SPEAKER DIRECTION FINDING: A COMPARISON OF PRACTICAL APPROACHES IN REVERBERANT ENVIRONMENT

Afsaneh Asaei[1], *Shirin Ghanbari*[1], *Mohammad Javad Taghizadeh*[2] *and Hossein Sameti*[3]

[1]{asaeiaf, sghanb}@itrc.ac.ir
[1]Multimedia Group, Iran Telecommunication Research Center, Tehran, Iran
[2]Iran Communication Industries, Tehran, Iran
[3]Computer Engineering Faculty, Sharif University of Technology, Tehran, Iran

ABSTRACT

Speaker direction finding techniques have aroused interests due to achieving the capability of receiving high-quality distant signals. Interesting concepts can be achieved through the comparison of such techniques whereby importance is in achieving high quality signals at reasonable complexity rates. With this aim in mind, this paper presents a critical comparison between two such traditional techniques; Time-Difference of Arrival (TDOA) estimation by Generalized Cross Correlation (GCC) and space scanning by Steered Response Power (SRP) of a beamformer. Each is analyzed under diverse conditions of noise and reverberation. Simulation results and experiments based on real data have been able to show that SRP with short data segments and due to its characteristic of averaging over the spatial dimension illustrate better accuracy results than that of GCC. These results have instigated a new method in the estimation of the source direction from a set of TDOAs based on spatial curvature collision. This paper discusses how this procedure reduces the computational cost more than 50 times compared to the conventional method of Root Mean Square (RMS) error minimization over the candidate locations.

1. INTRODUCTION

Applications such as automatic camera steering [1], video conferencing [2], hearing aids [2], hands-free speech recognition [3] and speaker identification are just a few applications that can benefit from speaker direction finding algorithms. The primary goal of such systems is in utilizing techniques that ensure accuracy. These techniques can be loosely classified into three general categories: (i) those adopting high resolution spectral concepts, (ii) techniques based upon maximizing the steered response power of a beamformer and (iii) approaches employing time difference of arrival information.

The first case characterizes any localization scheme dependent upon applications of the spatio-spectral correlation matrix. Interestingly, they are all designed for narrowband signals implying complexities within speaker localization [4]. The second strategy is based on maximizing the output power of a steered beamformer. In this case a beamformer can be used to scan over a predefined spatial region by adjusting its steering delays [5]. A filtering process can also be employed to increase accuracy whereby filters are designed in such a way to boost the power of the desired signal even if they cost distortion. That is the main distinction between the popular beamforming techniques in speech acquisition systems and that of localization.

The last category can be approached into two phases. Firstly it detects a set of Time-Difference of Arrival (TDOA) of the wave-front between different microphone pairs mostly based on the Generalized Cross Correlation (GCC) function. Then geometrical constraints are used to infer the source position [6]. Due to reduction of computational cost, this technique has aroused many interests. However, pair wise techniques suffer considerably from acoustic multipath propagation. As a solution, various weighting functions have been suggested; ML[1], PHAT[2], P[3] being the most prominent ones.

Organization of the paper is as follows: Section 2 presents the speaker direction finding strategies, including our new strategy. Section 3 discusses our simulation test scenario under which the various discussed techniques have been simulated and experimental results are presented. Finally, conclusions are given in section 4.

2. DIRECTION FINDING STRATEGIES

Source position in the spherical coordinate system is represented by azimuth, θ, elevation, φ and range, ρ. Whenever the source distance is much larger than its array aperture size, range becomes ambiguous and the

[1] Maximum Likelihood
[2] Phase Transform
[3] Pitch Harmonic

signal is received in a planar form. In this situation, a direction vector based on angles φ and θ as shown in "(1)," can be specified. Here the aim is to estimate angles φ, θ.

$$\vec{\zeta}_o^{(s)} = \begin{bmatrix} \sin\phi\cos\theta \\ \sin\phi\sin\theta \\ \cos\phi \end{bmatrix} \tag{1}$$

The signal received at the two microphones $x_1(t)$ and $x_2(t)$ can be modeled as:

$$\begin{aligned} x_1(t) &= s(t) * h_1(t) + v_1(t) \\ x_2(t) &= s(t) * h_2(t-\tau) + v_2(t) \end{aligned} \tag{2}$$

Where τ is the relative signal delay of interest, $h_1(t)$ and $h_2(t)$ are the impulse responses of the reverberant channels, $s(t)$ is the speech signal, $v_1(t)$ and $v_2(t)$ are uncorrelated noises.

