TOWARDS THE USE OF FULL COVARIANCE MODELS FOR MISSING DATA SPEAKER RECOGNITION

Marco Kühne[†]*, Daniel Pullella*[†]*, Roberto Togneri*[†] *and Sven Nordholm*[‡]

[†]School of Electrical, Electronic and Computer Engineering, The University of Western Australia
[‡]Western Australian Telecommunications Research Institute (WATRI)
{marco,daniel,roberto}@ee.uwa.edu.au, sven@watri.org.au

ABSTRACT

This work investigates the use of missing data techniques for noise robust speaker identification. Most previous work in this field relies on the diagonal covariance assumption in modeling speaker specific characteristics via Gaussian mixture models. This paper proposes the use of full covariance models that can capture linear correlations among feature components. This is of importance for missing data marginalization techniques as they depend on spectral rather than cepstral feature representations. Bounded and complete marginalization schemes are investigated both with diagonal and full covariance mixture models. Speaker identification experiments using stationary and non-stationary noise confirm that full covariance models are indeed superior compared to diagonal models.

Index Terms— Missing data, robustness, speaker recognition

1. INTRODUCTION

Speaker recognition technology is one of the key factors in controlling access to personalized communication devices like voice mail, voice dialing or telephone banking. It is the process of automatically recognizing who is speaking based on speaker specific characteristics extractable from speech signals. This work deals with text-independent speaker identification (TI-SID) where the objective is to find the speaker that best matches the incoming utterance. The use of Gaussian mixture models (GMMs) has long been established as state-of-the-art for TI-SID [1]. Each speaker's voice is modeled via a GMM representing broad phonetic classes (like vowels, nasals or fricatives) that correspond to speaker specific vocal tract configurations. However, in real world scenarios the utterance will most likely contain additional background noise or other competing speakers. Speaker models are usually trained on clean data and it is this mismatch between training and testing environments that is responsible for the low performance under practical conditions.

One promising solution to achieve noise robustness can be found in the missing data (MD) paradigm which was introduced for speech recognition [2, 3] and was also successfully applied for noise robust speaker recognition [4, 5]. Missing data recognition is based on the observation that under noisy conditions only parts of a spectral feature vector are corrupted while the remaining components stay relatively unaffected by the noise. The classification of a partly corrupted feature vector can then be performed on the reliable parts only, thus effectively ignoring noise contaminated components. If the decision about the reliability of the spectral components can be made with absolute certainty MD systems can achieve recognition performance close to clean conditions even under highly adverse signal-to-noise-ratios (SNR) [3].

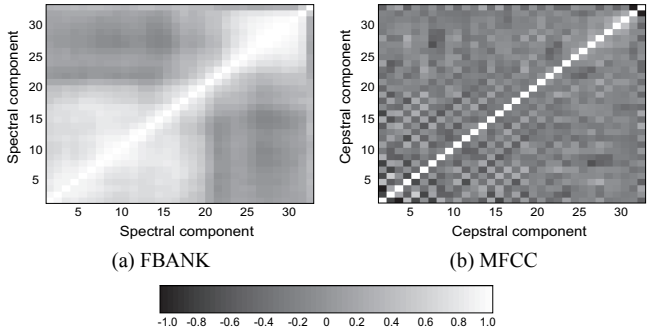


Fig. 1. Correlation maps for log-spectral FBANK (a) and cepstral MFCC (b) features for the TiDigit utterance "3o33951".

The use of full covariance models has not received much attention in the field of speaker recognition. Rather it is common to employ GMMs with diagonal covariance matrices in order to reduce the number of model parameters and save computation time [1, 4, 5]. However, in contrast to cepstral features for which the diagonal covariance assumption is justified log-spectral features exhibit a high correlation among their components (see Fig. 1). Because MD marginalization depends on spectral rather than cepstral features capturing the correlation among feature components is essential for an appropriate statistical modeling. So far, marginalization based on full covariance structures was only considered for isolated vowel classification [6] and speech recognition [2]. Unfortunately, the high computational cost associated with full covariances prohibits their use in speech recognition. On the other hand, the considerably simpler architecture of speaker recognition systems makes those models computationally feasible. However, the increase in model complexity is only acceptable if it is compensated for by gains in recognition accuracy. To the best of the author's knowledge this paper constitutes the first attempt to explore the use of full covariance models for speaker recognition within the MD framework. Building upon [2, 6] the focus is on quantizing the performance difference between diagonal and full covariance GMMs when scoring is performed on masked log-spectral features.

