4

Time-Frequency Masking: Linking Blind Source Separation and Robust Speech Recognition

Marco Kühne[1], Roberto Togneri[1] and Sven Nordholm[2]
[1]*The University of Western Australia*
[2]*Western Australian Telecommunications Research Institute*
[2]*Curtin University of Technology*
Australia

1. Introduction

In order to deploy automatic speech recognition (ASR) effectively in real world scenarios it is necessary to handle hostile environments with multiple speech and noise sources. One classical example is the so-called "cocktail party problem" (Cherry, 1953), where a number of people are talking simultaneously in a room and the ASR task is to recognize the speech content of one or more target speakers amidst other interfering sources. Although the human brain and auditory system can handle this everyday problem with ease it is very hard to solve with computational algorithms. Current state-of-the-art ASR systems are trained on clean single talker speech and therefore inevitably have serious difficulties when confronted with noisy multi-talker environments.

One promising approach for noise robust speech recognition is based on the missing data automatic speech recognition (MD-ASR) paradigm (Cooke et al., 2001). MD-ASR requires a time-frequency (T-F) mask indicating the reliability of each feature component. The classification of a partly corrupted feature vector can then be performed on the reliable parts only, thus effectively ignoring the components dominated by noise. If the decision about the reliability of the spectral components can be made with absolute certainty, missing data systems can achieve recognition performance close to clean conditions even under highly adverse signal-to-noise ratios (SNRs) (Cooke et al., 2001; Raj & Stern, 2005; Wang, 2005).

The most critical part in the missing data framework is the blind estimation of the feature reliability mask for arbitrary noise corruptions. The remarkable robustness of the human auditory system inspired researchers in the field of computational auditory scene analysis (CASA) to attempt auditory-like source separation by using an approach based on human hearing. CASA systems first decompose a given signal mixture into a highly redundant T-F representation consisting of individual sound elements/ atoms. These elementary atoms are subsequently arranged into separate sound streams by applying a number of grouping cues such as proximity in frequency and time, harmonicity or common location (Bregman, 1990; Brown & Cooke, 1994; Wang, 2005). The output of these grouping mechanisms can often be represented as a T-F mask which separates the target from the acoustic background. Essentially, T-F masking provides a link between speech separation and speech recognition (Cooke et al., 2001; Wang, 2005).

Source: Speech Recognition, Technologies and Applications, Book edited by: France Mihelič and Janez Žibert,
ISBN 978-953-7619-29-9, pp. 550, November 2008, I-Tech, Vienna, Austria

Most previous work related to missing data mask estimation is based on single-channel data (see Cerisara et al., (2007) for a review) and relies on SNR criteria (Cooke et al., 2001; Barker et al., 2000; El-Maliki & Drygajlo, 1999), harmonicity cues (Hu & Wang, 2004; van Hamme, 2004) or cue combinations (Seltzer et al., 2004). Alternatively, binaural CASA models (Harding et al., 2006; Roman et al., 2003; Kim & Kil, 2007) exploit interaural time and intensity differences (ITD)/ (IID) between two ears for missing data mask estimation. While used in the CASA community for quite some time, the concept of T-F masking has recently attracted some interest in the in the field of blind signal separation (BSS) (Yilmaz & Rickard, 2004; Araki et al., 2005). Similar to CASA, these methods exploit the potential of T-F masking to separate mixtures with more sources than sensors. However, the BSS problem is tackled from a signal processing oriented rather than psychoacoustic perspective. This, for instance, includes the use of multiple sensor pairs (Araki et al., 2007) and statistical approaches such as Independent Component Analysis (Kolossa et al., 2006; Hyvärinen, 1999).

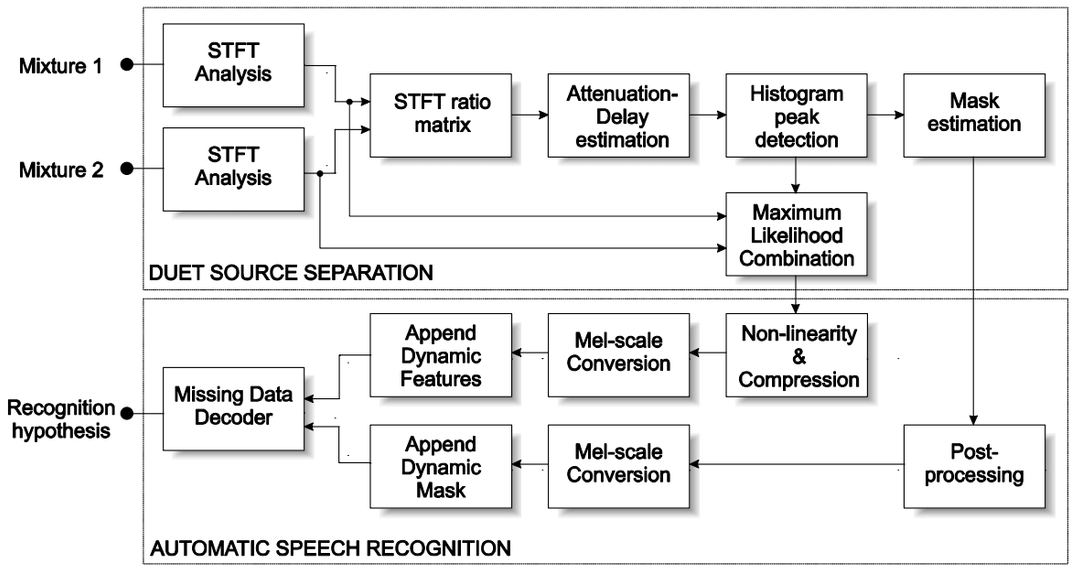


Fig. 1. Flowchart for proposed combination of DUET source separation and missing data speech recognition.

This chapter presents a scheme which combines BSS with robust ASR through the systematic application of T-F masking for both speech separation and speech recognition (Fig. 1). The outlined approach summarizes our previous work reported in Kühne et al. (2007; 2007a). In particular, we investigate the performance of a recently proposed BSS method called DUET (Yilmaz & Rickard, 2004) as front-end for missing data speech recognition. Since DUET relies on T-F masking for source demixing, this combination arises as a natural choice and is straightforward to implement. In Kühne et al. (2007) an approach was presented that avoids DUET's source reconstruction step and directly uses the mask together with the spectral mixture as input for the speech decoder. In subsequent work (Kühne et al., 2007a), a simple but effective mask post-processing step was introduced in order to remove spurious T-F points that can cause insertion errors during decoding. Our proposed combination fits seamlessly into standard feature extraction schemes (Young et al, 2006), but requires a modification of the decoding algorithm to account for missing feature components. It is particularly attractive for ASR scenarios where only limited space and resources for multi-channel processing are available (e.g., mobile phones).

The effectiveness of the proposed BSS-ASR combination is evaluated for a simulated cocktail party situation with multiple speakers. Experimental results are reported for a connected digits recognition task. Our evaluation shows that, when the assumptions made by DUET hold, the estimated feature reliability masks are comparable in terms of speech recognition accuracy to the oracle masks obtained with a prior knowledge of the sources. We further demonstrate that a conventional speech recognizer fails to operate successfully on DUET's resynthesized waveforms, which clearly shows the merit of the proposed approach.
The remainder of this chapter is organized as follows: Section 2 briefly reviews the DUET source separation method and outlines its main assumptions. Section 3 explains the methods used for feature extraction and missing data mask generation in more detail. Section 4 presents the experimental evaluation of the system. Section 5 gives a general discussion and illustrates the differences and similarities with a related binaural CASA segregation model. The section further comments on some of the shortcomings in the proposed approach. Finally, the chapter concludes in Section 6 with an outlook on future research.

2. Source separation

This section presents a short review of the DUET-BSS algorithm used in this study for blind separation of multiple concurrent talkers. We start with an introduction of the BSS problem for anechoic mixtures and highlight the main assumptions made by the DUET algorithm. After briefly outlining the main steps of the algorithm, the section closes with a short discussion on why the reconstructed waveform signals are not directly suitable for conventional speech recognition. For a more detailed review of DUET the reader is referred to Yilmaz & Rickard (2004) and Rickard (2007).

