Robust Speaker Localization Utilizing a Novel Beamforming Algorithm Based on Harmonic Structures

Afsaneh Asaei[1]*, Hossein Sameti*[2]*, Mohammad Shahram Moin*[3]
[1]Multimedia Research Group, Iran Telecommunication Research Center, Tehran, Iran
[2]Computer Engineering Faculty, Sharif University of Technology, Tehran, Iran
[3]IT Faculty, Iran Telecommunication Research Center, Tehran, Iran
E-mails: asaeiaf@itrc.ac.ir, sameti@sharif.edu, moin@itrc.ac.ir

ABSTRACT

Speaker localization by microphone array has recently received significant attention. Although various methods have been proposed; their performance with short data segments under noise and reverberation degrades considerably. Sound localization based on Steered Response Power (SRP) shows more robustness in practical situations especially with the use of short data segments. In SRP-PHAT algorithm by employing a filtering scheme based on phase transform, performance of this localization algorithm is improved in reverberant environments. There has been relatively little success at applying an SRP filtering scheme that is robust in both noisy and reverberant conditions.

This paper presents a novel and useful filter for SRP localization. The proposed approach employs estimated periodicity of harmonic spectral intervals of speech signal to design a filter for each channel of microphone array and is used within the filter and sum beamforming algorithm. Simulation results in various acoustical conditions (different SNR values and reverberation parameters) shows the superiority and robustness of the new algorithm compared to the conventional methods.

Keywords— *Speaker Localization, Microphone Array, Reverberation, Steered Response Power, Beamforming*

1. INTRODUCTION

Applications of the speaker localization are diverse ranging from automatic camera steering [1], hearing aids [2], and hands-free speech recognition [3] to speaker identification systems [4]. The primary goal of such systems is in utilizing techniques that ensure accuracy. These techniques can be classified into two general categories: direct localization and indirect localization.

The first strategy is based on maximizing the output power of a steered beamformer or Steered Response Power (SRP). In this case a beamformer is 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. Through extensive experiments we found that this category has the most robustness in source localization in practical situations and is preferable for reliable localization of speech signals [6].

The second category is approached into two phases. First 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 [7]. Then geometrical constraints are used to infer the source position. Due to the reduction of computational cost, this technique has aroused many interests. However, pair wise techniques suffer considerably from multipath propagation.

This paper considers speaker localization by beamforming method. A new filter is proposed for SRP algorithm that improves localization robustness to both noise and reverberation. Organization of the paper is as follows: Section 2 presents the speaker localization strategy based on the beamforming algorithm, including our new strategy. Section 3 discusses our simulation test scenario under which the different discussed techniques have been simulated and experimental results are presented. Finally, conclusions are given in section 4.

2. SOUND LOCALIZATION BY BEAMFORMING ALGORITHM

This method is based on steering the beam pattern of a microphone array towards candidate positions in space; also known as beamforming. Then the output power of a beamformer is computed and the source position is determined based on the beamformer's maximum power using the Maximum Likelihood (ML) estimation. The output of a filter and sum beamformer in frequency domain is defined in equation (1).

$$Y(\omega,\Delta_1 ... \Delta_M) \equiv \sum_{m=1}^{M} G_m(\omega)\, X_m(\omega)\, e^{-j\omega\Delta_m} \qquad (1)$$

Where $\Delta_1,\ldots,\Delta_M$ are steering delays which are computed for candidate positions in space. $X_m(\omega)$ is the received signal at microphone *m* which is filtered by $G_m(\omega)$. Steered Response Power (SRP) is then computed by equation (2).

$$P(\Delta_1..\Delta_M) \equiv \int_{-\infty}^{\infty} Y(\omega,\Delta_1..\Delta_M)Y'(\omega,\Delta_1..\Delta_M)d\omega \tag{2}$$

Although steering delays are continuous variables, the above equation is computed for sampled space locations. Source location in spherical domain is represented by a range ρ, an azimuth θ and an elevation $\varnothing$. If the source range compared with the array dimension is bigger than a specific threshold [8], its wavefront is received in a planar form and hence determining an accurate range becomes ambiguous. Therefore, due to this ambiguity, the source position will be specified by θ and $\varnothing$ with a vertical representation of

$$\vec{\zeta}_o^{(s)} = \begin{bmatrix} \cos\phi\sin\theta \\ \cos\phi\cos\theta \\ \sin\theta \end{bmatrix} \tag{3}$$

In this case, the steering delay of microphone *m* relative to a reference microphone is calculated by equation (4).

$$\Delta_m = \frac{d_m \cos\alpha}{c} \tag{4}$$

Where *d* is the distance between two microphones, *c* is the speed of sound and *α* is the wavefront angle relative to the microphones intersection line.

