

Comparaison des propriétés acoustiques de la parole lue, préparée et conversationnelle en français

Jean-Luc Rouas¹, Mayumi Beppu², Martine Adda-Decker¹

¹ LIMSI-CNRS UPR 3251, France

² Dept. Computer Science, Tokyo Institute of Technology, Japan
rouas@limsi.fr, beppu@ks.cs.titech.ac.jp, madda@limsi.fr

ABSTRACT

In this paper, we investigate the acoustic properties of vocalic phonemes in three speaking styles : Read speech, broadcast news and casual spontaneous speech. Our aim is to understand better why speech recognition systems still fail to achieve good performances on spontaneous speech. Using Nakamura's method [9], we use classical speech recognition features, MFCC, and try to represent the effects of the speaking styles on the spectral space. We happen to find some cues, and it also seems that phone duration also plays an important role regarding spectral reduction, especially for spontaneous speech.

Keywords: speaking styles, speech recognition

1. Introduction

Ce travail fait suite aux recherches effectuées par Nakamura et al. [9] sur les styles de parole en Japonais. Dans cet article, nous nous focaliserons sur les styles de parole en Français, tout d'abord en reproduisant certaines expériences de Nakamura et en proposant ensuite de nouvelles analyses. Les motivations pour ce travail sont décrites dans la section 2. La méthode d'analyse est détaillée dans la section 3. Les résultats obtenus par Nakamura sur le Japonais sont rappelés dans la section 4. Les données utilisées pour l'analyse du Français sont décrites dans la section 5. Les expériences effectuées sur les données en Français sont détaillées dans la section 6. Enfin, nous comparons les résultats obtenus sur le Français et sur le Japonais lorsque cela peut être fait.

2. Motivations

Après s'être principalement penchés sur la transcription de la parole lue, les systèmes de reconnaissance automatique de la parole obtiennent aujourd'hui de très bons résultats sur des données radiophoniques, habituellement légèrement au dessus de 90% [3]. Comme le montrent les évaluations NIST des deux dernières décennies, ces systèmes ont bénéficié de nombreuses années de recherches dédiées à la transcription automatique de données de parole lue ou radiophonique.

Toutefois, lorsque ces systèmes sont confrontés à de la parole conversationnelle spontanée, les performances se dégradent nettement. Cette chute drastique des performances peut s'expliquer par le fait que des dif-

férences majeures existent entre la parole lue et la parole spontanée, à la fois en termes linguistiques et acoustiques [4].

Un des challenges subsistant pour la reconnaissance automatique de la parole est d'apporter aux systèmes la possibilité de traiter toutes les qualités de parole, y compris la parole conversationnelle spontanée. Dans cet objectif, il est important d'analyser quelles sont les différences les plus notables entre les styles de parole en considérant les paramètres habituellement utilisés dans les systèmes de transcription automatique, les paramètres acoustiques.

De ce point de vue, les distributions spectrales des voyelles et des syllabes en parole continue sont bien plus réduites que lorsque celles-ci sont prononcées isolément. Ce phénomène est appelé réduction spectrale. Il a également été observé dans l'espace paramétrique lors d'une comparaison entre parole spontanée et parole lue, en utilisant des données formantiques provenant d'un locuteur [11]. L'étude de Nakamura [9] sur le Japonais a confirmé ce résultat en utilisant un corpus plus étendu.

En ce qui concerne le Français, nous savons que la parole lue et la parole spontanée sont structurellement différentes. Les syllabes complexes ont tendance à se simplifier, et la disparition des consonnes de fin de mot et des voyelles des syllabes non accentuées sont fréquentes pour la parole spontanée en Français [1].

C'est ce phénomène que nous estimons être responsable d'une partie des erreurs de transcription en parole conversationnelle que nous nous proposons d'étudier dans cet article.

Un autre phénomène pouvant être caractéristique de la parole spontanée est l'extension de variance spectrale. Ce phénomène a été caractérisé pour le Japonais dans [9].

Nous présentons dans la section suivante les méthodes de calcul permettant de quantifier l'intensité de ces phénomènes.

