HAL
open science

Uncertainty-based learning of Gaussian mixture models from noisy data

Alexey Ozerov, Mathieu Lagrange, Emmanuel Vincent

To cite this version:

Alexey Ozerov, Mathieu Lagrange, Emmanuel Vincent. Uncertainty-based learning of Gaussian mixture models from noisy data. [Research Report] RR-7862, 2012. hal-00660689v1

HAL Id: hal-00660689

https://inria.hal.science/hal-00660689v1

Submitted on 17 Jan 2012 (v1), last revised 15 Jul 2012 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Uncertainty-based learning of Gaussian mixture models from noisy data

Alexey Ozerov, Mathieu Lagrange, Emmanuel Vincent

RESEARCH REPORT
N° 7862
January 2012
Project-Teams Metiss

ISSN 0249-6399 ISRN INRIA/RR--7862--FR+ENG

Uncertainty-based learning of Gaussian mixture models from noisy data

Alexey Ozerov*, Mathieu Lagrange†, Emmanuel Vincent

Project-Teams Metiss

Research Report n° 7862 — January 2012 — 24 pages

Abstract: We consider the problem of Gaussian mixture model (GMM)-based classification of noisy data, where the uncertainty over the data is given by a Gaussian distribution. While this uncertainty is commonly exploited at the decoding stage via *uncertainty decoding*, it has not been exploited at the training stage so far. We introduce a new Expectation-Maximization (EM) algorithm called *uncertainty training* that allows to learn GMMs directly from noisy data while taking their uncertainty into account. We evaluate its potential for a speaker recognition task over speech data corrupted by real-world domestic background noise, using a state-of-the-art signal enhancement technique and various uncertainty estimation techniques as a front-end. Compared to conventional training, the proposed algorithm results in 3% to 4% absolute improvement in speaker recognition accuracy by training from either matched, unmatched or multi-condition noisy data. This algorithm is also applicable with minor modifications to maximum a posteriori (MAP) or maximum likelihood linear regression (MLLR) model adaptation and to the training of hidden Markov models (HMMs) from noisy data.

Key-words: noisy data, learning, uncertainty, classification, Gaussian mixture model, expectation-maximization

---

This is the preprint of an article submitted to Computer Speech and Language.

* Alexey Ozerov is with Technicolor Research & Innovation, France. This work was performed while A. Ozerov was with INRIA and partly supported by OSEO, the French State agency for innovation, under the Quaero program.

† Mathieu Lagrange is with STMS - IRCAM - CNRS - UPMC.

RESEARCH CENTRE
RENNES – BRETAGNE ATLANTIQUE

Campus universitaire de Beaulieu
35042 Rennes Cedex

Apprentissage de modèles de mélanges de gaussiennes sur des données bruitées basé sur l'incertitude

Résumé : Nous considérons le problème de la classification de données bruitées par des modèles de mélange de gaussiennes (GMM), où l'incertitude sur les données est décrite par une distribution gaussienne. Bien que cette incertitude ait été exploitée à l'étape de décodage par la méthode d'*uncertainty decoding*, elle n'a pas encore été exploitée à l'étape d'apprentissage. Nous introduisons un nouvel algorithme espérance-maximisation (EM) appelé *uncertainty training* qui permet d'apprendre des GMMs directement sur des données bruitées tout en prenant leur incertitude en compte. Nous évaluons son potentiel pour une tâche de reconnaissance du locuteur sur de la parole superposée à un bruit de fond domestique réel, en utilisant une technique de réhaussement de la parole de l'état de l'art et différentes techniques d'estimation de l'incertitude en entrée. Par rapport à l'algorithme classique d'apprentissage, l'algorithme proposé améliore de 3% à 4% la performance de reconnaissance du locuteur, que les données d'apprentissage correspondent à des conditions de bruit similaires ou dissimilaires de celles de test ou à plusieurs conditions. Cet algorithme est aussi applicable moyennant des modifications mineures à l'adaptation de modèles par maximum a posteriori (AMP) ou par régression linéaire par maximum de vraisemblance (MLLR) et à l'apprentissage de modèles de Markov cachés (HMM) à partir de données bruitées.

Mots-clés : données bruitées, apprentissage, incertitude, classification, modèle de mélange de gaussiennes, espérance-maximisation

1 Introduction

Classification and detection systems often face a variety of distortions (e.g., additive or convolutive) resulting in noisy data. In order to achieve noise robustness, three complementary approaches can be taken. At the signal level, one can apply enhancement techniques such as noise suppression (Ephraim, 1992), source separation (Vincent et al., 2012) or dereverberation (Delcroix et al., 2009). At the feature level, one can define features that are robust to the considered type of noise or to the residual noise after enhancement (Nadeu et al., 1997). Finally, at the classifier level, one can account for possible distortion of the features within the classifier itself. In this paper, we focus on the latter approach considering a Gaussian mixture model (GMM)-based classifier. While our approach is quite general, we mostly consider the classification of speech data in the experimental part and in the examples throughout the paper.

The most straightforward approach to increase the accuracy of the classifier is to train the models over *matched* training data exhibiting the same type and amount of noise as the test data (Droppo and Acero, 2008). Unfortunately, such data are not always available and one may be constrained to use *clean*, *multi-condition* or even *unmatched* training data whose noise characteristics do not match those of the test data. Another approach is to dynamically adapt the models so as to account for the *uncertainty* over the test data induced by noise. This uncertainty is typically encoded either by a set of binary flags indicating whether each data dimension is "observed" or "missing" (Cooke, 2001) or by a Gaussian distribution whose mean and covariance matrix represent, respectively, the estimated underlying clean data and noise covariance (Deng et al., 2005). The latter is more flexible than the former, since it allows to quantify the amount of noise and to account for noise correlation between different data dimensions.

While several algorithms have been derived that exploit either binary or Gaussian uncertainty over the test data (Cooke, 2001; Barker et al., 2005; Deng et al., 2005; Srinivasan and Wang, 2007; Delcroix et al., 2009; Shao et al., 2010; Kolossa et al., 2010), uncertainty over the training data has not been exploited so far. Most approaches (Cooke, 2001; Barker et al., 2005; Deng et al., 2005; Srinivasan and Wang, 2007; Shao et al., 2010) assume *conventional training from clean data*. This training strategy is not always applicable in the case of, e.g., field recording or mobile recording where the whole recording might be corrupted by noise. Also, even when sufficient clean data are available for training, the uncertainty over the test data is never perfectly estimated in practice such that some noise may remain that is not accounted for. Recently, Delcroix et al. (2011) and Kolossa et al. (2011) achieved better results by *conventional training from noisy data*. Nevertheless, this heuristic strategy remains sensitive to mismatched training and test noise conditions and, even in matched conditions, the noise variance is twice overestimated. Indeed, it is taken into account both at the training stage within the model parameters and at the decoding stage within the uncertainty and these two contributions add up.

In order to adress these issues, we propose to exploit uncertainty both over the training and the test data and introduce a new Expectation-Maximization (EM) algorithm that allows to learn GMMs directly from noisy data with Gaussian uncertainty. By analogy with the *uncertainty decoding* algorithm of Deng et al. (2005), we refer to this training strategy as *uncertainty training*. The proposed algorithm generalizes both the algorithm of Ghahramani and Jordan

(1994) for binary uncertainty and the algorithm of Arberet et al. (2012) for Gaussian uncertainty with diagonal covariance and zero-mean GMMs with diagonal covariances, which were applied in different contexts. Furthermore, it is also applicable with minor modifications to maximum a posteriori (MAP) (Gauvain and Lee, 1994) or maximum likelihood linear regression (MLLR) (Leggetter and Woodland, 1995) model adaptation and to the training of hidden Markov models (HMMs). This article expands our preliminary paper (Ozerov et al., 2011) by providing more insight about the proposed algorithm and by extensively evaluating it for a speaker recognition task with real-world data and uncertainties as opposed to toy data and oracle (i.e., ideal) uncertainty.

As a by-product, we also introduce the following two new uncertainty estimators. For the particular task and signal enhancement algorithm employed, we show that the best uncertainty estimator is obtained by computing the uncertainty resulting from multichannel Wiener filtering (MWF) (Fischer and Kammeyer, 1997) in the short term Fourier transform (STFT) domain and propagating it to the Mel Frequency Ceptral Coefficient (MFCC) domain using Vector Taylor Series (VTS) (Moreno et al., 1996). Moreover, for benchmarking purposes, we introduce an oracle *rank-1* uncertainty covariance estimator that outperforms the classical oracle diagonal covariance estimator.

The remaining of the paper is organized as follows. In Section 2, we introduce the notations and briefly recall the state-of-the-art GMM-based generative classification approach including uncertainty decoding. The proposed uncertainty training EM algorithm is then described in Section 3. An exhaustive evaluation of this algorithm is conducted in Section 4 for a speaker recognition task. Finally, we draw some conclusions in Section 5.

