International Research Journal of Advanced Engineering and Science
ISSN: 2455-9024

Human Voice Localization in Noisy Environment by SRP-PHAT and MFCC

Arjun Das H[1], Lisha Gopalakrishnan Pillai[2], Muruganantham C[3]
[1, 2, 3]Department of ECE, University of Calicut, Pambady, Thrissur, Kerala, India-680588

Abstract**—***This paper proposes an efficient method based on the steered response power (SRP) technique for sound source localization using microphone array: the SRP-PHAT and MFCC. As compared to the previous SRP technique, the proposed method using MFCC is more efficient in localizing the human speech. The previous method focuses only on the loudest sound source in a given area and such sound source may also contain noises which have high spectral peaks than human speech. The proposed method will recognize the speech through vowel corpus comparison and enhance the speaker speech. And finally by using MFCC feature extraction technique the human speech can be extracted from the noisy environment and this speech signal is used for sound source localization using SRP-PHAT.*

Keywords**—** *Mel-frequency cepstral coefficient (MFCC), peak valley difference (PVD, sound source localization (SSL), time difference of arrival (TDOA).*

I. INTRODUCTION

Nowadays, the sound source localization (SSL) had become an attractive field of research in human robot interaction. The problem of locating a human speech in space has become an important criterion in this approach. For this reason many investigations has been done in this research area and several alternative approaches have been proposed in the last few decades. Traditionally, the algorithm for the sound source localization begins with the estimation of direction of arrival of sound signals based on time differences at the microphone pairs through generalized cross-correlation with the phase transform. When several microphones are available for the source position estimation, a point in the space can be accurately estimated as source position that satisfies the set of Time Difference of Arrival (TDOA) values by applying SRP-PHAT algorithm.

SRP-PHAT is one of the robust methods used for sound source localization in noisy and reverberant environment. However, the direct use of SRP-PHAT in real life speech application has shown some negative impact on the performance because, the SRP-PHAT steers the microphone array towards the location having the maximum output power. Also the sum of the cross-correlation values for each microphone signal is measured as output power of the beamformer. Hence, for a given location the SRP-PHAT estimates the power of the voice signal by using only the cross correlation values of the input microphone signal. However if the source of noise is having maximum output power, then the noise signal is determined rather than the voice signal.

To handle such problems, higher weights should be assigned to the speech signal rather than the loud noises for this we consider human speech characteristics into account. The scheme that distinguishes human speech from noise is known as voice activity detection (VAD). This scheme utilizes formants present in the human vowels which are distinctive spectrals that are likely to be remained even after the noise has severely corrupted it.

Efficient sound source localization can be performed if we can extract and enhance the human voice signal accurately in all noisy environments. For this SRP-PHAT and VAD method is modified with MEL-frequency extraction technique which can more accurately derive the human voice.

The remainder of this paper is organized as follows description on Steered Response Voice Power (SRVP) based localization in noisy environment is presented in Section II. Section III describes about the proposed method and IV compares between both the localization methods while conclusion closes the paper in Section V

II. SRVP BASED LOCALIZATION

A. SRP-PHAT

SRP-PHAT based SSL uses an array of N microphones, where the source is given by signal, $x_n(t)$ received by the n-th microphone at time t, the output, y(t,s), of the delay-and-sum beamformer is defined as follows:

$$y(t,s) = \sum_{n=1}^{N} x_n(t + \tau_{n,s}) \quad (1)$$

Where $\tau_{n,s}$ is the direct time of travel from location s to the n^{th} microphone. To deal with complex noises, filter-and-sum beamformers using a weighting function may be used in reverberation cases. In the frequency domain, the filter-and-sum version of Eq. 8 can be written as:

$$Y(\omega, s) = \sum_{n=1}^{N} G_n(\omega) X_n(\omega) e^{j\omega\tau_{n,s}} \quad (2)$$

Where, are the Fourier transforms of the n^{th} microphone signal and its associated filter is represented by $X_n(\omega)$ and $G_n(\omega)$ respectively. In Equation (2) the microphone signals are phase-aligned by the steering delays and summed after the filter is applied.

