Comparison of spectral properties of read, prepared and casual speech in French

Jean-Luc Rouas[1], **Mayumi Beppu**[2], **Martine Adda-Decker**[1]

[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 phonemes in three speaking styles: read speech, prepared speech and spontaneous speech. Our aim is to better understand why speech recognition systems still fails to achieve good performances on spontaneous speech. This work follows the work of Nakamura et al. (Nakamura et al., 2008) on Japanese speaking styles, with the difference that we here focus on French. Using Nakamura's method, we use classical speech recognition features, MFCC, and try to represent the effects of the speaking styles on the spectral space. Two measurements are defined in order to represent the spectral space reduction and the spectral variance extension. Experiments are then carried on to investigate if indeed we find some differences between the three speaking styles using these measurements. We finally compare our results to those obtained by Nakamura on Japanese to see if the same phenomenon appears.

1. Introduction

This work follows the work of Nakamura et al. (Nakamura et al., 2008) on Japanese speaking styles, with the difference that we here focus on French. We first conduct the same experiments as Nakamura and then propose further analysis. Motivations for this work are expressed in section 2.. The section 3. describes the analysis method. In section 4., we recall the experiments and the results obtained by Nakamura on Japanese. The next section, section 5., details the French corpora we are using. The experiments achieved on this French data are detailed in section 6.. In the last section, we compare the results between Japanese and French data whenever they may be compared, and we propose further analyses.

2. Motivations

After having focused mainly on read speech, the automatic speech transcription systems also achieve nowadays very good performances on broadcast news data (usually around 10-15% of Word Error Rate (Fousek et al., 2008)). These systems benefit of years of research focused only on the automatic transcription of read speech and broadcast news data, which can be attested by looking at the numerous NIST evaluations carried on during the last decades.

However, when coping with spontaneous speech, even the best actual systems achieve quite poor performances, with an accuracy that may be as low as 40%. These results emphasise the fact that spontaneous speech and read speech are indeed very different both acoustically and linguistically (Furui, 2003).

One of the remaining challenges for automatic speech transcription is to give the systems the ability to deal with every kind of input data, including spontaneous speech. To fulfil this goal, it is crucial to analyse what are the most noticeable differences between speaking styles using features traditionally used by automatic speech transcription systems.

On the acoustic point of view, it is known that the spectral distribution of continuously spoken vowels or syllables is much smaller than that of isolated spoken vowels or syllables. This phenomenon is sometimes called spectral reduction. Similar reduction have been observed in the parametric space for spontaneous speech compared to read speech, using formant frequency data measured from one speaker (van Son and Pols, 1999). A study on Japanese confirmed this observation using large corpora (Nakamura et al., 2008). Considering French, we know that read speech and spontaneous speech are structurally different. Complex syllables tend to be simplified, and deletion of the word-final consonants and vowels of unstressed syllables frequently occurs in French spontaneous speech (Adda-Decker et al., 2005).

This phenomenon may be responsible for some of the automatic transcription errors on spontaneous speech. That is why we propose to study it in this paper.

Another phenomenon that may characterise spontaneous speech is the spectral variance extension. This phenomenon has been studied in (Nakamura et al., 2008) on Japanese data.

In the following section, we present the approaches used to compute means to assess these phenomena.

3. Acoustic analysis method

3.1. Features

Utterances are digitalised using 16kHz sampling, and segmented by silences of 400 ms duration or longer. The speech waveform are converted to 12 dimensional MFCC vectors using a 25 ms length window shifted every 10 ms. The MFCC vectors are augmented with their first and second derivatives, together with the first and second differential log-energy, resulting in 38-dimensional vectors.

3.2. Spectral reduction ratio

In order to characterise the spectral reduction phenomenon, we must use a reference corpus. This corpus, hereafter denoted R, is in our case a read speech corpus.

The spectral space extent is estimated by calculating the difference between the mean of the MFCC features for a given phoneme p and the averaged value over all the phonemes of the same corpus.

The ratio is then computed by dividing the spectral space extent for the corpus X with the one measured for the reference corpus R.

The distances used in this paper are euclidean distances.

