Détection automatique de frontières prosodiques dans la parole spontanée

Katarina Bartkova, Natalia Segal

France Télécom R&D/TECH/SSTP
2 av. Pierre Marzin, 22307 Lannion Cedex, France
katarina.bartkova@francetelecom.com, natalia.segal@francetelecom.com

ABSTRACT

The present study addresses the issue of the automatic detection of prosodic units in French. An analysis of two prosodic parameters, phone duration and F0 slope values, carried out on two spontaneous speech databases recorded by several thousands of speakers, revealed relevant deviations of these parameters at prosodic junctions. Vowel durations and F0 slopes are used to automatically detect prosodic units at the two data bases. The phone duration is modelled as the ratio of two subsequent vowel durations. Apart from the duration ratio, duration values are also modelled. The detection of prosodic units based on the F0 uses the value of the F0 slope and its standard deviation, recalculated after each pause. An evaluation of the automatic detection is carried out by comparing the prosodic border locations with the lexical boundaries and also with prosodic boundaries obtained by manual segmentation.

1. INTRODUCTION

La détection automatique des événements prosodiques à partir du signal acoustique constitue un pas important vers la construction d'une relation univoque entre le signal acoustique et la représentation abstraite de la prosodie. Plusieurs approches ont été proposées pour construire une telle relation dont l'importance serait primordiale pour toute application en traitement automatique de la parole. Certaines de ces approches sont basées sur des principes perceptifs [6], d'autres sur des principes articulatoires [3]. Toutefois, la plupart des modèles sont entrainés à partir d'un corpus plus ou moins important selon une méthode statistico-probabiliste [4,5].

Des tentatives d'utilisation des paramètres prosodiques en reconnaissance de la parole ont été entreprises avec plus ou moins de succès. Certaines études visaient la détection des unités prosodiques pour réduire l'espace de recherche des candidats lexicaux lors du décodage phonétique du signal de parole [7]. Le but recherché par cette démarche est la diminution de la perplexité de la tâche qui entraîne en général une amélioration des performances. L'utilisation des paramètres prosodiques a également été testée lors du post-traitement pour confirmer ou infirmer les hypothèses de reconnaissance proposées par le décodeur [2]. Même si certains systèmes montrent un progrès considérable, la construction d'un codage prosodique, général et suffisamment fiable pour les applications automatiques, reste à réaliser. Cette tâche est encore plus ardue pour la parole spontanée dont la prosodie est extrêmement variable.

Nous présentons dans cette étude deux techniques de détection des frontières des unités prosodiques à partir du signal acoustique de la parole, destinées à une utilisation en reconnaissance de la parole. L'une des méthodes est basée sur l'utilisation de la durée phonémique et l'autre sur l'évolution de la fréquence fondamentale.

2. CORPUS UTILISÉS

Dans cette étude, nous avons utilisé deux bases de données de la parole continue spontanée enregistrées à travers le réseau téléphonique. Le premier corpus est un corpus d'Enquêtes de Satisfaction (corpus ES) et le second un corpus de Messages Courts (corpus MC) destinés à être transmis sous forme de texto (SMS). Les deux corpus sont constitués de monologues de longueurs variables. Le premier corpus contient environ 1000 enregistrements (~50 mots par message) et le deuxième environ 9000 enregistrements (~18 mots par message). On suppose que chaque enregistrement est prononcé par un nouveau locuteur.

Les deux corpus ont été manuellement retranscrits. La forme phonétique des textes a été par la suite alignée automatiquement avec le signal de parole simulant ainsi la sortie d'une reconnaissance "parfaite", sans erreurs. Cet alignement nous a permis d'accéder aux paramètres prosodiques des phonèmes.

3. ANALYSE STATISTIQUE PRÉLIMINAIRE

L'étude statistique a été focalisée sur la durée syllabique et vocalique ainsi que sur les variations de la courbe de F0, car ces paramètres jouent un rôle primordial dans la perception de la structuration prosodique.