In calculating SRP, the choice of a suitable filter has a considerable effect on the robustness of localization to both noise and reverberation. In the familiar algorithm SRP-PHAT, the employed filter at each channel is:

$$G_m(\omega) \equiv \frac{1}{|X_m(\omega)|} \qquad for\ m = 1....M \tag{5}$$

2.1. Proposed Method

Room impulse response is illustrated in figure (1). The largest peak corresponds to the direct path and secondary peaks are due to the walls reverberation. Assuming that the total system of microphone array and room is a liner system [9]; the received signal at each microphone is the convolution of this impulse response and the source signal.

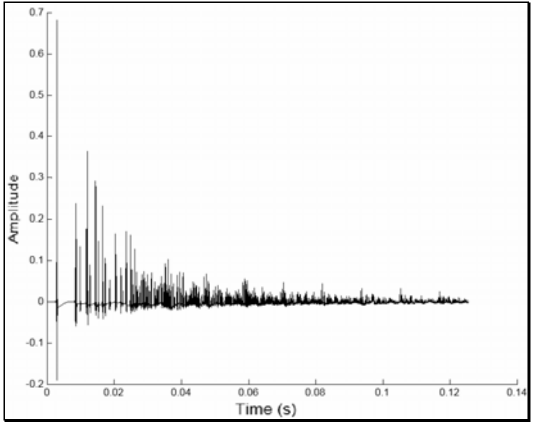


Figure 1. Room Impulse Response

Therefore, multipath effect and noise degrades the received signal at the microphone array and reduces the periodicity of voiced segments. This effect can be seen in Figure (2). Where T60 identifies room reverberation time. This parameter shows the time needed for the impulse response normalized power to become less that -60dB. This figure illustrates distortion of periodic structures at acoustical conditions of SNR = 5dB and T60 = 0.47s. Therefore those frames of speech signal with periodic structures are less influenced by reverberation and noise and must be weighted more in the localization process.

In order to use this idea we have to determine the degree of periodicity for each of the speech frames. Therefore the voicing decision as is used in Multi-Band Excitation (MBE) coder is utilized [10]. In MBE-based coders, a normalized error E_l between the original and modeled speech spectra is calculated in frequency bands as shown in equation (6).

$$E_l = \frac{\sum_{\omega=a_l}^{b_l} \left| X(\omega) - \hat{X}(\omega,\omega_0) \right|^2}{\sum_{\omega=a_l}^{b_l} \left| X(\omega) \right|^2} \tag{6}$$

where $X(\omega)$ is the original speech spectrum, ω_0 is the fundamental frequency, a_l and b_l are the first and last harmonic in l^{th} band, and $\hat{X}(\omega,\omega_0)$ is the estimated speech spectrum which is calculated through equation (7).

$$\hat{X}(\omega,\omega_0) = A_k(\omega_0)W(\omega) \quad 1 \le k \le K, \quad \lceil a_k \rceil \le \omega \le \lceil b_k \rceil \tag{7}$$

Where $a_k = (k-0.5)\omega_0$, $b_k = (k+0.5)\omega_0$, $\lceil . \rceil$ means the nearest integer greater than or equal to, *K* is the number of harmonics in the 4kHZ speech frequency bandwidth, $W(\omega)$ is the frequency response of the Hanning window centered at the k^{th} harmonic of the fundamental frequency and $A_k(\omega_0)$ is the k^{th} harmonic amplitude which is computed using:

$$A_k(\omega_0) = \frac{\sum_{\omega=\lceil a_k \rceil}^{\lceil b_k \rceil} X(\omega)W(\omega)}{\sum_{\omega=\lceil a_k \rceil}^{\lceil b_k \rceil} |W(\omega)|^2} \tag{8}$$

In order to model the spectrum at high frequency bands accurately, the fundamental frequency must be calculated with an error less than 1HZ. A practical way is to estimate the pitch in time domain with a fast algorithm and then to minimize the error criterion of equation (9) in the frequency domain to improve the accuracy.

$$\varepsilon_l = \sum_{\omega=a_l}^{b_l} |X(\omega) - A_l W(\omega)|^2 \tag{9}$$

The error criterion of equation (9) will then be computed by equation (10) for each of the estimated fundamental frequencies.

$$\varepsilon = \sum_L \varepsilon_l \tag{10}$$

Where *L* is the number of frequency bands. This error criterion is computed for a range of fundamental frequencies. The minimum value of this error corresponds to the most accurate estimation of the fundamental frequency.

