Hindawi Publishing Corporation
EURASIP Journal on Audio, Speech, and Music Processing
Volume 2010, Article ID 651420, 13 pages
doi:10.1155/2010/651420

Research Article

Independent Component Analysis and Time-Frequency Masking for Speech Recognition in Multitalker Conditions

Dorothea Kolossa, Ramon Fernandez Astudillo, Eugen Hoffmann, and Reinhold Orglmeister

Electronics and Medical Signal Processing Group, 10587 Berlin, Germany

Correspondence should be addressed to Dorothea Kolossa, d.kolossa@ee.tu-berlin.de

Received 25 September 2009; Revised 27 February 2010; Accepted 1 April 2010

Academic Editor: Li Deng

When a number of speakers are simultaneously active, for example in meetings or noisy public places, the sources of interest need to be separated from interfering speakers and from each other in order to be robustly recognized. Independent component analysis (ICA) has proven a valuable tool for this purpose. However, ICA outputs can still contain strong residual components of the interfering speakers whenever noise or reverberation is high. In such cases, nonlinear postprocessing can be applied to the ICA outputs, for the purpose of reducing remaining interferences. In order to improve robustness to the artefacts and loss of information caused by this process, recognition can be greatly enhanced by considering the processed speech feature vector as a random variable with time-varying uncertainty, rather than as deterministic. The aim of this paper is to show the potential to improve recognition of multiple overlapping speech signals through nonlinear postprocessing together with uncertainty-based decoding techniques.

1. Introduction

When speech recognition is to be used in arbitrary, noisy environments, interfering speech poses significant problems due to the ovelapping spectra and nonstationarity. If automatic speech recognition (ASR) is nonetheless required, for example for robust voice control in public spaces or for meeting transcription, the use of independent component analysis (ICA) can be important to segregate all involved speech sources for subsequent recognition. In order to attain the best results, it is often helpful to apply an additional nonlinear gain function to the ICA output to suppress residual speech and noise. After a short introduction to ICA in Section 2, this paper shows in Section 3 how such nonlinear gain functions can be attained based on three different principal approaches.

However, while source separation itself is greatly improved by nonlinear postprocessing, speech recognition results often suffer from artefacts and loss in information due to such masks. In order to compensate for these losses and to obtain results exceeding those of ICA alone, we suggest the use of uncertainty-of-observation techniques for the subsequent speech recognition. This allows for the utilization of a feature uncertainty estimate, which can be derived considering both artefacts and incorrectly suppressed components of target speech, and will be described in more detail in Section 4. From such an uncertain description of the speech signal in the spectrum domain, uncertainties need to be made available also in the feature domain, in order to be used for recognition. This can be achieved by the so-called "uncertainty propagation," which converts an uncertain description of speech from the spectrum domain, where ICA takes place, to the feature domain of speech recognition. After this uncertainty propagation, detailed in Section 5, recognition can take place under observation uncertainty, as shown in Section 6.

The entire process is vitally dependent on the appropriate estimation of uncertainties. Results given in Section 8 show that when the exact uncertainty in the spectrum domain is known, recognition results with the suggested approach are far in excess of those achievable by ICA alone. Also, a realistically computable uncertainty estimate is introduced, and experiments and results given in Sections 7 and 8 show that with this practical uncertainty measure, significant improvements of recognition performance can be attained for noisy, reverberant room recordings.

The presented method is closely related to other works that consider observation vectors as uncertain for decoding purposes, most often for noisy speech recognition [1–4], but in some cases also for speech recognition in multitalker conditions, as, for example, [5, 6], or [7] in conjunction with speech segregation via binary masking (see, e.g. [8, 9]).

The main novelty in comparison with the above techniques is the use of independent component analysis in conjunction with uncertainty estimation and with a piecewise approach of transforming uncertainties to the feature domain of interest. This allows for the suggested approach to utilize the combined strengths of independent component analysis and soft time-frequency masking, and to still be used with a wide range of feature parameterizations, often without the need for recomputing the uncertainty mapping function to the desired ASR-domain. Corresponding results are shown here for both MFCC and RASTA-PLP coefficients, but the discussed uncertainty transformation approach also generalizes well to the ETSI advanced front end, as shown in [10].

2. Independent Component Analysis for Reverberant Speech

Independent component analysis has been successfully employed for the separation of speech mixtures in both clean and noisy environments [11, 12]. Alternative methods include adaptive beamforming, which is closely related to independent component analysis when information-theoretic cost functions are applied [13], sparsity-based methods that utilize amplitude-delay histograms [6, 8, 14], or grouping cues typical of human stream segregation [15]. Here, independent component analysis has been chosen due to its inherent robustness to noise and its ability to handle strong reverberation by frequency-by-frequency optimization of the cost function.

In order to separate a number N of simultaneously active speech signals from M recordings, with $M \geq N$, the reverberant, noisy mixing process is modelled as

$$x_j(t) \approx \sum_{k=1}^{N} s_k(t) * h_{jk}(t) + d_j(t), \tag{1}$$

where the room impulse response $h_{jk}(t)$ from source k to sensor j is considered time-invariant.

Since convolutions are easily separable in the frequency domain, this expression is transformed by a short-time Fourier transform (STFT). Then, (1) becomes

$$\mathbf{X}(\Omega, \tau) \approx \mathbf{H}(\Omega)\mathbf{S}(\Omega, \tau) + \mathbf{D}(\Omega, \tau), \tag{2}$$

where $\mathbf{H}(\Omega)$ is composed of the room transfer functions $H_{jk}(\Omega)$ from all sources k to the sensors j, and $\mathbf{D}$ is the sensor noise. Here, Ω and τ denote the integer-valued frequency bin index and frame index, respectively.

In order to extract the original sources from the mixtures, ICA finds an unmixing matrix

$$\mathbf{W}(\Omega) \approx \mathbf{P}(\Omega)\Delta(\Omega)\mathbf{H}(\Omega)^{-1}, \tag{3}$$

for each frequency bin Ω, which by principle can only be known up to an arbitrary scaling and permutation described by the diagonal scaling matrix Δ and the permutation matrix $\mathbf{P}$. The unmixing matrix $\mathbf{W}$ is found by maximizing the statistical independence of the unmixed signals $\widehat{\mathbf{S}}$. Finally, unmixing is carried out separately in each frequency bin according to

$$\widehat{\mathbf{S}}(\Omega, \tau) = \mathbf{W}(\Omega) \cdot \mathbf{X}(\Omega, \tau). \tag{4}$$

To learn the matrix $\mathbf{W}$, the adaptive algorithm described in [16, Table 8.2, Equation 2] is used. In this algorithm, the demixing matrix $\mathbf{W}$ is calculated using a gradient descent. The update rule for the matrix in the ith iteration consists of two steps. At first, the current estimate of the source signals is computed by (4), using the result of the previous iteration $\mathbf{W}_{i-1}$ for unmixing. Then, the update of the unmixing matrix takes place according to

$$\mathbf{W}_i(\Omega) = \mathbf{W}_{i-1}(\Omega) + \eta\left(\Lambda - \left\langle \mathbf{f}\left(\widehat{\mathbf{S}}(\Omega, \tau)\right)\widehat{\mathbf{S}}(\Omega, \tau)^H \right\rangle\right)\mathbf{W}_{i-1}(\Omega), \tag{5}$$

where Λ is a diagonal matrix with

$$\lambda_{mm} = \left\langle \widehat{S}_m(\Omega, \tau)\widehat{S}_m^*(\Omega, \tau) \right\rangle. \tag{6}$$

Here, $\langle \cdot \rangle$ denotes the mean value and

$$f(x) = x \exp\left(-\frac{|x|^2}{2}\right). \tag{7}$$

Ideally, this optimization will result in independent output signals in each frequency bin. To obtain a complete spectrum of unmixed sources, it is additionally necessary to correctly sort the outputs, since their ordering after ICA is arbitrary and may vary from frequency bin to frequency bin. This so-called permutation problem can be solved in a number of ways; see, for example, [17, 18]. In all following work, permutations have been corrected by sorting outputs in accordance with the distance criterion

$$d_{m,n}(\Omega_r, \Omega_s) = \left(\sum_{\tau} |\nu_m(\Omega_r, \tau) - \nu_n(\Omega_s, \tau)|^p\right)^{1/p}, \tag{8}$$

described in [19]. Here, ν is defined by

$$\nu_m(\Omega, \tau) = \log\left|\widehat{S}_m(\Omega, \tau)\right|^2, \tag{9}$$

Ω_s is the frequency bin at which the permutation problem has to be solved and Ω_r denotes the frequency bin to be used as reference, and p is a constant. For this strategy, ordering permutations first at higher frequencies and proceeding downward has proven beneficial; therefore, the ordering at the maximum frequency bin was chosen as reference, and sorting according to (8) took place binwise in descending order.

