DIRECTIONALITY-BASED SPEECH ENHANCEMENT FOR HEARING AIDS

John Woodruff and DeLiang Wang

Department of Computer Science and Engineering
The Ohio State University
Columbus, OH, 43210-1277, USA
{woodrufj, dwang}@cse.ohio-state.edu

ABSTRACT

In this work we describe methods for using the directionality of sound energy as a criterion to estimate single- and multichannel linear filters for suppression of diffuse noise and reverberation in a hearing aid application. We compare conservative strategies where direction of arrival is unknown, and more aggressive strategies where the proposed methods can be used to derive a fast acting post-filter for the output of a beamformer. We show that in situations where a target of interest is near to the listener while interfering sources are more distant, simple features that capture the directionality of sound energy can be used to attenuate significant undesired signal energy and can be more effective than a strategy based on noise-floor tracking.

Index Terms— Directional sound, multichannel speech enhancement, hearing aids

1. INTRODUCTION

In complex acoustic environments, individuals with hearing impairment may struggle to isolate speech content of interest due to interfering sounds, background noise and reverberation. Speech segregation algorithms seek to improve the intelligibility of a desired speech source by attenuating undesired signal energy [1]. A design choice in any segregation algorithm is identifying a measurable and physically meaningful cue to help distinguish between signal energy that should be retained and energy that should be attenuated.

In multichannel speech enhancement, direction of arrival (DOA) is a powerful cue by which to segregate a desired signal. If the approximate DOA of the target signal is predetermined, procedures for estimation of optimal real-valued [2] or complex [3] spatial filters can be derived. In the absence of a predetermined DOA, tracking of the spatial and spectral statistics of the target and noise sources can be used to derive spatial filters [4]. However, using DOA may have undesirable implications in dynamic environments when the relative position of sound sources can change quickly (i.e. due to source or head movements), or when a novel source of interest appears from an unknown direction.

In this work we explore using inter-microphone features, independent of DOA, to estimate the *directionality* of sound energy and use this as a cue to estimate a real-valued gain function for noise suppression. We characterize directionality by the level of the dominant direct path component of any source relative to all other energy contained in a mixture. We use a supervised learning approach to build an estimator of the directionality of sound and treat this as an approximation of the input signal-to-noise ratio (SNR). We also consider the case in which the DOA is assumed known and derive a post-filter based on cues that measure the degree to which sound energy is consistent with the given DOA.

The proposed techniques are related to existing methods that calculate a noise suppression gain based on measures of signal coherence [5, 6]. However, in this study we consider additional inter-microphone features to characterize directionality and use a supervised learning method to combine multiple types of features as well as features measured across multiple microphone pairs.

In the experiments performed we assume two behind-the-ear (BTE) hearing aids, each with two omni-directional microphones. We also assume ideal binaural exchange of the front microphone signal from each hearing aid. We present results using measured room-impulse responses (RIRs) recorded with two BTE hearing aids mounted on a head and torso simulator (HATS) in a reverberant room.

In the following section, we introduce the signal model that will be used throughout the paper and briefly introduce two well-known optimal linear filters for enhancement. In Section 3, we introduce the proposed directionality analysis and the method used to approximate the input SNR. In Section 4, we analyze the capacity of a directionality-based strategy to drive the enhancement methods. We conclude with a discussion of our results in Section 5.

2. BACKGROUND AND DEFINITIONS

2.1. Signal model

For a mixture of N point sources in a natural environment, we can model the signal received at microphone m in an individual frequency band as,

$$Y_m = \underbrace{H^d_{m,n} S_n}_{X_{m,n}} + \underbrace{H^r_{m,n} S_n + \sum_{l \neq n} H_{m,l} S_l}_{V_{m,n}}, \quad (1)$$

where S_n denotes source signal n, $H_{m,n} = H^d_{m,n} + H^r_{m,n}$ denotes the transfer function from source n to microphone m, and $H^d_{m,n}$ and $H^r_{m,n}$ denote the direct and reflected components of $H_{m,n}$, respectively. We let $X_{m,n}$ denote the target signal received by microphone m and $V_{m,n}$ denote the interference signal received by microphone m. We then let $\mathbf{X}_n = [X_{1,n}, ..., X_{M,n}]^T$, $\mathbf{V}_n = [V_{1,n}, ...V_{M,n}]^T$, and $\mathbf{Y} = [Y_1, ...Y_M]^T$, where M is the number of microphones.

