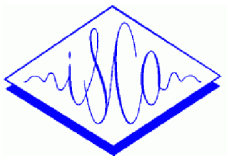

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Direction of Arrival Estimation Based on Subband Weighting for Noisy Conditions

Wei Xue[1], *Wenju Liu*[2]

National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

{wxue,lwj}@nlpr.ia.ac.cn

Abstract

In this paper, we present a novel DOA estimation method for human speech using subband weighting. Existing DOA estimation methods still can not perform quite reliably in low SNR condition. To improve the robustness of DOA estimator in noisy environment, we propose a novel DOA estimation approach. Firstly, the speech signal of each channel is passed through a Gammatone filterbank to obtain a set of time-domain subband signals. Secondly, we achieve TDOA estimation based on a new cost function in each subband. Subband weight is calculated to emphasis the estimation results of subbands with high probability containing speech signals. Finally, DOA is determined by the estimated TDOA and geometry of microphone array. Experimental results show that the proposed subband weighting based method outperforms SRP-PHAT and broadband MUSIC algorithm in highly noisy environment.

Index Terms: Direction of arrival estimation, array signal processing, Gammatone filters, subband weighting.

1. Introduction

Direction of arrival (DOA) estimation, which aims at determining the direction of sound source with microphone arrays, has received much attention due to its wide applications. It plays an important role in video conferences, hand free devices, or interactive robots. The estimation is often considered to be worked in noisy environment.

In general, DOA methods are classified into three categories: high-resolution spectral (HOS) estimation [1, 2], steered beamformer response power (SRP)[3, 4], and time difference of arrival (TDOA) estimation [5, 6]. A well-known HOS method is MUSIC algorithm [1]. The MUSIC algorithm is based on subspace analysis of spatial correlation matrix, and has achieved great success for narrowband signals. Although some broadband approaches [7, 8] are presented, it still suffers from performance degradation when applied to broadband signals. HOS methods also rely highly on the source number estimation accuracy and stationarity assumption of noise signal. SRP methods steer a spatial beamformer over all predefined directions to obtain a steered response, and estimate direction with the largest steered response as DOA. As a typical algorithm of SRP methods, SRP-PHAT [9] has been proved to be robust in moderate noisy environments, but the computational complexity is high when the element number is large. TDOA based estimation is a two step procedure. TDOAs are calculated firstly, then DOA is obtained according to the TDOAs and geometry of microphone array. The most widely used TDOA estimation algorithm is GCC methods [10]. They work well in moderately noisy and non-reverberant environments, but degrade much when noise level or reverberation is high.

In this paper, we propose a novel DOA method for human speech using subband weighting. The new approach is in framework of TDOA estimation. Firstly, the speech signal of each channel is passed through Gammatone filters to obtain time-domain subband signals. Secondly, TDOA estimation is performed on filtered signals in each subband. A new TDOA method for multichannel signals is used here. Finally, the subband weight is calculated to emphasis the estimation results of subbands which have high probability containing speech signals. In this way, subband estimation results with higher degree of confidence have a more prominent impact on the final estimation. Experimental results show that the proposed subband weighting based method outperforms SRP-PHAT and broadband MUSIC algorithm in highly noisy conditions.

The rest of the paper is organized as follows. Section 2 formulates the problem and gives some assumptions. In section 3 we introduce the proposed subband weighting based DOA algorithm. Experimental results are given in section 4, and finally section 5 concludes the paper.

2. Problem formulation and assumptions

Suppose there exists a uniform linear microphone array with N elements and the sound source is in the far field [11]. The speech source signal propagates radiatively and for the straight propagation path the sound level falls off as a function of distance from the source. The signal captured by the nth microphone at time k can be expressed as follows:

$$y_n(k) = a_n s(k - \tau_n) + v_n(k), n = 1, 2, \ldots, N \quad (1)$$

where a_n is the attenuation factor, τ_n is the propagation time between source and the nth microphone, and $v_n(k)$ is additive noise at the nth microphone. In ideal cases, a_n is set to be 1, and the additive noise is assumed to be uncorrelated with both source signal and noise signals received by other microphones.

The time difference of arrival between the ith and jth microphones is defined as

$$\tau_{ij} = \tau_i - \tau_j \quad (2)$$

For a linear and equispaced array, choosing the first microphone as reference, the TDOA between the first and nth microphone is calculated as

$$\tau_{i1} = (n - 1)\tau, n = 2, 3, \ldots, N \quad (3)$$

where τ is the TDOA between the first two microphones.

