A DOA Estimation Algorithm based on Equalization-Cancellation Theory

Duc Thanh Chau, Junfeng Li, Masato Akagi

School of Information Science, Japan Advanced Institute of Science and Technology

{ctduc,junfeng,akagi}@jaist.ac.jp

Abstract

Direction of arrival (DOA) estimation plays an important role in multi-channel (binaural) speech enhancement systems and auditory humanoid robots. A number of localization methods have been presented, however, most of them require a large array of microphones, or cannot adapt to some special conditions, e.g., humanoid robot with the effect of head-related transfer function (HRTF). In this paper, we propose a two-microphone DOA estimation algorithm, namely EC-Beam, which applies equalization-cancellation (EC) model to DOA estimation through beamformer-based technique. Specifically, the EC model is integrated into beamforming to remove the signal components from a given direction and yield the energy of the remaining signals from other directions. Through searching several DOA candidates in the space, the estimation of DOA is finally determined as the direction at which the energy of the remaining signal reaches the minimum. Interpolation method is further exploited in EC-Beam to estimate non-beamformed directions. Experimental results showed that the EC-Beam with only two microphones is able to estimate accurately the DOA of target signal in various noise conditions, and well adapted to binaural hearing systems.

Index Terms: DOA Estimation, Beamformer-Based Localization, Equalization-Cancellation Model.

1. Introduction

In recent years, DOA estimation has been widely exploited in binaural speech enhancement and humanoid robots [9, 14]. For binaural speech enhancement, there are some methods which performed well but require DOA information of target speech, since they enhance signal through a priori known direction [9]. DOA information of speech is also required to construct robots with human-like behaviors [14], for example, a robot should face the speaker during communication. Such systems require a DOA estimation method which uses two microphones, is robust under noisy conditions, and adapts to system effects (e.g. HRTF effect, body effect). So far, most DOA estimation methods rely on microphone array, few of them consider the effects caused by the systems. Recently, F. Keyrouz *et al.* proposed Inverse-HRTF method based on HRTF filters which can deal with HRTF effect [5]. Although the Inverse-HRTF gave a relatively highly-accurate estimation with two microphones, it is limited to artificial dummy head systems, since training inversion filters for an arbitrary system (not HRTF-like systems) is a complicated task.

For sound localization (SSL) in general, including DOA estimation, existing procedures can be loosely classified into three general categories: those based on steered response power (SRP) of a beamformer, techniques adopting high-resolution spectral estimation concepts, and approaches employing time-difference of arrival (TDOA) information [10]. Among them, the TDOA based methods have received extensive investigation with several well-known algorithms, for example, Generalized Cross Correlation (GCC), GCC with Phase Transform weighting (GCC-PHAT) [4], and SRP-PHAT [7]. However these methods just deal with low noise environments and do not consider the effects caused by the system's shape. While techniques in the second category are limited to the far-field, statistically stationary source and noise, and especially, less robust to source and sensor modeling errors [10], the beamforming-based category is considered to be a good choice, and widely used for practical systems. Although beamformer-based procedures are potentially robust under noisy conditions and can deal with multiple sources [11], they also suffer from some limitations. Conventional beamforming methods usually require a large array of microphones; this high complexity makes them impractical for real-time systems.

In psychoacoustics, human perception is simulated by Equalization-Cancellation (EC) Theory [12, 15], and based on it, many speech-processing applications have been created, especially beamforming-based applications. Nowadays, EC model is exploited widely in speech enhancement [9], and signal detection [12].

Motivated by EC theory and taking advantage of beamforming strategy, we proposed a DOA estimation algorithm with two microphones, namely EC-Beam. Basically, the EC-Beam integrates EC model into each beam in the searching process to remove the signal coming from beam's direction. The remaining energy of a beam should be smallest if its direction is toward the true sound source. After several pre-defined directions are beamformed, interpolation technique is applied to improve search resolution. Finally, the true DOA is realized as the direction at which the beam's power reaches the minimum. Experimental results confirmed that EC-Beam, with only two microphones, can accurately estimate the DOA of target signal under various kinds of noisy conditions.

