Séparation de sources à partir d'un seul capteur pour la reconnaissance robuste de la parole

Guillaume Gravier, Laurent Benaroya, Alexey Ozerov,
Rémi Gribonval et Frédéric Bimbot

IRISA (CNRS & INRIA), Equipe METISS
Campus de Beaulieu, F-35042 Rennes Cedex, France
Mél: prenom.nom@irisa.fr - http://www.irisa.fr/metiss

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, as it is often the case 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 in order to tackle the problem of 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

La transcription automatique est une étape clé pour l'indexation et la recherche de données dans des documents audio tels que les journaux radiodiffusés. Un problème typique de la transcription de journaux radio est la présence de musique en fond sonore. Les technologies de transcription automatique de la parole, bien qu'arrivées aujourd'hui à une certaine maturité, ont des performances qui se dégradent rapidement en présence de bruit.

Dans les systèmes de reconnaissance automatique de la parole (RAP), la compensation de l'effet du bruit est généralement effectuée à deux niveaux, d'une part au niveau du signal, d'autre part par adaptation non supervisée des modèles utilisées (par exemple, cf. [1]). Pour les systèmes de RAP basés sur une modélisation par modèles de Markov cachés, de nombreuses techniques d'adaptation non supervisée ont été proposées [2, 3]. La nécessité de supprimer l'effet du bruit est également prise en compte dans le choix des coefficients utilisés pour représenter le signal [10]. Par exemple, la soustraction de moyenne cepstrale et la normalisation de variance augmentent la robustesse au bruit. Dans cet article, nous nous intéressons à la première étape, *c.-à-d.* à la suppression de bruit au niveau du signal.

En reconnaissance de la parole, l'approche la plus généralement utilisée pour la compensation de bruit est la soustraction spectrale [4, 1]. Cette technique consiste à soustraire de chaque trame une estimation du spectre du bruit. La plupart des approches proposées dans la littérature pour estimer le spectre du bruit sont basées sur un détecteur parole/non-parole et requièrent un nombre relativement élevé de trames pour obtenir une bonne estimation. Elles sont donc peu appropriées au cas de bruits non-stationnaires tels que la musique. D'autres techniques classiques de débruitage telles que le filtrage de Wiener ou le seuillage en ondelettes [5] sont conçues pour traiter du bruit stationnaire gaussien, modèle qui ne s'applique clairement pas au fond sonore musical des journaux radiodiffusés.

Dans cet article, nous proposons une nouvelle approche pour débruiter des enregistrements de journaux radiodiffusés. Cette approche est basée sur des modèles probabilistes des signaux de parole et de bruit, et utilise des techniques de séparation de sources. Comme la suppression de bruit est un cas particulier de séparation de sources, les performances des algorithmes étudiés sont d'abord comparées avec les mesures de performance standard en séparation de sources, à savoir les rapports source/interférences (RSI) et source/artefacts (RSA). L'apport des différentes méthodes est ensuite étudié au niveau du système de transcription automatique de parole.

Dans la suite, nous rappelons tout d'abord le principe de quelques algorithmes standards de débruitage, à savoir le filtrage de Wiener et le seuillage temps-fréquence. Nous présentons ensuite deux méthodes de débruitage récentes, basées sur des algorithmes de séparation de sources avec un seul capteur : le filtrage de Wiener adaptatif et le seuillage adaptatif. Après une description rapide du corpus utilisé, nous évaluons les performance des différentes méthodes en terme de séparation de sources et de débruitage. Enfin, nous évaluons l'apport du débruitage sur notre système de transcription automatique.

2. ALGORITHMES CLASSIQUES DE DÉBRUITAGE

Le débruitage est un problème de traitement du signal assez courant et largement étudié dans la littérature. Soit $y(t) = x(t) + n(t)$ le signal bruité observé. Le problème est d'obtenir une estimation $\widehat{x}(t)$ du signal original $x(t)$. En termes probabilistes, x et n sont considérés comme des réalisations de variables aléatoires X et N. Selon le modèle employé pour ces variables aléatoires, plusieurs principes de débruitage ont été proposés. Dans cette section, nous rappelons deux des plus classiques méthodes de débruitage : le filtrage de Wiener et le seuillage temps-fréquence.

