Chapter 12
Recognition of Multiple Speech Sources using ICA

Eugen Hoffmann, Dorothea Kolossa and Reinhold Orglmeister

Abstract In meetings or noisy public places, often a number of speakers are active simultaneously and the sources of interest need to be separated from interfering speech in order to be robustly recognized. Independent component analysis (ICA) has proven to be a valuable tool for this purpose. However, under difficult environmental conditions, ICA outputs may still contain strong residual components of the interfering speakers. In such cases, time-frequency masking can be applied to the ICA outputs to reduce remaining interferences. In order to remain robust against possible resulting artifacts and loss of information, treating the processed speech feature vector as a random variable with time-varying uncertainty, rather than as deterministic, is a helpful strategy. This chapter shows the possibilities of improving recognition of multiple speech signals based on nonlinear postprocessing, applied together with uncertainty-based decoding techniques.

12.1 Introduction

In order for speech recognition to perform well in arbitrary, noisy environments, it is of special importance to suppress interfering speech, which poses significant problems for noise reduction algorithms due to the overlapping spectra and nonstationarity. In such cases, blind source separation can often be of great value, since it is applicable for any set of signals that is at least statistically independent and non-Gaussian, which is a fairly mild requirement. Blind source separation itself can, however, profit significantly from an additional nonlinear postprocessing, in order to suppress speech or noise which remains in the separated components. Such nonlinear postprocessing functions have been shown to result in SNR improvements in excess of 10dB, e.g. in [40].

Electronics and Medical Signal Processing Group, TU Berlin, Einsteinufer 17, 10587 Berlin.
e-mail: eugen.hoffmann.1@tu-berlin.de, dorothea.kolossa@gmx.de, reinhold.orglmeister@tu-berlin.de

However, while the results of source separation are greatly improved by nonlinear postprocessing, speech recognition results often suffer from artifacts and loss in information due to such postprocessing. In order to compensate for these losses and artifacts 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 the suppressed components of target speech, and will be described in more detail in Section 12.3. 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 uncertainty propagation, which converts an uncertain description of speech from the spectrum domain, where ICA takes place, to the feature domain of speech recognition, as described in Chapter 2. After this uncertainty propagation, recognition can take place under observation uncertainty, using uncertainty-of-observation techniques.

The entire process is vitally dependent on the appropriate estimation of uncertainties. Results given in Section 12.4.7 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 given, and the experiments and results in Section 12.4 show that with this practically available uncertainty measure, significant improvements of recognition performance can be attained for noisy and 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 [16, 18, 29, 32], but in some cases also for speech recognition in multi-talker conditions, as for example [10, 41], or [47] in conjunction with speech segregation via binary masking (see e.g. [9, 49]).

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 combine the strengths of independent component analysis and soft time-frequency masking, and to be still used with a wide range of feature parameterizations. Corresponding results are shown here for MFCC coefficients, but the discussed uncertainty transformation approach also generalizes well to the ETSI advanced front end, as shown in [48], and has been successfully used for time-frequency masking of ICA results in conjunction with RASTA-PLP features as well in [6].

12.2 Blind Source Separation

Blind source separation (BSS) is a technique of recovering the source signals of interest using only observed mixtures, when both the mixing parameters and the sources are unknown. Due to a large number of applications, for example in medical and speech signal processing, BSS has gained great attention. In the following

chapter, we will discuss the application of BSS for acoustic signals observed in a real environment, i.e. convolutive mixtures of multiple speakers recorded under mildly noisy and reverberant distant-talking conditions.

In recent years, this problem has been widely studied and a number of different approaches have been proposed [1–3]. Many existing unmixing methods of acoustic signals are based on Independent Component Analysis (ICA) in the frequency domain, where the convolutions of the source signals with the room impulse response are reduced to multiplications with the corresponding transfer functions. So for each frequency bin, an individual instantaneous ICA problem can be solved in order to obtain the unmixed sources in the frequency domain [3].

Alternative methods include adaptive beamforming, which is closely related to independent component analysis when information theoretic cost functions are applied [8], sparsity based methods that utilize amplitude-delay-histograms [9, 10], or grouping cues typical of human stream segregation [11]. 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.

12.2.1 Problem Formulation

This section provides an introduction into the problem of blind separation of acoustic signals.

At first, a general situation will be considered. In a reverberant room, N acoustic signals $\mathbf{s}(t) = [s_1(t), \ldots s_N(t)]$ are simultaneously present, where t represents the discrete time index. The vector of the source signals $\mathbf{s}(t)$ is recorded with M microphones placed in the room, so that an observation vector $\mathbf{x}(t) = [x_1(t), \ldots x_M(t)]$ results. Due to the time delay and to the signal reflections, the resulting mixture $\mathbf{x}(t)$ is a result of a convolution of the source signal vector $\mathbf{s}(t)$ with unknown filter matrices $\{\underline{\mathbf{a}}_1 \ldots \underline{\mathbf{a}}_K\}$ where $\underline{\mathbf{a}}_k$ is the k-th ($k \in [1 \ldots K]$) $M \times N$ matrix with filter coefficients and K is the filter length[1]. This problem can be summarized by

$$\mathbf{x}(t) = \sum_{k=0}^{K-1} \underline{\mathbf{a}}_{k+1} \mathbf{s}(t-k) + \mathbf{n}(t). \tag{12.1}$$

The term $\mathbf{n}(t)$ denotes the additive sensor noise. Now the problem is to find filter matrices $\{\underline{\mathbf{w}}_1 \ldots \underline{\mathbf{w}}_{K'}\}$ so that by applying them to the observation vector $\mathbf{x}(t)$ the source signals can be estimated via

$$\hat{\mathbf{s}}(t) = \mathbf{y}(t) = \sum_{k'=0}^{K'-1} \underline{\mathbf{w}}_{k'+1} \mathbf{x}(t-k') \tag{12.2}$$

[1] In all following work, only the situation with $M \geq N$ is considered. For $M < N$, the so-called underdetermined case, see e.g. [9].

with K' as the filter length. In other words, for the estimated vector $\mathbf{y}(t)$ and the source vector $\mathbf{s}(t)$, $\mathbf{y}(t) \approx \mathbf{s}(t)$ should hold.

This problem is also known as cocktail-party-problem. A common way to deal with the problem is to reduce it to a set of the instantaneous source separation problems, for which efficient approaches exist.

For this purpose, the time-domain observation vectors $\mathbf{x}(t)$ are transformed into a frequency domain time series by means of the short time Fourier transform (STFT)

$$\mathbf{X}(\Omega,\tau) = \sum_{t=-\infty}^{\infty} \mathbf{x}(t) w(t-\tau R) e^{-j\Omega t}, \tag{12.3}$$

where Ω is the angular frequency, τ represents the frame index, $w(t)$ is a window function (e.g., a Hanning window) of length N_{FFT}, and R is the shift size, in samples, between successive windows [12]. Transforming Eq. (12.1) into the frequency domain reduces the convolutions to multiplications with the corresponding transfer functions, so that for each frequency bin an individual instantaneous problem

$$\mathbf{X}(\Omega,\tau) \approx \mathbf{A}(\Omega)\mathbf{S}(\Omega,\tau) + \mathbf{N}(\Omega,\tau) \tag{12.4}$$

arises. $\mathbf{A}(\Omega)$ is the mixing matrix in the frequency domain, $\mathbf{S}(\Omega,\tau) = [S_1(\Omega,\tau), \ldots, S_N(\Omega,\tau)]$ represents the source signals, $\mathbf{X}(\Omega,\tau) = [X_1(\Omega,\tau),\ldots,X_M(\Omega,\tau)]$, denotes the observed signals, and $\mathbf{N}(\Omega,\tau)$ is the frequency domain representation of the additive sensor noise. In order to reconstruct the source signals, the unmixing matrix $\mathbf{W}(\Omega) \approx \mathbf{A}^{+}(\Omega)$ is derived[2] using a complex-valued unmixing algorithm, so that

$$\mathbf{Y}(\Omega,\tau) = \mathbf{W}(\Omega)\mathbf{X}(\Omega,\tau) \tag{12.5}$$

can be used for obtaining estimated sources in the frequency domain. Here, $\mathbf{Y}(\Omega,\tau) = [Y_1(\Omega,\tau), \ldots, Y_N(\Omega,\tau)]$ is the time frequency representation of the unmixed outputs.

