A Two Stage Mask Estimation Approach to Robust Speaker Verification

Yali Zhao, Lei Xie and Zhonghua Fu

Shaanxi Provincial Key Laboratory of Speech and Image Information Processing,
School of Computer Science, Northwestern Polytechnical University, Xi'an, China
ylzhao@nwpu-aslp.org, {lxie, mailfzh}@nwpu.edu.cn

Abstract

We propose a two-stage mask estimation approach to robust speaker verification (SV) in noise environments. We consider a practical semi-blind SV scenario: the location of the target speaker is fixed while the locations of all interferers are unknown. In the first stage, we use a dual-microphone and a semi-blind degenerate unmixing estimation technique (DUET) to estimate an initial binary mask. In the second stage, we refine the mask based on the time and frequency histograms of the initial mask. As a result, only highly reliable time-frequency components in the spectro-temporal features are kept for downstream verification. Experiments show that the proposed approach is superior to a baseline MFCC approach and a recent local SNR based mask estimation approach.

Index Terms: speaker verification, missing feature theory, dual-microphone, binary mask estimation

1. Introduction

Whereas speaker recognition systems can perform quite robustly in clean acoustic conditions, their recognition performance severely degrades in the presence of background noise. Recently, missing feature theory (MFT) [1] has demonstrated great potential for improving the noise robustness. According to the theory, recognition is performed only on the *reliable* parts of the spectro-temporal feature space and other *unreliable* parts contaminated by background noise are discarded. Therefore, accurate estimation of a so-called *binary mask* (that decides the reliable and unreliable parts) is essential to the success of an MFT-based recognizer. The mask can be estimated using various criteria, e.g., local SNR criterion [2] and auditory/perceptual criterion [3, 4]. The local SNR methods offer simplicity by direct estimation of the noise spectra from the contaminated speech signal. However, their performance remains poor in non-stationary noise environments, especially when the interferers are speech from others. Auditory approaches are able to use spatial information to assist reliability labeling and are thus more effective in adverse conditions. Roman *et. al.* [3] use binaural spatial cues to estimate the binary mask, assuming that the locations of all sound sources, including the target and the interferers, are known.

In this paper, we present a two-stage mask estimation approach for robust speaker verification (SV). Firstly, we use a dual-microphone and a semi-blind degenerate unmixing estimation technique (DUET) [5] to estimate an initial binary mask. Different from Roman's approach [3], our approach supposes that only the location of the target speaker is fixed while the locations of all interferers are unknown. We believe that this configuration is much closer to a real-world SV application. For example, in an access control application, users usually locate at a relatively fixed area while interferers remain unknown. Secondly, we perform mask refinement by reliable components selection based on the time and frequency histograms of the initial mask. As a result, only the highly reliable time-frequency (T-F) components in the spectral features are kept for downstream speaker verification. Experiments demonstrate that the proposed approach is superior to a baseline MFCC approach and a recent local SNR based mask estimation approach [2].

2. System Overview

Contaminated speech signal is first picked up by a dual microphone. Short-time Fourier transform (STFT) is then used to obtain time-frequency (T-F) representation, i.e., $X_i(t,f)$, of the recorded signal. Here, i, t and f denote microphone index ($i = 1, 2$), time frame and frequency bin index, respectively. Then in each frame, $X_i(t,f)$ is accumulated within K frequency subbands to obtain a K-dimension speaker spectral feature vector $\mathbf{x}_t$. We also use $X_i(t,f)$ to extract a spatial feature vector $O(t,f)$ and to estimate a binary mask $M_b(t,k)$, as described in Section 3. As introduced in Section 4, the mask is further refined to $M_b^r(t,k)$. Finally, the binary mask is applied to $\mathbf{x}_t$ to identify the reliable components $\mathbf{x}_t^r$ that are used for downstream speaker verification.

We approximate the speaker-dependent distribution of spectral features by Gaussian mixture models (GMMs). In the speaker verification phase, we use *marginalization* [2] to deal with the missing features. The reliable feature sub-vector $\mathbf{x}_t^r$ is used to estimate the likelihood of the speaker identityλ. The probability density becomes

$$p(\mathbf{x_t}|\lambda) = \sum_{m=1}^{M} w_m \prod_{x_{ti} \in \mathbf{x}_t^r} p(x_{ti}|\mu_{mi}, \sigma_{mi}^2), \quad (1)$$

where w_m is the weight of the mth Gaussian mixture, x_{ti} refers to the ith component in $\mathbf{x}_t^r$, μ_{mi} and σ_{mi}^2 are their corresponding mean and variance vectors in the mth Gaussian mixture, respectively.