3.1. Analyse de la durée phonémique

Afin d'analyser le comportement de la durée phonémique, nous avons comparé les distributions de la durée normalisée des noyaux syllabiques selon leur position dans le mot et dans des unités prosodiques. Nous considérons ici comme unité prosodique la portion de signal de parole délimitée par deux pauses. Cette simplification s'est avérée pratique pour le traitement de la parole spontanée où le nombre des pauses est relativement élevé. La comparaison des distributions avec le test ANOVA (***AN***alysis ***O***f ***VA***riance) a démontré, pour les deux corpus, un allongement très significatif ($p < 0,001$)

et très important de la durée à la fin de l'unité prosodique et un allongement un peu moins important mais toujours significatif à la fin du mot.

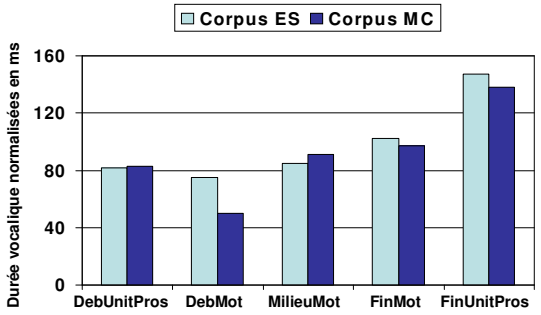

Figure 1 : Distribution de la durée vocalique selon la position du noyau vocalique.

L'interprétation possible de ces résultats est que les syllabes avant les pauses correspondent le plus souvent à une frontière prosodique, par conséquent leur durée s'allonge. Or, les syllabes finales des mots, non suivies de pause, ne coïncident pas toujours avec une frontière prosodique (d'où une plus grande dispersion de leurs durées).

3.2. Analyse de la fréquence fondamentale

Les distributions de la pente de F0 sont également comparées sur les syllabes selon leur position. Le mouvement de F0 a été représenté par sa direction, sa pente (la vitesse de changement) et son amplitude absolue mesurée en semi-tons [7]. L'amplitude absolue du mouvement s'est révélée être le paramètre le plus pertinent. Les résultats du test ANOVA, pour les amplitudes du mouvement de F0, ont montré que la valeur moyenne en fin d'unité prosodique est très significativement ($p < 0,001$) plus grande que dans toutes les autres positions et cela pour les deux corpus.

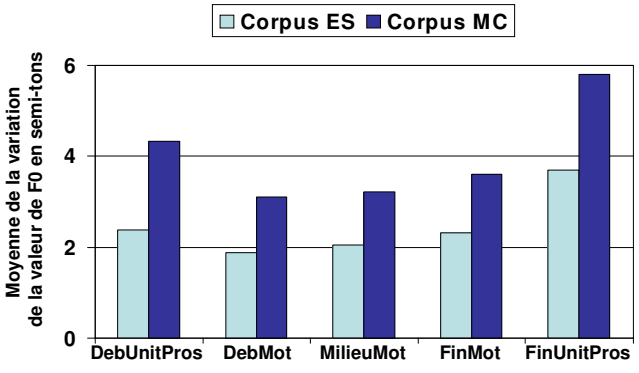

Figure 2 : Distribution de la variation de la valeur de F0 (amplitude absolue) selon la position de la syllabe.

Comme pour la durée vocalique, ici encore, nous pouvons émettre l'hypothèse que les mouvements importants de F0 marquent les frontières prosodiques : c'est presque toujours le cas avant une pause (sauf en cas d'hésitation [1]), et c'est également vrai pour certains mots à l'intérieur d'une unité prosodique.

4. DÉTECTION DES FRONTIÈRES PROSODIQUES

L'analyse statistique a démontré une spécificité de la valeur de la durée phonémique et du mouvement de F0 sur des frontières prosodiques. Pour capter cette spécificité nous avons entrepris la modélisation de ces paramètres afin de segmenter le signal de parole.