12.2.2 ICA

Independent Component Analysis (ICA) is an approach that can help to find optimal unmixing matrices $\mathbf{W}$. The main idea is to obtain statistical independence of the output signals, which is mathematically defined in terms of probability densities. The components of the vector $\mathbf{Y}$ are statistically independent if and only if the joint probability distribution function $f_{\mathbf{Y}}(\mathbf{Y})$ is equal to the product of the marginal distribution functions of each signal Y_i

$$f_{\mathbf{Y}}(\mathbf{Y}) = \prod_i f_{Y_i}(Y_i). \tag{12.6}$$

[2] $\mathbf{A}^{+}(\Omega)$ denotes the pseudo inverse of $\mathbf{A}(\Omega)$.

The process of finding the unmixing matrix $\mathbf{W}$ is now composed of two steps:

- the definition of a contrast function $\mathscr{J}(\mathbf{W})$, which is a quantitative measure of the statistical independence of all components in $\mathbf{Y}$ and
- the minimization of $\mathscr{J}(\mathbf{W})$ so that

$$\hat{\mathbf{W}} \stackrel{!}{=} \arg\min_{\mathbf{W}} \mathscr{J}(\mathbf{W}). \tag{12.7}$$

At this point, the definition of the contrast function $\mathscr{J}(\mathbf{W})$ is the key for the problem solution. For this purpose, it is possible to focus on different aspects of statistical independence, which results in the large number of ICA algorithms that have been proposed during the last decades [2]. The most common approaches use one of the following characteristics of independent signals:

- The higher order cross statistic tensor of independent signals is diagonal, so $\mathscr{J}(\mathbf{W})$ is defined as a sum of the off-diagonal elements of, e.g., the fourth order cross cumulant (JADE algorithm [13]).
- Each independent component remains independent in time, so the cross correlation matrix $\mathbf{C}(\tau) = \mathrm{E}[\mathbf{Y}(t)\mathbf{Y}(t+\tau)^T]$ remains diagonal, i.e.

$$\mathbf{C}(\tau) = \begin{pmatrix} \mathrm{R}_{Y_1Y_1}(\tau) & 0 & \cdots & 0 \\ 0 & \mathrm{R}_{Y_2Y_2}(\tau) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathrm{R}_{Y_NY_N}(\tau) \end{pmatrix} \tag{12.8}$$

  for each τ (SOBI algorithm [14]).
- The mutual information $\mathrm{I}(\mathbf{X}, g(\mathbf{Y}))$, with $\mathrm{g}(\cdot)$ as a nonlinear function, achieves its maximum when the components of $\mathbf{Y}$ are statistically independent. This assumption leads to a solution

$$\hat{\mathbf{W}} \stackrel{!}{=} \arg\max_{\mathbf{W}} \mathrm{H}(\mathrm{g}(\mathbf{Y})) \tag{12.9}$$

  where $H(g(\mathbf{Y}))$ is the joint entropy of $g(\mathbf{Y})$. This is known as the information maximization approach [15].
- If the distributions of the independent components are non-Gaussian, the search for maximal independence results in a search for maximal non-Gaussianity [17]. In this case, the negentropy

$$\mathscr{J}(\mathbf{W}) = \mathrm{H}(\mathbf{Y}_{\mathrm{Gauss}}) - \mathrm{H}(\mathbf{Y}) \tag{12.10}$$

  plays the role of the cost function, where $\mathbf{Y}_{\mathrm{Gauss}}$ is a vector valued Gaussian random variable of the same mean and covariance matrix as $\mathbf{Y}$.

The last assumption leads to an approach proposed by Hyvärinen et al. [17] also known as the FastICA algorithm. The problem of this approach is the computation of the negentropy. The calculation according to the Eq. (12.10) would require an

estimate of the probability density functions in each iteration, which is computationally very costly. Therefore, the cost function $\mathscr{J}(\mathbf{W})$ is approximated using a nonquadratic function G

$$\mathscr{J}(\mathbf{W}) \propto \left(\mathrm{E}\left[\mathrm{G}(\mathbf{Y})\right] - \mathrm{E}\left[\mathrm{G}(\mathbf{Y}_{\mathrm{Gauss}})\right]\right)^2. \tag{12.11}$$

Using the approximation of the negentropy from Eq. (12.11), the update rule for the unmixing matrix $\mathbf{W}$ in the i-th iteration can be derived as [1]

$$\tilde{\mathbf{W}}_i = \left\langle \mathrm{g}(\mathbf{Y})\mathbf{X}^H \right\rangle - \Lambda \mathbf{W}_{i-1} \tag{12.12}$$

$$\mathbf{W}_i = \left(\tilde{\mathbf{W}}_i \tilde{\mathbf{W}}_i^T\right)^{-1/2} \tilde{\mathbf{W}}_i \tag{12.13}$$

where Λ is a diagonal matrix with $\lambda_{ii} = \langle \mathrm{g}'(Y_i) \rangle$ and $\langle \cdot \rangle$ denotes the mean value. As for the function $\mathrm{g}(\cdot)$, the derivative of the function $\mathrm{G}(\cdot)$ from the Eq. (12.10) will be chosen, so setting

$$\mathrm{G}(x) = -\exp\left(-\frac{|x|^2}{2}\right) \tag{12.14}$$

the function $\mathrm{g}(\cdot)$ becomes

$$\mathrm{g}(x) = x \exp\left(-\frac{|x|^2}{2}\right) \tag{12.15}$$

and

$$\mathrm{g}'(x) = (1 - x^2) \exp\left(-\frac{|x|^2}{2}\right). \tag{12.16}$$

Ideally, these approaches will result in independent output signals in each frequency bin.

In order to obtain complete spectra of unmixed sources, it is additionally necessary to correctly sort the outputs, since their ordering after solving instantaneous ICA problems for each frequency is arbitrary and may vary from frequency bin to frequency bin. This so-called permutation problem can be solved in a number of ways and will be discussed in the following section.

12.2.3 Permutation Correction

Due to the principle of the ICA algorithms, it is highly unlikely to obtain a consistent ordering of the recovered signals for different frequency bins. In case of frequency domain source separation, this means that the ordering of the outputs may change in each frequency bin. In order to obtain correctly estimated source signals in the

time domain, however, all separated frequency bins have to be put in one consistent order. This problem is also known as the permutation problem.

There exist several classes of algorithms giving a solution for the permutation problem. Approaches presented in [4], [19], and [20] try to correct permutations by considering the cross statistics (such as cross correlation or cross cumulants) of the spectral envelopes of adjacent frequency bins. In [21], algorithms were proposed that make use of the spectral distance between neighboring bins and try to make the impulse response of the mixing filters short, which corresponds to smooth transfer functions of the mixing system in the frequency domain. The algorithm proposed by Kamata et al. [22] solves the problem using the continuity in power between adjacent frequency components of the same source. A similar method was presented by Pham et al. [23]. Baumann et al. [24] proposed a solution which works by comparing the directivity patterns resulting from the estimated demixing matrix in each frequency bin. Similar algorithms were presented in [25], [26] and [27]. In [28], it was suggested to use the direction of arrival (DOA) of source signals implicitly estimated by the ICA unmixing matrices W for the problem solution. The approach in [30] exploits the continuity of the frequency response of the mixing filter. A similar approach was presented in [31] using the minimum of the L_2-norm of the resulting mixing filter and in [33] using the minimum distance between the filter coefficients of adjacent frequency bins. In [34], the authors suggest to use the cosine between the demixing coefficients of different frequencies as a cost function for the problem solution. Sawada et al. [35] proposed an approach using basis vector clustering of the normalized estimated mixing matrices. In [36] the permutation problem was solved using a maximum-likelihood-ratio criterion between the adjacent frequency bins.

However, with a growing number of independent components, the complexity of the solution grows. This is true not only because of the factorial increase of permutations to be considered, but also because of the degradation of the ICA performance. Therefore, not all of the approaches mentioned above perform equally well for an increasing number of sources.

