ensemble de ce corpus était segmenté phonémiquement avec l'alignement forcé. Un sous-ensemble, d'environs 100 fichiers, était segmenté manuellement en mots prosodiques.

3. STRUCTURE PROSODIQUE ET DÉCOUPAGE EN MOTS PROSODIQUES

La représentation théorique de la prosodie de la langue française [3], utilisée comme base de nos algorithmes de découpage prosodique, présente la structure prosodique de chaque phrase comme une hiérarchie qui organise les unités prosodiques sous forme d'arbre. Cette structure prosodique est indépendante de la structure syntaxique bien qu'elle lui soit associée. Les mots prosodiques constituent les feuilles de l'arbre et la structure se manifeste par les paramètres prosodiques sur les syllabes accentuées à la fin des mots prosodiques (notamment par le contraste de pente).

Le mot prosodique (ou groupe prosodique) constitue l'unité prosodique minimale qui contient un accent lexical final et dans certains cas un accent optionnel sur la première syllabe. Un mot prosodique contient au moins un ou plusieurs mots lexicaux (mots de catégories grammaticales ouvertes) et optionnellement les mots grammaticaux rattachés (les mots des catégories fermées).

Nous avons élaboré deux algorithmes de détection de frontières de mots prosodiques : la première approche utilise des connaissances lexicales (hypothèses des mots et leur catégories) alors que la seconde approche fonctionne sans connaissances lexicales et utilise uniquement la segmentation phonémique [5, 6].

Les deux approches utilisent des paramètres prosodiques des syllabes finales (durée vocalique et F0) pour trouver les frontières prosodiques, ainsi que les contraintes rythmiques pour regrouper certains mots prosodiques trop courts (monosyllabiques) et interdire les mots prosodiques trop longs (pas plus de 8 syllabes par mot prosodique).

L'approche utilisant les connaissances de frontières des mots prend en compte en plus quelques contraintes lexicales sur l'organisation du mot accentué.

Les mots prosodiques ainsi identifiés sont organisés en une structure prosodique arborescente dont le nombre de niveaux n'est pas limité. Ici, les mots prosodiques n'ont pas de pattern mélodique standard, mais leurs paramètres prosodiques et les mouvements mélodiques sont imposés par la structure prosodique dont les deux règles principales sont :

- l'Inversion de la Pente Mélodique (**IPM**)
- l'Amplitude de Variation Mélodique (**AVM**)

L'algorithme de la construction d'arbres prosodiques a été présenté en détails dans [5, 6].

4. EVALUATIONS

4.1. Performance selon le type de segmentation phonémique

Le premier axe de tests avait pour but l'évaluation de performances des deux algorithmes de détection des frontières prosodiques en utilisant différents types de segmentation phonémique des données. Nous avons déjà mis en évidence une dégradation de performances entre l'alignement forcé et les résultats de la reconnaissance automatique dans [6].

Pour avoir plus d'information sur le rapport entre la qualité de segmentation phonémique et la performance de l'algorithme, nous avons testé l'approche sans connaissances lexicales sur un corpus de taille limitée (MC) avec trois types de segmentation différents.

Nous avons utilisé comme base de comparaison une segmentation phonémique et prosodique manuelle (SM). Cette segmentation, bien qu'elle puisse contenir quelques erreurs, peut néanmoins être considérée comme une segmentation de référence. Nous avons appliqué à cette segmentation manuelle les deux algorithmes de découpage prosodique (avec et sans connaissances lexicales). Nous cherchions à comparer

notamment les performances des deux approches sur cette segmentation de référence et effectuer une analyse des erreurs commises par les deux approches.

Le deuxième type de segmentation phonémique était réalisé par l'alignement forcé (AF). Le corpus a été transcrit manuellement sous forme orthographique, cette transcription a été ensuite phonétisée (avec variantes possibles) et alignée automatiquement avec le signal de parole. La segmentation par AF est a priori la plus proche de la segmentation phonémique manuelle, car les hypothèses des mots sont vérifiées manuellement.