With the estimated TDOA, the direction of arrival is given according to the geometry of microphone array:

$$\theta = \arcsin(\frac{\tau c}{f_s d}) \tag{4}$$

where c is the propagation speed of sound in the air, which is usually set to 343m/s, f_s is the sampling rate, d is the distance between two adjacent microphones, and θ is the angle from normal of array to the wave ray, ranging from -90° to 90° with interval of 1° .

3. DOA estimation based on subband weighting

The proposed algorithm is based on the fact that different subbands of the signals are not affected by noise equally, and those containing more speech tend to be more robust. Thus by emphasizing the effect of estimation results on these subbands, better performance in noisy environment can be expected. Methods of decomposing the broadband signal into subband signals, time difference of arrival estimation on each subband, and adaptive subband weight calculation will be discussed in this section.

3.1. Signal decomposition

In order to decompose the received broadband signal into a set of narrowband signals, a filterbank consisting of 64 overlapping bandpass Gammatone filters is used here. The reason for choosing Gammatone filterbank is that it simulates the cochlea analysis of human ear effectively, and has been widely used in many acoustic analysis systems. The center frequencies of these filters are uniformly spaced on the equivalent rectangular bandwidth (ERB) scale between 50 Hz and 8000 Hz [12]. By passing the broadband signal through these filters, 64 narrowband signals are obtained.

We know that, each Gammatone filter in the filterbank is uniform, so when passing the two channel broadband signals through the same Gammatone filter, the same phase shift is obtained for broadband signals. Therefore, the TDOA of broadband signals is preserved after being decomposed into a set of narrowband signals, and it is reasonable to estimate TDOA in each subband.

3.2. Time difference of arrival estimation

3.2.1. Subsample shifting

TDOA estimation algorithms always involve shifting signal of one channel with a hypothesized TDOA to align with signal of another channel. As we aim at estimating DOA ranging from -90° to 90° with interval of 1°, the TDOAs corresponding to some azimuths are not integer multiples of the sampling interval. However, the discrete signal cannot be shifted by decimal samples directly, as a result, subsample shifting is needed.

We cope with the subsample shifting problem based on Discrete Fourier Transform (DFT) and Inverse Discrete Fourier Transform (IDFT). According to the property of Fourier transform, time delay corresponds to phase shift in the frequency domain, so given a signal $s(k)$, subsample shifting formula can be simply expressed as:

$$\hat{s}(k) = s(k-\tau) = IDFT\left\{S(\omega)e^{-j\omega\tau}\right\} \tag{5}$$

where $\hat{s}(k)$ is the shifted signal of $s(k)$, τ is the TDOA, $IDFT$ denotes inverse Fourier transform, and $S(\omega)$ is the Fourier transformation of $s(k)$.

3.2.2. Cost function for TDOA

Time delay estimation is operated in each subband on multichannel subband signals. Similar to [9], a cost function using the spatial correlation matrix (SCM) for multichannel case is proposed here.

Define a $L \times N$signal matrix consisting of subband signals from different channels

$$\mathbf{y}(k,\tau,f) = [y_1(k,f)\ y_2(k+\tau,f)\ \ldots\ y_N[k+(N-1)\tau,f]]^T \tag{6}$$

where τ is the hypothesized TDOA, N is the number of channels in the microphone array, $y_i(k,f)$ is the fth subband signal vector with size $L \times 1$ from the ith channel at time k, $i = 1 \ldots N$, and L is the frame length of $y_i(k,f)$. Subband signals of channel $2 \sim N$ are shifted to align with channel 1 given τ.

The corresponding SCM is further defined as

$$\begin{aligned}
&\mathbf{R}(k,\tau,f) \\
&= E\left[\mathbf{y}(k,\tau,f)\mathbf{y}^T(k,\tau,f)\right] \\
&= \begin{bmatrix}
\sigma^2_{y_1} & r_{y_1y_2}(k,\tau,f) & \ldots & r_{y_1y_N}(k,\tau,f) \\
r_{y_2y_1}(k,\tau,f) & \sigma^2_{y_2} & \ldots & r_{y_2y_N}(k,\tau,f) \\
\vdots & \vdots & \ddots & \vdots \\
r_{y_Ny_1}(k,\tau,f) & r_{y_Ny_2}(k,\tau,f) & \ldots & \sigma^2_{y_N}
\end{bmatrix}
\end{aligned} \tag{7}$$

It is clear that only if two or more signals are perfectly aligned, $det[\mathbf{R}(\tau,f)]$ equals to 0, where $det[\cdot]$ stands for determinant, then the cost function for TDOA is defined as the reciprocal of SCM's determinant measuring the correlation among these aligned signals:

$$\mathbf{J}(k,\tau,f) = \frac{1}{det[\mathbf{R}(k,\tau,f)]} \tag{8}$$

As $\mathbf{R}(k,\tau,f)$ is a semi-positive definite matrix, the cost function is always non-negative. For a single subband, the TDOA is estimated as τ making the cost function reach the maximum.

