Speaker Localization in Noisy Environments
Using Steered Response Voice Power

Hyeontaek Lim, In-Chul Yoo, Youngkyu Cho, and Dongsuk Yook, Member, IEEE

Abstract — Many devices, including smart TVs and humanoid robots, can be operated through speech interface. Since a user can interact with such a device at a distance, speech-operated devices must be able to process speech signals from a distance. Although many methods exist to localize speakers via sound source localization, it is very difficult to reliably find the location of a speaker in a noisy environment. In particular, conventional sound source localization methods only find the loudest sound source within a given area, and such a sound source may not necessarily be related to human speech. This can be problematic in real environments where loud noises frequently occur, and the performance of speech-based interfaces for a variety of devices could be negatively impacted as a result. In this paper, a new speaker localization method is proposed. It identifies the location associated with the maximum voice power from all candidate locations. The proposed method is tested under a variety of conditions using both simulation data and real data, and the results indicate that the performance of the proposed method is superior to that of a conventional algorithm for various types of noises.1

Index Terms — sound source localization, speaker localization, human-robot interface.

I. INTRODUCTION

Speech has several benefits when used as communication media, mainly in that it is the basic interface that humans use to communicate with each other and that it does not require additional devices. More importantly, speech can travel over long distances, making it particularly useful for a variety of devices, including humanoid robots and smart TVs, since the user and the device are typically separated by a certain distance [1][2]. For example, a speech-based human-robot interface can provide a natural human-like interface without the need for external devices, such as remote controllers, and users can use familiar speech-based commands to control smart TVs from anywhere in the home. For such speech interfaces to be properly implemented, a method to process distant speech signals should be included [3]. Unlike speech signals detected from a close range, speech signals travelling over longer distances are usually degraded and corrupted by severe unrelated noise. A typical solution to this distant speech problem involves using microphone arrays to both enhance the speech signals coming from the desired direction and to reduce the noise signals coming from other directions. As a result, the quality of the speech signal improves.

However, the location of the speaker must be estimated before improving the speech signal. In addition to speech enhancements that use beamforming, information on the speaker’s location can be used to enable efficient and natural interfaces [4]-[7]. For example, when a user interacts with a humanoid robot, the robot can make use of the user’s location to turn and face him or her or smart doorbells can steer their cameras to focus on the visitor’s face. For most applications, the relative location of the speaker is not known, requiring some method to determine the position of the speaker.

Sound source localization (SSL) is one way to determine the location of the speaker, and this method is effective regardless of lighting conditions, allowing an estimation of the speaker’s location even in the dark. Several methods have been proposed for sound source localization [8]-[13], and steered response power with a phase transform filter (SRP-PHAT) is generally known to be one of the most robust of such methods when the room produces reverberation [12][13]. However, direct use of SRP-PHAT has been shown to negatively impact the performance of real-life speech-based applications. SRP-PHAT steers the microphone array to determine the location of the maximum output power, and the output power of the beamformer is typically measured as the sum of the cross correlation values for each pair of microphone signals. Since SRP-PHAT estimates the power of the voice signal for a given location by using only the cross correlation values of the input signal, a source of noise could be determined to be the maximum output power location if the noise is louder than the voice of the speaker. That is, conventional SRP-PHAT points to the direction of a source of noise if the unrelated noise has higher steered energy, even when the steered energy remains high from the location of the

1 This work was supported by the Korea Research Foundation (KRF) grant funded by the Korean government (MEST) (No. 2011-0002906).

Hyeontaek Lim is with the Speech Information Processing Laboratory, Korea University, 145 Anam-ro, Seongbuk-gu, Seoul, 136-701, Republic of Korea (e-mail: htlim@voice.korea.ac.kr).

In-Chul Yoo is with the Speech Information Processing Laboratory, Department of Computer and Communication Engineering, Korea University, 145 Anam-ro, Seongbuk-gu, Seoul, 136-701, Republic of Korea (e-mail: icyoo@voice.korea.ac.kr).

Youngkyu Cho is with LG Electronics Seocho R&D Campus, 19 Yangjae-daero 11-gil, Seocho-gu, Seoul, 137-130, Republic of Korea (email: youngkyu.cho@lge.com). This work was done when Youngkyu Cho was with Korea University.

