Separation of Underdetermined Reverberant Speech Mixtures by Monaural, Binaural and Statistical Cue Combination

By Atiyeh Alinaghi, Wenwu Wang & Philip Jackson

Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, Guildford, UK

Abstract

Underdetermined reverberant speech separation is a challenging problem in source separation that has received considerable attention in both computational auditory scene analysis (CASA) and blind source separation (BSS). Recent studies suggest that, in general, the performance of frequency domain BSS methods suffer from the permutation problem across frequencies which degrades in high reverberation, meanwhile, CASA methods perform less effectively for closely spaced sources. This paper presents a method to address these limitations, based on the combination of monaural, binaural and BSS cues for the automatic classification of time-frequency (T-F) units of the speech mixture spectrogram. By modeling the interaural phase difference, the interaural level difference and frequency-bin mixing vectors, we integrate the coherence information for each source within a probabilistic framework. The Expectation-Maximization (EM) algorithm is then used iteratively to refine the soft assignment of TF regions to sources and re-estimate their model parameters. It is observed that the reliability of the cues affects the accuracy of the estimates and varies with respect to cue type and frequency. As such, the contribution of each cue to the assignment decision is adjusted by weighting the log-likelihoods of the cues empirically, which significantly improves the performance. Results are reported for binaural speech mixtures in five rooms covering a range of reverberation times and direct-to-reverberant ratios. The proposed method compares favorably with state-of-the-art baseline algorithms by Mandel et al. and Sawada et al., in terms of signal-to-distortion ratio (SDR) of the separated source signals. The paper also investigates the effect of introducing spectral cues for integration within the same framework. Analysis of the experimental outcomes will include a comparison of the contribution of individual cues under varying conditions and discussion of the implications for system optimization.

1. Introduction

Hearing aids, automatic speech recognition (ASR) and many other communication systems work well when there is just one source with almost no echo, but their performance degrades in situations where there are more speakers talking simultaneously or the reverberation is high. Therefore, it is highly desirable to localize and separate the source signals as an auditory front-end. Many different solutions have been suggested to solve this problem which can be grouped into two major approaches known as blind source separation (BSS) (Makino, Lee & Sawada (2007) and Hyvarianen & Oja (2000)) and computational auditory scene analysis (CASA) (Wang & Brown (2006)). The former is based on the statistical properties of the signals whereas the latter is inspired by human auditory processes, and exploits various properties of the speech signals.

Since the convolutive mixing models are usually considered for real speech mixtures, applying the BSS techniques directly to the mixtures in the time domain such as ICA introduced by Hyvarianen & Oja (2000) would involve high computational cost. A computationally more efficient solution is to decompose the signal into its frequency components and separate the sources at frequency bins independently. However, as the BSS algorithms do not preserve the order of the sources, the recovered sources are not necessarily aligned over all the frequency channels, introducing a new challenge, known as the *permutation problem*. On the other hand, the performance of CASA techniques which are based on binaural cues such as interaural level difference (ILD) and interaural phase difference (IPD) degrades when the sources are close to each other. In this paper, the BSS, IPD and ILD cues are modeled and then combined to compensate for the limitations of each of them (Alinaghi, Wang & Jackson (2011)).

Moreover, to reduce the effect of reverberation we exploit the coherence between the left and right microphones to detect the T-F units containing more energy from the direct signals which show a high coherence and give more weights to those T-F regions. This approach resembles the *precedence effect* in the human auditory system which mainly considers the binaural cues of the first wave front as in Jeub et al. (2010).

The cues are extracted from the left and right recordings in the T-F domain and then employed to calculate the probability of each source at each T-F unit. Once the likelihood of the cues has been maximized, the results can be used to estimate a soft mask to extract the source signals from the mixtures. In this method, we also eliminate the permutation problem by an appropriate initialization using binaural information about the sources. We also reduce the effect of reverberation by weighting the TF units based on coherence information. Finally, it is shown that the proposed technique outperforms the two baselines by Mandel et al. (2010), and Sawada et al. (2011).