Les résultats de la segmentation en unités prosodiques sont évalués d'une part par rapport aux frontières lexicales et d'autre part par rapport aux frontières prosodiques annotées manuellement sur un sous-ensemble (20%) de données du corpus ES. La décision de tester le découpage automatique par rapport aux frontières lexicales a été motivée par le souci de relever des erreurs de découpage sur une base de données de taille conséquente. Dans ce test-là il ne s'agit pas d'examiner si l'emplacement d'une frontière prosodique est correct, mais de vérifier que la frontière prosodique hypothétique (proposée par le découpage automatique), coïncide avec la frontière du mot ou non. Si elle coïncide avec la frontière lexicale elle est acceptée comme potentiellement correcte, et si elle ne coïncide pas avec la frontière lexicale (dernière voyelle du mot), elle est comptabilisée comme erreur de découpage. Bien que cette dernière évaluation puisse être discutable, il faut remarquer qu'une segmentation manuelle effectuée par plusieurs experts ne fournit pas nécessairement les mêmes résultats. En effet dans beaucoup de cas la détection des frontières prosodiques dépend de l'appréciation subjective de l'expert qui segmente [5].

Lors de l'évaluation de l'emplacement des frontières prosodiques nous nous focaliserons sur les frontières non suivies de pauses; les frontières prosodiques suivies de pause sont exclues de l'analyse.

4.1. Durée phonémique

Une modélisation discrète de la durée phonémique a été réalisée pour deux positions pertinentes: position frontière prosodique et position non-frontière prosodique. Afin d'éviter de segmenter manuellement en unités prosodiques une base de données importante, nous avons utilisé pour l'apprentissage de nos modèles deux positions facilement repérables automatiquement et qui correspondaient néanmoins d'une façon relativement exacte à ces deux positions. Nous avons utilisée la position interne des mots pluri-syllabiques pour l'apprentissage du modèle "non-frontière-prosodique" et la position "fin de mot suivie d'une pause" pour l'apprentissage du modèle "frontière-prosodique"

La modélisation de la durée phonémique a été réalisée uniquement pour les voyelles. Cette démarche se justifie par le fait que les durées vocaliques sont plus facilement comparables entre elles que les durées syllabiques, car la syllabe possède une structure de complexité variable ayant un impact direct sur sa durée. La modélisation de la durée

a été réalisée à travers une modélisation discrète par la construction d'histogrammes normalisés des trois paramètres suivants : durée de la voyelle se trouvant sur la frontière prosodique hypothétique, durée de la voyelle qui suit l'hypothèse de frontière et le rapport de ces deux durées vocaliques. Le rapport des durées des deux voyelles adjacentes (voyelle courante et voyelle suivante) s'est avéré une méthode simple à implémenter et efficace dans la recherche des frontières prosodiques.

Utilisation et évaluation du modèle

Au moment du test, le critère de décision de l'occurrence d'une frontière prosodique (FP) basé sur un rapport de vraisemblance "**P(param|FP)/P(param|non_FP)**" , supérieur à un seuil prédéfini, et complété par un test sur le rapport des durées (pour une frontière prosodique, le rapport doit être supérieur à 1). Ainsi, pour qu'une frontière prosodique soit détectée par cette méthode, deux conditions doivent être réunies : le rapport des durées doit être plus grand que 1 (condition requise pour l'occurrence d'une frontière) et le score fourni par le modèle doit être plus élevé que le seuil de décision arbitraire utilisé. Si le rapport de vraisemblance est inférieur à ce seuil, la possibilité de l'occurrence d'une frontière prosodique est rejetée, sinon elle est acceptée.

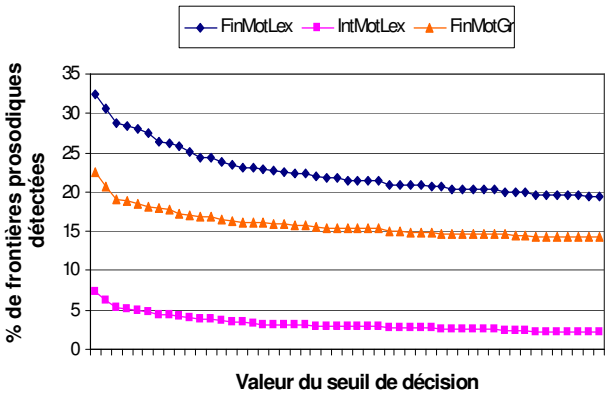

Figure 4 : Evolution des hypothèses de frontières prosodiques en fonction de la valeur du seuil choisi.