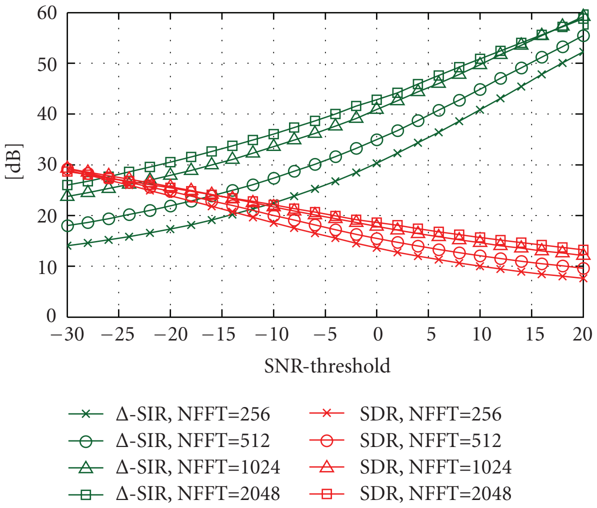

Figure 1: Performance of an ideal binary mask, tested on 12 pairs of same-and mixed-gender speakers. Performance is shown for frame lengths (NFFT) of 256, 512, 1024, and 2048 samples in terms of SDR and SIR-improvement. When the SNR-threshold is increased, the red SDR-curves are decreasing monotonically, while a more pronounced monotonic increase can be observed for the SIR-improvement, shown in green color.

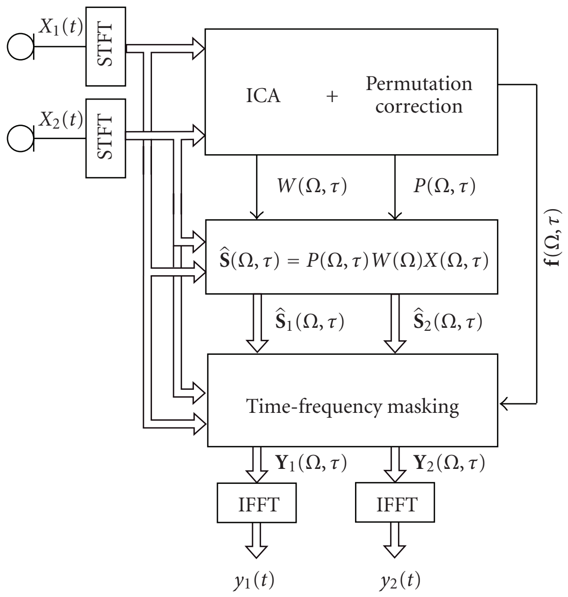
$X_1(t)$
STFT
$X_2(t)$
STFT
ICA
+
Permutation correction
$W(\Omega,\tau)$
$P(\Omega,\tau)$
$\hat{\mathbf{S}}(\Omega,\tau) = P(\Omega,\tau)W(\Omega)X(\Omega,\tau)$
$\hat{\mathbf{S}}_1(\Omega,\tau)$
$\hat{\mathbf{S}}_2(\Omega,\tau)$
$\mathbf{f}(\Omega,\tau)$
Time-frequency masking
$\mathbf{Y}_1(\Omega,\tau)$
$\mathbf{Y}_2(\Omega,\tau)$
IFFT
IFFT
$y_1(t)$
$y_2(t)$

Figure 2: Structure of ICA with Postmasking. Here, **f** stands for those ICA-based features that may be needed for mask estimation. All double arrows show the data flow of signal spectrograms, while single arrows indicate auxilliary information flow. The figure corresponds to the special case of two microphones and two estimated signals.

3. Time-Frequency Masking for ICA

However, some residual noise and interference are to be expected even after applying source separation, especially in reverberant environments. For removing these, postmasking of ICA-results has often been employed [14, 17, 20, 21] using

$$Y_k(\Omega,\tau) = M_k(\Omega,\tau) \cdot \hat{S}_k(\Omega,\tau). \tag{10}$$

This is motivated by the potential gains in Signal-to-Interference Ratio (SIR), which can already be attained by simple binary masking with an ideal mask. With such an, albeit practically elusive, mask, which is given by evaluating the true knowledge about the signal spectra via

$$M_k(\Omega,\tau) = \begin{cases} 1, & \text{for } |S_k(\Omega,\tau)| \geq \left|S_j(\Omega,\tau)\right| \quad \forall j, \\ 0, & \text{otherwise,} \end{cases} \tag{11}$$

it is possible to obtain more than 40 dB SIR improvement on two-speaker mixtures even without ICA, while remaining above 20 dB of Signal-to-Distortion Ratio (SDR) [22]. The results of one such exemplary experiment are shown in Figure 1. For this figure, an additional masking threshold T was introduced, and the mask in (11) was only set to 1, if the source of interest was greater than all other sources by at least T dB, that is, if

$$20\log_{10}|S_k(\Omega,\tau)| \geq 20\log_{10}\left|S_j(\Omega,\tau)\right| - T \quad \forall j. \tag{12}$$

However, an ideal mask is impossible to obtain realistically; thus, approximations to it are required. For obtaining such an approximation, mask estimation based on ICA results has been proposed and shown to be successful, both for binary and soft masks, see, for example, [17, 18, 20]. The motivation for this procedure lies both in the noise-robustness of ICA, which can therefore unmix signals even when large interferences make the estimation of a time-frequency mask extremely difficult, and also in the fact that ICA will unmix signals even in those time-frequency regions, where two or more of them are simultanously active to a significant extent.

The architecture of such systems is shown in Figure 2 for the exemplary case of two sources and microphones.

In the following, four types of masks are considered:

(i) amplitude-based masks,

(ii) phase-based masks,

(iii) two types of interference-based masks,

which will be described in the subsequent sections.

3.1. Amplitude-Based Masks. One of the simplest post-masks suitable for postprocessing of ICA results is based on comparing the magnitude of all ICA outputs [20]. Due to the sparsity of sources in an appropriate spectral representation [8], only one should be dominant; therefore, all others are discarded.

In order for the strategy to be independent of the source signal energies, all ICA output signals need to be normalized to equal variance via

$$\tilde{S}_k(\Omega,\tau) = \frac{\hat{S}_k(\Omega,\tau)}{\sqrt{\operatorname{var}\left(\hat{S}_k\right)}}, \qquad (13)$$

before the mask is computed.

Then, a hard amplitude mask can be obtained by comparing a local dominance ratio to an acceptance threshold T via

$$\begin{aligned} M_k(\Omega,\tau) &\\ = \Psi\left(\log\left(\left|\tilde{S}_k(\Omega,\tau)\right|^2\right) - \max_{\forall j\neq k}\log\left(\left|\tilde{S}_j(\Omega,\tau)\right|^2\right) - \frac{T}{10}\right), & \end{aligned} \qquad (14)$$

with Ψ defined by

$$\Psi(x) = \begin{cases} 0, & \text{for } -\infty \le x \le 0, \\ 1, & \text{for } 0 < x < \infty. \end{cases} \qquad (15)$$

This is a rather simple approach, which has been enhanced in the following by applying a sigmoid nonlinearity to reduce artefacts. This can be easily achieved by redefining Ψ to

$$\Psi(x) = \frac{1}{1+\exp(-gx)}, \qquad (16)$$

where g is the mask gain controlling its steepness.

3.2. Phase-Based Masks. The source separation performance of ICA can also be seen from a beamforming perspective. When the unmixing filters learned by ICA are viewed as frequency-variant beamformers, it can be shown that successful ICA effectively places zeros in the directions of all interfering sources [23]. Therefore, the zero directions of the unmixing filters should be indicative of all source directions. Thus, when the local direction of arrival (DOA) is estimated from the phase of any one given time-frequency bin, this should give an indication of the dominant source in this bin. This is the principle underlying phase-based time-frequency masking strategies.

Phase-based postmasking of ICA outputs was introduced in [17]. In this method, the angle $\theta_k(\Omega,\tau)$ between the k'th target basis vector of the unmixing matrix and the microphone signal vector is used in order to determine whether and to what degree a given channel should be masked.

According to (2), when noise is not considered, the mixing system can be modeled by

$$\mathbf{X}(\Omega,\tau) \approx \sum_{k=1}^{N} \mathbf{h}_k(\Omega) S_k(\Omega,\tau). \qquad (17)$$

Here, $\mathbf{h}_k(\Omega)$ denotes the k'th column of the mixing matrix, and $S_k(\Omega,\tau)$ is the value of source k in frequency Ω at frame τ.

ICA results in an unmixing matrix $\mathbf{W}$, which is used to obtain M estimated source signals according to (4). This corresponds to

$$\begin{aligned} \mathbf{X}(\Omega,\tau) &= \mathbf{W}(\Omega)^{-1}\hat{\mathbf{S}}(\Omega,\tau) \\ &= [\mathbf{a}_1(\Omega), \mathbf{a}_2(\Omega), \ldots, \mathbf{a}_M(\Omega)]\hat{\mathbf{S}}(\Omega,\tau), \end{aligned} \qquad (18)$$

where the estimated mixing matrix $\mathbf{W}^{-1}$ is given in terms of its constituent column vectors, $[\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_M]$. When comparing (18) and (2), and considering (3), it can be seen that the columns of $\mathbf{W}^{-1}$ correspond to the columns of $\mathbf{H}(\Omega)$, the matrix containing the values of the room transfer function for each frequency, up to an arbitrary scaling of column vectors and a reordering of sources, which is constant over frequencies after the permutation correction. Thus, in those time-frequency bins, where source k is dominant, the associated basis vector $\mathbf{a}_i(\Omega)$ should correspond to the column of the mixing matrix $\mathbf{H}(\Omega)$ associated with source k. In general, the index i may be different from the index k, due to possible permutations. However, as this change of indices will be consistent over frequency, it is disregarded in the following.