In all following work, permutations have been corrected by maximizing the likelihood-ratio criterion described in [36]. The correction algorithm from [36] was expanded for the case of more than two extracted channels. In order to solve the permutation problem, for each frequency bin a correction matrix $\hat{\mathbf{P}}(\Omega)$

$$\hat{\mathbf{P}}(\Omega) \stackrel{!}{=} \arg \min_{k=1\ldots K} \prod_{i,j \in \left\{ \mathbf{P}^k_{ij}(\Omega)=1 \right\}} \gamma_{ij}(\Omega) \tag{12.17}$$

has to be found, where $\mathbf{P}^k(\Omega)$ is the k-th among K possible permutation matrices, the parameter $\gamma_{ij}(\Omega)$ is

$$\gamma_{ij}(\Omega) = \frac{1}{T} \sum_{\tau} \frac{|Y_i(\Omega,\tau)|}{\beta_j(\tau)} \tag{12.18}$$

and

$$\beta_j(\tau) = \frac{1}{N}\sum_{\Omega}\left|Y_j(\Omega,\tau)\right|. \tag{12.19}$$

In this case, $\beta_j(\tau)$ is a scaling parameter of the signal envelope. $\beta_j(\tau)$ allows to consider the scaling of signals in permutation correction, so that the likelihood of an unmixed source at a given frequency will be weighted with the averaged magnitude of the current frame.

12.2.4 Postmasking

Even after source separation, in the majority of real-world cases, the extracted independent components are still corrupted by residual noise and interference, especially in reverberant environments. The residual disturbances are assumed to be a superposition of the other independent components and the background noise. Therefore, the quality of the recovered source signals often leaves room for improvement, which can be attained in a wide range of scenarios by applying a soft time-frequency-mask to the ICA outputs. While an ideal post processing function is impossible to obtain realistically, approximations to it are already advantageous. As one such approximation, mask estimation based on ICA results has been proposed and shown to be successful, both for binary and soft masks, see e.g. [35, 37, 40].

In this section, a brief review of four types of time-frequency masking algorithms, namely

- amplitude-based masks
- phase-based masks
- interference-based masks and
- a two stage noise suppression algorithm based masks

will be given. The data flow of the whole application is shown in Figure 12.1.

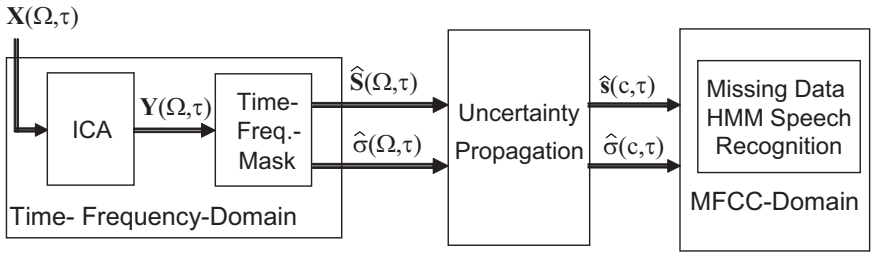

Fig. 12.1: Block diagram with data flow.

12.2.4.1 Amplitude based mask estimation

The time-frequency mask is calculated by comparing the local amplitude ratios between the output signal $Y_i(\Omega,\tau)$ and all other $Y_j(\Omega,\tau)$. With an additional sigmoid transition point T, the mask can be used to block all time-frequency components of Y_i which are not at least T dB above all other estimated sources in that time-frequency point. This corresponds to computing [37]

$$M_i(\Omega,\tau) = \Psi\left(\log\left(|Y_i(\Omega,\tau)|^2\right) - \max_{\forall j\neq i}\log\left(\left|Y_j(\Omega,\tau)\right|^2\right) - \frac{T}{10}\right) \tag{12.20}$$

and applying a sigmoid nonlinearity Ψ defined by

$$\Psi(x) = \frac{1}{1+\exp(-x)}. \tag{12.21}$$

An example of the resulting masks is shown in Fig. 12.2.

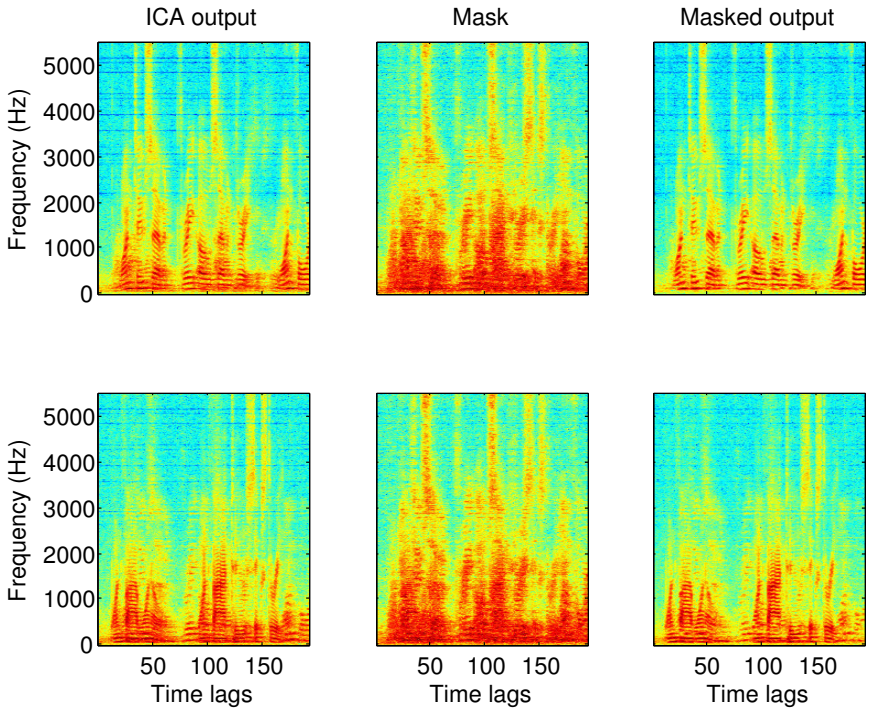

Fig. 12.2: Effect of Amplitude Mask for the case of $M = N = 2$. The spectrograms of 1) the output signals $Y_1(\Omega,\tau)$ and $Y_2(\Omega,\tau)$ obtained only with ICA (left column), 2) the estimated masks $M_1(\Omega,\tau)$ and $M_2(\Omega,\tau)$ (middle column), and 3) the output signals $\hat{S}_1(\Omega,\tau)$ and $\hat{S}_2(\Omega,\tau)$ obtained by a combination of ICA and T-F masking calculated with Eq. (12.20).

12.2.4.2 Phase angle based mask estimation

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 [39]. 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 post-masking of ICA outputs was introduced in [35]. This method considers closeness of the phase angle $\vartheta_i(\Omega,\tau)$ between a column of the mixing matrix $\mathbf{a}_i(\Omega)$ and the observed signal $\mathbf{X}(\Omega,\tau)$ calculated in the whitened space, with $\mathbf{V}(\Omega)=\mathbf{R}^{-1/2}(\Omega)$ as the whitening matrix that is obtained from the signal autocorrelation $\mathbf{R}(\Omega)=\left\langle \mathbf{X}(\Omega,\tau)\mathbf{X}(\Omega,\tau)^H \right\rangle$. ^{H} denotes the conjugate or Hermitian transpose. The phase angle is given by

$$\vartheta_i(\Omega,\tau)=\arccos\frac{\left|\mathbf{b}_i^H(\Omega)\mathbf{Z}(\Omega,\tau)\right|}{\|\mathbf{b}_i(\Omega)\|\,\|\mathbf{Z}(\Omega,\tau)\|}, \tag{12.22}$$

where $\mathbf{Z}(\Omega,\tau)=\mathbf{V}(\Omega)\mathbf{X}(\Omega,\tau)$ are whitened samples and $\mathbf{b}_i(\Omega)=\mathbf{V}(\Omega)\mathbf{a}_i(\Omega)$ is the basis vector i in the whitened space. Then the mask is calculated by

$$M_i(\Omega,\tau)=\frac{1}{1+\exp(g(\vartheta_i(\Omega,\tau)-\vartheta_T))} \tag{12.23}$$

where ϑ_T and g are parameters specifying the sigmoid transition point and steepness, respectively. Figure 12.3 shows exemplary results.

