Verified Speaker Localization Utilizing Voicing Level in Split-Bands

Afsaneh Asaei[1,2], Mohammad Javad Taghizadeh[3], Marjan Bahrololum[4] and Mohammed Ghanbari[5]

[1]IDIAP Research Institute, Martigny, Switzerland

[2]Swiss Federal Institute of Technology at Lausanne (EPFL), Switzerland

[3]Department of Information Technology, Iran Telecommunication Research Center, Tehran, Iran

[4]Department of Security, Iran Telecommunication Research Center, Tehran, Iran

[5]Department of Computing and Electronic Systems, University of Essex, Colchester, United Kingdom

Emails: afsaneh.asaei@idiap.ch, taghizadehmj@itrc.ac.ir, bahrololum@itrc.ac.ir, ghan@essex.ac.uk

ABSTRACT

This paper proposes a joint verification-localization structure based on split-band analysis of speech signal and the mixed voicing level. To address the problems in reverberant acoustic environments, a new fundamental frequency estimation algorithm is proposed based on high resolution spectral estimation. In the reconstruction of the distorted speech this information is utilized to reduce the side effect of acoustic noise on the voicing parts. A speaker verification system examines the features of the reconstructed speech in order to authorize the speaker before localization. This procedure prevents localization and beamforming for non-speech and specially the unwanted speakers in multi-speaker scenarios. The verification is implemented with the Gaussian Mixture Model and a new filtering scheme is proposed based on the voicing likelihood of each frequency band measured in the previous steps for efficient localization of the authorized speaker. The performance of the proposed VSL (Verified Speaker Localization) front-end is evaluated in various reverberant and noisy environments. The VSL is utilized in the development of distant-talking automatic speech recognition by microphone array where the system can lock on a specific source and hence the recognition quality improves noticeably.

***Keywords**: Microphone Array, Speaker verification, Speaker Localization, Reverberation, Beamforming, Speech Recognition*

1. INTRODUCTION

For a hands-free speech interface, it is very important to capture distant talking speech with high quality. An ideal solution for this purpose is sound acquisition by microphone array. A microphone array can acquire the desired speech signals selectively by steering the beam pattern directivity of the array towards the desired speaker. This process is called beamforming and due to the directivity pattern steering, it can spatially filter out noises from other directions regardless of the noise nature. The main obstacles to achieve reasonable performance in array based systems are the reverberation and the presence of ambient noise of acoustic environment. These parameters affect the accuracy of speaker localization and beamforming in capturing the desired spatial signal and suppressing the others. To tackle this problem, various methods have been proposed recently, but they all seem to give erroneous estimations in speaker direction finding under the presence of high noise and reverberation. These conventional algorithms in multi-speaker environments not only have difficulty in localizing the multiple sound sources accurately, but they also fail to localize the target talker among the known multiple speaker positions. These localization 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 class of these techniques, characterizes any localization scheme that is dependent upon applications of the spatio-spectral correlation matrix [1]. Interestingly, all of these methods are all designed for narrowband signals and are very sensitive to source and microphone modeling [2] implying complexities within the speaker localization process [3, 4]. The second class of the aforementioned strategies 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 may increase distortion. This is the main distinction between the popular beamforming techniques in speech acquisition systems and that of localization [6, 7]. This category has the most robustness in source localization in practical situations and is preferable in enabling reliable localization of speech signals with short frames [8]. The third category is realized in 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 maximization [9]. In computing the cross-correlation function, to increase accuracy, some weighting schemes are also employed. The most important weightings are ML (Maximum Likelihood) and PHAT (Phase Transform) [10, 11]. Second, geometrical constraints are used to infer the source position. Due to its low computational cost, this technique has attracted many interests. However, pair-wise techniques suffer considerably from multipath propagation [8]. Since the primary goal of microphone array based systems is practicality in the real environment, we have considered this subject for real applications. In the scenario which is the subject of this investigation, we have focused on steered response power (SRP) based localization.

All the above mentioned attempts were aimed to improve the localization accuracy in the presence of acoustic noise and reverberation and could not achieve satisfactory results in the presence of spurious speech sources such as the voice of unwanted speakers. In this scenario, speaker verification is needed to authorize the speech. This stage of speaker verification by microphone array is addressed in [12], where a microphone array is utilized to capture the speech and provide input for automatic speech identification. A 2-D matched filter microphone array is proposed to improve the identification scores in a reverberant environment. In this algorithm, the identification is addressed after the array-based analysis of the received signal. Investigations by Gianakopoulos et. al [13] are concentrated on the implementation of the front-end signal pre-processing tasks such as filtering, acquisition and beamforming to improve speaker recognition. This procedure suffers from over computation of localization and beamforming in the multi-speaker scenarios. In [14] an adaptive near-field beamformer is implemented for hands-free speaker recognition. In [15] speech enhancement techniques are utilized to reduce the acoustic degradation of source signal and improve speaker verification in the noisy environments. In [16] a speaker identification algorithm based on the angle of arrival of the speech is proposed. Since the convergence rate is large, the new algorithm has practical limitations and participants are required to remain seated during the experiment. Hence, limited number of investigators has studied speaker recognition and although the effectiveness of beamforming is proven in robust hands-free speaker recognition [17], but verification always comes after the localization, beamforming and other computational array processing algorithms.

In this paper, the idea of verification prior to localization is proposed. It has been observed through extensive testing that the quality of the voiced parts is very important for verification. Therefore, we have enhanced these parts and used them for verification. For the verified speech, localization is performed and the enhanced signal is acquired through sub-array beamforming. The verification result is tested again after beamforming to ensure a

high accuracy. We name this front-end block as Verified Speaker Localization (VSL). The multi-channel speech enhancement based on localization and beamforming is only run for the desired voices and the whole system becomes robust to unwanted noises as well as other spontaneous sources of energy. The over computation of beamforming and post processing for unwanted speech signals is also prevented which reduces the computational complexity of the front-end task in multi-speaker scenarios considerably.

Organization of the paper is as follows: The general architecture of the proposed VSL front-end is explained in section 2. It includes a brief overview of VSL components, details of the split-band reconstruction, speaker verification and localization. Scenario of testing and the results achieved are described in Section 3. A VSL based far-field Automatic Speech Recognition (ASR) is also introduced in this section and the effect of the VSL front-end on the performance of this system is evaluated. Finally, concluding remarks are given in section 4.

2. GENERAL OVERVIEW OF THE PROPOSED VSL FRONT-END

The main elements of the proposed front-end signal pre-processing block are: acquisition, reconstruction of the voiced parts, verification, localization and beamforming. The order in which they interact with each other is shown in Fig. 1.

The acquired speech is first analyzed in split-bands to measure the voicing level. For this purpose in the reverberant acoustic environments, a new fundamental frequency estimation algorithm is proposed based on the subspace approach in high resolution spectral estimation. A reconstruction stage for the degraded voiced bands is also proposed prior to the verification. The verification is implemented using Gaussian Mixture Model and a new SRP filtering scheme is proposed based on the voicing likelihood of each frequency band measured in the previous steps to effectively localize the authorized speaker.

In the traditional methods, as discussed in the introductory part, whenever a source of energy is detected by the localization algorithm, the beamforming will then be applied to acquire the enhanced signal. These two processes are computational intensive in the far-field interfaces. In the proposed Verified Speaker Localization (VSL) front-end, a new localization algorithm improves the speaker localization accuracy as well as the robustness against the reverberation and noise, while the verification which is performed prior to localization prevents the over computation of localization and beamforming for unwanted sources (specially transient or

unauthorized speakers). Therefore, the whole system will have the capability to update the location information of any specific individual. On the other hand, since the localization is based on short speech frames, it is also capable of tracking a moving speaker. These two capabilities indicate that the system can lock on a speaker, while ignoring other speech sources. Since localization and beamforming are highly computational demanding [11] and achieving an enhanced speech for far-field applications needs heavy processing, this lock on characteristic improves the front-end task both in terms of computation and robustness in far-field applications such as teleconferencing, voice control and speech recognition where the presence of unwanted speech signals is highly probable.

In the proposed VSL front-end, the received signal is first segmented based on detection of the non-speech activity for more than 2 seconds. Each segment is analyzed for voicing level measurement at speech sub-bands corresponding to the fundamental frequency harmonics. The voiced parts are then reconstructed at split bands regarding the harmonic bands of the speech spectrum and the signal is analyzed for authentication within a verification algorithm. For the verified speech, misdetection of source localization due to reverberation and acoustic noise is reduced through the voicing level measurement. The beamforming algorithm uses this information to steer the beam pattern towards the direction of the speaker to acquire the source signal while suppressing the noise from other directions. Details of each component are discussed in the following sections.