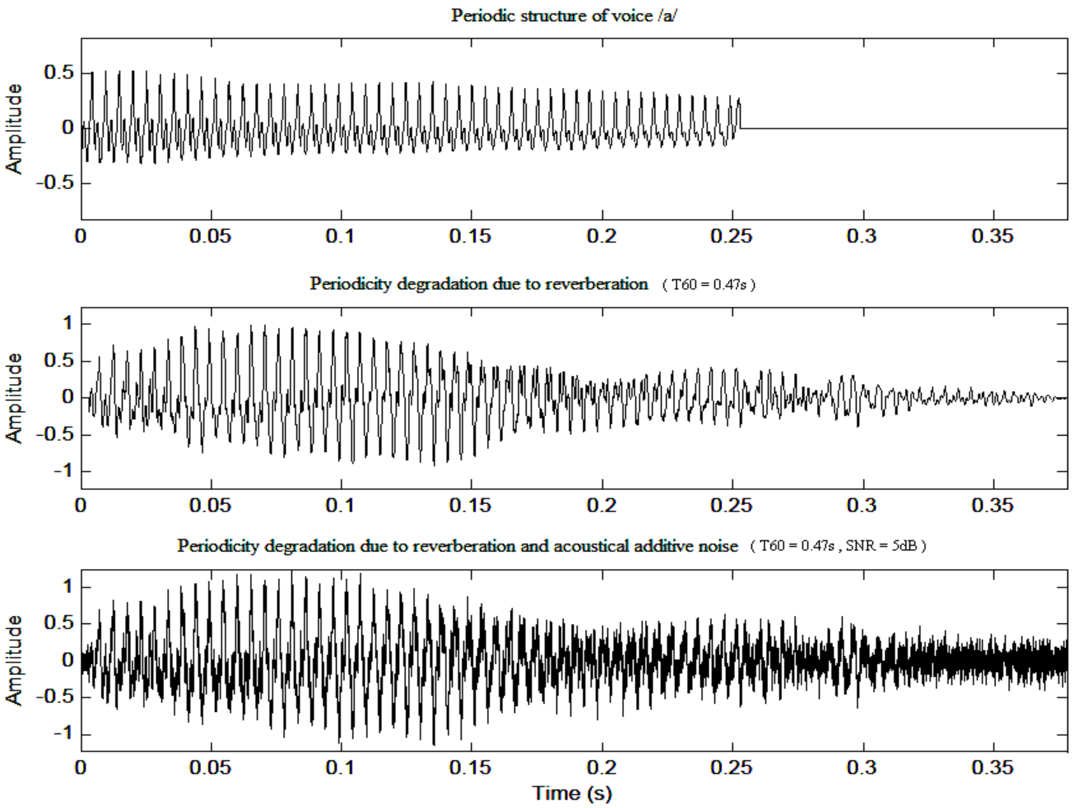


Figure 2. Periodicity degradation due to noise and reverberation. Upper graph: Initial signal of word /a/, Middle graph: Signal affected by reverberation, Lower graph: Signal affected by reverberation and noise

After the fundamental frequency is estimated, the similarity between the original spectrum and the modeled spectrum is measured at each harmonic band by equation (6). For voiced frames, E_l has a value close to zero, however the values close to 1 correspond to the noisy non-periodic intervals. Therefore, the calculated error from equation (6) is used to measure the degree of periodicity for each frequency band and can be employed as a filtering scheme for SRP localization. The proposed filter for each channel is:

$$G_{l,m}(\omega) = \frac{1 - E_{l,m}}{|X_m(\omega)|} \ , \ \omega \in [a_l, b_l] \qquad (11)$$

By employing the proposed filtering scheme, the voiced frames will be emphasized. Furthermore, the influence of the signal amplitude will be omitted and only phase information is used in the localization algorithm. This improves the robustness of the proposed algorithm to both noise and reverberation. In reality, for small arrays it is sufficient to compute the fundamental frequency harmonics at the reference microphone and then the error for each frequency band is calculated based on this frequency. Therefore, if a channel signal is for some reason degraded, its influence will be reduced. The employed filter in the beamforming algorithm and the output of the steered array is computed by equation (12) and then its power is calculated for that particular point in space. The proposed method is named SRP-H as it is based on beamforming and analyses of speech signal regarding the fundamental frequency harmonics.