The reminder of this paper is as follows. Section 2 briefly reviews GMM based speaker identification using the MD approach. In particular, the use of full and diagonal covariance models with and without bounded marginalization is discussed. Section 3 presents the experimental evaluation together with a short discussion on the obtained results. The paper closes in Section 4 and presents the main conclusions from this work.

2. ROBUST SCORING USING MISSING DATA THEORY

Let $p(\vec{x}|\lambda)$ denote a Gaussian mixture density given by

$$p(\vec{x}|\lambda) = \sum_{i=1}^{M} c_i \mathcal{N}(\vec{x}; \vec{\mu}_i, \Sigma_i) \quad (1)$$

where $\vec{x} = (x_1, x_2, \ldots, x_D)' \in \mathbb{R}^D$ is a D-dimensional feature vector, $c_i \in [0,1]$ are the mixture component weights and $\mathcal{N}$ is a D-variate Gaussian

$$\mathcal{N}(\vec{x}; \cdot) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}(\vec{x}-\vec{\mu}_i)'\Sigma_i^{-1}(\vec{x}-\vec{\mu}_i)\right\} \quad (2)$$

with mean vector $\vec{\mu}_i \in \mathbb{R}^D$ and covariance matrix $\Sigma_i \in \mathbb{R}^{D\times D}$. Each speaker is represented by such a GMM and we follow [1] in denoting the complete set of parameters by

$$\lambda = \{c_i, \vec{\mu}_i, \Sigma_i\}, \qquad i = 1, 2, \ldots, M. \quad (3)$$

The estimation of the parameter set λ is usually done on clean data via EM training and does not require any modifications with respect to the MD framework. However, during testing the computation of $p(\vec{x}|\lambda)$ has to be adapted in order to take missing components in $\vec{x}$ into account. Let $X = \{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_T\}$ be a possibly noise corrupted observation sequence and $\lambda_1, \lambda_2, \ldots, \lambda_S$ a set of trained speaker models representing a group of S speakers. The goal is to find the speaker that maximizes the log-likelihood given X and λ_s via

$$\hat{S} = \underset{1\le s\le S}{\operatorname{argmax}} \sum_{t=1}^{T} \log p(\vec{x}_t|\lambda_s). \quad (4)$$

To accommodate noise corrupted components the observation vector $\vec{x}$ as well as the model parameters $\vec{\mu}_i, \Sigma_i$ can be separated as

$$\vec{x} = (\vec{x}_r, \vec{x}_u)', \ \vec{\mu}_i = (\vec{\mu}_{r_i}, \vec{\mu}_{u_i})' \text{ and } \Sigma_i = \begin{bmatrix} \Sigma_{rr_i} & \Sigma_{ru_i} \\ \Sigma_{ur_i} & \Sigma_{uu_i} \end{bmatrix}, \quad (5)$$

where the time subscript t is dropped for convenience and reliable and unreliable components are marked by r and u respectively. Using the marginalization approach [2, 3, 7] $p(\vec{x}|\lambda)$ is replaced by the marginal density

$$p(\vec{x}_r|\lambda) = \sum_{i=1}^{M} c_i \mathcal{N}(\vec{x}_r; \vec{\mu}_{r_i}, \Sigma_{rr_i}) \int_{\vec{x}_l}^{\vec{x}_h} \mathcal{N}(\vec{x}_u; \vec{\mu}_{u|r_i}, \Sigma_{u|r_i}) d\vec{x}_u, \quad (6)$$

where the conditional mean $\vec{\mu}_{u|r_i} \in \mathbb{R}^{D_u}$ and covariance matrix $\Sigma_{u|r_i} \in \mathbb{R}^{D_u \times D_u}$ are:

$$\vec{\mu}_{u|r_i} = \vec{\mu}_{u_i} + \Sigma'_{ru_i}\Sigma_{rr_i}^{-1}(\vec{x}_{r_i} - \vec{\mu}_{r_i}) \quad (7)$$

$$\Sigma_{u|r_i} = \Sigma_{uu_i} - \Sigma'_{ru_i}\Sigma_{rr_i}^{-1}\Sigma_{ru_i}. \quad (8)$$

The computation of $p(\vec{x}_r|\lambda)$ depends on the employed covariance structure Σ and the choice of the integration bounds $\vec{x}_l$, $\vec{x}_h \in \mathbb{R}^{D_u}$. For log-spectral features in additive noise the clean speech value is confined to the interval between zero and the observed value which serves as motivation for the choice of the integration limits in bounded marginalization techniques [3]. In the following we discuss four cases based on complete and bounded marginalization with diagonal and full covariance models respectively.