We denote the target, interference and mixture covariances as $\Phi_{xx} = E[\mathbf{X}_n \mathbf{X}_n^H]$, $\Phi_{vv} = E[\mathbf{V}_n \mathbf{V}_n^H]$, and $\Phi_{yy} = E[\mathbf{Y}\mathbf{Y}^H]$, respectively, where H denotes conjugate transpose. Note that for clarity,

This research was supported by a grant from the Oticon Foundation and an AFOSR grant (FA9550-08-1-0155). The authors thank Ulrik Kjems and Michael Pedersen for their guidance and support in completing this work, and Jørn Skovgaard for measuring the room impulse responses used.

we have dropped the explicit reference to source n and will continue to do so when possible.

We denote the total power of the target and interference signals received at the microphones as ϕ_{xx} and ϕ_{vv}, respectively. The input SNR is then SNR $= \frac{\phi_{xx}}{\phi_{vv}}$. We assume $\Phi_{xx} = \phi_{xx} g g^H$, where g and ϕ_{xx} are calculated as the principal eigenvalue and eigenvector, respectively, of Φ_{xx}. In theory, g is a normalized version of $[H^d_{1,n}, ..., H^d_{M,n}]^T$. We also let $\Gamma_{vv} = \frac{1}{\phi_{vv}} \Phi_{vv}$ denote the interference coherence matrix, where $\phi_{vv} = \text{tr}(\Phi_{vv})$.

2.2. Enhancement methods

One enhancement approach is to build a single filter and apply it to both front microphones of the hearing aids. The single filter strategy has the advantage of preserving binaural cues for any sources that are not severely attenuated [2] and can be considered a more conservative enhancement approach. In this case we use a parametrized single-channel Wiener filter where the real-valued filter coefficient at microphone m is,

$$W_m = \frac{\text{SNR}}{\text{SNR} + \beta}, \tag{2}$$

where β controls the overall level of attenuation by the filter.

The parametrized multichannel Wiener filter (PMWF) [7] is a well-known multichannel strategy that can improve noise attenuation, but requires the estimation of spatial statistics for the target and interference signals and will distort the binaural cues of interfering sources (although cue preserving implementations have been proposed [8]). In this case, the multichannel linear filter coefficients can be expressed as the combination of a minimum variance distortionless beamformer (MVDR) and a post-filter,

$$W_m = \underbrace{\left(\frac{\text{SNR}}{\text{SNR} + \beta \left(g^H \Gamma_{vv}^{-1} g \right)^{-1}} \right)}_{post-filter} \underbrace{\frac{\Gamma_{vv}^{-1} g g^H u_m}{g^H \Gamma_{vv}^{-1} g}}_{MVDR}, \tag{3}$$

where u_m is a vector encoding the reference microphone m (e.g. $u_1 = [1, 0, 0, 0]^T$) and β is a parameter to control level of attenuation by the post-filter.

3. SNR ESTIMATION

In this section we describe a procedure for estimating the input SNR using features that seek to capture the level of directionality of sound energy. The estimation procedure can be used with either Equation (2) or (3). After time-frequency (T-F) analysis using a complex linear filterbank, we calculate the mixture covariance within each frequency band using,

$$\Phi_{yy}[k] = \alpha \Phi_{yy}[k-1] + (1-\alpha) \mathbf{Y}[k] \mathbf{Y}[k]^H, \tag{4}$$

where $\alpha \in [0, 1]$ controls the recursive averaging and k indexes time frames. Features are calculated from the mixture covariance in individual T-F units, although for convenience, we omit the time and frequency indices from our notation below.

3.1. Features to characterize directionality

A measure of the *diffuseness* of a sound field as calculated from a microphone pair is described in [9]. With access to two BTE hearing aids, due to the placement of the microphones and the effect of the head, diffuseness can only be measured along the front/back axis. Nevertheless, for certain DOAs using this feature may help to characterize the directionality of sound energy. We measure diffuseness using the microphone pairs at each hearing aid using Equation (2.43) in [9]. We denote these features as ψ_l and ψ_r for the left and right hearing aid, respectively.

