SPEECH SEPARATION WITH DEREVERBERATION-BASED PRE-PROCESSING INCORPORATING VISUAL CUES

Muhammad Salman Khan, Syed Mohsen Naqvi, and Jonathon Chambers

Advanced Signal Processing Group, School of Electronic, Electrical and Systems Engineering,
Loughborough University, Leicestershire, LE11 3TU, UK.
Email: {m.s.khan2*, s.m.r.naqvi, j.a.chambers}@lboro.ac.uk

1. INTRODUCTION

Humans are skilled in selectively extracting a single sound source in the presence of multiple simultaneous sounds. They (individuals with normal hearing) can also robustly adapt to changing acoustic environments with great ease. Need has arisen to incorporate such abilities in machines which would enable multiple application areas such as human-computer interaction, automatic speech recognition, hearing aids and hands-free telephony. This work addresses the problem of separating multiple speech sources in realistic reverberant rooms using two microphones.

Different monaural and binaural cues have previously been modeled in order to enable separation. Binaural spatial cues i.e. the interaural level difference (ILD) and the interaural phase difference (IPD) have been modeled [1] in the time-frequency (TF) domain that exploit the differences in the intensity and the phase of the mixture signals (because of the different spatial locations) observed by two microphones (or ears). The method performs well with no or little reverberation but as the amount of reverberation increases and the sources approach each other, the binaural cues are distorted and the interaural cues become indistinct, hence, degrading the separation performance. Thus, there is a demand for exploiting additional cues, and further signal processing is required at higher levels of reverberation.

2. PROPOSED APPROACH

There is evidence that visual cues contribute in enhancing intelligibility, specifically in adverse acoustic scenarios, such as with multiple sources and in the presence of larger levels of background noise [2] [3]. Motivated by this fact, this work explores one possible instance of incorporating the visual cues, namely utilizing the knowledge of the locations of the speakers in the audio source separation models. The speaker locations are estimated through video processing [4] to gain additional robustness over audio methods. These estimated locations are then used to calculate a direction vector towards each source. Fig. 1 shows the block diagram of the proposed approach.

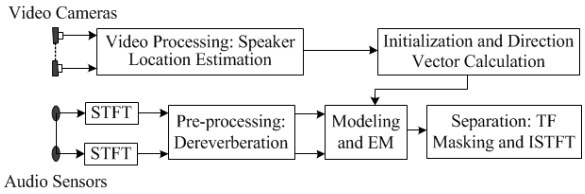

Fig. 1. Block diagram of the proposed processing.

As a pre-process, the observed reverberant mixtures are first dereverberated using a binaural spectral subtraction scheme. The late reverberant components are estimated using a state-of-the-art method [5] established for the monaural case. This monaural method is extended to the binaural form using a new gain derivation scheme. The mixtures are dereverberated and supplied to the second stage. The mixing vectors are modeled with Gaussian distributions [6] [7] with the aforementioned direction vector as its mean parameter. The ILD and IPD are also modeled as normal distributions in the TF domain. The mixing vector model is fused with the ILD and IPD models. The parameters of the models, apart from the mean of the mixing vector model, are estimated through the iterative expectation-maximization (EM) algorithm. The EM algorithm is also initialized with the source location estimates derived using the video process. The EM algorithm iterates between assigning regions in the TF spectrogram to individual sources based on the posterior probability of the combined models and refining the estimates of the parameters of these models. TF masks are obtained after a specified number of iterations of the EM algorithm. Each time and frequency component of the TF mask for each source indicates its probability of belonging to that source. The soft masks are then applied to the dereverberated mixtures from the first stage to separate all the sources in the mixture.

* Corresponding author

Table 1. SDR (dB) and PESQ for the case of two speakers at different levels of reverberation. The interferer was located at 15° azimuth. The proposed method is compared with the VIIMM and IIM methods.