2.1. Microphone Array Signal Model

In this paper, we assume the sound wave propagation follows a linear wave equation [18]. Hence, the acoustic path between the sound sources and microphones can be modeled as a linear system [19]. This assumption is plausible in small-room microphone array environments and is usually employed in the array-processing techniques [20]. With these assumptions the produced signal by the m^{th} microphone at location d_m can be expressed as:

$$x_m(t) = s(t) * h_s(d_m, d_s, t) + v_m(t) \qquad (1)$$

where $h_s(d_m,d_s,t)$ is the room impulse response from the speech source $s(t)$ at location d_s to microphone m. The operator * is convolution. v_m is a white Gaussian and is assumed to be uncorrelated to $s(t)$.

The impulse response h, characterizes all the acoustic paths from the source to location d_m, including the direct path. In general, h_s varies with environmental changes, such as temperature, humidity, furniture and people inside the room. It is reasonable to assume these factors to remain fixed in the period of each experiment. Separating the direct path component from the rest of the acoustic paths, the following expression can be defined for $h_s(d_m,d_s,t)$:

$$h_s(d_m, d_s, t) = \frac{a}{r_m}\delta(t - \tau_m) + u(d_m, d_s, t) \qquad (2)$$

where r_m is the distance between the source and the m^{th} microphone, τ_m is the propagation delay equal to the ratio of r_m to the speed of sound. The constant a depends on the medium and the system of units used. $u(d_m,d_s,t)$ characterizes all the acoustic paths except the direct path. Substituting this equation into (1), the signal model at microphone m is given by:

$$x_m(t) = \frac{a}{r_m}s(t - \tau_m) + s(t) * u\ (d_m, d_s, t) + v_m(t) \qquad (3)$$

The first term is the direct path component which is important for localization, the second term is the model of reverberation and the third term is the uncorrelated noise.

2.2. Split-band Reconstruction

A typical simulated room impulse response is illustrated in Fig. 2. The largest peak corresponds to the direct path and the other peaks are due to the surrounding walls reverberation. Assuming the total system of microphone array and room as a linear system [21], the received signal at each microphone is the convolution of this impulse response with the original source signal. This effect impairs the received signal quality at the microphone array and reduces the periodicity of the voiced segments. Hence we have considered this side-effect and have enhanced these harmonic parts through reconstruction.

The first step is the estimation of the fundamental frequency. However, due to the distortion of periodicity and harmonicity, conventional fundamental frequency extraction algorithms such as Autocorrelation Function (ACF), Average Magnitude Difference Function (AMDF), Cepstrum, Simple Inverse Filtering Tracking (SIFT) and Harmonic Product Spectrum (HPS) give erroneous results. Since the estimation accuracy of the fundamental frequency in the presence of noise and reverberation is very important for the performance of the

whole system, we have extracted the fundamental frequency on the subspace to benefit from the high resolution spectral estimation property of this technique.

The subspace based spectral estimation is an accurate method for detecting the discrete frequencies of a signal and hence we used the MUltiple Signal Classification (MUSIC) [22, 23] in our algorithm. The MUSIC algorithm detects complex sinusoids by performing eigendecomposition on the data vector covariance matrix of the received signal. Andrews et al. [24] have already proposed the pitch determination algorithm based on MUSIC. Here we have modified their approach for the reverberant signals. To find the fundamental frequency, the autocorrelation matrix of the speech signal is computed from its power spectrum via FT. Since the fundamental frequency of speech sources is less than 800 Hz [25], we have applied the MUSIC algorithm only to the lower frequency components of the speech spectrum. With an 800-point DFT of 20 ms of the speech signal at the sampling frequency of 16 KHZ, the frequency components of a MUSIC spectrum will be at 20 HZ, 40HZ, …, 800 HZ. The total number of these components is 40 and the eigenvalues are computed from the received signal autocorrelation matrix. The number of harmonics contained in the spectrum is an important parameter of the MUSIC algorithm. If it is set too large, the spectrum will be easily affected by the noise and if it is too small, the spectral estimation becomes inaccurate and the error will be increased. For our experiments, the set of dominant eigenvalues $\{\lambda_k\}$ which span over the signal subspace are chosen so as to satisfy $\lambda_1 \geq \lambda_k \geq \lambda_1 / 8$, where λ_1 is the eigenvalue of the first fundamental component. The FFT is applied to the logarithm of the MUSIC power spectrum and the peak location of the signal determines the estimated fundamental frequency. To reduce the computational cost, we have estimated the fundamental frequency at the precision of 20 Hz. This was done by searching the psuedospectrum of the signal with 1 Hz precision at the vicinity of 80 Hz around the pre-estimated fundamental frequency. The corresponding frequency of the local maxima is detected as the fundamental frequency.

Since the room can be modeled as a linear system, the frequency content of the received signal is similar to the original sound and it is only distorted in amplitude and phase. Therefore reverberation converts the global maximum of the spectrum to a local maximum with no frequency displacement.

Through a large number of experiments we have verified the robustness of the algorithm to different reverberant noisy environments. The algorithm was also verified for robustness to sudden closure, such as in a

vowel-to-nasal transition, where waveform periodicity is reduced but the fundamental frequency did not change.

After estimation of the fundamental frequency, the algorithm is used to measure the voicing level in each frequency band. An accurate measure of voicing level was applied to Multi-Band Excitation (MBE) coders [26]. The voicing decision was made by calculating the normalized error E_l between the original and the modeled speech spectrum in each frequency band of the fundamental frequency harmonics:

$$E_l = \frac{\sum_{\omega=a_l}^{b_l} \left| X(\omega) - \hat{X}(\omega, \omega_0) \right|^2}{\sum_{\omega=a_l}^{b_l} \left| X(\omega) \right|^2} \qquad (4)$$

where $X(\omega)$ is the speech spectrum of the received signal at the reference microphone channel (#5), ω_0 is the fundamental frequency, a_l and b_l are the first and last harmonics in the l^{th} band, and $\hat{X}(\omega, \omega_0)$ is the estimated speech spectrum calculated in each frequency band as the spectral shape of a Hanning window with a constant amplitude.

To determine the voicing decision, the normalized error, E_l, of the l^{th} frequency band is compared with an adaptive threshold [27]. If the normalized error is less than a threshold, the corresponding frequency band belongs to the target voice and it is reconstructed in the split-bands based on the fundamental frequency harmonics.

Since higher harmonics are more susceptible to reverberation and acoustic noise [28] decision on voicing for the frame was carried out on the majority of the lower half of the speech frequency band. For those intervals when all of the speakers are talking simultaneously, the speech frames loose their periodicity and these frames are not involved in the other phases of the VSL processing.

The speech signal due to acoustic noise is distorted. The distortion can be reduced in voiced parts by precise extraction of the fundamental frequency and then using it to reconstruct the speech spectrum. The split-band mixed voicing decision calculated for each frequency band is utilized to synthesize the voiced speech spectrum. Each harmonic band has a shape similar to the spectral shape of the window used prior to the Fourier transform, whereas the non-voiced bands are random in nature. Therefore, a voiced harmonic band can be finely synthesized as a multiplication of the frequency response of a suitable window centered at the harmonic of

fundamental frequency corresponding to that band with constant amplitude measured with respect to the original signal [29].

Reconstruction of the harmonic bands is given by equation (5). This reconstruction is performed up to the highest voiced band of the speech spectrum.

$$\hat{X}(\omega,\omega_0)=A_{k,\omega_0}W(\omega) \quad 1\le k\le K, \quad \lceil a_k\rceil\le\omega\le\lceil b_k\rceil \tag{5}$$

where $a_k = (k-0.5)\omega_0$, $b_k = (k+0.5)\omega_0$, $\lceil\cdot\rceil$ stands for the nearest integer greater than or equal to, K is the number of harmonics in the 8 kHz 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 defined as:

$$A_{k,\omega_0}=\frac{\sum_{\omega=\lceil a_k\rceil}^{\lceil b_k\rceil}X(\omega)W(\omega)}{\sum_{\omega=\lceil a_k\rceil}^{\lceil b_k\rceil}|W(\omega)|^2} \tag{6}$$

For concatenation of the reconstructed successive frames, we use linear interpolation to remove frequency mismatches [30]. Fig. 3 displays a clean speech, noisy signal and the synthesized speech from its noisy origin by spectrogram. This figure shows how reconstruction procedure reduces the acoustical noise and retrieves the harmonicity of voicing speech.

2.3. Speaker Verification

Mixture models belong to a family of density model that comprises of a number of component functions, usually Gaussian. The distribution of feature vectors was extracted from a speaker's speech modeled by a Gaussian mixture density. This is a method that has been proven to be one of the most successful approaches for text-independent speaker verification. Therefore we have implemented speaker modeling based on the Gaussian Mixture Models (GMM). In this algorithm Gaussian mixtures are used to model arbitrary densities of the speech signal [31, 32, 33].