2.1 Anechoic mixing model

The considered scenario uses two microphone signals $x_1(t)$ and $x_2(t)$ to capture $N \geq 2$ speech sources $s_1(t), \ldots, s_N(t)$ assuming the following anechoic mixing model

$$x_m(t) = \sum_{j=1}^{N} a_{mj} s_j(t - \delta_{mj}), \qquad m = 1, 2 \tag{1}$$

where a_{mj} and δ_{mj} are the attenuation and delay parameters of source s_j at microphone x_m. The goal of any BSS algorithm is to recover the source signals $s_j(t), j = 1, \ldots, N$ using only the mixture observations $x_m(t), m = 1, 2$. The mixing model can be approximated in the Short-Time-Fourier-Transform (STFT) domain as an instantaneous mixture at each frequency bin l through

$$X_m(k,l) \approx \sum_{j=1}^{N} a_{mj} e^{-il\omega_0 \delta_{mj}} S_j(k,l). \tag{2}$$

The STFT transform $S(k,l)$ for a time domain signal $s(t)$ is defined as

$$S(k,l) := \sum_{\tau=-T/2}^{T/2-1} w(\tau) s(\tau + k\tau_0) e^{-il\omega_0 \tau}, \tag{3}$$

where τ_0 and ω_0 specify the time-frequency grid resolution and $w(\tau)$ is a window function (e.g., Hamming) of size T which attenuates discontinuities at the frame edges.
The instantaneous BSS problem can be solved quite elegantly in the frequency domain due to the sparsity of time-frequency representations of speech signals. DUET proceeds by considering the following STFT ratio

$$\frac{X_2(k,l)}{X_1(k,l)} = \frac{\sum_{j=1}^{N} a_{2j} e^{-il\omega_0\delta_{2j}} S_j(k,l)}{\sum_{j=1}^{N} a_{1j} e^{-il\omega_0\delta_{1j}} S_j(k,l)}, \tag{4}$$

where the nominator and denominator are weighted sums of complex exponentials representing the delay and attenuation of the source spectra at the two microphones.

2.2 Assumptions

The key assumption in DUET is that speech signals satisfy the so-called W-disjoint orthogonality (W-DO) requirement

$$S_i(k,l)S_j(k,l) = 0, \forall i \neq j, \forall k,l, \tag{5}$$

also known as "sparseness" or "disjointness" condition with the support of a source S_j in the T-F plane being denoted as $\Omega_j := \{(k,l) : S_j(k,l) \neq 0\}$. The sparseness condition (5) implies that the supports of two W-DO sources are disjoint, e. g., $\Omega_i \cap \Omega_j = \emptyset$. This motivates a demixing approach based on time-frequency masks, where the mask for source S_j corresponds to the indicator function for the support of this source:

$$M_j(k,l) = \begin{cases} 1, & \text{if } (k,l) \in \Omega_j \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

It has previously been shown (Wang, 2005; Yilmaz & Rickard, 2004; Roman et al., 2003) that binary time-frequency masks exist that are capable of demixing speech sources from just one mixture with high speech fidelity. For example, Wang (2005) proposed the notion of an ideal/ oracle binary mask

$$O_j(k,l) := \begin{cases} 1, & \text{if } 20\log_{10}\frac{|S_j(k,l)|}{|\sum_{i\neq j} S_i(k,l)|}\text{dB} \geq 0\,\text{dB} \\ 0, & \text{otherwise,} \end{cases} \tag{7}$$

which determines all time-frequency points where the power of the source S_j exceeds or equals the power of the sum of all interfering sources (see Wang (2005) for a more detailed motivation of the ideal binary masks). Note that these masks can only be constructed if the source signals are known prior to the mixing process as they are defined by means of a SNR criterion. Instead, DUET relies on spatial cues extracted from two microphones to estimate the ideal binary mask. It solely depends on relative attenuation and delays of a sensor pair and assumes an anechoic environment where these cues are most effective. An additional assumption requires that the attenuation and delay mixing pairs for each source are unambiguous.

2.3 Estimation of relative mixing parameters using DUET

Due to (5) it follows, that only one arbitrary source S_j will be active at any T-F point such that (4) simplifies to

$$\frac{X_2(k,l)}{X_1(k,l)} = \frac{a_{2j}}{a_{1j}} e^{-il\omega_0(\delta_{2j}-\delta_{1j})} = a_j e^{-il\omega_0\delta_j}, \qquad \forall (k,l) \in \Omega_j \tag{8}$$

with a_j and δ_j denoting relative attenuation and delay parameters between both microphones and $(a_j \neq a_k)$ or $(\delta_j \neq \delta_k)$, $\forall j \neq k$. The goal is now to estimate for each source S_j the corresponding mixing parameter pair (a_j, δ_j) and use this estimate to construct a time-frequency mask that separates S_j from all other sources.

An estimate of the attenuation and delay parameter at each T-F point is obtained by applying the magnitude and phase operator to (8) leading to

$$\tilde{a}(k,l) := \left| \frac{X_2(k,l)}{X_1(k,l)} \right|, \qquad \tilde{\delta}(k,l) := -\frac{1}{l\omega_0} \arg\left(\frac{X_2(k,l)}{X_1(k,l)} \right). \tag{9}$$

If the sources are truly W-DO then accumulating the instantaneous mixing parameter estimates in (9) over all T-F points will yield exactly N distinct $(\tilde{a}, \tilde{\delta})$ pairs equal to the true mixing parameters:

$$\bigcup_{(k,l)} \left\{ (\tilde{a}(k,l), \tilde{\delta}(k,l)) \right\} = \{(a_j, \delta_j) : j = 1 \ldots, N\} \tag{10}$$

The demixing mask for each source is then easily constructed using the following binary decision

$$M_j(k,l) := \begin{cases} 1, & \text{if } \left(\tilde{a}(k,l), \tilde{\delta}(k,l)\right) = (a_j, \delta_j) \\ 0, & \text{otherwise.} \end{cases} \tag{11}$$

However, in practice the W-DO assumption holds only approximately and it will no longer be possible to observe the true mixing parameters directly through inspection of the instantaneous estimates in (9). Nevertheless, one can expect that the values will be scattered around the true mixing parameters in the attenuation-delay parameter space. Indeed, it was shown in Yilmaz & Rickard (2004) that T-F points with high power possess instantaneous attenuation-delay estimates close to the true mixing parameters. The number of sources and their corresponding attenuation-delay mixing parameters are then estimated by locating the peaks in a power weighted $(\tilde{\alpha}, \tilde{\delta})$-histogram (see Fig. 2a), where $\tilde{\alpha}(k,l) := \tilde{a}(k,l) - (\tilde{a}(k,l))^{-1}$ is the so-called symmetric attenuation (Yilmaz & Rickard, 2004). The peak detection was implemented using a weighted k-means algorithm as suggested in Harte et al. (2005).

2.4 Time-Frequency mask construction and demixing

Once the peak locations $(\hat{\alpha}_j, \hat{\delta}_j)$, $j = 1, \ldots, N$ have been determined, a second pass over the raw data set is required to assign each observation to one of the detected source locations. We used simple minimum distance classification to construct the binary T-F mask for source S_j as

$$\hat{M}_j(k,l) := \begin{cases} 1, & \text{if } j = \underset{z}{\operatorname{argmin}}\, d_z^2(k,l) \\ 0, & \text{otherwise,} \end{cases} \tag{12}$$

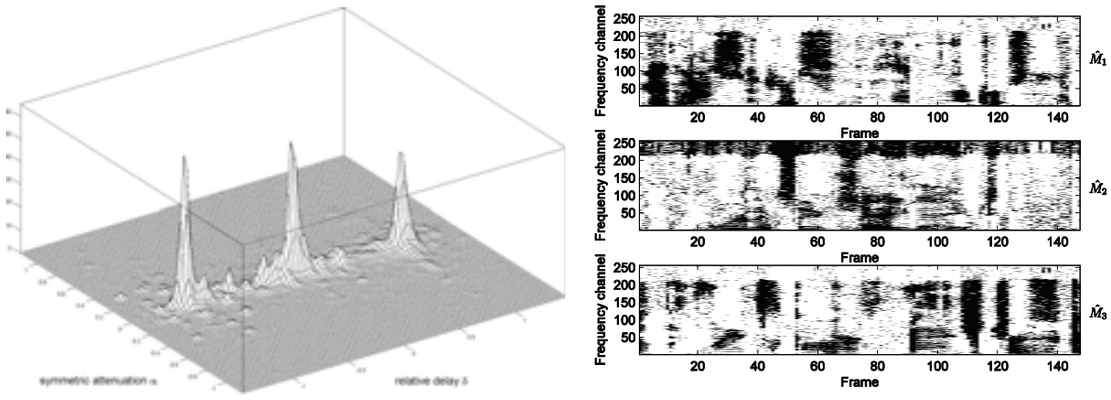
where d_z^2 is the squared Euclidean distance

$$d_z^2(k,l) = \left(\tilde{\alpha}(k,l) - \hat{\alpha}_z\right)^2 + \left(\tilde{\delta}(k,l) - \hat{\delta}_z\right)^2 \tag{13}$$

between the instantaneous mixing parameter estimate $(\tilde{\alpha}(k,l), \tilde{\delta}(k,l))$ and the histogram peak $(\hat{\alpha}_z, \hat{\delta}_z)$. The demixing then proceeds by masking the maximum likelihood combination $X_{\mathrm{ML}}(k,l)$ of both mixtures (Yilmaz & Rickard, 2004) to obtain the source estimate as

$$\hat{S}_j(k,l) = \hat{M}_j(k,l) \underbrace{\left(\frac{X_1(k,l) + \hat{a}_j e^{il\omega_0 \hat{\delta}_j} X_2(k,l)}{1 + \hat{a}_j^2} \right)}_{X_{\mathrm{ML}}(k,l)}. \tag{14}$$