Thus, after appropriate normalization, in frames with dominant source k, the associated basis vector $\mathbf{a}$ would also be equal to $\mathbf{X}(\Omega,\tau)$ of the current frame. If an anechoic model is appropriate for the mixing process at hand, the basis vectors should form clusters, one for each of the sources. For this purpose, the basis vectors need to be normalized regarding both their phases and amplitudes as detailed in [17]. For phasenormalization, they are first normalized with respect to a reference sensor J and secondly frequency-normalized, which gives

$$\begin{aligned} &\bar{a}_{jk}(\Omega) \\ &\quad = \left|a_{jk}(\Omega)\right| \exp\left(i\frac{\arg\left(a_{jk}(\Omega)/a_{Jk}(\Omega)\right)}{f(\Omega)4c^{-1}d_{\max}}\right); \quad j,k=1\ldots M, \end{aligned} \qquad (19)$$

as a normalized vector. Here, $f(\Omega)$ stands for the center frequency in Hz of frequency bin Ω; c is the velocity of sound and $d_{\max}$ stands for the distance between the reference sensor J and the farthest of all other microphones $j = 1\ldots M$. For this vector, the phase varies only between

$$-\frac{\pi}{2} \le \arg\left(\bar{a}_{jk}(\Omega)\right) \le \frac{\pi}{2}, \qquad (20)$$

which is important for computing a distance measure between vectors. Finally, amplitude-normalization is carried out by

$$\tilde{\mathbf{a}}_k(\Omega) = \frac{[\bar{a}_{1k}(\Omega), \bar{a}_{2k}(\Omega), \ldots, \bar{a}_{Mk}(\Omega)]^T}{\|\bar{\mathbf{a}}_k(\Omega)\|}. \qquad (21)$$

After the normalized basis vectors $\tilde{\mathbf{a}}_k(\Omega)$ are thus available, masking is carried out based on the angle $\theta_k(t,\tau)$ between the observed vector $\mathbf{X}(\Omega,\tau)$ and the basis vector $\tilde{\mathbf{a}}_k(\Omega)$. This angle is computed in a whitened space, where $\mathbf{X}(\Omega,\tau)$ and

$\tilde{\mathbf{a}}(\Omega)$ are premultiplied by the whitening matrix $\mathbf{V}$, which is the inverse square root of the sensor autocorrelation matrix, $\mathbf{V}(\Omega) = \mathbf{R}_{xx}^{-1/2}$.

The mask is a soft mask, which is determined from $\theta_k(\Omega, \tau)$ by the logistic function

$$\mathcal{M}_k(\Omega, \tau) = \frac{1}{1 + e^{g(\theta_k(\Omega,\tau) - \theta_T)}}. \tag{22}$$

The parameter g describes the steepness of the mask and θ_T is the transition point, where the mask takes on the value 1/2. More details on the mask computation can be found in [17].

3.3. Interference-Based Masks.

As an alternative criterion for masking, residual interference in the signal may be estimated and the mask may be computed as an MMSE estimator of the clean signal. This can be achieved with a number of approaches, two of which will be presented here in more detail.

3.3.1. Ephraim-Malah Filter-Based Post-Filtering.

The remaining noise components in the separated signals can be minimized based on the Ephraim-Malah filter technique. For this purpose, the following signal model is assumed

$$\hat{\mathbf{S}}(\Omega, \tau) = \mathbf{S}(\Omega, \tau) + \mathbf{D}(\Omega, \tau), \tag{23}$$

where the clean signal $\mathbf{S}(\Omega, \tau)$ is corrupted by a noise component $\mathbf{D}(\Omega, \tau)$, the remaining sum of the interfering signals and the background noise. The estimated clean signals are obtained by

$$\mathbf{Y}(\Omega, \tau) = \mathcal{M}_{SE}(\Omega, \tau)\hat{\mathbf{S}}(\Omega, \tau), \tag{24}$$

where $\mathcal{M}_{SE}(\Omega, \tau)$ is the amplitude estimator gain. For the calculation of the gain $\mathcal{M}_{SE}(\Omega, \tau)$, different speech enhancement algorithms can be used. In the following, we are using the log spectral amplitude estimator (LSA) as proposed by Ephraim and Malah [24].

For the algorithm, the a posteriori $\gamma_k(\Omega, \tau)$ and a priori SNR $\xi_k(\Omega, \tau)$ are defined by

$$\begin{aligned}\gamma_k(\Omega, \tau) &= \frac{\left|\hat{S}_k(\Omega, \tau)\right|^2}{\lambda_D(\Omega, \tau)},\\ \xi_k(\Omega, \tau) &= \max(\alpha(\gamma_k(\Omega, \tau - 1) - 1)\\ &\qquad +(1-\alpha)(\gamma_k(\Omega, \tau) - 1), 0).\end{aligned} \tag{25}$$

Here, α is a smoothing parameter, $\hat{S}_k(\Omega, \tau)$ is the kth ICA-output, and $\lambda_D(\Omega, \tau)$ is the noise power

$$\lambda_D(\Omega, \tau) = \alpha_D\lambda_D(\Omega, \tau - 1) + (1 - \alpha_D)|D_k(\Omega, \tau)|^2, \tag{26}$$

with the noise estimate $|D_k(\Omega, \tau)|$ given by

$$|D_k(\Omega, \tau)| = \max\left(\left|X_k(\Omega, \tau) - \hat{S}_k(\Omega, \tau)\right|, 0\right). \tag{27}$$

With these parameters, the log spectral amplitude estimator is given by

$$\mathcal{M}_{SE}(\Omega, \tau) = \frac{\xi(\Omega, \tau)}{1 + \xi(\Omega, \tau)} \exp\left(\int_{t=\nu(\Omega,\tau)}^{\infty} \frac{e^{-t}}{t} dt\right), \tag{28}$$

with $\xi(\Omega, \tau)$ denoting the local a priori SNR and

$$\nu(\Omega, \tau) = \left(\frac{\xi(\Omega, \tau)}{1 + \xi(\Omega, \tau)}\right)\gamma(\Omega, \tau). \tag{29}$$

3.3.2. Inclusion of Speech Presence Probabilities.

According to [25], the previous approach can be expanded using additional information for calculation of speech presence probabilities. The gain function of the Ephraim-Malah filter becomes

$$\mathcal{M}(\Omega, \tau) = \mathcal{M}_{SE}(\Omega, \tau)^{p(\Omega,\tau)}\mathbf{G}_{\min}^{(1-p(\Omega,\tau))}, \tag{30}$$

where $\mathbf{G}_{\min}$ is a spectral attenuation floor, $\mathcal{M}_{SE}$ the gain of the speech enhancement method, and $p(\Omega, \tau)$ the speech presence probability [26, 27]. The infomation needed for speech presence probability calculation is gained from a bin-wise noise dominance estimate, which can be computed in the spectrum domain by [18]

$$f_{N,k}(\Omega_0, \tau_0) = \frac{\left\|\Phi(\Omega, \tau)\left(\sum_{m \neq k} \hat{S}_m(\Omega, \tau) - \hat{S}_k(\Omega, \tau)\right)\right\|}{\left\|\Phi(\Omega, \tau)\hat{S}_k(\Omega, \tau)\right\|}. \tag{31}$$

A similar measure of speech dominance $f_{S,k}$ is needed in addition

$$f_{S,k}(\Omega_0, \tau_0) = \frac{\left\|\Phi(\Omega, \tau)\left(\hat{S}_k(\Omega, \tau) - \sum_{m \neq k} \hat{S}_m(\Omega, \tau)\right)\right\|}{\left\|\Phi(\Omega, \tau) \sum_{m \neq k} \hat{S}_m(\Omega, \tau)\right\|}. \tag{32}$$

Both measures utilize the difference between the estimated target spectrogram $\hat{S}_k(\Omega, \tau)$ and the sum of estimated nontarget signals $\sum_{m \neq k} \hat{S}_m(\Omega, \tau)$. The Euclidean norm operator $\|\cdot\|$ is applied to two-dimensional windowed spectrograms here by taking the sum over their squared entries, and

$$\Phi(\Omega, \tau) = \begin{cases} W(\Omega - \Omega_0, \tau - \tau_0), & |\Omega - \Omega_0| \leq R_\Omega/2, \\ & |\tau - \tau_0| \leq R_\tau/2, \\ 0, & \text{otherwise} \end{cases} \tag{33}$$

uses a two-dimensional window function W of size $R_\Omega \times R_\tau$, usually a two-dimensional Hanning window. The speech presence probability is then approximated by a soft mask via

$$\begin{aligned}\hat{p}_i(\Omega, \tau) &= \frac{1}{1 + \exp(g(f_{S,k}(\Omega_0, \tau_0) - \lambda_s))}\\ &\quad \cdot \left(1 - \frac{1}{1 + \exp(g(f_{N,k}(\Omega_0, \tau_0) - \lambda_n))}\right).\end{aligned} \tag{34}$$

Here, λ_s, λ_n and g are parameters specifying the two threshold points and the mask gain, respectively.

4. Estimation of Uncertainties

Due to the use of time-frequency masking, part of the information of the original signal might be eliminated along with the interfering sources. To compensate for this lack of information, each masked estimated source is considered as uncertain and described in the form of a posterior distribution of each Fourier coefficient of the clean signal $S_k(\Omega, \tau)$ given the available information.