3. Méthode de calcul et formules

3.1. Paramètres

Sur des fichiers audio numérisés avec un taux d'échantillonnage de 16 kHz, un jeu de 12 coefficients cep-

traux (MFCC) sont extraits en utilisant une fenêtre de 25 ms avec un recouvrement de 10 ms. Aux vecteurs MFCC sont ajoutées leurs dérivées premières et secondes, ainsi que les premières et secondes dérivées de la log-énergie. Les vecteurs résultants sont de dimension 38. Ce sont les paramètres les plus classiquement utilisés en reconnaissance de parole.

3.2. Ratio de réduction de l'espace spectral

Afin de caractériser la réduction de l'espace spectral, nous devons nous munir d'un corpus de référence. Ce corpus, dénoté dans les formules suivantes R , est dans notre cas un corpus de parole lue.

L'étendue de l'espace spectral est estimée en faisant la différence entre la moyenne des vecteurs MFCC pour un phonème p donné et la valeur moyenne sur l'ensemble des phonèmes du corpus.

Le ratio est mesuré en divisant l'estimateur de l'étendue de l'espace spectral d'un corpus X par celui du corpus de référence R . Nous utilisons ici des distances euclidiennes.

En d'autres termes, le calcul peut être résumé par la formule suivante :

$$red_p(X) = \frac{\|\mu_p(X) - Av(\mu_p(X))\|}{\|\mu_p(R) - Av(\mu_p(R))\|} \quad (1)$$

avec μ_p la valeur moyenne des vecteurs MFCC du phonème p du corpus X , $\mu_p(R)$ la moyenne des vecteurs MFCC pour le phonème p pour la parole lue (corpus R), Av indique la valeur moyenne.

3.3. Ratio d'extension de la variance spectrale

De la même manière, nous définissons le ratio d'extension de la variance spectrale par rapport à un corpus de référence R .

La variance spectrale est estimée comme étant la somme des variances de chacun des coefficients MFCC pour l'ensemble des réalisations d'un phonème p .

Le ratio d'extension de la variance est alors le rapport entre la variance spectrale du phonème p corpus X et celle du même phonème pour le corpus R .

Il est calculé en utilisant la formule suivante :

$$ext_p(X) = \frac{\sum_{k=1}^K \sigma_{pk}^2(X)}{\sum_{k=1}^K \sigma_{pk}^2(R)} \quad (2)$$

avec K la dimension des vecteurs MFCC, $\sigma_{pk}^2(X)$ le k ième élément du vecteur de variance obtenu pour le phonème p avec le style de parole X .

4. Caractérisation acoustique des styles de parole en Japonais [9]

Dans [9], les différents styles de parole sont étudiés grâce à de grands corpus. Les données utilisées lors de ces expériences proviennent du "Corpus of Sponta-

neous Japanese" (CSJ) [8] et du "Japanese Newspaper Article Sentence" (JNAS) [6].

Ces bases de données couvrent différentes conditions discursives, incluant des monologues, des dialogues et de la parole lue. Pour les expériences suivantes, seules quelques conditions, considérées comme les plus représentatives d'un style de parole, sont utilisées : Parole lue, Présentations académiques, Présentations informelles, Dialogue.

Les expériences menées en utilisant sur ces bases de données montrent que la réduction de l'espace acoustique des MFCC est observable pour quasiment tous les phonèmes dans les trois styles de parole par rapport à la parole lue. Ce phénomène est plus marqué lorsque l'on considère les situations de dialogue.

Les résultats obtenus avec la mesure du ratio d'extension de variance spectrale montrent bien une augmentation de la variance pour quasiment tous les phonèmes dans les trois styles de parole.

5. Données en Français

Trois bases de données comprenant de la parole lue, préparée et conversationnelle ont été utilisées. Pour la parole lue, nous utilisons le corpus BREF [7], qui est composé de textes journalistiques lus. Le corpus BREF comporte plus de 100 heures de parole lue par 120 locuteurs. Les textes ont été sélectionnés à partir du journal LE MONDE afin de couvrir un vocabulaire large (plus 20000 mots) et un nombre important de contextes phonétiques.

