REVERBERATED SPEECH SIGNAL SEPARATION BASED ON REGULARIZED SUBBAND FEEDFORWARD ICA AND INSTANTANEOUS DIRECTION OF ARRIVAL

Lae-Hoon Kim[1], Ivan Tashev[2] and Alex Acero[2]

[1]Departement of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA
lkim9@illinois.edu
[2]Microsoft Research, One Microsoft Way, Redmond, WA 98052, USA
{ivantash, alexac}@microsoft.com

ABSTRACT

In this paper, independent component analysis (ICA) in a subband domain has been extended into a feed-forward network. The feed-forward network maximizes mutual independence of separated current frames using information from both current and previous multi-channel frames of speech signals captured by a microphone array. To guide into a proper separation preventing permutation and arbitrary scaling, we not only rely on the steered response for the first tap of the demixing filter but also penalize on the direction thus drastically increasing the mean squared error with the spatial filtered output. After convergence, by applying instantaneous direction of arrival (IDOA) based post-processing, we can additionally suppress the leakage of interference as well as the reverberated target signal. The signal to interference ratio (SIR) is improved more than 20 dBC for distances up to 2.7 m and angle differences down to 26°.

Index Terms — Speech separation, Independent component analysis, Beamforming, Instantaneous direction of arrival, Feed-forward ICA

1. INTRODUCTION

Speech separation has been an active research topic for various interesting applications, ranging from speech enhancement to simultaneous capture and separation of human voices. One of the compelling applications is simultaneous voice control of multimedia equipment. In this scenario we should successfully handle various acoustic environments (noise levels and reverberation conditions) to achieve a robust separation of the simultaneous voices. Recently, combinations of beamforming and independent component analysis have been proposed [1]. Considering the fact that those two schemes are based on the different optimality criteria (minimizing power for non-look direction signal [2] versus maximizing non-Gaussianity or mutual independence [3]), we might be able to expect that combining two heterogeneous technologies would provide better results than each individual approach alone. However, most of the current approaches are reporting that converged ICA demixing filters are closer to a null-former on interference sources [1], [4], [5]. Although this observation is appealing because by nulling the interferences we can increase super-Gaussianity in such a way that we can suppress the unwanted speech signals. This also means that the ICA does not contribute much to the conventional beamforming combined with nullforming on the interferences (assuming known directions of arrival). In fact, the beamforming plus nullforming scheme seems to be the best we can achieve using only the information from the current frame.

Frequency domain ICA has been proposed to solve convolutive mixing with separated instantaneous demixing in each individual frequency bin [3]. Although the fact that convolution in time domain can be represented as multiplication in frequency domain has been a reasonable justification for the benefits of the frequency domain approach, this is true only when the frame length is large enough. In a typical frame length (10-20 ms), the reverberated target and interference cannot be compensated properly because reverberation time (typically 200-300 ms) exceeds the frame length [4]. Also, the permutation and arbitrary scaling among the separated sources per each frequency bin have been critical issues that still need to be solved.

In this paper, we overcome the fundamental limitation of the subband domain ICA in a reverberant acoustic environment by taking previous multi-channel frames into consideration as well in order to increase the super-Gaussianity of the separated speech signals rather than just using current frames for instantaneous demixing. Feed-forward demixing filter structure with several taps in the subband domain is accommodated with natural gradient update rules [6]. To prevent permutation and arbitrary scaling and guide the separated speech sources into the designated channel outputs, we not only use the estimated spatial information on the target and interference, but also add a regularization term on the update equation thus minimizing mean squared error between separated output signals and the outputs of spatial filters. After convergence of the regularized feed-forward demixing filter, we observe better separation of the speech signals, with audible late reverberation for both desired and interference speech signals. These reverberation tails can be substantially suppressed by using spatial filtering based on the instantaneous direction of arrival (IDOA), which gives us the probability for each frequency bin to be in the original source direction [7]. This post-processing also suppresses the remaining leakage of the interference speech coming from non-look directions. The proposed method is evaluated using two criteria: physical separation is measured by the signal to interference ratio (SIR), and separated speech quality is measured by the perceptual evaluation of speech quality (PESQ) algorithm [8]. Experiments are performed in a relatively adverse acoustic environment (T_{60} of 375 ms, SNR of 15 dBC, where dBC stands for dB in C-weighting [7]), a distance of 1.5 to 4.3 meters, and an angle between the two speakers ranging from 6° to 70°. The proposed algorithm achieves an improvement of 29 dBC in SIR and 0.6 in PESQ points for the best case. These improvements