Estimating the uncertainty in the spectrum domain has clear advantages, when contrasted with uncertainty estimation in the domain of speech recognition, since much intermediate information about the signal and noise process as well as the mask is known in this phase of signal processing, but is generally not available in the further steps of feature extraction. This has motivated a number of studies on spectrum domain uncertainty estimation, most recently for example [7, 10]. In contrast to other methods, the suggested strategy possesses two advantages: it does not need a detailed spectrum domain speech prior, which may require a large number of components or may incur the need for adaptation to the speaker and environment; and it gives a computationally very inexpensive approximation that is applicable for both binary and soft masks.

The model used here for this purpose is the complex Gaussian uncertainty model [28]

$$p(S_k(\Omega,\tau) \mid Y_k(\Omega,\tau)) = \frac{1}{\pi\sigma^2}\exp\left(-\frac{|S_k(\Omega,\tau)-Y_k(\Omega,\tau)|^2}{\sigma^2}\right), \tag{35}$$

where the mean is set equal to the Fourier coefficient obtained from post-masking $Y_k(\Omega, \tau)$ and the variance σ^2 represents the lack of information, or uncertainty. In order to determine σ^2, two alternative procedures were used.

4.1. Ideal Uncertainties. Ideal Uncertainties describe the squared difference between the true and the estimated signal magnitude. They are computed by

$$\sigma_T^2 = \left| |S_k(\Omega,\tau)| - |Y_k(\Omega,\tau)| \right|^2, \tag{36}$$

where S_k is the reference signal. However, these ideal uncertainties are available only in experiments where a reference signal has been recorded. Thus, the ideal results may only serve as a perspective of what the suggested method would be capable of if a very high quality error estimate were already available.

4.2. Masking Error Estimate. In practice, it is necessary to approximate the ideal uncertainty estimate using values that are actually available. Since much of the estimation error is due to the time-frequency mask, in further experiments such a masking error was used as the single basis of the uncertainty measure.

This uncertainty due to masking can be computed by

$$\sigma_E^2 = \alpha \left| |\hat{S}_k(\Omega,\tau)| - |Y_k(\Omega,\tau)| \right|^2. \tag{37}$$

If $\alpha = 1$, this error estimate would assume that the time-frequency mask leads to missing signal information with 100 certainty. The value should be lower to reflect the fact that some of the masked time-frequency bins contain no target speech information at all. To obtain the most suitable value for α, the following expression was minimized

$$\alpha = \arg\min_{\tilde{\alpha}} (\sigma_E(\tilde{\alpha}) - \sigma_T)^2. \tag{38}$$

In order to avoid adapting parameters to each of the test signals and masks, this minimization was carried out only once and only for a mixture not used in testing. After averaging over all mask types, the same value of α was used in all experiments and for all datasets. This optimal value was $\alpha = 0.71$.

5. Propagation of Uncertainties

When uncertain features are available in the STFT domain, they could in principle be used for spectrum domain speech recognition. However, as shown in [29], due to the less robust spectrum domain models, this does not provide for optimum results. Instead, a more successful approach is to transform the uncertain description of speech from the spectrum domain to the domain of speech recognition. This can in principle be achieved by two approaches, data-driven as in [7] or model-driven as in [5]. In the following, we only consider the model-driven approach, which can achieve very low propagation errors with small memory requirements and without the need for a training phase [10]. However, a detailed comparison of both principal methods remains an interesting target for future work.

In order to carry out the propagation through the feature extraction process, the uncertain spectrum domain description is considered as specifying speech as a random variable according to (35). If such an uncertain description of the STFT is used, the corresponding posterior distribution $p(\mathbf{S}_k \mid \mathbf{Y}_k)$ has to be propagated into the feature domain. For this purpose, the effect of all transformations in the feature extraction process on this probability distribution needs to be considered, which will result in an estimated feature domain random variable, describing both the mean of the speech features as well as the associated degree of uncertainty. Since this computation takes place for each feature and in each bin, subsequent recognition will have a maximally precise description of all uncertainties, allowing the algorithm to focus most on those features that are most reliable, and, if desired, to replace the uncertain ones by better estimates under simultaneous consideration of the recognizer speech model.

In conventional automatic speech recognition, only the STFT of each estimated source $\mathbf{Y}_k$ must be transformed into the feature domain of automatic speech recognition. Feature extractions involve multiple transformations, some of them nonlinear, which are performed jointly on multiple features of the same frame or by combining features from different time frames. Propagating an uncertain description of the STFT of each estimated source is therefore a complicated task that can be simplified by propagating only first- and second-order information. This section shows how this propagation can be attained by a piecewise approach in which the feature extraction is divided into different steps and the optimal method is chosen to perform uncertainty propagation in each step. Uncertainty propagation is used with two of the more robust speech recognition features, namely the Mel-cepstrum coefficients (MFCCs) [30] and the cepstral coefficients obtained from the RelAtive SpecTrAl Perceptual Linear Prediction (RASTA-PLP) feature extraction [31], here denoted as RASTA-LPCCs.

5.1. Mel-Cepstral Feature Extraction. The conventional Mel-cepstral feature extraction consists of the following steps.

(1) Extract the short-time spectral amplitude (STSA) from the STFT.

(2) Compute each filter output of a Mel-filterbank as a weighted sum of the STSA features of each frame.

(3) Apply the logarithm to each filter output.

(4) Compute the discrete cosine transform (DCT) from each frame of log-filterbank features.

In order to propagate random variables rather than deterministic signals, these steps were modified as follows.

Step (1) can be solved if we take into account that if a Fourier coefficient $S_k(\Omega, \tau)$ is complex Gaussian distributed as given by (35), its amplitude $|S_k(\Omega, \tau)|$ is Rice distributed. From the first raw moment of the Rice distribution, it is possible to compute the mean of the uncertain STSA features as [28]

$$\begin{aligned}\mu_k(\Omega, \tau)^{\mathrm{STSA}} &= \mathrm{E}\{|S_k(\Omega, \tau)|\} \\ &= \frac{\sqrt{\pi \lambda_k(\Omega, \tau)}}{2} \cdot \exp\left(-\frac{|Y_k(\Omega, \tau)|^2}{2\lambda_k(\Omega, \tau)}\right) \\ &\quad \times \left[\left(1 + \frac{|Y_k(\Omega, \tau)|^2}{\lambda_k(\Omega, \tau)}\right) I_0\left(\frac{|Y_k(\Omega, \tau)|^2}{2\lambda_k(\Omega, \tau)}\right)\right. \\ &\quad \left. + \frac{|Y_k(\Omega, \tau)|^2}{\lambda_k(\Omega, \tau)} I_1\left(\frac{|Y_k(\Omega, \tau)|^2}{2\lambda_k(\Omega, \tau)}\right)\right],\end{aligned} \tag{39}$$

where I_0 and I_1 correspond to the modified Bessel functions of order zero and one, respectively. The variance of the uncertain STSA features can be computed from the first and second raw moments as

$$\begin{aligned}\Sigma_k(\Omega, \tau)^{\mathrm{STSA}} &= \mathrm{E}\left\{|S_k(\Omega, \tau)|^2\right\} - \left(\mu_k(\Omega, \tau)^{\mathrm{STSA}}\right)^2 \\ &= \lambda_k(\Omega, \tau) + |Y_k(\Omega, \tau)|^2 - \left(\mu_k(\Omega, \tau)^{\mathrm{STSA}}\right)^2.\end{aligned} \tag{40}$$

Step (2) in the Mel-cepstral feature extraction corresponds to the Mel-filterbank, which is a linear transformation and bears no additional difficulty for the propagation of mean and covariance. In general, given a random vector variable $\mathbf{x}$ and a linear transformation defined by the matrix $\mathbf{T}$, the transformed mean and covariance correspond to

$$\begin{aligned}\mathrm{E}\left\{\mathbf{T}\mathbf{x}^T\right\} &= \mathbf{T}\mathrm{E}\{\mathbf{x}\}^T, \\ \mathrm{Cov}\{\mathbf{T}\mathbf{x}^T\} &= \mathbf{T}\mathrm{Cov}\{\mathbf{x}\}\mathbf{T}^T.\end{aligned} \tag{41}$$

Step (3) corresponds to the computation of the logarithm. Since the distribution of the Mel-STSA uncertain features has a relatively low skewness and the dimensionality of the features has been reduced by approximately one order of magnitude through the application of the Mel-filterbank, the use of the pseudo-Montecarlo method termed unscented transform [32] provides an acceptable trade-off between accuracy and computational cost. Details regarding the use of the unscented transform for uncertainty propagation can be found in [28].

Step (4), the DCT transform, completes the computation of the MFCC coefficients. Since this is a linear transformation like the Mel-filterbank, it can be computed according to (41).

5.2. Relative Spectral Perceptual Linear Prediction Feature Extraction. The obtention of the RASTA-Linear Prediction Cepstral Coefficients (RASTA-LPCCs) corresponds to the following steps.

(1) Extract the power spectral density (PSD) from the STFT.

(2) Compute each filter output of a Bark-filterbank as a weighted sum of the PSD features of each frame.

(3) Apply the logarithm to each filter output.

(4) Filter the resulting frames with the RASTA IIR filter.

