SPEECH RECOGNITION IN MULTISOURCE REVERBERANT ENVIRONMENTS WITH BINAURAL INPUTS

Nicoleta Roman[1], *Soundararajan Srinivasan*[2] *and DeLiang Wang*[3]

[1]Department of Mathematics, Statistics and Computer Science
The Ohio State University at Lima, Lima, OH, 45804, USA

[2]Biomedical Engineering Center
[3]Department of Computer Science and Engineering and Center for Cognitive Science
The Ohio State University, Columbus, OH, 43210, USA
{niki, srinivso, dwang}@cse.ohio-state.edu

ABSTRACT

We present a binaural solution to robust speech recognition in multi-source reverberant environments. 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. Our system estimates this ideal binary mask at the output of a target cancellation module implemented using adaptive filtering. This mask is used in conjunction with a missing-data algorithm to decode the target utterance. A systematic evaluation in terms of automatic speech recognition (ASR) performance shows substantial improvements over the baseline performance and better results over related two-microphone approaches.

1. INTRODUCTION

A typical auditory environment contains multiple concurrent sources that are also reflected by surfaces and may change their locations constantly. While human listeners are able to segregate and recognize a target signal under such adverse conditions, ASR remains a challenging problem [1]. ASR systems are trained on clean speech and face the problem of mismatch when tested in noisy and reverberant conditions. In this paper we address the problem of recognizing target speech from multi-source reverberant binaural recordings.

Microphone array processing techniques which enhance the target speech have been employed to improve the robustness of ASR systems in noisy environments [2]. These techniques are divided in two broad categories: beamforming and independent component analysis (ICA) [3]. To separate multiple sound sources, beamforming takes advantage of their different directions of arrival while ICA relies on their statistical independence. A fixed beamformer, such as that of the delay-and-sum, constructs a spatial beam to enhance signals arriving from the target direction independent of the interfering sources. A large number of microphones are however required in order to impose a constant beam shape across frequencies [3]. Adaptive beamforming techniques, on the other hand, attempt to null out the interfering sources in the mixture [4] [5]. While an adaptive beamformer with two microphones is optimal for canceling a single directional interference, additional microphones are required as the number of noise sources increases. Similarly, the drawbacks of ICA techniques include the requirement that the number of microphones be greater than or equal to the number of sources and poor performance in reverberant conditions [5]. Some recent sparse representations attempt to relax the former assumption but the performance is limited [6]. While the above techniques enhance target speech independently of the recognizer, Seltzer et al. optimize an adaptive filter based on recognition results [7].

Inspired by the robustness of the human auditory system, research in computational auditory scene analysis (CASA) has been devoted to build speech separation systems that incorporate known principles of auditory perception [8]. In particular, binaural CASA systems which utilize location information have shown very good recognition results in anechoic conditions. Reverberation, however, introduces potentially an infinite number of sources due to reflections from hard surfaces. As a result, the estimation of location cues in individual T-F units becomes unreliable and the performance of location-based segregation systems degrades. A notable exception is the binaural system proposed by Palomäki et al. [9] which includes an inhibition mechanism that emphasizes the onset portions of the signal and groups them according to common location. The system shows improved speech recognition results across a range of reverberation times with a single interference.

From an information processing perspective, the notion of an ideal T-F binary mask has been proposed as the computational goal of CASA [10]. Such a mask can be constructed from *a priori* knowledge of target and interference; specifically a value of 1 in the mask indicates that the target is stronger than the interference within a particular T-F unit and 0 indicates otherwise. Previously, we have proposed a binaural system that is capable of estimating the ideal binary mask under multi-source reverberant conditions [11] and reported results using a missing-data recognizer [12] trained on reverberant speech. Note that the missing-data recognizer treats the units labeled 1 in the mask as reliable data and the others as unreliable during recognition. To avoid using a different model for each reverberant condition, it is desirable to train the ASR on anechoic data. However, we find that the performance of the missing-data recognizer degrades considerably when obtained using anechoic training.

In this paper, we propose an alternate approach using a speech prior based spectrogram reconstruction technique [13]. In this technique, the target speech values in the unreliable T-F units are estimated by conditioning on the reliable ones. We observe that the reliable units in the mask correspond to regions in the spectrogram dominated by relatively clean target speech. Hence, the prior

speech model used in the reconstruction can be also trained using anechoic data. We show that the proposed system provides substantial improvement in speech recognition accuracy over baseline and other related two-microphone approaches.