Lors des évaluations préliminaires, des tests d'hypothèses ont été appliqués pour différentes valeurs du seuil de décision, d'une part sur les frontières des mots et d'autre part sur des positions internes aux mots. Une différence est faite entre mots grammaticaux (clitiques monosyllabiques) et mots lexicaux. La figure 4 trace les résultats obtenus. Selon la valeur du seuil, 20 à 30% des fins des mots lexicaux sont détectées comme frontières prosodiques possibles. En revanche, pour les positions internes aux mots (IntMotLex), qui ne peuvent pas être des frontières prosodiques et constituent donc des erreurs incontestables de détection, seules 3 à 5% d'entre elles sont détectées à tort comme des frontières prosodiques.

Ces mesures de taux de détection d'hypothèses de frontières prosodiques ont également été réalisées par des tests croisés entre les deux bases de données utilisées. L'apprentissage des paramètres a été effectué sur la moitié des données appartenant soit à la même base que les données de test, soit à l'autre base. Comme cela apparaît au vu des résultats de la Table I, le seuil de décision des frontières prosodiques est dépendant de la base d'apprentissage. Néanmoins, si nous options pour un seuil de décision fournissant un taux d'erreur comparable de **4%,** le nombre de frontières prosodiques hypothétiques, détectées sur le corpus **MC** serait alors de **26%** pour les mots lexicaux et **10%** pour les mots grammaticaux.

Table I: Détection des frontières prosodiques avec bases d'apprentissage et de tests croisées.

Corpus		Fin Mot Lexical	Fin Mot Gram	InterMot (Erreur)
Tst	**App**			
ES	ES	26,3%	17,8%	***4,3%***
	MC	17,1%	12,9%	***1,9%***
MC	ES	30,5%	10,7%	***4,0%***
	MC	19,9%	5,8%	***1,7%***

4.2. Fréquence fondamentale

L'évolution de la courbe de la fréquence fondamentale a été employée pour découper le signal de parole en unités prosodiques. Dans cette approche nous n'avons pas utilisé de modélisation proprement dite, ainsi elle est exempte d'apprentissage. La décision de détection d'une frontière prosodique est prise en fonction de la pente de la fréquence fondamentale observée et de l'écart-type de la variation de la valeur de la fréquence fondamentale sur la portion de signal de parole située entre deux pauses. Dans cette approche uniquement les pentes montantes sont considérées. La valeur de l'écart-type de F0 est remise à jour à chaque nouvelle pause. Le seul préalable pour l'utilisation de cette approche est la détection des pauses présentes sur le signal de parole (détection bruit/parole). Le seuil de détection des frontières prosodiques est exprimé en % de l'écart-type.

Evaluation de la méthode

Comme la technique du découpage prosodique par la pente de F0 n'utilise pas le décodage phonétique du signal de parole et exploite tous les segments voisés, certains critères d'évaluation sont différents. Le découpage prosodique obtenu par la pente de F0 est considéré correct quand il se situe sur le nucleus ou la coda de la dernière syllabe d'un mot – c'est-à-dire sur la fin du mot. Tous les autres emplacements sont considérés comme erronés. Quand la frontière prosodique est placée à tort alors l'erreur moyenne de détection est quantifiée en ms comme l'écart entre la frontière du mot et la frontière prosodique.

La Table II indique, pour les deux bases de données et pour un fonctionnement donné du seuil de décision, le taux de frontières prosodiques non-suivies de pauses et pour ces frontières-là le taux dont l'emplacement coïncidait avec une frontière lexicale. La dernière colonne comporte l'écart moyen d'erreur entre la frontière détectée et la frontière lexicale en ms (dont l'écart-type se situe à ~

70 – 80 ms).

Table II: Détection des frontières prosodiques (FP) par la pente de F0.