The sound source localization algorithm based on the SRP steers the microphone array to focus on each spatial point, s, and for the focused point s one can calculate the output power, P(s), for the microphone array as follows:

$$P(s) = \int_{-\alpha}^{\alpha} |Y(\omega)|^2 d\omega$$

$$= \left(\sum_{u=1}^{N} G_u(\omega) X_u(\omega) e^{j\omega\tau_{u,s}}\right)\left(\sum_{v=1}^{N} G_v^*(\omega) X_v^*(\omega) e^{-j\omega\tau_{v,s}}\right) d\omega$$

Arjun Das H, Lisha Gopalakrishnan Pillai, and Muruganantham. C, "Human voice localization in noisy environment by SRP-PHAT and MFCC," *International Research Journal of Advanced Engineering and Science*, Volume 1, Issue 3, pp. 33-37, 2016.

International Research Journal of Advanced Engineering and Science
ISSN: 2455-9024

$$= \sum_{u=1}^{N}\sum_{v=1}^{N}\int_{-\alpha}^{\alpha}\left(G_u(\omega)X_u(\omega)e^{j\omega\tau_{u,s}}\right)\left(G_v^*(\omega)X_v^*(\omega)e^{-j\omega\tau_{v,s}}\right)d\omega$$

$$= \sum_{u=1}^{N}\sum_{v=1}^{N}\int_{-\alpha}^{\alpha}\psi_{uv}(\omega)X_u(\omega)X_v^*(\omega)e^{j\omega(\tau_{u,s}-\tau_{v,s})}d\omega \quad (3)$$

where, $\psi_{uv}(\omega) = G_u(\omega)G_v^*(\omega)$. To reduce the effect of reverberation the filter used in SRP-PHAT is defined as follows

$$\psi_{uv}(\omega) = \frac{1}{|X_u(\omega)X_v^*(\omega)|} \quad (4)$$

After calculating the SRP, *P*(s), for the each candidate location, the point ŝ that has the highest output power is selected as the location of the sound source, i.e.,

$$\hat{s} = \arg\max_{s} \; p(s) \quad (5)$$

B. *Voice Activity Detection (VAD) using Formants in Vowel Sounds*

Though SRP-PHAT is the popular SSL technique it may not be adequate for human speech localization is noisy environment. Because of its inability to distinguish between the voice and the noise signals and it simply compute the output power of the input signals.

So to solve this problem VAD method is employed, which will boost the peaks from the speech signal and reduce peaks from the noise signals.

Human vowel sound contains formants which are well known distinctive spectral peaks and are likely to remain even after the noise has severally corrupted it. These are used for checking the similarities with human speech and distinguish with the noise.

C. *Peak valley difference (PVD)*

The method that combines two algorithms SRP-PHAT and VAD leads to SRVP [5]. Here to consider the content within the voice signal VAD is utilized. The method mentioned here composed three steps:

- SRP is calculated for every candidate. Then the top n- best candidates are selected
- By using an adaptive beamforming method the microphones are focused on the top n-best candidates
- Maximum voice similarity of the beamformed signal is evaluated by comparing it with the pre-trained samples present in the vowel corpus.

Therefore the algorithm computes the voice similarity between the input spectrums to the pre-trained spectrals present in the vowel corpus. So, if the spectral peak is present then the average energies of the spectral bands for that peak will be much higher than the average energies of other bands. It means the peak valley difference (PVD) will be higher. Using the PVD values one can measure the voice similarity. The similarity of the beamformed input spectrum Z_q and the signature S of the given binary spectral peak can be calculated as follows

$$\mathrm{PVD}(Z_q, S) = \frac{\sum_{j=0}^{N-1}(Z_q[j]\times S[j])}{\sum_{j=0}^{N-1}S[j]} - \frac{\sum_{j=0}^{N-1}(Z_q[j]\times(1-S[j]))}{\sum_{j=0}^{N-1}(1-S[j])} \quad (6)$$

Where N is the dimension of the spectrum. For every registered spectral peak signature the above similarity measurement is computed and the maximum value for the spectral peak energy location q can be determined as follows:

$$\mathrm{PVD}(Z_q) = \max_{s} \; \mathrm{PVD}(Z_q, S) \quad (7)$$

The location, $\hat{q}$ of the human voice can be determined by following simple linear algorithm

$$\hat{q} = \arg\max_{q}\left(\frac{P(q)}{P_{max}} + \frac{a.PVD(Z_q)}{PVD_{max}}\right) \quad (8)$$

Where PVD_{max} is the maximum value of PVD while P_{max} is the maximum value of the steered mean output power.