$\hat{S}_j(k,l)$ can then be converted back into the time domain by means of an inverse STFT transformation. Note that for the maximum likelihood combination of both mixtures the symmetric attenuation parameter was converted back to the relative attenuation parameter $\hat{a}_j$. However, here we are interested in evaluating the DUET demixing performance using an automatic speech recognizer. The reconstruced time domain signal $\hat{s}_j(t)$ will not be directly applicable for conventional speech recognition systems because non-linear masking effects due to $\hat{M}_j$ are introduced during waveform resynthesis. Conventional speech recognizers perform decoding on complete spectra and can not deal with partial spectral representations. Therefore, additional processing steps, either in the form of data imputation to reconstruct missing spectrogram parts (Raj & Stern, 2005) or missing data marginalization schemes (Cooke et al., 2001) that can handle partial data during decoding, are required before speech recognition can be attempted.



(a) (b)

Fig. 2. Power weighted attenuation-delay histogram (a) for a mixture of three sources with mixing parameters $\{(\alpha_1;\delta_1),(\alpha_2;\delta_2),(\alpha_3;\delta_3)\} = \{(-0.03;0.94),(0;0),(0.03;-0.94)\}$ and (b) the estimated time-frequency masks with selected points marked in black.

In this work the latter option was chosen allowing us to avoid source reconstrucion and directly exploit the spectrographic masks for missing data decoding. After source separation the missing data recognizer was informed which mask corresponded to the target speaker by comparing the detected histogram peaks with the true mixing parameters. However, the high STFT resolution is usually not suitable for statistical pattern recognition as it would

lead to very high-dimensional feature vectors. The following section explains how the results of the DUET separation can be integrated into standard feature extraction schemes and be utilized for missing data speech recognition.

3. Automatic speech recognition with missing data

A Hidden Markov Model (HMM) based missing data speech recognizer (Cooke et al., 2001) was used for all speech recognition experiments reported in this study. While the HMMs are trained on clean speech in exactly the same manner as in conventional ASR the decoding is treated differently in missing data recognition. Additionally to the feature vector sequence a mask is required to declare each feature component as reliable or unreliable using a hard or soft decision (Barker et al., 2000; Morris et al., 2001).
This section starts with a detailed description of the extracted acoustic features and how the DUET masks can be utilized for missing data recognition. A mask post-processing step is introduced in order to remove isolated mask points that can cause insertion errors in the speech decoding process. We then proceed with the missing data decoding and explain how observation likelihoods are computed in the presence of missing feature components.

3.1 Feature extraction

It is known that the human ear resolves frequencies by grouping several adjacent frequency channels into so-called critical bands (Moore, 2003). For speech recognition purposes the linear STFT frequency resolution is usually converted to a perceptual frequency scale, such as the bark or mel scale (Moore, 2003; Young et al., 2006). A widely used approximation of the non-linear frequency resolution of the human auditory system is the mel-frequency scale

$$\mathfrak{f}(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right), \tag{15}$$

where f denotes the linear frequency in Hz and $\mathfrak{f}$ is the corresponding non-linear frequency scale in mel. The grouping of individual frequency channels into critical bands can be accomplished by applying a triangular mel-filterbank to the magnitude or power FFT spectrum (Young et al., 2006). The triangular filters

$$\lambda_b(l) = \begin{cases} 0 & l\omega_0 < \omega_{c_{(b-1)}}, \\ \frac{l\omega_0 - \omega_{c_{(b-1)}}}{\omega_{c_b} - \omega_{c_{(b-1)}}} & \omega_{c_{(b-1)}} \leq l\omega_0 \leq \omega_{c_b}, \\ \frac{\omega_{c_{(b+1)}} - l\omega_0}{\omega_{c_{(b+1)}} - \omega_{c_b}} & \omega_{c_b} \leq l\omega_0 \leq \omega_{c_{(b+1)}}, \\ 0 & l\omega_0 > \omega_{c_{(b+1)}}, \end{cases} \tag{16}$$

with

$$\omega_{c_b} = 2\pi \cdot 700 \left(10^{\mathfrak{f}_{c_b}/2595} - 1\right) \tag{17}$$

are equally spaced along the mel-frequency scale through

$$\mathfrak{f}_{c_b} = \mathfrak{f}_l + b \cdot \frac{\mathfrak{f}_h - \mathfrak{f}_l}{B+1}, \qquad b = 1, \ldots, B. \tag{18}$$

Here B is the number of mel-frequency channels and f_l, f_h are the lower and higher cut-offs of the mel-frequency axis.

(A) **Acoustic feature extraction:** The preferred acoustic features employed in missing data speech recognition are based on spectral representations rather than the more common mel-frequency-cepstral-coefficients (MFCCs). This is due to the fact that a spectrographic mask contains localized information about the reliability of each spectral component, a concept not compatible with orthogonalized features, such as cepstral coefficients (see also de Veth et al. (2001) for a further discussion). For the scope of this study the extracted spectral features for missing data recognition followed the FBANK feature implementation of the widely accepted Hidden Markov Model Toolkit (Young et al., 2006).

Let $\boldsymbol{o}_k = (o_{k1}, \ldots, o_{kn})^T$ be the n-dimensional spectral feature vector at time frame k. The static log-spectral feature components (see Fig. 3.b) are computed as

$$o_{kb} = \log\left(\max\left\{\sum_l \lambda_b(l)|X_{\mathrm{ML}}(k,l)|, 1\right\}\right), \quad b = 1, \ldots, B, \tag{19}$$

where λ_b are the triangular mel-filterbank weights defined in (16) and X_{ML} is the maximum likelihood combination of both mixture observations as specified in (14). It is common to append time derivatives to the static coefficients in order to model their evolution over a short time period. These dynamic parameters were determined here via the standard regression formula

$$\Delta o_{kb} := o_{k(\frac{n}{2}+b)} = \frac{\sum_{\theta=1}^{\Theta} \theta\left(o_{(k+\theta)b} - o_{(k-\theta)b}\right)}{2\sum_{\theta=1}^{\Theta} \theta^2}, \quad b = 1, \ldots, B, \tag{20}$$

where Δo_{kb} is the regression coefficient at time frame k and mel-frequency subband b, computed over the corresponding static features using a temporal integration window of size Θ (Young et al., 2006). For this study, only first-order regression coefficients were used, thus producing a feature vector of dimension $n = 2B$.