The rest of the paper is organized as follows. The next section gives the details of our proposed binaural system for robust recognition. Section 3 gives evaluation results and comparisons with related two-microphone approaches and the last section concludes the paper.

2. MODEL DESCRIPTION

As described in the introduction, in the classical adaptive beamforming approach the filter learns to identify the differential acoustic transfer function of a particular noise source and thus perfectly cancels only one directional noise source. Systems of this type, however, are unable to cope well with multiple noise sources or diffuse background noise. As an alternative, we have proposed to use the adaptive filter only for target cancellation and then process the noise reference obtained using a nonlinear scheme in order to obtain an estimate of the ideal binary mask [11]. Specifically, we observe that the attenuation in a T-F unit due to target cancellation is systematically correlated with the relative strength between target and interference. Hence, the system estimates the ideal binary mask by imposing a threshold on the output-to-input energy ratio in each T-F unit [11]. In this work, we use a T-F decomposition consisting of 10 ms time frames with 256 DFT coefficients. The target is assumed to be fixed and the filter in the target cancellation module is trained in the absence of interference. However, no restrictions are imposed on the number, location, or content of the interfering sources.

Figure 1 demonstrates the performance of our segregation system for a mixture of a target male utterance at 0° location and four interfering speakers at -135°, -45°, 45°, 135°. The room conditions are reverberation time T_{60}=0.3 s and 5 dB input SNR. Observe that the estimated mask is able to estimate well the ideal binary mask especially in the target dominant high-energy T-F regions and to entirely suppress the multi-source interference. This highlights the capacity of our system to produce good segregation results.

Although subjective listening tests have shown that the signal reconstructed from the ideal binary mask is highly intelligible, the extraction of cepstral features for input to ASR systems from a signal reconstructed using such a mask is distorted due to the mismatch arising from the T-F units labeled 0, which smears the entire cepstrum via the cepstral transform [12]. One way to handle this problem is by estimating the original target spectral values in the T-F units labeled 0 using a prior speech model. This approach has been suggested by Raj et al. in the context of additive noise [13]. In this approach, a noisy log spectral energy vector Y at a particular frame is partitioned in its reliable Y_r and its unreliable Y_u components. The task is to reconstruct the underlying true spectral energy vector X. Assuming that the reliable features Y_r are approximating well the true ones X_r, a Bayesian decision is then employed to estimate the remaining X_u given only the reliable component. Hence, this approach works seamlessly with the T-F binary mask that our speech segregation system produces. Here, the reliable features are the T-F units labeled 1 in the mask while the unreliable features are the ones labeled 0. As seen in Fig. 1, the reliable units in the mask are relatively clean at moderate levels of reverberation. Hence, we train the prior speech model on anechoic

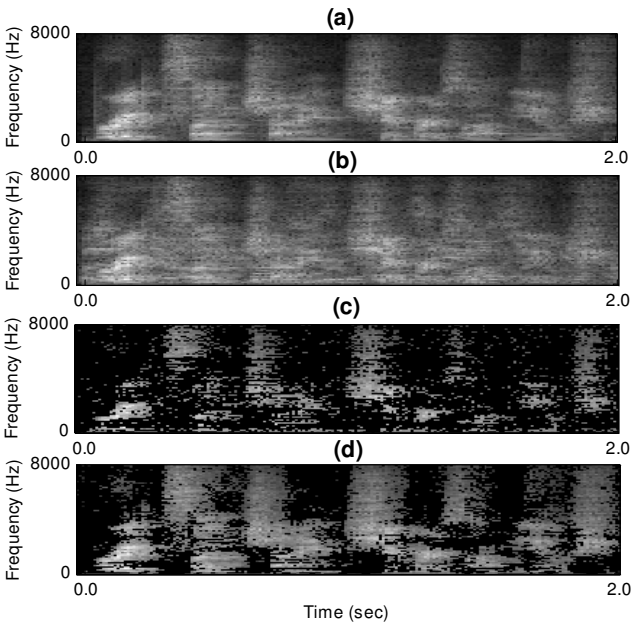

Figure 1. Comparison between the estimated mask and the ideal binary mask for a five-source configuration. (a) Reverberant target speech. (b) Reverberant mixture. (c) The mixture spectrogram overlaid by the estimated T-F binary mask. (d) The mixture spectrogram overlaid by the ideal binary mask. The recordings correspond to the left ear microphone.

data and thus avoid obtaining a prior for each deployment condition which is desirable for robust speech recognition.