Another well-known feature is the signal coherence [5, 6]. Directional sound energy will result in high coherence because two microphone signals will be similar (up to some scaling and phase shift), whereas diffuse energy will be less predictable between microphones. We calculate coherence between microphones m_0 and m_1 from the mixture covariance, Φ_{yy}, using Equation (5). We measure binaural coherence from the two front microphones, denoted c_b, and also measure coherence locally at each hearing aid, denoted c_l and c_r, respectively.

$$c = \frac{|\Phi_{yy}(m_0, m_1)|}{\sqrt{\Phi_{yy}(m_0, m_0)} \sqrt{\Phi_{yy}(m_1, m_1)}}. \tag{5}$$

We also observe that directional sound should be more likely to produce coordinated inter-microphone phase and level differences. With this in mind we measure the phase difference between microphones m_0 and m_1 using Equation (6) and the level difference using Equation (7). We measure the binaural phase difference from the front microphone pair, denoted τ_b, the binaural level difference, denoted λ_b, as well as the phase difference at each hearing aid, denoted τ_l and τ_r, respectively.

$$\tau = \angle \Phi_{yy}(m_0, m_1) \tag{6}$$

$$\lambda = 10 \log_{10} \left(\frac{\Phi_{yy}(m_0, m_0)}{\Phi_{yy}(m_1, m_1)} \right) \tag{7}$$

Finally, in the case where the spatial statistics of the target and interference are known (i.e. we have g and Γ_{vv} for Equation (3)), we consider a set of features to capture the similarity between an observed cue and the cues predicted by the known statistics. We calculate the inter-microphone phase and level differences from the look vector using Equations (6) and (7), where we replace Φ_{yy} with $g g^H$. We denote these values as τ_b^x, τ_l^x, τ_r^x, λ_f^x. Similarly, we calculate phase and level differences from the interference coherence matrix by replacing Φ_{yy} in Equations (6) and (7) with Γ_{vv}. We denote these values as τ_b^v, τ_l^v, τ_r^v, λ_f^v.

We then calculate 8 similarity values between each cue pair: $s(\tau_b, \tau_b^x)$, $s(\tau_b, \tau_b^v)$, $s(\tau_l, \tau_l^x)$, $s(\tau_l, \tau_l^v)$, $s(\tau_r, \tau_r^x)$, $s(\tau_r, \tau_r^v)$, $s(\lambda_b, \lambda_b^x)$, and $s(\lambda_b, \lambda_b^v)$. We use a Gaussian weight to capture similarity, $s(a, b) = \exp(-(a-b)^2 / 2\sigma^2)$. We calculate the normalization factor, σ^2, as the sample variance of the difference between the cues (e.g. $\sigma^2 = \text{Var}(\tau_b - \tau_b^x)$) using a set of training data (described in the following subsection). We calculate σ^2 separately for each of the 8 pairs.

3.2. Estimation method

The feature sets described above are used as predictors for the relative level of directional sound energy. We then treat this as an approximation of the input SNR. This approximation should be reasonable in cases where the source of interest has a strong direct path component while interfering sound is primarily due to diffuse and reverberant energy.

In this study, we use a multi-layer perceptron (MLP) as a generic function approximation tool because, with the exception of coherence, there is no straightforward relationship between feature values and the relative level of the target signal. We use a feed-forward

network in which all inputs (the selected set of feature values) are connected to a set of hidden nodes with tangential sigmoid transfer functions. The networks contain a single output node with a log-sigmoid transfer function. For simplicity, we use 4 hidden nodes for each feature provided to this system (i.e. a set of 4 features will be trained with 16 hidden nodes). Network training is performed using a gradient descent backpropagation algorithm.

We train the systems using a set of multi-talker babble mixtures different from those used in testing, but where mixtures were generated similarly to the test set described in Section 4.1. In each case there is a dominant target source and a set of interfering talkers. The number of interferers was randomly selected to be 3, 5, 10 or 15. Azimuth angles of all sources were randomly selected.

