MICROPHONE ARRAY POST-PROCESSOR USING INSTANTANEOUS DIRECTION OF ARRIVAL

Ivan Tashev and Alex Acero

{ivantash, alexac}@microsoft.com
Microsoft Research, One Microsoft Way, Redmond 98052, USA

ABSTRACT

In this paper we describe a novel algorithm for post-processing a microphone array's beamformer output to achieve better spatial filtering under noise and reverberation. For each audio frame and frequency bin the algorithm estimates the spatial probability for sound source presence and applies a spatio-temporal filter towards the look-up direction. It is implemented as a real-time post-processor after a time-invariant beamformer and it substantially improves the directivity of the microphone array. The algorithm is CPU efficient and adapts quickly when the listening direction changes. It was evaluated with a linear four element microphone array. The directivity index improvement is up to 8 dB, the suppression of a jammer 40° from the sound source is up to 17 dB.

1. INTRODUCTION

Microphone arrays can be used for hands-free real-time communication and speech recognition. A popular family of algorithms is a time-invariant beamformer followed by a set of adaptive filters, forming a generalized side-lobe canceller architecture [1][2]. Such processing is linear and doesn't introduce artifacts as musical noise, although adaptive algorithms can have some audible residuals and distortions. While time-invariant beamformers are designed under the assumption of isotropic noise, adaptive algorithms perform better with point noise sources. Other efforts to improve the overall efficiency of the microphone arrays are focused on post-processing (or post-filtering) techniques. They are based on multichannel Wiener noise suppressors [3], the Zelinski post-filter [4], or spatial noise models and suppressors [5].

In this paper we present a novel non-linear post-processing algorithm for microphone arrays, which improves the directivity and signal separation capabilities. The algorithm works in so-called instantaneous direction of arrival space, estimates the probability for sound coming from the look-up direction and applies a time-varying, gain based, spatio-temporal filter for suppressing sounds coming from other directions, resulting in minimal artifacts and musical noise.

2. MODELING

2.1. Sound capture model

Let vector $\vec{p} = \{ p_m \;\; m = 0,1,\ldots,M-1 \}$ denote the positions of the M microphones in the array, where $p_m = (x_m, y_m, z_m)$. This yields a set of signals that we denote by vector $\vec{x}(t, \vec{p})$. Each sensor m has known directivity pattern $U_m(f, c)$, where $c = \{\varphi, \theta, \rho\}$ represents the coordinates of the sound source in a radial coordinate system and f denotes the signal frequency. We consider signal processing algorithms in the frequency domain, because that can lead to efficient FFT-based implementations. For a sound source at location c the captured signal from each microphone is:

$$X_m(f, p_m) = D_m(f, c) S(f) + N_m(f) \qquad (1)$$

where the first term in the right-hand side,

$$D_m(f, c) = \frac{e^{-j2\pi f \frac{\|c - p_m\|}{v}}}{\|c - p_m\|} A_m(f) U_m(f, c) \qquad (2)$$

represents the delay and decay due to the distance to the microphone $\|c - p_m\|$, and v is the speed of sound. The term $A_m(f)$ is the frequency response of the system preamplifier/ADC, $S(f)$ is the source signal, and $N_m(f)$ is the captured noise.

2.2. Ambient noise model

We consider the captured noise $N_m(f, p_m)$ as containing two noise components: acoustic noise and instrumental noise. The acoustic noise $N_A(f)$ is correlated across all microphone signals. The instrumental noise in each channel is incoherent across the channels, and usually has a nearly white noise spectrum $N_I(f)$. Assuming isotropic ambient noise we can represent the signal, captured by any of the microphones, as a sum of infinite number of uncorrelated noise sources randomly spread in space:

$$N_m = N_A \sum_{l=1}^{\infty} D_m(c_l)\mathbb{N}\left(0, \lambda_l(c_l)\right) + N_l \mathbb{N}(0, \lambda_l) \qquad (3)$$

Indices for frame and frequency are omitted for simplicity. Estimation of all of these noise sources is impossible because we have a finite number of microphones. Therefore we model the isotropic ambient noise as one noise source in different position in the work volume for each frame, plus a residual incoherent random component, which incorporates the instrumental noise. The noise capture equation changes to:

$$N_m^{(n)} = D_m(c_n)\mathbb{N}\left(0, \lambda_N(c_n)\right) + \mathbb{N}(0, \lambda_{NC}) \qquad (4)$$

where c_n is the noise source random position for n^{th} audio frame, $\lambda_N(c_n)$ is the spatially dependent correlated noise variation ($\lambda_N(c_n) = const \ \ \forall c_n$ for isotropic noise) and λ_{NC} is the variation of the incoherent component.

