MEL-SPECTROGRAPHIC MASK ESTIMATION FOR MISSING DATA SPEECH RECOGNITION USING SHORT-TIME-FOURIER-TRANSFORM RATIO ESTIMATORS

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

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

ABSTRACT

This paper adopts the framework of DUET, a recently proposed blind source separation (BSS) method, for speech recognition. Based on the attenuation and delay estimation in stereo signals spectrographic masks are designed to extract a target speaker from a mixture containing multiple speech sources. Instead of using these masks for resynthesis we avoid source reconstruction and propose to combine the source separation with a missing data speech recognizer. The obtained results for connected digit experiments in a multi-speaker environment demonstrate the validity of the approach.

Index Terms— speech recognition, masks, missing data, attenuation and delay estimation

1. INTRODUCTION

To effectively apply automatic speech recognition (ASR) systems in real world scenarios it is necessary to handle hostile environments with multiple speech and noise sources. Beamforming and blind source separation (BSS) techniques have been applied with some success for distant speech recognition [1, 2]. These methods aim to filter out the desired target speech signal while suppressing noise and other interferences prior to recognition. Subsequent ASR systems then take the enhanced speech signal as input and perform speech recognition usually based on mel-frequency cepstral coefficients (MFCCs). While beamforming requires a large number of sensor elements to achieve a good separation, most BSS methods fail when the number of sources exceeds the number of microphones.

Recently, there has been an increased interest in underdetermined BSS methods that can deal with more sources then sensors [3, 4, 5]. In [3] the DUET algorithm was proposed to solve the underdetermined problem for the anechoic case. It was shown that under the so-called W-disjoint orthogonality (W-DO) assumption it is very unlikely that spectra of two or more speakers overlap. Similar observations for the single channel case have been made in the speech recognition community leading to the missing data automatic speech recognition (MD-ASR) paradigm [6]. To the best of our knowledge, DUET has been used solely in the context of BSS rather than for ASR. As the demixing relies on the estimation of a time-frequency mask we propose to adopt the DUET framework for MD-ASR.

The reminder of this paper is as follows. Section 2 illustrates how spectrographic masks suitable for MD-ASR can be estimated using Short-Time-FOURIER-Transform (STFT) ratios. Section 3 describes the missing data recognizer, followed by connected digit experiments reported in Section 4. A discussion of the results and relations to other approaches concludes the paper in Section 5.

2. SPECTROGRAPHIC MASK ESTIMATION

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

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

where a_{mj} and δ_{mj} are the attenuation and delay parameters of source s_j at microphone x_m. Because of the anechoic environment the attenuation and delay parameters of the first mixture can be merged with the source definitions and therefore only relative attenuation and delay parameters a_j and δ_j between both microphones are considered in the following. Taking the discrete STFT of (1) the mixing model equation in the time-frequency domain can be approximated as

$$\begin{bmatrix} X_1(k,l) \\ X_2(k,l) \end{bmatrix} \approx \begin{bmatrix} 1 & \cdots & 1 \\ a_1 e^{-il\omega_0\delta_1} & \cdots & a_N e^{-il\omega_0\delta_N} \end{bmatrix} \begin{bmatrix} S_1(k,l) \\ \vdots \\ S_N(k,l) \end{bmatrix}, \qquad (2)$$

where $X_m(k,l)$ and $S_j(k,l)$ are the STFT transforms using a time-frequency grid defined by $(k\tau_0, l\omega_0)$ [3]. We used a 25 ms HAMMING-window for the STFT with a frame shift of 10 ms to match the number of frames generated in the ASR feature extraction. Under the W-DO assumption exactly one source S_j will be active at any time-frequency point (k,l) and it becomes obvious that a set of instantaneous attenuation and delay parameter estimators

$$\tilde{a}(k,l) := \left| \frac{X_2(k,l)}{X_1(k,l)} \right|, \quad \tilde{\delta}(k,l) := -\frac{1}{l\omega_0} \arg\left(\frac{X_2(k,l)}{X_1(k,l)} \right) \qquad (3)$$

This work was supported in part by the University of Western 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.