The speech prior is modeled as a mixture of Gaussians:

$$p(X) = \sum_{k=1}^{M} p(k)\, p(X \mid k), \tag{1}$$

where M=1024 is the number of mixtures, k is the mixture index, $p(k)$ is the mixture weight and $p(X \mid k) = N(X; \mu_k, \Sigma_k)$.

Previous studies ([12], [13]) have shown that a good estimate of X_u is its expected value conditioned on X_r:

$$E_{X_u \mid X_r, 0 \le X_u \le Y_u}(X_u) = \sum_{k=1}^{M} p(k \mid X_r, 0 \le X_u \le Y_u) \cdot \int_0^{Y_u} X_u\, p(X_u \mid k,\ 0 \le X_u \le Y_u)\, dX_u, \tag{2}$$

where $p(k \mid X_r)$ is the *a posteriori* probability of the k'th Gaussian given the reliable data and the integral provides the expected value of the unreliable component X_u given the k'th mixture. Note that under the additive noise condition, the unreliable parts may be constrained as $0 \le X_u \le Y_u$ [12]. In our implementation, we have assumed that the prior can be modeled using a mixture of Gaussians with diagonal covariance. Theoretically, this is a good approximation if an adequate number of mixtures are used [12]. Additionally, our empirical evaluations have shown that for the case of M=1024 this approximation results in an insignificant degradation in recognition performance compared to a full-

covariance Gaussian model while the computational cost is greatly reduced. Hence, the expected value can now be computed as:

$$\tilde{X}_u = \begin{cases} \mu_{u,k} \ , \ 0 \le \mu_{u,k} \le Y_u \\ Y_u \ , \ \mu_{u,k} > Y_u \\ 0 \ , \ \mu_{u,k} < 0 \end{cases} \quad (3)$$

The *a posteriori* probability of the k'th mixture given the reliable data is estimated using the Bayesian rule from the simplified marginal distribution $p(X_r \mid k) = N(X_r; \mu_{r,k}, \sigma_{r,k})$ obtained from $p(X \mid k)$ without utilizing any bounds on X_u. While this simplification results in a small decrease in accuracy, it results in a substantially faster computation of the marginal. The reconstructed signal using the above method is used as input in the speech recognition experiments reported below.

3. RESULTS

We have evaluated our system on binaural stimuli, simulated using the room acoustic model described in Palomäki et al. [9]. The reflection paths of a particular sound source are obtained using the image reverberation model for a small rectangular room (6m×4m×3m). The resulting impulse response is convolved with the measured head related impulse responses (HRIR) of a KEMAR dummy head [14] in order to produce the two binaural inputs to our system. 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 when the source is in the median plane. For all our tests, target is fixed at 0° azimuth unless otherwise specified. To test the robustness of the system we have performed the following two tests: 1) an interference of rock music at 45° (Scene 1); and 2) four concurrent speakers (two female and two male utterances) at azimuth angles of -135°, -45°, 45° and 135° (Scene 2). The initial and the last speech pauses in the interfering utterances have been deleted in Scene 2 to make it more comparable with Scene 1. The signals are upsampled to the HRIR sampling frequency of 44.1 kHz and convolved with the corresponding left and right ear HRIRs to simulate the individual sources. Finally, the reverberated signals at each ear are added and then downsampled to 16 kHz which is the sampling frequency used for filter adaptation in the segregation system. In all our evaluations, the input SNR is calculated at the left ear using reverberant target speech as signal.

The task domain is speaker independent recognition of connected digits. Thirteen (the numbers 1-9, a silence, very short pause between words, zero and oh) word-level models are trained using an HMM toolkit, HTK [15]. All except the short pause model have 8 emitting states. The short pause model has a single emitting state, tied to the middle state of the silence model. The output distribution in each state is modeled as a mixture of 10 Gaussians. The HMM architecture is the same as the one used in Palomäki et al. [9]. The ASR and the prior used in our reconstruction are trained on the 4235 clean signals from the male speaker training dataset in the TIDigits database, downsampled to 16 kHz to be consistent with our model. Testing is performed on a subset of the testing set containing 229 utterances from 3 speakers which is similar to the test set used in [9]. The test signals are convolved with the corresponding left and right ear target impulse responses and noise is added as described above.