The GCC function for a given time-lag [6], $R_{x_1x_2}(\tau)$ is calculated as the inverse Fourier transform of the received signal cross-spectrum, $X_1(w)X_1'(w)$ scaled by a weighting function $\psi_{12}(w)$:

$$R_{12}(\tau) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \psi_{12}(w) X_1(w) X_2'(w) e^{jw\tau} dw \tag{3}$$

The delay estimate τ, corresponds to the TDOA; maximizes the GCC function. In general, "(3)," has multiple local maxima and can cause several erroneous estimates due to multipath effects and the background noise. The type of filtering or weighting, used within GCC is crucial to performance. The following ML-weighting function [7] is derived from the magnitude spectra of the microphone signals and noise signals and is equivalent to the signal SNR evaluated from a single frame of the observed data:

$$\hat{\psi}_{12}^{ML}(w) = \frac{|X_1(w)||X_2(w)|}{|V_1(w)|^2|X_2(w)|^2 + |V_2(w)|^2|X_1(w)|^2} \tag{4}$$

The noise power spectra, $|V_1(w)|^2$ and $|V_2(w)|^2$, are estimated during the silence intervals.

The Phase Transform (PHAT) employs the weighting function of "(5)," [8], removing the influence of spectral magnitudes and producing a GCC function dependent entirely on the phase of the cross spectrum.

$$\psi_{12}^{PHAT}(w) \equiv \frac{1}{|X_1(w)X_2'(w)|} \tag{5}$$

In another approach, GCC-P, the degree of periodicity in each frequency band is measured [9], and the weighting function based on harmonic modeling of its speech spectrum employed:

$$W_l(w) = \frac{(1-\max\{E_{l1}, E_{l2}\})^2}{|X_1(w)X'_2(w)|}, \quad w \in [a_l, b_l] \tag{6}$$

Where the interval $[a_l, b_l]$ is the frequency region centered on the l^{th} harmonic of the fundamental frequency and E_l is the normalized error associated with that harmonic.

In another approach, multiple spatial data can be exploited for direction finding algorithm. The direct procedure uses the idea of combining the data from multiple microphones earlier in the estimation process and is called steered beamformer. The general algorithm of the beamfoming is filter-and-sum method, in which the spatial-data are filtered using the frequency-filter [4]. Then all the signals are summed and hence with the choice of the most prominent filter the source signal is enhanced, as well as eradicating the un-correlated background noise. The simplest filter is a time-shift, known as the delay-and-sum beamforming. The SRP output is maximized under different factors. However, under desired conditions, if time-shift is equivalent to TDOA the maximum power is gained. The output of the filter-and-sum beamformer within the frequency domain is given in "(7),".

$$Y(w, \Delta_1 \ldots \Delta_M) \equiv \sum_{m=1}^{M} G_m(w) X_m(w) e^{-jw\Delta_m} \tag{7}$$

$\Delta 1, \ldots \Delta m$ represent steering delays and are calculated for a candidate direction in space. $X_m(w)$ is the received signal at microphone m, being filtered by $G_m(w)$. The output of the beamformer is used in "(8)," to form the power of the steered response.

$$P(\Delta_1 \ldots \Delta_M) \equiv \int_{-\infty}^{\infty} Y(w, \Delta_1 \ldots \Delta_M) Y'(w, \Delta_1 \ldots \Delta_M) dw \tag{8}$$

A specific filter for the purpose of localization is suggested in the SRP-PHAT algorithm. "Eq. (9)," shows the selection of filter for each channel.

$$G_m(w) = \frac{1}{|X_m(w)|} \tag{9}$$

2.1. Proposed Method

After time delay estimation, the second phase is to find the corresponding source direction. A new method is proposed here. Its idea is based on the fact that possible source direction due to far-field assumptions, as shown in "(10)," forms a cone, centered on the intersection line of the microphone pairs. It is assumed that the microphone panel is on the ZoY surface and the centre of the microphones is on the origin of the coordinate system.