A well-trained universal background model (UBM) λ_{ubm} is used to normalize the decision score. Finally, the decision likelihood is

$$P(\lambda_c|\mathbf{X}) = \frac{1}{T}\sum_{t=1}^{T} log\frac{p(\mathbf{x}_t|\lambda_c)}{p(\mathbf{x}_t|\lambda_{ubm})} \quad (2)$$

where λ_c denotes the claimed speaker model and $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_1, \cdots, \mathbf{x}_T)$ is the speaker feature set for a test utterance.

3. Initial Binary Mask Estimation

We use normalized inter-microphone phase differences (IPD) and inter-microphone amplitude differences (IAD) to represent the spatial difference of the two signals picked up by the dual-microphone. For each T-F unit $X_i(t,f)$, the spatial information can be represented by a 3-dimension spatial feature vector $O(t,f)$ as [5]

$$\{\frac{arg[X_2(f,t)/X_1(f,t)]}{2\pi f dc^{-1}}, \frac{[X_1(f,t)]}{A(f,t)}, \frac{[X_2(f,t)]}{A(f,t)}\} \quad (3)$$

where $A(f,t) = \sqrt{\sum_i |X_i(f,t)|^2}$, d represents the distance between the two microphones, and c denotes the sound speed.

Since the location of the target is fixed, a target spatial GMM model Ω_{tar} can be trained off-line using the EM algorithm. We use Gaussian white noise to train Ω_{tar} due to its wideband and uniform density. In online speaker verification when the target speech and the interfering noise are co-present, the spatial distribution will change because of the presence of new sound sources. As the location of the noise is unknown, we use online GMM adaptation to estimate the corrupted spatial model Ω_{in}, by several EM iterations of the target spatial model Ω_{tar}. Hence, the binary mask is obtained by likelihood comparison:

$$M(t,f) = \begin{cases} 1, & p(\Omega_{tar}|O(t,f)) > p(\Omega_{in}|O(t,f)) \\ 0, & otherwise. \end{cases} \quad (4)$$

Then we transform the estimated mask from frequency bins into frequency subbands by

$$M_b(t,k) = I(\sum_{M(t,f)\in B_k} M(t,f)), \quad (5)$$

where B_k represents the kth subband, $I(\cdot)$ denotes an indicator function, i.e., $I(x) = 1$ if $x > \delta$ and $I(x) = 0$ otherwise. δ is set empirically to one third of the number of bins within each subband.

4. Mask Refinement

The initial binary mask estimated from acoustic and spatial features contains many errors due to reverberation and misclassification. It is known that speech is sparse in the T-F domain, which means its energy is concentrated only in a few T-F regions. Hence, those isolated bins with weak power are possible unreliable. On the other hand, for some types of color noise, their energy is concentrated continuously in several subbands, which makes the surviving bins within those subbands are also unreliable. Therefore, we refine the initial mask by keeping those T-F portions with high confidence.

4.1. Refining Matrix

Firstly, two histograms are produced from the estimated initial binary mask. The frequency histogram $H_k(t)$ represents the number of '1's in $M_b(t,k)$ of frame t:

$$H_k(t) = \sum_{k=1}^{K} M_b(t,k). \quad (6)$$

The time histogram $H_t(k)$ represents the number of '1's in $M_b(t,k)$ of each subband k. We normalize the time histogram by the frame number:

$$H_t(k) = \frac{1}{T}\sum_{t=1}^{T} M_b(t,k). \quad (7)$$

Figure 1 shows an example of the mask refinement. From Figure 1(a), we can clearly see that both the time and frequency histogram curves are unevenly distributed.

To keep only the high confidence regions, we define a refining matrix:

$$R(t,k) = \begin{cases} 1, & H_k(t) > \theta_{H_k} \ \& \ H_t(k) > \theta_{H_t} \\ 0, & otherwise \end{cases} \quad (8)$$

where θ_{H_k} and θ_{H_t} represent the decision thresholds in frequency and time domain, respectively. The refined

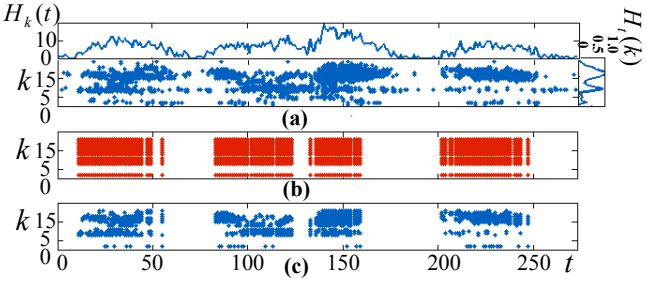

Figure 1: *An example showing mask refinement for a speech utterance. (a) Initial mask $M_b(t,k)$ and time and frequency histograms – $H_t(k)$ and $H_k(t)$; (b) the refining matrix $R(t,k)$; (c) the refined mask $M_b^r(t,k)$.*

mask can be achieved as follows:

$$M_b^r(t,k) = M_b(t,k) \cdot R(t,k), \tag{9}$$

where '$\cdot$' is the binary AND operator. The refined binary mask $M_b^r(t,k)$ is used for missing data recognition. Figure 1(b) and (c) show the refining matrix and the final mask.