Le corpus de parole préparée est constitué des parties de développement et de test des campagnes d'évaluation ESTER (2003-2004) et ESTER2 (2007-2008) [5]. Ce corpus a une durée d'environ 50 heures. Les données sont composées d'émission télévisées ou radiophoniques provenant de France mais également de plusieurs pays francophones (Maghred et Afrique) : France Inter, Radio France Internationale, Radio Télévision du Maroc, France Info.

Les données de parole spontanée proviennent du corpus NCCFr [10]. Ce corpus est composé de conversations informelles entre amis. D'une durée de 36 heures, ce corpus comprend 23 paires de locuteurs (24 hommes et 22 femmes), chaque conversation durant approximativement 90 minutes. Ce corpus a été enregistré en 2007 à Paris.

Chacun de ces corpus sera dénommé dans la suite de ce document par les acronymes BREF, ESTER, NCCFr respectivement. L'ensemble des corpus inclut des transcription orthographiques manuelles qui ont été phonétisées et alignées automatiquement.

6. Expériences

La première expérience est une mesure des durées des phonèmes automatiquement alignés pour les différents corpus. L'objectif est d'une part de vérifier la répartition des durées des phonèmes selon les styles de parole et, d'autre part, de sélectionner les phonèmes ayant les durées les plus représentées.

Le résultat de cette expérience est donné sur la figure 1.

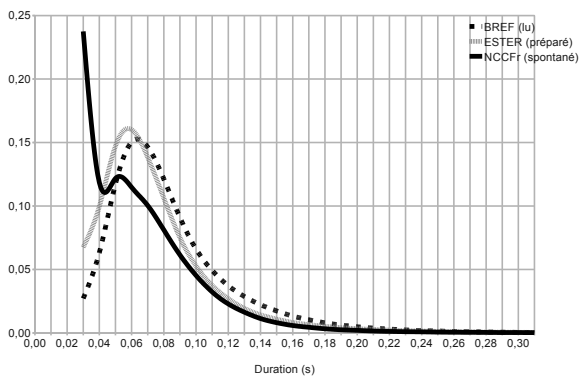


Fig. 1: Distribution des durées des phonèmes pour chaque corpus. Abscisses : durée en secondes, Ordonnées : estimation de la probabilité

Comme le montre cette figure, les distributions pour BREF et ESTER ont des formes semblables, avec un léger décalage vers la gauche pour ESTER. Ce décalage montre l'effet du débit de parole, normalement plus élevé dans le style journalistique et la parole spontanée que pour la lecture.

La forme de la distribution pour le corpus NCCFr est toutefois assez différente, avec un pic important vers les durées de l'ordre de 30 ms (durée minimale d'un phonème dans l'alignement automatique). Cela peut s'expliquer par le fait que les données considérées sont en parole conversationnelle spontanée, donc sujettes aux phénomènes de délétions de certains phonèmes. Une étude plus détaillée de ce phénomène est effectuée dans [2].

Les valeurs médianes des distributions sont cependant similaires pour chacun des corpus (autour de 60 à 70 ms). Pour la suite des expériences, nous avons dans un premier temps décidé de ne considérer que les segments phonétiques ayant des durées comprises entre 40 et 120 ms.

6.1. Ratio de réduction & extension de variance

Le ratio de réduction de l'espace spectral des phonèmes et le ratio d'extension de variance sont calculés sur nos données en utilisant les formules décrites dans les sections 3.2, équation 1 et 3.3, équation 2. Le corpus de référence R est pour l'ensemble des expériences menées ci-après le corpus BREF. Pour des raisons de clarté, nous avons décidé de ne représenter que les figures décrivant les résultats obtenus pour les voyelles.