12.2.4.3 Interference based mask estimation

Interference based mask estimation is introduced in detail in [40]. The main idea is to detect the time-frequency points in the separated signals, where the source signal and the interference are dominant, assuming them to be sparse in the time-frequency domain. The mask is estimated by

$$M_i(\Omega,\tau)=\frac{1}{1+\exp(g(\tilde{\mathbf{S}}_i(\Omega,\tau)-\lambda_s))}\times\left(1-\frac{1}{1+\exp(g(\tilde{\mathbf{N}}_i(\Omega,\tau)-\lambda_n))}\right) \tag{12.24}$$

where λ_s,λ_n and g are parameters specifying the threshold points and the steepness of the sigmoid function and $\tilde{\mathbf{S}}_i(\Omega,\tau)$ and $\tilde{\mathbf{N}}_i(\Omega,\tau)$ are speech and noise dominance measures given by

$$\tilde{\mathbf{S}}_i(\Omega,\tau,R_\Omega,R_\tau)=\frac{\left\|\Phi(\Omega,\tau,R_\Omega,R_\tau)(Y_i(\Omega,\tau)-\sum_{m\neq i}Y_m(\Omega,\tau))\right\|}{\left\|\Phi(\Omega,\tau,R_\Omega,R_\tau)\sum_{m\neq i}Y_m(\Omega,\tau)\right\|} \tag{12.25}$$

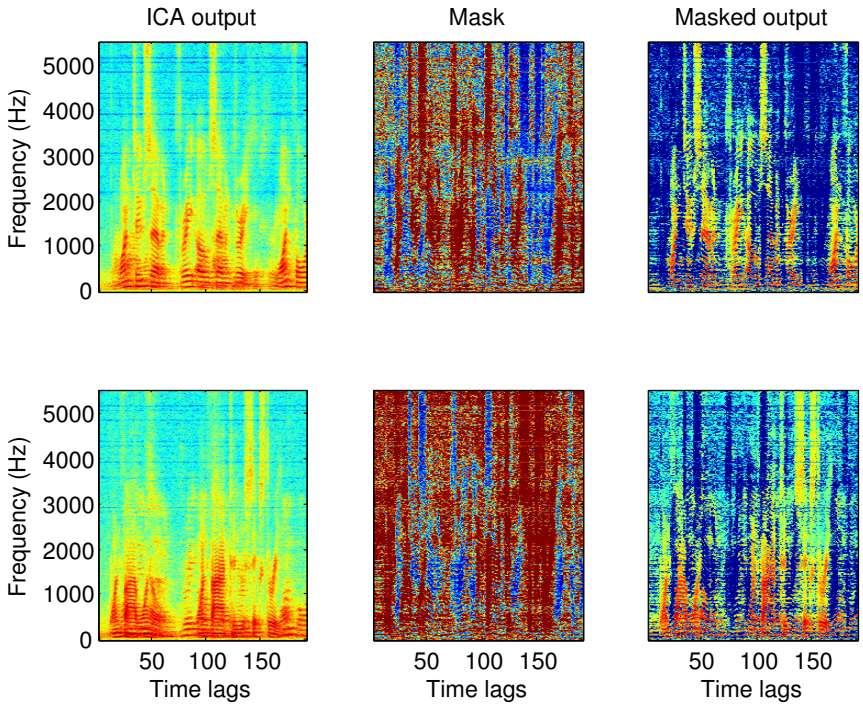

Fig. 12.3: Effect of Phase Mask for the case of $M = N = 2$. The spectrograms of 1) the output signals $Y_1(\Omega,\tau)$ and $Y_2(\Omega,\tau)$ obtained only with ICA (left column), 2) the estimated masks $M_1(\Omega,\tau)$ and $M_2(\Omega,\tau)$ (middle column), and 3) the output signals $\hat{S}_1(\Omega,\tau)$ and $\hat{S}_2(\Omega,\tau)$ obtained by a combination of ICA and T-F masking calculated with Eq. (12.23).

and

$$\tilde{\mathbf{N}}_i(\Omega,\tau,R_\Omega,R_\tau) = \frac{\left\|\Phi(\Omega,\tau,R_\Omega,R_\tau)(Y_i(\Omega,\tau) - \sum_{m\neq i} Y_m(\Omega,\tau))\right\|}{\|\Phi(\Omega,\tau,R_\Omega,R_\tau)Y_i(\Omega,\tau)\|}. \tag{12.26}$$

Here, $\|\cdot\|$ denotes the Euclidean norm operator and

$$\Phi(\Omega,\tau,R_\Omega,R_\tau) = \begin{cases} \mathscr{W}(\Omega-\Omega_0,\tau-\tau_0,R_\Omega,R_\tau), & |\Omega-\Omega_0| \le R_\Omega, \\ & |\tau-\tau_0| \le R_\tau \\ 0, & \text{otherwise} \end{cases} \tag{12.27}$$

utilizes a two dimensional window function $\mathscr{W}(\Omega-\Omega_0,\tau-\tau_0,R_\Omega,R_\tau)$ of the size $R_\Omega \times R_\tau$ (e.g. a two dimensional Hanning window) [40]. This mask tends to result in a very strong suppression of interferences, as can be already gleaned from its visual impression in Fig. 12.4.

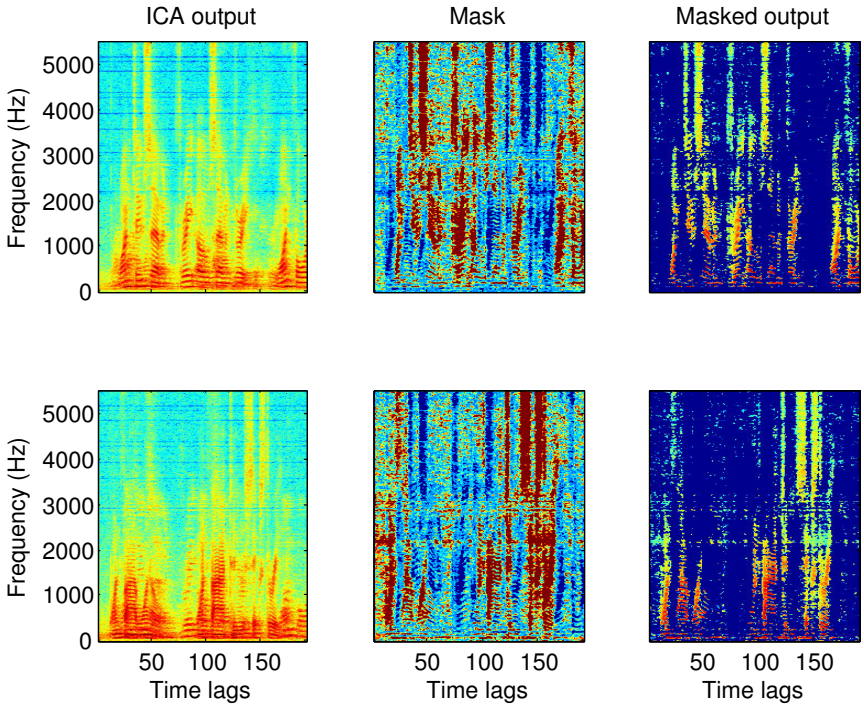

Fig. 12.4: Effect of Interference Mask for the case of $M = N = 2$. The spectrograms of 1) the output signals $Y_1(\Omega,\tau)$ and $Y_2(\Omega,\tau)$ obtained only with ICA (left column), 2) the estimated masks $M_1(\Omega,\tau)$ and $M_2(\Omega,\tau)$ (middle column), and 3) the output signals $\hat{S}_1(\Omega,\tau)$ and $\hat{S}_2(\Omega,\tau)$ obtained by a combination of ICA and T-F masking calculated with Eq. (12.24).

12.2.4.4 Two stage noise suppression

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. For this purpose, the following signal model is assumed

$$\mathbf{Y}(\Omega,\tau) = \mathbf{S}(\Omega,\tau) + \mathbf{N}(\Omega,\tau), \tag{12.28}$$

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

$$\hat{\mathbf{S}}(\Omega,\tau) = \mathbf{M}_{SE}(\Omega,\tau)\mathbf{Y}(\Omega,\tau), \tag{12.29}$$

where $\mathbf{M}_{SE}(\Omega,\tau)$ is the amplitude estimator gain. For the calculation of the gain $\mathbf{M}_{SE}(\Omega,\tau)$ in Eq. (12.34), different speech enhancement algorithms can be used. In [42] the method by McAulay and Malpass [43] has been used. In the following,

we use the log spectral amplitude estimator (LSA) as proposed by Ephraim and Malah [38].