The feature vectors for recognition in each frame consist of the 13 mel-frequency cepstral coefficients (MFCC) together with their first and second order temporal derivatives. Additionally, cepstral mean normalization (CMN) is applied to improve the robustness of the system under reverberant conditions. Frames are extracted using 25 ms windows with 15 ms overlap. The recognition accuracy using clean test utterances is 99%. On reverberated test utterances (T_{60} = 0.3 s), the accuracy is 94%.

Speech recognition results for the two-test conditions are reported separately in Fig. 2 and Fig. 3 for T_{60}=0.3 s at five SNR levels: –5 dB, 0 dB, 5 dB, 10 dB and 20 dB. Results are obtained using the same MFCC-based ASR as the back-end for the following approaches: fixed beamforming, adaptive beamforming, target cancellation through adaptive filtering followed by spectral subtraction, our proposed front-end ASR using the estimated mask and finally our proposed front-end ASR using the ideal binary mask. The baseline results correspond to the unprocessed left ear signal. Observe that our system achieves large improvements over the baseline performance across all conditions. Additionally, the excellent results reported for the ideal binary mask highlights the potential performance that can be obtained using this approach. As expected, the adaptive beamformer outperforms all the other algorithms in the case of a single interference (Scene 1). However, as the number of interferences increases, the performance of the adaptive beamformer degrades rapidly and approaches the performance of the fixed beamformer in the Scene 2 condition. Since the first stage of our system produces a noise estimate, alternatively we can combine our adaptive filtering stage with spectral subtraction to enhance the reverberant target signal (see also [16]). As illustrated by the recognition results in Fig. 3, this approach outperforms the adaptive beamformer in the case of multiple concurrent interferences. While spectral subtraction improves the SNR gain in target-dominant T-F units, it does not produce a good target signal estimate in noise-dominant regions. Note that our front-end ASR employs a better estimation of the spectrum in the unreliable T-F units and therefore results in large improvements over the spectral subtraction method. A similar pattern is observed when the reverberation time increases. Fig. 4 shows results for T_{60}=0.6 s in the Scene 2 condition.

We compare our system with the binaural system proposed by Palomäki et al. which was shown to produce significant recognition improvements on the same digit recognition task as used here [9]. Table 1 compares the two systems for the case of one interfering source of rock music. The recognition results for the Palomäki et al. system are the ones reported by the authors while the results for our system have been produced using the same configuration setup. Listener is located in the middle of the room while target and interfering sources are located at 20° and -20° respectively. T_{60} is 0.3 s and the input SNR is fixed before the binaural presentation of the signals at three SNR levels: 0 dB, 10 dB and 20 dB. Note that we obtain a marked improvement over the system of Palomäki et al., in the low SNR conditions. By utilizing location information only during acoustic onsets, the mask obtained by their system has a limited number of reliable units. This limits the amount of information available for recognition. This is probably the cause for the degradation in system performance at low SNRs.

4. CONCLUSION

We have proposed a binaural-based system for robust speech recognition in multi-source reverberant environments. In a

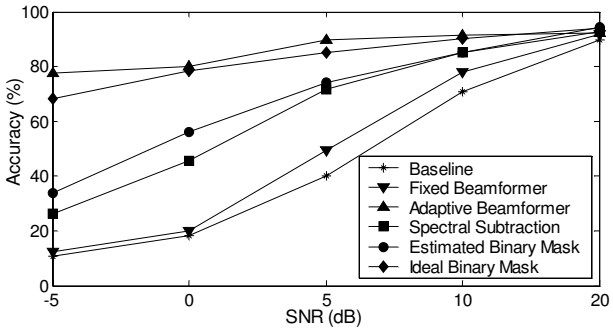

Figure 2. Recognition performance for Scene 1 at T_{60}=0.3 s and different SNR values for the reverberant mixture (∗), a fixed beamformer (▼), an adaptive beamformer (▲), a system that combines target cancellation and spectral subtraction (■), our front end ASR using the estimated binary mask (●), and our front-end ASR using the ideal binary mask (♦).

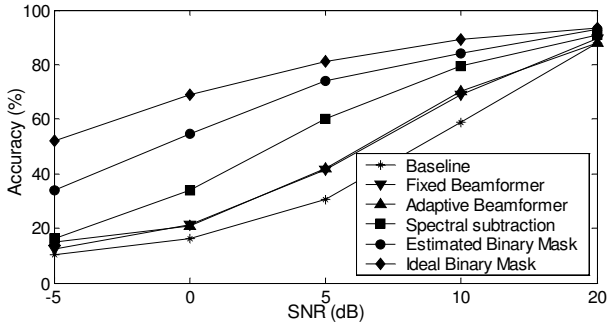

Figure 3. Recognition performance for Scene 2 at T_{60}=0.3 s and different SNR values. See Fig. 2 for notations.