$$\tilde{Y}^{SRP-H}(\omega, \Delta_1..\Delta_M) \equiv \sum_{m=1}^{M}\sum_{l=1}^{L} \frac{1-E_{l,m}}{|X_m(\omega)|} X_m(\omega) e^{-j\omega\Delta_m} \qquad (12)$$

3. EXPERIMENTAL ANALYSIS

The performance of the proposed method is compared with the other localization algorithms by a number of experiments in different acoustical conditions. Simulation test room is a 4m×4m×6m rectangular room, whereby a planar rectangular array of four microphones is placed on a wall. The surface reflection coefficients are assumed to be uniform and frequency independent. Room impulse response is calculated by the Image method for different reverberation parameters [11]. Details of this test scenario are listed in table 1.

Table 1. Simulation Parameters

Test Scenario for Direction Finding
Rectangular Array: 4 Omni directional Microphones
Length = 0.3m , Width = 0.25m
Speaker: ρ = 3m , θ = {40° , 60°} , Height = 1.5m
Noise: ρ = 3.5m , θ = 100° , Height = 1.5m
SNR = {5dB, 15dB, 30dB}, T60={0s, 0.17s , 0.47s}

170 frames of the speech signal sampled by the frequency at 22050HZ are convolved with a channel impulse response for each microphone. White Gaussian noise convolved with corresponding channel response is added to the signal. Then the signal is up-sampled to 96000 in order to scan the space with 1 degree accuracy in θ and φ angles. Low power frames are also removed from localization process. A voice activity detector estimates the background noise power and the frames with power near to background noise are omitted. Figure (3) presents the percentage of anomalies (which by definition is an absolute error value larger than 10 degrees) in direction finding for a speaker positioned at 40 degrees. Results are the same for a speaker positioned at 60 degrees [12].

Figure (3.a) shows the percentage of estimation anomalies for different signal to noise ratios in low reverberant environments. It can be seen that SRP-PHAT shows low performance in low SNR conditions due to removal of amplitude effects. Comparing the performance of this algorithm with SRP shows that SRP-PHAT is suboptimal at low reverberation times. The proposed algorithm SRP-H due to the employment of a filtering scheme based on periodic structures as well as Phase transform reduces the percentage of anomalies by 15% at signal to noise ratio of 5dB and a reduction of 7% at SNR=15dB of estimation anomalies is achieved.

Figure (3.b) presents anomaly percentage of localization algorithm at high reverberation time. By comparing the graphs in (3.b) with their correspondence in (3.a) it can be seen that multi-path effect in high noisy conditions increases anomalies of estimation by 20%. It also shows that SRP-PHAT and SRP-H performs much better than SRP in noisy condition. SRP-H by removing the influence of destroyed frames shows 10% less anomalies compared to SRP-PHAT in high reverberant, high noisy conditions.

Performance is improved by our proposed method, but enforces computation overhead of fundamental frequency estimation which can be reduced considerably, by a fast preliminary estimation algorithm and fundamental frequency tracking. Therefore, it can be implemented at real-time and is capable of speaker tracking with good accuracy results.

4. CONCLUSION

This paper proposed a new filtering scheme for speaker localization by beamforming, SRP-H, based on exploiting the harmonic structures of the speech signal. Simulation results in different noisy and reverberant conditions show that SRP-H due to the detection and removal of destroyed speech frames reduces the percentage of localization anomalies up to %15 compared to the result of the most common algorithm SRP-PHAT in low SNR reverberant condition.

Because the proposed method is based on fundamental frequency estimation of speech signals, by utilizing this information, the algorithm will be capable to localize the verified speakers based on fundamental frequency distinctions. This capability can be used in the applications such as distant-talking speech recognition to increase the system robustness to speech-like noises.

Furthermore, because the proposed algorithm is based on short data frames, it can be utilized in tracking speakers even under various environmental changes.