In practice, instead of only by the current subband signal data as in (7), we modify $\mathbf{R}(k,\tau,f)$ by smoothing the results over the past frame, in order to overcome the fluctuation of SCM estimation:

$$\hat{\mathbf{R}}(k,\tau,f) = \alpha\hat{\mathbf{R}}(k-L,\tau,f) + (1-\alpha)\mathbf{R}(k,\tau,f) \tag{9}$$

where $\hat{\mathbf{R}}(k,\tau,f)$ and $\hat{\mathbf{R}}(k-L,\tau,f)$ stand for modified SCM of current frame and last frame respectively, L is the frame length, and α is updating factor with range [0,1], which is set to 0.8 in this paper. By exploiting the smoothing, more reliable SCM estimation is achieved. As a result, the formula provides a much more stable TDOA estimation performance.

3.3. Subband weight calculation

We compute the cost function for TDOA on each subband. Each hypothesized TDOA corresponds to an azimuth within the range of $[-90^\circ, 90^\circ]$. A 64×181 dimension matrix which has one subband cost function stored in each row can be obtained. In order to emphasize the results of subbands containing more speech signals, the fullband TDOA cost function is calculated as the weighted sum of subband TDOA cost functions.

The subband weight is defined as

$$\omega(k,f) = (\frac{1}{N}\sum_{i=1}^{N}\{E[y_i(k,f)y_i(k,f)^T]\})^p \qquad (10)$$

where N is the number of channels of microphone array, and p is a control factor. When $p = 0$, $\omega(k,f)$ equals to 1, all subbands have the same effect. The higher the value of p, the more effect "speech" subbands will have on the final estimation. However, we can not ensure that subbands with high weight are all "speech" subbands. As a tradeoff between the effect of "speech" subbands and other subbands, p is set to be 2. Indeed, the subband weight indicates the energy of each subband. we assume that the noise signal has relatively flatter energy distribution on each subband than speech signal, therefore, higher energy implies higher possibility of speech.

Before the weighted sum operation, the cost function of each subband is normalized. The normalization procedure is expressed as

$$\hat{\mathbf{J}}(k,\tau,f) = \frac{\mathbf{J}(k,\tau,f)}{\min_\tau \mathbf{J}(k,\tau,f)} \qquad (11)$$

where $\min_\tau \mathbf{J}(\tau,f)$ denotes the minimum value of fth subband cost function. The cost function value of all subbands are normalized to the same scale after the procedure. Usually, subbands without significant peaks in cost functions are noise subbands, after normalization, the cost function values get close to 1, and these subbands will have little effect on the final estimation.

The fullband cost function $\bar{\mathbf{J}}(k,\tau)$ is defined as the weighted sum of subband cost function:

$$\bar{\mathbf{J}}(k,\tau) = \sum_{f=1}^{M}\omega(k,f)\hat{\mathbf{J}}(k,\tau,f) \qquad (12)$$

where M denotes the number of subbands.

Fig.1 shows the figure of a weighted cost function matrix. The bottom plot shows the fullband cost function. A sharp peak is observed in the fullband cost function. The final TDOA at time k is estimated as

$$\tilde{\tau}_k = arg\max_\tau \bar{\mathbf{J}}(k,\tau), \qquad (13)$$

Substituting τ in eq.(4) by the estimated TDOA $\tilde{\tau}_k$, we can obtain the estimated direction of arrival.

4. Experiment

To evaluate the performance of the proposed subband weighting based algorithm, the experiments are performed on synthetic data. The proposed algorithm is compared with the SRP-PHAT [9] and broadband MUSIC algorithm[8] in different noisy conditions.