	SDR (dB) [PESQ]			
	160 ms	300 ms	485 ms	600 ms
Proposed	8.30 [1.89]	7.99 [1.81]	6.29 [1.66]	4.71 [1.56]
VIIMM	7.26 [1.82]	6.67 [1.73]	5.26 [1.61]	3.60 [1.50]
IIM	3.10 [1.60]	2.88 [1.55]	2.17 [1.42]	0.83 [1.33]

3. EXPERIMENTAL RESULTS

Experiments were performed in different contexts ranging from varying the amount of reverberation, the number of source mixtures, and the proximity of the sources, to verify the improvement that can be achieved by exploiting visual cues. Results in terms of the signal-to-distortion ratio (SDR) [8] and the perceptual evaluation of speech quality (PESQ) highlight the advantage of the proposed approach over audio-only separation algorithms in multi-speaker highly reverberant scenarios. Speech sources were chosen from the TIMIT database and were convolved with room impulse responses generated using the image method [9]. The separation performance of the proposed algorithm was compared with the visually-aided ILD, IPD, and mixing vector model without the pre-processing, referred to as VIIMM, and the audio-only algorithm in [1] with ILD and IPD models, termed as IIM.

Table 1 summarizes the results in terms of SDR in dB and PESQ for mixtures of two sources. The target source was positioned at 0° azimuth and the interferer at 15°. This scenario is particularly challenging because of the relatively small separation angle between the sources. Performance was measured at different levels of reverberation. Considering the results at the highest reverberation time (RT60) of 600 ms, for instance, the proposed algorithm is 3.88 dB and 1.11 dB better in terms of SDR, and 0.23 and 0.06 better in terms of PESQ than the IIM and VIIMM methods respectively.

Experiments for mixtures of three sources were performed with the maskers placed symmetrically at 45° azimuth around the target. At RT60 of 485 ms, in terms of SDR, the proposed algorithm added an advantage of 4.35 dB to the IIM method and 1.29 to the VIIMM technique. While at 600 ms, the proposed scheme was 4.09 dB better than the IIM method and 1.33 dB than the VIIMM algorithm.

The results clearly indicate that utilizing visual cues, in terms of estimating speaker locations, is useful, specifically in challenging scenarios with higher level of reverberation, multiple speakers, and closely spaced sources. The additional resources required for the video processing will of course add to the overall complexity, but it is believed that given the advantage it can achieve over audio-only methods, multiple applications requiring superior performance within these difficult scenarios can afford the increased complexity. The pre-processing has been found to be useful in suppressing the late reverberant components before separation, adding useful gain to the overall output at higher levels of reverberation.

4. REFERENCES

[1] M. I. Mandel, R. J. Weiss, and D. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 382–394, 2010.

[2] W. H. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise," *Journal of the Acoustical Society of America*, vol. 26, no. 2, pp. 212–215, 1954.

[3] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, pp. 746–748, 1976.

[4] A. Rehman, S. M. Naqvi, W.Wang R. Phan, and J. A. Chambers, "MCMC-PF based multiple head tracking in a room environment," *4th UK Computer Vision Student Workshop (BMVW)*, 2012.

[5] K. Lebart, J. M. Boucher, and P. N. Denbigh, "A new method based on spectral subtraction for speech dereverberation," *Acta Acustica united with Acustica*, vol. 87, no. 3, pp. 359–366, 2001.

[6] P. D. O'Grady and B. A Pearlmutter, "Soft-LOST: EM on a mixture of oriented lines," *in Proc. ICA 2004, ser. Lecture Notes in Computer Science, Springer-Verlag*, pp. 430–436, 2004.

[7] H. Sawada, S. Araki, and S. Makino, "A two-stage frequency-domain blind source separation method for underdetermined convolutive mixtures," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY*, 21-24 October 2007.

[8] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[9] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *Journal of the Acoustical Society of America.*, vol. 65, no. 4, pp. 943–950, 1979.