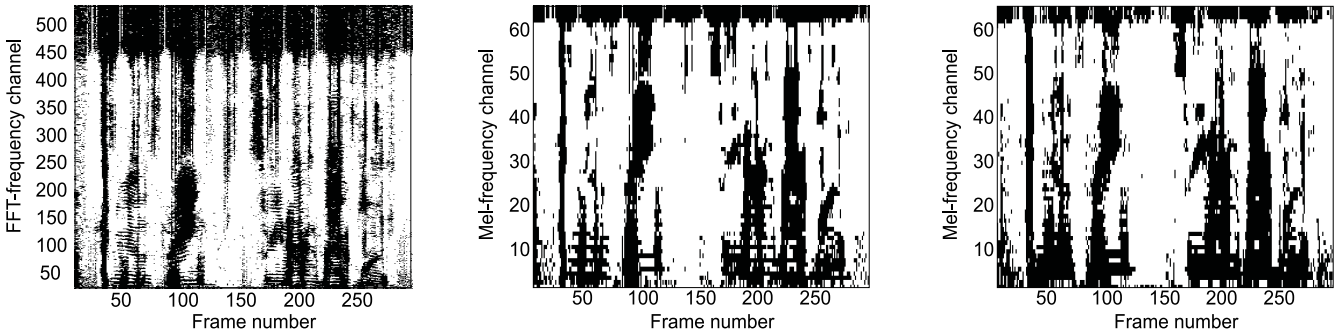


Fig. 1. Example of binary spectrographic masks for extracting a target source (black areas) from a mixture of two speakers. Left: FFT-frequency resolution mask M produced by the STFT ratio estimators; Middle: mel-frequency resolution mask $\mathfrak{M}$ produced by triangular mel-weighting; Right: corresponding oracle mel-frequency mask $\mathfrak{O}$

can be obtained by applying the magnitude and phase operator onto the complex STFT ratio of the two microphone signals. It has been shown in [3] that the number of sources and their corresponding mixing parameters and can be identified based on the number and location of the peaks in a 2-D power weighted $(\tilde{a}, \tilde{\delta})$ histogram. Once the mixing parameter estimates $(\hat{a}_j, \hat{\delta}_j)$ are obtained they can be used to label all $(\tilde{a}, \tilde{\delta})$ pairs resulting in N disjoint time-frequency masks. For a more detailed review of the original DUET method the reader is referred to [3]. Based on an EUCLIDean distance measure

$$D_j(k,l) := \sqrt{\big(\tilde{a}(k,l) - \hat{a}_j\big)^2 + \big(\tilde{\delta}(k,l) - \hat{\delta}_j\big)^2} \quad (4)$$

we construct for each source s_j a binary spectrographic mask

$$M_j := \mathbb{1}_{\left\{(k,l):j=\underset{z}{\operatorname{argmin}}\, D_z(k,l)\right\}} \quad (5)$$

where $\mathbb{1}$ denotes the indicator function assigning a 1 to all time-frequency points (k,l) with a minimum distance $D_j(k,l)$ and a 0 otherwise. Instead of using these masks for resynthesis as done in [3] we propose to avoid source reconstruction and directly exploit M_j through MD-ASR. However, for speech recognition purposes a perceptual frequency scale rather than the linear STFT frequency axis is preferred. Using a threshold $\theta \in [0,1]$ the STFT resolution mask can be converted to a binary mel-spectrographic mask

$$\mathfrak{M}_j := \mathbb{1}_{\left\{(k,b):\frac{\sum_l \lambda_b(l) M_j(k,l)}{\sum_l \lambda_b(l)} \geq \theta\right\}} \quad (6)$$

by applying for each subband b a triangular mel-weighting function

$$\lambda_b(l) = \begin{cases} 0 & l\omega_0 < \omega_{c_{(b-1)}}, \\ \frac{l\omega_0 - \omega_{c_{(b-1)}}}{\omega_{c_b} - \omega_{c_{(b-1)}}} & \omega_{c_{(b-1)}} \leq l\omega_0 \leq \omega_{c_b}, \\ \frac{\omega_{c_{(b+1)}} - l\omega_0}{\omega_{c_{(b+1)}} - \omega_{c_b}} & \omega_{c_b} \leq l\omega_0 \leq \omega_{c_{(b+1)}}, \\ 0 & l\omega_0 > \omega_{c_{(b+1)}}, \end{cases} \quad (7)$$

where ω_{c_b} specifies the center frequency

$$\omega_{c_b} = 2\pi \cdot 700 \left(10^{\mathfrak{f}_{c_b}/2595} - 1\right) \quad (8)$$

corresponding to the perceptual mel-frequency scale with

$$\mathfrak{f}_{c_b} = \mathfrak{f}_l + b \cdot \frac{\mathfrak{f}_h - \mathfrak{f}_l}{B+1}, \qquad b = 1, \ldots, B \quad (9)$$