2. Equalization-Cancellation Model

The equalization-cancellation (EC) model was originally developed by Durlach [13] and further improved by Culling and Summerfield [8]. In the original EC model, when subject is presented with a binaural-masking stimulus, the auditory system attempts to eliminate the masking components by transforming the signal arriving at one ear relative to the signal at the other ear to make the masker components "equalized" (the E process). Then part of the signal in each ear is cancelled by subtracting the signal in the other ear (the C process) [13]. This model was recently improved in [8] where the E and C processes were independently performed for the interfering signals in each channel. Research showed that these EC models can explain many psychoacoustic effects, such as *binaural masking level difference* (BMLD), etc [8, 13].

3. Proposed EC-Beam algorithm

Basically, like conventional beamformer-based algorithms, the EC-Beam also performs beamforming in several selected directions in searching plane and estimates DOA by finding a global peak. The speciality of EC-Beam is integrating EC-

model into beamforming process. The interpolation technique is exploited to reduce computational expense and increase resolution of searched space. The proposed algorithm is performed through three main stages: beamforming with EC-model; interpolation for non-beamformed directions; and DOA estimation by searching the minimum peak of beam values.

3.1. Beamforming with EC-model

The number of beams to be made on the search plane depends on the expected precision that the algorithm should have. It is clear that the higher the resolution of beams, the higher accuracy will be achieved. For each beam to a given direction, EC model is applied to remove the signal from that direction and yields *remaining signals* (which will be called *filtered-signal*) from other ones (**Figure 1**). Theoretically, when such a *null-beam* is pointing in the direction of the largest sound-source, the energy of *filtered-signal* should reach the minimum because the biggest signal has been removed.

In binaural hearing, the characteristics of sound coming from a direction are involved in the differences in amplitude and phase of signals at the left and right ears. To remove the sound from a direction, the Cancellation process needs two equalizers which have to be pre-trained in the Equalization process.

For a given direction θ, assume that the binaural signal model can be expressed as:

$$X_i(k,\ell,\theta) = S_i(k,\ell,\theta) + N_i(k,\ell,\theta), \quad i = L,R \qquad (1)$$

where k and ℓ respectively denote the frequency bin index and the frame index, $S_i(k,\ell,\theta)$ and $N_i(k,\ell,\theta)$, $i = L,R$, are the short-time Fourier transforms (STFTs) of signal of direction θ and signals from other directions. The EC processes will be performed as explained below.

3.1.1. Equalization Process

This process aims to construct two equalizers which equalize the signal components from the left input and those from the right input. After compensation for the differences in intensity and phase of target signal components from both ears, the two equalizers $W_R(k,\ell,\theta)$ and $W_L(k,\ell,\theta)$ should satisfy the following equations:

$$S_L(k,\ell,\theta) - W_R(k,\ell,\theta)S_R(k,\ell,\theta) = 0 \qquad (2)$$

$$S_R(k,\ell,\theta) - W_L(k,\ell,\theta)S_L(k,\ell,\theta) = 0 \qquad (3)$$

Specifically, these equalizers are obtained using the *normalized least mean square* (NLMS) algorithm, which is given as (θ is omitted for simplicity)

$$W_L(\ell+1) = W_L(\ell) + \mu\frac{X_L(\ell)}{\|X_L(\ell)\|^2}[X_R(\ell) - W_L^T(\ell)X_L(\ell)] \qquad (4)$$

$$\mathrm{W_R}(\ell+1) = \mathrm{W_R}(\ell) + \mu\frac{\mathrm{X_R}(\ell)}{\|\mathrm{X_R}(\ell)\|^2}[\mathrm{X_L}(\ell) - \mathrm{W_R^T}(\ell)\mathrm{X_R}(\ell)] \qquad (5)$$

where $W_i(\ell) = [W_i(1,\ell), W_i(2,\ell), \dots, W_i(K,\ell)]^T$ and $X_i(\ell) = [X(1,\ell), X(2,\ell), \dots, X_i(K,\ell)]^T$, $i = L,R$. K is the STFT length, superscript T denotes the transposition operator and μ is the step size.