A block diagram of the implemented speaker verification system is shown in Fig. 4. There are two steps in the process of speaker verification. In the first step, called the *training phase*, each registered speaker has to provide speech samples. After removal of silence intervals, the MFCC (Mel Frequency Cepstrum Coefficient) features

are extracted for speech frames and the effect of channel distortion is reduced by Cepstral Mean Subtraction (CMS). To build a reference model for every speaker, the parameters of a GMM are calculated for each speaker by determining the mean vector and covariance matrix of the Gaussian densities and the mixture weights. In addition, a threshold is also set from the training samples; the threshold is important for the final rejection or acceptance of a user, and it is independent of the acoustic characteristics of the environment. The second phase is the actual verification, where the input speech is compared with the stored reference models and the recognition decision is made to accept or reject a speaker.

It should be noted that to enhance speaker recognition, we have developed a pre-step gender recognition based on the maximum a posterior probability for a given observation sequence in the gender recognition phase. It is based on a GMM model and is separately trained for female and male speakers.

2.3.1. Decision parameter

The general approach proposed by Reynolds et al. [34] for speaker verification is to apply a likelihood ratio test to an input utterance to determine if the claimed speaker should be accepted or rejected. Given an utterance of speech signal x, a claimed speaker is identified with the corresponding model Ψ_c and an anti-model $\Psi_{\bar{c}}$. Discarding the constant prior probabilities for claimant and imposter speakers, the likelihood ratio in the log domain becomes:

$$\Lambda(x) = \log p(x|\Psi_c) - \log p(x|\Psi_{\bar{c}}) \qquad (7)$$

The term $p_i(x|\Psi_c)$ is the likelihood of the utterance if it is from the claimed speaker and $p(x|\Psi_{\bar{c}})$ is the likelihood of utterance if it is not from the claimed speaker. The likelihood ratio is compared to a threshold ϵ and the test speaker is accepted if $\Lambda(x) > \epsilon$, otherwise it is rejected.

To increase system accuracy, we first find the model $\hat{\Psi}$ which has the maximum a posteriori probability for the speaker sequence using equation (8). In case this model is the model of the claimed speaker, the likelihood ratio of the equation (7) is calculated and the verification decision is made. Otherwise, the claimed speaker will be rejected.

$$\hat{\Psi} = \arg\max_{1 \le i \le N} \{ \log(p(x \mid \Psi_i)) \} \qquad (8)$$

where $p(x|\Psi_i)$ is the likelihood of the input vector x for mixture model Ψ_i of speaker i. N is the total number of all registered speakers. A block diagram of the proposed speaker verification model is shown in Fig. 5. The model is able to prevent misdetection between two speakers of similar sounds by finding the associated model of the imposter speaker and rejects it prior to the computation of the threshold function. This sub-system has been tested on-line and proven to be highly robust and accurate in practice [35].

2.4. SRP Sound Localization

Source location in spherical coordinates is represented by range ρ, azimuth θ and elevation $\varnothing$. If the source range is larger than the array dimension within a specific threshold [36], its wave front is received in a planar form and the range accuracy becomes ambiguous. To alleviate the ambiguity, the source position could be specified within specific θ and $\varnothing$ and its directional vector is defined:

$$\vec{\zeta}_o^{(s)} \equiv \begin{bmatrix} \cos\phi \sin\theta \\ \cos\phi \cos\theta \\ \sin\theta \end{bmatrix} \qquad (9)$$

In this case the steering delay of microphone m relative to a reference microphone is calculated through equation (10).

$$\Delta_m = \left\lceil \frac{dr_m \cos\beta}{c} \times F_s \right\rceil \qquad (10)$$

where dr_m is the distance between the reference microphone and microphone m, c is the speed of sound, β is the ratio of wavefront angle to the microphones intersection line and F_s is the sampling frequency. $\lceil \cdot \rceil$ stands for the nearest integer greater than or equal to.

In the algorithm based on the Steered Response Power (SRP), a microphone array beam pattern is steered towards the candidate positions, the so called beamforming. The output power of a beamformer is then computed and the source position is determined based on the beamformer maximum power using the Maximum Likelihood (ML) estimation. The output of a filter-and-sum beamformer in the frequency domain is defined in equation (11).

$$Y(\omega) = \sum_{m=1}^{M} G_m(\omega)\, X_m(\omega)\, e^{-j\omega\Delta_m} \qquad (11)$$

where $X_m(\omega)$ is the discrete Fourier transform of the received signal from microphone m, $G_m(\omega)$ *is the corresponding filter for microphone m,* $\Delta_1,\ldots,\Delta_M$ are integer values representing the steering delays derived from equation (10) for candidate positions in the space. To avoid rounding up errors of Δ_m to integer values, prior to localization, the signal is upsampled to 96 KHz. The Steered Response Power (SRP) is calculated by:

$$P_{\Delta_1 \ldots \Delta_M} = Y(\omega) Y^T(\omega) \qquad (12)$$

where $Y(\omega)$ is the horizontal output vector of the beamformer and Y^T is its transpose. Although steering delays are continuous variables, equation (10) is computed for discrete locations in space.

In calculating SRP, choice of a suitable filter has a considerable impact on the robustness of localization to both noise and reverberation. In a well-known SRP-PHAT algorithm [10], the filter used at each channel is given by:

$$G_m(\omega) = \frac{1}{|X_m(\omega)|} \qquad for\ m = 1 \ldots M \qquad (13)$$

2.4.1. A New Filtering Scheme Exploiting Harmonic Structures

Since the received signal at microphone array is contaminated by multi-path and noise signals, the periodicity of the voiced segments is reduced. This phenomenon can be seen in Fig. 6, where T60 identifies the room reverberation time. It is measured as 10 times the logarithm of the normalized squared impulse response amplitude and is represented in the form of e.g. T60, implying the time needed for this value to decay from 0 to -60 dB [37].

The side effect of reverberation is presented in Fig. 6, where speech frames with periodic structures are less influenced by reverberation and noise and must have higher weights than the other parts in the localization algorithm.

An accurate measurement of voicing level in each frequency band was described in section 2.2. As mentioned, the voicing decision is made by calculating the normalized error E_l between the original and the modeled speech spectrum for each frequency band l. For voiced frames, E_l has a value close to zero; and values near to 1 correspond to the noisy and non-periodic intervals. Therefore, the calculated error from equation (4) is used to measure the degree of periodicity for each frequency band and can be employed in a filtering scheme for SRP localization. The filter to be employed at each channel is

$$G_{l,m}(\omega) = \frac{1 - E_l}{|X_m(\omega)|} \ , \ \omega \in [a_l, b_l] \qquad (14)$$

where E_l is calculated for the received signal at microphone m, $X_m(\omega)$. This filter will emphasize the voiced frames. Furthermore, the influence of the signal amplitude will be omitted and only the phase information will be used in the localization process which leads to improvement in the robustness of this algorithm to both noise and reverberation.

In practice, 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 by this pitch period. Therefore, if a channel signal for some reason is degraded, its influence will be reduced. The employed filter in beamforming algorithm and the output of the steered array is computed by equation (15) and its power is calculated for that particular point in space. We name the proposed method SRP-H, as it is based on beamforming and analyses of the fundamental frequency harmonics of the speech signal.

$$Y^{SRP-H}_{\Delta_1 \dots \Delta_M}(\omega) = \sum_{m=1}^{M} \sum_{l=1}^{L} \frac{1 - E_{l,m}}{|X_m(\omega)|} X_m(\omega) e^{-j\omega\Delta_m} \qquad (15)$$

3. TEST SCENARIO AND THE TENTATIVE RESULTS

The VSL front-end performance was evaluated under different conditions of noise and reverberation. The impact of these acoustic parameters on a VSL based Automatic Speech Recognition (ASR) is quantified in this section. Fig. 7 shows a VSL based ASR.

The original speech has been chosen from TIMIT database and the far-field signal received at each microphone channel was simulated based on the theories explained in section 2.1. The TIMIT speech data was recorded with a close-talking microphone of sampling frequency of 16 kHz. The New York City subset comprising of 13 females and 22 males were used. This database was divided into a training set and a testing set. The training set for each speaker comprised of ten sentences and these sentences were used in our evaluation. The first five sentences were concatenated and used as the training data for the speaker verification. The remaining five sentences were used to test the speaker verification.

Fig. 8 shows the relative positions of sources to the microphone array. In each scenario it was assumed one of the speakers was the wanted speaker and the total speeches including some segments from the others were the

subject of VSL testing for localization and beamforming. Our simulation test room has a dimensions of $4 \times 6 \times 4$ m^3 and the speech sources were at the same level as the linear array. Three speakers were positioned at three specific angles with respect to the array.