4.1. Experimental setup

A rectangular room with plane reflective boundaries is modeled in the experiment. The size of the room is $6 \times 4 \times 3$ meters. Four omnidirectional microphones are placed linearly and near the center of the small room, the location of these microphones are (3.00,2,2), (3.08,2,2), (3.16,2,2), (3.24,2,2) with space of 0.08m between two adjacent microphones. The speech source is located on a horizontal plane (x,y,2) with distance 2m to the center of the microphone array, ranging from -90° to 90°.

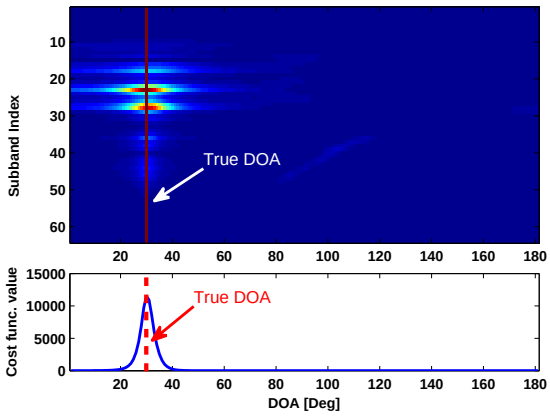

Figure 1: *Figure of a weighted cost function matrix, the true DOA is* 30°. *The bottom plot shows the fullband cost function.*

Room impulse responses from the location of speech source to microphones are modeled by image-source method [13], and a Matlab code implementation [14] is used to generate the microphone array data. The speech source is with durations of 10 seconds, microphone data is sampled with 16 bit resolution and sampling rate of 8KHz. We use two types of noise: white Gaussian noise and pink noise, to evaluate algorithms in white and colored noisy conditions. The noise added to each microphone is independent with each other, and scaled to control the SNR.

For all evaluated algorithms, the analysis frame size is set to be 256 samples with no frame shift. Algorithms are evaluated under different SNR conditions. For each SNR, 100 Monte Carlo simulations are conducted for DOA from -90° to 90° with 5° step size, resulting in 3600 simulations in total. The SNR changes from -10dB to 20dB, with a step size of 5dB.

4.2. Experimental results

Two frame level metrics, denoted as Accuracy and Root Mean Square Error (RMSE), are used to evaluate the performance of different algorithms. The estimation of one frame is considered to be correct if the difference between estimated DOA and real DOA does not exceed a certain threshold, which is commonly set to be 5°, then the Accuracy and RMSE are defined as below:

$$Accuracy = \frac{N_c}{N} \qquad (14)$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\theta - \hat{\theta})^2} \qquad (15)$$

Where N_c is the number of frames which have the correct estimation, N is the number of total frames, θ and $\hat{\theta}$ denote the estimated DOA and real DOA respectively.

Fig.2 shows the comparison results of broadband MUSIC, SRP-PHAT and the proposed subband weighting method. The performance under white Gaussian noise is illustrated in Fig.2(a) and Fig.2(c). It can be seen clearly that proposed algorithm yields great improvement in accuracy under low SNR conditions, and achieves performance similar to broadband MUSIC and better than SRP-PHAT when SNR is high. The proposed algorithm also gets the lowest RMSE under all SNR conditions considered. The performance under colored pink

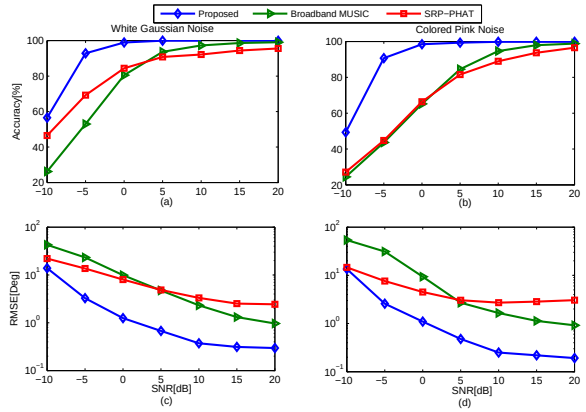

Figure 2: *Estimation performance of different algorithms. (a)(c) Accuracy and RMSE with white Gaussian noise. (b)(d) Accuracy and RMSE with colored pink noise. The error tolerance for Accuracy is* 5°