2. Methods

In stereo recordings there are two microphones representing right and left ears, and so two mixtures are available, *l(n)* and *r(n)*, where n is the discrete time index. Each recording is the combination of filtered source signals with additive or reverberant noise. It is found Mandel, Weiss & Ellis (2010) that a reverberant noise model works for both cases:

$$\begin{aligned} l(n) &= \sum_{i=1}^{N} s_i(n) * h_{il}(n) * n_l(n), \\ r(n) &= \sum_{i=1}^{N} s_i(n) * h_{ir}(n) * n_r(n), \end{aligned} \tag{2.1}$$

where N, known *a priori*, is the number of sources, $s_i(n)$, $h_{il}(n)$ and $h_{ir}(n)$ are the ith source signal and the room impulse responses from source i to the left and right ears, respectively. The signals $n_l(n)$ and $n_r(n)$ represent the effect of the background noise. The spectrogram of each signal can be computed using the short time Fourier transform (STFT). The main idea is to partition the T-F regions belonging to different sources exploiting various information at each T-F units. To extract the binaural features of the signal the interaural spectrogram is calculated by dividing the left and right spectrograms at each T-F unit. In addition, the left and right signal values are concatenated at each T-F unit to produce a 2 dimensional observation vector $\mathbf{X}$ as follows:

$$\frac{L(\omega,t)}{R(\omega,t)} = 10^{\alpha(\omega,t)/20} e^{j\phi(\omega,t)}, \mathbf{X}(\omega,t) = [L(\omega,t)R(\omega,t)]^T \tag{2.2}$$

where $L(\omega, t)$ and $R(\omega, t)$ are the transformed left and right signals at each frequency ω and time frame t, respectively.

At each T-F point (ω, t), three features are available, $\alpha(\omega, t)$, i.e. the ILD, $\phi(\omega, t)$, i.e. the IPD, and $\mathbf{X}(\omega, t)$, i.e. the observation vector. The mixing vector $\mathbf{a}_i$ gives a fit to the set of $\mathbf{X}(\omega, t)$ over time for each source i. Each feature can be modeled by a Normal distribution with parameters that maximize the likelihood of the measured values, as described in:

$$L(\hat{\Theta}) = \max_{\theta} \sum_{\omega,t} \log p(\phi(\omega, t), \alpha(\omega, t), \mathbf{x}(\omega, t)|\Theta) \tag{2.3}$$

where

$$\hat{\Theta} = \{\xi_i(\omega), \sigma_i(\omega), \mu_i(\omega), \eta_i(\omega), \mathbf{a}_i(\omega), \gamma_i(\omega), \psi_i(\omega)\}$$

and ξ_i, σ_i^2, μ_i, η_i^2, $\mathbf{a}_i$, and γ_i^2 are the mean and variance of the IPDs, the ILDs and the mixing vectors, respectively. Once the underlying parameters are estimated using the Expectation maximization (EM) algorithm, the probability of each TF unit belonging to each source can be calculated as described in Alinaghi, Wang & Jackson (2011).

We also introduced some weights to the cues to adjust their contribution:

$$\log(\nu) \propto W_P \cdot \log \psi p(\hat{\phi}|\xi, \sigma^2) + W_L \cdot \log p(\alpha|\mu, \eta^2) + W_B \cdot \log p(\mathbf{x}|\mathbf{a}, \gamma^2) \tag{2.4}$$

where W_P, W_L and W_B control the influence of IPD, ILD and basis vector cues, respectively.

In addition, as mentioned in Jeub, Schafer, Esch & Vary (2010), the precedence effect can be modeled by Interaural Coherence (IC). It is shown that the recordings corresponding to the direct sound are coherent over all microphones. The idea is to give more weight to the T-F units containing more energy from direct sound which can be performed by a Wiener filter.

$$G_{coh}(\omega, t) = \frac{\hat{\Phi}_{ss}(\omega, t)}{0.5(\hat{\Phi}_{\tilde{s_r}\tilde{s_r}}(\omega, t) + \hat{\Phi}_{\tilde{s_l}\tilde{s_l}}(\omega, t))} \tag{2.5}$$

where $\hat{\Phi}_{ss}$ is the auto-power spectral density (APSD) of the original signal which can be estimated using the correlation of the observed signals and a binaural coherence model as explained in Jeub, Schafer, Esch & Vary (2010). $\hat{\Phi}_{\tilde{s_r}\tilde{s_r}}$ and $\hat{\Phi}_{\tilde{s_l}\tilde{s_l}}$ are the (APSD) of the left and right ear recordings, respectively which can be calculated by recursive periodogram approach. The Wiener coefficients, G_{coh} are calculated in the first step and then multiplied by mixture spectrogram at each T-F unit. The signals should be time aligned for calculating the coefficients, so that it can be applied for sources at different azimuth. For start we only considered the target source at zero azimuth.