In formal terms, the spectral reduction ratio can be expressed by the following formula:

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

with μ_p being the mean vector of phoneme p uttered with the speaking style X, $\mu_p(R)$ is the mean vector for phoneme p for read speech, and Av indicates the average value.

3.3. Spectral variance extension ratio

Using a similar method, we now define the variance extension ratio, still using the reference corpus R.

The spectral variance is estimated as being the sum of the variances obtained for each MFCC coefficient for all the realisations of a phoneme p.

The spectral variance extension ratio is thus defined as being the ratio between the spectral variance of the phoneme p on the corpus X and the spectral variance of the same phoneme on the corpus R.

The formalised method used to compute this ratio is:

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

with K being the dimension of the MFCC vector, and $\sigma_{pk}^2(X)$ the kth dimensional element of the variance vector of MFCC for phoneme p uttered with speaking style X.

4. Acoustic characteristics of Japanese speaking styles (Nakamura et al., 2008)

In (Nakamura et al., 2008), various speaking styles are analysed using large-scale Japanese speech corpora. This work focused mainly on the analysis of the differences between acoustic characteristics for spontaneous and read speech.

The data used in these experiments come from the Corpus of Spontaneous Japanese (CSJ) (Maekawa, 2003) and the Japanese Newspaper Article Sentence (JNAS) (Itou et al., 1999). The CSJ corpus is composed of 660 hours of speech and was recorded from 1999 to 2004. The JNAS corpus consists in 60 hours of speech. It was recorded from 1995 to 1997.

These corpus covers several conditions, including monologue, dialogue and read speech. For the purpose of the following experiments, only some situations, considered most representative of a speaking style, are utilised. These are, ranking from the most formal to the most informal: Read speech, Academic presentations, Extemporeaneous presentations (informal presentations), Dialogue.

The experiments carried on using this framework show that reduction of MFCC space is observed for almost all the phonemes in the three speaking styles in comparison with read speech. This phenomenon is most noticeable for the dialogue utterances.

Results obtained using the spectral variance extension ratio indeed show that an extension of the variance is observed for almost all the phonemes in the three speaking styles.

5. French corpora

Three speech corpora with read, prepared and casual spontaneous speaking styles were used. The reference read speech corpus is the French BREF (Lamel et al., 1991) corpus composed of read newspaper speech. BREF includes over 100 h of speech from 120 speakers. The text materials were selected from the French newspaper Le Monde, so as to provide a large vocabulary (over 20,000 words) and a wide range of phonetic environments.

Prepared speech comes from the development and test sets of the ESTER 2003-2004 and ESTER2 2007-2008 evaluation campaigns (Galliano et al., 2006) totalling about 50 hours of speech. The journalistic speech were selected from various French and francophone (Maghreb and Africa) radios: France Inter, Radio France International, Radio Television du Maroc, France Info.

Casual spontaneous speech comes from the NCCFr corpus (Torreira et al., in press), composed of conversations between friends. NCCFr, totalling 36 hours of speech, comprises 23 pairs of speakers (24 males, 22 females) producing conversation of about 90 minutes each. recorded in 2007 in Paris

Each corpus is hereafter called BREF, ESTER, NCCFr respectively. All this corpus comprise manual orthographic transcriptions, which have been phonetised and automatically aligned.

6. Experiments

The first experiment aims at measuring the durations of the automatically aligned phonemes for the three corpora. This is done in order to visualise the distribution of phoneme durations for each speaking style and to select the phonemes according to their duration.

The result of this experiment is drawn in figure 1.

As seen on this figure, the distributions are quite similar for BREF and ESTER, with a slight shift to the left for ESTER. This shift may be due to the effect of speaking rate, which should be faster for prepared journalistic speech than for read speech.