III. SRP-PHAT AND MFCC BASED LOCALIZATION

A. *Mel- Frequency Extraction*

In this method the mel-frequency extraction technique is applied along with SRP-PHAT algorithm in order to accurately extract and enhance the human speech in noisy environment. The advantage of using mel-frequency scaling in sound source localization is it is very approximate to the frequency response of human auditory system and can be used to capture the important characteristics of speech.

The speech extraction and enhancement process can be done using a mel-frequency cepstral coefficient (MFCC) processor. This processing can be explained in five main steps.

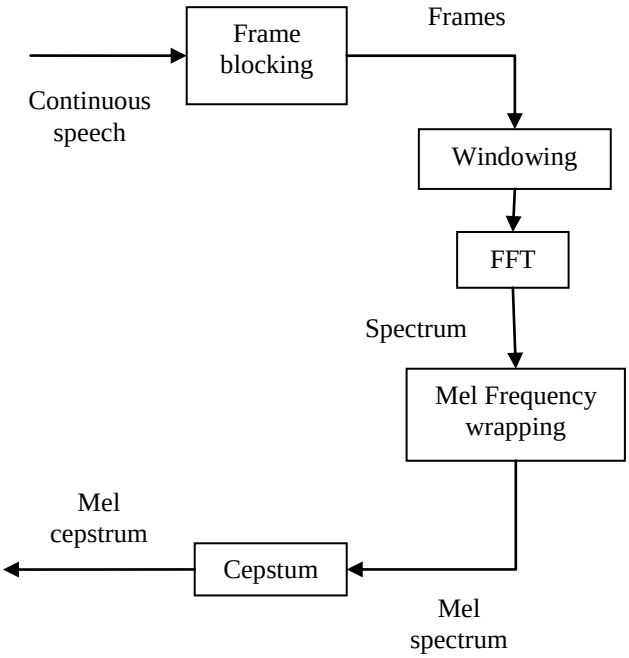

Fig. 1. Block diagram for MFCC processor.

- In the first step, the continuous speech signal is blocked into frames of N samples, where the adjacent samples are separated by M. (M<N)
- In the next step windowing is performed on each of the individual frame so as to minimize the signal discontinuities at the beginning and end of the each frame.

Arjun Das H, Lisha Gopalakrishnan Pillai, and Muruganantham. C, "Human voice localization in noisy environment by SRP-PHAT and MFCC," *International Research Journal of Advanced Engineering and Science*, Volume 1, Issue 3, pp. 33-37, 2016.

- The third step composed of Fast Fourier Transform, which converts each frames of N samples into frequency domain from time domain.
- The fourth step is called Mel-frequency wrapping, where we compute mels for a given frequency f in Hz by following formula

$$\text{mel}(f) = 2595 * \log_{10}(1 + f/700) \quad (9)$$

for simulating the subjective spectrum a filter bank can be used as an approach, one filter for each desired mel-frequency component. The filter bank used here has a triangular band pass frequency response, where the bandwidth and spacing of the filter is determined by a constant mel-frequency interval.
The last step converts the log mel spectrum to time domain. The result is known as mel-frequency cepstral coefficient (MFCC).

B. *Vector Quatization*

Vector quantization (VQ) is used for command identification in our system. VQ is a process of mapping vectors of a large vector space to a finite number of regions in that space. Each region is called a cluster and is represented by its center (called a centroid). A collection of all the centroids make up a codebook. The amount of data is significantly less, since the number of centroids is at least ten times smaller than the number of vectors in the original sample.