2.1. *Filtrage de Wiener*

Le filtre de Wiener est l'estimateur qui minimise l'espérance de la distortion quadratique entre le signal original non observé x et son estimation $\widehat{x}$. En pratique, on considère des trames du signal assez courtes pour que le signal et le bruit puissent y être considérés comme gaussiens stationnaires, de densités spectrales de puissances (DSP) respectives $\sigma_X^2(f)$ et $\sigma_N^2(f)$. Sous ces hypothèses, le filtre de Wiener s'écrit dans le domaine fréquentiel comme

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

où $Y(t,f)$ est le spectre du signal observé à la fréquence (discrète) f sur la trame t. Le signal estimé est reconstruit par addition des différentes composantes temps-fréquence ainsi estimées.

En d'autres termes, le filtre de Wiener atténue les composantes temps-fréquence du signal observé avec une pondération qui dépend du rapport signal/bruit à chaque fréquence.

2.2. *Seuillage temps-fréquence*

Une autre approche pour le débruitage est le seuillage temps-fréquence, souvent effectué sur la transformée en ondelettes du signal [5]. Le seuillage "doux" correspond à l'estimation de x au sens du maximum a posteriori (MAP), sous l'hypothèse que N est un bruit gaussien et que les composantes de la transformée temps-fréquence de X ont une distribution Laplacienne. Avec les même notations que précédemment, l'estimation se fait par seuillage doux des composantes : si $|Y(t,f)| > \beta(f)$, alors

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

sinon, $\widehat{X}(t,f) = 0$. Le seuil est défini par

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

où le paramètre λ contrôle la quantité de bruit supprimé.

Le débruitage par filtrage de Wiener ou par seuillage temps-fréquence nécessite une étape d'estimation les spectres de puissance $\sigma_X^2(f)$ et $\sigma_N^2(f)$.

3. MÉTHODES À BASE DE DICTIONNAIRE

Pour traiter le problème de la non stationnarité du bruit et du signal de parole, nous proposons d'utiliser des méthodes basées sur l'utilisation de dictionnaires de DSP, plutôt que d'une seule DSP par source comme dans le cas du filtre de Wiener ou du seuillage temps-fréquence. Ces méthodes adaptatives de séparation de sources à base de dictionnaires de DSP ont été développées dans [6] et nous en rappelons ici le principe de base.

3.1. *Principe*

Dans le contexte du débruitage, nous supposons que nous disposons d'un ensemble de DSP typiques, $\sigma_{N,k}^2(f)$, $k = 1,\ldots,d_N$ pour le bruit et $\sigma_{X,k}^2(f)$, $k = 1,\ldots,d_X$ pour le signal de parole. L'estimation de ces DSP est discutée au paragraphe 3.3.

Le principe de la méthode est d'estimer la contribution de chacune des DSP du dictionnaire au spectre de puissance d'une trame de signal bruité $|Y(t,f)|^2$. Pour cela, on estime un ensemble de coefficients d'amplitude $a_{N,k}(t)$ et $a_{X,k}(t)$ pour chacune des DSP. Ces coefficients sont estimés de manière à maximiser la vraisemblance de $|Y(t,f)|^2$ sous la contrainte que les coefficients sont positifs. Cette estimation peut être vue comme une approximation linéaire à coefficients positifs de $|Y(t,f)|^2$ à partir de l'ensemble des DSP disponibles, soit

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

3.2. *Débruitage*

A partir des facteurs d'amplitude correspondant à la décomposition (4), on construit deux algorithmes de débruitage : par filtrage ou par seuillage.

Dans le cas du filtrage, le signal de parole est alors estimé par le filtrage de Wiener généralisé [6]:

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

Cet estimateur, que l'on peut qualifier de filtrage de Wiener adaptatif (cf. Eq. (1)), correspond à un modèle sous-jacent où le signal y est la somme de $d_N + d_X$ sources gaussiennes stationnaires modulées en amplitude par des amplitudes $a_{X,k}(t)$ et $a_{N,k}(t)$ variant lentement au cours du temps.