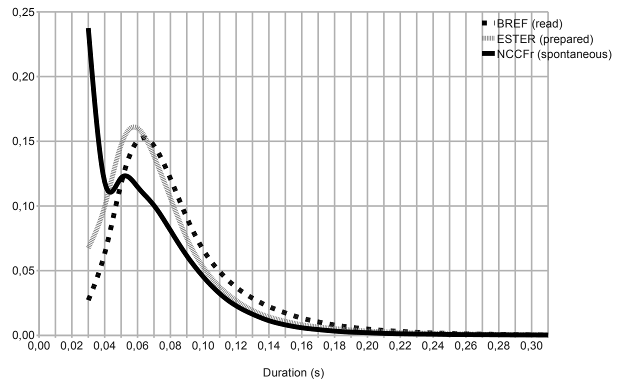


Figure 1: Duration distribution for each corpus. X-axis : duration in seconds, Y-axis : estimation of the probability

The shape of the distribution for the NCCFr corpus is however very different. We can observe a peak for phonemes of 30 ms duration, which is the shorter duration allowed by the automatic alignment. This may be an effect of conversational spontaneous speech, for which some phonemes might be subject to deletion (see (Adda-Decker et al., 2008) for more details on this phenomenon).
Although the shapes clearly differs, the median value for each distribution is almost the same (between 60 and 70 ms). Thus, we will first consider only "normal" length segments, which have a duration between 40 and 120 ms.

6.1. Phoneme reduction ratio & variance extension

The phoneme reduction ratio and variance extension ratio are computed on our data using the same formula respectively described in sections 3.2., equation 1 and 3.3., equation 2. The reference corpus R, used for all the following experiments, is the BREF corpus.
As we can observe on the first two columns of figure 2, the spectral space reduction is observed mainly for the ESTER corpus. For this corpus, almost all the phonemes have a reduced spectral space, except /i/ and /y/.
However, for NCCFr, the spectral reduction phenomenon is not as obvious. In fact, the spectral space seems almost identical to that of read speech, with a few exceptions.
The results on the spectral variance extension, shown on the last two columns of figure 2, are in accordance with our expectations. The spectral variance is indubitably increased for ESTER, and even more for ERNESTUS.

6.2. Influence of segment duration

As the results from the previous section are a bit surprising, we have therefore investigated if the observed reduction phenomenon may be emphasised or not for short or long segments, short segments being segments under 40 ms duration, long segments being over 120 ms.

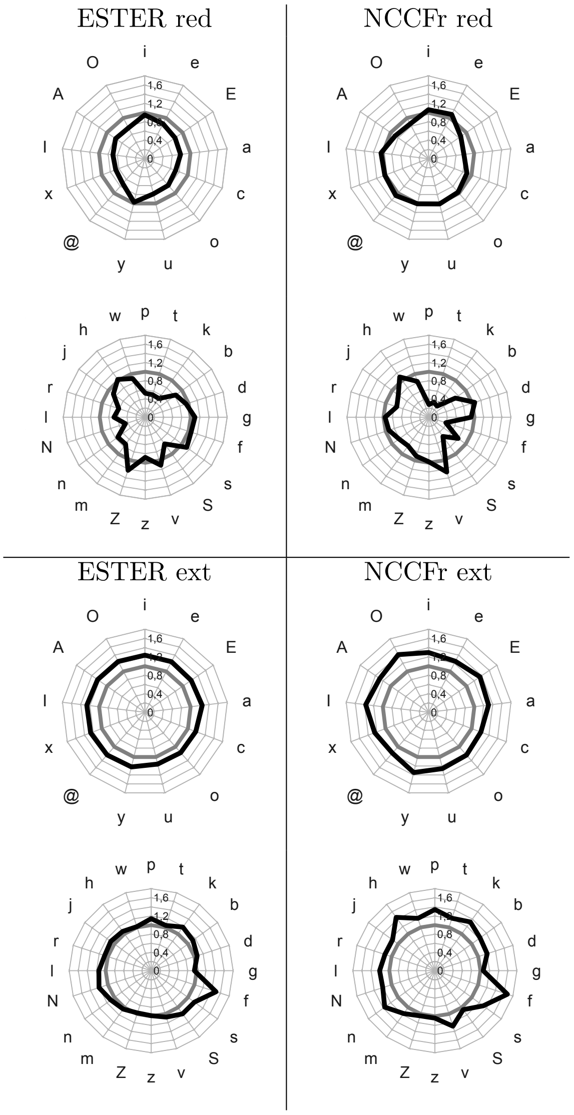
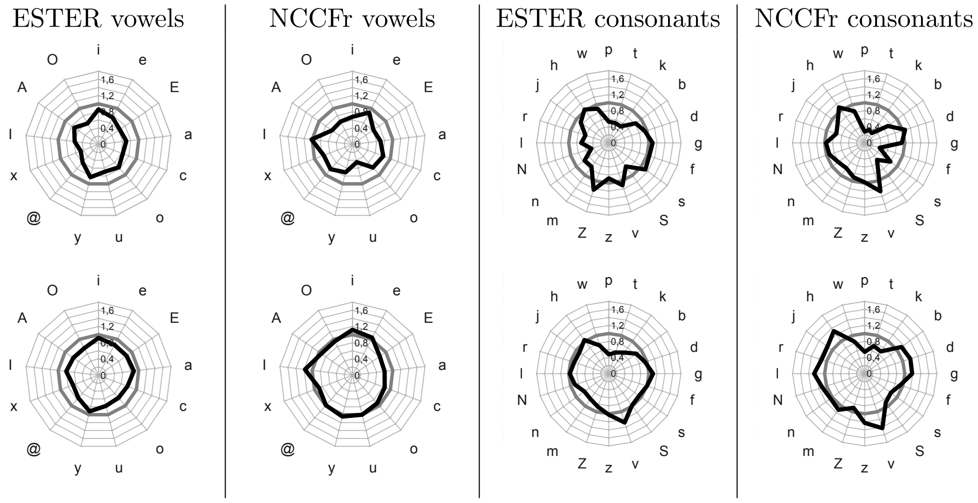