Room acoustic was simulated by the Image method [38] for different reflection coefficients corresponding to T60 = {0.27s, 0.47s}. Signal to noise ratio at the reference microphone was simulated by adding an uncorrelated noise simulated for each microphone channel by equation (1) for SNR = {5dB, 15dB, 25dB}. For better beam pattern steering, we had to capture different sub-bands from different arrays. This sub-array design is described in the following sub-section.

3. 1. VSL implementation

The received signal at the reference microphone was processed for the voice activity detection and segmentation. To detect speech frames, the received signal was analyzed for the stationary or non-stationary properties. Stationary frames were identified if the ratio of minimum to maximum powers was above a maximum threshold. The frame was assumed to be non-stationary if this ratio was below a minimum threshold. Since difference between the maximum and minimum powers is large, then for ratios between these two threshold levels, the algorithm keeps its previous decision [27]. Threshold values were adaptively determined based on the last 8 frames. Algorithm decisions were also followed for successive frames. Detection of one speech frame among the several non-speech frames can be erroneous. After the estimation of the background noise power from the non-speech frames, frames with power close to noise power were removed from the VSL procedure. In our experiments, speech intervals larger than two seconds of silence were to be gaps of speech segments.

The received signal was then analyzed every 20 ms to extract the fundamental frequency and the voiced parts were reconstructed. These parts were then matched against the trained speaker models. Prior to verification, the signal was pre-emphasized with a factor of 0.95 and the speech signal was framed in a Hanning window of 20 ms long with a 10 ms overlap. On each frame a 13^{th} order MFCC and a log energy analysis were performed [39] and the first and the second differentials were extracted from speech to form a 39 dimensional feature vector. There were 20 filters in the filter bank and to increase the accuracy, Cepstral coefficients at the lower and upper coefficients had smaller weights than the middle coefficients [40]. Finally, the Cepstral Mean Subtraction method was utilized to remove the channel effects on the Cepstral features [29, 41]. For the verified frames,

localization was performed by upsampling the input signal to 96 KHz, that would increase the accuracy and we were able to retrieve 0.23 sub-samples corresponding to one degree precision in space scanning. The steering delays of beamforming were set according to the direction of the desired source and the signal was acquired by a delay-and-sum beamforming over sub-arrays.

Array response which is a function of frequency and microphone distance is known as a beam-pattern. At higher frequencies, smaller arrays provide similar patterns. Therefore, the speech signal is divided into sub-bands and each frequency band is captured from different arrays called sub-array. The number and placement of microphones in each sub-array have important effects on the quality of the speech acquisition. Therefore, by investigating on the superdirective microphone array designs [42] and the spatial non-aliasing rules, 11 microphones were placed and the speech signal was divided into 5 sub-bands. These sub-bands are summarized in Table 1. Hence, the microphone array comprised of five nested sub-arrays where each sub-band was received from the assigned sub-array. Finally, all the sub-bands were combined to form the received signal.

For the scenario when all the speakers are talking, the intervals of simultaneous talks loose the periodicity nature of the voiced parts. Hence, these parts were not reconstructed and were not considered for verification and localization. Therefore, the whole system is robust to the speech of unwanted speakers and location information is not updated.

3.2. Speech Recognition engine

To evaluate the performance of the proposed VSL front-end, continuous speech recognition experiments were carried out through a variety of tentative settings. A continuous phoneme recognizer with Hidden Markov Model (HMM) [43] was used to model the speech phonemes. Modeling is based on continuous density HMMs with Gaussian compositions and only left to right jumping is permitted. An HMM with 6 statuses was trained for each phoneme and the number of compositions was set to 16. Generally, the system accepts an input speech and outputs a stream of phonemes with the best correspondence to the acoustic input stream [44, 45]. It should be mentioned that this system is speaker independent. Speech frames were 20 ms with 60% overlapping. The vector of the extracted features contained 12 Mel-Cepstrum coefficients and the first and second differentials. HMMs topology was the same for all phonemes.

3.3. VSL Results

The speaker verification accuracy is presented in Table 2. The verification result was recursively checked after beamforming. In case of ambiguity in decision, the system keeps the last decision. It can be seen that the verification accuracy in an average level of room reverberation (T60 = 0.27) is improved by up to %90 in low SNRs. For higher reverberation (e.g. T60=0.47), comparing Tables (2.a) to (2.b) reveals that due to reverberation the performance is reduced by almost %5 but this reduction is %26 when the signal to noise ratio is also reduced by 10dB (from SNR=15dB to SNR=5dB). Therefore, the acoustical additive noise is more destructive than the reverberation. However, this table shows that the accuracy of the proposed method of verification is %90 which is quite acceptable for practical applications. Reconstruction, particularly reduces the additive noise and use of channel compensation techniques such as CMS, reduces the reverberation side effect on the verification performance.

To measure the accuracy and robustness of direction estimation, anomaly statistics [46] were calculated over the ensemble of speech segments at various signal to noise ratios in high reverberant environments. Fig. 9, shows the percentage of estimates outside a 10° absolute error threshold as a function of SNR. It can be seen the SRP has much poorer performance than our two methods of SRP-PHAT and SRP-H, especially at lower SNRs. Also, through detection of periodic structures in the proposed localization algorithm of SRP-H (equation (16), section 2.4.1), the influence of destroyed frames is reduced by more than 10% over SRP-PHAT in high reverberant and high noisy conditions.

Having estimated the position of the speaker, the time delays for the microphone signals are set such that the beam pattern is directed towards the source and the speech sub-bands are received from the assigned sub-arrays. The SNR improvement over a single channel (reference microphone) is represented in Fig. 10. These results are for the scenario when the noise source is located right in front of the reference microphone at the distance of 3.5m and the speech source is the speaker 1 (Fig. 8). The relative angle of the speech source to the noise source is chosen so that the beampattern after beamforming has the minimum gain at the noise direction. The experiments of this scenario indicate the highest effectiveness of sub-array beamforming in SNR improvement by achieving similar beam pattern for the entire speech frequency band .

3.4. Speech Recognition Results

The enhanced signal through the above process was tested for feature extraction and recognition. Table 3 shows Phoneme Accuracy Rate or PAR[1] for different noise levels and reverberant conditions.

The first row of this table is the phoneme recognition accuracy of a single channel signal. The influence of location information on performance improvement of the proposed speech recognition system can be inferred by contrasting the first row against the other rows. In the second row, all sources are localized and beamforming is performed without verification. It can be seen that due to unwanted voices the recognition accuracy is considerably decreased. Negative values in the Table mean the number of phonemes which are erroneously inserted, deleted or replaced are greater than the number of all recognized phonemes.

The third row shows the outcome of the recognition after applying the VSL techniques. Since the whole speech band is received through all the microphones and due to the beam pattern variation at different frequencies, still some noise is received from the side lobes. However, sub-array and sub-band beamforming improves these results considerably (up to %52) as shown in the fourth row of the Table.

4. CONCLUSION

In this paper, we have proposed the VSL front-end design of microphone array interfaces for far-field applications. The idea is based on the realization that the most important effect of reverberation and acoustic noise is the voicing distortion. Therefore, by measuring the voicing level at split bands of speech signal through a novel approach of fundamental frequency extraction based on subspace decomposition, the voiced parts are reconstructed to reduce the effect of acoustic noise and reverberation as we are now able to verify the received signal prior to localization. For the verified speech, a new filtering scheme is employed for localization and the speech signal is acquired by sub-array beamforming. The verification result is tested again after beamforming to ensure a high accuracy and robustness. Due to the achievement of verification prior to localization, the multi-channel speech enhancement based on localization and beamforming is only run for the desired voices and hence the whole system becomes robust to speech noises and other spontaneous sources of energy. In addition, the over computation of localization and beamforming for the unwanted signals is prevented and the

computational complexity in multi-speaker scenarios is reduced considerably. Since the localization is not performed for unwanted speakers and its algorithm works favorably based on short frames, the proposed VSL can lock on a moving speaker while ignoring the other sources.

The VSL was then utilized in a far-field speech recognition system. It can be concluded that the use of VSL provides a considerable recognition accuracy improvement by up to 52% in the presence of speech noises. Reverberation however has an important destructive effect on our recognition engine performance. Therefore beampattern steering utilizing speaker localization is a very effective method but is not sufficient in high reverberant environments and dereverberation algorithms must be considered to recover the signal from channel effects for distant-talking speech recognition system.

ACKNOWLEDGEMENT

We would like to thank Dr. Mohammad Shahram Moin, director of the multimedia research group in the department of information technology in Iran Telecommunication Research Center for his support and Mr. Amir Hossein Keyhanipoor and Mr. Mehdi Hosseinpour in the department of information technology in Iran Telecommunication Research Center for their valuable comments.