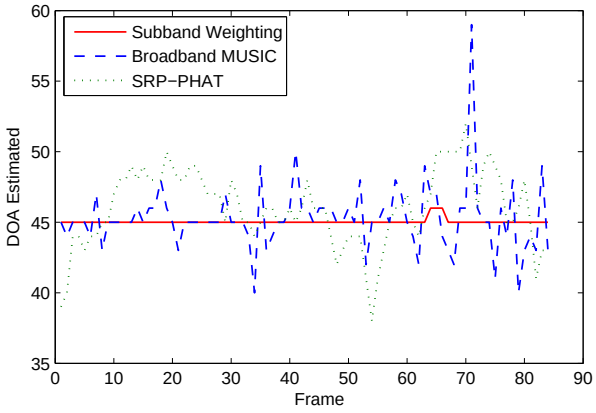

Figure 3: *DOA estimates for a continuous speech with 8dB SNR white Gaussian noise. The true DOA is* 45°.

noise is illustrated in Fig.2(b) and Fig.2(d). Similar comparison results are observed under colored noise condition. It should be noted that all the three algorithms suffer from performance degradation compared with white noise case in low SNR conditions, but the proposed algorithm degrades least. It is possibly due to the fact that we just assume the energy distribution of uncorrelated noise is flatter than speech, which is a relatively loose constraint in most conditions.

In Fig.3, the DOA estimates of three algorithms are shown for a continuous speech under white noise condition at 8dB SNR . The true DOA lies in 45°, 84 frames are contained in the speech. Obviously, the proposed algorithm yields the most stable and accurate estimation compared with the other two algorithms, which also confirm the efficiency of the proposed algorithm.

5. Conclusion

In this paper, we improve the performance of DOA estimation in noisy conditions based on subband weighting. Under the assumption that different subbands of the signals are not equally affected by noise, we estimate TDOA on each subband, and emphasize the estimation results of subbands which contains more speech signals. Experimental results on white and colored noise indicate that the proposed algorithm can achieve higher estimation accuracy and lower estimation error than the widely used SRP-PHAT and MUSIC algorithm.

6. Acknowledgements

This work was supported in part by the China National Nature Science Foundation (No.91120303, No.90820011, and No. 90820303), 863 China National High Technology Development Project (No.20060101Z4073, No.2006AA01Z194), and the National Grand Fundamental Research 973 Program of China (No. 2004CB318105).

7. References

[1] R. Schmidt, "Multiple emitter location and signal parameter estimation," IEEE Trans. Antennas and Propagation AP-34, vol. 3, pp. 276-280, 1986.

[2] M. McCloud and L. Scharf, "A new subspace identification algorithm for high resolution DOA estimation," IEEE Trans. Antennas Propag., vol. 50, pp.1382-1390, 2002.

[3] J.-M. Valin, F. Michaud, B. Hadjou, and J. Rouat, "Localization of simultaneous moving sound sources for mobile robot using a frequency-domain steered beamformer approach," in Proceedings IEEE International Conference on Robotics and Automation, 2004.

[4] J. Dmochowski, J. Benesty, and S. Affes, "A generalized steered response power method for computationally viable source localization," IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, no. 8, pp. 2510-2526, 2007.

[5] T. G. Dvorkind and S. Gannot, "Time difference of arrival estimation of speech source in a noisy and reverberant environment," Signal Processing, vol. 85, no. 1, pp. 177-204, 2005

[6] M. Brandstein, J. E. Adcock, and H. Silverman, "A practical time-delay estimator for localizing speech sources with a microphone array," Computer Speech and Language, vol. 9, no. 2, pp. 153-169, 1995.

[7] R. Jeffers, K. L. Bell, and H. L. Van Trees, "Broadband passive range estimation using MUSIC," in Proceedings IEEE International Conference on Acoustics, Speech, Signal Processing, vol. 3, pp. 2921-2924, 2002.

[8] J. P. Dmochowski , J. Benesty and S. Affes, "Broadband MUSIC: opportunities and challenges for multiple source localization," in Proceedings IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pp. 18-21, 2007.

[9] M. Brandstein, D. Ward, "Microphone Arrays: Signal Processing Techniques and Applications," Springer, 2001.

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

[11] Jacob Benesty, M. Mohan Sondhi, and Yiteng Huang, "Springer Handbook of Speech Processing," Springer, 2008.

[12] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," Hear. Res., vol. 47, pp. 103-138, 1990.

[13] E. Lehmann and A. Johansson, "Prediction of energy decay in room impulse responses simulated with an image-source model," Journal of the Acoustical Society of Amesrica, vol. 124, no. 1, pp. 269-277, 2008.

[14] Online: "http://www.eric-lehmann.com/ism_code.html," accessed on 19 Mar 2012.