5. REFERENCES

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

[2] J. G. Desloge, W. M. Rabinowitz, and P. M. Zurek, “Microphone-array hearing aids with binaural output. I. Fixed-processing systems,” *IEEE Transactions on Speech and Audio Processing, voume 5, no. 6, pp. 529–542*, 1997

[3] Hughes, T. B., Kim, H. J., DiBiase, H., Silverman, H. F., “Performance of an HMM Speech Recognizer Using a Real-Time Tracking Microphone Array as Input”, *IEEE Transaction on Speech Audio processing, pp. 346-349*, May 1999

[4] L. Qiguang, J. Ea-Ee, J. Flanagan, “Microphone Arrays and Speaker Identification”, *IEEE Transaction on Speech and Audio Processing, volume 2, pages 622-629*, Oct. 1994

[5] Ward, B., Kennedy, R. A., Williamson, R. C., “An Adaptive Algorithm for Broadband Frequency Invariant Beamforming”, *Proc. ICASSP97*, 1997

[6] A. Asaei, H. Sameti, "Speaker Direction Finding for Practical Systems: A Comparison of Different Approaches", *Proceeding of the third Annual IEEE BENELUX/DSP valley signal processing symposium, Metropolis, Antwerp, Belgium,* March 2007, *pp 129-133*

[7] Knapp, C. H., Carter, G. C., “The Generalized Correlation Method for Estimation of Time Delay”, *IEEE Transaction on Acoustic, Speech Signal Process., Vol. ASSP-24, pp. 320-327*, Aug. 1976

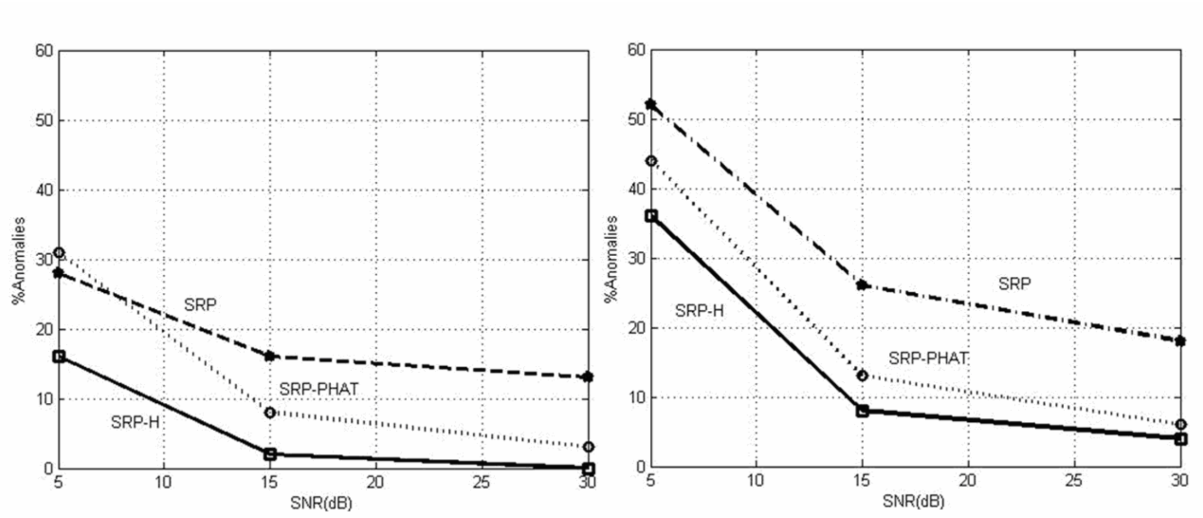
[8] McCowan, I. A., *Robust Speech Recognition Using Microphone Arrays*, PhD Thesis, *Queensland University of Technology, Australia*, 2001.

[9] Ziomek, L. J., *Fundamentals of Acoustic Field Theory and Space-Time Signal Processing*, CRC Press, 1995.

[10] D. Griffin and J. Lim, “Multiband Excitation Vocoder”, *IEEE Trans. Acoust., Speech, Signal Processing, vol. 36, no. 8, pages 1223-1235,* Aug. 1988.

[11] Allen, J. B., Berkley, D. A., “Image Method for Efficiently Simulating Small Room Acoustics”, *Journal of Acoustic Society of America, Vol. 6, pp. 943-950*, April 1979.

[12] A. Asaei, “Sound Source Localization by Beamforming Techniques for Robust Speech Recognition, Msc. Thesis, Computer Engineering Department, Sharif University of Technology, Novem 2006.



a. T60 = 0.17s

b. T60 = 0.47s

Figure 3. Anomaly percentage of location estimation by SRP, SRP-H and SRP-PHAT