Dongsuk Yook is with the Speech Information Processing Laboratory, Department of Computer Science and Engineering, Korea University, 145 Anam-ro, Seongbuk-gu, Seoul, 136-701, Republic of Korea (e-mail: yook@voice.korea.ac.kr).

Contributed Paper
Manuscript received 12/30/14
Current version published 03/30/15
Electronic version published 03/30/15.

0098-3063/15/$20.00 © 2015 IEEE
desired speaker because SRP-PHAT steers to the highest energy point, regardless of the characteristics or content of the sound signal.

When SRP-PHAT is implemented for speaker localization, speech characteristics must be taken into account in order to assign a higher weight to actual speech sources rather than to sources of loud noises [14]. Voice activity detection (VAD), which distinguishes human speech from noise [15][16], can be applied to handle such a problem. This paper proposes a robust speaker localization technique that utilizes VAD. The proposed method uses SRP-PHAT for sound source localization and adopts a VAD scheme to take into account the content of the sound signal rather than just the steered response power of the signal. As a result, the proposed method can compute the steered response voice power (SRVP) for the candidate speaker location. Since the proposed method can identify content within the signal and not just the power of the signal, the location of the voice source can be effectively localized, even under conditions with a 0dB signal-to-noise ratio (SNR). As a result, speech-based interfaces can be effectively localized, even under conditions with a 0dB signal-to-noise ratio (SNR). As a result, speech-based interfaces can be implemented for actual use with a variety of mobile devices, even where unrelated noise might frequently occur.

The rest of this paper is organized as follows. Section II analyzes the problem of conventional SSL using SRP-PHAT and then describes the speaker localization method that computes the SRVP by adopting SRP-PHAT and VAD. The proposed method is then evaluated in Section III. Finally, Section IV concludes the paper.

II. STEERED RESPONSE VOICE POWER

A. SRP-PHAT under Noisy Environments

In the frequency domain, the output $Y_q(\omega)$ of a filter-and-sum beamformer focused on location $q$ is defined as follows:

$$Y_q(\omega) = \sum_{m=1}^{M} G_m(\omega)X_m(\omega)e^{j\omega \tau_{mq}} , \quad (1)$$

where $M$ represents the number of microphones, $X_m(\omega)$ and $G_m(\omega)$ are respectively the Fourier transforms of the $m$-th microphone signal and its associated filter, and $\tau_{mq}$ is the direct time of travel from location $q$ to the $m$-th microphone. The output is obtained by phase-aligning the microphone signals with the steering delays and summing them after the filter is applied.

The sound source localization algorithm based on SRP-PHAT calculates the output power, $P(q)$, of the microphone array focused on location $q$ as follows:

$$P(q) = \int_{-\infty}^{\infty} |Y_q(\omega)|^2 \, d\omega = \int_{-\infty}^{\infty} \left( \sum_{m=1}^{M} G_m(\omega)X_m(\omega)e^{j\omega \tau_{mq}} \right) \left( \sum_{k=1}^{M} G_k(\omega)X_k(\omega)e^{-j\omega \tau_{kq}} \right) \, d\omega = \sum_{m=1}^{M} \sum_{k=1}^{M} \Psi_m(\omega)X_m(\omega)X_k^*(\omega)e^{j(\omega \tau_{mq} - \omega \tau_{kq})} \, d\omega , \quad (2)$$

where $\Psi_m(\omega) = G_m(\omega)G_m^*(\omega) = 1 / |X_m(\omega)X_m^*(\omega)|$. After calculating the steered response power, $P(q)$, for each candidate location, the point, $\hat{q}$, that has the maximum output power is selected as the location of the sound source.

$$\hat{q} = \arg \max_{q} P(q) \quad (3)$$

Although SRP-PHAT is one of the most popular techniques in use for sound source localization, it may not be adequate for speaker localization in noisy environments. Fig. 1 shows an example that is not unusual in many real-world scenarios where noise is louder than the voice. In such a case, SRP-PHAT does not distinguish between the voice and the noise and simply computes the output power of an input signal, so if the noise has greater power, the location of the noise is identified rather than that of the voice. It should be noted that a high level of energy for the desired speaker’s location can be still observed in Fig. 1, while unwanted noise has an even higher steered energy.