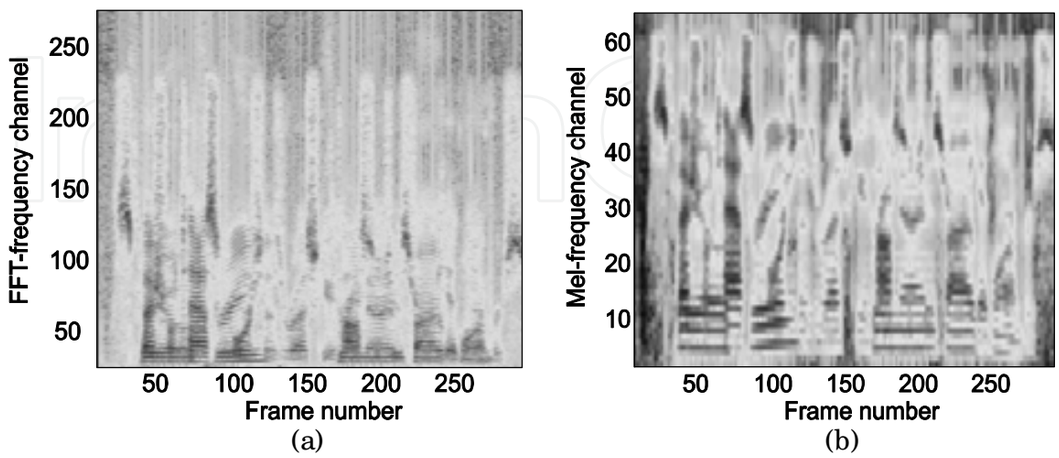


Fig. 3. Spectrograms for the TIDIGITS utterance "3o33951" mixed with three interfering speakers in anechoic condition. (a) linear FFT frequency scale; (b) non-linear mel-frequency scale

(B) Missing data reliability masks: The reliability of each feature component is indicated by a corresponding missing feature mask provided here by the source separation stage. Before converting the mask to the mel-frequency scale we introduce a mask post-processing step to eliminate spurious points in the mask. One important aspect that has not been considered so far is the high correlation of neighboring time-frequency points. That is, if a time-frequency point (k,l) is assigned to speaker S_j then it is very likely that points in the neighborhood of (k,l) are also belonging to S_j (see Fig. 5a).

The DUET method solely relies on the mask assignment in the attenuation-delay parameter space and does not take neighborhood information into account. We observed that for mixtures with more than two sources the time-frequency masks were overlaid by some scattered isolated "noise" points (compare Fig. 5a,c). This type of noise is similar to "shot-noise" known in the image processing community and can be dealt with effectively by means of a non-linear median filter (Russ, 1999). Similar smoothing techniques have been used previously for missing data mask post-processing (Harding et al., 2005). For this study, the median filter was preferred over other linear filters as it preserves the edges in the mask while removing outliers and smoothing the homogenous regions. The basic operation of a two-dimensional median filter consists in sorting the mask values of a T-F point (k,l) and its neighborhood and replacing the mask value with the computed median $\bar{M}(k,l)$. Several different neighborhood patterns exist in the literature ranging from 4-nearest neighbors over 3×3 or 5×5 square neighborhoods to octagonal regions (Russ, 1999). Here, we used a 5×5 plus sign-shaped median filter

$$\bar{M}(k,l) = \text{median}\left\{M(u,v) : (u,v) \in N_{(k,l)}\right\} \tag{21}$$

with the neighborhood pattern (Fig. 4) defined as

$$N_{(k,l)} := \{(n,m) : \max\{|n-k|, |m-l|\} \leq 2 \wedge \min\{|n-k|, |m-l|\} = 0\}\,.$$

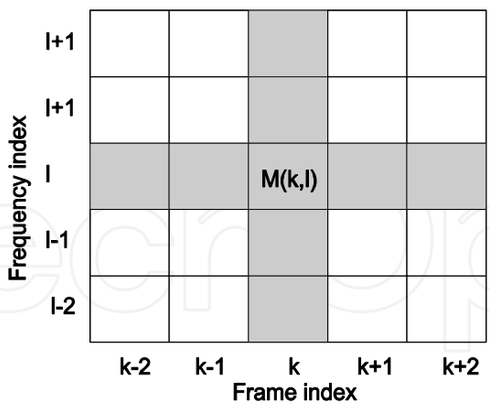


Fig. 4. Plus signed-shaped neighborhood pattern of size 5×5 used for the proposed two-dimensional median filtering of the DUET localization masks.

The filter is able to preserve vertical or horizontal lines that would otherwise be deleted by square neighborhoods. This is important in our application as these lines are often found at sound onsets (vertical, constant time) or formant frequency ridges (horizontal, constant frequency). Other more sophisticated rank filters like the hybrid median filter or cascaded median filters have not been considered here but can be found in Russ (1999). The effect of

the median filtering can be observed in Fig. 5e, where most of the isolated points have been removed while still preserving the main characteristics of the oracle mask (Fig. 5a).

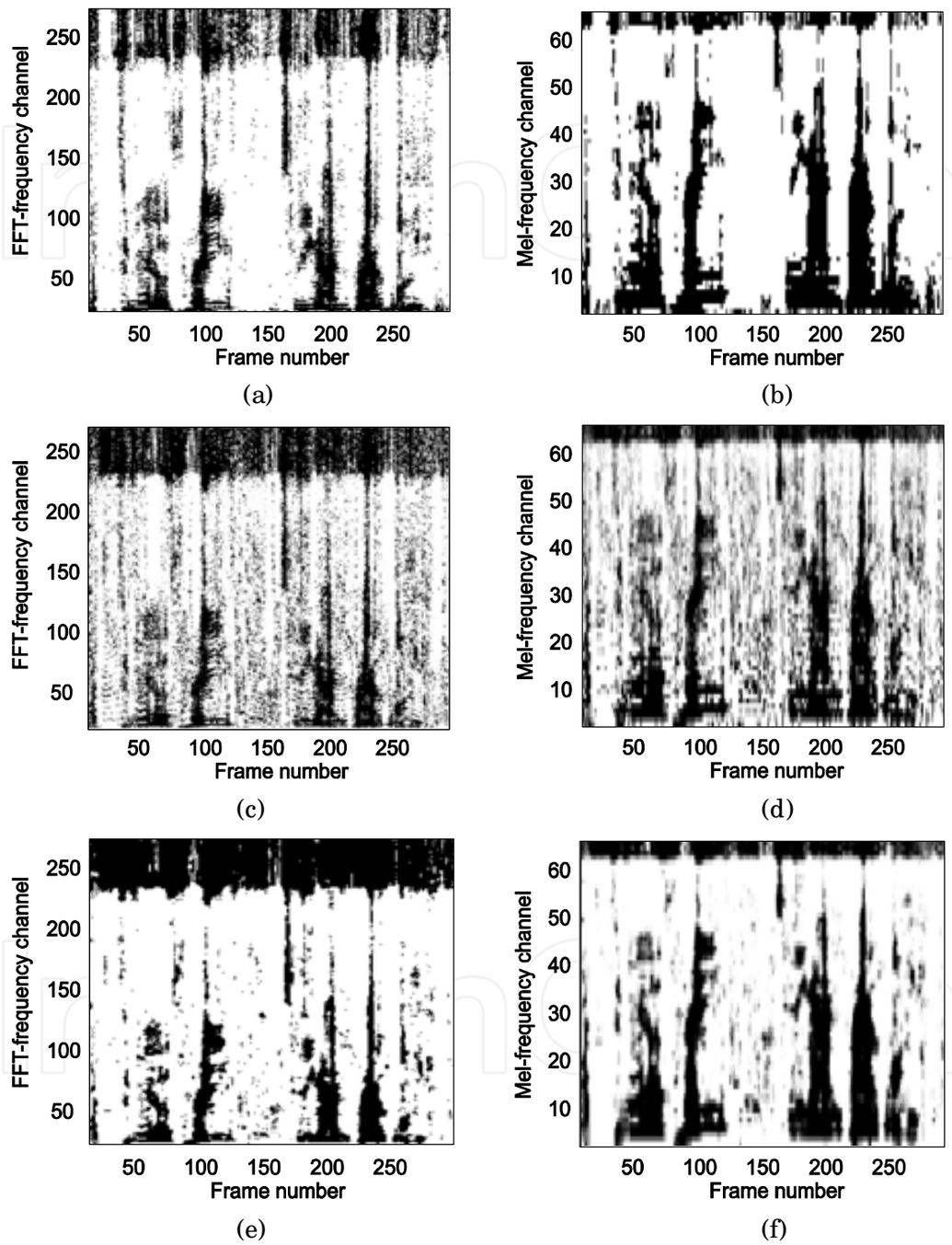


Fig.5. Example of localization masks for the TIDIGITS target source (black) "3o33951" in a mixture of three competing speakers (white). (a) oracle mask on linear FFT frequency scale; (b) oracle mask on non-linear mel-frequency scale; (c) DUET mask on linear FFT frequency scale; (d) DUET mask converted to non-linear mel-frequency scale; (e) median filtered mask of (c); (f) median filtered DUET mask from (e) converted to non-linear mel-frequency scale