3. SPATIO-TEMPORAL FILTER

3.1. Instantaneous Direction Of Arrival space

We can find the Instantaneous Direction Of Arrival (IDOA) for each frequency bin based on the phase differences of non-repetitive pairs of input signals. For M microphones these phase differences form a $M-1$ dimensional space, spanning all potential IDOA. If we define an IDOA vector in this space as

$$\Delta(f) \triangleq \left[\delta_1(f), \delta_2(f), \ldots, \delta_{M-1}(f)\right] \qquad (5)$$

where

$$\delta_l(f) = \arg(X_0(f)) - \arg(X_l(f)) \ \ l = \{1, \ldots, M-1\} \quad (6)$$

then the non-correlated noise will be evenly spread in this space, while the signal and ambient noise (correlated components) will lay inside a hypervolume that represents all potential positions $c = \{\varphi, \theta, \rho\}$ of a sound source in the real three dimensional space. For far field sound capture, this is an $M-1$ dimensional hypersurface as the distance is presumed approaching infinity. Linear microphone arrays can distinguish only one dimension – the incident angle, and the real space is represented by a $M-1$ dimensional hyperline. Figure 1 shows the distribution of 1000 audio frames at 750 Hz for a four element linear array. The solid line is the set of theoretical positions of sound sources in the range of –90° to +90°. The actual distribution is a cloud around the theoretical line due to the presence of an additive non-correlated component. Note that for each point in the real space we have a corresponding point in the IDOA space (may be not unique). The opposite is not true: there are points in the IDOA space without corresponding point in the real space.

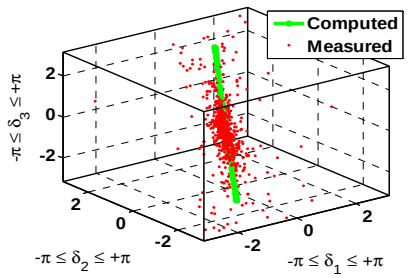

Figure 1. Distribution of frames in IDOA space.

3.2. Presence of a sound source

For simplicity and without any loss of generality, we will consider a linear microphone array, sensitive only to the incident angle θ – direction of arrival in one dimension. Let $\Psi_k(\theta)$ denote the function that generates the vector Δ for given θ and frequency bin k according to (1), (5) and (6). In each frame, the k^{th} bin is represented by one point Δ_k in the IDOA space. Let's have a sound source at θ_S with image in IDOA at $\Delta_S(k) = \Psi_k(\theta_S)$. With additive noise, the resultant point in IDOA space will be spread around $\Delta_S(k)$:

$$\Delta_{S+N}(k) = \Delta_S(k) + \mathbb{N}(0, \lambda_{IDOA}(k)) . \qquad (7)$$

3.3. Space conversion

The conversion from distance to the theoretical hyperline in IDOA space to distance into the incident angle space (real world, one dimensional in this case) is given by:

$$\Upsilon_k(\theta) = \frac{\left\|\Delta_k - \Psi_k(\theta)\right\|}{\left\|\frac{d\Psi_k(\theta)}{d\theta}\right\|} \qquad (8)$$

where $\left\|\Delta_k - \Psi_k(\theta)\right\|$ is the Euclidean distance between Δ_k and $\Psi_k(\theta)$ in IDOA space, $\frac{d\Psi_k(\theta)}{d\theta}$ are the partial derivatives, and $\Upsilon_k(\theta)$ is the distance of observed IDOA point to the points in the real world. Note that the dimensions in IDOA space are measured in radians as phase difference, while $\Upsilon_k(\theta)$ is measured in radians as units of incident angle.

3.4. Estimation of the variance in real space

Analytic estimation in real-time of the probability density function for a sound source in every bin is

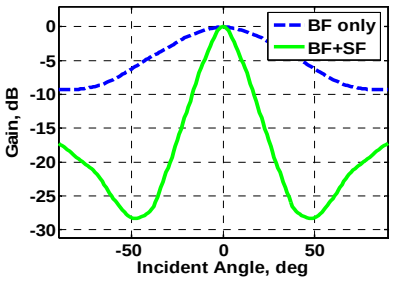

Figure 2. Directivity pattern for 1000Hz.