2 GMM-based classification and uncertainty decoding

2.1 Conventional training and decoding

Classification is the problem of assigning a sequence of M-dimensional real-valued vectors $\mathbf{y} = \{\mathbf{y}_n\}_{n=1}^N$ to a class C. In the context of audio, the observed vectors are typically feature vectors, e.g., MFCCs, each describing one frame of audio. Each class C is modeled by one GMM

$$\theta = \{\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \omega_i\}_{i=1}^I, \tag{1}$$

where $i = 1, \ldots, I$ are state indices, and $\boldsymbol{\mu}_i$, $\boldsymbol{\Sigma}_i$ and ω_i ($\sum_i \omega_i = 1$) are respectively the mean, the covariance matrix and the weight of the i-th state[1]. In other words, each vector $\mathbf{y}_n$ is modeled as follows:

$$(\mathbf{y}_n | q_n = i) \sim \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \qquad \mathbb{P}(q_n = i) = \omega_i, \tag{2}$$

where q_n denotes the state at time n. Under this model, the likelihood of the observed sequence $\mathbf{y}$ is given by

$$p(\mathbf{y}|\theta) = \prod_{n=1}^N p(\mathbf{y}_n|\theta), \tag{3}$$

[1]For the sake of brevity we omit here the class label C in the set of model parameters θ.

with

$$p(\mathbf{y}_n|\theta) = \sum\nolimits_{i=1}^{I} \omega_i \, \mathcal{N}(\mathbf{y}_n|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \tag{4}$$

where

$$\mathcal{N}(\mathbf{y}_n|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \triangleq \frac{1}{\sqrt{(2\pi)^M |\boldsymbol{\Sigma}_i|}} \left[-\frac{(\mathbf{y}_n - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1} (\mathbf{y}_n - \boldsymbol{\mu}_i)}{2} \right]. \tag{5}$$

Using this formulation, conventional GMM-based generative classification consists of the following two steps (Reynolds, 1995):

1. *Training (or adaptation):* For each class C the corresponding model parameters θ are estimated from some sequence of training vectors by maximizing the likelihood (3). This step may be replaced or completed by an adaptation step from some adaptation data, where the maximum likelihood (ML) criterion (3) is replaced by MAP or MLLR.

2. *Decoding:* For each test sequence $\mathbf{y}$, the likelihood (3) is computed for all classes C and the class is selected for which it is maximum.

Training is typically performed via the EM algorithm (Dempster et al., 1977), considering the state indices $\mathbf{q} = \{q_n\}_{n=1}^N$ as *latent data*. The resulting EM updates are summarized in Algorithm 1.

Algorithm 1 One iteration of the conventional EM algorithm (Dempster et al., 1977) for GMM training from clean or noisy data.

E step. Conditional expectations of natural statistics:

$$\gamma_{i,n} \propto \omega_i \, \mathcal{N}(\mathbf{y}_n|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \quad \text{and} \quad \sum\nolimits_i \gamma_{i,n} = 1. \tag{6}$$

M step. Update GMM parameters:

$$\omega_i = \frac{1}{N} \sum_{n=1}^{N} \gamma_{i,n}, \tag{7}$$

$$\boldsymbol{\mu}_i = \frac{1}{\sum_{n=1}^{N} \gamma_{i,n}} \sum_{n=1}^{N} \gamma_{i,n} \mathbf{y}_{i,n}, \tag{8}$$

$$\boldsymbol{\Sigma}_i = \frac{1}{\sum_{n=1}^{N} \gamma_{i,n}} \sum_{n=1}^{N} \gamma_{i,n} \mathbf{y}_{i,n} \mathbf{y}_{i,n}^T - \boldsymbol{\mu}_i \boldsymbol{\mu}_i^T. \tag{9}$$

2.2 Gaussian uncertainty decoding

In the case of noisy data, one may assume that the observed noisy data, denoted as $\bar{\mathbf{y}}_n$, are distributed as (Deng et al., 2005; Delcroix et al., 2009; Kolossa et al., 2010)

$$(\bar{\mathbf{y}}_n|\mathbf{y}_n) \sim \mathcal{N}(\mathbf{y}_n, \bar{\boldsymbol{\Sigma}}_{\mathbf{y},n}), \tag{10}$$

where $\mathbf{y}_n$ are the underlying clean data, which are themselves distributed according to a GMM, and $\bar{\boldsymbol{\Sigma}}_{\mathbf{y},n}$ is the noise covariance matrix, which is assumed

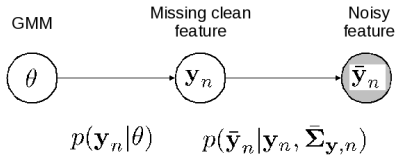

Figure 1: Bayesian network representing the distribution of noisy features with Gaussian uncertainty.

to be known or estimated[2]. The corresponding Bayesian network representation is shown in Fig. 1. Here and in the following, noise may refer either to the original acoustical noise corrupting the features or to the residual noise after signal enhancement as depicted in Fig. 2.

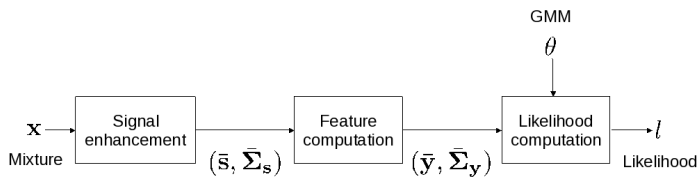

Figure 2: Block diagram of Gaussian uncertainty decoding.

Since the clean data $\mathbf{y}$ are unknown, one cannot directly compute the likelihood (3). It is hence modified by marginalizing over the clean data as (Deng et al., 2005; Kolossa et al., 2010):

$$\begin{aligned} p(\bar{\mathbf{y}}|\bar{\boldsymbol{\Sigma}}_{\mathbf{y}}, \theta) &= \int_{\mathbb{R}^{M\times N}} p(\bar{\mathbf{y}}|\mathbf{y}, \bar{\boldsymbol{\Sigma}}_{\mathbf{y}}) p(\mathbf{y}|\theta) d\mathbf{y} \\ &= \prod_{n=1}^{N} \sum_{i=1}^{I} \omega_i \mathcal{N}(\bar{\mathbf{y}}_n|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i + \bar{\boldsymbol{\Sigma}}_{\mathbf{y},n}) \end{aligned} \tag{11}$$

where $\bar{\mathbf{y}} = \{\bar{\mathbf{y}}_n\}_{n=1}^N$ and $\bar{\boldsymbol{\Sigma}}_{\mathbf{y}} = \{\bar{\boldsymbol{\Sigma}}_{\mathbf{y},n}\}_{n=1}^N$. This quantity can readily be used at the decoding stage, resulting in so-called uncertainty decoding.

3 Proposed uncertainty training algorithm

As discussed in the introduction, state-of-the-art approaches typically train the models either from clean data, as shown in Fig. 3A, or from noisy data, as shown in Fig. 3B, using the conventional training strategy in Section 2.1. By contrast, as shown in Fig. 3C, we propose to train the models over noisy data by maximizing the modified likelihood (11) that accounts for data uncertainty.

[2]Note that assuming a zero-mean noise does not reduce the generality of the approach. In the case of a noise with nonzero mean $\bar{\boldsymbol{\mu}}_{\mathbf{e},n}$ one may simply consider $\bar{\mathbf{y}}_n - \bar{\boldsymbol{\mu}}_{\mathbf{e},n}$ instead of $\bar{\mathbf{y}}_n$.

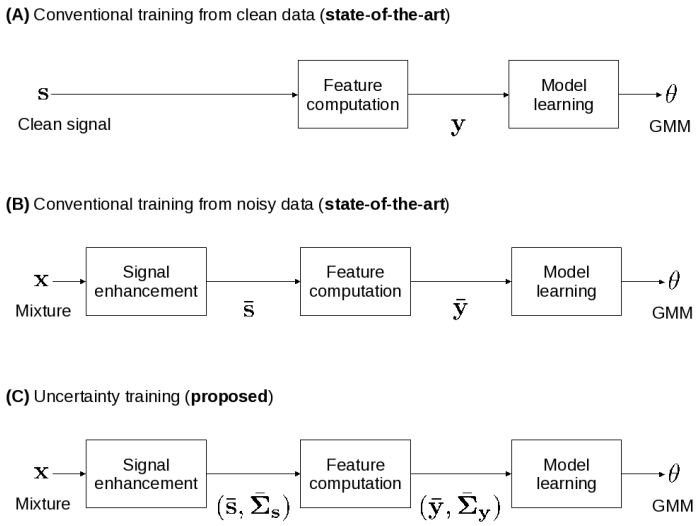

Figure 3: Block diagrams of the state-of-the-art and proposed training strategies.