Comme nous pouvons le voir sur la première ligne de la figure 2, la réduction de l'espace spectral est observée principalement pour le corpus ESTER. Pour ce corpus, quasiment tous les phonèmes voient leur espace spectral réduit, à part /i/ et /y/. Sur notre corpus de parole conversationnelle, NCCFr, l'espace

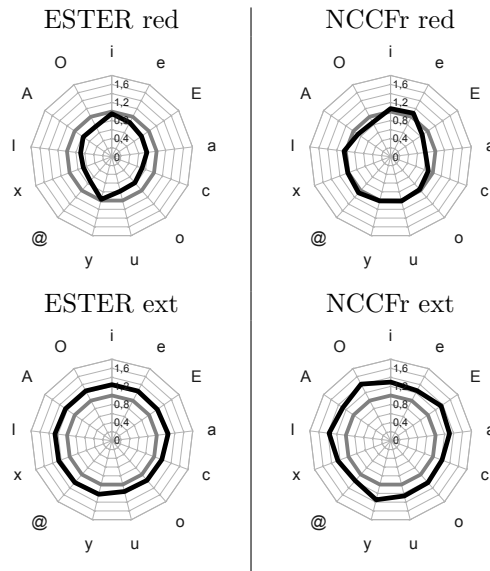


Fig. 2: ratio de réduction spectrale (red) and ratio d'extension de variance (ext) pour les corpus ESTER and NCCFr comparés avec BREF

spectral semble quasiment identique à celui du corpus de référence, à l'exclusion de certains phonèmes qui voient leur étendue spectrale légèrement diminuée.

Les résultats sur la mesure de l'extension de la variance spectrale sont en revanche cohérents avec nos prédictions. Nous pouvons observer, sur la deuxième ligne de la figure 2, une augmentation nette de la variance spectrale pour ESTER, et une augmentation encore plus importante pour NCCFr.

6.2. Influence de la durée des segments phonémiques

Au vu des résultats quelque peu étonnants de la section précédente, nous avons effectué des expériences supplémentaires pour évaluer la réduction de l'espace spectral pour les segments courts (en dessous de 40 ms) et les segments longs (au dessus de 120 ms).

La figure 3 montre les résultats des expériences effectuées selon ces deux conditions. Nous pouvons observer ici que les segments courts, notamment pour NCCFr, voient leur étendue spectrale très réduite. Malgré la durée des segments considérés, ce phénomène n'est pas à négliger car les segments courts sont nombreux dans notre corpus de parole conversationnelle. Pour les segments de durée importante, la réduction de l'espace spectral n'est pas très importante.

Le ratio d'extension de variance a également été calculé pour les deux conditions (figure 4). Ces figures montrent que la variance spectrale des phonèmes augmente pour les deux conditions de durée. L'extension paraît plus importante pour les segments courts de parole conversationnelle par rapport aux segments longs.

7. Discussion

Nakamura [9] obtient des résultats montrant des effets assez nets de la réduction de l'espace spectral et de l'extension de la variance spectrale en Japonais.

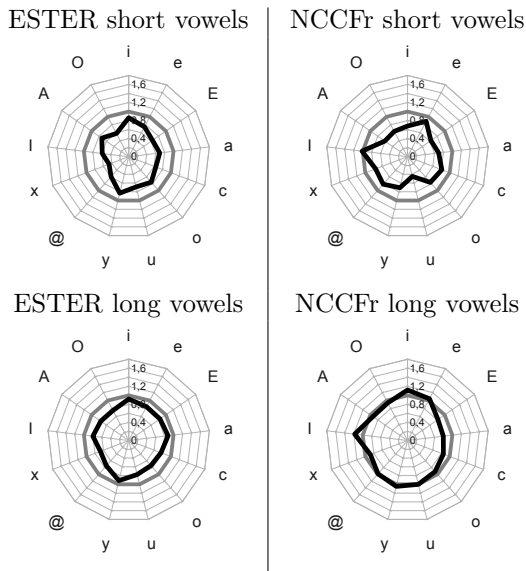


Fig. 3: Ratio de réduction spectrale pour les corpus ESTER et NCCFr en utilisant BREF comme référence. première ligne : segments courts (<40ms), deuxième ligne : segments longs (>120ms)

Sur ses données, l'espace spectral est de plus en plus réduit lorsqu'il analyse des données de plus en plus spontanées. De la même manière, la variance spectrale est toujours plus importante pour les corpus de parole spontanée.