Figure 2: spectral reduction ratio (red) and variance extension (ext) for the ESTER and NCCFr corpora compared with BREF

Figure 3 shows the spectral reduction ratio for the ESTER and NCCFr corpora compared to BREF for both duration conditions. We observe that for short segments, especially for the NCCFr corpus, the spectral space is quite reduced. Even given the short duration of those segments, this phenomenon should not be neglected as short segments are indeed frequent in our conversational speech corpus. We can note however that the spectral space reduction is not very important for longer segments.

The variance extension ratio has also been computed for both duration conditions (see figure 4). The figures show that the variance of phonemes are extended for both conditions. The extension appears to be stronger for short segments of conversational speech as opposed to long segments. This phenomenon is observed for both speaking styles.

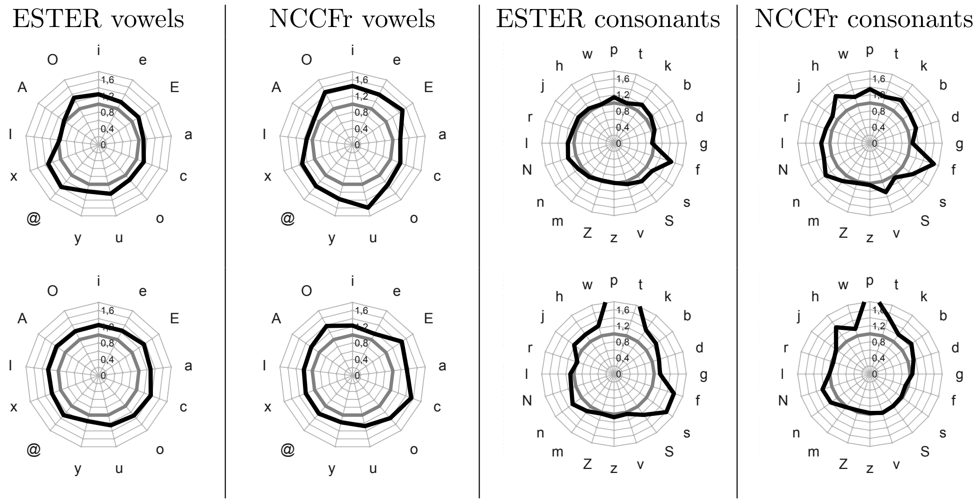

ESTER vowels

NCCFr vowels

ESTER consonants

NCCFr consonants

Figure 3: spectral reduction ratio for the ESTER and NCCFr corpora compared with BREF for different phoneme durations. first line : short phonemes, second line : long phonemes (>120ms)


ESTER vowels

NCCFr vowels

ESTER consonants

NCCFr consonants

Figure 4: spectral variance extension ratio for the ESTER and NCCFr corpora compared with BREF for different phoneme durations.first line : short phonemes, second line : long phonemes (>120ms)

6.3. Influence of lexical content

In this section, we will focus of acoustic differences that might be observed while considering the role of the words. We have distinguished function and content words, based on the word frequency in each corpus.