Contrary to the conventional likelihood (3), the modified likelihood (11) does not explicitly describe the clean data $\mathbf{y}$. Thus, in order to derive an EM algorithm maximizing this likelihood, the latent data including the GMM state indices $\mathbf{q}$ are completed with the clean data $\mathbf{y}$: $\mathcal{B} \triangleq \{\mathbf{y}, \mathbf{q}\}$. Denoting by $\mathcal{A} \triangleq \{\bar{\mathbf{y}}\}$ the observed data, it can be shown that the distribution of the *complete data* $\{p(\bar{\mathbf{y}}, \mathbf{y}, \mathbf{q}|\theta)\}_\theta$ belongs to the *exponential family* (Dempster et al., 1977) and that the set $\mathbf{t}(\mathbf{y}, \mathbf{q}) = \left\{t^0_{i,n}, \mathbf{t}^1_{i,n}, \mathbf{T}^2_{i,n}\right\}_{i,n}$ defined by

$$t^0_{i,n} \triangleq \delta(q_n, i), \qquad \mathbf{t}^1_{i,n} \triangleq \delta(q_n, i)\mathbf{y}_n, \qquad \mathbf{T}^2_{i,n} \triangleq \delta(q_n, i)\mathbf{y}_n\mathbf{y}_n^T, \tag{12}$$

where $\delta(i, j)$ is the Kronecker delta function, is a set of *natural (sufficient) statistics* (Ozerov et al., 2007) for this family.

One iteration of EM then consists of

- computing the expectation of the natural statistics conditionally on the current parameter estimates (E step), and
- re-estimating the parameters from the updated natural statistics by maximizing the conditional expectation of the complete data log-likelihood $Q(\theta|\theta') = \int_{\mathcal{B}} [\log p(\mathcal{A}, \mathcal{B}|\theta)]\, p(\mathcal{B}|\mathcal{A}, \theta') d\mathcal{B}$ (M step).

The resulting updates are given in Algorithm 2. For detailed derivation, please refer to A.

In this algorithm, the uncertainty covariances $\bar{\mathbf{\Sigma}}_{\mathbf{y},n}$ are exploited not only to compute the posterior state probabilities $\gamma_{i,n}$ in (13) as with uncertainty decoding, but also to compute the expectations $\hat{\mathbf{y}}_{i,n}$ and $\widehat{\mathbf{R}}_{\mathbf{yy},i,n}$ in (14) and (15) in the E step. These expectations are actually the first and second order moments of the underlying clean data, which are estimated by the Wiener filter

Algorithm 2 One iteration of the proposed uncertainty training EM algorithm for GMM training from noisy data.

E step. Conditional expectations of natural statistics:

$$\gamma_{i,n} \propto \omega_i \, \mathcal{N}(\bar{\mathbf{y}}_n | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i + \bar{\boldsymbol{\Sigma}}_{\mathbf{y},n}), \quad \text{and} \quad \sum\nolimits_i \gamma_{i,n} = 1, \tag{13}$$

$$\hat{\mathbf{y}}_{i,n} = \mathbf{W}_{i,n} \left(\bar{\mathbf{y}}_n - \boldsymbol{\mu}_i \right) + \boldsymbol{\mu}_i, \tag{14}$$

$$\widehat{\mathbf{R}}_{\mathbf{yy},i,n} = \hat{\mathbf{y}}_{i,n} \hat{\mathbf{y}}_{i,n}^T + \left(\mathbf{I} - \mathbf{W}_{i,n} \right) \boldsymbol{\Sigma}_i, \tag{15}$$

where

$$\mathbf{W}_{i,n} = \boldsymbol{\Sigma}_i \left[\boldsymbol{\Sigma}_i + \bar{\boldsymbol{\Sigma}}_{\mathbf{y},n} \right]^{-1}. \tag{16}$$

M step. Update GMM parameters:

$$\omega_i = \frac{1}{N} \sum_{n=1}^{N} \gamma_{i,n}, \tag{17}$$

$$\boldsymbol{\mu}_i = \frac{1}{\sum_{n=1}^{N} \gamma_{i,n}} \sum_{n=1}^{N} \gamma_{i,n} \hat{\mathbf{y}}_{i,n}, \tag{18}$$

$$\boldsymbol{\Sigma}_i = \frac{1}{\sum_{n=1}^{N} \gamma_{i,n}} \sum_{n=1}^{N} \gamma_{i,n} \widehat{\mathbf{R}}_{\mathbf{yy},i,n} - \boldsymbol{\mu}_i \boldsymbol{\mu}_i^T. \tag{19}$$

$\mathbf{W}_{i,n}$ in (16). This filter is characterized by the covariance $\boldsymbol{\Sigma}_i$ of the clean data, as modeled by the GMM, and the covariance $\bar{\boldsymbol{\Sigma}}_{\mathbf{y},n}$ of the noise, as modeled by the uncertainty. Given these moments, the M step is essentially the same as in Algorithm 1.

In other words, the proposed algorithm alternatingly re-estimates the underlying clean data and their distribution. Contrary to conventional training on noisy data, the estimated model parameters are therefore theoretically noise-free. In practice, they may still be affected by noise to a smaller extent, due to inaccurate estimation of the input uncertainty.

It can easily be shown that the updates in Algorithm 2 are "asymptotically" identical to those of the EM algorithm for binary uncertainty proposed in (Ghahramani and Jordan, 1994) in the case when the uncertainty covariances $\bar{\boldsymbol{\Sigma}}_{\mathbf{y},n}$ are diagonal with either zero entries for observed data or $+\infty$ entries for missing data, as well as to the conventional EM updates in Algorithm 1 in the case when all uncertainty covariances $\bar{\boldsymbol{\Sigma}}_{\mathbf{y},n}$ are zero. Moreover, although the proposed algorithm is presented in the context of GMM training, it can easily be modified to train HMMs or to perform MAP/MLLR adaptation, since only the M step should be modified as in (Rabiner, 1989; Gauvain and Lee, 1994; Leggetter and Woodland, 1995), while the E step remains unchanged for MAP/MLLR adaptation and should just be slightly modified for HMMs by taking transition probabilities into account in the computation of the posterior probabilities (13) as in (Rabiner, 1989).

4 Evaluation

We evaluate the proposed uncertainty training algorithm for a speaker recognition task on speech data corrupted by real-world domestic background noise, using a state-of-the-art signal enhancement technique and various uncertainty estimation techniques as a front-end. We mostly follow the methodology described in the well recognized work of Reynolds (1995) for clean data. We acknowledge that it does not constitute the state-of-the-art method for tackling speaker recognition today. However it provides a simple proof of concept and enables us to focus on the choice of the training, test and uncertainty estimation algorithms as opposed to the settings of the signal enhancer and the classifier. The data and the software used for this experiment are available at `http://www.irisa.fr/metiss/ozerov/Software/SP_REC_Uncrt_MFCC.zip` and `http://www.irisa.fr/metiss/ozerov/Software/GMMUL.zip`, respectively, together with a user guide and examples of use.

4.1 Test methodology

4.1.1 Data

We built a training set, a development set, and a test set by adding binaural clean speech and background noise from the CHiME training corpus (Christensen et al., 2010)[3]. Each of the three datasets involves 680 utterances of approximately 1.5 second duration spoken by 34 speakers (20 sentences per speaker) and continuous domestic background recordings including, e.g., interfering speakers, TV, outside traffic noise or footsteps. All signals are sampled at 16 kHz. The utterances and the background recordings were randomly selected in such a way that no utterance can be found in two datasets and that the background recordings in different datasets were recorded on different days. As such, the background noises in different datasets feature similar acoustical events but the actual signals are distinct.

For each clean speech utterance, different background excerpts were selected according to seven signal to noise ratios (SNRs)[4]: -6, -3, 0, 3, 6, and 9 dB and $+\infty$ (clean). The clean speech signal was then added to the selected background signals, resulting in $7 \times 680 = 4760$ mixtures per dataset. In line with the CHiME challenge (Christensen et al., 2010), we keep track of the temporal position of the selected background excerpts within the continuous background, which enables us to exploit the surrounding background signal for signal enhancement.

4.1.2 Signal enhancement

Signal enhancement is performed via the state-of-the-art algorithm of Ozerov et al. (2012), as implemented using the Flexible Audio Source Sepa-

[3]Our first intention was to use the CHiME test and development datasets as they are. However, clean speech, which is needed later on for benchmarking, was unavailable for the test dataset. We therefore built our own datasets such that they have almost the same characteristics as the CHiME test and development datasets. For that purpose, the CHiME training dataset was selected since it is much bigger than the CHiME development dataset.

[4]Note that, in line with the original CHiME data, no signal scaling was performed to achieve a desired SNR. Instead, for every utterance, we randomly browsed the background until we found a time interval leading to an SNR within a ± 1 dB of the desired SNR.

ration Toolbox (FASST)[5]. This toolbox allows the user to specify the desired spectral and spatial signal models for each sound source from a library of models. Contrary to the use of speaker-dependent models in (Ozerov et al., 2011), target speech is modeled here by a 256-component speaker-independent nonnegative matrix factorization (NMF) spectral model. Background noise is modeled as the sum of 4 sources, each of which follows an 8-component NMF spectral model. In addition, all sources are assumed to follow a rank-1 spatial model. The NMF spectral patterns and the parameters of the spatial model are first trained either on clean speech from the development set or on 20 s of surrounding background noise from the test set (10 s before and 10 s after each utterance). The former are then kept fixed, while the latter are adapted to the test mixture in an unsupervised fashion. The NMF temporal activations are randomly initialized and inferred from the test mixture. Finally, the binaural target speech signal is extracted by MWF. The effectiveness of this signal enhancement algorithm is evaluated in D.1 using standard source separation metrics.