2.1. Full covariances with complete marginalization

Let the integration bounds be $\vec{x}_l = (-\infty, \ldots, -\infty)'$ and $\vec{x}_h = (\infty, \ldots, \infty)'$ and let Σ_{rr_i} be a full covariance matrix. The unreliable components are then completely marginalized by integrating over the entire conditional Gaussian which simplifies Eq. (6) to

$$p(\vec{x}_r|\lambda) = \sum_{i=1}^{M} c_i \mathcal{N}(\vec{x}_r; \vec{\mu}_{r_i}, \Sigma_{rr_i}), \quad (9)$$

where $\mathcal{N}$ is a D_r-variate Gaussian as defined in Eq. (2).

2.2. Diagonal covariances with complete marginalization

Let the integration bounds be $\vec{x}_l = (-\infty, \ldots, -\infty)'$ and $\vec{x}_h = (\infty, \ldots, \infty)'$ and let Σ_i be a diagonal covariance matrix. Then, Σ_{rr_i} is also diagonal with variances $\sigma^2_{r_{ij}}$ along its main diagonal. By exploiting that for multivariate Gaussians zero correlation implies independence [7], Eq. (9) further simplifies to

$$p(\vec{x}_r|\lambda) = \sum_{i=1}^{M} c_i \prod_{x_j \in \vec{x}_r} \mathcal{N}\left(x_j; \mu_{r_{ij}}, \sigma^2_{r_{ij}}\right), \quad (10)$$

where

$$\mathcal{N}(x_j; \cdot) = \frac{1}{\sqrt{2\pi}\sigma_{ij}} \exp\left\{-\frac{1}{2}\frac{(x_j - \mu_{ij})^2}{\sigma^2_{ij}}\right\} \quad (11)$$

is now a univariate Gaussian with mean μ_{ij} and variance σ^2_{ij}.

2.3. Full covariances with bounded marginalization

Let the integration bounds be $\vec{x}_l = (0, \ldots, 0)'$ and $\vec{x}_h = \vec{x}_u$ and let Σ_{rr_i} and $\Sigma_{u|r_i}$ be full covariance matrices. In order to restrict the computation time for the evaluation of the multivariate integral in Eq. (6), $\Sigma_{u|r_i}$ was approximated as a diagonal matrix $\hat{\Sigma}_{u|r_i}$ with variances $\hat{\sigma}^2_{u|r_{ik}}$. Taking advantage of the diagonal structure of $\hat{\Sigma}_{u|r_i}$ Eq. (6) can be approximated as

$$p(\vec{x}_r|\lambda) \approx \sum_{i=1}^{M} c_i \mathcal{N}(\vec{x}_r; \vec{\mu}_{r_i}, \Sigma_{rr_i}) \prod_{\tilde{x}_k \in \vec{x}_u} \int_0^{x_k} \mathcal{N}(\tilde{x}_k; \mu_{u|r_{ik}}, \hat{\sigma}^2_{u|r_{ik}}) d\tilde{x}_k, \quad (12)$$

where $\mathcal{N}(\cdot; \vec{\mu}_{r_i}, \Sigma_{rr_i})$ is a D_r-variate and $\mathcal{N}(\cdot; \mu_{u|r_{ik}}, \hat{\sigma}^2_{u|r_{ik}})$ is a univariate Gaussian as defined in Eq. (2) and (11). Although simply ignoring the off-diagonal elements in $\Sigma_{u|r_i}$ is a very crude approximation the bounded univariate integrals still contain sufficient information to improve recognition accuracy (see Section 3.2). More sophisticated strategies for the evaluation of the multivariate integral can be employed at the expense of a higher computation time [2, 7].

