A NEW COST FUNCTION FOR DIRECTION-OF-ARRIVAL ESTIMATION OF MULTIPLE SOUND SOURCES USING TWO MICROPHONES

Minh The Dang and Seung-Hyon Nam

shnam@pcu.ac.kr
Paichai University
Dept. of Electronic Engineering
439-6 Doma-dong, Seo-gu, Daejeon, Korea

ABSTRACT

In this paper, a new cost function for estimating the direction-of-arrival (DOA) using two microphones is proposed. The proposed cost function is cross-correlation of two microphone signals after canceling a sound source at each microphone by steering the null. The signal cancellation mechanism resembles null bemaforming except a scaling factor. It is shown that the null beamformer effectively cancels early reflections as well as the direct part of a signal. Due to this property, the estimation is shifted from the true direction in a reverberant environment. Cancellation of signals provides robustness of estimation in multi-source environments. Theoretical investigation shows that local/global minima of the cost function correspond to directions of sound sources. Simulation results using real world recordings in a highly reverberant room demonstrates the effectiveness of the proposed cost function for multiple sound sources.

Index Terms— direction-of-arrival, Jeffress model, null beamforming, early reflections

1. INTRODUCTION

Direction-of-Arrival (DOA) estimation of multiple sound sources using two microphones in a reverberant environment is a challenging task. Two classical models of binaural hearing are Jeffress and Equalization-Cancelation (EC) models [1, 2, 3]. The dual-delay line structure of the Jeffress model is considered as a key concept in binaural hearing of humans. Signal cancellation by null steering in the EC model provides robustness in multi-source environments.

The time domain method by Bodden applies the Jeffress model to 24 critical bands and performs the running integration to estimate DOAs [4]. It has been reported that it successfully estimates DOAs of up to two sources in an anechoic environment. Its performance in the reverberant environments or for more than two sources is unknown. Another interesting method is to detect coincidence in a dual-delay line with the modified EC model in each frequency bin [5]. It uses time and frequency domain integrations and estimates DOAs of up to four sources even in the reverberant environments. It exploits sparseness of speech signals and detects only one dominant signal in each frequency bin.

In this paper, the EC model is further extended to improve the performance of DOA estimation in multi-source reverberant environments. Signal cancellation in the new model is analyzed in the presence of early reflections. The proposed model is then analyzed for multiple sound sources in the anechoic environment. Finally, simulation results using real world recordings are presented to confirm the analytical expectations.

2. A NEW MODEL FOR DOA ESTIMATION

2.1. Binaural Cross-Correlation Models for DOA Estimation

A basic binaural DOA estimation model is the Jeffress model which consists of a dual-delay line and a correlator at each delay time. The Jeffress model is represented by the equation [1, 3]

$$R_{\text{Jeffress}}(\tau) = \int_0^T x_L(t)x_R(t-\tau)dt \qquad (1)$$

Here, $R(\tau)$ is maximum when τ is the delay time corresponds to the source direction. The Jeffress model, however, becomes inaccurate as the number of sources increases. To accommodate the multiple sources, the Jeffress model has been extended by Colburn to the EC (Equalization-Cancellation) model described by [2]

$$R_{\text{EC}}(\alpha,\tau) = E\{x_L(t) - \alpha x_R(t-\tau)\}^2 \qquad (2)$$

where α plays a role to equalize the intensity difference between two sensors. In a free space, the source direction is dominantly governed by the interaural time difference (ITD) τ and we may omit the interaural intensity difference (IID) α. In (2), a signal is canceled by placing a null to each direction and the DOA is defined by the delay time that gives the maximum cancellation. The EC model is closely related

to the Jeffress model. That is, minimization of $R_{\rm EC}(\tau)$ corresponds to maximization of $R_{\rm Jeffress}(\tau)$. In general, however, the EC model provides more robustness against multi-source environment than the Jeffress model.