4.1.3 Feature computation

After enhancement, both the mixture signals and the enhanced target speech signals are downmixed to mono by adding both channels together and converted into the time-frequency domain using the STFT with a window size of 1024 samples and 512 samples overlap. 19 MFCCs (2nd to 20th coefficients) are computed for each time frame using the Auditory Toolbox (Slaney, 1998) with default settings. The first MFCC was excluded since it is strongly affected by noise and contains little information about speaker identity.

4.1.4 Uncertainty estimation

The uncertainty over the MFCC features is then estimated using alternative state-of-the-art or novel estimators that we present here. Uncertainty estimation techniques typically consist of the following two steps shown in Fig. 2:

1. estimate uncertainty ($\bar{\mathbf{\Sigma}}_{\mathbf{s}}$) in the complex-valued STFT domain, and
2. propagate it through the corresponding (usually non-linear) feature transform.

STFT-domain uncertainty estimation In the STFT domain, Kolossa et al. (2010) define the uncertainty covariance as a diagonal matrix $\bar{\mathbf{\Sigma}}_{\mathbf{s},n} = \mathrm{diag}\left\{[\bar{\sigma}^2_{\mathbf{s},fn}]_f\right\}$ whose entries are given by

$$\bar{\sigma}^2_{\mathbf{s},fn}(\beta) = \beta\,|\bar{s}_{fn} - x_{fn}|^2\,, \tag{20}$$

where the scaling factor β is optimized on ground truth data as

$$\beta = \arg\min_{\beta'} \sum\nolimits_{f,n} \left(\bar{\sigma}_{\mathbf{s},fn}(\beta') - |\bar{s}_{fn} - s_{fn}|\right)^2 \tag{21}$$

and x_{fn}, $\bar{s}_{fn}$ and s_{fn} denote respectively the STFT coefficients of the mixture signal, the target speech signal and the ground truth clean speech signal in time

[5] `http://bass-db.gforge.inria.fr/fasst/`

frame n and frequency bin f. We denote this estimator as *Beta*, and propose a novel variant denoted *Beta_f*, where the scaling factor β_f depends on f and is optimized according to the frequency-dependent counterpart of (21). The optimal scaling factors for our development set are represented in Fig. 4. Note that, due to the use of different signal enhancement algorithms, $\beta = 0.073$ is 10 times smaller than the optimal β reported in (Kolossa et al., 2010).

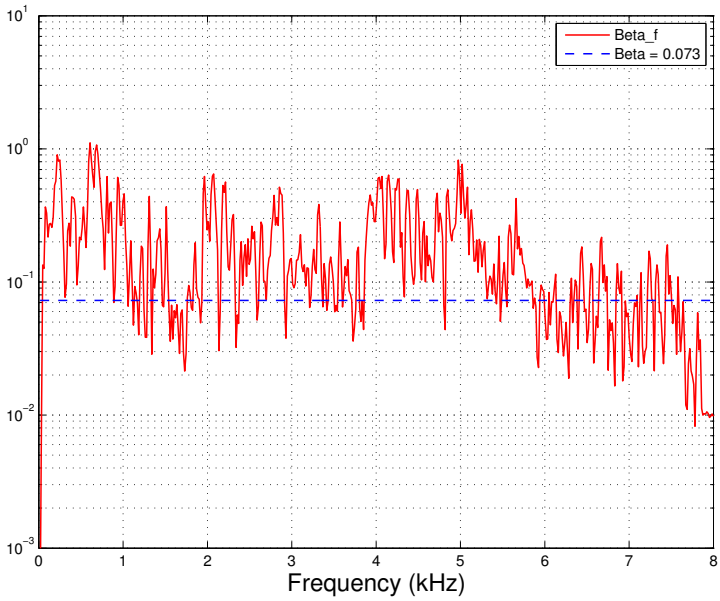

Figure 4: Scaling factors for the STFT-domain uncertainty estimation technique of Kolossa et al. (2010) and the proposed frequency-dependent variant, as optimized on the development set.

In some other work, Kolossa et al. (2011) estimate $\bar{\sigma}^2_{\mathbf{s},fn}$ as the variance of a single-channel Wiener filter applied to the output of a beamformer. This estimator is not directly applicable here due to the use of a MWF in FASST. Instead, we consider the covariance of the MWF whose computation is detailed in B and refer to this estimator as *Wiener*.

Feature-domain uncertainty propagation In order to propagate Gaussian uncertainty from the STFT to the MFCCs, Kolossa et al. (2010) and Adiloğlu and Vincent (2011) use *moment matching (MM)* techniques. The computation of the MFCCs involves two nonlinearities, namely the magnitude of the STFT coefficients and the logarithm of the Mel filterbank outputs. A closed-form solution is derived to match the moments through the first nonlinearity, based on the statistics of the Rice distribution. As for the second nonlinearity, Kolossa et al. (2010) use the *unscented transform*, which is a simplified and efficient version of Monte-Carlo sampling detailed in (Astudillo, 2010), while Adiloğlu and Vincent (2011) use the log-normal transformation of Gales (1995). In our experiments, we call these estimators *MM (unsc.)* and *MM (Gales)*, respectively.

As an alternative to MM, we propose to consider the *VTS* technique that was introduced by Moreno et al. (1996) in the context of feature-domain enhancement. To the best of our knowledge, this technique has not yet been applied in the context of STFT-domain enhancement considered here. Given the non-linear STFT-to-MFCC transform, VTS consists of linearizing this transform by its first-order vector Taylor expansion in the neighborhood of $\bar{\mathbf{s}}_n$. The resulting MFCC uncertainty estimator is detailed in C.

Overall, this results in 9 possible uncertainty estimators including all possible combinations of

- STFT-domain uncertainty estimation: *Beta*, *Beta_f* or *Wiener*, and
- feature-domain uncertainty propagation: *MM (unsc.)*, *MM (Gales)* or *VTS*.

The accuracy of the resulting estimated mean MFCCs $\bar{\mathbf{y}}_n$ is assessed in D.2.

4.1.5 GMM-based classification

Finally, the classifier is built as follows (Reynolds, 1995). The speaker models are 32-state GMMs with diagonal covariance matrices. For each speaker, the GMM parameters are initialized by clustering the corresponding training data (with or without noise) using a hierarchical K-means algorithm and subsequently trained from the same data using either conventional training by Algorithm 1 or uncertainty training by Algorithm 2. For each test utterance, the speaker is selected that maximizes either the conventional likelihood (3) or the uncertainty decoding likelihood (11).

4.2 Main results for the best uncertainty estimator

When running the above experiment on clean training and test data without signal enhancement, 100% recognition accuracy is achieved. This confirms the suitability of the considered classifier as a baseline. In the case of noisy data, we perform a number of experiments specified by

- whether the signal was *enhanced or not*,
- the *decoding strategy:* conventional decoding or uncertainty decoding,
- the *training strategy:* conventional training or uncertainty training.

Furthermore, each experiment is conducted for all possible combinations of the following 8 training and 6 test SNRs:

- training SNR (dB): -6, -3, 0, 3, 6, 9, $+\infty$ (clean), all except $+\infty$ (multi-condition),
- test SNR (dB): -6, -3, 0, 3, 6, 9.

Note that no signal enhancement is applied when training from clean data (see Fig. 3), which corresponds to the state of the art (Deng et al., 2005; Delcroix et al., 2009; Kolossa et al., 2010). Finally, the recognition accuracies are averaged according to four typical *training conditions*:

Enhanced signal	Training strategy	Decoding strategy	Training condition			
			Clean	Matched	Unmatched	Multi
No	Conventional	Conventional	65.17	71.81	69.34	84.09
Yes	Conventional	Conventional	55.22	82.11	80.91	90.12
Yes	Conventional	Uncertainty	**83.48**	87.92	87.19	90.12
Yes	Uncertainty	Uncertainty	**83.48**	**91.79**	**90.61**	**94.04**

Table 1: Main results: average speaker recognition accuracy (in %) for all training and decoding strategies in all training conditions with the Wiener+VTS uncertainty estimator (for detailed results see D.3).

- clean training (training on clean data then average over all test SNRs),
- matched condition training (average over all pairs of equal training and test SNRs),
- unmatched condition training (average over all pairs of distinct training and test SNRs),
- multi-condition training (train on multi-condition data then average over all test SNRs).

Table 1 summarizes the average results obtained for all training and decoding strategies in all training conditions using the Wiener+VTS uncertainty estimator. The corresponding detailed results for all pairs of training and test SNRs are given in D.3. This estimator performed best over all estimators, as will be shown in Section 4.3.

One can see from Table 1 that, in the clean training condition, signal enhancement with conventional training and decoding degrades the speaker recognition accuracy by 10% absolute. However, in all noisy training conditions, signal enhancement with conventional training and decoding systematically improves the performance over "no enhancement" by 6% to 12% absolute. Uncertainty decoding further improves the performance compared to conventional decoding by 0% to 28% absolute depending on the training condition. Finally, uncertainty training combined with uncertainty decoding further increases the accuracy by 3% to 4% absolute in all noisy training conditions[6], compared to the use of uncertainty for decoding alone.