First, words are ranged in decreasing frequencies. Function words are then isolated using the following paradigm: they are the 100 most frequent words, as long as their cumulated count is less than 50% of all the words in the corpus. Some content words occuring in this list are manually removed: as we deal here with french news corpora (BREF and ESTER), some of the most frequent words are for example “France”, “Monsieur”, “président”.

Moreover, function words do not cover all the french phonemes equally. We have thus removed from the analysis the phonemes that occur less than 1000 times. Those omitted phonemes are : /c o ə/ and /b g ʃ v ŋ r h w/.

The results of the analysis are shown in figure 5 for the spectral space reduction, and in figure 6 for the variance extension.

Figure 5 shows that the spectral space is always more conserved for the content words than for the function words.

The variance extension ratio is larger for the vowels contained in function words than for content words.

7. Discussion

Nakamura, in (Nakamura et al., 2008) seems to provide results showing quite clearly the effects of spectral space reduction and spectral variance extension in Japanese. On his data, the spectral space is more reduced when the speaking style gets more spontaneous. Similarly, the spectral variance is increased for spontaneous data.

In our experiments on French data, we have been able to observe the same phenomenon for the spectral variance. The tests carried on “normal” duration phonemes do not however show the same behaviour for the spectral reduction. The spectral space is more reduced for

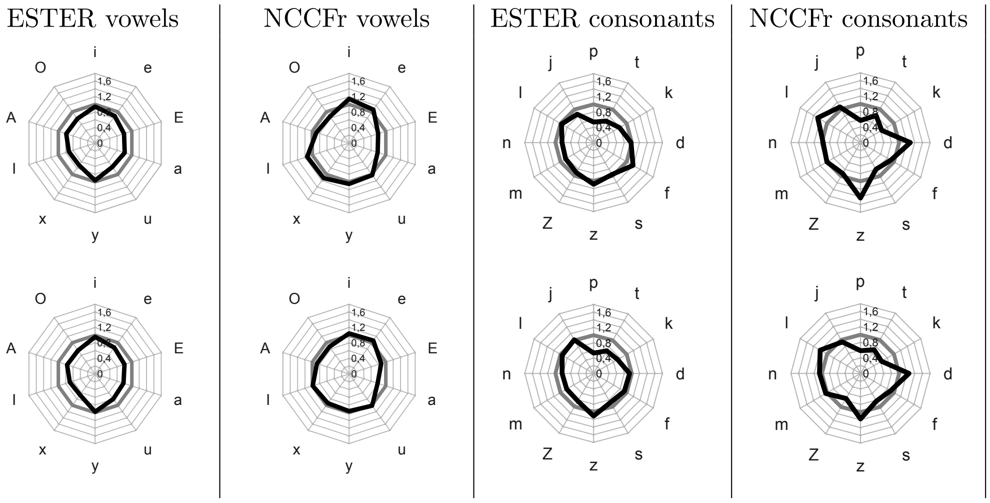


Figure 5: spectral reduction ratio for the ESTER and NCCFr corpora compared with BREF for different word classes. first line : function words, second line : content words