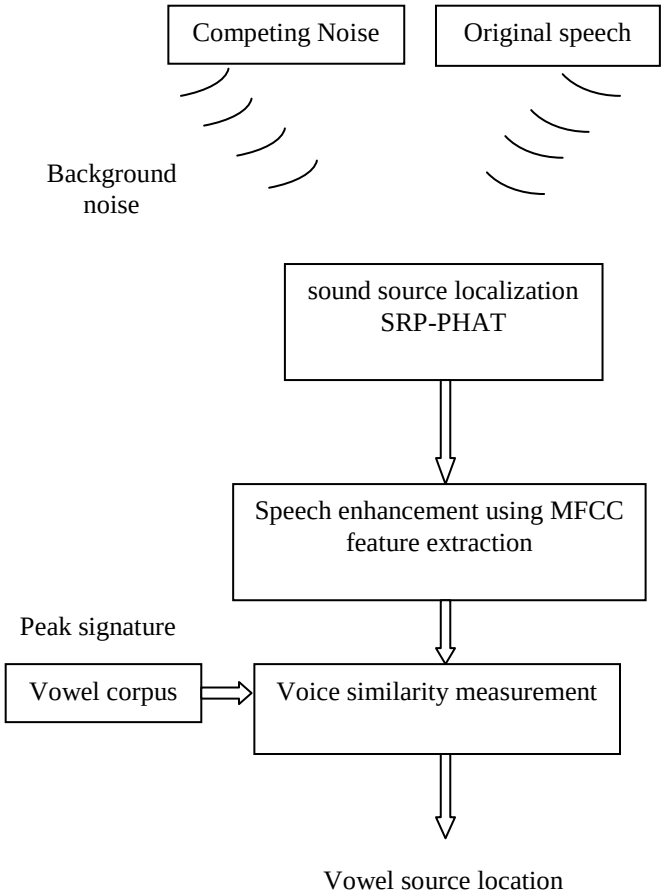

Fig. 2. Functional block diagram of system model.

This will reduce the amount of computations needed when comparing in later stages. Even though the codebook is smaller than the original sample, it still accurately represents command characteristics. The only difference is that there will be some spectral distortion.

C. *Codebook Generation*

Since command recognition depends on the generated codebooks, it is important to select an algorithm that will best represent the original sample. There is a well-known algorithm, namely LBG algorithm [Linde, Buzo and Gray, 1980], for clustering a set of *L* training vectors into a set of *M* codebook vectors (also known as the binary split algorithm) is used. The algorithm is implemented by the following recursive procedure:
1. Design a 1-vector codebook; this is the centroid of the entire set of training vectors (hence, no iteration is required here).
2. Double the size of the codebook by splitting each current codebook according to the rule: where n varies from 1 to the current size of the codebook, and is the splitting parameter.

$$Y_n^+ = Y_n(1 + \varepsilon) \quad (10)$$

3. Nearest-Neighbor Search: for each training vector, find the centroid in the current codebook that is closest (in terms of similarity measurement), and assign that vector to the corresponding cell (associated with the closest centroid). This is done using the K-means iterative algorithm.
4. Centroid Update: update the centroid in each cell using the centroid of the training vectors assigned to that cell.
5. Iteration 1: repeat steps 3 and 4 until the average distance falls below a preset threshold
6. Iteration 2: repeat steps 2, 3, and 4 until a codebook of size is reached.
Intuitively, the LBG algorithm designs an *M*-vector codebook in stages. It starts first by designing a 1-vector codebook, then uses a splitting technique on the codewords to initialize the search for a 2-vector codebook, and continues the splitting process until the desired *M*-vector codebook is obtained.
In the recognition phase the features of unknown command are extracted and represented by a sequence of feature vectors {x1… xn}. Each feature vector in the sequence X is compared with all the stored codewords in codebook and the codeword with the minimum distance from the feature vectors is selected as proposed command. For each codebook a distance measure is computed, and the command with the lowest distance is chosen.

D. *Direction of arrival*

We know that the near field situation, the microphones are situated very close to the acoustic effective source and in the same time for the element of microphone array, there are *M* directions of arrival which is commonly used in DOA (Direction of Arrival). Each and every DOA is the direct path from the microphones point to the acoustic source. Mathematically, we can express the full process by a point on the unit vector expression.

$$\vec{D}_m = \frac{\vec{d_m} - \vec{d_s}}{|\vec{d_m} - \vec{d_s}|} \quad (11)$$

Arjun Das H, Lisha Gopalakrishnan Pillai, and Muruganantham. C, “Human voice localization in noisy environment by SRP-PHAT and MFCC,” *International Research Journal of Advanced Engineering and Science*, Volume 1, Issue 3, pp. 33-37, 2016.