Note that, whatever the chosen training and decoding strategies, clean training performs worse than the other training conditions due to the presence of residual background noise in the test data that is not perfectly accounted for by the estimated uncertainties. Moreover, the best recognition accuracy is achieved for multi-condition training thanks to the fact that the multi-condition training set contains 6 times as many data as the other training sets. If the datasets were balanced, matched condition training would probably have performed better than multi-condition training. However, we believe that our comparison is fair in the sense that it is always much easier to build a huge multi-condition set (virtually any data can be used) than a small matched set.

More detailed analysis is provided in Fig. 5, where the average accuracy resulting from the Wiener+VTS estimator together with uncertainty decoding is

[6]Recall that the uncertainty is zero in the case of clean training, so that conventional training and uncertainty training are equivalent in this case.

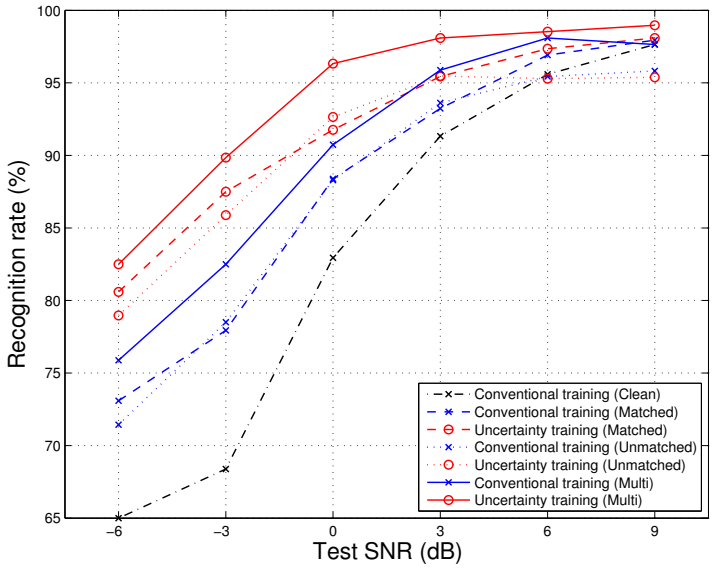

Figure 5: Average speaker recognition accuracy (in %) as a function of the test SNR for all training strategies in all training conditions with the Wiener+VTS uncertainty estimator. The results are averaged over the training SNRs corresponding to each training condition. Uncertainty decoding is performed in all cases.

plotted as a function of the test SNR for all training strategies in all training conditions. Uncertainty training is shown to systematically outperform conventional training for all test SNRs.

4.3 Results for other uncertainty estimators

The goal of the following experiment is to assess the results with different uncertainty estimators. To this aim, we compare the recognition accuracy resulting from the 9 uncertainty estimators in Section 4.1.4 for all training strategies in all training conditions.

Table 2 shows the average accuracies, where uncertainty decoding was performed in all cases. In all training conditions, the best results are obtained by the Wiener+VTS estimator together with uncertainty training. Moreover, uncertainty training outperforms conventional training for each estimator, except for Beta+VTS and Beta_f+VTS in matched or unmatched conditions, and on average over all estimators in all noisy conditions. These results further support the proposed uncertainty training approach and indicate that it is reasonably robust to the choice of the estimator. Note also that MM (unsc.) and MM (Gales) lead to similar performance, the former being slightly better in almost all cases. The proposed Beta_f estimator outperforms the conventional Beta estimator of Kolossa et al. (2010) only in combination with VTS. Finally, it should be noted that the recognition results are loosely correlated with the accuracy of the estimated MFCCs measured in D.2.

Uncertainty estimator	Training condition	Clean	Matched		Unmatched		Multi	
	Training strategy	Conv.	Conv.	Uncrt.	Conv.	Uncrt.	Conv.	Uncrt.
Beta+MM (unsc.)		75.96	78.70	**82.60**	77.79	**81.69**	85.17	**91.18**
Beta_f+MM (unsc.)		73.28	80.07	**81.64**	79.06	**80.32**	84.83	**90.37**
Wiener+MM (unsc.)		41.40	30.27	**68.38**	30.51	**65.89**	37.97	**81.05**
Beta+MM (Gales)		75.51	78.60	**82.87**	77.58	**81.52**	85.02	**91.13**
Beta_f+MM (Gales)		72.70	79.88	**81.32**	78.84	**80.26**	84.73	**90.39**
Wiener+MM (Gales)		41.35	30.20	**68.04**	30.36	**65.01**	37.77	**79.90**
Beta+VTS		77.99	**88.95**	86.69	**88.52**	86.30	92.06	**92.52**
Beta_f+VTS		81.37	**90.15**	88.80	**89.65**	87.98	92.38	**93.80**
Wiener+VTS		83.48	87.92	**91.79**	87.19	**90.61**	90.12	**94.04**
Average over all estimators		69.23	71.64	**81.35**	71.06	**79.95**	76.67	**89.38**

Table 2: Average speaker recognition accuracy (in %) for all training strategies in all training conditions as a function of 9 different uncertainty estimators with full covariance. Uncertainty decoding is performed in all cases.

Uncertainty estimator	Uncertainty covariance	Full		Diagonal	
	Training strategy	Conv.	Uncrt.	Conv.	Uncrt.
Beta+MM (unsc.)		85.17	**91.18**	84.63	90.71
Beta_f+MM (unsc.)		84.83	**90.37**	84.19	89.61
Wiener+MM (unsc.)		37.97	**81.05**	36.81	77.08
Beta+MM (Gales)		85.02	**91.13**	84.58	90.78
Beta_f+MM (Gales)		84.73	**90.39**	84.22	89.80
Wiener+MM (Gales)		37.77	**79.90**	36.89	76.32
Beta+VTS		92.06	**92.52**	91.25	91.42
Beta_f+VTS		92.38	**93.80**	90.76	91.96
Wiener+VTS		90.12	**94.04**	80.47	85.69
Average over all estimators		76.67	**89.38**	74.87	87.04

Table 3: Average speaker recognition accuracy (in %) for all training strategies as a function of 18 different uncertainty estimators, including 9 with full covariance and 9 with diagonal covariance. Multicondition training and uncertainty decoding are assumed in all cases.

All uncertainty estimators presented in Section 4.1.4 lead to full covariance matrices which require the inversion of an $M \times M$ non-diagonal matrix per time frame at each EM iteration (see (16)). Given that the GMM covariances are diagonal, considering diagonal uncertainty covariance matrices would significantly reduce the computational load. Thus, we also evaluate the 9 estimators with diagonal covariances obtained by simply setting to zero the off-diagonal elements of the full uncertainty covariance matrix estimators considered above. As shown in Table 3, using diagonal uncertainty covariances leads to a systematic loss in recognition accuracy in a multi-condition setting. This clearly indicates that the correlation of errors between different feature dimensions is an important point that must be taken into account. Similar behaviour is observed in other training conditions.

4.4 Benchmarking results for oracle uncertainty estimators

In order to demonstrate that the proposed uncertainty training strategy will remain useful in the future even with improved uncertainty estimators, we redo the same experiments with two different oracle uncertainty estimators. By oracle, we mean that the optimal uncertainties are computed from the clean data $\mathbf{y}$ in the ML sense in a benchmarking context.

Deng et al. (2005) constrain the oracle uncertainty covariance to be *diagonal*. In this case the oracle uncertainty is given by $\bar{\mathbf{\Sigma}}_{\mathbf{y},n} = \text{diag}\left\{[\bar{\sigma}^2_{\mathbf{y},mn}]_m\right\}$ with

$$\bar{\sigma}^2_{\mathbf{y},mn} = |\bar{y}_{mn} - y_{mn}|^2. \tag{22}$$

We have found that relaxing this constraint leads to the new oracle estimator

$$\bar{\mathbf{\Sigma}}_{\mathbf{y},n} = (\bar{\mathbf{y}}_n - \mathbf{y}_n)(\bar{\mathbf{y}}_n - \mathbf{y}_n)^T, \tag{23}$$

which is a full matrix of rank 1. This oracle *rank-1* estimator is more informative, since it encodes exactly the direction of the noise $\bar{\mathbf{y}}_n - \mathbf{y}_n$ in $\mathbb{R}^M$ and the only remaining uncertainty is about its position on this line.

Table 4 reports the average speaker recognition accuracy for these two oracle estimators for all training strategies in all training conditions, in a similar way as the two bottom lines of Table 2. We see that uncertainty training outperforms conventional training in all cases and that the oracle rank-1 estimator achieves similar performance to conventional training on clean data. These results provide again a systematic confirmation of the superiority of uncertainty training compared to conventional training and full uncertainty covariances compared to diagonal covariances.