Dans nos expériences sur le français, nous avons pu reproduire les mêmes résultats pour les mesures de variance spectrale. Cependant, les tests effectués sur les segments de durée "normale" vont quelque peu à l'encontre des résultats de Nakamura : l'espace spectral est moins étendu pour la parole journalistique que pour la parole spontanée.

Nous avons pu toutefois mesurer une réduction spectrale importante pour les segments courts, majoritaires dans le cas de la parole spontanée.

C'est certainement sur ces segments, déjà difficiles à reconnaître de manière automatique à cause de leur faible durée, que nous devons focaliser nos efforts afin d'améliorer les performances des systèmes de transcription automatique.

Références

- [1] M. Adda-Decker, P. Boula de Mareuil, G. Adda, and L. Lamel. Investigating syllabic structures and their variation in spontaneous french. *Speech Communication*, 46(2) :119–139, 2005.
- [2] M. Adda-Decker, C. Gendrot, and N. Nguyen. Contributions du traitement automatique de la parole à l'étude des voyelles orales du français. *Traitement Automatique des Langues*, 49, 2008.
- [3] P. Fousek, L. Lamel, and J.-L. Gauvain. Transcribing Broadcast Data Using MLP Features. In *InterSpeech'08*, pages 1433–1436, Brisbane, Australia, September 22–26 2008.
- [4] S. Furui. Recent advances in spontaneous speech recognition and understanding. In *ISCA & IEEE*

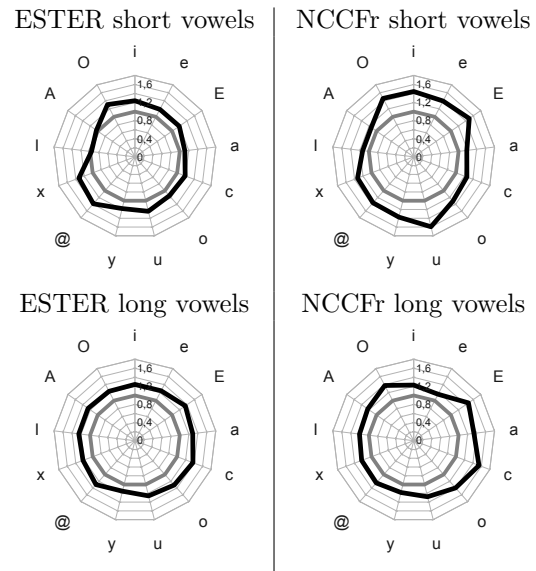


Fig. 4: Ratio d'extension de variance spectrale pour les corpus ESTER et NCCFr en utilisant BREF comme référence. première ligne : segments courts (<40ms), deuxième ligne : segments longs (>120ms)

workshop on Spontaneous Speech Processing and Recognition (SSPR), 2003.

- [5] S. Galliano, E. Geoffrois, G. Gravier, J.-F. Bonastre, M. Mostefa, and K. Choukri. Corpus description of the ester evaluation campaign for the rich transcription of french broadcast news. In *Language Evaluation and Resources Conference*, 2006.
- [6] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi. Jnas : Japanese speech corpus for large vocabulary continuous speech recognition research. *Journal of the acoustical society of Japan*, 20(3) :199–206, 1999.
- [7] L. Lamel, J. L. Gauvain, and M. Eskenazi. Bref, a large vocabulary spoken corpus for french. In *Eurospeech*, 1991.
- [8] K. Maekawa. Corpus of spontaneous japanese : its design and evaluation. In *ISCA & IEEE workshop on Spontaneous Speech Processing and Recognition (SSPR)*, 2003.
- [9] M. Nakamura, K. Iwano, and S. Furui. Differences between acoustic characteristics of spontaneous and read speech and their effects on speech recognition performance. *Computer Speech and Language*, 22 :171–184, 2008.
- [10] F. Torreira, M. Adda-Decker, and M. Ernestus. The nijmegen corpus of casual french. *Speech Communication*, in press.
- [11] R. J. J. H. van Son and L. C. W. Pols. An acoustic description of consonant reduction. *Speech Communication*, 28(2) :125 – 140, 1999.