In case of the LSA estimator, first the a posteriori SNR $\gamma_i(\Omega,\tau)$ and the a priori SNR $\xi_i(\Omega,\tau)$ are defined by

$$\gamma_i(\Omega,\tau) = \frac{|Y_i(\Omega,\tau)|^2}{\lambda_{D,i}(\Omega,\tau)} \tag{12.30}$$

and

$$\xi_i(\Omega,\tau) = \alpha\xi_i(\Omega,\tau-1) + (1-\alpha)\frac{\lambda_{X,i}(\Omega,\tau)}{\lambda_{D,i}(\Omega,\tau)}, \tag{12.31}$$

where α is a smoothing parameter that controls the trade-off between the noise reduction and the transient distortions [45], $Y_i(\Omega,\tau)$ is the i-th ICA-output, $\lambda_{D,i}(\Omega,\tau)$ is the noise power and $\lambda_{X,i}(\Omega,\tau)$ is the approximate clean signal power. With these parameters, the log spectral amplitude estimator is given by:

$$\mathbf{M}_{SE}(\Omega,\tau) = \frac{\xi(\Omega,\tau)}{1+\xi(\Omega,\tau)} \exp\left(\int_{t=\nu(\Omega,\tau)}^{\infty} \frac{e^{-t}}{t} dt\right) \tag{12.32}$$

and

$$\nu(\Omega,\tau) = \left(\frac{\xi(\Omega,\tau)}{1+\xi(\Omega,\tau)}\right)\gamma(\Omega,\tau) \tag{12.33}$$

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

According to [44], this approach can be generalized by using additional information for calculation of speech presence probabilities. The speech presence probability $\mathbf{p}(\Omega,\tau)$ can then be used to modify the spectral gain function

$$\mathbf{M}(\Omega,\tau) = \mathbf{M}_{SE}(\Omega,\tau)^{\mathbf{p}(\Omega,\tau)}\mathbf{G}_{min}^{(1-\mathbf{p}(\Omega,\tau))}, \tag{12.34}$$

where $\mathbf{G}_{min}$ is a spectral floor constant [44, 45]. Since the probability functions are not known, the masks from Sections 12.2.4.1-12.2.4.3 can be used at this point as an approximation. Considering $\mathbf{p}(\Omega,\tau) = \mathbf{M}(\Omega,\tau)$ from Eq. (12.24) as the approximate speech presence probability, we estimate the noise power $\lambda_{D,i}(\Omega,\tau)$ as

$$\begin{aligned}\lambda_{D,i}(\Omega,\tau) &= p_i(\Omega,\tau)\lambda_{D,i}(\Omega,\tau-1) \\ &\quad + (1-p_i(\Omega,\tau))\left[\alpha_D\lambda_{D,i}(\Omega,\tau-1) + (1-\alpha_D)\,|Y_i(\Omega,\tau)|^2\right]\end{aligned} \tag{12.35}$$

with α_D as as a smoothing parameter and the approximate clean signal power $\lambda_{X,i}(\Omega,\tau)$ as

$$\lambda_{X,i}(\Omega,\tau) = \left(|Y_i(\Omega,\tau)|\,p_i(\Omega,\tau)\right)^2. \tag{12.36}$$

The effect of the two-stage mask is again a strong interference suppression, however, the spectral distortion is reduced compared to that of the interference mask. This can also be observed from the associated spectrographic representation in Fig. 12.5.

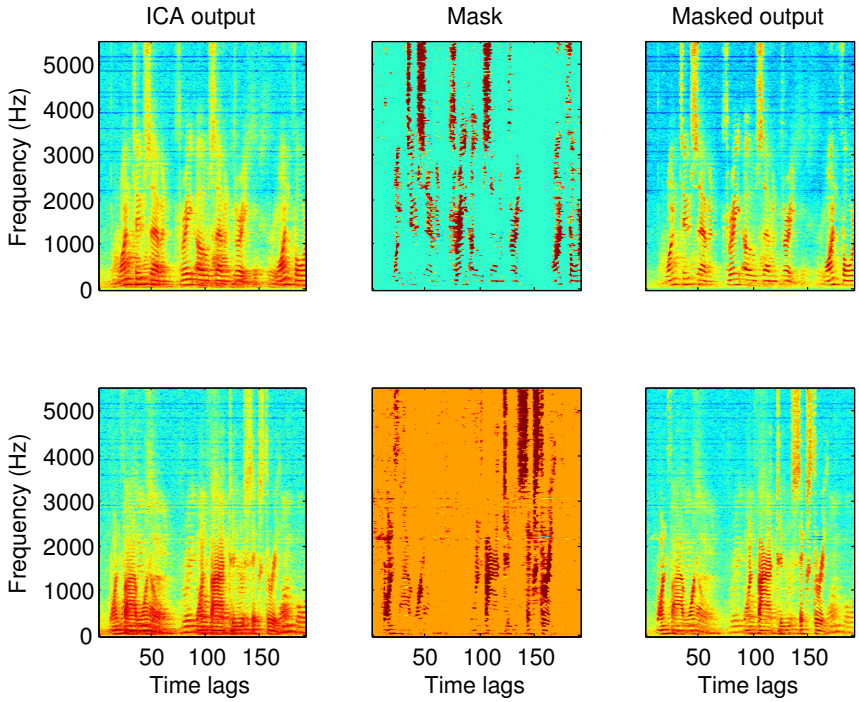

Fig. 12.5: Effect of 2-stage noise suppression for the case of $M = N = 2$. The spectrograms of 1) the output signals $Y_1(\Omega,\tau)$ and $Y_2(\Omega,\tau)$ obtained only with ICA (left column), 2) the estimated masks $M_1(\Omega,\tau)$ and $M_2(\Omega,\tau)$ (middle column), and 3) the output signals $\hat{S}_1(\Omega,\tau)$ and $\hat{S}_2(\Omega,\tau)$ obtained by a combination of ICA and T-F masking calculated with Eq. (12.34).

12.3 Uncertainty Estimation

Because of the use of time-frequency masking, a part of the information of the original signal is often 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, as described in more detail e.g. in [6].

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 [47] and [48]. 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 [50]

$$p(S_k(\Omega,\tau)|\hat{S}_k(\Omega,\tau)) = \frac{1}{\pi\sigma^2(\Omega,\tau)} \exp\left(-\frac{|S_k(\Omega,\tau)-\hat{S}_k(\Omega,\tau)|^2}{\sigma^2(\Omega,\tau)}\right), \quad (12.37)$$

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

12.3.1 Ideal Uncertainties

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

$$\sigma_I^2 = \left||S_k(\Omega,\tau)| - |\hat{S}_k(\Omega,\tau)|\right|^2, \quad (12.38)$$

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.

12.3.2 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

$$\sigma_E^2 = \alpha\left||\hat{S}_k(\Omega,\tau)| - |Y_k(\Omega,\tau)|\right|^2. \quad (12.39)$$

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

$$\alpha = \arg\min_{\tilde{\alpha}} (\sigma_E(\tilde{\alpha}) - \sigma_T)^2. \tag{12.40}$$

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, at which point stereo data was also necessary in order to compute σ_T according to (12.38). After averaging over all mask types, the same value of α was used in all experiments and for all datasets. This optimal value was $\alpha = 0.71$.

12.3.3 Uncertainty Propagation

Once the clean speech features and their uncertainties have been estimated in the STFT domain, the uncertain features need to be made available in that feature domain where speech recognition takes place. In all subsequent experiments and results, this ASR feature domain was the mel-frequency cepstrum.

Therefore, after uncertainty estimation, an additional step of uncertainty propagation was necessary, as it is also shown in Fig. 12.1. For this purpose, the estimated speech signal $\hat{S}_k(\Omega,\tau)$ and its variance $\sigma^2(\Omega,\tau)$ are interpreted as mean and variance of a complex Gaussian distributed random variable. Then, the effect that subsequent MFCC feature extraction stages have on these random variables can be determined. This uncertainty propagation was carried out as described in detail in Chapter 3, and its outputs are the approximate mean and variance of the uncertain speech features, after they have been nonlinearly transformed to the mel-frequency cepstrum domain.

12.4 Experiments and Results

12.4.1 Recording Conditions

For the evaluation of the proposed approaches, different real room recordings were used. In these recordings, audio files from the TIDigits database [46] were used and mixtures with up to 3 speakers were recorded in a mildly reverberant ($T_R \approx 160$ ms) and noisy lab room at TU Berlin. The distances L_i between loudspeakers and microphones were varied between 0.9 and 3 m.