2.4. Diagonal covariances with bounded marginalization

Let the integration bounds be $\vec{x}_l = (0, \ldots, 0)'$ and $\vec{x}_h = \vec{x}_u$ and let Σ_i be a diagonal covariance matrix. Then, Σ_{rr_i} and $\Sigma_{u|r_i}$ are also diagonal with variances $\sigma^2_{r_{ij}}$ and $\sigma^2_{u_{ik}}$ respectively. Hence, by exploiting the independence between feature components Eq. (6) gives

$$p(\vec{x}_r|\lambda) = \sum_{i=1}^{M} c_i \prod_{x_j \in \vec{x}_r} \mathcal{N}(x_j; \mu_{r_{ij}}, \sigma^2_{r_{ij}}) \prod_{\tilde{x}_k \in \vec{x}_u} \int_0^{x_k} \mathcal{N}(\tilde{x}_k; \mu_{u_{ik}}, \sigma^2_{u_{ik}}) d\tilde{x}_k \quad (13)$$

with univariate Gaussians $\mathcal{N}$ as defined in Eq. (11). All univariate integrals in Eq. (12) and (13) can be evaluated using the standard Gaussian error function [3].

3. EXPERIMENTAL EVALUATION

3.1. Setup

Speaker identification experiments were conducted using the TiDigits database. All speech utterances contained connected digits sampled at 20 kHz. Note that short utterances with only one digit were not removed from the data set making it a challenging task for speaker identification. A subset of 31 speakers (21 male, 10 female) was randomly selected for evaluation purposes. For each speaker 50 of the available 77 utterances were randomly chosen for model training while the remaining 27 samples were used for testing. The Hidden Markov Toolkit (HTK) was employed to learn 1,2,8,16 and 32 mixture component GMMs for each speaker. The feature extraction used a 25 ms Hamming window with a frame step of 10 ms. A 48-channel HTK mel-filterbank was used to produce 48 log-spectral energy vectors (FBANK) for the MD system. To obtain a baseline speaker identification system similar to [5] 24 cepstral features (MFCC_Z) were derived from the log-spectral energy vectors followed by cepstral mean normalization (CMN). All test utterances were corrupted by white and factory noise taken from the NOISEX database with SNRs of -5 dB to 35 dB to simulate additive noise conditions. Factory noise possesses significant energy below 2 kHz and because of its non-stationary nature resembles practical conditions more realistic than white noise. Two methods were tested for constructing the MD mask needed to perform the segmentation in Eq. (5). The first technique called oracle masking (Oracle) is impractical as it utilizes the clean speech and the noise signal [5]. It is used here to demonstrate the upper performance limit of the MD approach given a highly accurate mask. Let x^s_{tf} and x^n_{tf} denote the spectral energies at time t and feature component f for the clean and the noise signal respectively. The oracle mask is defined as

$$M^{\text{o}}(t,f) := \begin{cases} 1 & \text{if } x^s_{tf} \geq x^n_{tf}, \\ 0 & \text{otherwise.} \end{cases} \tag{14}$$

The second technique called spectral substraction (SS) can be implemented in practice but is mainly applicable for stationary-noise types [3, 4]. Let $\hat{x}^s_{tf}$ and $\hat{x}^n_{tf}$ denote estimates for the clean and noise spectral energies and let x_{tf} be the noise corrupted observation. The SS mask is then defined as

$$M^{\text{ss}}(t,f) := \begin{cases} 1 & \text{if } \hat{x}^s_{tf} \geq \hat{x}^n_{tf}, \\ 0 & \text{otherwise,} \end{cases} \tag{15}$$

where the estimates for the noise and clean speech energy are

$$\hat{x}^n_{tf} = \frac{1}{T_{\text{avg}}} \sum_{\tau=1}^{T_{\text{avg}}} x_{\tau f}, \quad \text{and} \quad \hat{x}^s_{tf} = x_{tf} - \hat{x}^n_{tf} \tag{16}$$

and $\hat{x}^n_{tf}$ is estimated over the first $T_{\text{avg}} = 10$ frames of an utterance.

3.2. Results

Model order

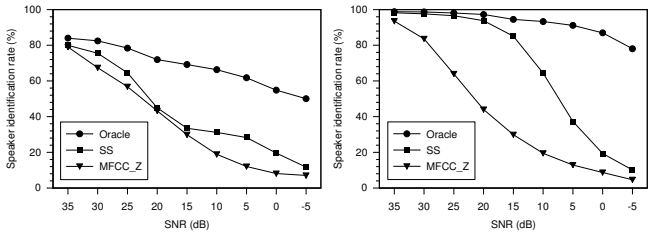
The first experiment determined the speaker identification rate in clean conditions for different GMM configurations and feature parameterizations (see Tab. 1). The best result was achieved using a full covariance model with 16 mixtures and log-spectral features. For both feature types and diagonal covariances increasing the number of mixtures had a positive impact on performance. However, the boost was more distinctive for the diagonal case than for its full counterpart where the one mixture case already outperformed any other diagonal model. In addition, full covariances were more beneficial for log-spectral features compared to cepstral coefficients which is intuitive as FBANKs produce high values in the off-diagonal covariance elements (see Fig. 1). However, as MFCCs are orthogonalized they can achieve better performance using diagonal models independently of the number of mixtures used. For comparison purposes all GMMs in the following experiments used 16 mixture components.