The final missing data mask is then obtained by converting the high STFT resolution to the mel-frequency domain. Similar to (19), we apply the triangular mel-weighting function λ_b to obtain a soft mel-frequency mask

$$w_{kb} = \frac{\sum_l \lambda_b(l)\bar{M}(k,l)}{\sum_l \lambda_b(l)}. \tag{22}$$

While the mask (22) is valid for static features only a reliability mask is also required for the dynamic feature coefficients in (20). The corresponding mask for Δo_{kb} was determined based on the static mask values as

$$\Delta w_{kb} := w_{k(\frac{n}{2}+b)} = \prod_{\substack{\theta=-\Theta,\\ \theta\neq 0}}^{\Theta} w_{(k+\theta)b}. \tag{23}$$

3.2 HMM observation likelihoods with missing features

In this study a HMM based missing data recognizer was used for scoring the n-dimensional spectro-temporal feature vectors described in Section 3.1. The HMM state output distributions were modeled via Gaussian mixture models (GMMs) with diagonal covariance matrices. Let the GMM model parameters for a particular HMM state q be denoted as $\Lambda_q = \{\boldsymbol{c}_q, \boldsymbol{\mu}_q, \boldsymbol{\sigma}_q^2\}$, where the three components represent the mixture weights, mean and variance vectors of the Gaussian mixture probability density function. For a GMM with R mixtures the emission likelihood of $\boldsymbol{o}_k$ for HMM state q is given by

$$p(\boldsymbol{o}_k|\Lambda_q) = \sum_{r=1}^{R} c_{qr} \prod_{i=1}^{n} p\left(o_{ki}|\mu_{qri}, \sigma_{qri}^2\right), \tag{24}$$

where in the case of missing or uncertain features $p\left(o_{ki}|\mu_{qri}, \sigma_{qri}^2\right)$ is evaluated as

$$p\left(o_{ki}|\mu_{qri}, \sigma_{qri}^2\right) = w_{ki}\mathcal{N}\left(o_{ki}; \mu_{qri}, \sigma_{qri}^2\right) + (1-w_{ki})\frac{1}{b_{ki}-a_{ki}}\int_{a_{ki}}^{b_{ki}} \mathcal{N}\left(\tilde{o}_{ki}; \mu_{qri}, \sigma_{qri}^2\right) d\tilde{o}_{ki}, \tag{25}$$

with w_{ki} denoting the value of the missing data mask at T-F point (k,i), a_{ki} and b_{ki} being the lower and upper integration bound and $\mathcal{N}(o_{ki}; \mu_{qri}, \sigma_{qri}^2)$ being a univariate Gaussian

$$\mathcal{N}\left(o_{ki}; \mu_{qri}, \sigma_{qri}^2\right) = \frac{1}{\sqrt{2\pi\sigma_{qri}^2}} \exp\left[-\frac{1}{2}\frac{(o_{ki}-\mu_{qri})^2}{\sigma_{qri}^2}\right], \tag{26}$$

with mean μ_{qri} and variance σ_{qri}^2. The value of the missing data mask w_{ki} weights the present and missing data contributions with a soft "probability" between 0 and 1 (Harding et al., 2006; Barker et al., 2000). The likelihood contribution in (25) for the missing static features is evaluated as a bounded integral over the clean static feature probability density by exploiting the knowledge that the true clean speech value is confined to the interval between zero and the observed noisy spectral energy, e.g. $o_{ki}^{\text{clean}} \in [a_{ki}, b_{ki}] = [0, o_{ki}], \forall i = 1, \ldots, \frac{n}{2}$. Past research (Cooke et al., 2001; Morris et al., 2001) has shown that bounding the integral in (25) is beneficial as it provides an effective

mechanism to incorporate counter-evidence by penalizing models with insufficient spectral energy. However, no bounds on dynamic feature components were utilized here, thus $a_{ki} \rightarrow -\infty$ and $b_{ki} \rightarrow \infty, \forall i = \frac{n}{2} + 1, \ldots, n$.

4. Experimental evaluation

4.1 Setup

(A) Recognizer architecture and HMM model training: The proposed system was evaluated via connected digit experiments on the TIDIGITS database (Leonard, 1984) with a sample frequency of $20\,\mathrm{kHz}$. The training set for the recognizer consisted of 4235 utterances spoken by 55 male speakers. The HTK toolkit (Young et al., 2006) was used to train 11 word HMMs ('1'-'9','oh','zero') each with eight emitting states and two silence models ('sil','sp') with three and one state. All HMMs followed standard left-to-right models without skips using continuous Gaussian densities with diagonal covariance matrices and $R = 10$ mixture components. Two different sets of acoustic models were created. Both used $25\,\mathrm{ms}$ Hamming-windows with $10\,\mathrm{ms}$ frame shifts for the STFT analysis. Note that Yilmaz & Rickard (2004) recommend a Hamming window size of 64 ms for a sampling frequency of 16 kHz in order to maximize the W-DO measure for speech signals. However, for the ASR application considered here, the chosen settings are commonly accepted for feature extraction purposes. The first set of HMMs was used as the cepstral baseline system with 13 MFCCs derived from a $B = 32$ channel HTK mel-filterbank plus delta and acceleration coefficients ($\Theta = 2$) and cepstral mean normalization. This kind of baseline has been widely used in missing data ASR evaluations (Cooke et al., 2001; Morris et al., 2001; Harding et al., 2006). The second model set was used for the missing data recognizer and used spectral rather than cepstral features as described in Section 3.1. In particular, acoustic features were extracted from a HTK mel-filterbank with $B = 64$ channels and first order delta coefficients ($\Theta = 2$) were appended to the static features according to (19) and(20).

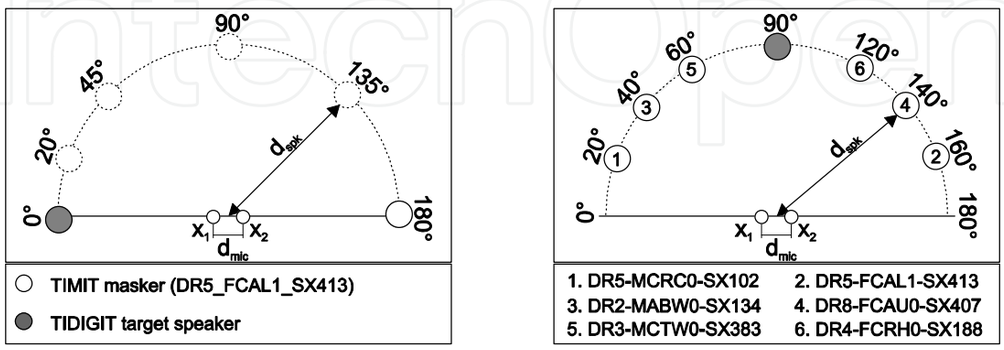
(B) Test data set and room layout: The test set consisted of 166 utterances of seven male speakers containing at least four digits mixed with several masking utterances taken from the TIMIT database (Garofolo et al., 1993; see Table 1).

TIMIT ID code			Utterance transcription
Dialect	Speaker	Sentence	
DR5	MCRC0	SX102	"Special task forces rescue hostages from kidnappers."
DR5	FCAL1	SX413	"Daphne's Swedish needlepoint scarf matched her skirt."
DR2	MABW0	SX134	"December and January are nice months to spend in Miami."
DR8	FCAU0	SX407	"Laugh, dance, and sing if fortune smiles upon you."
DR3	MCTW0	SX383	"The carpet cleaners shampooed our oriental rug."
DR4	FCRH0	SX188	"Who authorized the unlimited expense account?"

Table 1. Transcription for six utterances taken from the test section of the TIMIT database.

The signal-to-interferer ratio (SIR) for each masker was approximately $0\,\mathrm{dB}$. Stereo mixtures were created by using an anechoic room impulse response of a simulated room of size

$4\,\mathrm{m} \times 6\,\mathrm{m} \times 3\,\mathrm{m}$ (length x width x height). Two microphones were positioned in the center of the room, $2\,\mathrm{m}$ above the ground, with an interelement distance of $d_{\mathrm{mic}} = 1.72\,\mathrm{cm}$ to guarantee accurate phase parameter estimates (Yilmaz & Rickard, 2004). Fig. 6a shows the setup for a single masker scenario and Fig. 6b for a multi-speaker scenario with up to six different speech maskers (three male, three female) placed at a distance of $d_{\mathrm{spk}} = 1\,\mathrm{m}$ to the microphones. For testing, the HTK decoder (HVite) was modified according to (25) to incorporate the missing data marginalization framework.