(5) Add the equal loudness curve and multiply by 0.33 to simulate the power law of hearing.

(6) Apply the exponential to invert the effect of the logarithm.

(7) Compute an all-pole model of each frame to obtain the linear prediction coefficients (LPCs).

(8) Compute cepstral coefficients from each LPC frame.

This feature extraction also requires a set of modifications and approximations in order to be applicable for uncertain features. An overview of these is shown in Figure 3 and the necessary computational steps are given in detail below.

Step (1) can be solved similarly to the case of the STSA. The propagated mean and covariance can be computed from the second and fourth raw moments of the Rice distribution as [33]

$$\begin{aligned}\mu_k(\Omega, \tau)^{\mathrm{PSD}} &= \mathrm{E}\left\{|S_k(\Omega, \tau)|^2\right\} = \lambda_k(\Omega, \tau) + |Y_k(\Omega, \tau)|^2, \\ \Sigma_k(\Omega, \tau)^{\mathrm{PSD}} &= \mathrm{E}\left\{|S_k(\Omega, \tau)|^4\right\} - \left(\mu_k(\Omega, \tau)^{\mathrm{PSD}}\right)^2 \\ &= 2\lambda_k(\Omega, \tau)|Y_k(\Omega, \tau)|^2 + \lambda_k(\Omega, \tau)^2.\end{aligned} \tag{42}$$

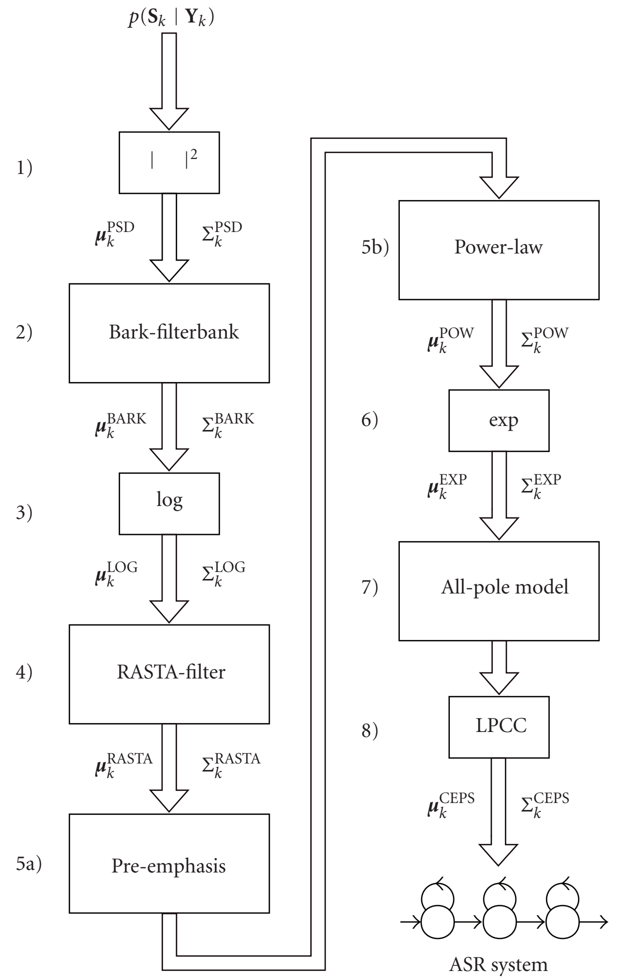
$p(\mathbf{S}_k \mid \mathbf{Y}_k)$
1)
$|\ \ |^2$
$\boldsymbol{\mu}_k^{\mathrm{PSD}}$
Σ_k^{PSD}
2)
Bark-filterbank
$\boldsymbol{\mu}_k^{\mathrm{BARK}}$
Σ_k^{BARK}
3)
log
$\boldsymbol{\mu}_k^{\mathrm{LOG}}$
Σ_k^{LOG}
4)
RASTA-filter
$\boldsymbol{\mu}_k^{\mathrm{RASTA}}$
$\Sigma_k^{\mathrm{RASTA}}$
5a)
Pre-emphasis
5b)
Power-law
$\boldsymbol{\mu}_k^{\mathrm{POW}}$
Σ_k^{POW}
6)
exp
$\boldsymbol{\mu}_k^{\mathrm{EXP}}$
Σ_k^{EXP}
7)
All-pole model
8)
LPCC
$\boldsymbol{\mu}_k^{\mathrm{CEPS}}$
Σ_k^{CEPS}
ASR system

Figure 3: Block diagram of the RASTA-LPCC feature extraction extended to encompass uncertainty propagation. Arrows indicate the propagation of mean and covariance at each step. The figure shows the propagation of the posterior corresponding to the kth estimated source.

Step (2), which corresponds to the Bark-filterbank, can be resolved identically to the case of the Mel-filterbank of the MFCCs by using (41).

Step (3) of the RASTA-PLP transformation consists of the computation of the logarithm as in the case of the Mel-cepstral feature extraction. However, the distribution of the Bark-PSD uncertain features presents a much higher skewness compared to the case of the Mel-STSA features. Consequently, the propagation through this step is more accurately computed using the assumption of log-normality of the Bark-PSD features, also used in other propagation approaches like [3, 5, 34]. The covariance under this assumption can be approximated by [34, equation 5.47], yielding

$$\Sigma_k(i,j,\tau)^{\mathrm{LOG}} \approx \log\left(\frac{\Sigma_k(i,j,\tau)^{\mathrm{BARK}}}{\mu_k(i,\tau)^{\mathrm{BARK}}\mu_k(j,\tau)^{\mathrm{BARK}}}+1\right), \tag{43}$$

where i, j are the filterbank indices and $\mu_k(i,\tau)^{\mathrm{BARK}}$ and $\Sigma_k(i,j,\tau)^{\mathrm{BARK}}$ correspond to the mean and covariance after the Bark-filterbank transformation. The mean can be approximated by [34, equation 5.46]

$$\mu_k(j,\tau)^{\mathrm{LOG}} \approx \log\left(\mu_k(j,\tau)^{\mathrm{BARK}}\right) - \frac{1}{2}\Sigma_k(j,j,\tau)^{\mathrm{LOG}}. \tag{44}$$

Step (4) corresponds to the RASTA filter. The RASTA filter is an IIR filter that imitates the preference of humans for sounds with a certain rate of change. It realizes the transfer function

$$H(z) = 0.1\frac{2+z^{-1}-z^{-3}-2z^{-4}}{1-0.94z^{-1}}. \tag{45}$$

This can also be expressed by the following difference equation

$$\mathbf{y}(\tau) = \sum_{d=0}^{4} b_d\,\mathbf{x}(\tau-d) - a_1\mathbf{y}(\tau-1), \tag{46}$$

where $\mathbf{y}(\tau)$ is a column vector containing the τth frame of RASTA-filtered features, and $\mathbf{x}(\tau)\cdots\mathbf{x}(\tau-4)$ and $\mathbf{y}(\tau-1)$ correspond to previous logarithm domain input and RASTA domain output frames, respectively. The scalars $b_0\cdots b_4$ and a_1 are the normalized feedforward and feedback coefficients. Computing the propagation of the mean $\boldsymbol{\mu}_k(\tau)^{\mathrm{RASTA}}$ through this transformation is identical to the case of the Mel or Bark filterbanks. The computation of the covariance is, however, more complex due to the created time correlation between inputs and outputs. The correlation matrix for the τth filter output $\mathbf{y}(\tau)$ can be computed from (46) as

$$\begin{aligned}\mathrm{E}\left\{\mathbf{y}(\tau)\mathbf{y}(\tau)^T\right\} &= \sum_{d=0}^{4} b_d^2\,\mathrm{E}\left\{\mathbf{x}(\tau-d)\mathbf{x}(\tau-d)^T\right\}\\ &\quad + a_1^2\mathrm{E}\left\{\mathbf{y}(\tau-1)\mathbf{y}(\tau-1)^T\right\}\\ &\quad + 2\sum_{d=0}^{4}\sum_{q=1}^{d}(-1)^q b_d a_1^q b_{d-q}\\ &\quad \cdot\mathrm{E}\left\{\mathbf{x}(\tau-d)\mathbf{x}(\tau-d)^T\right\},\end{aligned} \tag{47}$$

where the last summand accounts for the input output correlation. The corresponding covariance of the RASTA features can be obtained as

$$\boldsymbol{\Sigma}_k(\tau)^{\mathrm{RASTA}} = \mathrm{E}\left\{\mathbf{y}(\tau)\,\mathbf{y}(\tau)^T\right\} - \boldsymbol{\mu}_k(\tau)^{\mathrm{RASTA}}\left(\boldsymbol{\mu}_k(\tau)^{\mathrm{RASTA}}\right)^T. \tag{48}$$

Steps (5a) and (5b) correspond to conventional linear transformations in the logarithm domain, and therefore the propagation through them can be solved by applying (41) to obtain the means $\boldsymbol{\mu}_k^{\mathrm{POW}}$ and covariances $\boldsymbol{\Sigma}_k^{\mathrm{POW}}$. Furthermore, since the assumption of log-normality in the Bark-PSD domain implies that the log-domain features are normally distributed, RASTA, preemphasis, and power-law transformations do not alter this condition.

