Audio source separation with one sensor for robust speech recognition

Laurent Benaroya, Frédéric Bimbot, Guillaume Gravier, Rémi Gribonval

AUDIO SOURCE SEPARATION WITH ONE SENSOR FOR ROBUST SPEECH RECOGNITION

L. Benaroya, F. Bimbot, G. Gravier and R. Gribonval

IRISA (INRIA & CNRS), Equipe METISS
Campus de Beaulieu, 35760 Rennes Cedex, France
E-mail : {lbenaroy,bimbot,ggravier,remi}@irisa.fr

ABSTRACT

In this paper, we address the problem of noise compensation in speech signals for robust speech recognition. Several classical denoising methods in the field of speech and signal processing are compared on speech corrupted by music, which correspond to a frequent situation in broadcast news transcription tasks. We also present two new source separation techniques, namely adaptive Wiener filtering and adaptive shrinkage. These techniques rely on the use of a dictionary of spectral shapes to deal with the non stationarity of the signals. The algorithms are first compared on the source separation task and assessed in terms of average distortion. Their effect on the entire transcription system is eventually compared in terms of word error rate. Results show that the proposed adaptive Wiener filter approach yields a significant improvement of the transcription accuracy at signal/noise ratios greater than 15 dB.

1. INTRODUCTION

Automatic transcription is a key step for the indexing and retrieval of data from audio documents such as radio broadcast news. A problem with the transcription of broadcast news is the presence of background music which is often superimposed to the voice of the speaker(s). While automatic speech recognition is a rather mature technology, its performance quickly degrades in noisy conditions, hence the need for a noise compensation scheme in the recognizer.

In automatic speech recognition (ASR) systems, noise compensation is typically a two step process. The first step consists in suppressing noise at the waveform level while the second step consists in an unsupervised adaptation of the models to the acoustic condition of the document (see *e.g.* [1]). For systems based on hidden Markov models (HMM), many model adaptation techniques have been proposed (see, *e.g.*, [2, 3]). To some extent, the choice of the features used to represent the speech signal is also designed to reduce the sensitivity to the noise [4]. For example, cepstral mean subtraction and variance normalization increase the robustness to noise. In this paper, we focus on the first step, *i.e.* noise suppression at the waveform level.

In speech recognition, the most commonly used approach for noise compensation is spectral subtraction [5, 1], which consists in removing from each frame of speech signal an estimate of the noise spectrum. Most of the existing approaches to estimate the latter rely on a speech/non-speech detector and require a relatively large number of frames to provide a good estimate of the spectrum. Therefore, such methods are not very well suited to a rapidly varying noise such as music. Moreover, many other classical denoising techniques such as Wiener filtering or "wavelet" shrinkage [6] are essentially designed to deal with Gaussian noise, a model which clearly does not fit the background music in broadcast news.

In this paper, we propose a novel approach to denoise radio broadcast news recordings. The approach is based on probabilistic models of the speech and noise signals and relies on techniques borrowed from audio source separation. Since noise suppression is a particular case of source separation, the performance of the proposed algorithms are first compared using standard performance measures for the latter field, namely source to distortion ratio (SDR), source to interference ratio (SIR) and source to artifact ratio (SAR). The methods are then compared from the "end-user" point of view, in terms of word error rate (WER).

The paper is organized as follows: in section 2, we recall the principles of some standard noise suppression algorithms such as Wiener filtering and time/frequency shrinkage. Then, in section 3, we give some details on two recently proposed denoising methods which are based on single sensor source separation algorithms: adaptive Wiener filtering and adaptive shrinkage. After a brief description of the corpus used for these experiments, the performance of the various methods is assessed in terms of the quality of source separation/denoising. Last, we evaluate the impact of the methods on the automatic speech recognition task.

2. CLASSICAL DENOISING ALGORITHMS

Denoising a signal is a widely studied problem in signal processing. Given an observed signal $y(t) = x(t) + n(t)$, the problem is to get an estimate $\widehat{x}(t)$ of the original signal $x(t)$. In a probabilistic setting, x and n are considered as observations of two random processes X and N and, according to the statistical models used, several denoising principles have been proposed. In this section, we recall two classical denoising methods: Wiener filtering and time-frequency shrinkage.