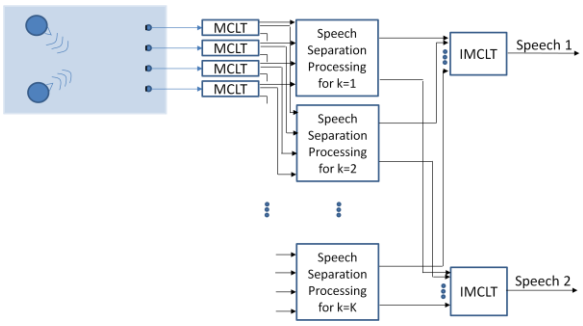

Figure 1. Diagram of speech separation in subband domain

remain above 10 dBC and 0.084 PESQ points in the most adverse condition. For distances up to 2.7 m and separation angles down to 26° the SIR stays above 20 dBC.

2. PROBLEM FORMULATION AND BACKGROUND

Figure 1 shows a block diagram for separation of two independent speeches in the subband domain. Time-domain signals captured using multiple microphones are converted to the subband domain using a modulated complex lapped transform (MCLT) that can produce better separation between frequency bands in an efficient manner [9]. The source separation can be performed using a demixing filter in each individual frequency bin $k = 1, 2, \cdots, K$ where K is the number of the frequency bins. Then the resulting signal can be converted back to the time domain using inverse MCLT. Source separation per each bin can be formulated as following:

$$\mathbf{S} = \mathbf{WY} \tag{1}$$

where $\mathbf{S}$ is the separated speech vector, $\mathbf{W}$ is the demixing matrix, and $\mathbf{Y}$ is the measured speech vector in a reverberant and noisy environment. We omitted a bin index for clarity of presentation.

2.1. Beamforming

One of the most commonly used beamformers is the minimum variance distortionless response (MVDR) beamformer, which in the frequency domain can be described as:

$$\mathbf{W}^H = \frac{\mathbf{D}^H \mathbf{R}_n^{-1}}{\mathbf{D}^H \mathbf{R}_n^{-1} \mathbf{D}} \tag{2}$$

Here $\mathbf{D}$ is a steering vector, $\mathbf{R}_n$ is a noise covariance matrix, and $\mathbf{W}$ is a weights matrix. Often the noise only covariance $\mathbf{R}_n$ is replaced by $\mathbf{R}$, which is the covariance matrix of the input (signal plus noise). This is more convenient as we avoid using a voice activity detector. This beamformer is known as minimum power distortionless response (MPDR). To prevent instability due to the direction of arrival mismatch, a regularization term is added to the sample covariance matrix [2], [7]. In our case, we also add an additional null constraint in the direction of the interference. The beamformer with extra nullforming constraint can be formulated as:

$$\mathbf{W}^H = [1 \quad 0]\left([\mathbf{D}_t \,|\, \mathbf{D}_i]^H [\mathbf{R} + \lambda \mathbf{I}]^{-1} [\mathbf{D}_t \,|\, \mathbf{D}_i]\right)^{-1} [\mathbf{D}_t \,|\, \mathbf{D}_i]^H [\mathbf{R} + \lambda \mathbf{I}]^{-1} \tag{3}$$

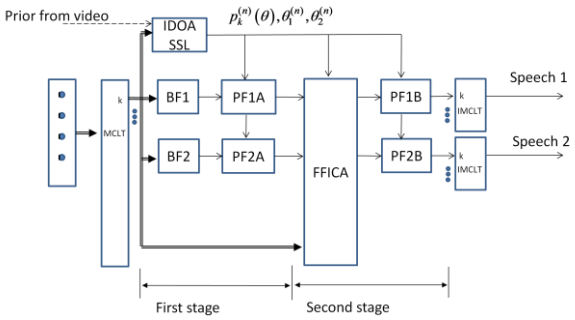

Figure 2. Proposed regularized feed-forward ICA with IDOA based post-processing

where $\mathbf{D}_t$ and $\mathbf{D}_i$ are steering vectors toward the target and interference direction respectively, and λ is the regularization term for diagonal loading. With the beam on the target and null on the interference directions, we can initialize the first-tap of the feed-forward ICA filter for appropriate channel assignment.