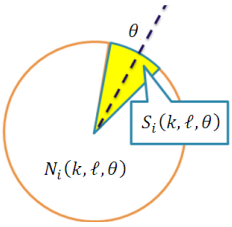

Figure 1: *Cancellation of signal from direction θ*

The Equalization is performed beforehand with clean signal (each signal contains only one speech from only one direction) in a *training process* (offline). This process yields the equalizers which will be used in the Cancellation process.

3.1.2. The Cancellation Process

This process applies two equalizers to cancel signals from the given direction (target signal) out of observed signals. Since the equalizers have been trained to satisfy the equations (2) and (3), it is expected that the left (right) *filtered-signal* will be equivalent (or at least approximate) to the right (left) input signal. Consequently, the target-cancelled signals are derived by the following formulas:

$$\begin{aligned} Z_L(k,\ell,\theta) &= X_L(k,\ell,\theta) - W_R(k,\ell,\theta)X_R(k,\ell,\theta) \\ &\approx N_L(k,\ell,\theta) - W_R(k,\ell,\theta)N_R(k,\ell,\theta) \end{aligned} \qquad (6)$$

$$\begin{aligned} Z_R(k,\ell,\theta) &= X_R(k,\ell,\theta) - W_L(k,\ell,\theta)X_L(k,\ell,\theta) \\ &\approx N_R(k,\ell,\theta) - W_L(k,\ell,\theta)N_L(k,\ell,\theta) \end{aligned} \qquad (7)$$

Equations (6) and (7) indicate that, in the *filtered-signals*, the signal from direction θ was completely removed. These *filtered signals* may not be exactly equal to the remaining signals from all other directions received at left and right ears, but it is clear that their energy will be significantly reduced if θ is the true direction to largest sound source.

After cancellation process, the remaining energy of the *null-beam* to θ is computed by following formula

$$\mathcal{P}(\theta) = \sum_{k,l}[Z_L(k,\ell,\theta)^2 + Z_R(k,\ell,\theta)^2] \qquad (8)$$

3.2. Interpolation for non-beamformed directions

One disadvantage of beamforming methods is the high complexity because of their exhaustive search. In a searching plane, beamforming for all possible directions is not practical for real-time systems. Hence, one of possible solutions is interpolation method which, given two beamforming values of two neighboring directions, could correctly produce a good number of beam values between them. The purpose of this stage is to obtain beam values for DOAs that were not covered by the training set, with little computational expense.

Within this research, the cube spline interpolation [3] has been applied. After the interpolation process, a grid of beamforming values in the space with a suitable resolution is obtained. This grid will be used in DOA estimation stage.

3.3. DOA estimation

The problem of DOA estimation now can be considered as finding a direction whose equalizers "match" with the true sound-source. The equalizers of a direction will match to the true sound-source when the beam to that direction gets to minimum power. Once a full grid of beam values is obtained, by finding its minimum value, the direction of the sound source is also determined.

$$\hat{\theta} = \arg\min_\theta \mathcal{P}(\theta) \qquad (9)$$

4. Experiments and results

The proposed EC-Beam algorithm was examined under various conditions and compared with the well-known GCC-PHAT algorithm. To evaluate the performance of EC-Beam and its robustness under noisy environments as well, a number of experiments have been carried out with clean and noisy conditions. The adaptability of the proposed algorithm is also

evaluated by comparing the estimation results of *in-ear* signals with those of *behind-the-ear* signals.