2.1. Wiener filtering

The Wiener filter is the estimator which minimizes the average quadratic distortion between the original signal x and its estimate $\widehat{x}$. In practice, Wiener filtering is performed on short-term frames of signal which are supposed short enough so that the signal can be assumed stationary and Gaussian on each frame, with power spectral density (psd) $\sigma_X^2(f)$ and $\sigma_N^2(f)$. Under these assumptions, the Wiener filter can be expressed in the frequency domain as

$$\widehat{X}(t,f) = \frac{\sigma_X^2(f)}{\sigma_X^2(f) + \sigma_N^2(f)} \; Y(t,f) \;\; , \qquad (1)$$

where $Y(t,f)$ is the spectrum at the (discrete) frequency f for the frame t. The estimate $\widehat{x}$ is obtained by summing of the estimated time-frequency components.

In other words, the Wiener filter weights the frequency components of the noisy speech signal according to the signal to noise ratio at each frequency.

2.2. Time-frequency shrinkage

Another approach to denoising is time-frequency shrinkage, where the most popular time-frequency representation is a wavelet representation [6]. Shrinkage with soft thresholding corresponds to estimating x according to the maximum a posteriori (MAP) criterion, under the assumption that N is a white Gaussian noise and that the time-frequency components of X have a Laplacian distribution. With the same notations as before, the estimation is obtained by soft-thresholding of the observed components: if $|Y(t,f)| > \beta(f)$ then

$$\widehat{X}(t,f) = \frac{Y(t,f)}{|Y(i,f)|} \left(|Y(i,f)| - \beta(f) \right) \;\; , \qquad (2)$$

else $\widehat{X}(t,f) = 0$. The threshold $\beta(f)$ is defined as

$$\beta(f) = \lambda \; \frac{\sigma_N^2(f)}{\sigma_X(f)} \qquad (3)$$

where λ is a parameter that controls the amount of noise which should be suppressed. Note that spectral subtraction, traditionally used in speech recognition, is very similar to the shrinkage algorithm detailed above.

Wiener filtering as well as time-frequency shrinkage require a preliminary step where the PSD $\sigma_X^2(f)$ and $\sigma_N^2(f)$ are estimated.

3. DICTIONARY-BASED METHODS

In order to deal with the non-stationary nature of music and speech, we propose to use denoising methods which explicitly model the psd of X and N as mixtures of several psd's rather than a single one, in contrast with the underlying models in Wiener filtering or time-frequency shrinkage. These adaptive source separation methods based on a dictionary of psd's were introduced in [7] and we only recall here their basic principles.

3.1. Principle

In the context of the denoising task under consideration in this work, we assume that we have a set of typical psd vectors, $\sigma_{N,k}^2(f)$ $(k = 1, \ldots, d_N)$ and $\sigma_{X,k}^2(f)$ $(k = 1, \ldots, d_X)$, for the noise and speech signal respectively. Estimation of psd dictionaries is discussed further in section 3.3.

The principle of the method is to estimate the contribution of each of these PSD to the short-time power spectrum $|Y(t,f)|^2$ of the noisy signal. For each frame t, we estimate positive amplitude coefficients $a_{N,k}(t)$ and $a_{X,k}(t)$ for each PSD. These coefficients are estimated according to the maximum Likelihood criterion under a positivity constraint, which roughly corresponds to the non-negative linear decomposition of the $|Y(t,f)|^2$ on the dictionary of available psd's, that is to say:

$$|Y(t,f)|^2 \approx \sum_{k=1}^{d_X} a_{X,k}(t) \; \sigma_{X,k}^2(f) + \sum_{k=1}^{d_N} a_{N,k}(t) \; \sigma_{N,k}^2(f) \;\; . \qquad (4)$$

3.2. Denoising

We propose two denoising approaches based on the amplitude factors corresponding to decomposition (4). The first approach is based on filtering while the second one exploits the principle of shrinkage.

In the filtering approach, the speech signal is estimated through the generalized Wiener filter formula:

$$\widehat{X}(t,f) = \frac{\sum_{k=1}^{d_X} a_{X,k}(t) \; \sigma_{X,k}^2(f)}{\sum_{k=1}^{d_X} a_{X,k}(t) \; \sigma_{X,k}^2(f) + \sum_{k=1}^{d_N} a_{N,k} \; \sigma_{N,k}^2(f)} \; Y(t,f) \;\; .$$

This estimator can be considered as an adaptive Wiener filter (cf. (1)) and corresponds to an underlying model where the observed signal is a sum of $d_N + d_X$ stationary Gaussian sources which are modulated by slowly time-varying amplitude parameters $a_{X,k}(t)$ and $a_{N,k}(t)$.