International Research Journal of Advanced Engineering and Science
ISSN: 2455-9024

Where m=1,2,3...M,

In far field condition, we know that the microphones are located far away from the acoustic sources and all microphones in the array design maintained the same Direction of arrival (DOA), which is commonly chosen as the path of the system from the origin of the array design to the active acoustic source. We can express the full process mathematically, where the origin is expressed as O in the coordinate system.

$$\overrightarrow{D}_m = \frac{\overrightarrow{d_o} - \overrightarrow{d_s}}{|\overrightarrow{d_o} - \overrightarrow{d_s}|} \quad (12)$$

In this process, we can express the Direction of Arrival (DOA) through the standard Azimuth Angle and the Elevation Angle. In this angle measurement, we can express it educationally in the following way

$$\overrightarrow{d_o} = \begin{pmatrix} \cos\emptyset\sin\theta \\ \cos\emptyset\cos\theta \\ \sin\theta \end{pmatrix} \quad (13)$$

In the far filed condition, the distance or the range between the microphone array and the effective acoustic source cannot be determined in the acoustic source localization problems. The Direction of Arrival is the only spatial information about the Source.

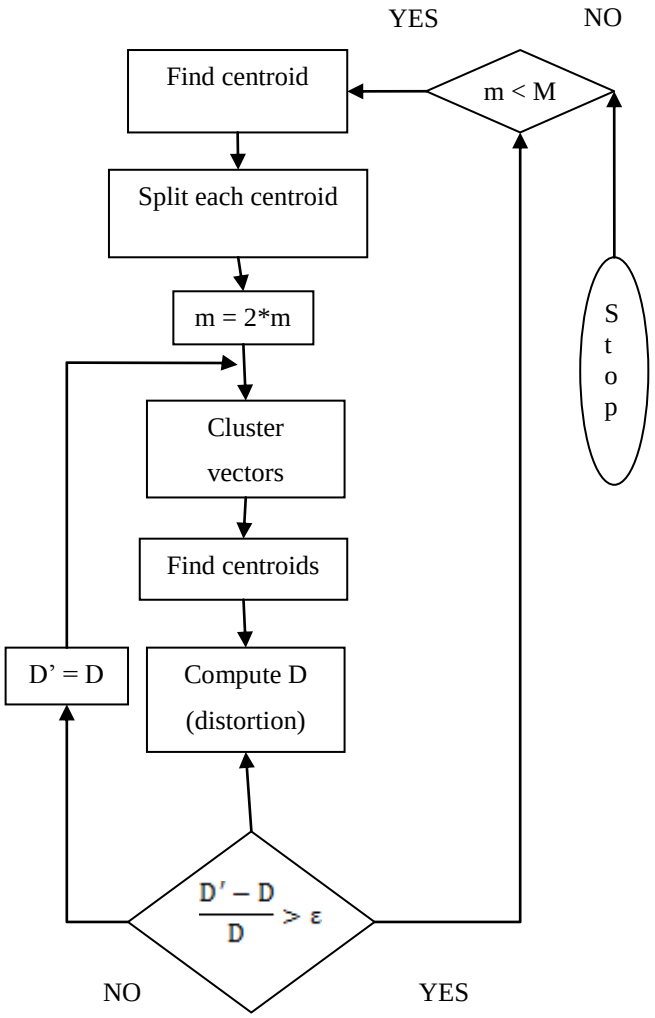

Fig. 3. Flow diagram for LBG algorithm.

IV. Comparison Between SRP-PHAT Based SSL and SRP-PHAT and MFCC Based SSL

The comparison between these two methods is done based on the peak signal to noise ratio (PSNR), otherwise known as PSNR it is the ratio between the maximum possible power of a signal and the power of corrupting noise that affects the fidelity of its representation. PSNR is most easily defined via mean squared error (MSE). Given a noiseless m×n monochrome image I and its noisy approximation K, MSE is defined as,

$$MSE = \frac{1}{mn} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} [I(i,j) - k(i,j)]$$

The PSNR value in DB is defined as:

$$PSNR = 10.\log_{10}\left(\frac{MAXi}{MSE}\right)$$

$$= 20.\log_{10}\left(\frac{MAXi}{\sqrt{MSE}}\right)$$

$$= 20.\log_{10}(MAXi) - 10.\log_{10}(MAXi) \quad (14)$$

TABLE I. PSNR value obtained for both methods at various noisy environment.