Step (6) corresponds to the transformation through the exponential. Since this transformation is the inverse of the logarithm, the corresponding features are log-normally distributed with mean and covariance computable from [34, equations 5.44, 5.45]

$$\begin{aligned}\mu_k(j,\tau)^{\text{EXP}} &\approx \exp\left(\mu_k(j,\tau)^{\text{POW}} + \frac{\Sigma_k(j,j,\tau)^{\text{POW}}}{2}\right),\\ \Sigma_k(i,j,\tau)^{\text{EXP}} &\approx \mu_k(i,\tau)^{\text{POW}}\mu_k(j,\tau)^{\text{POW}}\\ &\quad\cdot\left(\exp\left(\Sigma_k(i,j,\tau)^{\text{POW}}\right)-1\right).\end{aligned}\tag{49}$$

The final steps of the RASTA-LPCC feature extraction, Steps (7) and (8), correspond to the computation of the all-pole model to obtain the LPC coefficients, described in the conventional PLP technique [35], and the computation of the cepstral coefficients from the LPCs using [30, equation 3] Due to the complex nature of these transformations and the low skewness of the uncertain features after the exponential transformation, the propagation is computed using the unscented transform, similarly to the case of the logarithm transformation for the Mel-cepstral features.

6. Recognition of Uncertain Features

When features for speech recognition are given not as point estimates, but rather in the form of a posterior distribution $\mathrm{p}(\mathbf{o}_k|\mathbf{Y}_k)$ with estimated mean $\boldsymbol{\mu}_k^{\text{CEPS}}$ and covariance $\boldsymbol{\Sigma}_k^{\text{CEPS}}$, the speech decoder must be modified in order to take this additional information into account. A number of approaches exist, both for binary and for continuous-valued uncertainties, for example, [2, 36, 37].

Here, two missing feature approaches were applied, which are capable of considering real-valued uncertainties. These methods, modified imputation [5] and HMM variance compensation [2], have been implemented for the Hidden Markov Model Toolkit (HTK) [38] and were used in the tests.

Both methods are appropriate for HMM-based systems, where recognition takes place by finding the optimum HMM state sequence $[q_1,\ldots,q_E]$, which gives the best match to the feature vector sequence $[\mathbf{o}(1),\ldots,\mathbf{o}(E)]$ when each HMM state has an associated output probability distribution $\mathrm{p}(\mathbf{o}\mid q)$.

6.1. HMM Variance Compensation. In HMM variance compensation, the computation of state output probabilities is modified to incorporate frame-by-frame and feature-by-feature uncertainties [2]. This is formulated as an averaging of the output probability distribution $\mathrm{p}(\mathbf{o}_k(\tau)\mid q)$ over all possible unseen cepstra defined by the posterior $\mathrm{p}(\mathbf{o}_k(\tau)\mid\mathbf{Y}_k(\tau))$

$$\hat{\mathrm{p}}\left(\boldsymbol{\mu}_k^{\text{CEPS}}(\tau)\mid q\right)=\int_{-\infty}^{\infty}\mathrm{p}(\mathbf{o}_k(\tau)\mid\mathbf{Y}_k(\tau))\mathrm{p}(\mathbf{o}_k(\tau)\mid q)d\mathbf{o}_k(\tau)\tag{50}$$

which leads to

$$\hat{\mathrm{p}}\left(\boldsymbol{\mu}_k^{\text{CEPS}}(\tau)\mid q\right)=N\left(\boldsymbol{\mu}_k^{\text{CEPS}}(\tau);\boldsymbol{\mu}_q,\boldsymbol{\Sigma}_q+\boldsymbol{\Sigma}_k^{\text{CEPS}}(\tau)\right).\tag{51}$$

Here, q denotes the HMM state, with mean $\boldsymbol{\mu}_q$ and covariance $\boldsymbol{\Sigma}_q$. For Gaussian mixture models, the same procedure can be applied to each mixture component. This yields

$$\hat{\mathrm{p}}\left(\boldsymbol{\mu}_k^{\text{CEPS}}(\tau)\mid q\right)=\sum_{m=1}^{M}w_mN\left(\boldsymbol{\mu}_k^{\text{CEPS}}(\tau);\boldsymbol{\mu}_{q,m},\boldsymbol{\Sigma}_{q,m}+\boldsymbol{\Sigma}_k^{\text{CEPS}}(\tau)\right),\tag{52}$$

for an M-component mixture model with weights w_m.

6.2. Modified Imputation. In modified imputation, the idea is to replace the imputation equation, originally proposed for completely missing features in [36], with an alternative formulation, which also allows for real-valued degrees of uncertainty. Thus, whereas missing parts of feature vectors are replaced by the corresponding components of the HMM model mean $\boldsymbol{\mu}_q$ in classical imputation, modified imputation finds the maximum a posteriori estimate

$$\hat{\mathbf{o}}_k(\tau)=\arg\max_{\mathbf{o}_k(\tau)}\mathrm{p}(\mathbf{o}_k(\tau)\mid\mathbf{Y}_k(\tau),q).\tag{53}$$

Assuming a flat prior for $\mathbf{o}_k(\tau)$, as shown in [5], (53) leads to

$$\hat{\mathbf{o}}_k(\tau)=\arg\max_{\mathbf{o}_k(\tau)}\mathrm{p}(\mathbf{o}_k(\tau)\mid\mathbf{Y}_k(\tau))\mathrm{p}(\mathbf{o}_k(\tau)\mid q)\ .\tag{54}$$

Finally, the modified imputation estimate of the feature vector $\hat{\mathbf{o}}_k$ in state q

$$\begin{aligned}&\hat{\mathbf{o}}_{k,q}(\tau)\\&\quad=\left(\boldsymbol{\Sigma}_k^{\text{CEPS}}(\tau)^{-1}+\boldsymbol{\Sigma}_q^{-1}\right)^{-1}\cdot\left(\boldsymbol{\mu}_q\boldsymbol{\Sigma}_q^{-1}+\boldsymbol{\mu}_k^{\text{CEPS}}(\tau)\boldsymbol{\Sigma}_k^{\text{CEPS}}(\tau)^{-1}\right),\end{aligned}\tag{55}$$

can be obtained. This estimate is used to evaluate the pdf of the HMM state q at time τ, as in conventional recognition or classical imputation.

For mixture-of-Gaussian (MOG) models, (55) is evaluated separately for each mixture component m to obtain separate estimates $\hat{\mathbf{o}}_{k,q,m}(\tau)$, and all mixture component probabilities $\mathrm{p}(\hat{\mathbf{o}}_{k,q,m}(\tau)\mid\boldsymbol{\mu}_{q,m},\boldsymbol{\Sigma}_{q,m})$ are finally added to obtain the feature likelihood for state q via

$$\mathrm{p}(\mathbf{o}_k\mid q)=\sum_{m=1}^{M}w_m\mathrm{p}\left(\hat{\mathbf{o}}_{k,q,m}(\tau)\mid\boldsymbol{\mu}_{q,m},\boldsymbol{\Sigma}_{q,m}\right),\tag{56}$$

where w_m stands for the mixture weight of component m. This, again, is analogous to the process in conventional recognition or classical imputation.

7. Experiments

7.1. Room Recordings. For the evaluation of the proposed approaches, recordings were made in a noisy lab room with a reverberation time of $T_{60}\approx 160$ ms. In these recordings, audio files from the TIDigits database [39] were used and mixtures with two and three speakers were recorded at

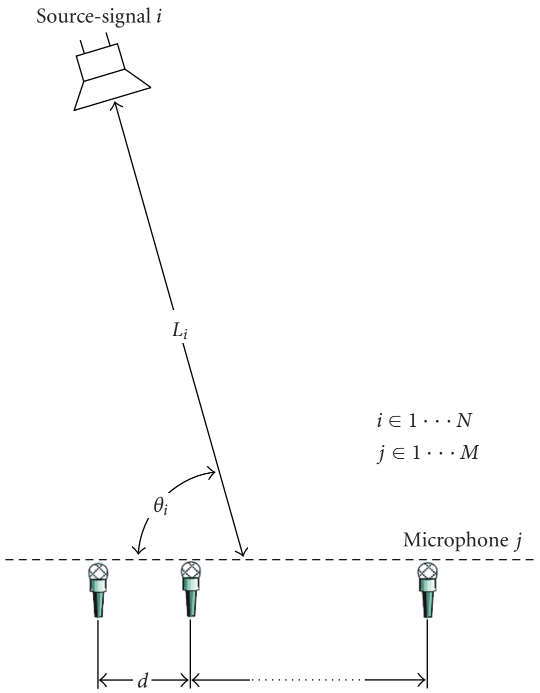

Figure 4: Experimental Setup.

f_s =11 kHz. The distance L_i between the loudspeakers and the center of the microphone array was varied between 0.9 and 3 m. The experimental setup is shown schematically in Figure 4. The distance d between two sensors was 3 cm and a linear array of four microphones was used in all experiments. The recording conditions for all mixtures are summarized in Tables 1 and 2.

7.2. Model Training. The HMM speech recognizer was trained with the HTK toolkit [38]. The trained HMMs comprised phoneme-level models with 6-component MOG emitting probabilities and a conventional left-right structure. The training data was mixed and it comprised the 114 speakers of the TI-DIGITS clean speech database along with the room recordings for speakers sa and rk used for adaptation. Speakers used for adaptation were removed from the test set. The feature extractions presented in Section 5 were also complemented with cepstral mean subtraction (CMS) for further reduction of convolutive effects. Since CMS is a linear operation, it poses no additional difficulty for uncertainty propagation.