where B is the number of mel-frequency channels and $\mathfrak{f}_l, \mathfrak{f}_h$ are the lower and higher cut-offs of the mel-frequency axis. In this paper a threshold $\theta = 0.7$ was used for all experiments. Similarly, the mel-frequency spectrum $\mathfrak{S}_j$ of a source s_j can be computed via $\mathfrak{S}_j(k,b) = \sum_l \lambda_b(l)|S_j(k,l)|$ [7]. As we are interested in performing recognition experiments only for the target speaker s_i we select the mask that is closest to the oracle binary mel-frequency mask of this speaker which is defined by

$$\mathfrak{O}_i := \mathbb{1}_{\left\{(k,b):20\log_{10}\left(\frac{\mathfrak{S}_i(k,b)}{\sum_{j\neq i}\mathfrak{S}_j(k,b)}\right)\geq 0\right\}}. \quad (10)$$

The oracle mask determines all time-frequency points where the power of the target speaker exceeds or equals the power of the interferences. Please note that $\mathfrak{O}_i$ can only be computed if the source signals are known prior to the mixing process. Figure 1 shows an example for the mask resolution conversion and the corresponding oracle mask designed using a priori knowledge of the sources.

3. MISSING DATA SPEECH RECOGNITION

In this paper an Hidden MARKOV Model (HMM) based missing data speech recognizer [6] is used as it can directly exploit the spectrographic masks discussed above. While the HMMs are trained on clean speech in exactly the same manner as in conventional ASR the decoding is treated differently in MD-ASR. Additionally to the feature vector sequence $\mathbf{o}$ a time-frequency mask is required. The mask declares each feature component as reliable $\mathbf{o}_r$ or unreliable $\mathbf{o}_u$ using a hard or soft decision [8]. As this study employed mel-filterbank energy features we used the bounded marginalization technique [6] to compute the observation probability density function (PDF) in HMM state q as

$$\phi_q(\mathbf{o}) = \sum_{p=1}^{P} c_{qp} \mathcal{N}_q(\mathbf{o}_r; \boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p) \int_{\mathbf{o}_{u_{\text{low}}}}^{\mathbf{o}_{u_{\text{high}}}} \mathcal{N}_q(\tilde{\mathbf{o}}_u; \boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)\, d\tilde{\mathbf{o}}_u, \quad (11)$$

where P denotes the number of mixture components, c_{qp} is the corresponding mixture weight and $\mathcal{N}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is a multivariate GAUSSean with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Conventional ASR systems often append so-called delta coefficients to static feature vectors [7]. In MD-ASR an additional decision about the reliability of these appended components is required. Here we follow [8] and declare delta components as unreliable if one of the static features involved in their calculation was declared as unreliable via

$$\Delta\mathfrak{M}(k,b) = \prod_{\kappa=-\Theta}^{\Theta} \mathfrak{M}(k+\kappa, b), \tag{12}$$

where Θ denotes the size of the temporal integration window used for the delta component calculation. While for the static energy features the integral in (11) can be evaluated over $[\mathbf{o}_{u_{\text{low}}}(b), \mathbf{o}_{u_{\text{high}}}(b)] = [0, \mathbf{o}_u(b)]$, no bounds on dynamic feature components were utilized here, thus we set $[\mathbf{o}_{u_{\text{low}}}(b), \mathbf{o}_{u_{\text{high}}}(b)] = [-\infty, \infty]$ for all delta feature components.

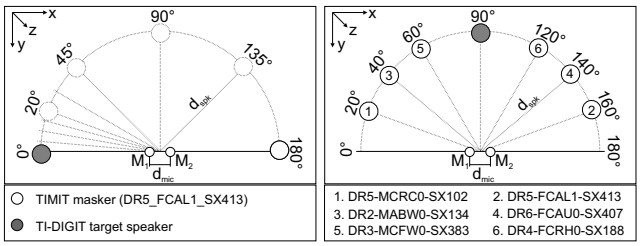
4. EVALUATION

The proposed system was evaluated via connected digit experiments on the TI-DIGIT database sampled at 20 kHz. The training set consisted of 4235 utterances spoken by 55 male speakers. The Hidden MARKOV Model Toolkit (HTK) [7] was used to train 11 word HMMs ('1'-'9','oh','zero') each with 8 emitting states and 2 silence models ('sil','sp') with 3 and 1 state. All HMMs followed standard left-to-right models without skips using continuous GAUSSean PDFs with diagonal covariance matrices and 10 mixture components. Two different sets of acoustic models were created. Both used 25 ms HAMMING-windows with 10 ms frame shifts for the STFT analysis. The first set of HMMs was used as the single channel baseline system employing 13 MFCCs derived from a 32-channel HTK mel-filterbank plus delta and acceleration coefficients ($\Theta = 2$) and cepstral mean normalization (CMN). The second model set was used for the missing data speech recognizer and used spectral rather than cepstral features. In particular, acoustic features were extracted from a 64-channel HTK mel-filterbank and their first order delta coefficients ($\Theta = 2$) were appended to the static feature vector.