B. Steered Response Voice Power

One method that can be used to manage such a problem involves applying VAD values as weights for the SRP-PHAT energy maps. Since the VAD values are high for the speech signals and low for the noise signals, this method can effectively boost the peak from the speech signals while also reducing the peaks from the noise signals. However, the SRP-PHAT algorithm already requires a huge amount of computation, and computing the VAD values for every candidate location significantly increases the computational load. This section presents a robust speaker localization method that can distinguish between the location of voice sources and noise sources with a little additional computational costs.

The proposed method finds a point associated with the maximum voice similarity [14] instead of the maximum

![Fig. 1. An example of the steered response power of a noisy voice signal where the noise is louder than the voice.](image)
output power for SRP-PHAT. It extracts $n$-best candidate locations and applies VAD algorithms to these candidate locations to determine the position where the voice similarity is the highest.

Fig. 2 illustrates the steps of the proposed method. In the first step, the usual SRP-PHAT algorithm is applied to compute the steered response power for every candidate location. The top $n$-best candidate locations are then detected for further computations. A simple smoothing method with a moving average is applied to minimize the effect of the serrated peaks surrounding the main peaks. The value of the output power for each location in the energy map is substituted by the mean of the neighboring output power values. The mean output power $\overline{P}$ is defined as follows:

$$\overline{P}(q_{a,e}) = \frac{1}{(2\theta+1)^2} \sum_{\alpha=-\theta}^{\theta} \sum_{\epsilon=-\theta}^{\theta} P(q_{\alpha,\epsilon}),$$

where $q_{a,e}$ is a point with azimuth $a$ and elevation $e$, and $\theta$ is the number of neighbors that are considered. Fig. 3 shows an energy map smoothed by using (4). This simple smoothing scheme effectively helps identify multiple sources by discarding serrated peaks surrounding the main peaks.

In the second step, the microphone array is focused on the selected $n$-best candidate locations by using an adaptive beamforming method. An adaptive beamformer, such as a generalized sidelobe canceller (GSC) [17][18], boosts the signals from the desired locations and reduces the signals from other locations.

In the third step, the voice similarity of the beamformed signal is evaluated. Since the proposed method targets situations where speech and background noise are captured simultaneously, it is crucial for the VAD algorithm to work reliably under conditions with a mixed signal. If vowel sounds are utilized for the VAD algorithm, it can operate well under these conditions [16]. Human vowel sounds have formants which are distinctive spectral peaks and are likely to remain even after noise has caused severe corruption [19]. However, non-relevant spectral peaks caused by noise corruption are major obstacles to utilizing these spectral peaks in noisy situations. Direct computation of pre-trained spectral peak templates effectively avoids the problem caused by non-relevant spectral peaks [16]. This makes it possible to detect the presence of speech signals even when there is a simultaneous presence of noise. Thus, the characteristic spectral peaks for human vowels are utilized to compute voice similarity, and the training data from several speakers are then used to extract the characteristic spectral peaks [16]. The algorithm does not extract spectral peaks during recognition, but rather directly computes the similarity of the input spectrum to the pre-trained spectral peak signatures. The main idea is that if a spectral peak is present, the average energies of the spectral bands for that peak will be much higher than the average energies of other bands. That is, the peak valley difference (PVD) will be higher. The positions of the spectral peaks are obtained during training and are stored as binary peak signatures consisting of values of ‘1’ for spectral peak bands and ‘0’ for the other bands. During training, similar spectral peak signatures can be clustered to reduce the computational overhead. The PVD is then used as a measure of voice similarity. The similarity of a given binary spectral peak signature $S$ and the beamformed input spectrum $Y_q$ can be calculated as follows [16]:

$$\text{PVD}(Y_q, S) = \frac{\sum_{k=0}^{N-1} [Y_q[k] \times S[k]]}{\sum_{k=0}^{N-1} S[k]} - \frac{\sum_{k=0}^{N-1} [Y_q[k] \times (1 - S[k])]}{\sum_{k=0}^{N-1} (1 - S[k])},$$
where \( N \) is the dimension of the spectrum. The similarity measurement is performed for every registered spectral peak signatures, and the maximum value is determined to be the spectral peak energy of location \( q \) as follows:

\[
PVD(Y_q) = \max_s \text{PVD}(Y_q,S) .
\] (6)