$$x^2 + (y\cos\alpha + z\sin\alpha)^2 = tg^2\beta(y\sin\alpha - z\cos\alpha)^2$$
$$\beta = \cos^{-1}(\frac{c\tau}{d}) \quad (10)$$

Where α is the angle between z axes and the microphone pair intersection line and β, which is calculated from TDOA; representing the cone head angle. The cones collide two by two in a line with specific φ, θ and the source direction is evaluated from their averaged values.

In order to simplify and solve the collision equations, appropriate rotation of the coordinate system is employed ensuring that the cone is centered on the z axis. Then φ and θ are calculated in spherical system and inverse rotation is performed. This new procedure seems complex in concept but the computation is limited to a few matrix multiplications and is much less than (more than 50 times) the computation of direction finding by minimization of the root mean square error [4] between the estimated delays and those of candidate directions.

For example if our simulation test room is searched in 0.2m×0.2m blocks. The new method requires only 800 multiplications compared to 48000 required by our traditional RMS method. Another point is that the computational complexity of the new method is independent of room dimensions while traditional RMS method's computational cost grows considerably by increasing room dimensions.

3. EXPERIMENTAL RESULTS

The performance of the direction-finding algorithm was evaluated and analyzed through a series of experiments. The test scenario was within a rectangular room (4m×4m×6m) with plane reflective surfaces and uniform, frequency independent reflection coefficients. Room impulse responses were generated utilizing the Image method [10] for the reverberation times T60 = {0s, 0.17s, 0.37s, 0.46s}. The speed of sound is assumed to be 346m/s. Four microphones were positioned on its wall within 4 corners of a specified rectangle. The origin of the coordinate system was assumed to be the centre of the microphone array and is 1.5 meters high. Details of this test scenario are shown in Table I.

For each combination of parameters, 2s of 22050Hz sampled speech signal and zero mean white Gaussian noise were convolved with the proper channel impulse responses and added together. With the intension of simulating the direction-finding algorithm, the array signal is processed in 25ms blocks with 50% overlapping. After employing a hanning window, DFT of the blocks are formed.

Three defined weighting schemes are used for TDOA estimation by GCC function. With the use of estimates in six different microphone pairs the direction of the source is evaluated from the collision of the cones. First, the number of the estimated τ is reduced to four by averaging between the parallel pairs forming a cone from the origin of the coordinate system. Second, non-linear equations of cones collision are formed.

Each TDOA estimate is the integer sample delay corresponding to the maximum of the GCC. The input signal is up sampled to 96000Hz to achieve 0.23 sub-sample accuracy. A Voice Activity Detector (VAD) detects the background noise and the blocks of pseudo noise speech signal spectrum with the power close to the background noise are removed from the direction finding procedure. In SRP method a steered beamformer searches the space with one degree accuracy based on ϕ, θ.

Four different direction finding methods were evaluated: the GCC-ML, the GCC-PHAT, the GCC-P, and the SRP-PHAT. To measure the accuracy and robustness, anomaly statistics were calculated over the ensemble of speech segments with each of the source positions, reverberation times, and SNR conditions. Referring to "Fig. 1," these plots represent the percentage of estimates outside a 10° absolute error threshold as a function of SNR and reverberation time. The general results observed here are consistent with those obtained with the other source position.

TABLE I
SIMULATION PARAMETERS

Test Scenario for Direction Finding
Rectangular Microphone Array: Omni directional
Length = 0.3m , Width = 0.25m
Speaker: ρ = 3m , θ = {40° , 60°} , φ = 70°
Noise: ρ = 3.5m , θ = 80° , φ = 110°
SNR = {5dB, 15dB, 30dB}
T60={0s, 0.17s , 0.37s , 0.46s}