2.2. Combination of conventional subband domain ICA and beamforming

With a proper choice of the non-linear mapping function based on the speech signal statistic [3], [10], maximizing of the mutual independence results in the maximization of the super-Guassianity of the separated speech signals. The original ICA for the instantaneous mixing case can be extended to the convolutive mixing such as mixing of multiple simultaneous speech signals in a room. This conventional ICA approach performed in the subband domain converges towards the beamformer plus nullformer solution (3). This fact has been used as an additional component in the ICA filter update [1] to prevent misleading of the convergence and to speed up the convergence itself. Also, the steered response on the converged ICA filter per each frequency bin has been utilized as a cue for solving the permutation and scaling problem.

3. REGULARIZED FEED-FORWARD ICA WITH IDOA BASED POST-PROCESSING

Figure 2 shows a block diagram of the proposed two step method for one subband. The first step is the beamforming plus nullforming, followed by an IDOA based spatial filter, which produces additional suppression of the interference signals. The second step is feed-forward ICA, which uses the output of the first step for regularization. Here we maximize the mutual independence of the separated speeches by using both current and previous multi-channel frames. A secondary spatial filter is applied at the end of the second step.

3.1. Beamforming followed by a spatial filter

To determine the direction of arrival (DOA) of the desired and interference speech signals we use an IDOA based sound source localizer. Instantaneous Direction of Arrival (IDOA) space is $M-1$ dimensional with the axes being the phase differences between the non-repetitive pairs [7]. Here M is the number of microphones. This space allows estimation of the probability density function $p_k(\theta)$ as a function of the direction θ for

each subband. The results from all subbands are aggregated and clustered [7]. At this stage additional cues (from a video camera, for example) can be applied to improve the localization and tracking precision. The sound source localizer provides directions to desired θ_1 and interference θ_2 signals. Given the proper estimation on the DOAs for the target and interference speech signals we apply the constrained beamformer plus nullformer according to (3).

The consequent spatial filter applies a time-varying real gain for each subband, acting as a spatio-temporal filter for suppressing sounds coming from non-look directions. The suppression gain is computed as:

$$G_k^{(n)} = \int_{\theta_1-\Delta\theta}^{\theta_1+\Delta\theta} p_k(\theta)d\theta \Big/ \int_{-\pi}^{+\pi} p_k(\theta)d\theta, \quad (4)$$

where $\Delta\theta$ is the range around the desired direction θ_1 from which we want to capture the sound.

3.2. Regularized feed-forward ICA followed by IDOA based post-processing

In this paper, we utilize the virtue of the time-domain source separation approach [6] in the subband domain case by allowing multiple taps in the demixing filter structure in each subband. The proposed update rule for the regularized feed-forward ICA (RFFICA) is given below:

$$\mathbf{W}_i = \mathbf{W}_i + \mu\left((1-\alpha)\cdot\Delta_{\text{ICA},i} - \alpha\cdot\Delta_{\text{First stage},i}\right) \quad (5)$$

where $i = 0,1,\cdots,N-1$, N is the number of taps. $\Delta_{\text{ICA},i}$ and $\Delta_{\text{First stage},i}$ represent the portion of the ICA update and the regularized portion on the first stage output.

$$\Delta_{\text{ICA},i} = \mathbf{W}_i - \left\langle g\left(\mathbf{S}\left(\cdot-(N-1)\right)\right)\mathbf{Y}^H(\cdot-i)\right\rangle_t \quad (6)$$

$$\mathbf{S}(\cdot) = \sum_{n=0}^{N-1}\mathbf{W}_n(\cdot)\mathbf{Y}(\cdot-n) \quad (7)$$

$$\mathbf{Y}(\cdot) = \sum_{n=0}^{N-1}\mathbf{W}_{N-1-n}^H(\cdot)\mathbf{S}(\cdot-n) \quad (8)$$

$$\Delta_{\text{First stage},i} = \left\langle \mathbf{Y}(\cdot-i)\big|_{\text{Ref}}\left(\mathbf{S}(\cdot)\big|_{\text{Ref}} - \mathbf{S}_{\text{First stage}}(\cdot)\right)\right\rangle_t \quad (9)$$

where $\langle\cdot\rangle_t$ represents time averaging, $(\cdot-i)$ represents i sample delay, $\mathbf{S}_{\text{First stage}}$ is the first stage output vector for regularization and $|_{\text{Ref}}$ represents the reference channels. The penalty term has been only applied to the channel where the references are assigned; the other entries for the mixing matrix are set to zero so that the penalty term vanishes on those channel updates. For initialization of the subsequent filters, we modeled the reverberation process as exponential attenuation:

$$\mathbf{W}_i = \exp(-\beta i)\cdot\mathbf{I} \quad (10)$$

where $\mathbf{I}$ is an identity matrix, β is selected to model the average reverberation time, and i is the tap index. Note that we initialized the first tap of RFFICA for the reference channels as a pseudo-inversion of the steering vector stack for the current experiment so that we can assign 1 to the target direction and null to the interference direction:

$$\mathbf{W}_{0,\text{ini}}|_{\text{ref}} = \left(\left[e(\theta_t)\,|\,e(\theta_i)\right]^H\left[e(\theta_t)\,|\,e(\theta_i)\right]\right)^{-1}\left[e(\theta_t)\,|\,e(\theta_i)\right]^H. \quad (11)$$