The setup is shown schematically in Figure 12.6 and the experimental conditions are summarized in Table 12.1.

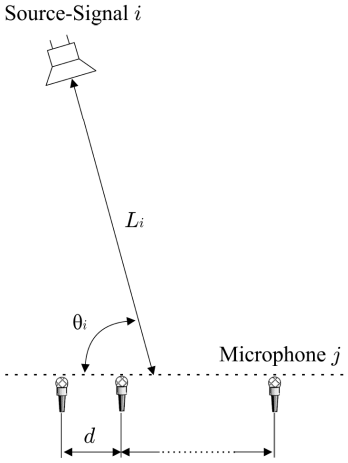

Fig. 12.6: Experimental setup used in recording of mixtures. The distance d between adjacent microphones was 3cm for all recordings.

Table 12.1: Mixture description.

Mixture	Mix. 1	Mix. 2	Mix. 3	Mix. 4	Mix. 5
Number of speakers N	2	3	2	2	3
Number of microphones M	2	3	2	2	3
Speaker Codes	ar,ed	pg,ed,cp	fm,pg	cp,ed	fm,ga,ed
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	$L_1 = L_2 =$ 0.9 m	$L_1 = L_2 =$ $L_3 = 0.9$ m
Angular position of the speaker i (as shown in Fig. 12.6)	$\theta_1 = 75^\circ$ $\theta_2 = 165^\circ$	$\theta_1 = 30^\circ$ $\theta_2 = 80^\circ$ $\theta_3 = 135^\circ$	$\theta_1 = 50^\circ$ $\theta_2 = 100^\circ$	$\theta_1 = 50^\circ$ $\theta_2 = 115^\circ$	$\theta_1 = 40^\circ$ $\theta_2 = 60^\circ$ $\theta_3 = 105^\circ$

12.4.2 Parameter Settings

The algorithms were tested on the room recordings, which were first transformed to the frequency domain at a resolution of $N_{\text{FFT}} = 1024$. For calculating the STFT, the signals were divided into overlapping frames using a Hanning window with an overlap of $3/4 \cdot N_{\text{FFT}}$. For the BSS, the FastICA algorithm (Eq. (12.12)-(12.13)) with the nonlinearity $g(\cdot)$ from Eq. (12.15) and $g'(\cdot)$ from Eq. (12.16) was used.

The parameter settings for the different evaluated time-frequency masks are summarized in Table 12.2.

Table 12.2: Parameter Settings.

Algorithm	Settings
Amplitude based mask	$T = 0$ dB
Phase angle based mask	$\vartheta_T = 0.3 \cdot \pi$ $g = 50$
Interference based mask	$\lambda_s = 1.3$ $\lambda_n = 1.3$ $g = 20$
Two stage noise suppression	$\alpha = 0.7$ $\alpha_D = 0.3$ $\mathbf{G}_{min} = 0.1$

12.4.3 Performance Measures

For determining of the performance of ICA and time-frequency masking, the signal to interference ratio (SIR) was used as a measure of the separation performance and the signal to distortion ratio (SDR) as a measure of the signal quality. The SIR improvement ΔSIR is obtained from

$$\Delta\mathrm{SIR}_i = 10\log_{10}\frac{\sum_n y_{i,s_i}^2(n)}{\sum_{j\neq i}\sum_n y_{i,s_j}^2(n)} - 10\log_{10}\frac{\sum_n x_{i,s_i}^2(n)}{\sum_{j\neq i}\sum_n x_{i,s_j}^2(n)}, \tag{12.41}$$

and the SDR is computed by

$$\mathrm{SDR}_i = 10\log_{10}\frac{\sum_n x_{k,s_i}^2(n)}{\sum_n (x_{k,s_i}(n) - \alpha y_{i,s_i}(n-D))^2}, \tag{12.42}$$

where y_{i,s_j} is the i-th separated signal with only the source signal s_j active, and x_{k,s_j} is the observation obtained by microphone k, again when only s_j is active. α and D are parameters for phase and amplitude which are chosen automatically to optimally compensate the difference between y_{i,s_j} and x_{k,s_j}.

12.4.4 Separation Results

All the mixtures from Table 12.1 were separated with the FastICA algorithm and subsequently the time frequency masking from Sections 12.2.4.1-12.2.4.4 was performed using parameter settings as shown in Section 12.4.2. For each result, the performance was calculated using Eq. ((12.41))-((12.42)). Table 12.3 shows the results of the applied methods.

As can be seen in Table 12.3, the best SIR improvements were achieved by the two stage approach. Still, the results of all time-frequency masks depend on the

performance of the preceding BSS algorithm, which in turn depends on the experimental setup. As can be seen, the best BSS results were generally achieved when the microphones were placed near the source signals. Thus, given a low ICA performance for large microphone distances (in terms of SIR and SDR), a stronger signal distortion should be expected from subsequent masking as well. Furthermore, the higher the SIR improvement gained with a time-frequency mask, the lower the SDR value tends to become. The consequence of this for speech recognition will be discussed further in Section 12.4.7.

Table 12.3: Experimental results (mean value of output ΔSIR/SDR in dB).

Scenario	none	Amplitude	Phase	Interference	2-stage
Mix. 1	3.48 / 5.13	6.35 / 3.98	4.93 / **4.38**	8.43 / 2.48	**8.57** / 2.84
Mix. 2	9.06 / 4.23	11.99 / **4.10**	13.76 / 3.86	16.88 / 2.68	**17.25** / 2.87
Mix. 3	6.14 / 6.33	11.20 / 5.39	9.11 / **5.88**	14.11 / 3.78	**14.14** / 4.14
Mix. 4	8.24 / 8.68	14.56 / 7.45	11.32 / **7.91**	**19.04** / 4.88	18.89 / 5.33
Mix. 5	3.93 / 2.92	5.24 / 2.41	6.70 / **2.66**	9.31 / 0.84	**9.55** / 1.11

12.4.5 Model Training

The HMM speech recognizer was trained with the HTK toolkit [51]. The HMMs were trained at the phoneme-level with 6-component mixture-of-Gaussian 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. The speakers that had been used for adaptation were removed from the test set. The features were Mel-Frequency Cepstral Coefficients with deltas and accelerations, which were postprocessed with cepstral mean subtraction (CMS) for further reduction of convolutive effects.

12.4.6 Recognition of Uncertain Data

In the following experiments, the clean cepstrum domain speech features $s_k(c,\tau)$ are assumed to be unavailable, with only an estimate $\hat{s}_k(c,\tau)$ and its associated uncertainty or variance $\sigma^2(c,\tau)$ as the available data according to Fig. 12.1.

In the recognition tests, we compare three strategies to deal with this uncertainty.

- All estimated features $\hat{s}_k(c,\tau)$ are treated as reliable observations and recognized by conventional likelihood evaluation. This is labeled by *no Uncertainty* in all following tables.

- Uncertainty decoding is used, as described in [16]. This will be labeled by *Uncertainty Decoding (UD)*.
- Modified imputation according to [41] is employed, which will be denoted by *Modified Imputation (MI)*.

The implementation that was used for the experiments with both considered uncertainty-of-observation techniques is also described in more detail in Chapter 13 of this book.

12.4.7 Recognition Results

Table 12.4 shows the baseline result, attained after some adaptation to the reverberant room environment, as well as the word error rate on the noisy mixtures and on the ICA output signals. Here, the word error rate is computed via

$$\mathrm{WER} = 100\frac{D+S+I}{N}, \tag{12.43}$$

with D as the number of deletions, S as the substitutions, I as the insertions, and N as the number of reference output tokens, and error rates are computed over all 5 scenarios.

Table 12.4: Word error rate (WER) for reverberant data, noisy mixtures and ICA results.

	reverberant data	mixtures	ICA output
WER	9.28	72.55	26.34

The recognition results in Table 12.5 and 12.7 are achieved with true squared errors used as uncertainties. As it can be seen here, all considered masks lead to a greatly reduced average word error rate under these conditions. However, since only uncertainties estimated from the actual signal should be considered, Tables 12.6 and 12.8 show the error rate reductions that can easily be attained in practice by setting the uncertainty to a realistically available estimate as described in Section 12.3.2.

In each of the tables, the numbers in parentheses give the word error rate reduction, relative to that of the ICA outputs, which are achieved by including time-frequency masking with observation uncertainties.