In the shrinkage approach, denoising is performed by thresholding (cf. (2)) each frame based on the estimated contributions of the noise and the speech signal, with the adaptive threshold

$$\beta(t,f) = \lambda \frac{\sum_{k=1}^{d_N} a_{N,k}(t)\ \sigma_{N,k}^2(f)}{\sqrt{\sum_{k=1}^{d_X} a_{X,k}(t)\ \sigma_{X,k}^2(f)}} \ . \qquad (5)$$

This method is called adaptive shrinkage as the threshold depends on the observed signal y and is therefore time-dependent.

3.3. Estimation of psd dictionaries

The dictionaries of normalized psd's are estimated from a set of training data. The dictionary is first initialized by vector quantization. The estimation algorithm then proceed so as to maximize the likelihood of the training data psd's. The maximization algorithm is similar to the algorithm used to estimate the amplitude factors with an additional step to estimate the psd, given the amplitude factors. This algorithm is described in details in [7].

4. CORPUS

The task considered in this paper is the transcription of read sentences in French. Speech recognition experiments are carried out on a subset of the BREF corpus [8] which contains sentences from the French newspaper "Le Monde", read in a clean studio environment and recorded with a high quality microphone. The test set contains 300 sentences[1], to which background music (jingle) was added at various levels of signal/noise ratios (SNR). In these experiments, the jingle is an instrumental loop of a few seconds which essentially contains low frequency components between 0 and 800 Hz (bass guitar) and transients (drums), as illustrated figure 1.

The use of such a corpus with artificially added music, as opposed to real broadcast news data, is necessary for two reasons. First, it enables a control of the SNR. Second, it also enables to evaluate the denoising performances of the algorithm in terms of distortion, as this evaluation requires the original (noiseless) data. Typical SNRs in real broadcast news data is between 15 and 5 dB, depending on the radio station.

[1] Test set from the AUPELF ILOR-B1 evaluation campaign [9]

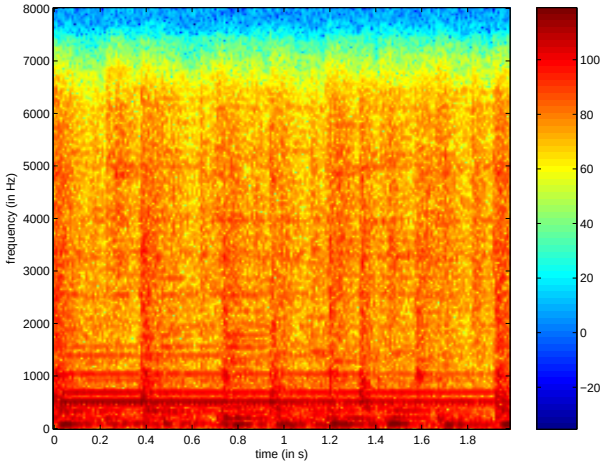

Fig. 1. Spectrogram of (a portion of) the music added to the test set.

In the experiments discussed in the next sections, the psd dictionaries for music were estimated from the jingle itself. For adaptive methods, a dictionary of $d_N = 64$ psd's was estimated. For speech, distinct dictionaries with $d_X = 256$ spectral shapes were estimated for male and female speakers. Both dictionaries were estimated on a subset of 50 utterances different from the test data (different sentences and speakers). Speaker gender and SNR are assumed to be known during tests. Note however that adaptive methods do not make use of the knowledge of the SNR.

5. DENOISING PERFORMANCES

In this section, the methods described previously are compared in terms of source separation and in terms of spectral distortion.

5.1. Source discrimination

In order to evaluate the performance of the various denoising methods in terms of separation of the speech and music, we computed the source to interference ratio (SIR) and source to artifacts ratio (SAR) [10]. The goal of these criteria is to measure separately the distortion level due to the remaining interferences of the unwanted source (music) and that due to artifacts of the algorithm such as nonlinearities. The larger the SIR or SAR figures, the better the performance.