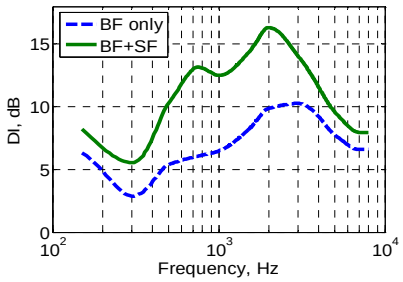

Figure 3. Directivity Index vs. frequency.

computationally expensive due to the non-linearity of (2) and complexity of the analytic form of $U_m(f,c)$. Therefore we estimate indirectly the variation $\lambda_k(\theta)$ of the sound source position in presence of $\mathbb{N}(0,\lambda_{IDOA}(k))$ from Eq. (7). Let $\lambda_k(\theta)$ and $\Upsilon_k(\theta)$ be a *K*x*N* matrices, where *K* is the number of frequency bins and *N* is the number of discrete values of the direction angle θ. Variation estimation goes trough two stages. During the first stage we build rough variation estimation matrix $\hat{\lambda}(\theta,k)$. If $\theta_{\min}$ is the angle that minimizes $\Upsilon_k(\theta)$, only corresponding to the minimum values in the rough model are updated:

$$\hat{\lambda}_k^{(n)}(\theta_{\min}) = (1-\alpha)\hat{\lambda}_k^{(n-1)}(\theta_{\min}) + \alpha\Upsilon_k(\theta_{\min})^2 \quad (9)$$

where Υ is estimated according to Eq. (8), $\alpha = \frac{T}{\tau_A}$ (τ_A is the adaptation time constant, *T* is the frame duration). During the second stage a direction-frequency smoothing filter $H(\theta,k)$ is applied after each update to estimate the spatial variation matrix $\lambda(\theta,k) = H(\theta,k) * \hat{\lambda}(\theta,k)$. Here we assume a Gaussian distribution of the non-correlated component, according to Eqns. (4) and (7), which allows us to assume the same deviation in the real space towards θ.

3.5. Likelihood estimation

With known spatial variation $\lambda_k(\theta)$ and distance $\Upsilon_k(\theta)$, the probability density for frequency bin *k* to originate from direction θ is given by:

$$p_k(\theta) = \frac{1}{\sqrt{2\pi\lambda_k(\theta)}} \exp\left\{-\frac{\Upsilon_k(\theta)^2}{2\lambda_k(\theta)}\right\}, \quad (10)$$

and for a given direction θ_S the likelihood is:

$$\Lambda_k(\theta_S) = \frac{p_k(\theta_S)}{p_k(\theta_{\min})}, \quad (11)$$

where $\theta_{\min}$ is the value which minimizes $p_k(\theta)$.

3.6. Spatio-temporal filtering

Besides spatial position, the speech signal has temporal characteristics and consecutive frames are highly correlated due to the fact that the speech signal changes slowly relatively to the frame duration. Rapid change of the estimated spatial filter can cause musical noise and distortions in the same way as in gain based noise suppressors. To reflect the temporal characteristics of the speech signal we apply time smoothing. For a given direction, we model the absence/presence of speech with two states: S_0 and S_1. The sequence of frequency bin states is modeled as first-order Markov process. Then the pseudo-stationarity property of the speech signal can be represented by $P(q_n = S_1 \mid q_{n-1} = S_1)$ with the following constraint: $P(q_n = S_1 \mid q_{n-1} = S_1) > P(q_n = S_1)$, where q_n denotes the state of *n*-th frame as either S_0 or S_1. By assuming that the Markov process is time invariant, we can use the notation $a_{ij} \triangleq P(q_n = H_j \mid q_{n-1} = H_j)$. Based on the formulations above, a recursive formula for signal presence likelihood for given look-up direction in n^{th} frame $\Lambda_k^{(n)}$ is obtained as:

$$\Lambda_k^{(n)}(\theta_S) = \frac{a_{01} + a_{11}\Lambda_k^{(n-1)}(\theta_S)}{a_{00} + a_{10}\Lambda_k^{(n-1)}(\theta_S)} \Lambda_k(\theta_S), \quad (12)$$

where a_{ij} are the transition probabilities [6], $\Lambda_k(\theta_S)$ is estimated by Eq. (11) and $\Lambda_k^{(n)}(\theta_S)$ is the likelihood of having a signal at direction θ_S for n^{th} frame. Conversion to probability gives us the estimated probability to have speech signal in this direction:

$$P_k^{(n)}(\theta_S) = \frac{\Lambda_k^{(n)}(\theta_S)}{1 + \Lambda_k^{(n)}(\theta_S)}. \quad (13)$$

The spatio-temporal filter to compute the post-processor output $Z_k^{(n)}$ from the beamformer output $Y_k^{(n)}$ is:

$$Z_k^{(n)} = P_k^{(n)}(\theta_S).Y_k^{(n)}, \quad (14)$$

i.e. we use the signal presence probability as a suppression rule which is an MMSE solution according to [7].