It is visible from the results that modified imputation clearly tends to give the best results for true uncertainties, whereas uncertainty decoding is the superior strategy for the estimated uncertainty that was tested here. This is indicative of a high sensitivity of modified imputation to uncertainty estimation errors.

However, since a good uncertainty estimation is vital in any case for optimal performance of uncertain feature recognition, it will be interesting to further compare the performance of both uncertainty-of-observation techniques in conjunction with

Table 12.5: Word error rate (WER) for true squared error used as uncertainty. The relative error rate reduction in percent is given in parentheses.

	none	2-stage	Phase	Amplitude	Interference
Mixtures	72.55	n.a.	n.a.	n.a.	n.a.
ICA, no Uncertainty	26.34	55.53	91.79	29.91	96.92
Modified Imputation	n.a.	**11.48 (56.4)**	13.84 (47.5)	16.42 (37.7)	16.80 (36.2)
Uncertainty Decoding	n.a.	12.35 (53.1)	18.59 (29.4)	16.50 (37.4)	22.20 (15.7)

Table 12.6: Word error rate (WER) for estimated uncertainty. The relative error rate reduction in percent is given in parentheses.

	none	2-stage	Phase	Amplitude	Interference
ICA, no Uncertainty	26.34	55.53	91.79	29.91	96.92
Modified Imputation	n.a.	24.78 (5.9)	20.87 (20.8)	26.53 (-0.01)	23.22 (11.9)
Uncertainty Decoding	n.a.	20.79 (21.1)	**19.95 (24.3)**	23.41 (11.1)	21.55 (18.2)

more precise uncertainty estimation techniques, which are an important target for future work.

As for mask performance, the lowest word error rate with estimated uncertainty values is quite clearly achieved for the phase masking strategy. This corresponds well to the high SDR that has been achieved with this strategy in Table 12.3.

On the other hand, the lowest word error rate for ideal uncertainties is almost always reached using the two-stage mask. Again, it is possible to draw a conclusion from comparing with Table 12.3, which now shows that best performance is apparently possible when the largest interference suppression is reached, i.e. when the ΔSIR takes on its largest values.

Also, a more detailed analysis of results is provided in Tables 12.7 and 12.8, which each show the word error rates separately for each mixture recording. Here, it can be seen how the quality of source separation influences the overall performance gained from the suggested approach. For the lower quality separation observable in mixtures #1 and #5, the relative performance gains are clearly lower than average, especially for estimated uncertainties. In contrast, the mean performance improvement for the best separated mixtures #2 and #4 is 36.9% for estimated uncertainties with uncertainty decoding and phase masking, and 61.6% for true squared errors with uncertainty decoding and the 2-stage mask.

A special case is presented by mixture #3. Here, the separation results are comparatively good, however, due to the large microphone distance of 3m, the recognition performance is not ideal. Thus, the rather small performance improvement for estimated uncertainties in this case can also be understood from the fact that much of the recognition error is likely due to mismatched conditions, and hence cannot be expected to be overly influenced by uncertainties derived only from the value of the time-frequency mask.

Table 12.7: Detailed results (WER) for true squared error used as uncertainty.

Algorithm	none	Amplitude	Phase	Interference	2-stage
Mix. 1					
no Unc.	31.54	34.65	92.12	96.89	53.73
MI		21.16	**13.90**	18.26	15.77
UD		22.20	20.54	22.61	17.84
Mix. 2					
no Unc.	18.38	18.86	94.45	96.83	48.65
MI		11.57	9.35	11.25	8.87
UD		11.89	10.78	15.21	**7.45**
Mix. 3					
no Unc.	29.93	33.67	88.03	96.76	52.62
MI		18.45	17.71	21.45	**12.22**
UD		16.21	23.94	25.94	14.71
Mix. 4					
no Unc.	14.73	22.86	90.77	97.14	55.60
MI		10.99	9.01	9.45	6.37
UD		9.89	10.33	13.85	**5.27**
Mix. 5					
no Unc.	35.95	39.58	91.99	96.98	65.11
MI		20.09	19.03	23.26	**13.90**
UD		21.45	27.04	32.02	16.47

12.5 Conclusions

We have discussed the application of ICA for the recognition of multiple, overlapping speech signals, which have been recorded with distant talking microphones in noisy and mildly reverberant environments. Independent component analysis can segregate these multiple sources also in such realistic environments, and can thus lead to significantly improved robustness of automatic speech recognition.

In order to gain more performance even when the ICA outputs still contain residual interferences, the use of time-frequency masking has proved beneficial. However, it improves results only in conjunction with uncertainty-of-observation techniques, in which case, a further 24% relative reduction of word error rate has been shown possible, on average, for datasets with 2 or 3 simultaneously active speakers.

Even greater error rate reductions, about 39% when averaged over all tested time-frequency masks and decoding strategies, have been achieved for ideal uncertainties, i.e. when the true squared estimation error is utilized as the feature uncertainty. This indicates the need for further work on reliable uncertainty estimation as a step for greater robustness with respect to highly instationary noise and interferences. This should also include an automatized parameter adaptation for the uncertainty compensation, e.g. via an EM-style unsupervised adaptation.

Another important target for further work is the source separation itself. As the presented experimental results have shown, overall performance both of time-frequency masking and of subsequent uncertain recognition depends strongly on

Table 12.8: Detailed results (WER) for estimated uncertainty.

Algorithm	none	Amplitude	Phase	Interference	2-stage
Mix. 1					
no Unc.	31.54	34.65	92.12	96.89	53.73
MI		32.99	24.90	27.80	29.25
UD		27.59	**23.86**	24.48	24.07
Mix. 2					
no Unc.	18.38	18.86	94.45	96.83	48.65
MI		19.81	14.58	16.32	18.70
UD		16.16	**11.89**	13.79	12.36
Mix. 3					
no Unc.	29.93	33.67	88.03	96.76	52.62
MI		32.92	**24.19**	32.67	35.91
UD		27.68	26.18	27.93	27.68
Mix. 4					
no Unc.	14.73	22.86	90.77	97.14	55.60
MI		15.16	11.21	11.21	11.87
UD		12.75	**9.01**	10.55	9.45
Mix. 5					
no Unc.	35.95	39.58	91.99	96.98	65.11
MI		32.18	**28.55**	29.00	29.46
UD		32.02	**28.55**	30.51	30.06

the quality of preliminary source separation by ICA. Thus, more successful source separation in strongly reverberant environments would be of great significance for attaining the best overall results from the suggested approach.

References

1. A. Hyvärinen, J. Karhunen, E. Oja, “Independent Component Analysis,” New York: John Wiley, 2001.
2. A. Mansour and M. Kawamoto, “ICA papers classified according to their applications and performances,” in *IEICA Trans. Fundamentals*, vol. E86-A, No. 3, pp. 620-633, March 2003.
3. M. S. Pedersen, J. Larsen, U. Kjems and L. C. Parra, “Convolutive blind source separation methods”, in *Springer Handbook of Speech Processing and Speech Communication*,pp. 1065-1094, Springer Verlag Berlin Heidelberg, 2008.
4. J. Anemüller and B. Kollmeier, “Amplitude modulation decorrelation for convolutive blind source separation”, in *Proc. ICA 2000*, Helsinki, pp. 215-220, 2000.
5. 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”, in *IEEE Trans. Speech and Audio Processing*, vol. 13, no. 3, pp. 412-421, May 2005.
6. D. Kolossa, R. F. Astudillo, E. Hoffmann and R. Orglmeister, “Independent Component Analysis and Time-Frequency Masking for Speech Recognition in Multitalker Conditions”, in *EURASIP J. on Audio, Speech, and Music Processing*, vol. 2010, Article ID 651420, 2010.