4.1. Experiments Part I

The purpose of this section is to test EC-Beam with signals affected by HRTF recorded from microphones placed *in ear* of artificial dummy head, and test its robustness under noisy conditions. In these experiments, the KEMAR HRTF database measured at 44.1 KHz of MIT [2] was applied to synthesize speech signals. Regarding DOA estimation, we just used the HRTF measurement in horizontal plane (0^0 elevation) with 5^0-intervals in an azimuth range from -90^0 to 90^0. For speech, we collected 110 recorded raw audio samples from 11 Japanese speakers in which each speaker has 10 samples with 5 vowel samples (e.g. /a/) and 5 phrase (or short sentence) samples. For each sample, by convoluting with the HRIR, we created 37 signals for 37 directions from -90^0 to 90^0. In total, 4070 signals were created, of which 370 signals were used for training to obtain 37 pairs of equalizers (left and right) corresponding to those directions. The 3700 remaining signals then were used to produce testing data for each experiment below.

4.1.1. Clean condition

In this experiment, the testing data are the original dataset without interference signals. To confirm whether the EC-Beam can well estimate the DOA in cases in which the directions of observed signals have not been trained, we just used the trained equalizers at 10^0-intervals from [-90^0, -80^0,…, 90^0] and applied interpolation to get 1^0-interval grid. In the result, as shown in **Figure 2**, although the equalizers at azimuth -85^0, -75^0, …,85^0 had not been applied, these directions were also correctly estimated. Consider that the estimation is correct if the difference between estimated DOA and real DOA does not exceed 5^0, the accuracy (Table 1) in this case is relatively high, 98.21%. Also in Table 1, the *Standard Deviation* is only 1.29, which means that the error does not change so much among those directions.

4.1.2. Noisy conditions

In these experiments, the signals in the original dataset were mixed together to obtain noisy data. For one-source noise, once a signal of a speaker was considered as target, another signal from another speaker was added as noise. The direction of noise was fixed to 60^0, while the direction of the target varied from -90^0 to 90^0 (5^0-intervals). When mixing these signals, the amplitude of noise was controlled to make the Signal-to-Noise Ratios (SNR) of 5dB, 10dB, 15dB and 20dB. At each SNR level, a total 3700 signals were created. For two-source noise signals, mixing method was also performed in the same way, but the directions of noises were fixed at -30^0 and 60^0.

Table 2 and Table 3 show that the accuracy of estimation in these noisy conditions does not decrease very much compared to clean condition, and when the SNR exceeds 10dB, the estimation result becomes closer and closer to that of clean data.

To evaluate EC-Beam in real noisy conditions, another experiment was designed with real noise recorded in the cafeteria of Japan Advanced Institute of Science and Technology (JAIST). In mixing process, the noise amplitude was also controlled to get SNR of 5dB, 10 dB, 15 dB and 20 dB. The detailed results are shown in Table 4, in which there is almost no difference between the results in this case and those of two-source noise condition.

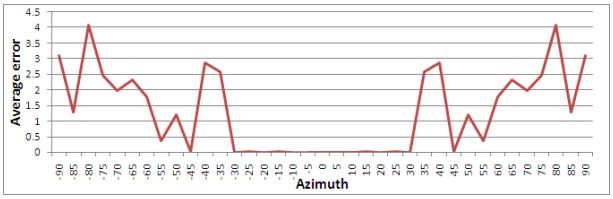

Figure 2: *Average estimation error of clean signals*

Table 1: *EC-Beam results with clean signals*

Average Error	Average Std.	Accuracy
1.29	1.29	98.21%

Table 2: *EC-Beam with one-source noisy signals*

SNR	Average Error	Average Std.	Accuracy
5 dB	4.02	3.47	89.95%
10 dB	1.55	1.48	97.27%
15 dB	1.33	1.32	98.08%
20 dB	1.28	1.29	98.19%

Table 3: *EC-Beam with two-source noisy signals*

SNR	Average Error	Average Std.	Accuracy
5 dB	2.84	2.94	93.19%
10 dB	1.45	1.35	97.70%
15 dB	1.31	1.31	98.14%
20 dB	1.29	1.29	98.14%

Table 4: EC-*Beam with In-ear cafeteria noisy signals*

SNR	Average Error	Average Std.	Accuracy
5 dB	3.05	3.41	92.11%
10 dB	1.43	1.46	97.43%
15 dB	1.31	1.31	98.11%
20 dB	1.29	1.29	98.11%

In general, the results for the cases of two-source noisy and cafeteria noisy signals (Table 3, 4) are higher than those of one-source noisy signals (Table 2). That means, at the same SNR, EC-Beam has higher accuracy when noise is diffused.