La troisième segmentation phonémique était obtenue avec un modèle trigramme de phonèmes (PHO). En l'absence de contraintes lexicales, cette segmentation est a priori moins fiable que la segmentation fournie par le système de reconnaissance, par l'alignement forcé, ou la segmentation manuelle. Mais dans certaines applications du traitement automatique de la parole, telle l'indexation du signal, les seules hypothèses fournies par le système de traitement sont les frontières phonémiques. Par conséquent il nous est apparu important d'appliquer notre algorithme de découpage prosodique sans connaissances lexicales à une telle segmentation afin d'évaluer dans quelle mesure la qualité de segmentation phonémique peut influencer la performance de la segmentation prosodique.

Le Tableau 1 résume les taux d'erreur de détection des frontières prosodiques automatiques par rapport aux frontières prosodiques manuelles pour chacune de segmentations phonémiques.

Tableau 1 : Taux d'erreurs selon l'approche de découpage prosodique (ACL – avec connaissances lexicales, SCL – sans connaissances lexicales) et les différents types de segmentation phonémique.

Découpage prosodique	**ACL**	**SCL**		
Segmentation phonémique	**SM**	**SM**	**AF**	**PHO**
Frontières détectées	1104	1267	1253	1335
Frontières insérées	77	206	238	654
Frontières omises	127	93	139	473
Rappel (%)	89	93	88	59
Précision (%)	93	84	81	51
F_1-mésure	91	88	84	55

Le Tableau 1 montre la dégradation des performances entre les différents types de segmentation phonémique des données.

La meilleure performance est naturellement obtenue pour la segmentation phonémique manuelle (segmentation de référence). La majorité des erreurs pour cette segmentation est due à l'algorithme de découpage prosodique et non pas aux fautes de segmentation.

Pour l'alignement forcé, la performance n'est que légèrement dégradée par rapport à la segmentation phonémique manuelle.

Par contre, pour la segmentation fournie par le modèle de phonèmes, la dégradation de performance est importante. Une analyse manuelle des erreurs a montré que la plupart des erreurs est due à une mauvaise détection (mauvaise segmentation) de frontières des voyelles aboutissant à des durées vocaliques incorrectes, or la durée vocalique est un des paramètres prosodiques principaux utilisés par l'algorithme du découpage prosodique. Le modèle trigramme de phonèmes n'est donc pas suffisamment contraignant pour que la segmentation prosodique à la sortie puisse être aussi performante que celle obtenue avec les deux autres types de segmentation phonémique.

Il apparaît que l'approche avec connaissances lexicales présente une très bonne performance dans le cas de la segmentation phonémique manuelle. Les erreurs sont peu nombreuses et leur analyse manuelle a montré qu'elles correspondaient, dans la plupart des cas, aux anomalies importantes soit dans la structure syntaxique des phrases soit dans la prononciation, aux cas où le placement d'une frontière prosodique s'avère difficile même pour un expert humain.

Il est important de souligner que lorsqu'une frontière était détectée par les deux approches (avec et sans connaissances lexicales), dans 95% des cas elle correspondait à une frontière manuelle. Nous pouvons donc réduire le taux d'insertion en utilisant les deux approches simultanément et obtenir ainsi une meilleure précision.

4.2. Evaluation des deux approches de découpage prosodique sur un grand corpus

Dans [6] nous avons comparé les performances de deux algorithmes de découpage prosodique (avec et sans connaissances lexicales) sur les données résultant de l'alignement forcé et sur les données fournies par un système de reconnaissance de la parole. Dans chacun des cas la performance a été évaluée par rapport à la segmentation manuelle en mots prosodiques sur un sous-ensemble du corpus ES (100 fichiers).

Il a été montré que l'approche avec connaissances lexicales donne de meilleurs résultats (notamment une meilleure précision) sur des données provenant de l'alignement forcé, où les hypothèses des mots sont a priori bonnes. Quant à la sortie de la reconnaissance, où les hypothèses des mots peuvent être en partie erronées, l'approche sans connaissances lexicales fournit de meilleurs résultats que l'approche avec connaissances lexicales (la performance générale est naturellement dégradée pour les deux approches par rapport à l'alignement forcé).