PSNR values obtained for SRP-PHAT	PSNR values obtained for enhanced MFCC
60	65
61	67
65	72
68	76
71	80
73	84
77	88
81	92
82	96
88	100

V. Conclusion

Especially under the higher noise conditions, spoken languages in real environments where speech and noise can occur simultaneously. Computation of Steered Response Power Phase Transformation (SRP-PHAT) is important to localize the active sound source. The increase in the accuracy of sound source localization allows location-based interactions between the user and various speech operated devices. The voice similarity of the enhanced signals is computed and combined with the steered response power. The final output is then interpreted as a steered response voice power, and the maximum SRVP location is selected as the speaker location. The computational cost is relatively low because only the top *n* best candidate locations are considered for similarity measurements. The method focuses only on vowel sounds, so a possible extension for the proposed method to improve the accuracy of the voice activity detection involves the use of additional features suitable for consonants. Further extraction and enhancement of the speech signal using MFCC helps in accurately locating the sound source.

Arjun Das H, Lisha Gopalakrishnan Pillai, and Muruganantham. C, "Human voice localization in noisy environment by SRP-PHAT and MFCC," *International Research Journal of Advanced Engineering and Science*, Volume 1, Issue 3, pp. 33-37, 2016.

REFERENCES

[1] Y. Oh, J. Yoon, J. Park, M. Kim, and H. Kim, "A name recognition based call-and-come service for home robots," *IEEE Transactions on Consumer Electronics*, vol. 54, no. 2, pp. 247-253, 2008.

[2] K.Kwak and S.Kim, "Sound source localization with the aid of excitation source information in home robot environments," *IEEE Transactions on Consumer Electronics*, vol. 54, no. 2, pp. 852-856, 2008.

[3] J. Park, G. Jang, J. Kim, and S. Kim, "Acoustic interference cancellation for a voice-driven interface in smart TVs," *IEEE Transactions on Consumer Electronics*, vol. 59, no. 1, pp. 244-249, 2013.

[4] T. Kim, H. Park, S. Hong, and Y. Chung, "Integrated system of face recognition and sound localization for a smart door phone," *IEEE Transactions on Consumer Electronics*, vol. 59, no. 3, pp. 598-603, 2013

[5] Y. Cho, D. Yook, S. Chang, and H. Kim, "Sound source localization for robot auditory systems," *IEEE Transactions on Consumer Electronics*, vol. 55, no. 3, pp. 1663-1668, 2009.

[6] Hyeontaek Lim, In-Chul Yoo, Youngkyu Cho, and Dongsuk Yook,"Speaker Localization in Noisy Environments Using Steered Response Voice Power" *IEEE Transactions on Consumer Electronics*, vol. 61, no. 1, 2015.

[7] A. Sekmen, M. Wikes, and K. Kawamura, "An application of passive human-robot interaction: Human tracking based on attention distraction," *IEEE Transactions on Systems, Man, and Cybernetics - Part A*, vol. 32, no. 2, pp. 248-259, 2002.

[8] X. Li and H. Liu, "Sound source localization for HRI using FOC-based time difference feature and spatial grid matching," *IEEE Transactions on Cybernetics*, vol. 43, no. 4, pp. 1199-1212, 2013.

[9] T. Kim, H. Park, S. Hong, and Y. Chung, "Integrated system of face recognition and sound localization for a smart door phone," *IEEE Transactions on Consumer Electronics*, vol. 59, no. 3, pp. 598-603, 2013.

[10] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-24, no. 4, pp. 320-327, 1976.

[11] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, pp. 276-280, 1986.

[12] B. Mungamuru and P. Aarabi, "Enhanced sound localization," *IEEE Trans. Systems, Man, and Cybernetics - Part B*, vol. 34, no. 3, pp. 1526-1540, 2004.

[13] V. Willert, J. Eggert, J. Adamy, R. Stahl, and E. Korner, "A probabilistic model for binaural sound localization," *IEEE Transactions on Systems, Man, and Cybernetics - Part B*, vol. 36, no. 5, pp. 982-994, 2006.

[14] Y. Cho, "Robust speaker localization using steered response voice power," Ph.D. Dissertation, Korea University, 2011.

[15] J. Sohn, N. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1-3, 1999.

[16] I. Yoo and D. Yook, "Robust voice activity detection using the spectral peaks of vowel sounds," *ETRI Journal*, vol. 31, no. 4, pp. 451-453, 2009.