REFERENCES

[1] B. Mungamuru, P. Arabi, "Enhanced sound localization", IEEE Transaction on Systems, Man and Cybernetics-Part B: Cybernetics, Vol. 34, no. 3, June 2004

[2] B. Allen, M. Ghavami, Adaptive array systems, fundamentals and applications, Wiley Publisher, 2005

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

[4] H. F. Silverman, "Some analysis of microphone arrays for speech data acquisition", Transaction on Acoustics, Speech, and Signal Processing, 35(12):1699-1711, December 1987

[5] D. B. Ward, R. A. Kennedy, R. C. Williamson, "An Adaptive Algorithm for Broadband Frequency Invariant Beamforming", Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing, pp. 3737-3740 vol.5 April 1997.

[6] M. M. Goodwin and G. W. Elko, "Constant beam width beamforming", Proc. IEEE Int. Conference on Acoustic,

[1] *PAR = (Number of all recognized phonemes – Number of erroneously inserted phonemes – Number of erroneously deleted phonemes - Number of erroneously replaced phonemes) / (Number of all phonemes)*

Speech, Signal Processing, May 1993,. pp. I-169–172

[7] J. C. Chen, K. Yao, and R. E. Hudson, “Source localization and beamforming”, IEEE Signal Processing Magazine, March 2002

[8] A. Asaei, M. J. Taghizadeh, S. Ghanbari, 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, pp 129-133, March 2007

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

[10] M. Omologo and P. Svaizer, “Use of the crosspower-spectrum phase in acoustic event localization”, IEEE Transaction on Speech Audio Processing, vol. 5, pp. 288-292, May 1997

[11] M. Brandstein, D. Ward, Microphone Arrays Signal Processing Techniques and Application, Springer, 2001

[12] Q. Lin, E. Jan, J. Flanagan, “Microphone Arrays and Speaker Identification”, IEEE Transaction on Speech and Audio Processing, vol. 2, no. 4, October 1994

[13] T. Giannakopoulos, N. Tatlas, T. Ganchev and I. Potamitis, “A Practicalm, Real-Time Speech-Driven, Home Automation Front-end”, IEEE Transactions on Consumer Electronics, vol. 51, no. 2, May 2005

[14] I. McCowan, J. Pelecanos, S. Sridharan, “Robust Speaker Recognition using Microphone Arrays”, In Proceeding of 2001: A speaker odyssey, June 2001

[15] J. Gonzalez-Rodriguez, J. Ortega-Garcia, C. Martin, L. Hernandez, “Increasing Robustness In GMM Speaker Recognition Systems For Noisy and Reverberant Speech With Low Complexity Microphone Arrays“, Proceeding of the Fourth International Conference in Spoken Language, ICSLP, vol. 3, pp. 1333-1336, October 1996

[16] J. W. Stokes, J. C. Platt and S. Basu, “Speaker Identification Using A Microphone Array and a Joint HMM With Speech Spectrum and Angle of Arrival”, IEEE International Conference on Multimedia and Expo, pp. 1381-1384, July 2006

[17] R. Xu, G. Mei, Z. Ren and C. Kwan, “A Real Time Speaker Verification Demonstration on The Smart Flow System”, Proceedings of 2004 International Symposium on Intelligent Multimedia, Video, and Speech Processing, pp. 226-229, October 2004

[18] L. E. Kinsler, Fundamentals of Acoustics. John Wiley & Sons, New York, 3rd edition, 1982

[19] . Ziomek. Fundamentals of Acoustic Field Theory and Space-Time Signal Processing. CRC Press, Inc., 2000 Corporate Blvd., N. W., Boca Raton, Florida 33431, 1995

[20] D. H. Johnson and D. E. Dudgeon. Array Signal Processing: Concepts and Techniques. P T R Prentice Hall, Englewood Cliffs, New Jersey 07632, 1993

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

[22] P. Stoica, R. Moses, Spectral Analysis of Signals, Pearson Prentice Hall, 2005

[23] M. G. Christensena, P. Stoicab, A. Jakobssonc, S. H. Jensen “Multi-pitch Estimation”, Elsevier Journal on Signal Processing, vol. 88, 972–983, 2008

[24] M. S. Andrews, J. Picone and R. D. Degroat, ‘‘Robust pitch determination via SVD based Cepstral methods’’, ICASSP90, pp. 253–256

[25] W. Hess, Pitch Determination of Speech Signals, Springer-Verlag, New York, 1983

[26] D. Griffin and J. Lim, “Multiband Excitation Vocoder”, IEEE Transaction on Acoustic, Speech and Signal Processing, vol. 36, no. 8, pp. 1223-1235, August 1988

[27] A. M. Kondoz, Digital speech coding for low bit rate communication systems, Wiley Publisher, 2004

[28] N. Roman, D. wang, “Pitch-based monaural segregation of reverberant speech”, Journal of the Acoustical Society of America, vol. 120, pp. 458-469, July 2006

[29] Y. Tokhura, “A Weighted Cepstral Distance Measure for Speech Recognition” IEEE Transactions on Acoustic, Speech and Signal Processing, vol. 35, no. 10, pp. 1414-1422, October 1987

[30] Y. Stylianou, “Applying the Harmonic Plus Noise Model in Concatenative Speech Synthesis”, IEEE Transaction on Speech And Audio Processing, vol. 9, no. 1, January 2001

[31] Chun-Nan Hsu, Hau-Chung Yu, Bo-Hou Yang, "Speaker Verification without Background Speaker Models", ICASSP 2002

[32] D. A. Reynold, R. C. Rose, “Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models”, IEEE Transactions on Speech & Audio Processing, vol.3, no.1, 1995

[33] G. Singh, A. Panda, S. Bhattacharyya, T. Srikanthan, “Vector Quantization Techniques For GMM Based Speaker Verification”, ICASSP 2003

[34] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. “Speaker Verification using Adapted Gaussian Mixture Models”, Digital Signal Processing, vol. 10, pp. 19–41, 2000

[35] M. Bahrololum and M.S Moin, "Speaker Identification based on Gaussian Mixture Models mixed by Gender Recognition", in Proc. Iranian Conference on Electrical Engineering, May 2004, Shiraz, Iran

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

[37] J. H. DiBiase, A High-Accuracy, Low-Latency Technique for Talker Localization in Reverberant Environments Using Microphone Arrays, PhD Thesis, Brown University, May 2000

[38] 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

[39] S. B. Davis and P. Mermelstein, “Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences”, IEEE Transaction on Acoustic, Speech and Signal Processing, vol. 28, pp 357-366, 1980

[40] B.H. Juang , L.R. Rabiner & J.G. Wilpon , “On The Use Of Band-pass Liftering In Speech Recognition “, IEEE Transactions on Acoustic, Speech and Signal Processing, vol. 35, no. 7, pp. 954-974, July 1987

[41] M. J. F. Gales, Model-Based Techniques for Noise Robust Speech Recognition, PhD Thesis, University of Cambridge, 1995

[42] W. Tager, “Near field superdirective (NFSD)”, ICASSP98, pp 2045-2048

[43] A. P. Varga, and R. K. Moore, “Hidden Markov Model decomposition of speech and noise”, ICASSP90, pp. 845-848, April 1990

[44] B. Babaali, H. Sameti, “The Sharif Speaker-Independent Large Vocabulary Speech Recognition System“, The 2nd Workshop on Information Technology & Its Disciplines (WITID 2004), Feb. 24-26, 2004, Kish Island, Iran.

[45] A. Asaei, Sound Source Localization by Beamforming Techniques for Robust Speech Recognition, M.Sc. thesis, Sharif University of Engineering, Tehran-Iran, November 2007

[46] M. Brandstein, “A pitch based approach to time delay estimation of reverberant speech”, Proc. IEEE ASSP Workshop Appls. Signal Processing Audio Acoustics, 1997.