4.2. Experiments Part II

This experiment aims to evaluate the adaptability of EC-Beam with different systems. The database used in this experiment is HRIR database from University of Oldenburg [6], in which recording system had 8 microphones, with 2 inside-ear microphones and 6 behind-the-ear mikes. In order to test EC-Beam with signals under effects other than in-ear HRTF (like KEMAR Database), we used the recorded signals from the first 2 of 6 behind-the-ear microphones in “Anechoic” set. The dataset was created in the same way as in Experiment Part I. We also used 370 signals (of one speaker) for training, and the 3700 remaining signals for testing. Since the robustness of the proposed algorithm under noisy conditions was confirmed by experiments in 4.1.2, this experiment was carried out with only clean data.

The result shown in Table 5 indicates that although the accuracy decreased a little, the average error and its standard deviation remained low. Moreover, as shown in Figure 3, the average errors of all directions are still low (less than 4).

Table 5: *EC-Beam with behind-the-ear clean signals*

Average Error	Average Std.	Accuracy
1.66	1.19	92.97%

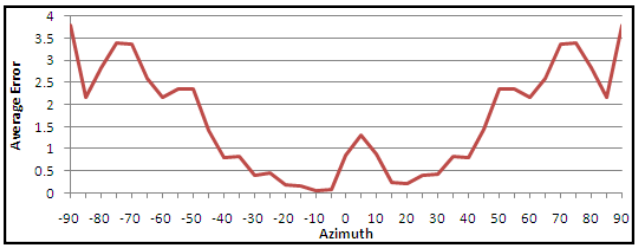

Figure 3: *Average estimation error of behind-the-ear clean signals*

4.3. Comparison to GCC-PHAT

Reported in many studies, SRP-PHAT is considered to be one of the most effective algorithms in sound source localization [1, 10]. Mathematically, SRP-PHAT is a version of GCC-PHAT in which the system has more than two microphones. In order to compare with EC-Beam, an experiment was carried out using the GCC-PHAT with clean signals discussed in section 4.1.1.

Figure 4 shows the comparison of EC-Beam and GCC-PHAT. Overall, the average estimation error of EC-Beam is much lower than GCC-PHAT especially in the ranges of [-90^0 ~ 65^0] and [65^0 ~ 90^0]. There are two main reasons that GCC-PHAT performed poorly (accuracy 45.08%) in this case:

- The first and most important is the effect of HRTF. Because GCC-PHAT is designed for sound localization in the general case (microphone array), and not for the case in the presence of HRTF, the more HRTF effect there is, the more estimation error there is using GCC-PHAT. And it is clear that the two azimuth ranges above are the regions which are most affected by HRTF.
- The second reason is that the distance between two microphones was too short (for a dummy head, it is usually shorter than 0.2 m). For normal speech with medium wave length, the short distance between microphones causes low resolution in performance of GCC algorithms.

5. Conclusions

Recently, a number of sound localization methods have been presented, but few of them can be implemented widely in reality because of some limitations: the number of microphones required, the robustness under noisy conditions, and the adaptability to practical systems. In this paper, we proposed a DOA estimation algorithm with only two microphones, called EC-Beam, based on Equalization-Cancellation model and beamforming strategy. In EC-Beam, the main steps are quite similar to conventional beamformer-based methods, but it is specialized by integrating EC-model into the beamforming process to remove the energy of target signal from the direction of each beam and yield energy of remaining signals from other directions. The interpolation method is further applied to reduce computational cost and increase search resolution. Experimental results confirmed that the proposed method, with two microphones, can estimate DOA of speech accurately even in high noise conditions, and is potentially well-adapted to practical systems.