Table 1. Speaker identification results in % for full and diagonal covariance models in clean conditions.

	Diagonal Covariance		Full Covariance	
Mixtures	FBANK	MFCC_Z	FBANK	MFCC_Z
1	49.22	77.30	98.21	97.73
8	84.47	89.96	99.16	97.85
16	87.22	91.88	99.28	96.89
32	88.17	91.88	99.04	96.65

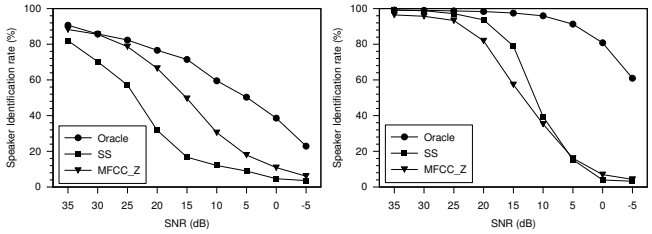
Complete marginalization

The second experiment determined the speaker identification rate under noisy conditions for the two MD systems using complete marginalization and the cepstral baseline (see Fig. 2 and 3).



(a) diagonal covariance (b) full covariance

Fig. 2. Speaker identification rates for white noise and complete marginalization of missing feature components.



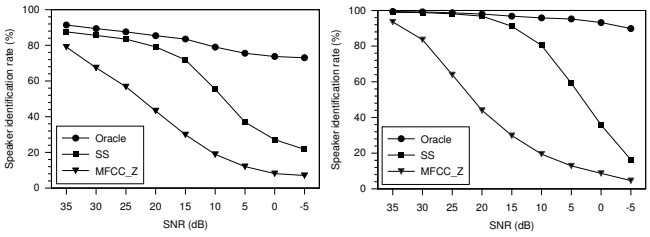
(a) diagonal covariance (b) full covariance

Fig. 3. Speaker identification rates for factory noise and complete marginalization of missing feature components.

For complete marginalization the SS mask outperformed the cepstral baseline for all SNRs with white noise corruption. This is expected as these masks are designed for stationary noise environments. Consequently, for non-stationary factory noise and diagonal covariances the SS mask failed completely (Fig.3a). The performance of the oracle mask for diagonal covariances was better than baseline but disappointingly low for such an accurate mask (Fig.2a,3a). In contrast, the oracle mask results for full covariance models demonstrated a high robustness against both types of noise (Fig.2b,3b). In addition, modeling the feature correlation enabled the SS mask to significantly outperform the cepstral baseline even for factory noise for some SNRs.

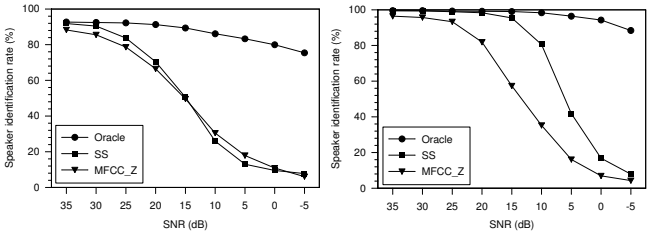
Bounded marginalization

The last experiment determined the speaker identification rate under noisy conditions for the two MD systems using bounded marginalization and the cepstral baseline (see Fig. 4 and 5).



(a) diagonal covariance (b) full covariance

Fig. 4. Speaker identification rates for white noise and bounded marginalization of missing feature components.



(a) diagonal covariance (b) full covariance

Fig. 5. Speaker identification rates for factory noise and bounded marginalization of missing feature components.

By integrating over unreliable components information in missing features is exploited and therefore higher speaker identification rates were achieved compared to simply ignoring these components. The outcomes regarding diagonal and full covariance models basically follow the same trends as observed for the complete marginalization in terms of recognition accuracy improvements.