Corpus	FP non suivi de pause	FP-Frontière Lexical	Erreur
ES	43%	80%	30 ms
MC	23%	77%	36 ms

4.3. Comparaison à la segmentation manuelle

Afin d'estimer la précision du découpage prosodique automatique, 20% du corpus ES a été segmenté manuellement en unités prosodiques. 48% des frontières prosodiques, placées manuellement, étaient des frontières non-suivies de pauses. La comparaison entre segmentation manuelle et automatique est effectuée uniquement pour les frontières prosodiques non suivies de pause. Afin d'effectuer cette comparaison, le point de fonctionnement des méthodes automatiques a été choisi de sorte qu'il représente un bon compromis entre le nombre de frontières prosodiques placées à tort et le nombre de frontières prosodiques placées sur des frontières lexicales.

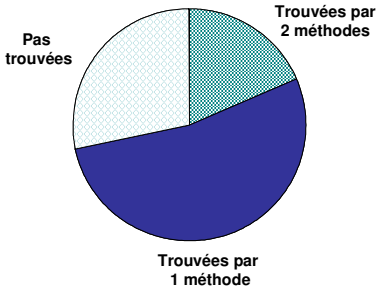

Figure 5 : Evaluation de la détection des frontières prosodiques en fonction des frontières manuelles.

Pour le point de fonctionnement choisi, la segmentation automatique utilisant la durée phonémique a abouti à un nombre de frontières prosodiques non suivies de pause (1242) inférieur à la segmentation manuelle (1446) alors que la segmentation par F0 a abouti à un nombre supérieur (2001). Une correspondance entre découpage manuel et découpage automatique obtenu par la durée phonémique concernait 44% des unités prosodiques (non suivies de pause), et entre découpage manuel et découpage par F0 concernait 46% des unités prosodiques (non suivies de pause). La répartition de la complémentarité et de la correspondance entre les 3 méthodes de détection des frontières prosodiques est illustrée sur la Figure 5. Ainsi, 18,5% des frontières prosodiques manuelles correspondaient avec les frontières prosodiques détectées par les deux paramètres, tandis que 53% des frontières prosodiques placées manuellement correspondaient avec une frontière prosodique détectée par une des deux méthodes de découpage automatique. Cela démontre qu'une complémentarité entre les deux paramètres existe quant à la détection des frontières prosodiques. Ainsi, par exemple, 80% des hésitations présentes sur le signal de parole ont été détectées comme frontière prosodique par la durée vocalique. Or, une frontière prosodique a été placée sur moins de 10% des hésitations (8 % pour la base ES et 6,7% pour la base MC) quand la détection de la frontière prosodique a été réalisé par la pente de F0.

5. CONCLUSION

Nous avons présenté dans cette étude une analyse prosodique de deux bases de données de parole spontanée enregistrées par plusieurs milliers de locuteurs. Une méthode de découpage automatique du signal de parole en unités prosodiques a été développée en utilisant la durée vocalique et la pente de F0. L'évaluation des résultats de cette méthode a été effectuée par rapport aux frontières lexicales et aux frontières prosodiques placées manuellement. Le découpage automatique a donné des résultats plus qu'encourageants. Par ailleurs, une complémentarité des paramètres dans la détection des frontières prosodiques a été observée.

La suite de ce travail devrait s'orienter vers l'étude d'une combinaison des deux paramètres tendant vers leur utilisation conjointe dans la détection des frontières prosodiques. Par ailleurs, la fiabilité des paramètres devrait être renforcée par leur apprentissage sur une base segmentée manuellement en unités prosodiques.

BIBLIOGRAPHIE

[1] K. Bartkova. Prosodic cues of spontaneous speech in French. Dans *DISS'05,* pages 21-25, 2006.

[2] K. Bartkova, D. Jouvet. Usefulness of phonetic parameters in a rejection procedure of an HMM-based speech recognition system. Dans *EUROSPEECH-1997*, pages 267-270, 1997.

[3] H. Fujisaki. Information, Prosody, and modeling with emphasis on tonal features of Speech. Dans *Speech Prosody 2004,* pages 1–10, 2004.

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

[5] 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

[6] J. 't Hart, R. Collier, A. Cohen. *A perceptual study of intonation. An experimental-phonetic approach to speech melody*, Cambridge University Press, 1990.

[7] A. Waibel. *Prosody and speech recognition*, Morgan Kaufmann, London, 1988