Figure 2 displays the average SIR (top) and SAR (bottom) over 20 sentences from the corpus, for SNRs between 20 and 0 dB. For all the methods, the SAR is smaller than the SIR at all SNRs. Even though all the algorithms e-

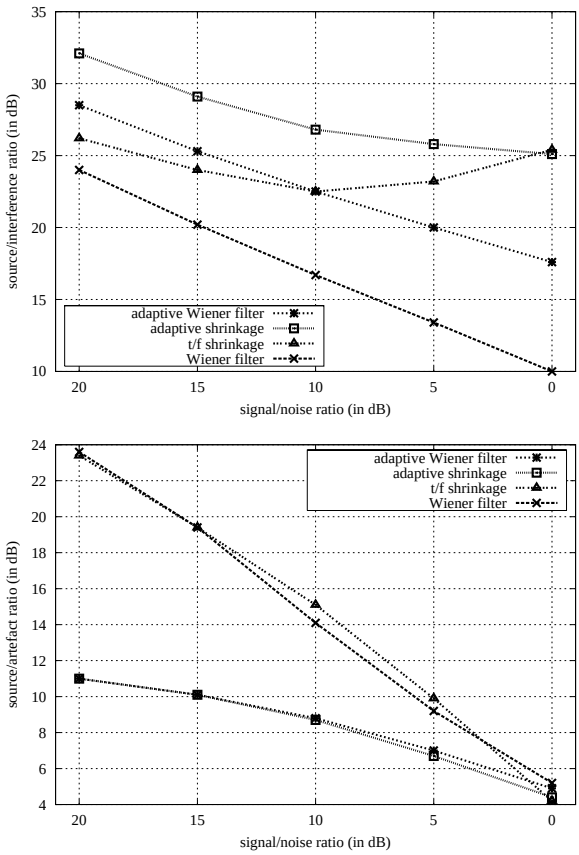

Fig. 2. Source / Interference Ratio (top) and Source / Artifact Ratio (bottom) for the various algorithms.

liminate quite well music, they introduce quite strong non-linearities which generate artifacts that are not inconsiderable.

Unsurprisingly, whatever the denoising method, the SAR decreases as the SNR decreases, and one can clearly observe two groups of methods in terms of SAR. Adaptive methods globally generate more artifacts than the non-adaptive ones. As far as the SIR is concerned, it also decreases when the SNR decreases, except for the time-frequency shrinkage algorithm for which the SIR increases again at very low SNRs. This behavior corresponds to the fact that time-frequency shrinkage suppresses well the music by setting to zero the time-frequency components that are below the threshold. However, this implies that some speech components are also set to zero and the price is that the level of artifacts is quite high, as can be seen on the corresponding SAR curve: for this algorithm, the SAR is very low for SNRs around 0 dB.

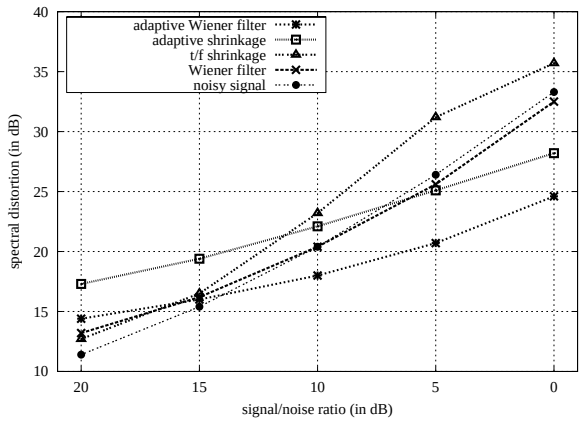

Fig. 3. Spectral distortion for the various noise suppression techniques.

5.2. Spectral distortion

Since the target application is denoising and speech recognition, we also measured the performance of the denoising methods in terms of spectral distortion between the original signal and its estimate. Spectral distortion is a relevant measure of the potential impact of the denoising method of the ASR system, since the latter uses a spectral representation of the signal. The results are depicted figure 3. As a reference, spectral distortion between the original and noisy signals is also given.

At high SNRs, methods based on a Gaussian model of the noise, *i.e.* Wiener filtering and time-frequency shrinkage, result in less spectral distortion than the proposed adaptive methods. However the noisy signal is even less spectrally distorted. On the contrary, for SNRs below 10 dB, one can observe the opposite phenomenon with a clear advantage for the adaptive Wiener filter approach. Generally speaking, filtering methods introduce less spectral distortion than shrinkage methods.

Typical spectrograms before and after denoising are illustrated figure 4 at 10 dB.

6. SPEECH RECOGNITION PERFORMANCES

Speech recognition was carried out with a 20k word vocabulary with a trigram language model and context-independent phone models. Speech signal is represented using 12 cepstral coefficients plus energy, along with their first and second order derivatives. In order to improve robustness, short-term cepstral mean subtraction and variance normalization is used [4], with a 3 s time span. No blind model adaptation is performed.