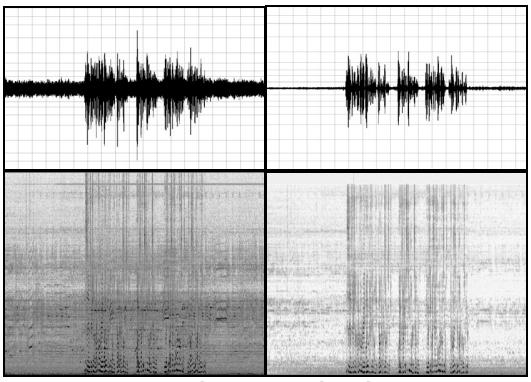

Figure 4. Input and output signals and spectrograms.

4. EXPERIMENTAL RESULTS

The proposed microphone array post-processor was evaluated with a four-element linear microphone array with length of 190 mm. The sampling rate is 16kHz, the processing is preformed in MCLT domain [8]. We used the time-invariant beamformer described in [9]. Wide-band chirp signals were recorded in an anechoic chamber, and the microphone array was rotated 10° after each recording. The recorded signals were processed twice: using only the time-invariant beamformer, and using the time-invariant beamformer and the spatial filter. The results were used to compute the directivity pattern and directivity index in eight logarithmically spaced frequency subbands. Figure 2 shows the measured directivity patterns for the band with center frequency of 1000Hz, figure 3 – the directivity index as function of the frequency. The improvement of the directivity index in the band of 500–3000Hz is 3–8dB. In another experiment we recorded a human speaker positioned 1.5m in front of the microphone array at 0°. The recording was done in a normal office. The spatial filter improved the output SNR by 6.9dB, compared with the time-invariant beamformer. In the next experiment we added a radio 2.0m from the microphone array at a -40° angle. The spectrograms of the post-processor input and output signals (estimated direction 0°) are shown on Figure 4. The post-processor suppressed the sound from the radio by 17dB. Figure 5 presents the spatial distribution of the overall likelihood of the signal as function of time and direction. The vertical axis is the sum of $\left|Y_k^{(n)}\right| P_k^{(n)}(\theta)$ for the frequency bins between 300 and 3500Hz. Multiple other experiments with varying positions of the sound source and the jammer confirmed the improvement in the suppression capabilities. Listening examinations showed minimal distortions and low level of musical noise.

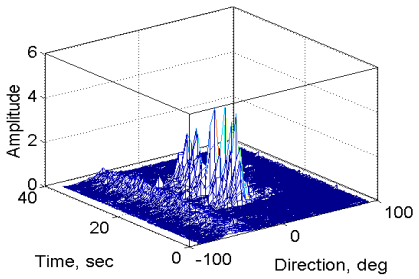

Figure 5. Spatial distribution of the signal.

5. CONCLUSIONS

We presented a novel post-processing algorithm for microphone arrays that further improves the microphone array directivity and signal separation capabilities. It works in IDOA space and is a non-linear, spatial gain based suppressor. The algorithm was implemented as real-time program and evaluated in a regular office. It showed excellent suppression capabilities while maintaining good sound quality.

6. REFERENCES

[1] H. L. Van Trees, *Detection, Estimation and Modulation Theory: Part 4, Optimum Array Processing*, John Wiley & Sons, Inc., New York, 2002.

[2] M. Brandstein, D. Ward. *Microphone Arrays.* Springer, 2001. ISBN 3-540-41953-5.

[3] C. Marro, Y. Mahieux, K. U. Simmer. "Analysis of Noise Reduction and Dereverberation Techniques Based on Microphone Arrays with Postfiltering", *Trans. on Speech and Audio Processing*, vol. 6, no. 3, pp. 240–259, May 1998.

[4] I. McCowan, H. Boulard. "Microphone Array Post-Filter Based on Noise Field Coherence". *IEEE Trans. on Speech and Audio Processing*, vol. 11, no. 6, pp. 709–716, Nov. 2003.

[5] I. Tashev, M. L. Seltzer, A. Acero. "Microphone Array for Headset with Spatial Noise Suppressor". *Proc. of IWAENC* 2005, Eindhoven, Netherlands, Sep. 2005.

[6] J. Sohn, N. S. Kim, W. Sung. "A Statistical Model-Based Voice Activity Detection", *IEEE Signal Processing Letters*, vol.6, no. 1. Jan. 1999.

[7] R. J. McAulay and M. L. Malpass, "Speech enhancement using a soft-decision noise suppression filter", *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. ASSP-28, no. 2, pp. 137–145, Apr. 1980.

[8] H. S. Malvar, "A modulated complex lapped transform and its applications to audio processing", *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Phoenix, AZ, pp. 1421–1424, March 1999.

[9] I. Tashev, H. S. Malvar. "A new beamformer design algorithm for microphone arrays", in *Proc. of ICASSP 2005*, Philadelphia, PA, USA, March 2005.