For the feature sets with unknown DOA, we consider the SNR of the dominant source for each T-F unit as the *ideal* measure of directionality. In other words, we measure the SNR (using ϕ_{xx} and ϕ_{vv}) from the perspective of each of the sources contained in the mixture and calculate $\text{SNR}_{\text{dom}} = \max_n(\text{SNR}_n)$. We then train the MLPs using $\frac{\text{SNR}_{\text{dom}}}{\text{SNR}_{\text{dom}}+1}$ in order to have a bounded target value between 0 and 1. The output of the MLP, denoted $\widehat{R}$, can then be transformed to an SNR estimate using, $\widehat{\text{SNR}} = \frac{\widehat{R}}{\widehat{R}-1}$. As stated above, the assumption is that in the case where a single, near-field source corrupted by a number of distant and reverberant interfering sources, SNR_{dom} will be a good approximation of SNR_n. For the case when the DOA of the target source is known, training is more straightforward and we use SNR_n rather than SNR_{dom}. In all cases, a separate MLP is trained for each frequency channel.

We train function estimators using five feature sets: "coh": $\{c_b, c_l, c_r\}$, "diff": $\{\psi_l, \psi_r\}$, "imd": $\{\tau_b, \tau_l, \tau_r, \lambda_b\}$, "dir": $\{\tau_b, \tau_l, \tau_r, \lambda_b, c_b, c_l, c_r, \psi_l, \psi_r\}$, "$\text{dir}_{\text{doa}}$": $\{s(\tau_b, \tau_b^x), s(\tau_b, \tau_b^v), s(\tau_l, \tau_l^x), s(\tau_l, \tau_l^v), s(\tau_r, \tau_r^x), s(\tau_r, \tau_r^v), s(\lambda_b, \lambda_b^x), s(\lambda_b, \lambda_b^v)\}$

4. EXPERIMENTAL RESULTS

In this section we describe two experiments to measure the noise reduction capacity, using either Equation (2) or (3), of the different feature sets. For the post-filter case, we show results assuming known and fixed spatial configuration of the target and noise signals, where the average spatial statistics of the target and noise signals, g and Γ_{vv}, are calculated from the mixture with ideal specification of noise-only frames. We generate the beamformer output assuming exchange of only the front microphone signal, so for the left hearing aid, we use the two local microphone signals and the front microphone of the right aid. Similarly for the right hearing aid.

4.1. Setup and Test Database

We generate a database of 20 mixtures for 5 target azimuths (0°, 45°, 90°, 135° and 180°) and 3 input SNR conditions (0, 3 and 6 dB), where we use individual speech utterances for both target and interfering signals. Speech signals were recorded monaurally in a dry environment and consist of both male and female talkers.

We create a mixture by convolving a speech utterance with a RIR recorded in a reverberant room with two BTE hearing aids placed on a HATS. The room dimensions are $10.4 \times 12.1 \times 4.1$ m and the HATS was placed 5 m from the left-most wall and 4.8 m from the rear-most wall. RIRs were measured at 8 angles (0°, 45°, 90°,..., 315°) at distances of 1 m and either 4 or 4.5 m (depending on the distance available in the measurement setup). The reverberation time (T_{60}) of the room is 741 ms, as calculated using the method in [10] from the left-front microphone measurement made at 0° and 1 m.

Table 1. ΔSIW-SNR (in dB) for 5 single filter and 3 post-filter methods, averaged over 60 mixtures (20 for each of 3 SNR conditions) and both left and right ear signals.

	Single Filter					Post-filter		
	coh	**diff**	**imd**	**dir**	**ste**	**dir_{doa}**	**ste**	**lte**
0°	4.7	4.2	4.2	**4.9**	4.4	**3.0**	1.7	1.4
45°	5.0	5.1	5.0	**5.4**	4.0	**2.8**	1.2	1.3
90°	**5.1**	3.2	4.4	**5.1**	4.1	**3.1**	1.6	1.6
135°	4.8	3.6	4.6	**4.9**	3.9	**3.1**	1.4	1.3
180°	**4.9**	3.4	3.7	4.6	4.2	**3.3**	1.8	1.3
Avg.	4.9	3.9	4.4	**5.0**	4.1	**3.1**	1.6	1.4

Each test mixture consists of 1 target talker placed at a selected azimuth and at a distance of 1 m. A multi-talker babble is created for the interference where azimuths are selected randomly (from the 8 possibilities) at either 4 or 4.5 m. In all cases we consider the direct path signal of the predetermined target source as the desired signal, and we treat the noise signal as the mixture after the desired signal has been removed. We take the first 20 ms of each RIR to be the direct-path component. The mixture SNR is set by balancing the level of the target signal relative to the noise signal, averaged over the two front microphones.