(a) single masker (b) multiple masker

Fig. 6. Source configurations for the TIDIGITS target speaker and (a) a single TIMIT masker placed at different angles and (b) for corruption with multiple TIMIT maskers.

(C) Performance measures: The following standard performance measures were computed based on the decoder output (Young et al., 2006). The percentage correctness score is defined as

$$\mathrm{COR} = \frac{\mathrm{NUM} - \mathrm{DEL} - \mathrm{SUB}}{\mathrm{NUM}} \times 100\%, \tag{27}$$

where NUM is the total number of digits in the test set and DEL and SUB denote the deletion and substitution errors, respectively. The second performance measure, the percent accuracy is defined as

$$\mathrm{ACC} = \frac{\mathrm{NUM} - \mathrm{DEL} - \mathrm{SUB} - \mathrm{INS}}{\mathrm{NUM}} \times 100\% \tag{28}$$

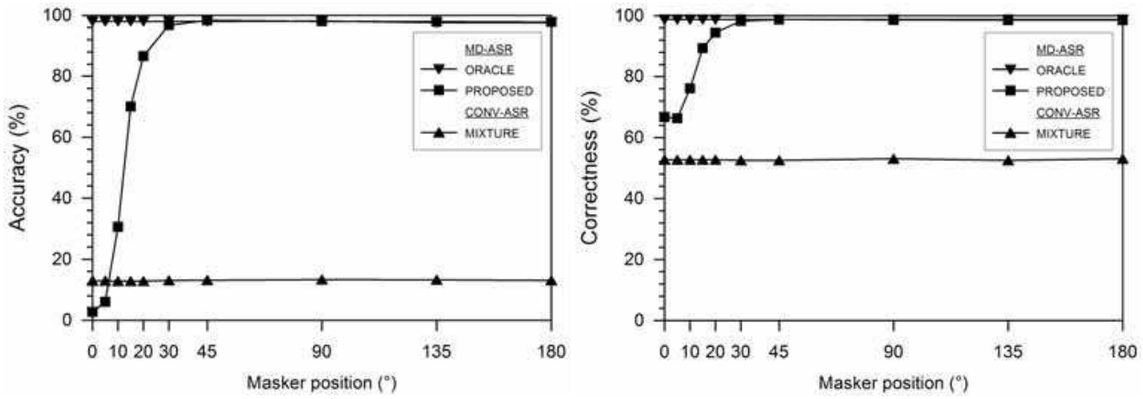
and in contrast to (27) additionally considers insertion errors denoted as INS. The accuracy score is therefore considered the more representative performance measure.

4.2 Results

A number of experiments were conducted to investigate the DUET separation in terms of speech recognition performance. The cepstral baseline measured the decoder's robustness against speech intrusions by scoring directly on the speech mixture. The missing data system reported the improvements over this baseline obtained by ignoring the spectral parts that are dominated by interfering speakers as indicated by the missing data masks. The performance in clean conditions (zero maskers) was 99.16% for the cepstral baseline and 98.54% for the spectral missing data system using the unity mask.

(A) Angular separation between target and masker: The first experiment used a female TIMIT speech masker to corrupt the target speech signal. The speaker of interest remained

stationary at the 0° location while the speech masker was placed at different angles but identical distance to the microphone pair (see Fig. 6a).



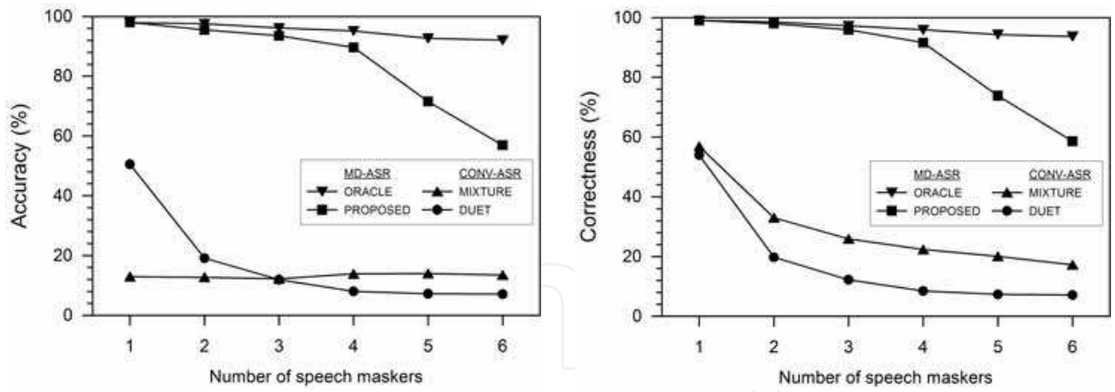
(a) (b)

Fig. 7. Speech recognition performance in terms of (a) accuracy and (b) correctness score for different masker positions. The target remained stationary at the 0° location. A conventional decoder using MFCC features was used to score on the speech mixtures. The spectral missing data system performed decoding with the proposed soft reliability mask (DUET+post-processing+mel-scale conversion) and the binary oracle mask.

The recognition performance was evaluated for a conventional recognizer and the missing data system using the oracle and estimated soft masks (Fig. 7).

Not surprisingly, the oracle mask performed best marking the upper performance bound for the missing data system while the conventional recognizer represented the lower bound. When the speech masker was placed between 45° to 180° angle relative to the target speaker, the estimated mask almost perfectly matched the oracle mask and hence achieved very high recognition accuracy. However, once the spatial separation between masker and target fell below 30° the accuracy score rapidly started to deteriorate falling below that of the cepstral baseline at the lowest separation angles (0°-5°). The correctness score followed the same trend as the accuracy score but performed better than the baseline for closely spaced sources. For these small angular separations the assumption that the sources possess distinct spatial signatures becomes increasingly violated and the DUET histogram localization starts to fail. The more the sources move together the less spatial information is available to estimate the oracle mask leading to large mask estimation errors. Nevertheless, the oracle masks (7) still exist even when target and masker are placed at identical positions because they depend on the local SNR rather than spatial locations.

(B) Number of concurrent speech maskers: The second experiment recorded the recognition performance when the target speaker was corrupted by up to six simultaneous TIMIT maskers (Fig. 8). Accuracy and correctness score were measured for the conventional recognizer using as input the speech mixture or the demixed target speaker as generated by DUET. As before, the missing data recognizer used the oracle and estimated soft masks. The number of simultaneously active speech maskers was increased by successively adding one masker after another according to the order shown in Fig. 6b.



(a) (b)

Fig. 8. Speech recognition performance in terms of (a) accuracy and (b) correctness score for different numbers of concurrent speech maskers. A conventional decoder using MFCC features was used to score on the speech mixtures and DUET's reconstructed target signal. The spectral missing data system performed decoding with the proposed soft reliability mask (DUET+post-processing+mel-scale conversion) and the binary oracle mask.

As expected, the conventional recognizer performed very poorly when scoring on the speech mixture. Performance dropped from 99% in clean conditions to 13% for the single speech masker case. Clearly, state-of-the-art cepstral feature extraction alone provides no protection against additive noise intrusions. For all but the single masker case, it also failed to produce significant improvements for the demixed DUET speech signal. In fact, for most conditions scoring on the speech mixture was better than decoding with the demixed DUET output. As discussed in Section 2.4 and 3.1, conventional speech recognizers require complete data and can not deal with masked spectra such as produced by DUET.

In contrast, the missing data system is able to handle missing feature components and provided the upper performance bound when using the oracle mask. Performance degraded very gradually with only a 6% decrease between clean conditions and corruption with six speech maskers. The estimated soft missing data masks closely matched the performance of the oracle masks for up to three simultaneously active speech maskers before starting to fall behind. The more speakers are present in the mixture the more the sparseness assumption (5) becomes invalid making an accurate peak detection in the attenuation-delay histogram increasingly difficult. Indeed, closer inspection of the 5 & 6 masker scenarios revealed that often peaks were overlapping and the peak detection algorithm failed to identify the locations correctly. For example, once the fifth masker was added, we observed in some cases that the histogram showed only four distinct peaks instead of five. This occasionally led the peak detection algorithm to place the fifth peak near the target speaker location. Due to DUET's minimum distance classification the wrongly detected speaker location absorbed some of the T-F points actually belonging to the target speaker. Consequently, performance dropped significantly for the 5 & 6 masker configurations, as evident from Fig. 8. Results can be improved somewhat by using soft assignments (Araki et al., 2006a; Kühne et al., 2007a) instead of the winner-takes-it-all concept utilized for the mask construction in (12).