3. Experiments and Results

For each T_{60} and configuration, 15 pairs from those 15 selected utterances from Garofalo et al. (1993) were chosen in such a way that no signal would be mixed with itself. They were then convolved by the room impulse responses measured in Hummersone (2011). The mixtures were then generated by simply adding the reverberant target and interferer signals which is equivalent to assuming the superposition of their respective sound fields. The target source was always located at the zero azimuth while the interferer's azimuth varied from 10° to 90° with steps of 5°, 1.5 m away from the head (this defines 6

Case	Methods	X	A	B	C	D	Mean
2-Src	Sawada	12.52	9.11	6.19	8.63	4.36	7.07
	Mandel	13.27	10.14	7.10	9.51	5.42	8.04
	Unweighted	**14.57**	10.65	7.27	9.79	5.93	8.41
	Weighted	14.03	10.80	7.61	10.05	6.31	8.69
	Dereverb	-	**10.90**	**7.70**	**10.15**	**6.41**	**8.70**
3-Src	Sawada	4.95	6.43	4.13	6.03	3.30	4.97
	Mandel	8.78	7.81	4.93	7.40	3.97	6.03
	Unweighted	9.58	8.31	5.21	7.69	4.20	6.35
	Weighted	**9.61**	**8.49**	**5.52**	**8.03**	**4.73**	**6.69**
	Dereverb	-	8.49	5.52	8.03	4.73	6.69

TABLE 1. Results of baseline methods and proposed method without ($W_P = W_L = W_B = 1$) and with weighting ($W_P = 0.8, W_L = 0.1, W_B = 0.5$) for anechoic, X, and reverberant mixtures with the average over A ($T_{60} = 0.32s$), B ($0.47s$), C ($0.68s$) and D ($0.89s$) in SDR [dB].

different configurations). Table 1. shows improvement for 2-src mixtures, while for 3-src mixtures dereverberation has not been effective. However, the dereverberation algorithm contains some parameters that can be adjusted for various conditions to achieve better performance which can be investigated in our future work.

REFERENCES

MAKINO, S., LEE, T. -W. & SAWADA, H. 2007 *Blind Speech Separation,* 1 ST ED., SER. SIGNAL AND COMMUNICATION TECHNOLOGY

HYVARIANEN, A. & OJA, E. 1991 Independent Component analysis: algorithms and applications *Neural Networks,* vol. 13, no.4-5, 411–430.

WANG, D. L. & BROWN, G. J. 2006 *Computational Auditory Scene Analysis: Principles, Algorithms and Applications.* (ed. Wang, D. L. & Brown, G. J.). Wiley interscience and IEEE press.

SAWADA, H., ARAKI, S. & MAKINO, S. 2011 Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment. *IEEE Trans. Audio, Speech, Lang. Process.,* vol. 19, no. 3, 516–527.

MANDEL, M. I., WEISS, R. J. & ELLIS, D. P. W. 2011 Model-based expectation maximization source separation and localization. *IEEE Trans. Audio, Speech, Lang. Process.,* vol. 18, no. 2, 382–394.

JEUB, M., SCHAFER, M., ESCH, T. & VARY, P. 2010 Model-Based Dereverberation Preserving Binaural Cues. *IEEE Trans. Audio, Speech, Lang. Process.,* vol. 18, no. 7, 1732–1745.

GAROFALO, J. S., LAMEL, L. F., FISHER, W. M., FISCUS, J. G., PALLETT, D. S. & DAHLGREN, N. L. 1993 The DARPA TIMIT acoustic-phonetic continuous speech corpus cdrom Linguistic Data Consortium. [Online] Available: http://www.idc.upenn.edu/Catalog/LDC93S1.html

HUMMERSONE, C. 2011 A psychoacoustic ngineering approac to machine sound source separation in reverberant environments. Ph.D. dissertation, Music and Sound Recording, University of Surrey, UK.

ALINAGHI, A., WANG, W. & JACKSON, P. J. B. 2011 Integrating binaural cues and blind source separation method for separating reverberant speech mixtures. in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.(ICASSP),* 209–212.