4.2. Decision Thresholds

For decision of θ_{H_t}, we want to preserve the bands with high reliability. Hence for an utterance, $H_t(k)$ is sorted in descending order and θ_{H_t} is set so that the components with N highest values are preserved. For decision of θ_{H_k}, with the same purpose, we adopt a sliding window as does in the minimum statistics [6] in noise reduction. We set θ_{H_k} dynamically by moving average of $H_k(t)$ within the sliding window $[t-L, t)$:

$$\theta_{H_k}(t) = \frac{1}{L} \sum_{i=t-L}^{t-1} H_k(t) \tag{10}$$

where L is the length of the window.

5. Experiments

5.1. Experiment Setup

We record the experimental data in a quite room as shown in Fig. 2. Two omni-directional microphones are deployed in parallel with 4cm spacing. Thirteen high-fidelity loudspeakers are located at 13 different positions to simulate a target speaker and 12 interfering sources. The clean speech from AURORA-2 corpus [7] is played through the loudspeaker at the target position. The noise is played through the loudspeaker at the 12 interfering positions. Noise type includes female speech, pure music, car noise and white noise. These noise signals are played at the 12 positions with the same average sound level, respectively, and are mixed with the target speech signal at different SNR levels.

We use the AURORA-2 "TRAIN" set for UBM training. The clean speech from all 110 speakers in the "TRAIN" set is played by the loudspeaker at the target position and recorded by the dual microphone. We use the speech recorded by Microphone L to train the UBM with 64 Gaussian components. All 104 speakers from the "TESTA" set (52 males and 52 females) are used as the target speakers for SV experiments. For each speaker, 30 of the 36 utterances are used to obtain his/her target model by adapting the UBM and the rest 6 are used for testing. We run 6 rounds of SV test at each SNR. At each round, the testing utterances (104×6) are played at the target position and mixed with noise randomly chosen from one interfering position. The SV evaluation results are averaged over the 6 rounds.

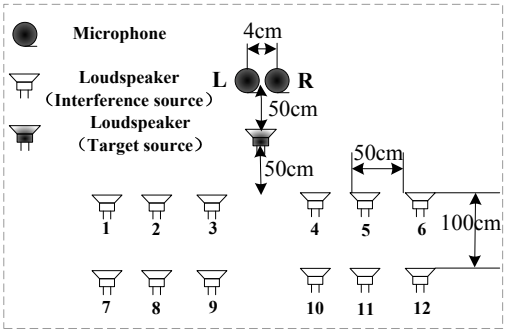

Figure 2: *Experimental data collection in a quite room.*

The speech signal is 8KHz, 16bit and the speech frame is 20ms with 10ms shift in feature extraction. The length of STFT is 256. We use 24 subbands (i.e.,K=24) equally distributed in linear frequency scale [8] and the raw speaker acoustic feature vector consists of 24 conventional linear frequency log power spectra components.

For comparison purposes, we build a baseline SV system using a 39-dimensional feature vector consisting of 13 static MFCC coefficients, including energy, and first and second order temporal derivatives. Cepstral mean normalization (CMN) is applied for improving robustness. We also compare our approach with another missing data recognizer that uses a recent mask estimation approach, namely MCMM [2]. In order to see the upper bound performance, we run a recognizer using an ideal binary mask (IBM). The IBM is estimated regarding that the prior recordings of the target speech and the interfering noise are known. As a sanity check, we also implement our mask refinement method to MCMM and IBM. In the experiments, N is 21, and L is 40.

5.2. Results

The experimental results in terms of equal error rate (EER) are summarized in Table 1. It is expected that the MCMM method shows inferior performance as compared with the MFCC baseline in the presence of speech interferer. This is because MCMM uses local SNR criterion that fails to perform in non-stationary noise environments. We notice that the proposed mask estimation methods (M_b and M_b^r) show substantial performance gain as compared with MCMM in the presence of speech interferer, pure music and car noise under most SNRs. However, in the white noise corrupted environment, the performance of our method is worse than that of MCMM. This is because the white noise is wideband and uniform density that affect all the frequency bands of speech, resulting in more estimation errors in the binary mask. Moreover, we can clearly see the effectiveness of mask refinement (M_b^r). It can bring EER reduction to MCMM, M_b and IBM in most noise conditions. The refined mask M_b^r only under-performs in speech interferer conditions. This may be explained as follows. The interfering speech has the same property of sparseness as the target speech.