In the DOA estimation method proposed in [5], location that minimizes absolute difference between two microphone signals $|x_L(f_n,t) - x_R(f_n, t-\tau_n)|$ is detected at each frequency bin. The locations are then accumulated over time frames and frequency bins. One advantage of this method is the usage of the delta function that replaces the signal level by a location. The cost function does not depend on the energy of sound sources and weaker signals are not masked by higher energy signals if they occupy different frequency bins. However, we may have difficulty if they occupy the same frequency bins.

In this paper, we propose a new cost function for multi-source DOA estimation as follows:

$$R_{\rm NS}(\tau',\tau'') = E\{|x_L(t) - x_R(t-\tau')| \cdot |x_R(t) - x_L(t-\tau'')|\} \quad (3)$$

Unlike the above mentioned models, the new model (3) computes τ' and τ'' that minimize cross-correlation of absolute differences of dual-delay line outputs. The dual-delay line in (3) is symmetric in the sense that the difference is taking in each channel. Except the scaling factor, the dual-delay line in (3) resembles the two-channel null beamformer (NBF) that places the spatial null in the direction of an unwanted signal in each channel.

Taking absolute values in (3) is similar to the half-wave rectification process in human ears [3]. The rectification process makes the cost function nonnegative and reduces sensitivity of cross-correlation to phase ambiguity so that spurious local minima are suppressed. Although the two-dimensional cost function $R_{\rm NS}(\tau',\tau'')$ is computationally heavier than (2), we may expect the improved performance in the presence of multiple sound sources and reverberations.

2.2. The NBF output in the presence of reflections

Figure 1 shows a two-channel NBF in the presence of reflections. We consider only the direct part to express the NBF output and the result will be generalized for the case with delays and attenuations. The NBF filter at frequency ω is given by

$$\mathbf{w} = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} e^{j\omega\tau} \\ e^{-j\omega\tau} \end{bmatrix} \quad (4)$$

where the τ is zero at the phase center. The received signals at two microphones can be expressed as

$$\mathbf{x} = \begin{bmatrix} e^{-j\omega\tau_1} & e^{-j\omega\tau_2} \\ e^{j\omega\tau_1} & e^{j\omega\tau_2} \end{bmatrix} \begin{bmatrix} s_1 \\ s_2 \end{bmatrix}. \quad (5)$$

The NBF output is then given by, ignoring a constant scale factor,

$$u(\tau) = s_1 \sin(\omega(\tau-\tau_1)) + s_2 \sin(\omega(\tau-\tau_2)). \quad (6)$$

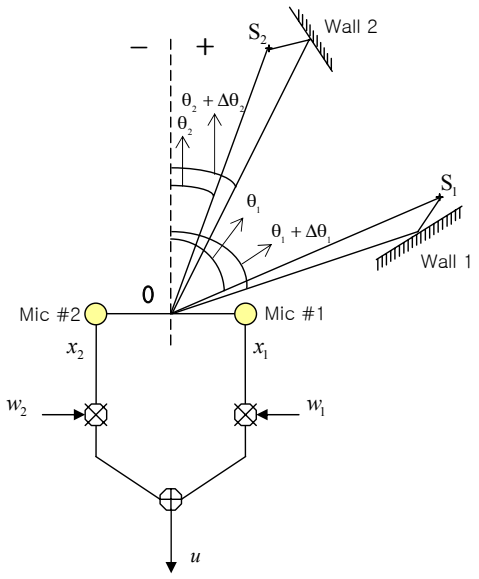

Fig. 1. *Two channel NBF in the presence of an echo.*

That is, s_1 is removed if we set $\tau = \tau_1$.