Figure 5 displays the word error rates (WER) for SNRs between 20 and 0 dB. The error rate obtained on the clean test set (no noise added), around 28%, is also reported on the

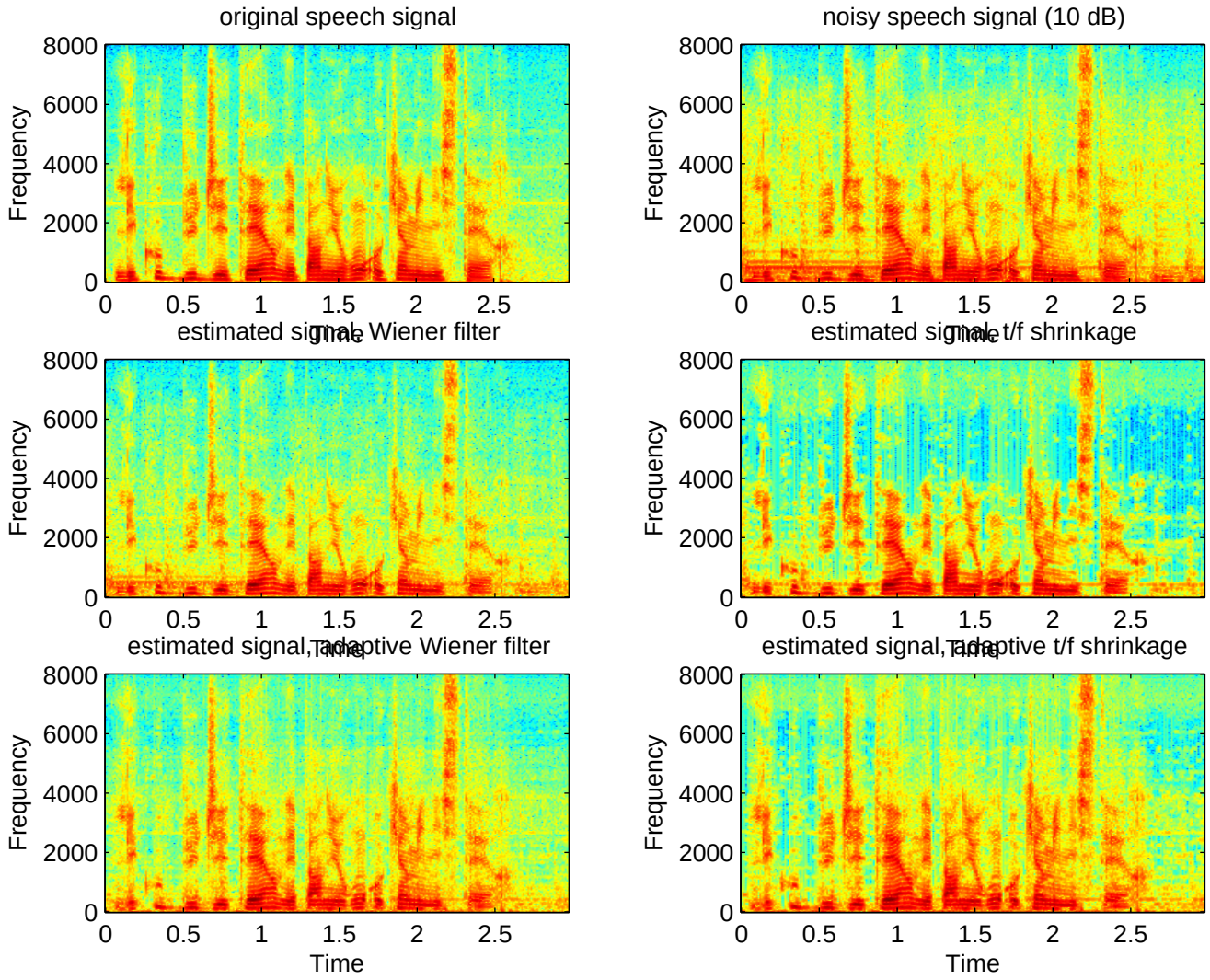

Fig. 4. Spectrograms for one sentence at 10 dB: clean and noisy speech signal (upper rows), non-adaptive methods (middle rows), adaptive methods (lower rows).

figure (SNR = ∞). The WER plots are very similar in shape to the spectral distortion ones. For SNRs below 10 dB, all denoising methods result in an improvement compared to a recognition without prior denoising. Moreover, as soon as the SNR falls below 15 dB, denoising by adaptive Wiener filtering outperforms all other techniques. Indeed, methods based on a stationary Gaussian model of the noise are not very efficient for such a complex noise as music, and thus improving the WER only marginally. As far as adaptive shrinkage is concerned, the recognition process suffers from the high level of artifacts introduced by the denoising algorithm.