When no reverberation exists ("Fig. 1.a,") each of the direction estimators performs well at the high SNR levels. These experiments show that GCC-ML outperforms under noisy condition due to the use of noise spectrum in its filtering procedure. However, if reverberation exists the basic assumption of uncorrelated noise no longer exists and the performance of this algorithm reduces considerably. This result is clear from other graphs ("Fig. 1.b-d,") where the channel multipath is included. Therefore, SNR based weighting function does not perform well in practical condition where reverberation is a critical problem. It can also be seen that the anomaly of GCC-ML is relatively independent of SNR and increases with reverberation. In algorithm GCC-PHAT, by whit

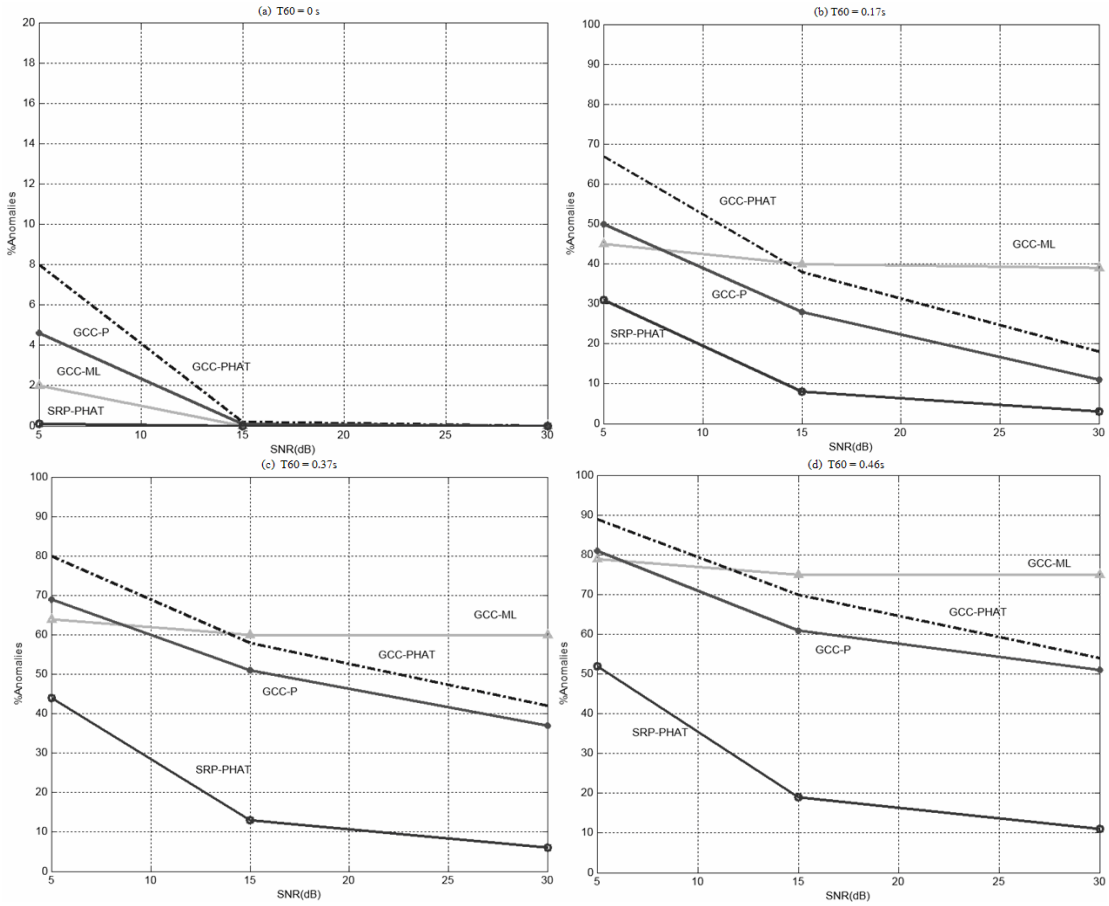

Figure 1. *Direction Finding Accuracy Comparison for Different Techniques*

ening the speech signal spectrum a higher accuracy is gained under reverberant conditions. The GCC-P algorithm by investigating part of a periodic structure shows more robustness compared to GCC-PHAT, in particular in the presence of high noise. On the other hand, SRP-PHAT with multiple spatial data achieve high robustness (up to 40% reduction of estimation anomalies) to both noise and reverberation.