7. 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 *Proc. Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 82–85, Oct. 2005.
8. K. Kumatani, J. McDonough, D. Klakow, P. Garner, and W. Li, "Adaptive beamforming with a maximum negentropy criterion," in *Proc. HSCMA*, 2008.
9. O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Processing*, vol. 52, no. 7, pp. 1830–1847, July 2004.
10. M. Kühne, R. Togneri, and S. Nordholm, "Time-frequency masking: Linking blind source separation and robust speech recognition," in *Speech Recognition, Technologies and Applications*. I-Tech, 2008.
11. G. Brown and M. Cooke, "Computational auditory scene analysis," *Computer Speech and Language*, vol. 8, pp. 297–336, 1994.
12. J. B. Allen and L. R. Rabiner, "A unified approach to short-time Fourier analysis and synthesis," Proc. IEEE, vol. 65, pp. 1558-1564, Nov. 1977.
13. J.-F. Cardoso and A. Souloumiac, "Blind beamforming for non-Gaussian signals," *Radar and Signal Processing, IEEE Proceedings*, F 140(6), pp. 362370, Dec. 1993.
14. A. Belouchrani, K. Abed Meraim, J.-F. Cardoso and E. Moulines, "A blind source separation technique based on second order statistics," in *EEE Trans. on Signal Processing*, vol. 45(2), pp. 434-444, 1997.
15. A. Bell and T. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," in *Neural Computation,*, vol. 7, pp. 1129-1159, 1995.
16. L. Deng and J. Droppo and A. Acero, "Dynamic compensation of HMM variances using the feature enhancement uncertainty computed from a parametric model of speech distortion," in *IEEE Trans. Speech and Audio Processing*, vol. 13, pp. 412-421, 2005.
17. A. Hyvärinen and E. Oja. A Fast Fixed-Point Algorithm for Independent Component Analysis. in *Neural Computation*, vol. 9, pp. 1483-1492, 1997.
18. T. Kristjansson and B. Frey. Accounting for Uncertainty in Observations: A new Paradigm for Robust Automatic Speech Recognition, in *Proc. ICASSP*, 2002.
19. C. Mejuto, A. Dapena and L. Castedo, "Frequency-domain infomax for blind separation of convolutive mixtures", in *Proc. ICA 2000*, pp. 315-320, Helsinki, 2000.
20. N. Murata, S. Ikeda, and A. Ziehe, "An approach to blind source separation based on temporal structure of speech signals," *Neurocomputing*, vol. 41, no. 1-4, pp. 124, Oct. 2001.
21. L. Parra, C. Spence and B. De Vries, "Convolutive blind source separation based on multiple decorrelation." in *Proc. IEEE NNSP workshop*, pp. 23-32, Cambridge, UK, 1998.
22. K. Kamata, X. Hu, and H. Kobatake, "A new approach to the permutation problem in frequency domain blind source separation," in *Proc. ICA 2004*, pp. 849-856, Granada, Spain, September 2004.
23. D.-T. Pham, C. Servière, and H. Boumaraf, "Blind separation of speech mixtures based on nonstationarity" in *IEEE Signal Processing and Its Applications*, Proceedings of the Seventh International Symposium, pp. 73-76, 2003.
24. W. Baumann, D. Kolossa and R. Orglmeister, "Maximum likelihood permutation correction for convolutive source separation," in *ICA 2003*, pp. 373-378, 2003.
25. S. Kurita, H. Saruwatari, S. Kajita, K. Takeda, and F. Itakura, "Evaluation of frequency-domain blind signal separation using directivity pattern under reverberant conditions," in *ICASSP2000*, pp. 3140-3143, 2000.
26. M. Ikram and D. Morgan, "A beamforming approach to permutation alignment for multichannel frequency-domain blind speech separation," in *ICASSP02*, pp. 881-884, 2002.
27. N. Mitianoudis and M. Davies, "Permutation alignment for frequency domain ICA Using Subspace Beamforming Methods", in *Proc. ICA 2004*, LNCS 3195, pp. 669-676, 2004.
28. H. Sawada, R. Mukai, S. Araki, S. Makino, "A robust approach to the permutation problem of frequency-domain blind source separation," in *Proc. ICASSP*, vol. V, pp. 381-384, Apr. 2003.
29. V. Stouten and 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 *Proc. ICASSP*, vol. 1, May 2006.

30. D.-T. Pham, C. Servière, and H. Boumaraf, "Blind separation of convolutive audio mixtures using nonstationarity," in *Proc. ICA2003*, pp. 981-986, 2003.
31. P. Sudhakar, and R. Gribonval, "A Sparsity-Based Method to Solve Permutation Indeterminacy in Frequency-Domain Convolutive Blind Source Separation," in *Independent Component Analysis ans Signal Separation: 8th International Conference, ICA 2009, Proceedings*, Paraty, Brazil, March 2009.
32. M. Van Segbroeck and H. Van Hamme, "Robust Speech Recognition using Missing Data Techniques in the Prospect Domain and Fuzzy Masks," in *Proc. ICASSP*, pp. 4393–4396, 2008.
33. W. Baumann, and B.-U. Khler, and D. Kolossa, and R. Orglmeister, "Real Time Separation of Convolutive Mixtures." in: *Independent Component Analysis and Blind Signal Separation: 4th International Symposium, ICA 2001, Proceedings*, San Diego, USA, 2001.
34. F. Asano, S. Ikeda, M. Ogawa, H. Asoh, and N. Kitawaki, "Combined approach of array processing and independent component analysis for blind separation of acoustic signals," in *IEEE Trans. Speech Audio Proc.*, vol. 11, no. 3, pp. 204-215, May 2003.
35. H. Sawada, S. Araki, R. Mukai and S. Makino, "Blind extraction of a dominant source from mixtures of many sources using ICA and time-frequency masking," in *ISCAS 2005*, pp. 5882-5885, May 2005.
36. N. Mitianoudis, and M. E. Davies, "Audio Source Separation of Convolutive Mixtures." in: *IEEE Transactions on Audio and Speech Processing*, vol 11(5), pp. 489497, 2003.
37. D. Kolossa and R. Orglmeister, "Nonlinear Post-Processing for Blind Speech separation," in *Proc. ICA (LNCS 3195)*, Sep. 2004, pp. 832839.
38. Y. Ephraim and D. Malah, "Speech Enhancement using a Minimum Mean Square Error Log-Spectral Amplitude Estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, pp. 443-445, Apr. 1985.
39. 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 convolutivemixtures," in *EURASIP Journal on Applied Signal Processing*, vol. 11, p. 1157-1166, 2003.
40. E. Hoffmann, D. Kolossa and R. Orglmeister, "A Batch Algorithm for Blind Source Separation of Acoustic Signals using ICA and Time-Frequency Masking," in *Proc. ICA (LNCS 4666)*, Sep. 2007, pp. 480488.
41. 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)*, pp. 82-85, New Paltz, NY, 2005.
42. E. Hoffmann, D. Kolossa, and R. Orglmeister, "A Soft Masking Strategy based on Multichannel Speech Probability Estimation for Source Separation and Robust Speech Recognition", *In: Proc. WASPAA*, New Paltz, NY, 2007.
43. R. J. McAulay and M. L. Malpass, "Speech Enhancement using a Soft-Decision Noise Suppression Filter," *IEEE Trans. ASSP-28*, pp. 137-145, Apr. 1980.
44. I. Cohen, "On Speech Enhancement under Signal Presence Uncertainty," *International Conference on Acoustic and Speech Signal Processing*, pp. 167-170, May 2001.
45. Y. Ephraim and I. Cohen, "Recent Advancements in Speech Enhancement", *The Electrical Engineering Handbook*, CRC Press, 2006.
46. R. G. Leonard, "A Database for Speaker-Independent Digit Recognition", *Proc. ICASSP 84*, Vol. 3, p. 42.11, 1984.
47. S. Srinivasan and D. Wang, "Transforming Binary Uncertainties for Robust Speech Recognition", in *IEEE Trans. Audio, Speech and Language Processing, IEEE Transactions on Speech and Audio Processing* vol. 15, pp. 2130-2140, 2007.
48. R. F. Astudillo, D. Kolossa, P. Mandelartz and R. Orglmeister, "An Uncertainty Propagation Approach to Robust ASR using the ETSI Advanced Front-End", accepted for publication in *IEEE Journal of Selected Topics in Signal Processing*, 2010.
49. G. Brown and D. Wang, "Separation of Speech by Computational Auditory Scene Analysis", *Speech Enhancement*, eds. J. Benesty, S. Makino and J. Chen, Springer, pp. 371-402, 2005.

50. R. F. Astudillo, D. Kolossa and R. Orglmeister, “Propagation of Statistical Information through non-linear Feature Extractions for Robust Speech Recognition”, in *Proc. MaxEnt*, 2007.
51. S. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, P. Woodland, “The HTK Book (for HTK Version 3.4)”, Cambridge University Engineering Department, 2006.