Oracle uncertainty estimator	Training condition	Clean	Matched		Unmatched		Multi	
	Training strategy	Conv.	Conv.	Uncrt.	Conv.	Uncrt.	Conv.	Uncrt.
diagonal		92.92	91.96	**94.71**	92.44	**94.68**	95.32	**97.70**
rank-1		99.66	96.10	**99.46**	96.37	**99.36**	98.75	**99.68**

Table 4: Average speaker recognition accuracy (in %) with uncertainty training and decoding for the two considered oracle uncertainty estimators.

For a closer look at the performance obtained with oracle uncertainty estimators, we display the results as a function of the test SNR in the same way as in Fig. 5. We only show the results for the diagonal estimator, since for the rank-1 estimator the results are very similar to each other and almost reach 100 % accuracy. It appears from Fig. 6 that the qualitative behaviour of these results is very similar to that with the blind estimator in Fig. 5. The absolute improvement of uncertainty training over conventional training is naturally smaller in this oracle setting, but the relative improvement remains comparable.

5 Conclusion

In this paper, we have argued that, when classifying noisy data, uncertainty should be taken into account both during training and decoding. We have introduced a new EM algorithm that allows to learn GMMs from noisy data with

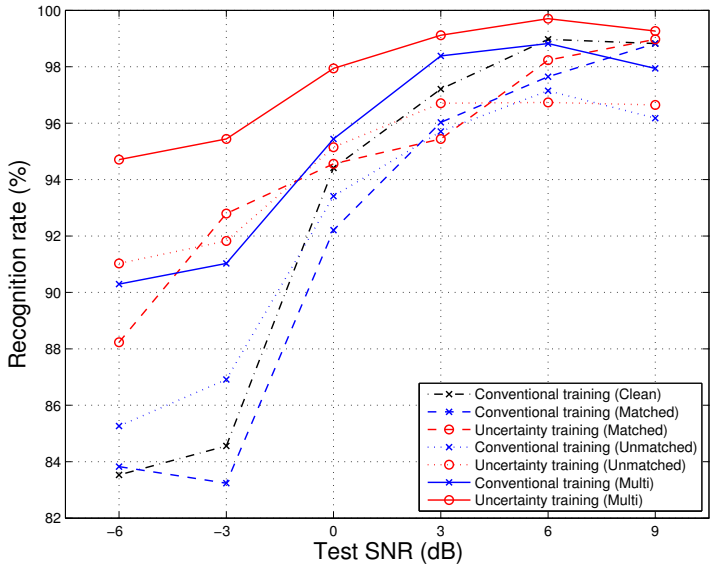

Figure 6: Average speaker recognition accuracy (in %) as a function of the test SNR for all training strategies and training conditions with the oracle diagonal uncertainty estimator. The results are averaged over the training SNRs corresponding to each training condition. Uncertainty decoding is performed in all cases.

Gaussian uncertainty and shown that it outperforms conventional training in both blind and oracle settings for a speaker recognition task in a real-world multisource environment using a state-of-the-art signal enhancement front-end. Extensive evaluation has shown that this algorithm is robust to the training condition (matched, unmatched, or multicondition) and to the choice of the uncertainty estimator. The proposed algorithm performed best when used in conjunction with the VTS uncertainty propagation scheme fed with STFT-domain uncertainty estimates stemming from the MWF.

As already mentioned, it is straightforward to extend this algorithm to the adaptation of GMMs via, e.g., MAP or MLLR, and to the training of HMMs. Thus, it exhibits a great potential for other applications, such as noise-robust speaker diarization or automatic speech recognition. It is also particularly promising for a variety of Music Information Retrieval (MIR) tasks, e.g., singer identification within polyphonic music recordings, where the target sound source is never available in isolation so that clean training is impossible.

Rather then considering binary or Gaussian uncertainty, both the learning and decoding strategies could also be extended to other types of uncertainty. For example the uncertainty could be encoded by a GMM on each time frame.

Finally, since this study constitutes to the best of our knowledge the first use of VTS in the context of STFT-domain speech enhancement, it would be interesting to study its behavior more deeply, e.g., as a function of the SNR.

A Derivation of the proposed uncertainty training algorithm

Let us consider $\mathcal{A} = \{\bar{\mathbf{y}}\}$ as observed data, $\mathcal{B} = \{\mathbf{y}, \mathbf{q}\}$ as latent data, and $\mathcal{C} = \{\bar{\mathbf{y}}, \mathbf{y}, \mathbf{q}\}$ as the complete data. Using (2), (10) and some algebra the negative log-likelihood of the complete data can be written as

$$\begin{aligned}
&-\log p(\bar{\mathbf{y}}, \mathbf{y}, \mathbf{q}|\theta) = -\log p(\bar{\mathbf{y}}|\mathbf{y}) - \log p(\mathbf{y}|\mathbf{q}, \theta) - \log p(\mathbf{q}|\theta) \\
&\stackrel{\mathrm{c}}{=} -\log p(\bar{\mathbf{y}}|\mathbf{y}) + \frac{1}{2}\sum_{i,n} \delta(q_n, i) \left\{\log|\boldsymbol{\Sigma}_i| + (\mathbf{y}_n - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\mathbf{y}_n - \boldsymbol{\mu}_i) - 2\log\omega_i\right\} \\
&= -\log p(\bar{\mathbf{y}}|\mathbf{y}) + \frac{1}{2}\sum_{i,n} \{\log|\boldsymbol{\Sigma}_i|\delta(q_n, i) - 2\log\omega_i\delta(q_n, i) \\
&+\mathrm{trace}\left[\boldsymbol{\Sigma}_i^{-1}(\mathbf{y}_n\mathbf{y}_n^T - \mathbf{y}_n\boldsymbol{\mu}_i^T - \boldsymbol{\mu}_i\mathbf{y}_n^T + \boldsymbol{\mu}_i\boldsymbol{\mu}_i^T)\delta(q_n, i)\right]\} \\
&= -\log p(\bar{\mathbf{y}}|\mathbf{y}) + \frac{1}{2}\sum_{i,n} \left\{(\log|\boldsymbol{\Sigma}_i| - 2\log\omega_i)t_{i,n}^0\right. \\
&\left.+\mathrm{trace}\left[\boldsymbol{\Sigma}_i^{-1}\mathbf{T}_{i,n}^2 - \mathbf{t}_{i,n}^1\boldsymbol{\mu}_i^T - \boldsymbol{\mu}_i(\mathbf{t}_{i,n}^1)^T + \boldsymbol{\mu}_i\boldsymbol{\mu}_i^T t_{i,n}^0)\right]\right\} \quad (24)
\end{aligned}$$

with $t_{i,n}^0$, $\mathbf{t}_{i,n}^1$ and $\mathbf{T}_{i,n}^2$ defined by (12). This expression shows that the log-likelihood of the complete data can be represented in the following form:

$$\log p(\mathcal{A}, \mathcal{B}|\theta) = \langle \eta(\theta), \mathcal{T}(\mathcal{B})\rangle + \nu(\theta) + \phi(\mathcal{A}, \mathcal{B}), \quad (25)$$

where $\mathcal{T}(\mathcal{B})$ is the vector of all scalar elements of $\mathbf{t}(\mathcal{B}) = \left\{t_{i,n}^0, \mathbf{t}_{i,n}^1, \mathbf{T}_{i,n}^2\right\}_{i,n}$, $\eta(\theta)$ and $\nu(\theta)$ are some vector and scalar functions of the parameters θ, and $\phi(\mathcal{A}, \mathcal{B})$ is a scalar function of the complete data. This means that the distribution of the complete data $\{p(\mathcal{A}, \mathcal{B}|\theta)\}_\theta$ belongs to the *exponential family* (Dempster et al., 1977) and that the statistics $\mathbf{t}(\mathcal{B})$ are *natural (sufficient) statistics* (Ozerov et al., 2007) for this family. To derive an EM algorithm in this special case one needs to (i) maximize the likelihood of the complete data (thanks to (25) the ML solution can be always expressed as a function of the natural statistics $\mathbf{t}(\mathcal{B})$), and (ii) replace $\mathbf{t}(\mathcal{B})$ in the ML solution by its conditional expectation $\hat{\mathbf{t}}(\mathcal{A}, \theta') \triangleq \int_{\mathcal{B}} \mathbf{t}(\mathcal{B})p(\mathcal{B}|\mathcal{A}, \theta')d\mathcal{B}$ given the parameters θ' estimated at the previous iteration. By doing so, we obtain the update equations of Algorithm 2.