In all testing scenarios, signals are sampled at 20 kHz and passed through a linear filterbank with 128 frequency channels. The filterbank output is decimated by a factor of 64 (hop size of $\sim$ 3 ms) and a 256-sample Hamming window is used. We use a 20 ms time constant to set α in Equation (4).

4.2. Enhancement results

We first measure the performance of 5 alternative methods that use a single filter (Equation (2)). We measure performance using the change in speech-intelligibility weighted SNR (ΔSIW-SNR) relative to the unprocessed signals [11]. To compare to the directional methods, we use a system that estimates SNR by noise floor tracking, using the method proposed in [12]. We denote this system as "ste" for short-term energy. We calculate the SNR in this case as $\text{SNR} = \frac{\phi_{yy}}{\phi_{vv}} - 1$, where ϕ_{yy} is the low-pass filtered mixture energy, calculated with a time constant of 20 ms, and we set any negative SNR values to 0. The noise floor tracker uses a 500 ms time constant.

On the left side of Table 1, we show ΔSIW-SNR using the 5 alternatives to derive a single filter. Results are averaged over 120 signals (2 for each of 20 mixtures in 3 input SNR conditions) and shown for 5 azimuth angles, 0°, 45°, 90°, 135° and 180°. The final row of the table shows results averaged over azimuth. For each system, β was selected to maximize ΔSIW-SNR.

We can see that in general, directionality provides an improvement over the noise tracking approach. The feature set that includes all directional cues achieved the best performance in all but 1 condition and achieved between 0.4 and 1.4 dB higher ΔSIW-SNR than the "ste" system. The key difference between the directionality and the noise floor tracking strategies is replacing the assumption of spectral stationarity of the noise signal with the assumption that the noise signal is (approximately) diffuse. This leads to a significant difference in the character of the noise suppression achieved by the different systems.

Among the directional cues, coherence proved the most pow-

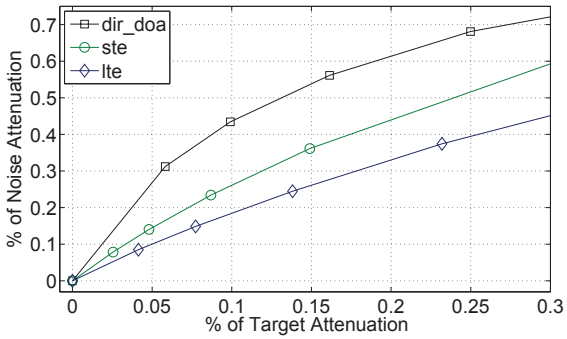

Fig. 1. Percentage of target attenuation versus noise attenuation as the parameter β is varied for 3 post-filter methods.

erful and as the target sound was moved closer to the side of the head, the coherence system improved. Further investigation showed that this is due to improved estimates on the ipsilateral side, relative to contralateral or binaural coherence. We found the combination of all 3 coherence estimates to perform better than binaural coherence or bilateral coherence measurements alone, which have been previously used for noise suppression [5, 6]. The diffuseness cues are reliable (again on the ipsilateral side) when the influence of the head shadow is minimized and when the signal is not arriving from a direction perpendicular to the array.

For a strategy that includes a spatial filter (i.e. using Equation (3)), we compare the post-filter generated using the "dir_{doa}" system to post-filters generated using the "ste" system and a common approach of using the long-term input SNR estimate captured by Φ_{xx} and Φ_{vv} [4]. We denote this approach as "lte" for long-term energy estimate. As stated above, Φ_{xx} and Φ_{vv} are calculated with ideal knowledge of noise-only frames, so both the beamformer and the long-term "lte" post-filter can be considered ideal.

We show the ΔSIW-SNR achieved by the 3 alternative post-filters in the rightmost columns of Table 1. Results are measured with reference to the output of the MVDR beamformer and again, results are shown with the optimal β for each system. We can see that the directional post-filter is able to achieve an average improvement of about 1.5 dB relative to the energy-based alternatives. This is due to achieving a high-quality estimate of the input SNR that varies quickly over time, as opposed to an estimate the averages over a large amount of data (as in the "lte" approach).