3.3. Discussion

One of the appealing properties of GMMs is their ability to approximate arbitrarily-shaped densities [1, 7]. In [1] it is also argued that any set of full covariance matrices can be equally replaced by a larger set of diagonal covariance models . While this work confirmed that accuracy of diagonal models improves with increasing mixture components there remains a large performance gap of 11 % for FBANKs and 6 % for MFCC_Z between the best diagonal and full covariance model (Tab. 1). It is unlikely that further increasing the number of mixtures will close this difference. Moving from 16 to 32 mixtures resulted in a minor 1 % increase for FBANKs while the performance of the remaining feature sets started to degrade.

Several other observations can be made from the results of the recognition experiments. The cepstral baseline (MFCC_Z) with diagonal covariance models performed very poorly even for high SNRs. This is due to the applied CMN which led to a performance degradation in clean conditions but improved accuracy under noise. The results for MFCC without CMN were omitted here due to space constraints. However, the performance consistently improved when full covariance GMMs were employed (compare Fig. 2-5 (a) and (b)). This is somewhat unexpected as cepstral features are assumed to be approximately uncorrelated. On the other hand, the significant improvements for log-spectral features are not surprising given their high correlation among feature components. In particular the results for oracle masks illustrate this point clearly as diagonal covariance GMMs performed relatively poorly on these highly accurate masks.

Bounded marginalization led to an improved recognition accuracy for both covariance structures, especially in low SNRs. Utilizing the information in unreliable components made the performance of oracle masks highly resistent to noises with SNRs between 35 dB and 0 dB (see Fig. 4b, 5b). However, the poor performance in low SNRs of simple SS masks underlines the need for more accurate mask estimation techniques. This is the crucial point in MD recognition and is still an area of active research [8, 9, 10].

One point of concern when considering feature correlation is the increase in model complexity that comes as a price for the improved statistical modeling. Although this issue has not been addressed here several options exist to speed up computation time. Referring to Tab. 1 it is clear that for the full covariance case the number of mixtures could be reduced without sacrificing performance. Also, other schemes involving grand and global covariances [1, 6] or block-diagonal structures [11] can be employed to further reduce the number of parameters and speed up computation.

4. CONCLUSIONS

The following conclusions about the use of full covariance models for missing data speaker recognition can be drawn.

Firstly, diagonal covariances are inferior to full covariances in terms of speaker identification rates when considering log-spectral features in a MD framework. The overall best results for white and factory noise were obtained using bounded marginalization with full covariances. Comparisons for oracle masks suggest that full covariance models lead to very high noise robustness revealing the full potential of MD techniques. Secondly, further research is needed to overcome computational issues related to the high number of model parameters when modeling feature correlation. This remains a challenging task as traditional matrix orthogonalization techniques are not applicable within the MD framework.

5. REFERENCES

[1] D. Reynolds and R. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 1, 1995.

[2] A. Morris, M. Cooke, and P. Green, "Some solution to the missing feature problem in dataclassification, with application to noise robust asr," in *Proc. ICASSP*, Seattle, USA, 1998.

[3] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Communication*, vol. 34, 2001.

[4] A. Drygajlo and M. El-Maliki, "Speaker verification in noisy environments with combined spectral substraction and missing feature theory," in *Proc. ICASSP*, Seattle, USA, 1998.

[5] Y. Shao and D. Wang, "Robust speaker recognition using binary time-frequency masks," in *Proc. ICASSP*, Toulouse, France, 2006.

[6] M. Cooke, A. Morris, and P. Green, "Missing data techniques for robust speech recognition," in *Proc. ICASSP*, Munich, Germany, 1997.

[7] Y. Tong, *The multivariate normal distribution*, Springer Series in Statistics. Springer-Verlag, New York, 1990.

[8] M. Kühne, R. Togneri, and S. Nordholm, "Smooth soft mel-spectrographic masks based on blind sparse source separation," in *Interspeech*, Antwerp, Belgium, 2007.

[9] M. van Segbroeck and H. van Hamme, "Vector-quantization based mask estimation for missig data automatic speech recognition," in *Interspeech*, Antwerp, Belgium, 2007.

[10] D. Pullella, M. Kühne, and R. Togneri, "Robust speaker identification using combined feature selection and missing data recognition," in *Proc. ICASSP*, Las Vegas, USA, 2008, (accepted for publication).

[11] R. Wang et al., "Fast likelihood computation method using block-diagonal covariance matrices in hidden markov model," in *Proc. ISCSLP*, Taipei, Taiwan, 2002.