FIGURES

Fig. 1. A general Architecture of the proposed VSL front-end

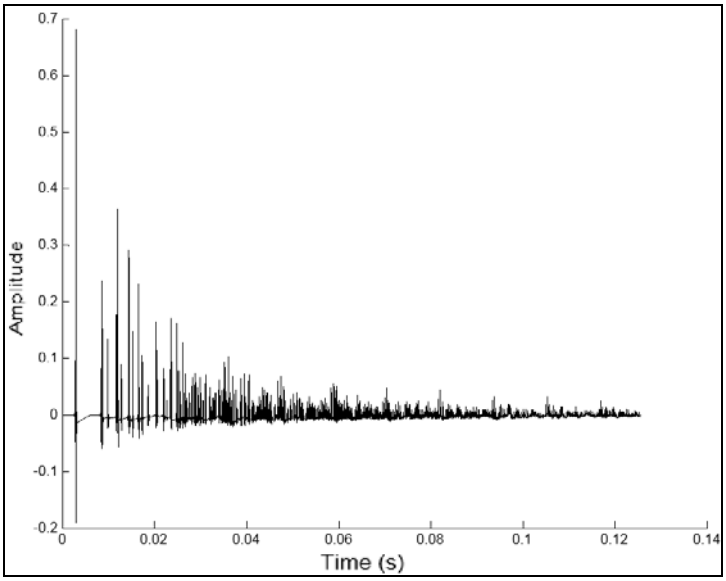
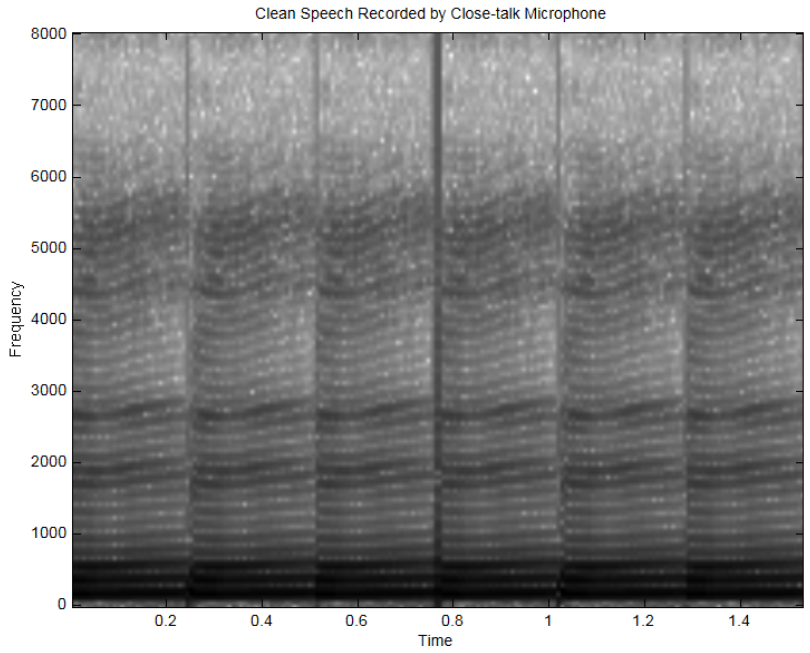
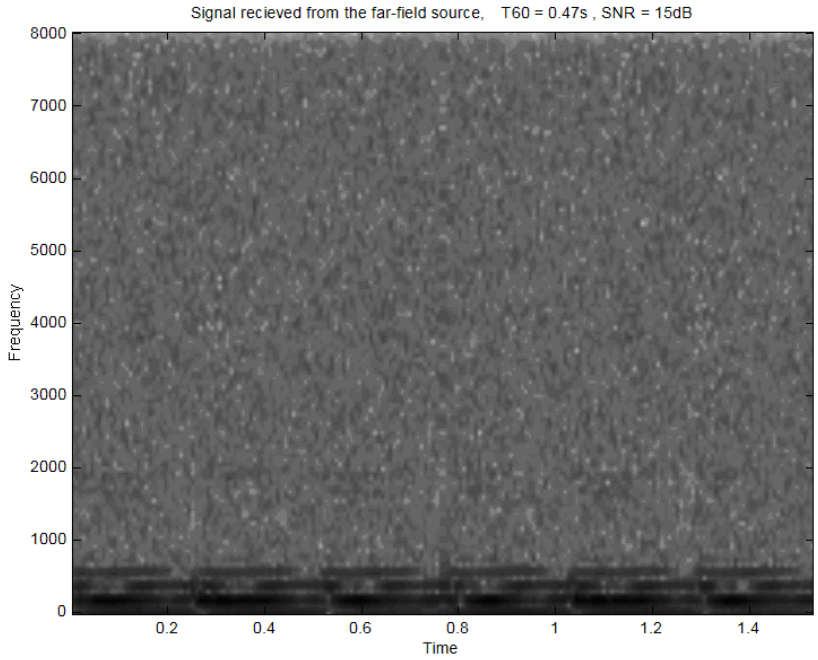


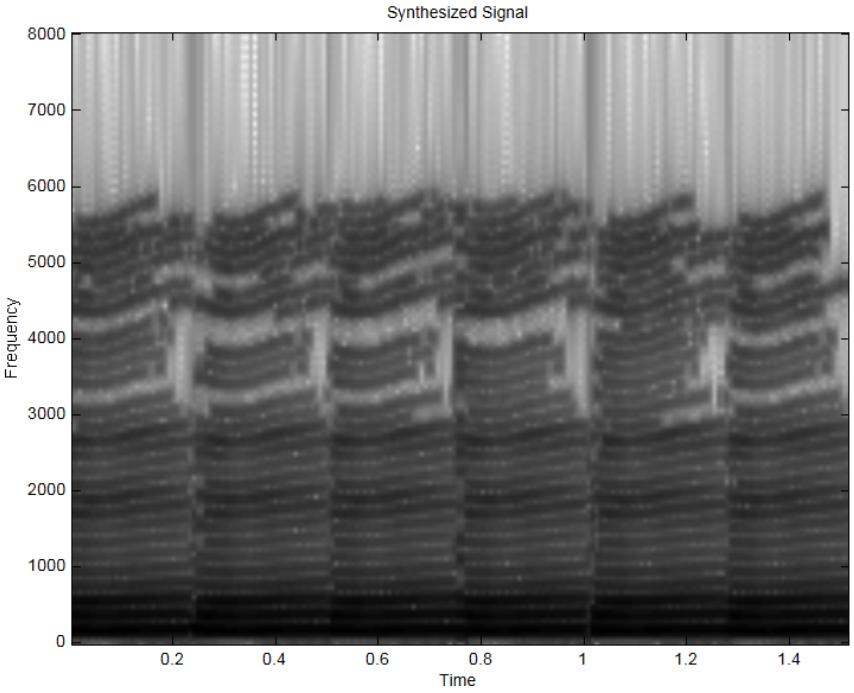
Fig. 2. Room Impulse Response


Clean Speech Recorded by Close-talk Microphone
8000
7000
6000
5000
4000
3000
2000
1000
0
Frequency
0.2
0.4
0.6
0.8
1
1.2
1.4
Time

(a)


Signal recieved from the far-field source, T60 = 0.47s , SNR = 15dB
8000
7000
6000
5000
4000
3000
2000
1000
0
Frequency
0.2
0.4
0.6
0.8
1
1.2
1.4
Time

(b)



(c)

Fig. 3. Spectrogram of (a) clean close-talk speech, (b) Far field signal, (c) Synthesized speech