Because we update the initialized filter using ICA, a slight mismatch with actual DOA can be adjusted in the updating procedure. For the current experiment, we set α as 0.5 just to penalize the larger deviation from the first stage output. As a nonlinear function $g(\cdot)$ we used a polar-coordinate based tangent hyperbolic function, suitable to the super-Gaussian sources with a good convergence property [10]:

$$g(\mathbf{X}) = \tanh(|\mathbf{X}|)\exp(j\sphericalangle\mathbf{X}) \quad (12)$$

where $\sphericalangle\mathbf{X}$ represents the phase response of the complex value X. To deal with the permutation and scaling we also used the steered response of the converged first tap demixing filter as following:

$$S_l = \frac{S_l}{F_l}\cdot\left(\frac{|F_l|}{\max|\mathbf{F}|}\right)^{\gamma} \quad (13)$$

where l is the designated channel number, F_l is the steered response for the channel output, $\mathbf{F}$ is the steered response to the candidate DOAs. To penalize the non-look direction in the scaling process we added the non-linear attenuation with the normalization using the steered response. For our current experiment, we set γ as 1. The spatial filter also penalizes on the non-look directional sources in each frequency bin.

4. EXPERIMENTAL RESULTS

4.1. Experimental setup

For sound capture we used a four element microphone array with unidirectional microphones and a length of 225 mm. In a large office room (T_{60} of 375 ms) we measured the room impulse responses (RIRs) between a rectangular grid of points and each of the microphones. This was done by using a mouth simulator playing a wideband chirp signal from each point of the grid. Using the same microphone array we recorded ambient noise in the same space (air conditioning plus five computers). Using a clean speech corpus convolved with the corresponding RIRs and adding the natural noise, we generated 18 different evaluation cases with two speech sources. The distance from the microphone array to both speech sources was the same and varied from 1.3 to 4.3 meters. The distance between the sound sources varied from 0.6 to 1.8 meters, resulting in a distance angle between 6º and 70º from the microphone array point of view.

For evaluation we used two different measures: signal-to-interference-ratio (SIR) and perceptual sound quality measured with the PESQ algorithm [8]. While the first is the intuitive measure for a speech separation algorithm, the second allows us to keep track of the output signal quality. SIR is defined as the following:

$$\text{SIR} = 10\log 10\frac{\text{Target Portion Energy}}{\text{Interference Portion Energy}}\ \text{dBC} \quad (14)$$

where target and interference portion can be estimated precisely using the clean speech signals. Although PESQ is not a typical measure for a source separation purpose, it allows keeping track

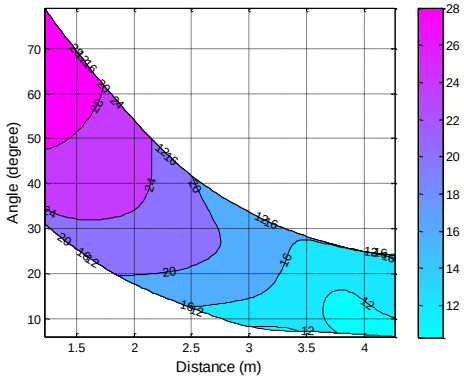

Figure 3. Contour plot of the improvement in SIR (dBC) as a function of the distance and separation angle.

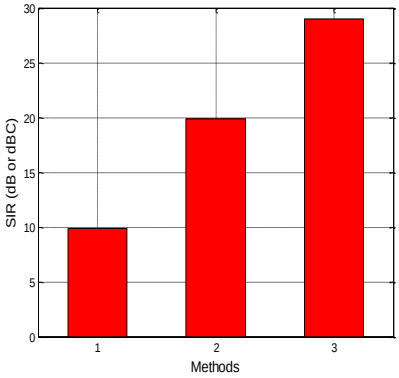

Figure 4. SIR (dBC) improvement comparison: 1.22 meters distance and 55°angle. 1 (BF+NF), 2 (BF+NF+SF), and 3 (Proposed).

of the distortions in the desired speech signal. The Matlab implementation of the proposed algorithm was built based on the audio stack and microphone array processing implementation provided in [7].