Table 1: *Experimental results in terms of EERs.*

Noise		Methods						
Type	SNR (dB)	MFCC baseline	MCMM [2]	MCMM+M_b^r	Proposed M_b	Proposed M_b^r	IBM	IBM+M_b^r
female speech	5	10.76	13.73	11.67	5.99	6.62	2.92	3.78
	10	6.84	8.08	7.88	5.35	5.24	2.36	2.83
	15	5.10	6.3	6.94	5.2	5.13	2.2	2.83
pure music	5	23.01	21.46	20.77	14.51	13.88	8.51	7.51
	10	17.67	11.67	11.04	9.53	8.83	5.67	4.73
	15	10.44	7.88	7.17	6.76	6.15	3.78	3.15
car noise	5	32.82	14.82	14.45	17.66	17.01	7.57	6.94
	10	24.17	10.24	9.85	9.56	8.98	6.32	5.34
	15	13.05	7.58	7.24	6.62	6.09	3.78	3.06
white noise	5	40.71	36.08	35.77	47.88	47.31	15.47	14.56
	10	33.67	19.87	19.35	36.59	35.63	15.15	13.51
	15	21.89	11.35	10.86	15.77	15.15	8.3	7.82

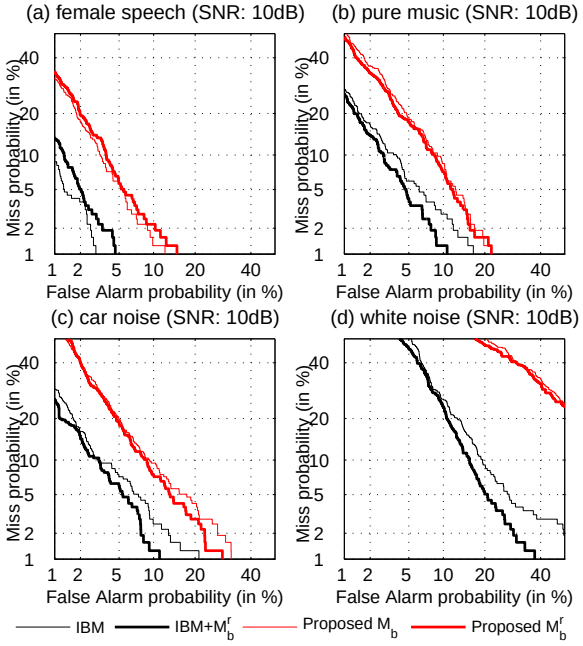

Figure 3: *DET curves before and after mask refinement.*

As a result, mask refinement produces more serious errors that result in weak discrimination of reliable speaker feature components. The DET curves in Figure 3 also show mask refinement achieves performance improvement in music, car noise and white noise conditions.

6. Conclusions

We have proposed a two-stage mask estimation approach to MFT-based robust speaker verification. The proposed approach first estimates a raw mask by a semi-blind DUET method under the dual-microphone setup. A mask refinement strategy is used to preserve the highly reliable T-F components in the spectral features for missing data recognition. Experiments have demonstrated the effectiveness of the proposed approach. In future work, we plan to use a soft masking decision strategy in missing data recognition.

7. Acknowledgements

This work was supported by the National Natural Science Foundation of China (61175018, 60901077), the Natural Science Basic Research Plan of Shaanxi Province (2011JM8009), the Key Science and Technology Program of Shaanxi Province (2011KJXX29), the Doctoral Program of Higher Education in China (20096102120044) and Fok Ying Tung Education Foundation (131059).

8. References

[1] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Communication*, vol. 34(3), pp. 267–285, 2001.

[2] T. May, S. van de Par, and A. Kohlrausch, "Noise-robust speaker recognition combining missing data techniques and universal background modeling," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 20(1), pp. 108–121, 2011.

[3] N. Roman and D. Wang, "Speech segregation based on sound localization," *J. Acoust. Soc. Am*, vol. 114, pp. 2236–2252, 2003.

[4] S. Harding, J. Barker, and G. Brown, "Mask estimation for missing data speech recognition based on statistics of binaural interaction," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14(1), pp. 58–67, 2006.

[5] Z.-H. Fu, L. Xie, and D.-M. Jiang, "Dual-microphone noise reduction based on semi-blind DUET," in *Proc. of ICSLP*, Taiwan, 2010.

[6] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 9(5), pp. 504–512, 2001.

[7] H. Hirsch and D.Pearce, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. of ISCA ASR2000 Workshop*, Beijing, China, 2000, pp. 181–188.

[8] X. Lu and J. Dang, "An investigation of dependencies between frequency components and speaker characteristics for text-independent speaker identification," *Speech Communication*, vol. 50(4), pp. 312–322, 2008.