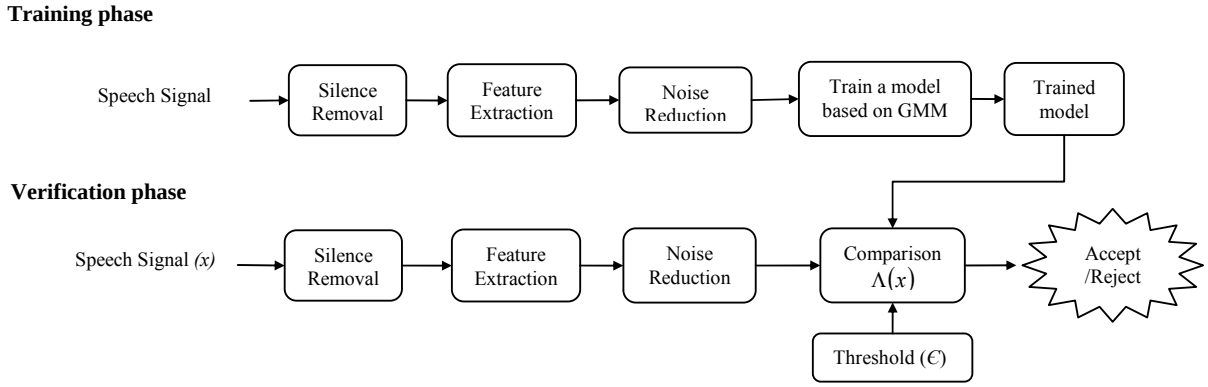


Fig. 4. The block diagram of the two phases in a speaker verification system

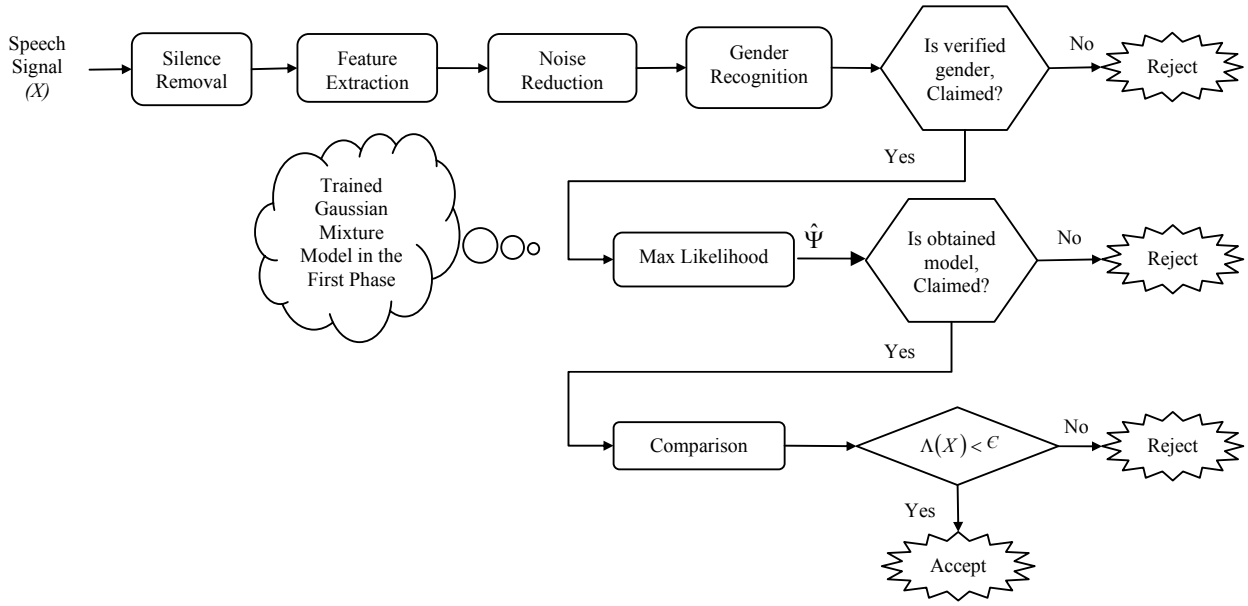


Fig. 5. The proposed speaker verification model

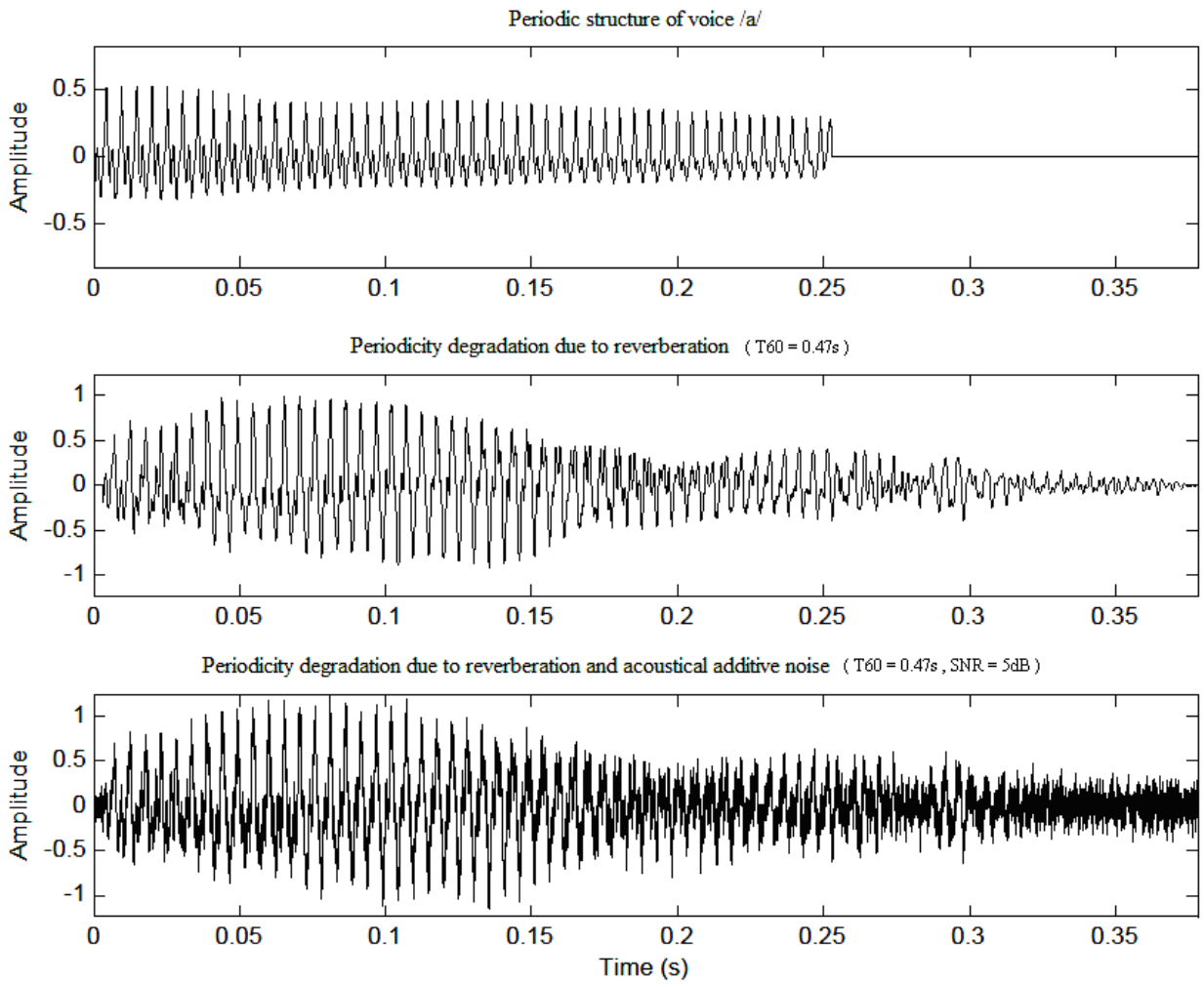


Fig. 6. Periodicity degradation due to noise and reverberation: Upper graph: Initial signal of word /a/, Middle graph: Signal affected by reverberation, Lower graph: Signal affected by reverberation and noise