(C) Mask post-processing: The last experiment investigated the influence of the proposed mask post-processing for a four speaker configuration (three maskers). To underline the importance of the mask smoothing the recognition performance with and without the

proposed two-dimensional median filtering was measured (see Table 2). In order to eliminate the effect of the histogram peak detection the true mixing parameters were directly passed to the mask construction and no source localization was performed.

Mask type	COR %	ACC %	DEL	SUB	INS
Without mask smoothing	88.62	75.37	17	92	127
With mask smoothing	94.57	93.53	12	40	10

Table 2. Recognition results in terms of HTK correctness (COR) and accuracy (ACC) score for missing data masks with and without median smoothing. The number of insertions (INS), deletions (DEL) and substitutions (SUB) is also given.

Clearly, if no median smoothing is applied to the DUET masks the recognized digit hypotheses contained a high number of insertion and substitution errors. Over 70% of the observed insertions were caused by the digit models "oh" and "eight". With the proposed median smoothing technique both the insertion and substitution errors were dramatically reduced resulting in an improved recognition performance.

5. Discussion

The experimental results reported here suggest that DUET might be used as an effective front-end for missing data speech recognition. Its simplicity, robustness and easy integration into existing ASR architecture are the main compelling arguments for the proposed model. It also fundamentally differs from other multi-channel approaches in the way it makes use of spatial information. Instead of filtering the corrupted signal to retrieve the sources (McCowan et al., 2000; Low et al., 2004, Seltzer et al., 2004a) the time-frequency plane is partitioned into disjoint regions each assigned to a particular source.

A key aspect of the model is the histogram peak detection. Here, we assumed prior knowledge about the number of speakers which should equal the number of peaks in the histogram. However, for a high number of simultaneous speakers the sparseness assumption becomes increasingly unrealistic and as a consequence sometimes histogram peaks are not pronounced enough in the data set. Forcing the peak detection algorithm to find an inadequate number of peaks will produce false localization results. Ultimately, the algorithm should be able to automatically detect the number of sources visible in the data which is usually denoted as unsupervised clustering. This would indeed make the source separation more autonomous and truly blind. However, unsupervised clustering is a considerably more difficult problem and is still an active field of research (Grira et al., 2004). Other attempts to directly cluster the attenuation and delay distributions using a statistical framework have been reported elsewhere (Araki et al., 2007; Mandel et al., 2006) and would lead to probabilistic mask interpretations.

A point of concern is the microphone distance d_{mic} that was kept very small to avoid phase ambiguities (Yilmaz & Rickard, 2004). Clearly, this limits the influence of the attenuation parameter (see Fig. 2a). Rickard (2007) has offered two extensions to overcome the small sensor spacing by using phase differentials or tiled histograms. Another option to consider

is the use of multiple microphone pairs or sensor arrays allowing for full three-dimensional source localization (Araki et al., 2006; Araki et al., 2007).

While the proposed median smoothing was highly successful in reducing spurious points in the time-frequency masks the filter was applied as a post-processing step only. Other more sophisticated methods that incorporate neighborhood information already into the mask assignment or the peak detection itself might be more appropriate. In particular, Markov Random Fields (Li, 2001) have been quite successful in the field of image processing but tend to be more complex and demanding in terms of computational resources. Other schemes for incorporating neighborhood information into clustering or mixture model learning are also readily available (Ambroise et al., 1997; Chuang et al., 2006). The advantage of the proposed post-processing scheme lies in its simplicity and relatively fast computation. Nevertheless, careful selection of the size of the median filter is required as otherwise the filter tends to remove too much energy of the target signal.

In regards to related work the overall architecture of our system is in line with previously proposed binaural CASA models. However, the DUET separation framework differs in some key aspects as it models human hearing mechanisms to a much lesser degree. Whereas Harding et al. (2006) and Roman et al. (2003) perform mask estimation for each critical band using supervised learning techniques, DUET blindly estimates these masks based on a simple frequency independent classification of attenuation and delay parameters. The spatial cues are extracted from STFT ratios which offer significant speedups over computationally expensive cross-correlation functions commonly used to compute binaural ITDs (see also Kim & Kil (2007) for an efficient method of binaural ITD estimation using zero-crossings). More importantly, Roman et al. (2003) need to recalibrate their system for each new spatial source configuration which is not required in our model. DUET also directly operates on the mixture signals and does not employ Head-Related-Transfer-Functions (HRTFs) or gammatone filterbanks for spectral analysis. However, we expect supervised source localization schemes to outperform DUET's simple histogram peak detection when angular separation angles between sources are small (0°-15°).

In terms of ASR performance we achieved comparable results to Roman et al. (2003), in that the estimated masks matched the performance of the oracle masks. Recognition accuracy remained close to the upper bound for up to three simultaneous speech maskers. While other studies (Roman et al., 2003; Mandel et al., 2006) have reported inferior localization performance of DUET even for anechoic, two or three source configurations we can not confirm these observations based on the experimental results discussed here. Mandel et al. (2006) offer a possible explanation for this discrepancy by stating that DUET was designed for a closely spaced omni-directional microhone pair and not the dummy head recordings used in binaural models.

Finally, we acknowledge that the results presented here were obtained under ideal conditions that met most of the requirements of the DUET algorithm. In particular the noise-free and anechoic environment can be considered as strong simplifications of real acoustic scenes and it is expected that under more realistic conditions the parameter estimation using DUET will fail. Future work is required to make the estimators more robust in hostile environments. To this extent, it is also tempting to combine the DUET parameters with other localization methods (Kim & Kil, 2007) or non-spatial features such as harmonicity cues (Hu & Wang, 2004). However, the integration of additional cues into the framework outlined here remains a topic for future research.

6. Conclusion

This chapter has investigated the DUET blind source separation technique as a front-end for missing data speech recognition in anechoic multi-talker environments. Using the DUET attenuation and delay estimators time-frequency masks were constructed by exploiting the sparseness property of speech in the frequency domain. The obtained masks were then smoothed with a median filter to remove spurious points that can cause insertion errors in the speech decoder. Finally, the frequency resolution was reduced by applying a triangular mel-filter weighting which makes the masks more suitable for speech recognition purposes. The experimental evaluation showed that the proposed model is able to retain high recognition performance in the presence of multiple competing speakers. For up to three simultaneous speech maskers the estimated soft masks closely matched the recognition performance of the oracle masks designed with a priori knowledge of the source spectra. In our future work we plan to extend the system to handle reverberant environments through the use of multiple sensor pairs and by combining the T-F masking framework with spatial filtering techniques that can enhance the speech signal prior to recognition.

7. Acknowledgments

This work was supported in part by The University of Western Australia, Australia and in part by National ICT Australia (NICTA). NICTA is funded through the Australian Government's Backing Australia's Ability Initiative, in part through the Australian Research Council.

8. References

Ambroise, C. ; Dang, V. & Govaert, G. (1997). Clustering of Spatial Data by the EM Algorithm, In: *geoENV I - Geostatistics for Environmental Applications,* Vol. 9, Series: Quantitative Geology and Geostatistics, pp. 493-504, Kluwer Academic Publisher

Araki, S.; Sawada, H.; Mukai, R. & Makino, S. (2005). A novel blind source separation method with observation vector clustering, *International Workshop on Acoustic Echo and Noise Control*, Eindhoven, The Netherlands, 2005

Araki, S.; Sawada, H.; Mukai, R. & Makino, S. (2006). DOA Estimation for Multiple Sparse Sources with Normalized Observation Vector Clustering, *IEEE International Conference on Acoustics, Speech and Signal Processing,* Toulouse, France, 2006

Araki, S.; Sawada, H.; Mukai, R. & Makino, S. (2006a). Blind Sparse Source Separation with Spatially Smoothed Time-Frequency Masking, *International Workshop on Acoustic Echo and Noise Control,* Paris, France, 2006

Araki, S.; Sawada, H.; Mukai, R. & Makino, S. (2007). Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors, *Signal Processing,* Vol. 87, No. 8, pp. 1833-1847

Barker, J.; Josifovski, L.; Cooke, M. & Green, P. (2000). Soft decisions in missing data techniques for robust automatic speech recognition, *Proccedings of the 6th International Conference of Spoken Language Processing,* Beijing, China, 2000

