Smooth soft mel-spectrographic masks based on blind sparse source separation

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 investigates the use of DUET, a recently proposed blind source separation method, as front-end for missing data speech recognition. Based on the attenuation and delay estimation in stereo signals soft time-frequency masks are designed to extract a target speaker from a mixture containing multiple speech sources. A postprocessing step is introduced in order to remove isolated mask points that can cause insertion errors in the speech decoder. The results for connected digit experiments in a multi-speaker environment demonstrate that the proposed soft masks closely match the performance of the oracle mask designed with a priori knowledge of the source spectra.

Index Terms: speech recognition, missing data, attenuation and delay estimation

1. Introduction

The concept of time-frequency (TF) masking has recently attracted some interest in the field of blind signal separation (BSS) [1, 2]. Demixing via TF-masks has the potential to separate mixtures with more sources than sensors as it does not rely on matrix inversion. Instead the TF-plane is partitioned into disjoint regions each assigned to a particular source. The sources are then recovered by converting each region back into the time domain. It seems promising to use BSS systems as front-ends for automatic speech recognition (ASR). In [3] we have proposed such a combination using a BSS technique called DUET and a missing data (MD) speech recognizer. The proposed system uses DUET to estimate TF-masks in the sparse Short-Time-Fourier-Transform (STFT) domain before converting the high STFT frequency resolution to a perceptual mel-frequency scale suitable for ASR. In this way we can avoid source reconstruction and directly exploit the spectrographic masks for MD-ASR. This paper extends our previous work in two regards. Firstly, we replace binary masks with soft masks which have been proven to be beneficial for both speech recognition and delay estimation.


delay estimation

2. Spectrographic mask estimation

2.1. Parametric 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

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

where $a_{mj}$ and $\delta_{mj}$ are the attenuation and delay parameters of source $s_j$ at microphone $x_m$. The mixing model equation can be approximated in the TF-domain as

$$\begin{bmatrix} X_1(k, l) \\ X_2(k, l) \end{bmatrix} \approx \begin{bmatrix} a_{11} e^{-il\omega_0 \delta_{11}} \cdots a_{1N} e^{-il\omega_0 \delta_{1N}} \\ a_{21} e^{-il\omega_2 \delta_{21}} \cdots a_{2N} e^{-il\omega_2 \delta_{2N}} \end{bmatrix} \begin{bmatrix} S_1(k, l) \\ \vdots \\ S_N(k, l) \end{bmatrix} \tag{2}$$

where $X_m(k, l)$ and $S_j(k, l)$ are STFT transforms using a TF-grid defined by $(k \tau_j, l \omega_j)$. The objective is now to estimate the mixing parameter pairs $(a_{mj}, \delta_{mj})$, $\forall j = 1, \ldots, N$ via DUET.

2.2. Estimation of mixing parameters via DUET

The DUET method proceeds in constructing a ratio matrix

$$\frac{X_2(k, l)}{X_1(k, l)} = \frac{\sum_{j=1}^{N} a_{2j} e^{-il\omega_2 \delta_{2j}} S_j(k, l)}{\sum_{j=1}^{N} a_{1j} e^{-il\omega_1 \delta_{1j}} S_j(k, l)} \tag{3}$$

and assumes that only one arbitrary source $S_j$ will be active at any TF-point such that (3) simplifies to

$$\frac{X_2(k, l)}{X_1(k, l)} = \frac{a_{2j} e^{-il\omega_2 \delta_{2j}}}{a_{1j} e^{-il\omega_1 \delta_{1j}}} \tag{4}$$

where $a_j$ and $\delta_j$ are relative attenuation and delay parameters. A set of instantaneous attenuation and delay estimators

$$\hat{a}(k, l) := \frac{X_2(k, l)}{X_1(k, l)}, \quad \hat{\delta}(k, l) := -\frac{1}{l \omega_0} \arg \left( \frac{X_2(k, l)}{X_1(k, l)} \right) \tag{5}$$

are obtained by applying the magnitude and phase operator onto (4). The mixing parameters are then estimated by locating the peaks in a power weighted $(\hat{a}, \hat{\delta})$-histogram, where

$$\hat{a}(k, l) := \hat{a}(k, l) - \frac{1}{l \omega_0} \arg \left( \frac{X_2(k, l)}{X_1(k, l)} \right)$$

For a more detailed description of DUET the reader is referred to [1].