Fig. 4 shows the energy map of the steered response voice power obtained using (6). Fig. 4 clearly shows that the VAD weights effectively boosted the peak from the speaker location while reducing the peak from the noise.

The proposed method selects a point \( \hat{q} \) that is associated with high steered energy and voice similarity by combining the values obtained from SRP-PHAT with those obtained from PVD. Since both values have different ranges, a kind of normalization must be applied.

The location, \( \hat{q} \), of the speaker is determined by using a simple linear combination algorithm as follows:

\[
\hat{q} = \arg\max_q \left( \frac{\overline{P}(q)}{P_{\max}} + a \cdot \frac{\text{PVD}(Y_q)}{\text{PVD}_{\max}} \right) .
\] (7)

where \( P_{\max} \) is the maximum value of the steered mean output power and \( \text{PVD}_{\max} \) is the maximum of the PVD values. Unlike conventional SSL based on SRP-PHAT, the proposed method considers the content of the input signal – that is, the voice similarity and its power – rather than looking for the maximum output energy of an input signal as in (3). Therefore, the location of the speaker can be effectively found even when the interfering noise is louder than the voice. The effectiveness of the proposed method in noisy environments is evaluated in the next section.

III. EXPERIMENTS

A. Simulation Data Experiments

In order to analyze the performance of the proposed method under various noisy environments, noisy speech data was created using the image method [20]. In this paper, a circular microphone array of 25cm radius with eight sensors was used. The proposed method is based on SRP-PHAT, and therefore, the algorithm can be used for various other microphone array configurations as well [12][13]. The microphones were placed with the same intervals around the circular array. The array was located at (250cm, 300cm, 80cm) in a room with dimensions of 600cm \( \times \) 500cm \( \times \) 240cm, and the voice source was located at (250cm, 500cm, 80cm) in the room. Noise sources were placed at the same distance with varying degrees ranging from 0 to 180 degrees at an interval of 10 degrees (except 90 degrees), resulting in 18 different positions. Fig. 5 illustrates this configuration.

The types of noise used include a car, factory, channel, music, subway, train, white, and pink noises. The sampling rate was set to 16kHz and the frame length was 128ms. The performance was measured as the percentage of the estimated speaker locations that lie within \( \pm 5 \) degrees of the true voice source location.

The performance of the speaker localization was analyzed for five different SNR levels (-5, 0, 5, 10 and 15dB). Fig. 6 shows that the proposed method always yielded better performance when compared to that of SRP-PHAT for various levels of SNR. For the 0dB SNR condition, conventional
SRP-PHAT showed only 18.8% accuracy for speaker localization while the proposed method achieved an accuracy of 30.6% (for an absolute error reduction of 11.8%). The proposed method showed a speaker localization accuracy of 49.2% when the SNR was 5dB. When compared to conventional SRP-PHAT, the proposed method achieved an absolute error reduction of 12.6% on the average.

Fig. 7 shows the localization performance for SRP-PHAT and for the proposed method under various types of noise at 0dB. The proposed method can be seen to exhibit better performance when compared to SRP-PHAT for all types of noise. The accuracy of SRP-PHAT was severely degraded in an environment with broad band noise such as white noise.

This can be attributed to the fact that SRP-PHAT calculates the output power using only the phase information over all frequency bands. Most of the gains in performance came from the factory, channel, music, subway, white, and pink noises. For the car and train noise environments, relatively small improvements were made when compared to the other noises given above. It may be that some peak signatures from the car and train noises were very similar to those of some vowel sounds. If so, some low energy noise was boosted, causing a higher SRVP.