We can generalize (6) for $N > 2$ sources. Assume that $s_n, n = 1, \cdots, N$, have N_n reflection paths (early reflections) which are characterized by N_n time delays $\tau_n + \Delta\tau_{ni}$ and attenuation factors α_{ni} for $i = 1, \cdots, N_n$. Then we may express the NBF output as

$$u(\tau) = \sum_{n=1}^{N} s_n \sqrt{A_n^2 + B_n^2}\, \sin(\omega(\tau - \tau_n - \gamma_n) \quad (7)$$

where A_n, B_n, and γ_n are defined for $0 \le \omega\gamma_n \le 2\pi$ as

$$\begin{aligned} A_n &= 1 + \sum_{i=1}^{N_n} \alpha_{ni} \cos(\omega\Delta\tau_{ni}) \\ B_n &= \sum_{i=1}^{N_n} \alpha_{ni} \sin(\omega\Delta\tau_{ni}) \\ \cos(\omega\gamma_n) &= \frac{A_n}{A_n^2 + B_n^2}, \quad \sin(\omega\gamma_n) = \frac{B_n}{A_n^2 + B_n^2}. \end{aligned}$$

Therefore, the direct part and early reflections of a signal are regarded as one direct signal coming from a shifted direction $\theta_n + \Delta\theta_n$ corresponding to time delay $\tau_n + \Delta\tau_n$. Thus we may eliminate both the direct part and early reflections of s_n by setting the NBF to the shifted direction. This result is closely related to the simulations observed in the beamformer steered by eigenvectors [6].

2.3. Cross-correlation of NBF outputs

For simplicity, we slightly modify the model in (3) into

$$g(\tau',\tau'') = E\{|u(\tau')||u(\tau'')|\} \quad (8)$$

where $u(\tau) = x_L(t) - x_R(t-\tau)$ is the NBF output. Cross-correlation $g(\tau', \tau'')$ of the NBF outputs is lowered when τ' and τ'' are matched to the delays correspond to source directions θ_1 and θ_2, respectively. In the sequel, we show that delays correspond to source directions are global or local minima of the cost function $g(\tau', \tau'')$ in (8).

Case of a single source

For the case of single source, the cost function $g \equiv g(\tau', \tau'')$ is given by, from (6) and (8),

$$g = E\{|s_1 \sin(\omega(\tau' - \tau_1))||s_1 \sin(\omega(\tau'' - \tau_1))|\} \quad (9)$$

We see that $g \geq 0$ with equality holds when $\omega(\tau' - \tau_1) = k\pi, k = 0, \pm 1, \pm 2, \cdots$. The true DOA of the signal correspond to the case of $k = 0$ and non-integer k's correspond to aliased cases. The contribution from the aliased part increases as the microphone spacing increases for a fixed sampling frequency. Therefore, closer microphone spacing is desirable for this model since it reduces spatial aliasing and fulfils the requirement of the far-field assumption better.

Case of two sources

Assume that two sources s_1 and s_2 are at θ_1 and θ_2 corresponding to τ_1 and τ_2, respectively, as in Fig.1. The cost function is then given by

$$g = E\{|s_1 \sin(\omega(\tau' - \tau_1)) + s_2 \sin(\omega(\tau' - \tau_2))| \\ |s_1 \sin(\omega(\tau'' - \tau_1)) + s_2 \sin(\omega(\tau'' - \tau_2))|\}. \quad (10)$$

Using the assumption that s_1 and s_2 are independent with zero mean or $E\{s_1 s_2\} = 0$, (10) can be rewritten as

$$g = E\{|s_1^2 \sin(\omega(\tau' - \tau_1)) \sin(\omega(\tau'' - \tau_1)) \\ + s_2^2 \sin(\omega(\tau' - \tau_2)) \sin(\omega(\tau'' - \tau_2))|\}. \quad (11)$$

Notice that $g = 0$ identically at $(\tau', \tau'') = \{(\tau_1, \tau_2), (\tau_2, \tau_1)\}$. Thus these points are two global minima since the cost function g is symmetric about the straight line DOA2 = DOA1. However, $(\tau', \tau'') = \{(\tau_1, \tau_1), (\tau_2, \tau_2)\}$ may or may not be local minima in general.

Fig.2 shows the cost function for two speech sources of 1sec long at $(-50°, 30°)$ in an anechoic environment. Time delay is replaced by the corresponding angle in the plot for clarity. Microphone spacing is 15cm and the sources are located 1m apart from the microphones. Sampling rate of 16kHz has been used. It is clear that two global minima and two local minima correspond to source locations as expected. This enables us to estimate DOAs of weak signals masked by a strong dominant signal.