The above experiments show that by increasing spatial data used in the estimation process, performance rates will increase. In order to investigate the effect of more temporal data on the procedure, accumulation time of cross-spectra is increased by averaging GCC for successive blocks. The number of blocks equaling to 1, 5, 10, and 20 are chosen. Results of this experiment can be seen in Table II; presenting anomalies (percentage) regarding the GCC-PHAT algorithm.

It can be seen that in low reverberant conditions by utilizing 20 frames and taking the time-delay as 10 times the frame delay, the estimation anomaly reduces up to 34%. This improvement is up to 41% in more reverberant situations. Based on these results, it is obvious that if long data segments are available, GCC-PHAT is capable of speaker direction finding at a high accuracy level. Similar effects can also be seen within the other GCC-based direction estimation techniques.

TABLE II

BLOCK AVERAGING EFFECT ON GCC-PHAT

T60	SNR	N = 1	N = 5	N = 10	N = 20
0.17	15	38%	23%	11%	4%
	30	18%	8%	6%	2%
0.37	15	58%	36%	24%	17%
	30	42%	30%	21%	11%

4. CONCLUSION

In this paper two direction finding approaches based on the steered response power of beamformer and time delay estimation by generalized cross correlation function have been simulated and analyzed, under different conditions. A new method to find the source direction from a set of estimated TDOAs is proposed. This method is based on collision of possible source directions corresponding to each TDOA. Therefore by eliminating the searching procedure reduces the computational cost of direction finding more than 50 times and is independent of room dimensions. However under extreme acoustic conditions, TDOA estimation by GCC takes a considerable time to output a reasonable performance rate. Hence applications that require high update rates to follow the dynamic conditions can no longer benefit from such an approach. The underlying pair wise process incorporates the data from only two microphones. Rather than increasing the accumulation time, an increase in data may be achieved by incorporating the data from several microphones. That is averaging over the spatial dimension instead of the temporal one. Hence, SRP with short data segments illustrates up to 40% reduction of estimation anomalies compared to GCC.

Under favourable conditions, direction finding with the use of a moderate number of microphones can achieve reasonable results. On the other hand integrating data from a multitude of microphones can be exploited to source localization under the notorious conditions of reverberation and background noise.

5. REFERENCES

[1] H. Wang and P. Chu. "Voice source localization for automatic camera pointing system in videoconferencing", *ICASSP*, volume 1, pages 187-190. IEEE, 1997.

[2] H. Krim and M. Viberg, "Two Decades of Array Signal Processing", *IEEE Signal Processing Magazine*, July 1996.

[3] T. B. Hughes, H. Kim, J. H. DiBiase, and H. F. Silverman, "Per formance of an HMM speech recognizer using a real-time tracking microphone array as input", *IEEE Trans. on Speech Audio proc. 7(3)*: 346-349, May 1999.

[4] M. S. Brandstein, Microphone arrays, signal processing techniques and applications, Springer Verlag, 2001.

[5] J. C. Chen, K. Yao, and R. E. Hudson, "Source localization and beamforming", *IEEE Signal Processing Magazine*, March 2002.

[6] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay", *IEEE Trans. Acoust., Speech Signal Process., vol. ASSP-24*, pp. 320-327, Aug. 1976.

[7] B. Champagne, S. Bedard and A. Stephenne, "Performance of time-estimation in the presence of room reverberation", *IEEE Trans. Speech Audio Proc.*, 4(2): 148-152, Mar. 1996.

[8] P. Svaizer, M. Mattassoni and M. Omologo, "Acoustic source localization in a three-dimensional space using crosspower spectrum phase", *ICASSP97*, 1997.

[9] M. Brandstein, "Time delay estimation of reverberated speech exploiting harmonic structure", *Journal of Acoustic Society of America*, vol. 105, no. 5, pp. 2914-2919, 1999.

[10] J. B. Allen and D. A. Berkley, Image Method for Efficiently Simulating Small Room Acoustics, *Journal of Acoustic Society of America*, vol. 6, no. 4, pp. 943-950, April 1979.