The test set consisted of 166 utterances of 7 male speakers containing at least 4 connected digits mixed with several masking utterances taken from the TIMIT database each with a signal-to-interferer ratio (SIR) of 0 dB. Stereo mixtures were created by using an anechoic room impulse response of a simulated room of size 4 m x 6 m x 4 m (length x width x height). The two microphones were positioned in the center of the room, 1 m above the ground, with an inter-element distance of $d_{mic} = 1.72\,\text{cm}$ to guarantee accurate phase parameter estimates [3].

4.1. Experiment 1: single masking source

In this experiment a single speech masking source was used to corrupt the target speech signal. Figure 2a demonstrates the setup for the two speaker scenario. The speaker of interest remained stationary at the 0°location while the speech masker



(a) (b)

Fig. 2. Anechoic room configuration for a single masker in various positions (a) and multiple TIMIT maskers (b).

simulated by a female TIMIT speaker was placed in different angles but identical distance $d_{spk} = 1\,\text{m}$ to the microphone pair. The recognition performance was evaluated for the missing data system using the oracle mask $[\mathfrak{O}, \Delta\mathfrak{O}]$, the estimated mask $[\mathfrak{M}, \Delta\mathfrak{M}]$ and a conventional recognizer as baseline. The recognizers used the appropriate acoustic models described in the previous section. The obtained results are presented in Figure 3.

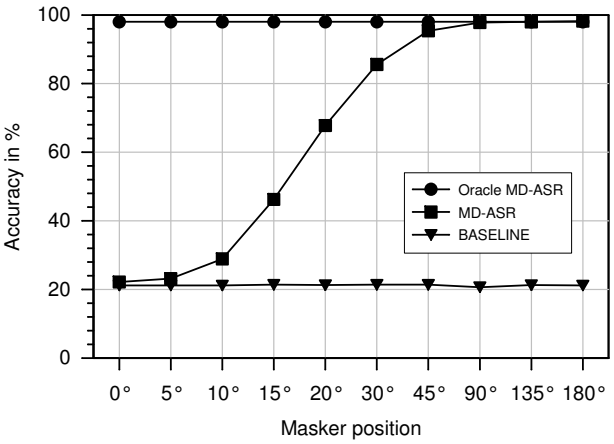


Fig. 3. Recognition accuracy of the target speaker depending on the location of a single speech masker.

As expected, the oracle mask performed best marking the upper performance bound for the MD-ASR system while the single-channel ASR baseline represented the lower bound using only one microphone and no spatial information. 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 a very high accuracy. However, once the masker was placed below the 30°angle the performance rapidly started to deteriorate merging with that of the single-channel baseline at 0°. The more the sources move together the less spatial information is available to estimate the oracle mask which nevertheless still exists even when target and masker are placed at identical positions.

4.2. Experiment 2: multiple masking sources

In this experiment up to 6 different speech maskers (3 male, 3 female) were used to test the ability of the mask estimation algorithm. The recognition performance for the target speaker was recorded for the missing data recognizer and the single channel baseline using identical models as in the first experiment. The number of simultaneously active speech sources was increased by successively adding one masker after another according to the order shown in Figure 2b. The obtained results are presented in Figure 4.

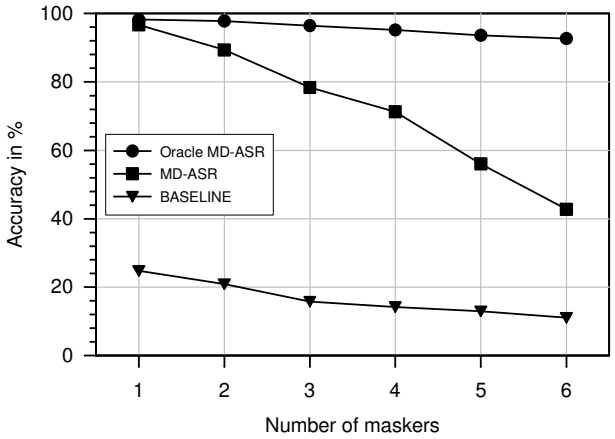


Fig. 4. Recognition accuracy of the target speaker depending on the number of simultaneously active speech maskers.