6. References

[1] A. Badali, J.-M. Valin, F. Michaud and P. Aarabi, "Evaluating real-time audio localization algorithms for artificial audition on mobile robots," IEEE Int. Conf. on Intelligent Robots and Systems, 2009.

[2] B. Gardner and K. Martin, "HRTF Measurements of a KEMAR Dummy Head Microphone", Available at http://sound.media.mit.edu/KEMAR.html, Accessed April, 2010.

[3] C. Boor, A Practical Guide to Splines, Springer-Verlag, 1978.

[4] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," IEEE Trans. Acoustic Speech Signal Processing, vol. ASSP-24, pp. 320-327, 1976.

[5] F. Keyrouz, Y. Naous and K. Diepold. "A new method for binaural 3D localization based on HRTFs," ICASSP, vol. 5, pp. 341-344, Toulouse, France, May 2006.

[6] H. Kayser et al., "Database of Multichannel In-Ear and Behind-the-Ear Head-Related and Binaural Room Impulse Responses," EURASIP Journal on Advances in Signal Processing, Volume 2009.

[7] J. DiBiase, A High-Accurate, Low-Latency Technique for Talker Localization in Reverberation Environments Using Microphone Array, PhD thesis, Brown University, Providence RI, USA, 2000.

[8] J. Culling and Q. Summerfield, "Perceptual segregation of concurrent speech sounds: absence of across-frequency grouping by common interaural delay," JASA, vol. 98, pp. 785-797, 1995.

[9] J. Li, S. Sakamoto, S. Hongo, M. Akagi and Y. Suzuki, "Two-stage binaural speech enhancement with Wiener filter for high-quality speech communication," In Press, Speech Communication, 2008.

[10] M. Brandstein and D. Ward, Microphone Arrays, Digital Signal Processing, Springer, ISBN 3-540-41953-5, pp.157-201, 2001.

[11] M. Wax and T. Kailath, "Optimum localization of multiple sources by passive arrays," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-31, pp. 1210-1217, 1983.

[12] N. Durlach, "Binaural signal detection: Equalization and cancellation theory", In J. V. Tobias Editor, Foundations of Modern Auditory Theory, vol.2 ,pp. 369-462, Academic Press, New York, 1972.

[13] N. Durlach, "Equalization and cancellation theory of binaural masking level differences," JASA, vol. 35, no. 8, pp. 1206-1218, 1963.

[14] V. Trifa, A. Koene, J. Moren and G. Cheng. "Real-time acoustic source localization in noisy environments for human-robot multimodal interaction," The 16th IEEE International Symposium, Jeju Island, Korea, 393-398. 2007.

[15] W. Kock, "Binaural localization and masking," Journal of the Acoustical Society of America, vol. 22, pp. 1-804, 1950.

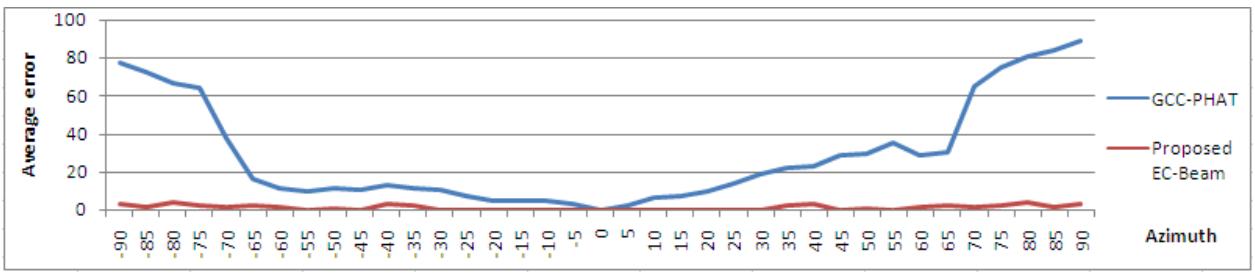

Figure 4: *Comparison of EC-Beam and GCC-PHAT*