7. PERSPECTIVE

In this paper, we proposed signal denoising methods based on single sensor source separation algorithms. The results show that the methods based on time-frequency shrinkage are efficient at removing the noise but introduce important spectral distortions which perturb the recognition system. On the contrary, adaptive Wiener filtering yield a clear improvement of the word error rate in presence of music.

For moderate to low SNRs, we have shown that the proposed adaptive methods allow for an efficient processing of non-stationary noises, with no prior knowledge of the SNR. However, experiments were limited to a well-controlled experimental setup with rather strong hypotheses. In particular, the noise is assumed to be known, which is rarely the

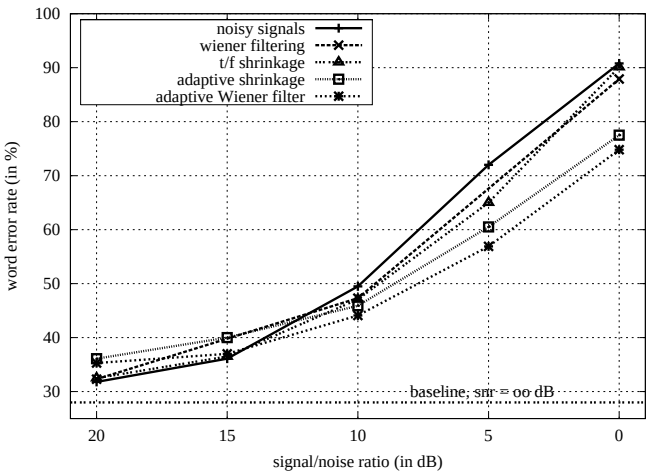

Fig. 5. Performance of an ASR algorithm as a function of the noise level

case in practice! Practically, it will be necessary to estimate the psd dictionary for music on external data or on the result of an automatic music/non-music detector.

Moreover, there is little difference in acoustic quality of the speech between the training and the test data. Thus, the speech psd dictionaries estimated on the training data are well suited to the test data. In practice, recording conditions may vary greatly from one document to another. Also, the jingle can be quite different too. To keep the adaptive Wiener filter efficient, unsupervised adaptation of the psd dictionaries is desirable.

8. REFERENCES

[1] J. Nolazco Flores and S. Young, "Continuous speech recognition in noise using spectral subtraction and HMM adaptation," in *IEEE Conf. on Acoustic, Speech and Signal Processing*, pp. 409–412, 1994.

[2] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. on Speech and Audio Processing*, 2(2), April 1994.

[3] Mark Gales, "Model-based techniques for noise robust speech recognition," Ph. D. Thesis, University of Cambridge, September 1995.

[4] S. Molau, F. Hilger, and H. Ney, "Feature space normalization in adverse acoustic conditions," in *Proc. Intl. Conf. on Speech and Audio Processing*, 2003.

[5] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. on Acoustic, Speech and Signal Processing*, 28(2), 1979.

[6] D. Donoho, "Denoising by soft-thresholding," *IEEE Trans. Inform. Theory*, 41, pp. 613–627, 1995.

[7] Laurent Benaroya, "Séparation de plusieurs sources sonores avec un seul microphone," Ph. D. Thesis, Université de Rennes 1, 2003.

[8] L. Lamel, J.-L. Gauvain, and M. Eskénazi, "BREF, a large vocabulary spoken corpus for French," in *European Conference on Speech Communication and Technologies*, pp. 505–508, 1991.

[9] J.M. Dolmazon, F. Bimbot, G. Adda, M. El-Bèze, J.C. Caërou, J. Zeiliger, M. Adda-Decker, "Organisation de la première campagne AUPELF pour l'évaluation des systèmes de dictée vocale," in *Actes des 1ères Journée Scientifique et Techniques Francil*, pp. 13–18, 1997.

[10] R. Gribonval, L. Benaroya, E. Vincent, and C. Févotte, "Proposals for performance measurement in source separation," in *Proc. 4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003)*, pp 763–768, 2003.