For the general case ($N > 2$), g in (11) is generalized as

$$g = \sum_{i=1}^{N} E\{|s_i^2 \sin(\omega(\tau' - \tau_i)) \sin(\omega(\tau'' - \tau_i))|\}. \quad (12)$$

It is not clear whether $(\tau', \tau'') = (\tau_i, \tau_j)$ for $i, j = 1, \cdots, N$ $i \neq j$, correspond to global/local minima or not. However, at

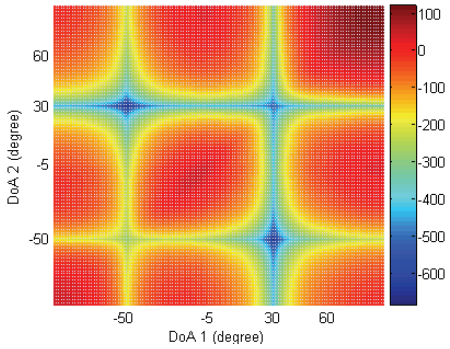

Fig. 2. *Plot of $g(\tau', \tau'')$ for two speech sources at $(-50°, 30°)$ in an anechoic room.*

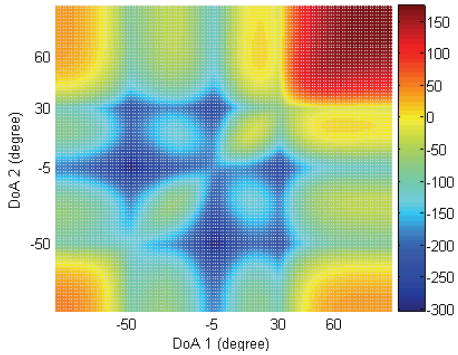

Fig. 3. *Plot of $g(\tau', \tau'')$ for three speech sources at $(-50°, -5°, 30°)$ in an anechoic room.*

$(\tau', \tau'') = (\tau_i, \tau_j), i \neq j$, two signals s_i^2 and s_j^2 are canceled simultaneously and g is lowered. Therefore, these are good candidates for global/local minima. As an example, Fig. 3 for three sources at $(-50°, -5°, 30°)$ reveals that any combination of two DOAs out of three corresponds to global or local minimum.

Reverberation of a room hinders DOA estimation procedures and deteriorates accuracies. To see the effect of the reverberation, we computed the cost function in a real environment with reverberation time of 500msec for the same sources used in Fig. 2. Comparing Fig. 2 and 4, we observe that global minimum is shifted to $(28°, -46°)$ from $(30°, -50°)$ as expected in (7). In addition, the cost function for the echoic case is much noisier than that of the anechoic case, which implies the existence of spurious local minima. Fortunately, however, g at source directions are much smaller than those of spurious minima and we may devise a good estimation procedure.

3. EXPERIMENTS

We investigated the performance of the proposed cost function in an office with reverberation time of 500msec. Four sources were placed at 1m apart from two microphones in the

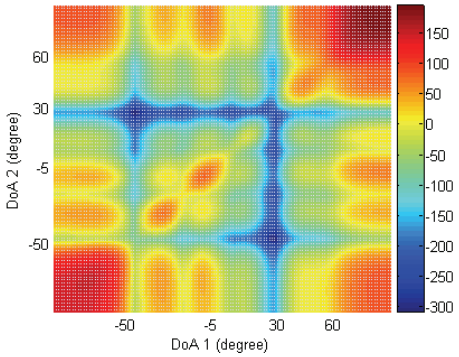

Fig. 4. *Plot of $g(\tau', \tau'')$ for two speech sources at $(-50°, 30°)$ in a real room with reverberant time of 500msec.*