Dans le cas du seuillage, l'algorithme consiste à effectuer le seuillage (2) pour chaque trame à partir des contributions estimées du bruit et de la parole, avec le seuil adaptatif

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

Cette méthode est qualifiée de seuillage adaptatif par la suite, dans la mesure où le seuil dépend du signal observé et varie donc en fonction du temps.

Notons que, tant pour le filtrage que pour le seuillage, la connaissance a priori du rapport signal/bruit n'est pas nécessaire puisque les amplitudes respectives sont adaptées automatiquement dans (4).

3.3. *Estimation des DSP*

Les dictionnaires de DSP sont estimés à partir de données d'apprentissage. Après initialisation des dictionnaires par quantification vectorielle des DSP normalisées, chaque dictionnaire est estimé de manière à maximiser la vraisemblance des DSP d'apprentissage. L'algorithme de maximisation suit un schéma similaire à celui de l'estimation des

facteurs d'amplitudes avec une étape supplémentaire d'estimation des DSP connaissant les facteurs d'amplitudes. Cet algorithme est décrit en détails dans [6].

4. CORPUS

La tâche étudiée dans cet article est la transcription de phrases lues, en français. Nous avons effectué des expériences de reconnaissance de la parole sur un sous-ensemble du corpus BREF [7] qui contient des phrases lues en studio dans un environnement non bruité et enregistrées avec un microphone de haute qualité. Les phrases lues sont tirées du journal "Le Monde". Le corpus de test contient 300 phrases[1] auxquelles on a ajouté un fond musical de type *jingle* à divers niveaux de rapport signal / bruit (RSB). Dans ces expériences, le *jingle* est une boucle de quelques secondes de musique comportant essentiellement une composante basse fréquence entre 0 et 800 Hz (basse) et des impulsions (batterie).

L'utilisation d'un tel corpus (plutôt que de données réellement enregistrées lors de journaux radiodiffusés) est dictée par la nécessité de pouvoir contrôler le RSB, d'une part, et de mesurer les performances de l'algorithme de débruitage en termes de distortion, d'autre part, ce qui implique de pouvoir disposer des originaux non bruités et des signaux de bruit. Les RSB rencontrés sur des données réelles de radio se situent entre 15 et 5 dB selon la radio.

Dans les expériences présentées par la suite, les DSP pour la musique sont estimées à partir du *jingle* lui-même. Pour les méthodes adaptatives, un dictionnaire de $d_N = 64$ DSP est utilisé. Pour la parole, nous avons estimé deux dictionnaires de DSP distincts pour les hommes et les femmes, à partir d'un sous-ensemble de 50 phrases (pour chaque sexe) distinctes des données de test. Ces dictionnaires contiennent $d_X = 256$ formes spectrales chacun. Le sexe du locuteur est supposé connu dans toutes les expériences. Pour le filtrage de Wiener et le seuillage temps-fréquence, le RSB est supposé connu tandis que les méthodes adaptatives sont indépendantes du RSB.

5. PERFORMANCES EN DÉBRUITAGE

Pour évaluer les performances des différentes méthodes de débruitage en termes de séparation de la parole et de la musique, nous avons calculé le rapport source / interférences (RSI) et le rapport source / artefacts (RSA) [8]. Ces critères ont pour but de mesurer séparément le niveau de distortion dû à l'interférence résiduelle de la source non désirée (ici la musique) et celui dû à des artefacts introduits, par exemple, par les non-linéarités de l'algorithme de séparation. La performance est d'autant meilleure que le RSI et le RSA sont élevés.

La figure 1 donne la moyenne sur 20 phrases du corpus (10 hommes et 10 femmes) du RSI (en haut) et du RSA (en bas), pour des RSB en entrée allant de 20 à 0 dB. Pour toutes les méthodes et tous les niveaux de RSB, le RSA est inférieur au RSI : bien que les algorithmes éliminent relativement bien la source musique, cela se fait au prix de non-linéarités qui génèrent des artefacts non négligeables.