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.
2.3. Soft mask construction

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. Previously we obtained a hard partition by assigning each TF-point in the \((\hat{\alpha}, \hat{\delta})\)-space exclusively to one source position using a minimum distance classification [3]. However, there exists a certain degree of uncertainty in the decision of which TF-point belongs to which source. In particular at the cluster boundaries the chance of a misclassification increases. It might therefore be appropriate to reflect this uncertainty with soft mask values \(M^s_j \in [0, 1]\) defined by the fractional assignment

\[
M^s_j(k, l) = \frac{e^{-\zeta d^2_j(k, l)}}{\sum_{j=1}^{N} e^{-\zeta d^2_j(k, l)}},
\]

where \(d_j(k, l)\) is the Euclidian distance

\[
d_j(k, l) = \sqrt{\left(\hat{\alpha}_j(k, l) - \hat{\alpha}_j\right)^2 + \left(\hat{\delta}_j(k, l) - \hat{\delta}_j\right)^2}
\]

and \(\zeta \in \mathbb{R}_+\) is a decay parameter determining the softness at the cluster boundaries. In this study we fixed the decay parameter to \(\zeta = 12\) which was determined empirically to give good results. Figure 1 illustrates the differences of hard and soft masks in the mixing parameter space for a three cluster example.

![Figure 1: Example of hard (left) and soft (right) time-frequency masks in the feature space with three sources located at \((\alpha_1, \delta_1), (\alpha_2, \delta_2), (\alpha_3, \delta_3)\) = \{(-0.03, 0.94), (0, 0), (0.03, -0.94)\}.

If we knew for each TF-point which source is dominant we could construct an oracle mask

\[
O_j(k, l) = \begin{cases} 1, & \text{if } 20 \log_{10} \left( \frac{|S_j(k, l)|}{\sum_{i \neq j} |S_i(k, l)|} \right) \geq 0 \\ 0, & \text{otherwise.} \end{cases}
\]

which determines all TF-points where the power of the speaker \(S_j\) exceeds or equals the power of the interferences (Figure 2e) [1]. Note that \(O_j\) can only be computed if the source signals are known prior to the mixing process.

2.4. Mask postprocessing

DUET solely relies on the mask assignment in the feature space and does not take the correlation of neighboring TF-points into account. We observed that for mixtures with more than two sources the DUET masks were overlaid by scattered isolated ”noise” points when compared with the oracle mask (see Figure 2a,e). 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 edge preserving median filter [6]. Here we used a 5 x 5 plus sign-shaped median filter

\[
\tilde{M}^s_j(k, l) = \text{med} \left\{ M^s_j(u, v) : (u, v) \in N(k, l) \right\}
\]

where the neighborhood \(N(k, l)\) of a TF-point \((k, l)\) is

\[
N(k, l) := \left\{ (u', v') : \max \left\{ |u' - k|, |v' - l| \right\} \leq 5 \land \min \left\{ |u' - k|, |v' - l| \right\} = 0 \right\}.
\]

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 (constant time) or formant frequency ridges (constant frequency). The effect of the median filtering can be observed in Figure 2b where most of the "noise" has been successfully removed.

2.5. Mel-spectrographic mask conversion

It is known that the human ear resolves frequencies by grouping several adjacent frequency channels into so-called critical bands. For speech recognition purposes the high STFT frequency resolution is usually converted down to a perceptual frequency scale [7]. A common approximation of this non-linear frequency resolution is the mel-frequency scale

\[
f(f) = 2595 \log_{10} \left( 1 + \frac{f}{700} \right)
\]

where \(f\) denotes the linear frequency in Hz and \(\lambda\) 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 [7]. The triangular filters