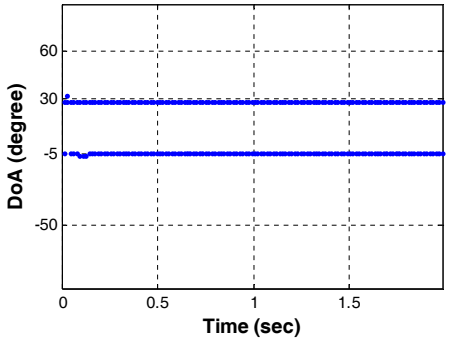

Fig. 5. *DOA estimation by the proposed method when the energy of the source at $30°$ is 5dB larger than that of the source at $-5°$.*

Table 1. Performance of the proposed cost function in a real reverberant environment.

no.src	Setup DOA	MUSIC(each)	Proposed
1	-5	-5	-5
	60	54	55
2	(-5,30)	(-5,28)	(-5,28)
	(-50,30)	(-46,28)	(-46,28)
3	(-50,-5,30)	(-46,-5,28)	(-45,-5,28)
4	(-50,-5,30,60)	(-46,-5,28,54)	(-44,-5,28,55)

directions of $-50°$, $-5°$, $30°$ and $60°$. Microphone spacing is set to 15cm. Room impulse responses were measured and convolved with speech signals of 2sec long taken from the TIMIT data sampled at 16kHz. Each signal was set to have the same power.

We used the MUSIC algorithm as a reference [7]. Since the MUSIC requires more microphones than the the number of sources, we computed each DOA one by one. To obtain the DOA estimates using the proposed cost function, we computed accumulated the cost function (8) at each frame of 512 samples with 50% overlap with a forgetting factor. Then a histogram on DOA candidates is used to find the final DOA estimates at each frame.

Table. 1 shows the results for various number of sources. Each result is the average of eight different combinations of speech signals. It is observed that the estimates with the proposed method are equal or very close to those with the MUSIC algorithm for all cases.

To test the performance of the proposed method for weak signals, we considered two signals at $30°$ and $-5°$ in the same environment as in Table 1 but with different signal power. We set the signal at $30°$ larger than that of the signal at $-5°$ by 5dB. As shown in Fig. 5, the proposed method accurately estimate the DOA of the weak signal in the presence of the strong signal. It is also observed that the proposed method provides accurate estimates even if a signal duration is very short.

4. CONCLUSION

We proposed a new cost function for DOA estimation using two microphones in reverberant multi-source environments. The cost function is a cross-correlation of the outputs of a dual-delay line that resembles a two-channel null beamformer. It is shown that the two-channel null beamformer can capture early reflection as well as the direct part of a signal with a shifted steering vector. Signal cancellation in the cost function provides robustness of estimation in the presence of multiple sound sources. Simulation results using real recordings confirmed that proposed method can estimate DOAs of up to four sources very accurately even in a highly reverberant environment. It also works well for weak signals.

5. REFERENCES

[1] Jeffress L.A., "A place theory of sound localization," *J. of Comp. Physiol. Psychol.*, pp. 35–39, 1948.

[2] Colburn H.S. and Durlach N.I., "Models of binaural interaction," in *Handbook of Perception: Hearing*, Carterette E. and Friedman M., Eds., vol. 4. Academic Press, 1978.

[3] Brown G.Y. Stern R.M. and Wang D., "Binaural sound localization," in *Computational Auditory Scene Anaylsys*, Wang D. and Brown G.J., Eds., pp. 147–185. Wiley Interscience, 2006.

[4] Bodden M., "Modeling human sound source locaization and the cocktail-party-effect," *Acta Acust. (China)*, pp. 43–55, 1993.

[5] Liu C. *et. al.*, "Localization of multiple sound sources with two microphones," *JASA*, pp. 1888–1905, 2000.

[6] Warsitz E. and Haeb-Umbach R., "Blind acoustic beamforming based on generalized eigenvalue decomposition," *IEEE Trans. Audio Speech Lang. Processing*, pp. 1529–1539, July 2007.

[7] Schmidt R.O., "Multiple emiter location and signal parameter estimation," in *Proc. RADC Spectral Estimation Workshop*, 1979, pp. 243–258.