Assez logiquement, quelle que soit la méthode de débruitage, le RSA se dégrade lorsque le RSB en entrée

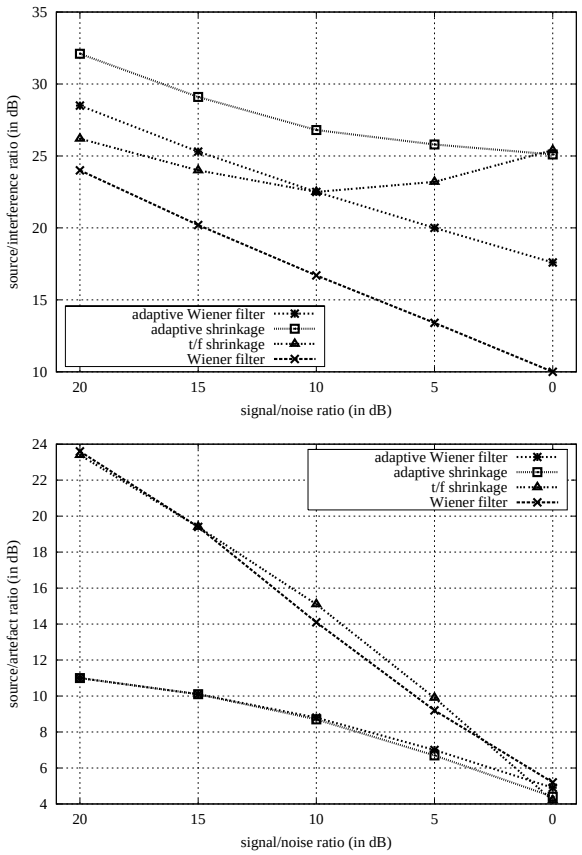

FIG. 1: Rapport source / interférences (en haut) et source / artefacts (en bas) pour les différents algorithmes.

diminue, et l'on observe clairement deux groupes de méthodes en termes de RSA : les méthodes adaptatives engendrent en général plus d'artefacts que les autres. En ce qui concerne le RSI, il décroît également avec le RSB en entrée, à une exception près pour l'algorithme de seuillage temps-fréquence où le RSI croît de nouveau lorsque le RSB d'entrée est faible. Ce comportement correspond au fait que le seuillage temps-fréquence élimine bien la musique en mettant à zéro les composantes temps-fréquence qui ne dépassent pas le seuil. Cependant, cela se fait au prix de l'introduction d'artefacts assez forts, comme on peut l'observer sur la partie correspondante de la courbe de RSA de cet algorithme : pour un RSB en entrée de 0 dB, le RSA se dégrade fortement.

L'application visée étant le débruitage et la reconnaissance automatique de la parole, on évalue également les performances des méthodes de débruitage en terme de distortion spectrale entre le signal d'origine et le signal débruité. La distortion spectrale permet aussi de mesurer l'impact du débruitage sur le système de RAP, ce dernier travaillant sur la base d'une représentation spectrale du signal. Les résultats sont illustrés figure 2, où l'on a également affiché comme référence la distortion spectrale entre le signal d'origine et le signal bruité.

A faible niveau de bruit, les méthodes basées sur le modèle gaussien du bruit (filtrage de Wiener et seuillage temps-fréquence) entraînent moins de distortion spectrale que les méthodes adaptatives proposées, mais c'est encore le signal non débruité qui est le moins distordu spectralement par rapport au signal d'origine. En revanche, pour un RSB inférieur à 10 dB, on observe le phénomène inverse

[1] Corpus de test de l'évaluation AUPELF ILOR-B1.

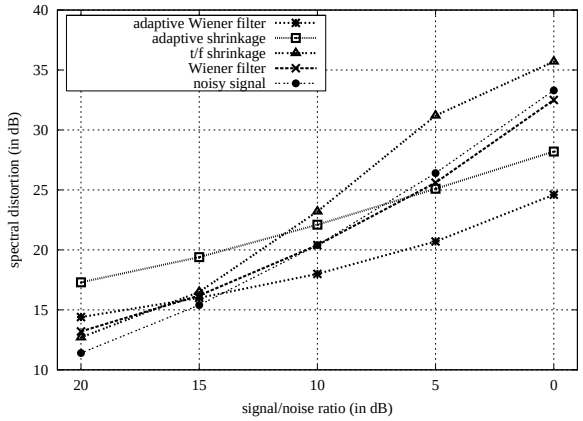

FIG. 2: Distortion spectrale pour les différents algorithmes.