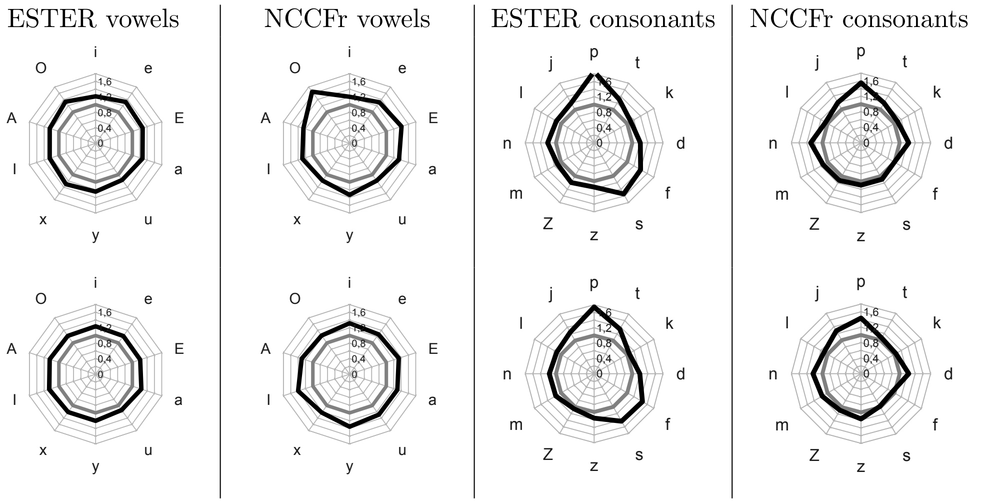


Figure 6: spectral variance extension ratio for the ESTER and NCCFr corpora compared with BREF for different word classes. first line : function words, second line : content words

prepared journalistic speech than for casual spontaneous speech.

We have nevertheless been able to observe a strong spectral space reduction for short segments, quite frequent in our spontaneous speech data.

Considering the role of words, we have seen that function words tend to be more subject to spectral reduction than content words. This is also true for the variance extension phenomenon, which is more present on function words.

However interesting these results may seem, they have to be taken with some caution: the features are indeed extracted at the middle of each phonetic segment, which might lead to false results in some cases, especially when considering consonantal phonemes.

Nevertheless, this approach provides a simple way of representing variations occurring between speaking styles, and we plan to carry on further experiments using different features, amongst them prosodic features.

8. References

Martine Adda-Decker, Philippe Boula de Mareuil, Gilles Adda, and Lori Lamel. 2005. Investigating syllabic structures and their variation in spontaneous french. *Speech Communication*, 46(2):119–139.

M. Adda-Decker, Cédric Gendrot, and Noël Nguyen. 2008. Contributions du traitement automatique de la parole à l'étude des voyelles orales du français. *Traitement Automatique des Langues*, 49.

Petr Fousek, Lori Lamel, and Jean-Luc Gauvain. 2008. Transcribing Broadcast Data Using MLP Features. In *InterSpeech'08*, pages 1433–1436, Brisbane, Australia, September 22-26.

Sadaoki Furui. 2003. Recent advances in spontaneous speech recognition and understanding. In *ISCA & IEEE workshop on Spontaneous Speech Processing and Recognition (SSPR).*

S. Galliano, E. Geoffrois, Guillaume Gravier, J.-F. Bonastre, M. Mostefa, and K. Choukri. 2006. Corpus description of the ester evaluation campaign for the rich transcription of french broadcast news. In *Language Evaluation and Ressources Conference*.

K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi. 1999. Jnas: Japanese speech corpus for large vocabulary continuous speech recognition research. *Journal of the acoustical society of Japan*, 20(3):199–206.

L. Lamel, J. L. Gauvain, and M Eskenazi. 1991. Bref, a large vocabulary spoken corpus for french. In *Eurospeech*.

K. Maekawa. 2003. Corpus of spontaneous japanese: its design and evaluation. In *ISCA & IEEE workshop on Spontaneous Speech Processing and Recognition (SSPR)*.

Masanobu Nakamura, Koji Iwano, and Sadaoki Furui. 2008. Differences between acoustic characteristics of spontaneous and read speech and their effects on speech recognition performance. *Computer Speech and Language*, 22:171–184.

F. Torreira, M. Adda-Decker, and M. Ernestus. in press. The nijmegen corpus of casual french. *Speech Communication*.

R. J. J. H. van Son and Louis C. W. Pols. 1999. An acoustic description of consonant reduction. *Speech Communication*, 28(2):125 – 140.