7.3. Parameter Settings of Time-Frequency Masks. Parameters of all masks were set manually for good performance on all datasets, and were kept consistent throughout all experiments.

7.3.1. Amplitude-Based Masking. For amplitude-based masking, a soft mask according to (14) and (16) was used. Thus, there are two parameters, the mask threshold T and the gain g, which were set to $T = 0$ and $g = 1$, respectively.

Table 1: Mixture description.

Mixture	Mix. 1	Mix. 2	Mix. 3
Number of speakers N	2	3	2
Speaker Codes	ar,ed	pg,ed,cp	fm,pg
Distance between speaker i and array center	$L_1 = L_2 = 2.0$ m	$L_1 = L_2 = L_3 = 0.9$ m	$L_1 = 1.0$ m $L_2 = 3.0$ m
Angular position of the speaker i (as shown in Figure 4)	$\theta_1 = 75^\circ$ $\theta_2 = 165^\circ$	$\theta_1 = 30^\circ$ $\theta_2 = 80^\circ$	$\theta_1 = 50^\circ$ $\theta_2 = 100^\circ$

Table 2: Mixture description.

Mixture	Mix. 4	Mix. 5
Number of speakers N	2	3
Speaker Codes	cp,ed	fm,ga,ed
Distance between speaker i and array center	$L_1 = L_2 = 0.9$ m	$L_1 = L_2 = L_3 = 0.9$ m
Angular position of the speaker i (as shown in Figure 4)	$\theta_1 = 50^\circ$ $\theta_2 = 115^\circ$	$\theta_1 = 40^\circ$ $\theta_2 = 60^\circ$ $\theta_3 = 105^\circ$

7.3.2. Phase-Based Masking. In phase-based masking according to (22), there are two free parameters as well, again a mask gain g and also a mask threshold, the angle threshold θ_T. However, optimum performance was reached for different parameter values depending on the recognizer parameterization. For optimal performance on MFCC features, they were set to $g = 20$ and $\theta_T = 0.2\pi$, which will be refered to as *Phase1* in the results. In contrast, for RASTA-PLP-based recognition, better results were generally achieved with $g = 15$ and $\theta_T = 0.2\pi$ (*Phase2*), that is, the same threshold but less steep of a mask gain.

7.3.3. Interference-Based Masking. For the first interference-based mask, defined in Section 3.3.1, the two smoothing parameters defining the algorithm are set to $\alpha = 0.1$ and $\alpha_D = 0.9$. This algorithm will be denoted by *IB* in the following.

The second interference-based algorithm additionally includes the speech probability estimate defined in Section 3.3.2. Thus, in addition to the parameters $\alpha = 0.9$ and $\alpha_D = 0.9$, there are additional parameters in the weighting function (34). These are λ_s, λ_n and g, parameters specifying the two threshold points and the mask gain. They are defined to correspond to the mean absolute value of the estimated signal Fourier coefficients $\lambda_s = \overline{f_{S,k}}$, the mean absolute value of the noise estimate Fourier coefficient $\lambda_n = \overline{f_{N,k}}$; and the mask gain is set to $g = 10$. For windowing in (33), a Hanning window of size 3 × 3 is used. For this algorithm, the abbreviation *IBPE* will be used.

8. Results

8.1. Recognition Performance Measurement. To evaluate recognition performance, the number of reference labels

Table 3: Word accuracy (WA) of ASR tests for RASTA-PLP features, estimated uncertainties. Here, the algorithms *Phase1* and *Phase2* utilize the parameters defined in Section 7.3.2, the entries with the heading *Amplitude* correspond with the mask given in Section 7.3.1, and the two interference-based strategies *IB* and *IBPE* are specified in Section 7.3.3. The two robust recognition strategies are abbreviated by *MI* for modified imputation and *UD* for uncertainty decoding.

	Phase1		Phase2		Amplitude		IB		IBPE	
No. of speakers	2	3	2	3	2	3	2	3	2	3
no ICA	31.4	6.3	31.4	6.3	31.4	6.3	31.4	6.3	31.4	6.3
only ICA	58.0	48.0	63.1	52.7	63.1	52.7	63.1	52.7	63.1	52.7
ICA + Mask	15.1	16.7	16.4	17.1	52.1	51.9	02.2	00.9	38.9	30.7
ICA + Mask + UD	52.8	50.7	66.0	59.1	67.2	60.3	66.0	56.5	67.7	59.5
ICA + Mask + MI	60.0	58.0	**73.4**	69.1	72.9	68.8	71.4	67.1	72.9	**69.2**

Table 4: Word accuracy (WA) of ASR tests for MFCC+Δ + $\Delta\Delta$ features, estimated uncertainties.

	Phase1		Phase2		Amplitude		IB		IBPE	
No. of speakers	2	3	2	3	2	3	2	3	2	3
no ICA	37.7	17.2	37.7	17.2	37.7	17.2	37.7	17.2	37.7	17.2
only ICA	69.9	67.4	70.0	68.6	70.0	68.6	70.0	68.6	70.0	68.6
ICA + Mask	03.5	03.7	03.3	03.6	62.1	60.6	04.3	04.2	38.7	35.7
ICA + Mask + UD	72.9	73.2	75.2	73.7	74.2	72.5	75.1	73.7	**76.7**	74.7
ICA + Mask + MI	72.3	**75.4**	71.4	70.7	70.0	68.9	69.9	70.1	72.0	72.2

(N), of substitutions (S), insertions (I) and deletions (D) are counted. From these values, the recognition accuracy PA is defined as

$$\text{PA} = 100 \cdot \frac{N - D - S - I}{N}. \tag{57}$$

The value of PA, output by the HTK scoring tool, corresponds with 100 − WER, where WER is the word error rate that is also commonly used in the evaluation of speech recognition performance.

8.2. Multispeaker Recognition Results. At first, results are given for the estimated uncertainty values and RASTA-PLP features in Table 3 and for MFCC features in Table 4. Especially for RASTA-PLP features, results are improved notably by masking and missing feature recognition by modified imputation, averaging an absolute improvement of more than 10% over all tested masks and experiments. For MFCCs, significant improvements can also be achieved by the suggested strategy. This is true especially for the two strategies of phase masking and interference-based filtering with speech probability estimation. In both cases, an absolute improvement of about 6% can be achieved. It is also clearly visible that here, uncertainty decoding performs better on average.

When true rather than estimated uncertainties are used, results are again improved greatly, both for RASTA-PLP and for MFCC features, as shown in Tables 5 and 6. Compared to the use of ICA alone, a relative error rate reduction of 59% for uncertainty decoding and of 69% for modified imputation is achieved in the case of RASTA features.

Similar performance gains can be observed in the case of MFCC features, where word error rates can be reduced by 64% and 62% for uncertainty decoding and modified imputation, respectively. Comparing the uncertain recognition strategies, again, modified imputation is on average the better performer for RASTA-PLPs, whereas uncertainty decoding leads to better performance gains for MFCCs. Concerning the masking strategies, it is clear that the IB-mask, which has fairly aggressive parameter settings and an extremely low recognition rate without missing feature approaches, is the best for this case of ideal uncertainties.

9. Conclusion

An overview of the use of independent component analysis for speech recognition under multitalker conditions has been given. As shown by the presented results, the conventional strategy of purely linear source separation can be improved by post-masking in the time-frequency domain, if this is accompanied by missing-feature speech recognition. Especially for three-speaker scenarios, this improves the recognition rate notably. Interestingly, the optimal decoding strategy is apparently dependent on the features that are used for recognition. Whereas modified imputation was clearly superior for RASTA features, better results for MFCC features have almost consistently been achieved by uncertainty decoding, even though uncertainties were estimated in the spectrum domain for both features and propagated to the recognition domain of interest. Further work will be necessary to determine how these results correspond to the degree of model mismatch in both domains, with the aim

Table 5: Word accuracy (WA) of ASR tests for RASTA-PLP features, true uncertainties.

	Phase1		Phase2		Amplitude		IB		IBPE	
No. of speakers	2	3	2	3	2	3	2	3	2	3
no ICA	31.4	6.3	31.4	6.3	31.4	6.3	31.4	6.3	31.4	6.3
only ICA	58.0	48.0	63.1	52.7	63.1	52.7	63.1	52.7	63.1	52.7
ICA + Mask	15.1	16.7	16.4	17.1	52.1	51.9	02.2	00.9	38.9	30.7
ICA + Mask + UD	82.7	76.1	88.0	83.7	81.2	74.0	91.5	84.5	86.2	75.3
ICA + Mask + MI	86.7	80.6	89.8	86.3	85.6	79.0	**92.6**	**88.3**	89.5	82.9

Table 6: Word accuracy (WA) of ASR tests for MFCC+Δ + $\Delta\Delta$ features, true uncertainties.

	Phase1		Phase2		Amplitude		IB		IBPE	
No. of speakers	2	3	2	3	2	3	2	3	2	3
no ICA	37.7	17.2	37.7	17.2	37.7	17.2	37.7	17.2	37.7	17.2
only ICA	69.9	67.4	70.5	68.6	70.5	68.6	70.5	68.6	70.5	68.6
ICA + Mask	03.5	03.7	03.3	03.6	62.1	60.6	04.3	04.2	38.7	35.7
ICA + Mask + UD	89.9	82.3	91.3	88.6	89.6	83.8	**93.9**	**89.6**	92.8	88.1
ICA + Mask + MI	90.3	87.4	89.5	87.3	87.0	83.4	90.3	88.2	92.1	87.5

of determining an optimal decoding strategy depending on specific application scenarios.