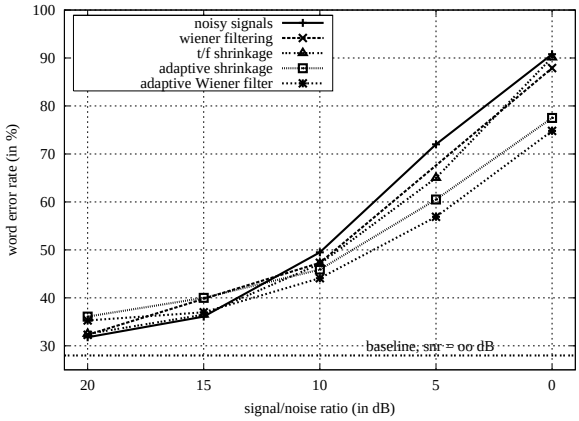

FIG. 3: Performance du système de reconnaissance en fonction du RSB.

avec un net avantage pour le filtrage de Wiener adaptatif. De manière générale, on observe que les méthodes de filtrage introduisent moins de distortions que les méthodes de seuillage.

6. PERFORMANCES EN RECONNAISSANCE

Le système de reconnaissance de la parole [9] utilise un vocabulaire de 20 000 mots avec un modèle de langage trigramme et une modélisation par modèles de Markov cachés de phones hors contexte. Afin d'améliorer la robustesse au bruit et aux distortions introduites par les algorithmes de débruitage, on utilise une représentation cepstrale du signal avec centrage et réduction à court terme [10].

La figure 3 donne le taux d'erreur de mots pour des RSB allant de 20 à 0 dB ainsi que le taux d'erreur pour le système de référence en l'absence de bruit (RSB=∞). Les courbes d'erreur suivent un profil similaire au profil des courbes de distortions spectrales. Pour un RSB inférieur à 10 dB, toutes les méthodes de débruitage apportent une amélioration par rapport à une reconnaissance sans débruitage. De plus, dès que le RSB est inférieur à 15 dB, le débruitage par filtre de Wiener adaptatif donne les meilleurs résultats. En effet, les méthodes qui se basent sur l'hypothèse de gaussianité et de stationnarité du bruit sont peu efficaces pour un bruit complexe comme la musique et n'améliorent que très peu la reconnaissance. Quant au seuillage adaptatif, il introduit de nombreuses distortions.

7. PERSPECTIVES

Dans cet article, nous avons proposé des méthodes de débruitage basées sur des algorithmes de séparation de sources avec un seul capteur. Les résultats montrent que les méthodes basées sur le principe de seuillage du spectre sont efficaces pour enlever le bruit mais introduisent des distortions au niveau du spectre qui pénalisent ensuite le système de reconnaissance. En revanche, la méthode de filtrage de Wiener adaptatif permet une nette amélioration de la transcription en présence de musique.

Pour les RSB modérés ou faibles, nous avons montré que les méthodes adaptatives proposées permettent de traiter efficacement des bruits non stationnaires, sans aucune connaissance du rapport signal / bruit. Cependant, les expériences réalisées ici se limitent à un cadre expérimental contrôlé avec des hypothèses fortes. En particulier, le bruit est supposé connu, ce qui n'est que très rarement le cas dans la pratique! En pratique, il faudra donc estimer des dictionnaire de DSP pour la musique sur des données externes ou sur le résultat d'une détection automatique des zones de musique.

Par ailleurs, il y a peu de différences de qualité acoustique pour la parole entre les données d'apprentissage et de test utilisées. Le dictionnaire de DSP de parole estimé sur les données d'apprentissage est donc bien adapté aux données de test. En pratique, les conditions d'enregistrement peuvent varier fortement d'un document à l'autre. De même, la musique utilisée peut être de nature différente. Pour assurer l'efficacité du filtrage de Wiener adaptatif, il est donc souhaitable de pouvoir adapter de manière non supervisée les dictionnaires de DSP.

RÉFÉRENCES

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

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

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

[7] 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.

[8] 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.

[9] G. Gravier, F. Yvon, B. Jacob and F. Bimbot, "Sirocco, un système ouvert de reconnaissance de la parole," in *Journées d'étude sur la parole*, pp. 273–276, 2002.

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