Par la suite, nous avons entrepris des évaluations complémentaires, cette fois-ci sur la totalité de la base de données ES segmentée par l'alignement forcé, afin d'estimer si les tendances des performances des deux approches testées (avec et sans connaissances lexicales) restent les mêmes pour l'intégralité du corpus. Comme nous ne disposions pas de découpage prosodique manuel pour la totalité du corpus, nous avons comparé les performances des deux approches entre elles sur le sous-ensemble découpé manuellement en mots prosodiques et sur l'ensemble du corpus ES. Le découpage prosodique provenant de l'approche avec connaissances lexicales a été utilisé ici comme référence pour évaluer l'approche sans connaissances lexicales, car nous avons montré que la précision de l'approche avec connaissances lexicales par rapport au découpage prosodique manuel était meilleure pour ce type de données [6].

Le Tableau 2 présente les performances des deux méthodes de découpage prosodique automatique par rapport au découpage prosodique manuel sur le sous-ensemble du corpus ES.

Tableau 2 : Performance des approches sans connaissances lexicales (SCL) et avec connaissances lexicales (ACL) par rapport au découpage prosodique manuel (sous-corpus ES).

Découpage prosodique	**ACL**	**SCL**
Segmentation phonémique	**AF**	
Frontières détectées	2365	2992
Frontières insérées	261	618
Frontières omises	657	387
Rappel (%)	76	86
Précision (%)	89	79
F_1-mésure	82	82

Le Tableau 3 résume les performances de l'approche sans connaissances lexicales par rapport à l'approche avec connaissances lexicales pour le sous-ensemble et l'ensemble du corpus ES.

Tableau 3 : Performance de l'approche sans connaissances lexicales (SCL) par rapport à l'approche avec connaissances lexicales (ACL).

	Sous-ensemble du corpus ES	**Ensemble du corpus ES**
Nombre de frontières pros. ACL	2365	22539
Nombre de frontières pros. SCL	2992	28910
Nombre de frontières communes pour les deux approches	2127	19937
Précision SCL par rapport à ACL	71	69
Rappel SCL par rapport à ACL	90	88

Il est à constater que la tendance générale de correspondance de frontières entre les deux approches reste stable, avec néanmoins une petite dégradation sur la totalité du corpus. Cela permet de supposer que la performance générale de deux approches obtenues sur la segmentation manuelle peut être extrapolable sur la totalité du corpus.

5. CONCLUSION

L'évaluation du découpage prosodique sur les différents types de segmentation phonémique a montré une forte dépendance entre la qualité de segmentation et la performance de l'algorithme de découpage. La dégradation la plus importante est due à la mauvaise détection des frontières des voyelles et donc de la durée vocalique, le paramètre auquel l'algorithme est très sensible.

Les deux approches de découpage prosodique (avec et sans connaissances lexicales) ont donné un bon taux de détection de frontières de mots prosodiques pour la parole spontanée avec la segmentation phonémique manuelle ou proche de manuelle telle que l'alignement forcé. Ces approches peuvent donc être utilisées pour le découpage prosodique automatique de grands corpus de la parole spontanée.

BIBLIOGRAPHIE

[1] D. Hirst, A. Di Cristo and R. Espesser. Levels of representation and levels of analysis for the description of intonation systems. In *Prosody: Theory and Experiment.* Kluwer Academic Press, Dordrecht, 2000.

[2] Y. Liu, E. Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, M. Harper. Enriching Speech Recognition with Automatic Detection of Sentence Boundaries and Disfluencies. In *IEEE Trans. Audio, Speech and Language Processing*, 14(5):1526-1540, 2006.

[3] P. Martin. Prosodic and rhythmic structures in French. *Linguistics*, 25:925-949, 1987.

[4] M. Rossi. *L'intonation. Le système du français : description et modélisation*. Gap: Ophrys, 1999.

[5] N. Segal and K. Bartkova. Prosodic structure representation for boundary detection in spontaneous French. In *Proceedings of ICPhS'2007*. Saarbrücken, pages 1197-1200, 2007.

[6] N. Segal, P. Martin and K. Bartkova. Prosodic trees for boundary detection in ASR in French. In *Proceedings of Speech Prosody 2008.* Campinas, 2008 (à paraître).

[7] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, J. Hirschberg. TOBI: A standard for labeling English prosody. Dans *Proceedings ICSLP 92*, pages 867-870, 1992.