Figure 1 shows the percentage of target attenuation versus the percentage of noise attenuation as β is varied. Percentages are the amount of additional attenuation due to the post-filter. In this plot we can see that, although not apparent in the ΔSIW-SNR numbers, the "ste" system is able to achieve more noise attenuation for a fixed amount of target attenuation relative to the "lte" system. This suggests that, despite the common use of the "lte" approach, basing the post-filter on locally derived SNR estimates rather than long-term statistics can be advantageous. The "dir_{doa}" system further improves the amount of noise reduction that is possible for a fixed amount of target attenuation by incorporating awareness of the beam pattern.

5. CONCLUDING REMARKS

In this work we have proposed multichannel speech enhancement methods based on directionality. We have shown that in a conservative processing design, where one may want to avoid a spatial filter, estimating the directionality of sound energy independent of DOA can be a reliable indicator of the relative level of near-field sources in a reverberant space. This suggests that the proposed method, even in the absence of a spatial filter, can be used to attenuate diffuse noise and reverberation and enhance sources that are close to the listener.

In a design where a spatial filter can be robustly used to attenuate noise energy, our results indicate that using directionality to drive an accurate local estimate of the input SNR can be used to construct a better post-filter than simply using the long-term statistics in the PMWF formulation, as is commonly done.

Future work should consider non-linear smoothing and gain functions that are common in single-channel methods to improve SNR estimates. Alternative techniques to construct a post-filter with known DOA should also be considered (e.g. [2]). The use of an MLP was motivated by a desire to analyze the capacity of different directional cues to detect directional energy, and the generalization capacity of the proposed methods requires investigation.

6. REFERENCES

[1] D. L. Wang and G. J. Brown, Eds., *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, Wiley/IEEE Press, Hoboken, NJ, 2006.

[2] K. Reindl, Y. Zheng, and W. Kellermann, "Analysis of two generic wiener filtering concepts for binaural speech enhancement in hearing aids," in *Proc. EUSIPCO*, 2010, pp. 988–993.

[3] S. Doclo, A. Spriet, J. Wouters, and M. Moonen, "Frequency-domain criterion for the speech distortion weighted multichannel wiener filter for robust noise reduction," *Speech Commun.*, vol. 49, pp. 636–656, 2007.

[4] B. Cornelis, M. Moonen, and J. Wouters, "A QRD-RLS based frequency domain multichannel wiener filter algorithm for noise reduction in hearing aids," in *Proc. EUSIPCO*, 2010, pp. 1953–1857.

[5] T. Wittkop and V. Hohmann, "Strategy-selective noise reduction for binaural digital hearing aids," *Speech Commun.*, vol. 39, pp. 111–138, 2003.

[6] G. Grimm, V. Hohmann, and B. Kollmeier, "Increase and subjective evaluation of feedback stability in hearing aids by a binaural coherence-based noise reduction scheme," *IEEE Trans. Audio, Speech, Lang. Proc.*, vol. 17, pp. 1408–1419, 2009.

[7] M. Souden, J. Benesty, and S. Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE Trans. Audio, Speech, Lang. Proc.*, vol. 18, pp. 260–275, 2010.

[8] T. Klasen, T. Van den Bogaert, M. Moonen, and J. Wouters, "Binaural noise reduction algorithms for hearing aids that preserve interaural time delay cues," *IEEE Trans. Signal Process.*, vol. 55, pp. 1579–1585, 2007.

[9] J. Merimaa, *Analysis, Synthesis and Perception of Spatial Sound*, Ph.D. thesis, Helsinki University of Technology, 2006.

[10] M. R. Schroeder, "New method of measuring reverberation time," *J. Acoust. Soc. Am.*, vol. 37, pp. 409–412, 1965.

[11] J. E. Greenberg, P. M. Peterson, and P.M. Zurek, "Intelligibility-weighted measures of speech-to-interference ratio and speech system performance," *J. Acoust. Soc. Am.*, vol. 94, pp. 3009–3010, 1993.

[12] R. C. Hendriks, R. Heusdens, and J. Jensen, "MMSE based noise PSD tracking with low complexity," in *Proc. ICASSP*, 2010, pp. 4266–4269.