# ROBUST SPEAKER RECOGNITION USING BINARY TIME-FREQUENCY MASKS

*Yang Shao* and *DeLiang Wang*

Department of Computer Science and Engineering
& Center for Cognitive Science
The Ohio State University
Columbus, OH 43210-1277, USA
{shaoy, dwang}@cse.ohio-state.edu

## ABSTRACT

Conventional speaker recognition systems perform poorly under noisy conditions. In this paper, we evaluate binary time-frequency masks for robust speaker recognition. An ideal binary mask is *a priori* defined as a binary matrix where 1 indicates that the target is stronger than the interference within the corresponding time-frequency unit and 0 indicates otherwise. We perform speaker identification and verification using a missing data recognizer under cochannel and other noise conditions, and show that the ideal binary mask provides large performance gains. By employing a speech segregation system that estimates the ideal binary mask, we achieve significant improvements over alternative approaches. Our study, thus, demonstrates that the use of binary masking represents a promising direction for robust speaker recognition.

## 1. INTRODUCTION

Cochannel speech comprises speech mixtures from two talkers who, unlike conversations, are unaware of each other. Consequently, speech from both channels has large overlap, which presents a considerable challenge for automatic speaker recognition. Previously, we proposed the use of estimated pitch contours for automatic extraction of usable speech segments [12], defined as consecutive frames of speech that are dominated by one speaker [7]. Furthermore, a model-based sequential grouping method organizes the segments into corresponding streams using available speaker models. It achieves speaker identification (SID) performance close to that when the usable segments are assigned ideally.

However, some of the usable frames still contain sound energy from both speakers. Instead of extracting usable speech at the frame level, it may be desirable to identify usable speech at the level of time-frequency (T-F) units so that recognition performance could be further improved under noisy conditions. For such purposes, we study binary T-F masks for robust speaker recognition in this paper.

To use binary time-frequency masks for recognition, we employ a missing data method [4] for robust SID and speaker verification (SV) tasks. The basic idea is to treat the noise-dominant T-F units as missing during recognition. To apply a missing data recognizer requires a binary mask to provide information about whether a specific T-F unit is reliable or missing. Drygajlo and El-Maliki proposed a robust SV method by combining spectral subtraction and missing data recognition [4], [5]. Spectral subtraction is used to obtain the needed binary mask. Their method works well for stationary noise, but it degrades severely in nonstationary noise conditions, such as cochannel speech.

In this paper, we evaluate binary T-F masks for robust speaker recognition using the missing data method. Given premixing recordings, the ideal binary mask is easily conducted which labels a T-F unit as reliable if it contains more energy from target than interference, and labels it as unreliable otherwise [13]. We find that the ideal binary mask achieves substantial performance gains under cochannel and additive noise conditions. We also employ a voiced speech segregation system to estimate the ideal binary mask [6]. Our results are compared with those using binary masks estimated by spectral subtraction for conditions where target speech is corrupted by cocktail party noise or rock music. The comparison shows that our estimated binary masks perform significantly better.

The rest of the paper is organized as follows. Section 2 describes the notion of the ideal binary mask. Ideal mask estimation using speech segregation is presented in Section 3. Section 4 describes our missing data recognizer. Evaluation results are given in Section 5. Section 6 concludes the paper.

## 2. IDEAL BINARY TIME-FREQUENCY MASK

A two-dimensional time-frequency representation is widely used in speech processing. Within this representation, the binary T-F mask furnishes the information about whether a T-F unit is reliable or not. The ideal binary T-F mask [13] is a binary matrix, defined *a priori* as follows:

$$M(f,t) = \begin{cases} 1, & \text{if } S(f,t) > N(f,t) \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

$M(f,t)$ is the T-F mask indexed by frequency $f$ and time $t$. $S(f,t)$ is the energy from the target source in the frequency channel centered at $f$ and frame $t$; $N(f,t)$ is the corresponding energy from the interfering source. If a T-F unit contains stronger energy from target than interference, the corresponding mask element is labeled 1; it is assigned 0 otherwise. This implies a local signal-to-noise ratio (SNR) criterion of 0 dB. Given premixing target and interfering signals, the ideal binary mask can be readily constructed.

The ideal binary mask is directly motivated by the human auditory masking phenomenon [8], and it has many desirable properties. The ideal binary mask provides the maximum SNR gain of all the binary masks [6]. Moreover, such masks have been

applied to robust speech recognition and shown to be highly effective as a front-end [3], [10]. Besides, depending on what the target is, the ideal binary mask can be constructed differently. Under cochannel conditions, the binary values of a T-F unit in the mask correspond to the two underlying speakers in the mixture. If one speaker is of interest to the user, it can be designated as target, and the other speaker will be the interferer. If both speakers are desired, after selecting one as the target, the other speaker could be identified through the complement mask. Under non-speech noisy conditions, the speech signal to be recognized can be regarded as target, and the ideal binary mask can be defined accordingly.

## 3. ESTIMATION OF IDEAL BINARY MASK

Construction of the ideal binary mask requires prior recordings of target and interferer sources. To estimate the ideal binary mask, we adapt and apply a pitch-based speech segregation system [6].

This system produces a binary T-F mask after four stages of processing. In the first stage, the input signal is transformed into T-F domain by passing through a Gammatone filterbank in consecutive time frames. Here a T-F unit corresponds to the response of a frequency channel within a time frame. Then the system extracts temporal and frequency features of the units.

In the second stage, the T-F units are merged into segments using the extracted features. The units that establish similar responses are supposedly from the same source and therefore grouped together. Thus, the resulting segments are deemed to be homogeneous.

In the third stage, the units are labeled target or background according to the estimated target pitch period. (see Section 5.2) Specifically, for low-frequency channels where harmonics are resolved, if a unit shows similar response at the target pitch period, the corresponding T-F unit is labeled as target-dominant; it is labeled background otherwise. For high-frequency channels that respond to several harmonics, an amplitude modulation model is used to determine whether a unit response shows beating at the target pitch period and thus considered as target-dominant.

In the final stage, the segments from the second stage are further grouped into target or background streams according to the unit labels. Target units are labeled 1 in the mask, and background units are labeled 0. Figure 1 gives an illustration of a mask estimate from noisy speech.

## 4. MISSING DATA RECOGNITION

To utilize the binary masks for robust speaker recognition, we employ a missing data method in the T-F domain.

In a typical speaker identification or verification system, the probability distribution of an extracted feature vector, $\boldsymbol{x}$, produced by a speaker, $\lambda$, is modeled as a Gaussian mixture model (GMM) [9]. GMM is a weighted linear combination of $M$ unimodal multivariate Gaussian densities, typically parameterized with diagonal covariance matrices [3]. Given a binary mask showing whether a feature component is reliable or missing, the feature vector can be split into reliable components, $\boldsymbol{x_r}$, or unreliable ones, $\boldsymbol{x_u}$, and its probability density becomes,

$$p(\boldsymbol{x} \mid \lambda) = \sum_{k=1}^{M} w_k \prod_{x_i \in \boldsymbol{x_r}} p(x_i \mid \mu_{ki}, \sigma_{ki}^2) \prod_{x_j \in \boldsymbol{x_u}} p(x_j \mid \mu_{kj}, \sigma_{kj}^2). \qquad (2)$$

$w_k$ is the weight of the $k$th Gaussian mixture. $x_i$ and $x_j$ refer to a reliable and unreliable feature component in $\boldsymbol{x}$, respectively; $\mu_{k.}$ and $\sigma_{k.}^2$ are their corresponding means and variances in the $k$th Gaussian mixture.

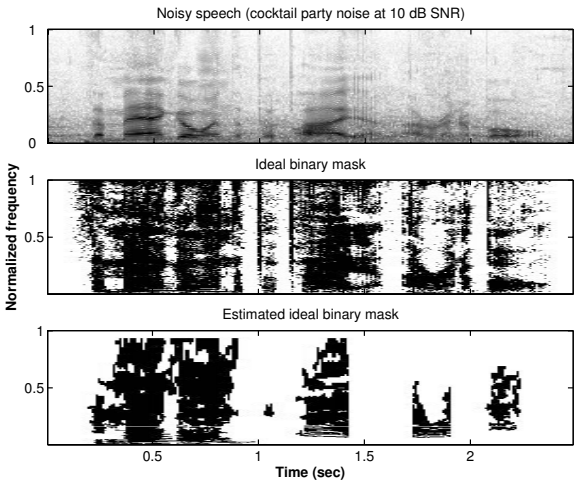


Figure 1. Illustrations of noisy speech, its ideal binary mask and estimated mask. Noisy speech is created by mixing clean speech and cocktail party noise at 10 dB.

The first likelihood term on the right-hand side of eq. (2) can be easily obtained from training since the features are considered reliable (clean). However, the second likelihood term is hard to compute because the feature component is regarded as missing and its distribution is unknown. There are two approaches to handle this situation, marginalization and imputation [3]. The latter seeks to impute the missing features and replace them with estimates. The former reduces the distribution by integrating over the missing components. Imputation increases computational complexity but does not necessarily produce better verification results [5]. In this paper, we use marginalization and compute the overall likelihood as,

$$p(\boldsymbol{x} \mid \lambda) = \sum_{k=1}^{M} w_k \prod_{x_i \in \boldsymbol{x_r}} p(x_i \mid \mu_{ki}, \sigma_{ki}^2). \qquad (3)$$

The likelihood of a noisy utterance given a specific speaker model is computed as the likelihood product of feature vectors of individual frames. For SID tasks, the speaker model that gives the maximum likelihood value is selected as the identified speaker. For SV tasks, we use universal background models (UBM) for score normalization. This missing data method, i.e. eq. (3), can be naturally extended to include the UBM.

## 5. EVALUATION

### 5.1 Cochannel SID evaluation

This experiment demonstrates the SID performance when the noise source is a speaker. To have a consistent comparison with previous studies [7], [11], [12], we use the same evaluation data from the TIMIT corpus. Specifically, the speaker set consists of 38 speakers from the "DR1" dialect region, 14 of which are female and the rest are male. For each speaker, 5 out of 10 utterances are randomly selected for speaker model training and the remaining 5 are used for testing purpose. The training data for a speaker averages 10 *sec* of clean speech, and 16-mixture GMMs are trained using the EM algorithm [9].

Cochannel speech is simulated by mixing the utterances in the testing corpus. When one speaker is deemed as target, every other

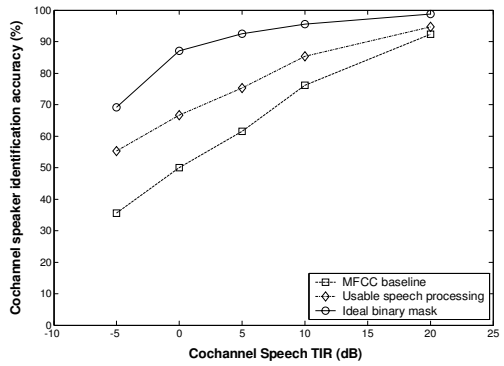


Figure 2. Speaker identification performance under cochannel conditions. The square line shows the performance when MFCCs are used. The diamond line shows the results of extracted usable speech segments after they are *a priori* assigned. The circle line gives performance achieved by the ideal binary mask using the missing data method.

speaker could be considered as interferer. For each pair, 1 out of 5 test files is randomly selected from the target and mixed with randomly selected files from the interferer. The utterances are aligned in length and mixed at target-to-interferer ratio (TIR) of -5 dB, 0 dB, 5 dB, 10 dB and 20 dB respectively. TIR is calculated as the ratio of target speech energy to interfering speech energy. Hence for each TIR level, 1406 cochannel mixtures are created.

Figure 2 presents the results of this experiment. As a baseline, we extract 12 mel-frequency cepstral coefficients (MFCCs) and their first-order derivatives as the feature vector. To compare with usable speech processing, we apply the usable speech extraction method and ideally assign the extracted segments into the target stream using *a priori* pitch information [11]. The same type of MFCCs is derived and identification is performed on the target stream. To evaluate the binary masks, we implement the missing data recognizer with 255-coefficient DFT feature vectors. Specifically, vectors are extracted from the log-compressed power spectrum of 20 ms frames with 10 ms overlap. The frames are extracted by applying a running Hamming window on the signal.

It can be observed from the figure that the ideal binary mask performs significantly better than the usable speech method, which in turn is much better than the baseline performance. This is to be expected since the ideal binary mask provides reliable/unreliable information at a finer level than usable speech, and the T-F redundancy facilitates identification when features are partially missing. Evaluation of the estimated binary mask is not performed in this task because it is hard to determine target pitch contours for the speech segregation system under cochannel conditions.

### 5.2 SID evaluation in noisy background

In this experiment, we demonstrate the effectiveness of binary masks in adverse environments when the intrusion source is not a speaker. Two types of noise are selected from a noise database collected by Cooke [2], cocktail party noise and rock music. Both are wide-band and non-stationary, containing significant energy below 2 kHz. It is also observed that both noises have some harmonic structure because the cocktail party noise contains speech-like sounds and the rock music contains musical instruments. The noisy speech utterances are simulated by mixing all the test files with the selected noises at -5 dB, 0 dB, 5 dB, 10 dB and 20 dB SNRs. For each pair of SNR and noise, 190 mixtures are created for testing.

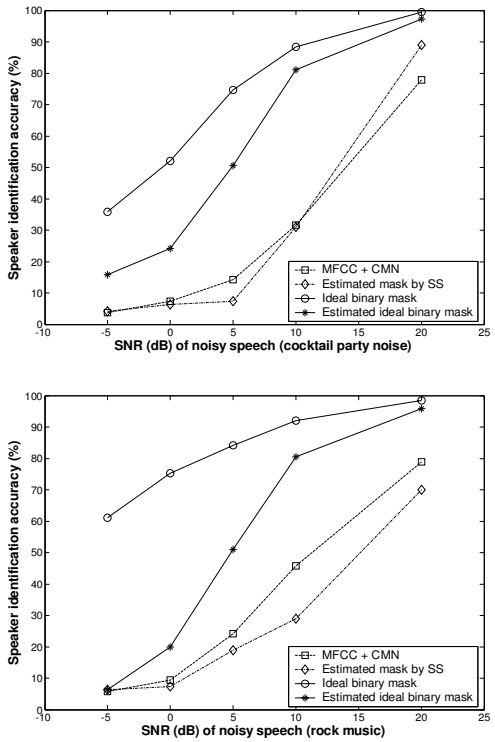


Figure 3. Speaker identification performance under noisy conditions. The top plot shows the results for cocktail party noise, and the bottom one for rock music. The square line represents baseline results of the GMM recognizer using cepstral mean normalized (CMN) MFCCs. The diamond line shows the missing data recognition results using binary masks estimated by spectral subtraction (SS). The circle line gives performance achieved by the ideal binary mask. The star line shows the results of the estimated ideal binary mask.

Figure 3 shows the SID results for both noise conditions at various SNRs. The baseline system uses MFCCs and their first-order derivatives. Cepstral mean normalization (CMN) is applied for robustness. We also employ spectral subtraction to estimate binary mask for the missing data recognizer as proposed by Drygajlo and El-Maliki [4]. Specifically, the average noise spectrum is estimated from the initial 10 frames of the mixture, and subtracted from each subsequent mixture spectrum. If the resulting component is greater than the noise estimate, the corresponding mask element is labeled 1 and 0 otherwise. The implied 0 dB SNR criterion is preferred over the negative energy criterion because it produces better results [3].

To estimate the ideal binary mask, target pitch contours are determined by applying the widely-used Praat toolkit [1] on the noisy speech. Please note that an estimated mask is obtained using the auditory filterbank that models human's auditory response and it has large overlaps between neighboring filters. Directly using the filterbank energy gives SID accuracy of 94.2% on clean speech, which is significantly lower than that using the DFT coefficients, 99.5%. Thus, we transform the estimated mask from Gammatone frequency bands into DFT domain by labeling the corresponding frequency bins. Subsequently, the same missing data recognizer is used as in the previous experiment.

It can be observed from the figure that the estimated binary mask performs significantly better than the baseline system using MFCC-CMN. As both noises are non-stationary, spectral subtraction is unable to provide a good mask estimate, and its performance degrades sharply with decreasing SNR. The ideal binary mask produces best performance. Since this is a preliminary study, the performance gap between the ideal binary mask and estimated mask leaves much room for improvement by adopting the binary mask approach.

### 5.3 SV evaluation in noisy background

In a similar configuration as the preceding experiment, we evaluate binary masks for speaker verification tasks. Here, only the mixtures with the cocktail-party noise are tested on the 38-speaker set. One mixture file contributes 1 true score for the target speaker and 37 imposter scores for the other speakers in the set. For each SNR, there are 190 true scores and 7030 imposter scores. The scores are normalized using a UBM of 4096 mixtures, which is trained from the entire TIMIT training set, excluding the above 38 speakers.

Evaluation results are given in Figure 4. The ideal binary mask yields substantial performance gains over the baseline in the entire range of SNR levels. The estimated mask achieves significant improvement from 10 dB to -5 dB. It under-performs only at the 20 dB condition largely due to the segregation strategy that attempts to reconstruct the target signal by grouping harmonic components [6]. Consequently, inharmonic target components are removed even when interference is very weak.

## 6. CONCLUSION

We have evaluated the utility of the ideal binary time-frequency mask for robust speaker recognition. Our evaluation under cochannel and noisy conditions shows that the ideal binary mask produces superior performance. We have also employed a speech segregation system that estimates the ideal binary mask. The resulting system produces significant performance gains over alternative approaches.

**Acknowledgements**. This research was supported in part by an AFOSR grant (FA9550-04-1-0117) and an AFRL grant (FA8750-04-1-0093). We thank G. Hu for his assistance in speech segregation.

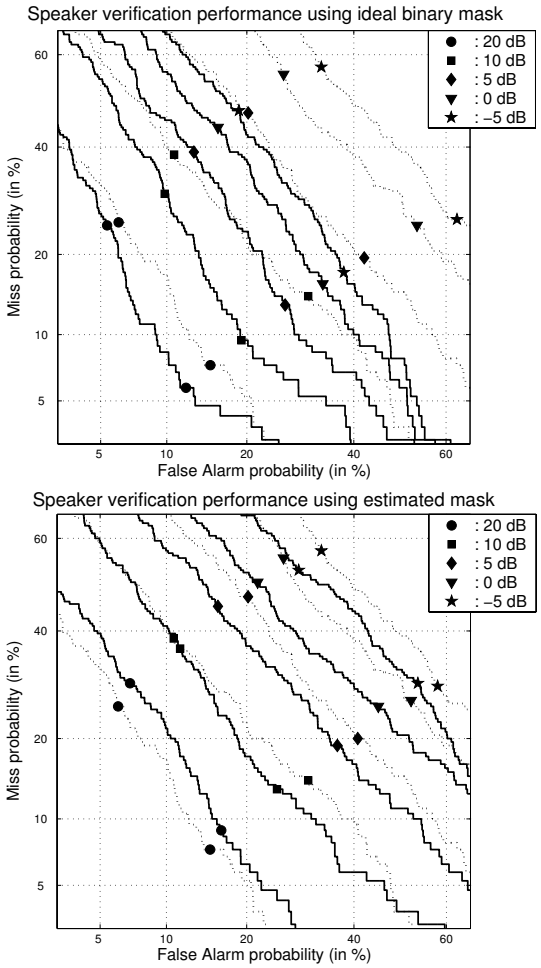


Figure 4. Speaker verification performance under cocktail party noise. The top plot shows the results for the ideal binary mask, plotted in solid curves against MFCC baseline in dotted curves. The bottom one shows performance of the estimated binary mask in solid curves against the same baseline in dotted curves.

## 7. REFERENCE

[1] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot International*, (5:9/10), pp. 341-345, 2001.

[2] M.P. Cooke, *Modelling auditory processing and organization.* Cambridge U.K.: Cambridge University Press, 1993.

[3] M.P. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Comm.*, vol. 34, pp. 267-385, 2001.

[4] A. Drygajlo and M. El-Maliki, "Speaker verification in noisy environments with combined spectral subtraction and missing feature theory," in *Proc. ICASSP*, pp. 121-124, 1998.

[5] A. Drygajlo and M. El-Maliki, "Integration and imputation methods for unreliable feature compensation in GMM based speaker verification," in *Proc. 2001: A Speaker Odyssey - The Speaker Recognition Workshop*, pp. 107-112, 2001.

[6] G. Hu and D.L. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Trans. Neural Net.*, vol. 15, pp. 1135-1150, 2004.

[7] J.M. Lovekin, R.E. Yantorno, K.R. Krishnamachari, D.S. Benincasa, and S.J. Wenndt, "Developing usable speech criteria for speaker identification," in *Proc. ICASSP*, pp. 421-424, 2001.

[8] B.C.J. Moore, *An introduction to the psychology of hearing.* 5th ed., San Diego: Academic, 2003.

[9] D.A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Comm.*, vol. 17, pp. 91-108, 1995.

[10] N. Roman, D.L. Wang, and G.J. Brown, "Speech segregation based on sound localization," *J. Acoust. Soc. Am.*, vol. 114, pp. 2236-2252, 2003.

[11] Y. Shao and D.L. Wang, "Co-channel speaker identification using usable speech extraction based on multi-pitch tracking," in *Proc. ICASSP*, vol. 2, pp. 205-208, 2003.

[12] Y. Shao and D.L. Wang, "Model-based sequential organization in cochannel speech," *IEEE Trans. Audio Speech and Language Process.*, vol. 14(1), pp. 289-298, 2006.

[13] D.L. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines,* P. Divenyi, Ed., Norwell MA: Kluwer Academic, pp. 181-197, 2005.