\[
\lambda_b(l) = \begin{cases} 0 & \omega_0 < \omega_{c(b-1)} \\ \omega_{c(b-1)} - \omega_{c(b-1)} & \omega_{c(b-1)} \leq \omega_0 \leq \omega_{c(b)} \\ \omega_{c(b)} - \omega_{c(b)} & \omega_{c(b)} \leq \omega_0 \leq \omega_{c(b+1)} \\ 0 & \omega_0 > \omega_{c(b+1)} \end{cases}
\]

are equally spaced along the mel-frequency scale via

\[
f_c = f_0 + b \cdot \frac{f_0 - f_1}{B + 1} \quad b = 1, \ldots, B
\]

with

\[
\omega_{c(b)} = 2\pi \cdot 700 \left( 10^{b_0/2595} - 1 \right)
\]

where \(B\) is the number of mel-frequency channels and \(f_0, f_1\) are the lower and higher cut-offs of the mel-frequency axis. We apply for each mel-frequency subband \(b\) the triangular mel-weighting function \(\lambda_b\) to obtain a soft mel-frequency mask

\[
M^s_j(k, l) = \frac{\sum_b \lambda_b(l) \tilde{M}^s_j(k, l)}{\sum_b \lambda_b(l)}
\]

shown in Figure 2c,d). Analogous to (8) we define the oracle mask on the mel-frequency scale (Figure 2f) as

\[
O_j(k, l) = \begin{cases} 1, & \text{if } 20 \log_{10} \left( \frac{|\mathcal{S}_j(k, l)|}{\sum_{i \neq j} |\mathcal{S}_i(k, l)|} \right) \geq 0 \\ 0, & \text{otherwise.} \end{cases}
\]

where the mel-frequency spectrum \(\mathcal{S}_j\) of a source \(S_j\) is given as

\[
\mathcal{S}_j(k, l) = \sum_b \lambda_b(l) |S_j(k, l)|.
\]
Δ determines the mask for the delta coefficients of the target source (black areas) from a mixture of three speakers. The corresponding mask vector of the target speaker where the likelihood of the delta coefficients \( M(k) = \{50, 100, 150, 200, 250\} \) for each channel \( r \) is calculated as

\[
M_r(k) = \sum_{\theta=1}^{\Theta} \left( o_{k+r,\theta} - o_{k-r,\theta} \right)^2 + \sum_{\theta=1}^{\Theta} \sigma_r^2 \tag{22}
\]

with mean \( \mu_r \) and standard deviation \( \sigma_r \). Note that in (21) statistical independence of the \( o_k \) is assumed. The value of the missing data mask \( m_{nk} \) basically weights the present and missing data contributions with a soft "probability" between 0 and 1. While the integral in (21) can be evaluated over \( [0, \infty] \), \( \forall i = 1, \ldots, n \), no bounds on delta features were utilized, thus we set \( (o_{low}, o_{high}) = (-\infty, \infty) \) and neglected the normalization with \( (o_{high} - o_{low}) \), \( \forall i = 1, \ldots, n \).

4. Experimental evaluation

4.1. Setup

The proposed system was evaluated via connected digit experiments on the TI-DIGIT database with a sample frequency of 20 kHz. The training set for the recognizer 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’, ‘0’, ‘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 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. The second model set was used for the MD recognizer and used spectral rather than cepstral features. 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 (18) and (19). 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. The signal-to-interferer ratio (SIR) for each masker was 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). Two microphones were positioned in the center of the room, 1 m above the ground, with an inter-element distance of \( d_{mic} = 1.72 \) cm to guarantee accurate phase parameter estimates [1]. Figure 3 shows the setup for the multi-speaker scenario with up to six different speech maskers (three male, three female) placed at a distance \( d_{phant} = 1 \) m to the microphones. After the source separation the MD recognizer was informed of the detected histogram peaks with the true mixing parameters.

![Figure 3: Anechoic room layout.](image-url)
4.2. Results

In a first experiment the recognition performance for the target speaker was recorded for the MD recognizer using the oracle mask $\Omega$, the estimated median filtered mask $\mathcal{M}$ and the MFCC baseline system. The performance in clean conditions (zero maskers) was 99.16 % for the MFCC baseline and 98.54 % for the spectral feature based MD system using the unity mask. The number of simultaneously active speech sources was increased by successively adding one masker after another according to the order shown in Figure 3. The obtained results for both hard [3] and soft masks are compared with the oracle mask and the MFCC baseline in Figure 4.