B Estimation of the STFT-domain Wiener filter covariance with FASST

As mentioned in Section 4.1.2, both the mixture signals $\underline{\mathbf{x}}_{fn}$ and the enhanced target speech signals $\underline{\mathbf{s}}_{fn}$ considered in our experiments are binaural. Denoting by $\tilde{\boldsymbol{\Sigma}}_{\underline{\mathbf{x}},fn}$ and $\tilde{\boldsymbol{\Sigma}}_{\underline{\mathbf{s}},fn}$ the respective prior covariance matrices of these signals as estimated by FASST, the MWF posterior mean and covariance of the multichannel target are given by (Ozerov et al., 2012)

$$\bar{\underline{\mathbf{s}}}_{fn} = \tilde{\boldsymbol{\Sigma}}_{\underline{\mathbf{s}},fn}\tilde{\boldsymbol{\Sigma}}_{\underline{\mathbf{x}},fn}^{-1}\underline{\mathbf{x}}_{fn}, \quad (26)$$

$$\bar{\boldsymbol{\Sigma}}_{\underline{\mathbf{s}},fn} = \left(\mathbf{I} - \tilde{\boldsymbol{\Sigma}}_{\underline{\mathbf{s}},fn}\tilde{\boldsymbol{\Sigma}}_{\underline{\mathbf{x}},fn}^{-1}\right)\tilde{\boldsymbol{\Sigma}}_{\underline{\mathbf{s}},fn}. \quad (27)$$

As the first step towards MFCC extraction, the above signals are downmixed into single-channel mixture x_{fn} and target s_{fn} signals as

$$x_{fn} = \frac{1}{J}\sum_j \underline{x}_{j,fn}, \qquad s_{fn} = \frac{1}{J}\sum_j \underline{s}_{j,fn}, \tag{28}$$

where j denotes the channel index and J the number of channels (in our case $J = 2$). The posterior mean and variance of the single-channel target are then easily derived as

$$\bar{s}_{fn} = \frac{1}{J}\sum_j \bar{\underline{s}}_{j,fn}, \qquad \bar{\sigma}^2_{\mathbf{s},fn} = \frac{1}{J^2}\sum_{j,j'} \bar{\mathbf{\Sigma}}_{\underline{\mathbf{s}},fn}[j,j']. \tag{29}$$

C Vector Taylor series uncertainty estimator for MFCCs

Denoting by $\mathcal{F}(\cdot)$ be the nonlinear transform used to compute a given feature vector (here MFCCs), VTS (Moreno et al., 1996) consists of linearizing this transform by its first-order vector Taylor expansion in the neighborhood of the source estimate $\bar{\mathbf{s}}_n$:

$$\mathbf{y}_n = \mathcal{F}(\mathbf{s}_n) \approx \mathcal{F}(\bar{\mathbf{s}}_n) + J_{\mathcal{F}}(\bar{\mathbf{s}}_n)\,(\mathbf{s}_n - \bar{\mathbf{s}}_n), \tag{30}$$

where $J_{\mathcal{F}}(\bar{\mathbf{s}}_n)$ is the Jacobian matrix of $\mathcal{F}(\mathbf{s}_n)$ computed in $\mathbf{s}_n = \bar{\mathbf{s}}_n$. This leads to the following estimates of the noisy feature value $\bar{\mathbf{y}}_n$ and its uncertainty covariance $\bar{\mathbf{\Sigma}}_{\mathbf{y},n}$, as propagated through this (now linear) transform:

$$\bar{\mathbf{y}}_n = \mathcal{F}(\bar{\mathbf{s}}_n), \qquad \bar{\mathbf{\Sigma}}_{\mathbf{y},n} = J_{\mathcal{F}}(\bar{\mathbf{s}}_n)\bar{\mathbf{\Sigma}}_{\mathbf{s},n}J_{\mathcal{F}}(\bar{\mathbf{s}}_n)^T. \tag{31}$$

In the case of MFCC, $\mathcal{F}(\cdot)$ can be expressed as (see, e.g., (Adiloğlu and Vincent, 2011))

$$\mathbf{y}_n = \mathcal{F}(\mathbf{s}_n) = \mathbf{D}\log(\mathbf{M}|\mathbf{s}_n|), \tag{32}$$

where $\mathbf{D}$ is the $M \times M$ DCT matrix, $\mathbf{M}$ is the $M \times F$ matrix containing the Mel filter coefficients, and $|\cdot|$ and $\log(\cdot)$ are both element-wise operations. With these notations the Jacobian matrix appearing in (31) can be expressed as

$$J_{\mathcal{F}}(\bar{\mathbf{s}}_n) = \mathbf{D}\frac{\mathbf{M}}{\mathbf{M}|\bar{\mathbf{s}}_n|\mathbf{1}_{1\times F}}, \tag{33}$$

where $\mathbf{1}_{1\times F}$ is an $1 \times F$ vector of ones and the magnitude $|\cdot|$ and the division are both element-wise operations.

D Supplementary material and detailed results

D.1 Signal enhancement results

To evaluate the effectiveness of the considered signal enhancement algorithm, we first evaluate speech source separation performance in terms of the SDR, ISR, SIR, and SAR metrics in (Vincent et al., 2012). As suggested in (Ozerov et al.,

2007), we compare these results to reference results obtained by so-called "do nothing separation". These reference results are simply equal to the mixture divided by two, as we separate two sources (target speech and background). The average results over the test set are reported in Table 5. We see that the considered signal enhancement algorithm improves all source separation metrics except the SAR[7]) w.r.t. "do nothing separation" for all SNRs.

SNR		-6 dB	-3 dB	0 dB	3 dB	6 dB	9 dB	Avg.
Source separation	SDR (dB)	2.62	4.23	5.53	6.52	7.18	7.63	5.62
	ISR (dB)	7.63	7.84	8.08	8.45	8.58	8.67	8.21
	SIR (dB)	5.03	7.70	10.79	13.34	15.98	18.58	11.90
	SAR (dB)	9.93	10.91	12.11	13.13	14.00	14.78	12.48
"Do nothing" separation (target = 1/2 mix)	SDR (dB)	-0.90	1.30	3.02	4.26	5.04	5.50	3.04
	ISR (dB)	5.62	5.78	5.89	5.95	5.99	6.00	5.87
	SIR (dB)	-5.41	-2.54	0.31	3.26	6.18	9.16	1.83
	SAR (dB)	$+\infty$	$+\infty$	$+\infty$	$+\infty$	$+\infty$	$+\infty$	$+\infty$

Table 5: Average source separation metrics for the target speech source over the test set.

D.2 Feature enhancement results

Here we evaluate whether the estimation of the mean MFCCs can be improved by signal enhancement alone or whether it must be cascaded with uncertainty propagation. In order to evaluate the quality of feature enhancement we use the Feature to Noise Ratio (FNR) measure we introduced in (Ozerov et al., 2011). The average results over the test set are reported in Table 6. We see that signal enhancement slightly improves the FNR, except for high SNRs (3 and 6 dB), and that the Beta and Beta_f estimators improve the FNR over both the features computed from the mixture and those computed from the enhanced speech, while the reverse is observed for the Wiener estimator.

SNR		-6 dB	-3 dB	0 dB	3 dB	6 dB	9 dB	Avg.
Mixture		4.589	4.770	5.739	7.140	8.181	9.381	6.633
Signal enhancement or VTS		4.987	5.315	6.328	7.419	8.138	8.891	6.846
Uncertainty estimator	Wiener+MM (unsc.)	4.254	4.360	4.987	5.981	6.602	7.133	5.553
	Wiener+MM (Gales)	4.252	4.358	4.984	5.978	6.599	7.130	5.550
	Beta+MM (unsc.)	5.538	5.886	6.989	8.270	9.229	10.302	7.702
	Beta+MM (Gales)	5.538	5.885	6.988	8.270	9.229	10.302	7.702
	Beta_f+MM (unsc.)	5.426	5.722	6.762	8.073	9.025	10.114	7.520
	Beta_f+MM (Gales)	5.425	5.719	6.759	8.071	9.023	10.113	7.518

Table 6: Average FNR (dB) for the mean MFCC features of target speech over the test set. Note that, by definition of VTS, the mean features estimated by Wiener+VTS, Beta+VTS or Beta_f+VTS are equal to those estimated from the enhanced speech signal without uncertainty propagation.

[7]As the "do nothing separation" approach is a linear separation method, it does not introduce any artifacts (SAR = $+\infty$), while the considered non-linear separation method does.

D.3 Detailed speaker recognition results

Table 7 lists the speaker recognition results obtained for every considered pair of training and test SNR conditions and for all training and decoding strategies with the Wiener+VTS uncertainty estimator. These detailed results correspond to the main average results reported in Table 1. One can note that in all the cases the best results lie usually near the diagonal of the 6×6 matrix corresponding to different training and test SNRs, i.e., the matched conditions.

Acknowledgments

The authors would like to thank Kamil Adiloğlu for providing us with some parts of the necessary code for computing the MFCC uncertainty, as well as Ramón Fernandez Astudillo and Dorothea Kolossa for detailed explanation of their work.

References

Adiloğlu, K., Vincent, E., 2011. An uncertainty estimation approach for the extraction of individual source features in multisource recordings. In: EUSIPCO, 19th European Signal Processing Conference.

Arberet, S., Ozerov, A., Bimbot, F., Gribonval, R., 2012. A tractable framework for estimating and combining spectral source models for audio source separation. Signal Processing (to appear).

Astudillo, R. F., 2010. Integration of short-time Fourier domain speech enhancement and observation uncertainty techniques for robust automatic speech recognition. Ph.D. thesis, Technical University Berlin.

Barker, J. P., Cooke, M. P., Ellis, D. P. W., 2005. Decoding speech in the presence of other sources. Speech Communication 45 (1), 5–25.