Similar to the first experiment the performance bounds were marked by the oracle mask and the single channel baseline. The accuracy of the oracle mask only slightly degraded from 98.2 % for the 1 masker case down to 92.4 % for the 6 masker scenario. The MD-ASR system using the estimated mask showed less robustness when the number of simultaneous speaker was increased. However, considering the fact that only two microphones were used the 69 % recognition accuracy obtained for the 4 masker scenario is very promising.

5. DISCUSSION

The conducted experiments demonstrate that STFT ratio estimators can be successfully used to obtain spectrographic masks suitable for MD-ASR. The results of the oracle mask show that even for the 7 speaker scenario there exist spectrographic masks leading to a high recognition accuracy. The drop in performance for the estimated masks can be explained by the fact that once the number of speakers increases the SIR improvements start to decrease due to misassigned mask points. Also it gets more difficult to accurately estimate the histogram peaks in these scenarios because the sparseness assumption becomes increasingly unrealistic. Nevertheless the achieved performance is far superior to that of the single channel baseline system.

However, the proposed method is not free from any drawbacks. The requirement that the microphone distance has to be small enough to avoid phase ambiguities limits the influence of the attenuation parameter. As speech contains its main information in the frequency range of 100 Hz to 4 kHz, the problem can be relaxed by lowpass filtering which in turn would allow us to increase the microphone spacing. Probably the biggest limitation is the assumption of an anechoic mixing model that prohibits the use in reverberant environments.

Recently, binaural computational auditory scene analysis (CASA) systems based on interaural time and intensity differences (ITD)/(IID) have been used in conjunction with MD-ASR [9, 10]. These models use computationally intensive cross-correlation methods to derive ITD estimates for each frequency band. Similar to DUET they use joint ITD-IID histograms or PDFs to construct time-frequency masks that are able to suppress unwanted interferences. However, unlike DUET the histograms/PDFs are constructed via training data using a priori information about the target source and/or interferences. For example, the system in [10] requires a retraining for every new spatial configuration. Although the attenuation and delay parameters can be interpreted as a crude approximation of the IID and ITD values DUET is a purely engineering-computational model.

Both, binaural CASA and DUET fundamentally differ from other multi-channel approaches in the way they make use of spatial information. Instead of filtering the corrupted signal to retrieve the sources the time-frequency plane is partitioned into disjoint regions each assigned to a particular source. The proposed system in this paper can be seen as a low-computational alternative to binaural CASA systems. It is in particular attractive for ASR scenarios where only limited resources for multi-channel processing are available (e.g., mobile phones).

In our future work we want to test the proposed method on real data, extend it to handle reverberant environments and introduce soft decisions into the mask estimation. Also it remains to be seen whether the algorithm stays robust if non-sparse noise interferers are present in the mixture.

6. REFERENCES

[1] I.A. McCowan, C. Marro, and L. Mauuary, "Robust speech recognition using near-field superdirective beamforming with post-filtering," in *ICASSP*, Istanbul, Turkey, 2000.

[2] S.Y. Low, R. Togneri, and S. Nordholm, "Spatio-temporal processing for distant speech recognition," in *ICASSP*, Montreal, Canada, 2004.

[3] Ö. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on signal processing*, vol. 52, no. 7, pp. 1830–1847, 2004.

[4] R. Nickel and A. Iyer, "A novel approach to automated source separation in multispeaker environments," in *ICASSP*, France, 2006.

[5] M. Puigt and Y. Deville, "A time-frequency correlation-based blind source separation method for time-delayed mixtures," in *ICASSP*, Toulouse, France, 2006.

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

[7] S. Young et al., *The HTK Book*, Cambridge University Engineering Department, 2005.

[8] J. Barker, L. Josifovski, M.P. Cooke, and P.D. Green, "Soft decisions in missing data techniques for robust automatic speech recognition," in *ICSLP*, Beijing, China, 2000.

[9] S. Harding, J. Barker, and G.J. Brown, "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, pp. 58–67, 2006.

[10] N. Roman, D.L. Wang, and G.J. Brown, "Speech segregation based on sound localization," *JASA*, vol. 114, no. 4, pp. 2236–2252, 2003.