Fig. 8 illustrates the speaker localization performance for various noise positions. The result was not symmetrical around 90 degrees because the circular array was rotated slightly. The overall performance of the proposed method can be seen to have increased when compared to that of the SRP-PHAT algorithm. As the angle between the voice and the noise locations increased, the gains in performance became larger. The relatively higher performances for SRP-PHAT between 80 and 100 degrees can be explained by the fact that the noise signal was so close to the voice signals that the sidelobes of the SRP-PHAT from the noise signals also affected the steered energy for the voice signals.

### B. Real Data Experiments

The performance of the proposed method under real use was verified by using actual sound data collected using the robot prototype shown in Fig. 9. The configuration of the microphones, the room dimensions, and the location of the microphone arrays were the same as those for the simulation data. It should be emphasized that as noted in simulation data experiment, the proposed method does not restrict the configuration of microphone arrays to a circular form. Noise sources were placed at 0, 30, and 60 degrees at a distance of 200cm. Eight types of noises with five levels of SNR were used for the experiment as for the simulation data.
Fig. 10 summarizes the performance of SRP-PHAT and of the proposed method for five different levels of SNR over real data. The results shown are similar to those of the simulation data from Fig. 6. For the condition with 0dB SNR, SRP-PHAT showed a localization accuracy of 23.7% while the proposed method showed an accuracy of 42.9%, which is a decrease of 19.2% in absolute error rates. Similar results were obtained for the 5 dB SNR conditions, where the absolute error rates were reduced by 22.9%.

Fig. 11 summarizes the performance for various types of noise over real data. The results are also similar to those of the simulation data, where the increase in performance was larger for a factory, channel, music, subway, white, and pink noises, with a smaller increase for the train noise.

IV. CONCLUSION

This paper proposed a robust speaker localization method that uses the voice similarity of the input signal instead of the simple output power of the beamformer. The proposed method uses SRP-PHAT to find several candidate locations and then uses GSC to enhance the signals coming from the top \( n \) candidate locations. 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 GSC and voice similarity measurements. The experimental results showed that the proposed method significantly outperformed SRP-PHAT, which is a conventional sound source localization method, for very low SNR conditions where the noise signals have equal or higher energies than the voice signals. When compared to the conventional SRP-PHAT method, the proposed method achieved an absolute localization error reduction of 19.3% on average for real data environments with various kinds of noise.

The proposed method can be used for interfaces based on spoken languages in real environments where speech and noise can occur simultaneously. The increase in the accuracy of sound source localization allows location-based interactions between the user and various devices. For example, the camera of a smart doorbell system can be steered. The proposed method can also be used to increase the accuracy of speech-based interfaces. The naturalness and long-distance characteristics of speech can provide for useful interfaces for various devices, including smart TVs and humanoid robots.
REFERENCES


BIOGRAPHIES

Hyeontaek Lim received a B.S. degree in Computer Engineering from Yonsei University, and an M.S. degree in Computer and Communication Engineering from Korea University, Korea, in 2007 and 2010, respectively. He is currently in the Ph.D. program at the Speech Information Processing Laboratory in Korea University. His research interests are speech recognition for mobile devices and parallel speech recognition.

In-Chul Yoo received B.S. and M.S. degrees in computer science from Korea University, Seoul, Korea, in 2006 and 2008, respectively. He is currently pursuing the Ph.D. degree at the Speech Information Processing Laboratory in Korea University. His research interests include robust speech recognition and speaker recognition.

Youngkyu Cho received M.S. and Ph.D. degrees in computer science and engineering from Korea University, Korea, in 2002 and 2011, respectively. Currently, he is working for LG Electronics. His current research interests are acoustic modeling, speaker recognition, and sound source localization using a microphone array.

Dongsuk Yook (M’02) received B.S. and M.S. degrees in computer science from Korea University, Seoul, Korea, in 1990 and 1993, respectively, and a Ph.D. degree in computer science from Rutgers University, New Jersey, U.S., in 1999. He worked on speech recognition at IBM T.J. Watson Research Center, New York, USA, from 1999 to 2001. Currently, he is a professor in the Department of Computer Science and Engineering, Korea University, Seoul, Korea. His research interests include machine learning and speech processing.