Christensen, H., Barker, J., Ma, N., Green, P., 2010. The CHiME corpus: a resource and a challenge for computational hearing in multisource environments. In: Proc. Interspeech'10. pp. 1918–1921.

Cooke, M., Jun. 2001. Robust automatic speech recognition with missing and unreliable acoustic data. Speech Communication 34 (3), 267–285.

Delcroix, M., Kinoshita, K., Nakatani, T., Araki, S., Ogawa, A., Hori, T., Watanabe, S., Fujimoto, M., Yoshioka, T., Oba, T., Kubo, Y., Souden, M., Hahm, S.-J., Nakamura, A., 2011. Speech recognition in the presence of highly non-stationary noise based on spatial, spectral and temporal speech/noise modeling combined with dynamic variance adaptation. In: Proc. 1st Int. Workshop on Machine Listening in Multisource Environments (CHiME). pp. 12–17.

Delcroix, M., Nakatani, T., Watanabe, S., 2009. Static and dynamic variance compensation for recognition of reverberant speech with dereverberation preprocessing. IEEE Transactions on Audio, Speech, and Language Processing 17 (2), 324–334.

(A) Conventional training and decoding without signal enhancement

		Test SNR						Average	
		-6 dB	-3 dB	0 dB	3 dB	6 dB	9 dB	Matched	Unmatch.
Training SNR	-6 dB	51.32	49.41	61.32	78.68	84.85	90.00	71.81	69.34
	-3 dB	44.41	48.82	62.94	78.97	86.18	91.91		
	0 dB	45.29	49.85	63.24	78.82	87.94	90.59		
	3 dB	45.29	50.59	65.00	79.85	88.09	92.50		
	6 dB	45.74	48.82	67.06	82.35	91.47	94.71		
	9 dB	41.18	41.47	63.09	81.03	92.06	96.18		
	Clean	40.44	41.32	58.09	74.12	84.71	92.35	65.17	
Multi-condition		63.97	68.38	82.06	93.53	97.94	98.68	84.09	

(B) Conventional training / Conventional decoding with signal enhancement

		Test SNR						Average	
		-6 dB	-3 dB	0 dB	3 dB	6 dB	9 dB	Matched	Unmatch.
Training SNR	-6 dB	62.35	68.09	79.71	89.26	91.76	93.53	82.11	80.91
	-3 dB	61.32	69.41	79.41	89.56	91.03	93.38		
	0 dB	60.59	66.76	79.12	88.68	93.09	93.09		
	3 dB	62.65	69.26	84.56	90.44	93.53	95.88		
	6 dB	58.24	64.41	82.06	92.21	93.68	96.62		
	9 dB	55.44	62.50	81.18	92.50	97.06	97.65		
	Clean	33.53	34.26	47.50	60.74	72.06	83.24	55.22	
Multi-condition		74.41	82.21	90.59	96.62	98.24	98.68	90.12	

(C) Conventional training / Uncertainty decoding with signal enhancement

		Test SNR						Average	
		-6 dB	-3 dB	0 dB	3 dB	6 dB	9 dB	Matched	Unmatch.
Training SNR	-6 dB	73.09	77.94	87.94	94.12	95.59	95.88	87.92	87.19
	-3 dB	72.79	77.94	87.65	91.62	92.94	94.12		
	0 dB	72.06	77.50	88.38	92.06	95.15	95.00		
	3 dB	70.59	78.97	87.50	93.24	96.18	96.76		
	6 dB	69.71	79.26	88.24	94.41	96.91	97.35		
	9 dB	72.06	78.82	90.15	95.88	97.35	97.94		
	Clean	65.00	68.38	82.94	91.32	95.59	97.65	83.48	
Multi-condition		75.88	82.50	90.74	95.88	98.09	97.65	90.12	

(D) Uncertainty training / Uncertainty decoding with signal enhancement

		Test SNR						Average	
		-6 dB	-3 dB	0 dB	3 dB	6 dB	9 dB	Matched	Unmatch.
Training SNR	-6 dB	80.59	85.44	91.62	94.41	93.68	93.97	91.79	90.61
	-3 dB	79.71	87.50	91.32	94.26	94.71	95.00		
	0 dB	79.12	85.74	91.76	94.41	94.56	94.12		
	3 dB	80.15	87.35	92.94	95.44	95.74	96.03		
	6 dB	77.06	85.74	93.68	96.47	97.35	97.79		
	9 dB	78.82	85.15	93.68	97.79	97.79	98.09		
	Clean	65.00	68.38	82.94	91.32	95.59	97.65	83.48	
Multi-condition		82.50	89.85	96.32	98.09	98.53	98.97	94.04	

Table 7: Detailed speaker recognition accuracy (in %) for conventional vs. uncertainty training and decoding after signal enhancement. Both uncertainty training and decoding are based on the Wiener+VTS uncertainty estimator.

Dempster, A. P., Laird, N. M., Rubin., D. B., 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B (Methodological) 39, 1–38.

Deng, L., Droppo, J., Acero, A., 2005. Dynamic compensation of HMM variances using the feature enhancement uncertainty computed from a parametric model of speech distortion. IEEE Transactions on Speech and Audio Processing 13 (3), 412–421.

Droppo, J., Acero, A., 2008. Environmental robustness. In: Benesty, J., Sondhi, M. M., Huang, Y. (Eds.), Handbook of Speech Processing. Springer, pp. 653–680.

Ephraim, Y., 1992. Statistical-model-based speech enhancement systems. Proceedings of the IEEE 80 (10), 1526–1555.

Fischer, S., Kammeyer, K.-D., 1997. Broadband beamforming with adaptive postfiltering for speech acquisition in noisy environments. In: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'97). Vol. 1. pp. 359–362.

Gales, M. J. F., September 1995. Model-based techniques for noise robust speech recognition. Ph.D. thesis, University of Cambridge, UK.

Gauvain, J.-L., Lee, C.-H., 1994. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. IEEE Transactions on Speech and Audio Processing 2 (2), 291–298.

Ghahramani, Z., Jordan, M., 1994. Supervised learning from incomplete data via an EM approach. In: Advance on Neural Information Processing Systems. pp. 120–127.

Kolossa, D., Astudillo, R. F., Abad, A., Zeiler, S., Saeidi, R., Mowlaee, P., da Silva Neto, J., Martin, R., 2011. CHIME challenge: approaches to robustness using beamforming and uncertainty-of-observation techniques. In: Proc. 1st Int. Workshop on Machine Listening in Multisource Environments (CHiME). pp. 6–11.

Kolossa, D., Astudillo, R. F., Hoffmann, E., Orglmeister, R., 2010. Independent component analysis and time-frequency masking for speech recognition in multitalker conditions. EURASIP Journal on Audio, Speech, and Music Processing 2010, 1–14.

Leggetter, C., Woodland, P., 1995. Flexible speaker adaptation using maximum likelihood linear regression. In: ARPA Spoken Lang. Technol. Workshop. pp. 104–109.

Moreno, P. J., Raj, B., Stern, R. M., 1996. A vector Taylor series approach for environment-independent speech recognition. In: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'96). Vol. 2. pp. 733 – 736.

Nadeu, C., Pachès-Leal, P., Juang, B.-H., 1997. Filtering time sequences of spectral parameters for speech recognition. Speech Communication 22, 315–332.

Ozerov, A., Lagrange, M., Vincent, E., September 2011. GMM-based classification from noisy features. In: Proc. 1st Int. Workshop on Machine Listening in Multisource Environments (CHiME). Florence, Italy, pp. 30–35.

Ozerov, A., Philippe, P., Bimbot, F., Gribonval, R., 2007. Adaptation of Bayesian models for single-channel source separation and its application to voice/music separation in popular songs. IEEE Trans. on Audio, Speech and Language Proc. 15 (5), 1564–1578.

Ozerov, A., Vincent, E., Bimbot, F., 2012. A general flexible framework for the handling of prior information in audio source separation. IEEE Transactions on Audio, Speech and Language Processingss (to appear).

Rabiner, L., 1989. A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE 77 (2).

Reynolds, D., 1995. Large population speaker identification using clean and telephone speech. IEEE Signal Processing Letters 2 (3), 46–48.

Shao, Y., Srinivasan, S., Jin, Z., Wang, D., 2010. A computational auditory scene analysis system for speech segregation and robust speech recognition. Computer Speech & Language 24 (1), 77–93.

Slaney, M., 1998. Auditory toolbox version2. Tech. rep., Interval Research Corporation.

Srinivasan, S., Wang, D., 2007. Transforming binary uncertainties for robust speech recognition. IEEE Transactions on Audio, Speech and Language Processing 15 (7), 2130–2140.

Vincent, E., Araki, S., Theis, F. J., Nolte, G., Bofill, P., Sawada, H., Ozerov, A., Gowreesunker, B. V., Lutter, D., Duong, N. Q. K., 2012. The signal separation evaluation campaign (2007–2010): Achievements and remaining challenges. Signal Processing (to appear).

RESEARCH CENTRE
RENNES – BRETAGNE ATLANTIQUE

Campus universitaire de Beaulieu
35042 Rennes Cedex

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-6399