A vital aspect of missing feature recognition is still the estimation of the feature uncertainty. Here, an ideal uncertainty estimate will result in superior recognition performance for all considered test cases and all applied post masks. Since such an ideal uncertainty is not available in practice, the value needs to be estimated from available data. In the presented cases, this measure has been derived from the ICA output signal and the applied nonlinear gain function. The resulting uncertainty estimate has a correlation coefficient of 0.45 with the true uncertainties, leading to superior and consistent performance among all tested uncertainty estimates.

However, uncertainty estimation for the ICA output signals should be improved further, in order to approximate more closely the ideally achievable performance of this strategy. For this purpose, it will be interesting to compare the proposed uncertainty estimation to other approaches. Specifically, the uncertainty estimation described in [7] is of interest for use with any type of recognition feature and preprocessing method, but it requires learning of a regression tree for the given specific feature set and environment. In contrast, feature-specific methods described for example in [2, 3] are applicable only to the feature domain they have been derived for, but can be used without the need for additional training stages.

Since none of the above methods is designed specifically for use with ICA, another direction of research is a better use of the statistical information gathered during source separation. Further research can thus focus on an optimal use of this intermediate data, and on its combination with more detailed prior models in the spectrum domain, as those in [29], for arriving at more accurate uncertainty estimates which utilize all avaliable data from multiple microphones.

References

[1] T. T. Kristjansson and B. J. Frey, “Accounting for uncertainty in observations: a new paradigm for robust automatic speech recognition,” in *Proceedings of the IEEE International Conference on Acustics, Speech, and Signal Processing (ICASSP ’02)*, 2002.

[2] L. Deng, J. Droppo, and A. Acero, “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*, vol. 13, no. 3, pp. 412–421, 2005.

[3] V. Stouten, H. Van Hamme, and P. Wambacq, “Application of minimum statistics and minima controlled recursive averaging methods to estimate a cepstral noise model for robust ASR,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP ’06)*, vol. 1, May 2006.

[4] M. Van Segbroeck and H. Van Hamme, “Robust speech recognition using missing data techniques in the prospect domain and fuzzy masks,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP ’08)*, pp. 4393–4396, 2008.

[5] D. Kolossa, A. Klimas, and R. Orglmeister, “Separation and robust recognition of noisy, convolutive speech mixtures using time-frequency masking and missing data techniques,” in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA ’05)*, pp. 82–85, 2005.

[6] M. Kühne, R. Togneri, and S. Nordholm, “Time-frequency masking: linking blind source separation and robust speech recognition,” in *Speech Recognition: Technologies and Applications*, pp. 61–80, IN-TECH, Vienna, Austria, 2008.

[7] S. Srinivasan and D. Wang, “Transforming binary uncertainties for robust speech recognition,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 7, pp. 2130–2140, 2007.

[8] Ö. Yilmaz and S. Rickard, “Blind separation of speech mixtures via time-frequency masking,” *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, 2004.

[9] G. Brown and D. Wang, “Separation of speech by computational auditory scene analysis,” in *Speech Enhancement*, J. Benesty, S. Makino, and J. Chen, Eds., pp. 371–402, Springer, New York, NY, USA, 2005.

[10] R. F. Astudillo, D. Kolossa, P. Mandelartz, and R. Orglmeister, “An uncertainty propagation approach to robust ASR using the ETSI advanced front-end,” *IEEE Journal of Selected Topics in Signal Processing*. In press.

[11] H. Buchner, R. Aichner, and W. Kellermann, “TRINICON: a versatile framework for multichannel blind signal processing,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP ’04)*, vol. 3, pp. 889–892, 2004.

[12] S. Makino, T.-W. Lee, and H. Sawada, Eds., *Blind Speech Separation*, Springer, New York, NY, USA, 2007.

[13] K. Kumatani, J. McDonough, D. Klakow, P. N. Garner, and W. Li, “Adaptive beamforming with a maximum negentropy criterion,” in *Proceedings of the Hands-Free Speech Communication and Microphone Arrays (HSCMA ’08)*, pp. 180–183, Trento, Italy, January 2008.

[14] N. Roman, D. Wang, and G. J. Brown, “Speech segregation based on sound localization,” *Journal of the Acoustical Society of America*, vol. 114, no. 4 I, pp. 2236–2252, 2003.

[15] G. J. Brown and M. Cooke, “Computational auditory scene analysis,” *Computer Speech and Language*, vol. 8, no. 4, pp. 297–336, 1994.

[16] A. Cichocki and S. Amari, *Adaptive Blind Signal and Image Processing*, John Wiley & Sons, New York, NY, USA, 2002.

[17] H. Sawada, S. Araki, R. Mukai, and S. Makino, “Blind extraction of dominant target sources using ICA and time-frequency masking,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 6, pp. 2165–2173, 2006.

[18] E. Hoffmann, D. Kolossa, and R. Orglmeister, “A batch algorithm for blind source separation of acoustic signals using ICA and time-frequency masking,” in *Proceedings of the 7th International Conference on Independent Component Analysis and Signal Separation (ICA ’07)*, pp. 480–487, London, UK, 2007.

[19] K. Kamata, X. Hu, and H. Kobatake, “A new approach to the permutation problem in frequency domain blind source separation,” in *Proceedings of the 5th International Conference on Independent Component Analysis and Blind Signal Separation (ICA ’04)*, pp. 849–856, Granada, Spain, September 2004.

[20] D. Kolossa and R. Orglmeister, “Nonlinear postprocessing for blind speech separation,” in *Proceedings of the 5th International Conference on Independent Component Analysis and Signal Separation (ICA ’04)*, pp. 832–839, Granada, Spain, 2004.

[21] M. S. Pedersen, D. Wang, J. Larsen, and U. Kjems, “Overcomplete blind source separation by combining ICA and binary time-frequency masking,” in *Proceedings of the IEEE Workshop on Machine Learning for Signal Processing*, pp. 15–20, September 2005.

[22] D. Kolossa, *Independent component analysis for environmentally robust speech recognition*, Ph.D. dissertation, TU Berlin, Berlin, Germany, 2007.

[23] S. Araki, S. Makino, Y. Hinamoto, R. Mukai, T. Nishikawa, and H. Saruwatari, “Equivalence between frequency-domain blind source separation and frequency-domain adaptive beamforming for convolutive mixtures,” *EURASIP Journal on Applied Signal Processing*, vol. 2003, no. 11, pp. 1157–1166, 2003.

[24] Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean-square error-log-spectral amplitude estimator,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 443–445, 1985.

[25] I. Cohen, “Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator,” *IEEE Signal Processing Letters*, vol. 9, no. 4, pp. 113–116, 2002.

[26] I. Cohen, “On speech enhancement under signal presence uncertainty,” in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP ’01)*, pp. 167–170, Salt Lake City, Utah, USA, May 2001.

[27] Y. Ephraim and I. Cohen, “Recent advancements in speech enhancement,” in *The Electrical Engineering Handbook*, CRC Press, Boca Raton, Fla, USA, 2006.

[28] R. F. Astudillo, D. Kolossa, and R. Orglmeister, “Propagation of statistical information through non-linear feature extractions for robust speech recognition,” in *Proceedings of the 27th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering (MaxEnt ’07)*, vol. 954, pp. 245–252, November 2007.

[29] B. Raj, M. L. Seltzer, and R. M. Stern, “Reconstruction of missing features for robust speech recognition,” *Speech Communication*, vol. 43, no. 4, pp. 275–296, 2004.

[30] S. B. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[31] H. Hermansky and N. Morgan, “RASTA processing of speech,” *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, 1994.

[32] S. Julier and J. Uhlmann, “A general method for approximating nonlinear transformations of probability distributions,” Tech. Rep., University of Oxford, Oxford, UK, 1996.

[33] R. F. Astudillo, D. Kolossa, and R. Orglmeister, “Uncertainty propagation for speech recognition using RASTA features in highly nonstationary noisy environments,” in *Proceedings of the Workshop for Speech Communication (ITG ’08)*, 2008.

[34] M. Gales, *Model-based techniques for noise robust speech recognition*, Ph.D. thesis, Cambridge University, Cambridge, UK, 1996.

[35] H. Hermansky, B. A. Hanson, and H. Wakita, “Perceptually based linear predictive analysis of speech,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP ’85)*, pp. 509–512, 1985.

[36] J. Barker, P. Green, and M. Cooke, “Linking auditory scene analysis and robust ASR by missing data techniques,” in *Proceedings of the Workshop on Innovation in Speech Processing (WISP ’01)*, 2001.

[37] J. Arrowood and M. Clements, “Using observation uncertainty in HMM decoding,” in *Proceedings of the International Conference on Spoken Language Processing (ICSLP ’02)*, 2002.

[38] S. Young, “The HTK Book (for HTK Version 3.4),” Cambridge University, Engineering Department.

[39] R. G. Leonard, “Database for speaker-independent digit recognition,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP ’84)*, vol. 3, 1984.