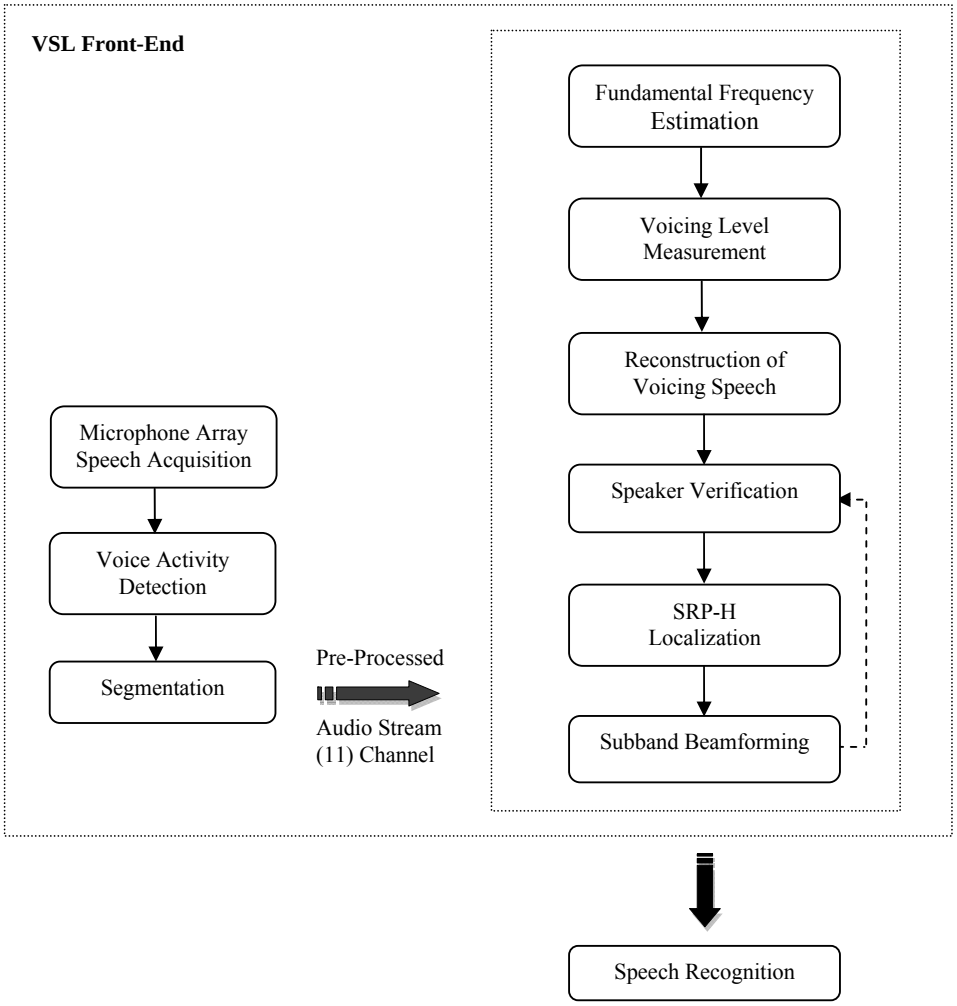


Fig. 7. VSL based Speech Recognition block-diagram

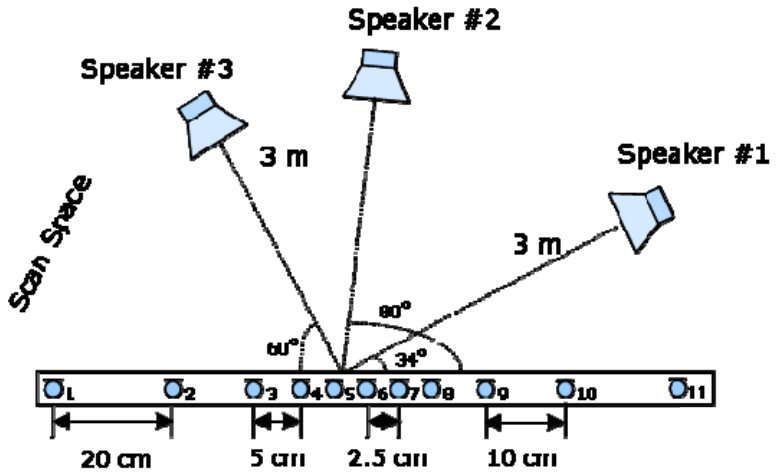


Fig. 8. Microphones and speech sources positions in the simulated scenario with reference microphone at channel 5

Signal to Noise Ratio

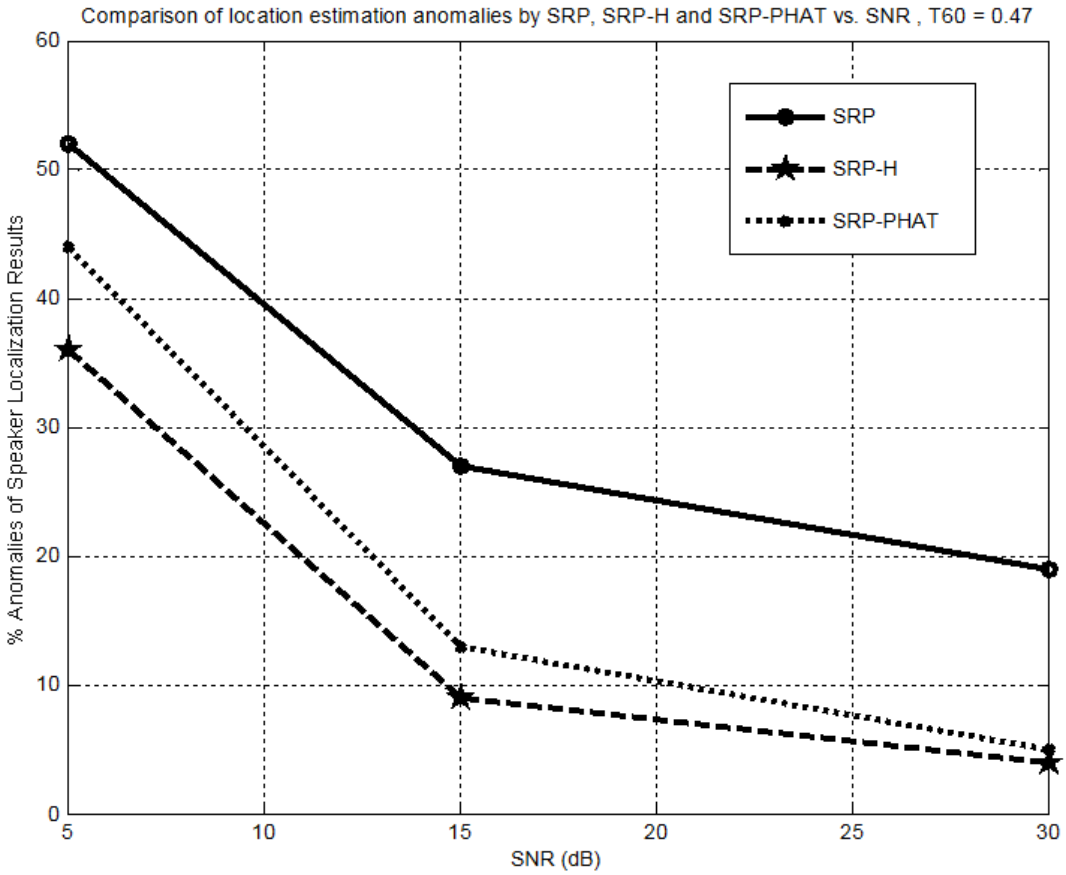


Fig. 9. Anomaly percentage of location estimation by SRP, SRP-H and SRP-PHAT vs. SNR at T60 = 0.47s

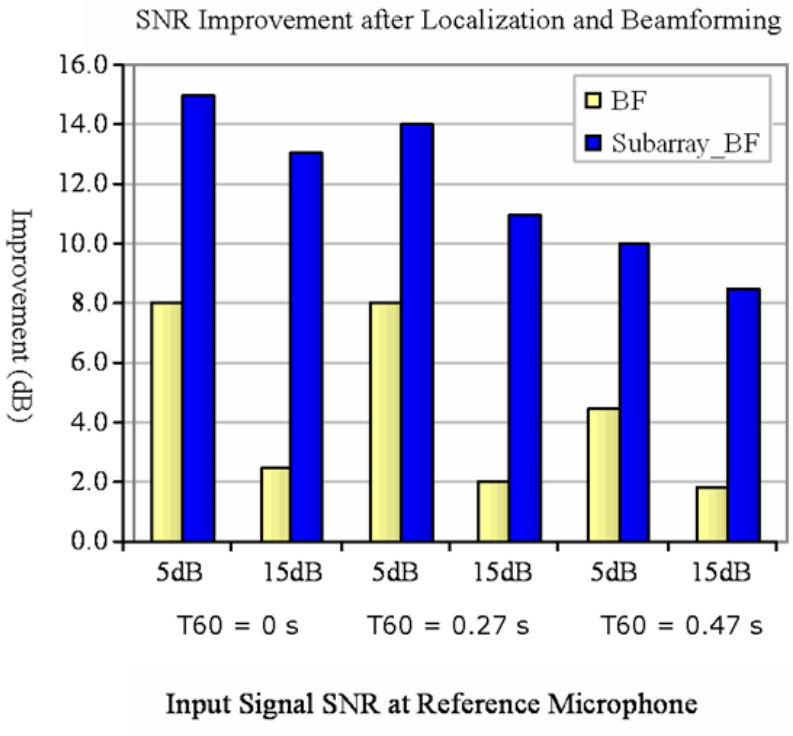


Fig. 10. SNR improvement after beamforming (BF) and subarray beamforming (Subarray_BF)

TABLES

Table 1. Speech Sub-bands and assigned sub-arrays

Frequency Band	Microphone Index	Microphone Distance
Less than 500	1 - 11	-
500 - 1000	1, 2, 6, 10, 11	20 cm
1000 - 2000	2, 3, 6, 9, 10	10 cm
2000 - 4000	3, 4, 6, 8, 9	5 cm
4000 - 8000	4, 5, 6, 7, 8	2.5 cm

Table 2. Speaker verification accuracy rate for various SNRs, a) large reverberation and b) average reverberation

a) Verification accuracy rate vs. SNR (T60=0.47s)

Far-Filed		**VSL**	
Signal to Noise Ratio	**accuracy rate**	**Signal to Noise Ratio**	**accuracy rate**
5 db	30%	5 db	84%
15 db	58%	15 db	88%
25 db	61%	25 db	91%

b) Verification accuracy rate vs. SNR (T60=0.27s)

Far-Filed		**VSL**	
Signal to Noise Ratio	**accuracy rate**	**Signal to Noise Ratio**	**accuracy rate**
5 db	36%	5 db	89%
15 db	59%	15 db	92%
25 db	65%	25 db	97%

Table 3. Speech recognition in the presence of speech noises and verified speaker localization

SNR (dB)	**5dB**	**25dB**

T60 (s)	0.27	0.47	0.27	0.47
1. Single (%PAR)	-12	-18	37	30
2. BF (%PAR)	-25	-30	8	4
3. BF + VSL (%PAR)	-4	-9	42	36
4. Sub_BF + VSL (%PAR)	0	-3	61	56