Bregman, A. (1990). Auditory Scene Analysis, MIT Press, Cambridge MA., 1990

Brown, G. & Cooke, M. (1994). Computational auditory scene analysis, *Computer Speech and Language,* Vol. 8, No. 4, pp. 297-336

Cerisara, C.; Demangea, S. & Hatona, J. (2007). On noise masking for automatic missing data speech recognition: A survey and discussion, *Speech Communication*, Vol. 21, No. 3, (2007), pp. 443-457

Cherry, E. (1953). Some experiments on the recognition of speech, with one and with two ears, *Journal of Acoustical Society of America,* Vol. 25, No. 5, (1953), pp. 975-979

Chuang, K.; Tzeng, H.; Chen, S.; Wu, J. & Chen, T. (2006). Fuzzy c-means clustering with spatial information for image segmentation, *Computerized Medical Imaging and Graphics,* Vol. 30, No. 1, pp. 9-15

Cooke, M.; Green, P.; Josifovski, L. & Vizinho, A. (2001). Robust Automatic Speech Recognition with missing and unreliable acoustic data, *Speech Communication,* Vol. 34, No. 3, (2001), pp. 267-285

de Veth, J ; de Wet, F., Cranen, B. & Boves, L. (2001). Acoustic features and a distance measure that reduces the impact of training-set mismatch in ASR, *Speech Communication,* Vol. 34, No. 1-2, (2001), pp. 57-74

El-Maliki, M. & Drygajlo, A. (1999). Missing Features Detection and Handling for Robust Speaker Verification, *Proceedings of Eurospeech,* Budapest, Hungary, 1999

Garofolo, J.; Lamel, L.; Fisher, W.; Fiscus, J.; Pallett, D.; Dahlgren, N. & Zue, V. (1993). TIMIT Acoustic-Phonetic Continuous Speech Corpus, *Linguistic Data Consortium,* Philadelphia, USA

Grira, N.; Crucianu, M. & Boujemaa, N. (2004). Unsupervised and Semi-supervised Clustering: a Brief Survey, In: *A Review of Machine Learning Techniques for Processing Multimedia Content,* MUSCLE European Network of Excellence, 2004

Harding, S.; Barker, J. & Brown, G. (2005). Mask Estimation Based on Sound Localisation for Missing Data Speech Recognition, *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Philadelphia, USA, 2005

Harding, S.; Barker, J. & Brown, G. (2006). Mask estimation for missing data speech recognition based on statistics of binaural interaction, *IEEE Transactions on Audio, Speech, and Language Processing,* Vol. 14, No. 1, (2006), pp. 58-67

Harte, N.; Hurley, N.; Fearon, C. & Rickard, S. (2005). Towards a Hardware Realization of Time-Frequency Source Separation of Speech, *European Conference on Circuit Theory and Design,* Cork, Ireland, 2005

Hu, G. & Wang, D. (2004). Monaural Speech Segregation Based on Pitch Tracking and Amplitude Modulation, *IEEE Transactions on Neural Networks,* Vol. 15, No. 5, (2004), pp. 1135-1150

Hyvärinen, H. (1999). Survey on Independent Component Analysis, *Neural Computing Surveys,* Vol. 2, (1999), pp. 94-128

Kim, Y. & Kil, R. (2007). Estimation of Interaural Time Differences Based on Zero-Crossings in Noisy Multisource Environments, *IEEE Transactions on Audio, Speech, and Language Processing,* Vol. 15, No. 2, (2007), pp. 734-743

Kolossa, D.; Sawada, H.; Astudillo, R.; Orglmeister, R. & Makino, S. (2006). Recognition of Convolutive Speech Mixtures by Missing Feature Techniques for ICA, *Asilomar Conference on Signals, Systems and Computers,* Pacific Grove, CA, 2006

Kühne, M.; Togneri, R. & Nordholm, S. (2007). Mel-Spectrographic Mask Estimation for Missing Data Speech Recognition using Short-Time-Fourier-Transform Ratio Estimators, *IEEE International Conference on Acoustics, Speech and Signal Processing,* Honolulu, USA, 2007

Kühne, M.; Togneri, R. & Nordholm, S. (2007a). Smooth soft mel-spectrographic masks based on blind sparse source separation, *Proceedings of Interspeech 2007*, Antwerp, Belgium, 2007

Leonard, R. (1984). A database for speaker-independent digit recognition, *IEEE International Conference on Acoustics, Speech, and Signal Processing,* San Diego, USA, 1984

Li, S. (2001). Markov Random Field Modeling in Image Analysis, Springer-Verlag, 2001

Low, S.; Togneri, R. & Nordholm, S. (2004). Spatio-temporal processing for distant speech recognition, *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Montreal, Canada, 2004

Mandel, M.; Ellis, D. & Jebara, T. (2006). An EM Algorithm for Localizing Multiple Sound Sources in Reverberant Environments, *Twentieth Annual Conference on Neural Information Processing Systems,* Vancouver, B.C., Canada, 2006

McCowan, I.; Marro, C. & Mauuary, L. (2000). Robust Speech Recognition Using Near-Field Superdirective Beamforming with Post-Filtering, *IEEE International Conference on Acoustics, Speech and Signal Processing,* Istanbul, Turkey, 2000

Moore, B. (2003). An introduction to the psychology of hearing, Academic Press, San Diego, CA

Morris, A.; Barker, J. & Bourlard, H. (2001). From missing data to maybe useful data: soft data modelling for noise robust ASR, *WISP,* Stratford-upon-Avon, England, 2001

Raj, B. & Stern, R. (2005). Missing-feature approaches in speech recognition, *IEEE Signal Processing Magazine,* Vol. 22, No. 5, (2005), pp. 101-116

Rickard, S. (2007). The DUET Blind Source Separation Algorithm, In: *Blind Speech Separation,* Makino, S.; Lee, T.-W.; Sawada, H., (Eds.), Springer-Verlag, pp. 217-237

Roman, N.; Wang, D. & Brown, G. (2003). Speech segregation based on sound localization, *Journal of the Acoustical Society of America,* Vol. 114, No. 4, (2003), pp. 2236-2252

Russ, J. (1999). The Image Processing Handbook, CRC & IEEE, 1999

Seltzer, M.; Raj, B. & Stern, R. (2004). A Bayesian classifier for spectrographic mask estimation for missing feature speech recognition, *Speech Communication*, Vol. 43, No. 4, (2004), pp. 379-393

Seltzer, M.; Raj, B. & Stern, R. (2004a). Likelihood-Maximizing Beamforming for Robust Hands-Free Speech Recognition, *IEEE Transactions on Speech and Audio Processing,* Vol. 12, No. 5, (2004), pp. 489-498

Van Hamme, H. (2004). Robust speech recognition using cepstral domain missing data techniques and noisy masks, *IEEE International Conference on Acoustics, Speech, and Signal Processing,* Montreal, Canada, 2004

Wang, D. (2005). On Ideal Binary Mask As the Computational Goal of Auditory Scene Analysis, In: *Speech Separation by Humans and Machine,* Divenyi, P., pp. 181-197, Kluwer Academic

Yilmaz, Ö. & Rickard, S. (2004). Blind Separation of Speech Mixtures via Time-Frequency Masking, *IEEE Transactions on Signal Processing,* Vol. 52, No. 7, (2004), pp. 1830-1847

Young, S.; Evermann, G.; Gales, M.; Hain, T.; Kershaw, D.; Liu, X.; Moore, G.; Odell. J; Ollason, D.; Povey, D.; Valtchev, V. & Woodland, P. (2006). The HTK Book, *Cambridge University Engineering Department*, 2006

Speech Recognition

Edited by France Mihelic and Janez Zibert

ISBN 978-953-7619-29-9

Hard cover, 550 pages

Publisher InTech

Published online 01, November, 2008

Published in print edition November, 2008

Chapters in the first part of the book cover all the essential speech processing techniques for building robust, automatic speech recognition systems: the representation for speech signals and the methods for speech-features extraction, acoustic and language modeling, efficient algorithms for searching the hypothesis space, and multimodal approaches to speech recognition. The last part of the book is devoted to other speech processing applications that can use the information from automatic speech recognition for speaker identification and tracking, for prosody modeling in emotion-detection systems and in other speech processing applications that are able to operate in real-world environments, like mobile communication services and smart homes.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen