BINAURAL SOUND SEGREGATION FOR MULTISOURCE REVERBERANT ENVIRONMENTS

Nicoleta Roman and *DeLiang Wang*

Department of Computer and Information
Science and Center for Cognitive Science
The Ohio State University
Columbus, OH 43210, USA
{niki,dwang}@cis.ohio-state.edu

ABSTRACT

We present a novel method for binaural sound segregation from acoustic mixtures contaminated by both multiple interferences and reverberation. We employ the notion of an ideal time-frequency binary mask, which selects the target if it is stronger than the interference in a local time-frequency (T-F) unit. As opposed to classical adaptive filtering which focuses on the suppression of noise, our model employs an adaptive filter that performs target cancellation. T-F units dominated by target are largely suppressed at the output of the cancellation unit when compared to units dominated by noise. Consequently, the actual input-to-output attenuation level in each T-F unit is used to estimate an ideal binary mask. A systematic evaluation in terms of automatic speech recognition performance shows that the resulting system produces masks close to ideal binary ones.

1. INTRODUCTION

Human listeners are able to segregate and recognize a sound signal from a background of acoustic interference even under adverse conditions. The *auditory scene analysis* (ASA) process described by Bregman involves grouping elements that are likely to have originated from the same source into a perceptual structure called an auditory stream. Our focus is on computational approaches to ASA that exploit the information to detect time-frequency (T-F) units that preserve reliable information for the target sound. Therefore, the computational goal is to estimate an *ideal* binary mask which selects T-F units where the target energy dominates [1]. Such ideal binary masks have been shown to be effective front-ends for robust automatic speech recognition [2][3] and to provide speech intelligibility improvements for normal listeners under noisy conditions [3].

Location-based segregation algorithms have reported very good results for multi-talker scenarios under anechoic conditions [2], [4], [5]. The underlying assumption in those systems is that time delays and attenuation levels are reliable cues for sound segregation which show location-based characteristic clustering. In the reverberant conditions, anechoic modeling of time delayed and attenuated mixtures is inadequate. Reverberation introduces potentially an infinite number of sources due to reflections of sound sources against the surfaces of an enclosure. As a result, the location cues estimated in individual T-F units become unreliable when reverberation increases and the system performance degrades under realistic conditions. A notable exception is the system proposed by Palomaki et al. that combines both binaural-based grouping and reverberation masking and have shown improved speech recognition results in reverberant conditions [6].

Other approaches to sound separation include the classical two-microphone adaptive beamformers which can cancel almost perfectly one interference under optimal conditions. The performance, however, degrades rapidly when the number of sources and the reverberation level increase. A subband adaptive scheme has been proposed by Liu et al. [7] to address the multi-source problem. Their two-microphone system exploits the location information to steer independent nulls that suppress the strongest interference in each time-frequency unit. However, the underlying signal model is still anechoic and performance gradually degrades in echoic conditions. As an extension to Liu's method, one could potentially switch adaptive filters to obtain better suppression in individual T-F units, except that it requires exact knowledge for each individual interference.

The model proposed here is motivated by the need to find a two-microphone solution to target segregation in real world conditions when the acoustic scene is contaminated by both strong reverberation and the presence of multiple concurrent sound sources. We propose a method for sound segregation centered on target cancellation through adaptive filtering. We observe a correlation between the amount of cancellation produced in individual T-F units and the relative strength between target and interference. Consequently, a threshold on the input-to-output attenuation level is employed to estimate an ideal binary mask. Related work includes the system proposed by Alvarez et al. [8], which combines a first-order differential beamformer to obtain a noise estimate by suppressing one of two sources and spectral subtraction to simultaneously enhance the sound sources.

The rest of the paper is organized as follows. The next section defines the problem and contains a detailed presentation of the model. Section 3 gives simulation results and the last section concludes this paper.

Proc. ICASSP2004, Volume 2, Pages 373-376, May 17-21, Montreal

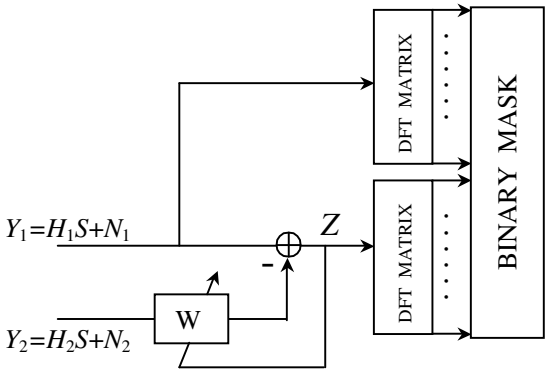

Figure 1. Schematic diagram of the model proposed. The input signal is a mixture of reverberant target sound and acoustic interference. At the core of the system is an adaptive filter for target cancellation. The output of the system is an estimation of the ideal binary mask.

2. MODEL ARCHITECTURE

The proposed model consists of two stages as shown in Fig. 1. In the first stage, the system performs target cancellation through adaptive filtering. In the second stage, the system labels as 1 those T-F units that have been largely attenuated in the first stage since those T-F units are likely to have originated from the target location.

The model of the input signal as shown in Fig. 1, assumes that a desired speech source S has been uttered in a reverberant enclosure and recorded by two microphones to produce the signal pair (X_1, X_2). It is assumed that the transmission path from the target location to the microphones is a linear system and is modeled as ($X_1=H_1S$, $X_2=H_2S$). In this problem formulation, the challenge arises when an unwanted interference pair is also present at the input of the microphones (N_1, N_2). The interference here is a combination of multiple reverberant sources and diffuse background noise. The target is assumed fixed but there are no restrictions imposed on the number, location or content of the interfering sources. In realistic conditions, the interference can suddenly change its location and may contain short impulsive sounds. Under these conditions, it is hard to localize each individual source in the scene. The goal is therefore to remove the noise part and recover the reverberant target speech based only on the spatial information of the target source. Dereverberation can be potentially employed as a post-processing stage of our model in order to recover the speech source S [9].

The objective here is to develop an efficient mechanism for estimating an ideal binary mask, which selects the T-F units where the *a priori* local SNR exceeds a 0-dB threshold. The relative strength between the target signal and interference for a T-F unit (ω, t) is defined as:

$$R(\omega,t)=\frac{|X_1(\omega,t)|}{|X_1(\omega,t)|+|N_1(\omega,t)|} \quad (1)$$

where $X_1(\omega,t)$ and $N_1(\omega,t)$ are the spectral estimates for signal and noise, respectively, at microphone 1 (assume microphone 1 as the primary microphone). Thus, a T-F unit is set 1 in the ideal binary mask if the relative strength exceeds 0.5, otherwise it is set 0.

In the classical noise cancellation configuration, the filter learns to identify the differential acoustic transfer function of a particular noise source. This kind of system is unable to cope well under multiple concurrent noise sources and diffuse background noise. As an alternative, we propose to use the adaptive filter to produce target cancellation and denote the module as the "target cancellation module" (TCM). In the experiments reported here, we assume a fixed target location and the system is trained in the absence of interference on calibration sequences of white noise of 10 s duration. We implement the adaptation using the Fast-Block LMS algorithm with an impulse response of 120 ms length (6000 samples at 44.1 kHz sampling rate) [10]. The signals are resynthesized and a shorter time-frequency analysis is applied to both the TCM output $Z(\omega, t)$ and the input at the primary microphone $Y_1(\omega, t)$. The time-frequency resolution is 40-ms time frames with a 20-ms frame shift, and 512 DFT outputs. Frames are extracted by applying a running Hamming window to the signal.

As a measure of signal suppression at the output of the TCM unit, we define the output-to-input energy ratio as follows:

$$OIR(\omega,t)=\frac{|Z(\omega,t)|^2}{|Y_1(\omega,t)|^2} \quad (2)$$

Consider a T-F unit for which the noise is zero. Ideally, the TCM module cancels perfectly the target source resulting in a zero output and therefore $OIR(\omega,t)\to 0$. On the other hand, T-F units dominated by noise are not suppressed by the TCM and thus $OIR(\omega,t)>>0$. Thus, a simple binary decision can be implemented by imposing a threshold on the estimated output-to-input energy ratio. The estimated binary mask is 1 in those T-F units where $OIR(\omega,t)>\theta(\omega)$, otherwise is 0.

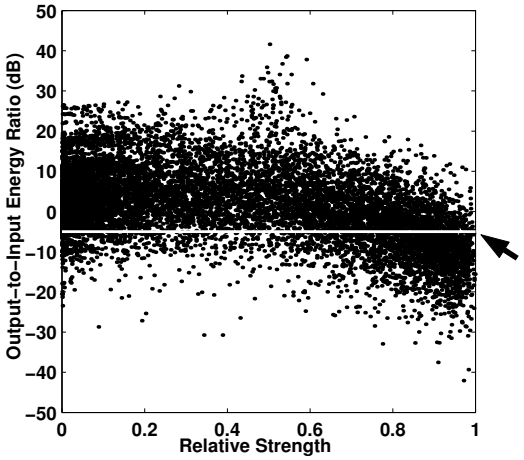

Figure 2. Scatter plot of the estimated relative strength and output-to-input attenuation for a frequency bin with center frequency of 1 kHz. The white dotted line corresponds to the –6 dB threshold used in the binary mask estimation.

Figure 2 shows a scatter plot of the estimates of relative strength *R* and output-to-input ratios *OIR* obtained for individual T-F units corresponding to a frequency bin of 1 kHz. The results are extracted from 100 mixtures of target speech mixed in the four speaker noise condition at 5 dB SNR. Observe that there exists a correlation between the amount of cancellation in the individual T-F units and the relative strength between target and interference. Throughout this paper, we have used a fixed threshold of –6 dB for the output-to-input energy ratio as suggested by our informal listening experiments. As indicated in the figure this threshold achieves almost complete noise removal at the expense of some speech energy loss. This threshold can be optimized for a particular application and SNR presentation.

3. RESULTS

We have evaluated the system on a binaural stimulus, simulated using a room acoustic model as described in [6]. The reflection paths of a sound source are obtained using the image reverberation model for a small rectangular room (6m×6m×3m). The resulting impulse response is convolved with the measured head related impulse responses (HRIR) from a KEMAR dummy head in order to produce two binaural inputs to the system [11]. The position of the listener was fixed asymmetrically at (2.5m×2.5m×2m) to avoid obtaining near identical impulse responses at the two microphones. All sound sources are presented at different angles at a distance of 1.5 m from the listener. For all our tests, target is fixed at 0° azimuth. Two noise configurations are tested: 1) an interference of rock music at 45° (condition NC1) and 2) four concurrent speakers (two female and two male utterances) at azimuth angles of -135°, -45°, 45° and 135° (condition NC2). To make the two situations more comparable, the interference containing the four superimposed speakers has been edited and the initial and last speech pauses deleted. In all our tests, the input SNR corresponds to the global SNR using the reverberated target speech as the signal.

An example of speech target extraction is presented in Fig. 3. The interference consists of four simultaneous speakers at a 0 dB SNR presentation. As seen from the figure, the extracted signal is very similar to the original one. Informal listening shows that the system filters out the acoustical interference and preserves a highly intelligible target signal.

We perform an SNR evaluation for the two conditions using 10 speech signals from the TIMIT database as target, and results are given in Table I and Table II. In order to completely assess the system performance, both the output SNR and the retained speech ratio (RSR) are computed as follows:

$$Output\ \ SNR = 10\log 10\left(\sum_t s^2(t) \Big/ \sum_t n^2(t)\right) \tag{3}$$

$$RSR = \sum_t s^2(t) \Big/ \sum_t o^2(t) \tag{4}$$

where $o(t)$ corresponds to the original target signal, $s(t)$ the target signal resynthesized through the estimated mask and $n(t)$ the noise signal resynthesized through the same mask. Results show SNR improvements in the range of 7-15 dB while preserving much of the target energy (70-90%). Observe that the system performance degrades at lower SNR values because of the increased overlap between the target signal and the noise. The speech energy loss may be improved by imposing a higher threshold on the output-to-input attenuation level at the expense of increasing the residual noise.

Table I: SNR evaluation for a one-source interference.

MIXTURE SNR	**-5**	**0**	**5**	**10**
OUTPUT SNR	10.5	14.5	17.0	20.32
RSR(%)	77	83	88	90

Table II: SNR evaluation for a four-speaker interference.

MIXTURE SNR	**-5**	**0**	**5**	**10**
OUTPUT SNR	2.7	8.5	13.2	17.6
RSR(%)	57	75	84	89

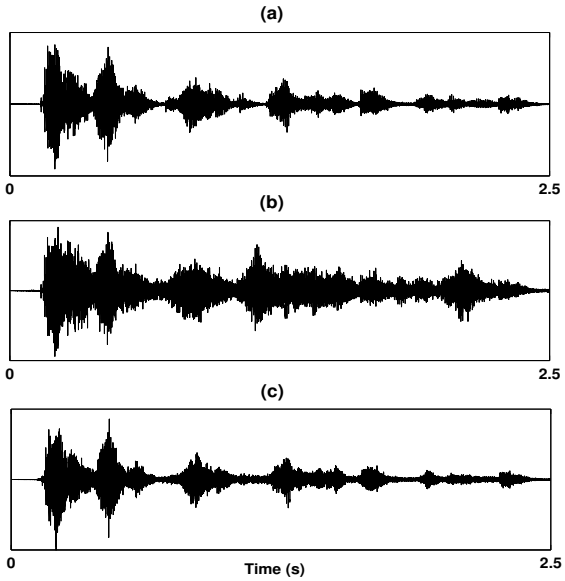

Figure 3: (a) Target speech recorded at the primary microphone. (b) Mixture of target speech presented at 0° and four interfering speakers at locations -135°, -45°, 45° and 135° at the primary microphone. SNR presentation level of 0dB. (c) Segregated target signal.

We also evaluate the performance of our system in a speech recognition task using the missing-data technique as described in [3]. In this approach, a hidden Markov model recognizer is modified such that only those acoustic features indicated as reliable in a binary mask are used during decoding. Hence, it works seamlessly with the output from our speech segregation system. We have implemented the missing data algorithm with a 512 coefficient DFT feature vector. More specifically, each feature vector is extracted by computing the log-compressed energy in frames of 20 ms with 10 ms overlap. Frames are extracted by applying a running Hamming window to the signal.

We use the bounded marginalization method for classification [3]. The task domain is recognition of connected digits, and both training and testing are performed on acoustic features from the right ear signal using the male speaker dataset in the TIDigits database.

Figure 4 shows the speech recognition results based on the binary masks estimated by our system for condition NC1 (Fig. 4A) and condition NC2 (Fig. 4B). For all tests, the same male target speaker is located at 0°. Both training and testing of the system are performed on acoustic features from the right ear signal. The performance of our model is compared against the ideal masks systematically for five SNR levels: –5 dB, 0 dB, 5 dB, 10 dB and 20 dB. Also shown in the figure is the baseline performance where the recognition is conducted on the unprocessed mixtures from the right ear. As seen previously in the SNR evaluation, the system estimates well the ideal binary mask for positive SNR levels but the estimates degrade gradually at lower SNR levels. However, observe that large improvements over baseline performance are obtained across all conditions. Also, the system produces a significant improvement compared with the system by Palomaki et al., especially at low SNRs. For example, a word error reduction of 50% is obtained at 0 dB SNR. This shows the strong potential of applying our model for robust speech recognition.

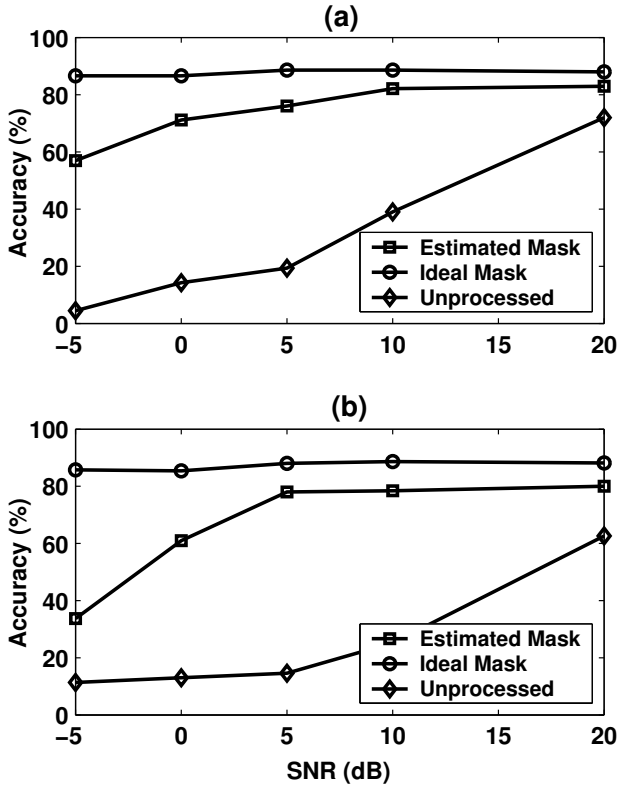

Figure 4. Recognition performance at different SNR values for original mixture (diamond), ideal binary mask (circle) and estimated mask (square). (a) Accuracy score for NC1 condition. (b) Accuracy score for NC2 condition.

4. CONCLUSION

We have presented a novel two-microphone sound segregation system that performs well under realistic conditions. The system can be applied to spatial configurations with multiple interfering sources and strong reverberation. We have also carried out experiments in a real office setting and found similar performance with the simulations presented here. Our approach is based on target cancellation through adaptive filtering followed by an analysis of the output-to-input attenuation level in individual T-F units. Therefore, there are no restrictions imposed on the number and location of the interfering sources and moving interferences are easily accommodated by the system. The output of the system is an estimate of a binary mask which labels the T-F components of the acoustic scene dominated by the target sound. A systematic evaluation using an automatic speech recognizer shows that the resulting system produces masks close to the ideal ones in a variety of conditions.

Acknowledgements. This research was supported in part by an NSF grant (IIS-0081058) and an AFOSR grant (F49620-01-0027).

5. REFERENCES

[1] G. Hu and D. Wang, "Speech segregation based on pitch tracking and amplitude modulation," *Proc. WASPAA*, pp. 79-82, 2001.

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

[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. 267-285, 2001.

[4] A. Jourjine, S. Rickard and O. Yilmaz, "Blind separation of disjoint orthogonal signals: demixing N sources from 2 mixtures," *Proc. IEEE ICASSP*, vol. 5, pp. 2985-2988, 2000.

[5] H. Glotin, F. Berthommier and E. Tessier, "A CASA-labelling model using the localisation cue for robust cocktail-party speech recognition," *Proc. EUROSPEECH*, 1999.

[6] K. J. Palomaki, G. J. Brown and D. L. Wang, "A binaural processor for missing data speech recognition in the presence of noise and small-room reverberation," *Speech Comm.*, to appear, 2004.

[7] C. Liu *et al.*, "A two-microphone dual delay-line approach for extraction of a speech sound in the presence of multiple interferers," *J. Acoust. Soc. Am.,* vol. 110, pp. 3218-3230, 2001.

[8] A. Alvarez *et al.*, "Speech enhancement and source separation supported by negative beamforming filtering," *Proc. ICSP*, pp. 342-345, 2002.

[9] M. Wu, "Pitch tracking and speech enhancement in noisy and reverberant environments," Ph.D. Dissertation, Department of Computer and Information Science, The Ohio State University, 2003.

[10] S. Haykin. *Adaptive Filter theory*, 4th ed., Prentice Hall, 2002.

[11] W. G. Gardner and K. D. Martin, "HRTF measurements of a KEMAR dummy-head microphone," *MIT Media Lab Perceptual Computing Technical Report* #280, 1994.