![Figure 4: Recognition accuracy for the target speaker depending on the number of simultaneously active speech maskers.](image)

As expected, the oracle mask performed best marking the upper performance bound for the MD-ASR system while the MFCC baseline represented the lower bound using no spatial information. The soft mask closely matched the performance of the oracle mask for up to three simultaneously active speech maskers before starting to fall behind. However, the soft mask clearly outperformed the hard mask for all but the two speaker case where both masks achieved equal performance. To underline the importance of the mask postprocessing the performance with and without the proposed 2-D median filtering was measured in a second experiment for three simultaneous speech maskers (Table 1). In order to eliminate the effect of the histogram peak detection the true mixture parameters were directly passed to the mask construction and no source localization was performed.

Table 1: Recognition results in terms of HTK correctness (COR) and accuracy (ACC) score as well as number of insertions (INS), deletions (DEL) and substitutions (SUB) for soft and hard masks with and without median filtering.

<table>
<thead>
<tr>
<th>Mask type</th>
<th>Median filter</th>
<th>COR in %</th>
<th>ACC in %</th>
<th>DEL</th>
<th>SUB</th>
<th>INS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hard</td>
<td>no</td>
<td>85.49</td>
<td>63.88</td>
<td>13</td>
<td>126</td>
<td>207</td>
</tr>
<tr>
<td>Soft</td>
<td>no</td>
<td>89.87</td>
<td>73.17</td>
<td>15</td>
<td>82</td>
<td>160</td>
</tr>
<tr>
<td>Hard</td>
<td>yes</td>
<td>93.95</td>
<td>87.27</td>
<td>14</td>
<td>44</td>
<td>64</td>
</tr>
<tr>
<td>Soft</td>
<td>yes</td>
<td>95.62</td>
<td>94.36</td>
<td>12</td>
<td>30</td>
<td>12</td>
</tr>
</tbody>
</table>

Clearly, if no median filtering is applied the recognized digit hypotheses contained a high number of insertion and substitution errors. Over 60 % of the observed insertions were caused by the digit model “oh”. Both the insertion and substitution errors were dramatically reduced by the 2-D median filtering which resulted in an improved recognition accuracy score.

4.3. Discussion

The proposed method depends on a number of parameters that have to be set prior to the application of the algorithm. The decay parameter $\zeta$ is used here to control the softness at the cluster boundaries. Ideally, the parameter should be related to the cluster size and shape. It might therefore be worthwhile to estimate not only the cluster center but also a covariance matrix to better characterize the cluster distribution which would eventually lead to Gaussian mixture model learning [8]. Although the proposed 2-D median smoothing was successful in reducing isolated points in the TF-masks the filter was applied as a postprocessing step only. Other more sophisticated methods that incorporate neighborhood information directly into the mask assignment or the clustering itself might be more appropriate. In particular, Markov Random Fields [9] have been quite successful in the field of image processing but tend to be more complex and demanding in terms of computational resources. The advantage of the proposed postprocessing 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 from the target signal. Finally, we have to acknowledge that the results were obtained under ideal conditions that met most of the DUET requirements. In particular the noise-free and anechoic environment can be considered as a strong simplification of real acoustic scenes and it is expected that under more realistic conditions the source localization using instantaneous DUET estimates will degrade or fail. Future work is required to make the estimators robust in hostile environments.

5. Conclusion

This paper has presented a novel method to estimate soft mel-spectrographic masks using a BSS-DUET front-end. Our experiments confirmed previous reports suggesting that soft masks outperform hard masks. For up to three simultaneous speech maskers the soft masks closely matched the recognition performance of the oracle mask designed with a priori knowledge of the source spectra. We further demonstrated that the inclusion of neighborhood information via simple 2-D median filtering can remove isolated TF-points that cause insertion and substitution errors during ASR decoding. We conclude that more work is needed for an adequate comparison with related approaches and to test the system on real data. Our future work will concentrate on extending the source separation to handle room reverberation and investigate the influence of background noise.

6. References