4.2. Results and discussion

We have evaluated all 18 two-speech cases. With a conservative setting of 1000 iterations and 20 taps filters for each subband, the proposed approach improves SIR in the range of 10 to 29 dBC and PESQ 0.1 to 0.6 points. Figure 3 represents contour plots of the improvement in SIR as a function of the distance to and the angle between the sound sources. Note that the contour plot has been generated by interpolating the results of 18 actual measurements. Assuming 20 dBC separation as good enough for practical purposes, we can say that the proposed algorithm is good enough for distances up to 2.7 meters for two speakers at 26°, i.e. standing shoulder to shoulder. Published papers report an SIR around 8 dB at a distance of 1.15 m, 70°, T_{60} 300 ms [1], or 13 dB for a distance of 1.7 m, 75°, T_{60} 200 ms [5]. In the most difficult condition of 6° between speakers at 4.23 meters distance, we can still maintain a 10 dBC SIR and 0.1 improvement in PESQ points.

Figure 4 provides a comparison of the results for several methods for a distance of 1.22 meters, 55° angle between speakers, and T_{60} 375 ms. These conditions are close, but more difficult than those published in [1] and [5]. Just the conventional beamformer plus nullformer provides a nearly 10 dBC improvement, which corresponds to the conclusions in these papers. Adding the spatial filter (the first stage of the proposed algorithm) increases the suppression close to 20 dBC, which is practically the maximum a spatial separation can achieve. Adding the second stage increases the SIR to 29 dBC while keeping an acceptable quality of the separated speech signal.

5. CONCLUSION AND FUTURE WORK

In this paper, we proposed a subband domain based, two stage approach utilizing spatial filtering, and regularized feed-forward ICA with multi-taps demixing filter. This approach produces substantial improvement in terms of SIR, while maintaining sound quality, measured with the PESQ algorithm. The proposed approach can be interpreted from both points of view independently. From the beamforming point of view, with proper prior knowledge of DOAs for the target and interferences, we can construct a subband domain filter structure augmented with higher-order independence maximization criterion for better suppression of the unwanted interference and noise. From the ICA point of view, with a prior knowledge on the DOAs for the target and interference, we can expand the conventional instantaneous demixing in the subband domain into the feed-forward network which turned out to increase the mutual independence without additional processing to solve the permutation problems. Many different ways of solving the permutation problem can be combined without hurting current schemes. In our case we used the IDOA-based sound source localization information.

6. REFERENCES

[1] H. Saruwatari, T. Kawamura, T. Nishikawa, A. Lee and K. Shikano, "Blind Source Separation Based on a Fast-Convergence Algorithm Combining ICA and Beamforming," IEEE Trans. on Audio, Speech, and Language Processing, vol. 14, no. 2, pp. 666-678, 2006.

[2] H. L. Van Trees, *Optimum Array Processing* Part IV of *Detection, Estimation, and Modulation Theory*, chapter 6 and 7, Wiley, 2002.

[3] T.-W. Lee, *Independent Component Analysis Theory and Applications*, Kluwer Academic Publishers, 1998.

[4] S. Araki, R. Mukai, S. Makino, T. Nishikawa and H. Saruwatari, "The Fundamental Limitation of Frequency Domain Blind Source Separation for Convolutive Mixtures of Speech," IEEE Trans. on Speech and Audio Processing, vol. 11, no. 2, pp. 109-116, 2003.

[5] H. Sawada, S. Araki, and S. Makino, "Frequency-Domain Blind Source Separation," in S. Makino, T.-W. Lee and H. Sawada Eds. *Blind Speech Separation*, Springer, pp. 47-78, 2007.

[6] S. Amari, S. C. Douglas, A. Cichocki and H. H. Yang, "Novel on-line adaptive learning algorithms for blind deconvolution using the natural gradient approach," Proc. IEEE 11th IFAC SYSID-97, pp. 1057-1062, 1997.

[7] I. Tashev, *Sound Capture and Processing Practical Approaches*, pp. 33-45, Wiley, 2009.

[8] ITU-T Recommended P. 862, "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone network and speech codecs," May 2000.

[9] H. S. Malvar, "A modulated complex lapped transform and its applications to audio processing," Proc. ICASSP'99, 1999.

[10] H. Sawada, R. Mukai, S. Araki, and S. Makino, "Polar Coordinate Based Nonlinear Function for Frequency-Domain Blind Source Separation," Proc. ICASSP'02, pp. 1001-1004, 2002.