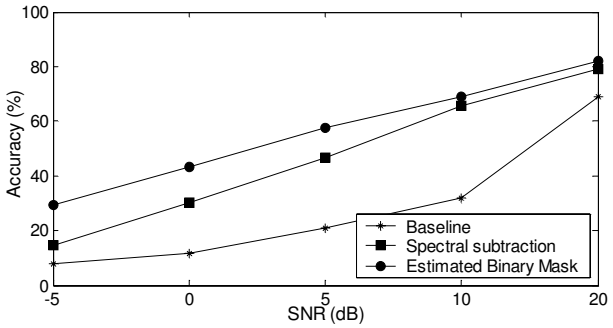

Figure 4. Recognition performance for Scene 2 at T_{60}=0.6 s and different SNR values. See Fig. 2 for notations.

systematic comparison, we have shown that the system yields substantial performance gains over baseline and related approaches. A key observation is that the segregation stage is able to preserve the high-energy target dominant regions and therefore our target reconstruction using an anechoic prior speech model performs well. In addition, CMN performed on the reconstructed target provides additional robustness to reverberation. The main advantage for our system is that the prior and ASR models are trained on clean speech and hence our algorithm is applicable for recognition in changing reverberant environments.

Table 1. Comparison with the Palomäki et al. system in terms of speech recognition accuracy (%)

Input SNR	0 dB	10 dB	20 dB
Baseline (MFCC+CMN)	13.04	43.01	81.85
Palomäki et al.	32.7	78.8	91.9
Proposed system	47.58	81.59	91.80

Acknowledgements. This research was supported in part by an AFOSR grant (FA9550-04-1-0117), an AFRL grant (FA8750-04-1-0093) and an AFRL contract via Veridian. We thank B. Raj for helpful discussions.

5. REFERENCES

[1] X. Huang, A. Acero and H-W. Hon, *Spoken Language Processing*, Upper Saddle River, NJ: Prentice Hall PTR, 2001.

[2] M. Omologo, M. Matassoni and P. Svaizer, "Speech recognition with microphone arrays," in *Microphone Arrays: Signal Processing Techniques and Application*, M. Brandstein and D. Ward, eds., Berlin: Springer, pp. 331-353, 2001.

[3] M. Brandstein and D. Ward, Eds., *Microphone Arrays: Signal Processing Techniques and Application*, Berlin: Springer, 2001.

[4] D. Van Compernolle, "Switching adaptive filters for enhancing noisy and reverberant speech from microphone array recordings," Proc. ICASSP, pp. 833-836, 1990.

[5] A. Hyväarinen, J. Karhunen and E. Oja, *Independent Component Analysis*, New York: Wiley, 2001.

[6] M. Zibulevsky, B. A. Pearlmutter, P. Bofill and P. Kisilev, "Blind source separation by sparse decomposition", in *Independent Component Analysis: Principles and Practice*, Roberts, S. J., and Everson, R.M., Eds., Cambridge University Press, 2001.

[7] M. L. Seltzer, B. Raj and R. M. Stern, "Likelihood-maximizing beamforming for robust hands-free speech recognition," IEEE Trans. Speech and Audio Proc., vol. 12, pp. 489-498, 2004.

[8] G. J. Brown and D. L. Wang, "Separation of speech by computational auditory scene analysis," in Speech Enhancement, J. Benesty, S. Makino and J. Chen, Eds. New York: Springer, pp. 371-402, 2005.

[9] 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.*, vol. 43, pp. 361-378 , 2004.

[10] 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.

[11] N. Roman and D. L. Wang, "Binaural sound segregation for multisource reverberant environments," Proc. ICASSP, vol.2, pp. 373-376, 2004.

[12] 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.

[13] B. Raj, M. L. Seltzer, R. M. Stern, "Reconstruction of missing features for robust speech recognition," Speech Comm., vol. 43, pp. 275-296, 2004.

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

[15] S. Young, D. Kershaw, J. Odell, V. Valtchev and P. Woodland, The HTK Book (for HTK Version 3.0), Microsoft Corporation, 2000.

[16] A. Álvarez, P. Gómez, V. Nieto, R. Martínez and V. Rodellar, "Speech enhancement and source separation supported by negative beamforming filtering," Proc. ICSP, pp. 342-345, 2002.