REAL TIME SPEAKER LOCALIZATION AND DETECTION SYSTEM FOR CAMERA STEERING IN MULTIPARTICIPANT VIDEOCONFERENCING ENVIRONMENTS

Amparo Marti, Maximo Cobos, Jose J. Lopez
Instituto de Telecomunicaciones y Aplicaciones Multimedia (iTEAM), Universidad Politécnica de Valencia

ABSTRACT
A real time speaker localization and detection system for videoconferencing environments is presented. In this system, a recently proposed modified Steered Response Power - Phase Transform (SRP-PHAT) algorithm has been used as the core processing scheme. The new SRP-PHAT functional has been shown to provide robust localization performance in indoor environments without the need for having a very fine spatial grid, thus reducing the computational cost required in a practical implementation. Moreover, it has been demonstrated that the statistical distribution of location estimates when a speaker is active can be successfully used to discriminate between speech and non-speech frames by using a criterion of peakedness. As a result, talking participants can be detected and located with significant accuracy following a common processing framework.

Index Terms— SRP-PHAT, source localization, speaker detection, microphone arrays

1. INTRODUCTION
Many applications, ranging from teleconferencing systems to artificial perception, hands-free speech acquisition, digital hearing aids, video-gaming, autonomous robots and remote surveillance require the localization of one or more acoustic sources. Since the boost of new generation videoconferencing environments, there has been growing interest in the development of automatic camera-steering systems using microphone arrays [1],[2]. In this work, we present a microphone array system for camera-steering to be used in a multi-participant videoconferencing environment based on the well-known SRP-PHAT algorithm [3]. The SRP-PHAT method has been shown to be one of the most robust sound source localization approaches operating in noisy and reverberant environments. It is commonly interpreted as a beamforming-based approach that searches for the candidate source position that maximizes the output of a steered delay-and-sum beamformer. However, the computational requirements of the method are large, making its real-time implementation considerably difficult. Since the SRP-PHAT method was proposed, there have been several attempts to reduce the computational cost of the method, such as those presented in [4],[5]. Recently, the authors proposed a new strategy based on a modified SRP-PHAT functional that, instead of evaluating the SRP at discrete positions of a spatial grid, it is accumulated over the Generalized Cross Correlation (GCC) lag space corresponding to the volume surrounding each point of the grid [6]. The benefits of following this approach are twofold. On the one hand, it incorporates additional spatial knowledge at each point for making a better final decision. On the other hand, the proposed modification achieves the same performance as SRP-PHAT with fewer functional evaluations, relaxing the computational demand required for a practical application.

In this paper, we analyze the distribution of location estimates obtained with the modified SRP-PHAT functional with the aim of establishing a speaker detection rule to be used in a videoconferencing environment involving multiple participants. The analysis shows that location estimates follow different distributions when speakers are active, allowing to discriminate between speech and non-speech frames under a common localization framework. Moreover, the distribution of an active speaker remains almost the same for different positions inside the room, which makes easier to select a candidate location following a maximum-likelihood criterion, thus simplifying the camera-steering task.

The paper is structured as follows. Section 2 describes the conventional SRP-PHAT algorithm and our modified functional. Section 3 explains the proposed localization-based approach to speech/non-speech discrimination and speaker detection. Experiments with real-data are discussed in Section 4. Finally, the conclusions of this work are summarized in Section 5.

2. SRP-BASED SOURCE LOCALIZATION
Consider the output from microphone $l$, $m_l(t)$, in an $M$ microphone system. Then, the SRP at the spatial point $\mathbf{x} = [x, y, z]$ for a time frame $n$ of length $T$ is defined as: ...
as
\[ P_n(x) \equiv \int_{-\infty}^{\infty} \left[ \sum_{l=1}^{M} w_l m_l(t - \tau(x,l)) \right]^2 dt, \quad (1) \]

where \( w_l \) is a weight and \( \tau(x,l) \) is the direct time of travel from location \( x \) to microphone \( l \). DiBiase [7] showed that the SRP can be computed by summing the GCCs for all possible pairs of the set of microphones. The GCC for a microphone pair \((k,l)\) is computed as
\[ R_{m_k,m_l}(\tau) = \int_{-\infty}^{\infty} \Phi_{kl}(\omega) M_k(\omega) M^*_l(\omega) e^{j\omega\tau} d\omega, \quad (2) \]

where \( \tau \) is the time lag, \( ^* \) denotes complex conjugation, \( M_k(\omega) \) is the Fourier transform of the microphone signal \( m_k(t) \), and \( \Phi_{kl}(\omega) = W_k(\omega) W^*_l(\omega) \) is a combined weighting function in the frequency domain. The phase transform (PHAT) [8] has been demonstrated to be a very effective GCC weighting for time delay estimation in reverberant environments:
\[ \Phi_{kl}(\omega) = \frac{1}{|M_k(\omega) M^*_l(\omega)|}. \quad (3) \]

Taking into account the symmetries involved in the computation of Eq.(1) and removing some fixed energy terms [7], the part of \( P_n(x) \) that changes with \( x \) is isolated as
\[ P'_n(x) = \sum_{k=1}^{M} \sum_{m_k=1}^{M} R_{m_k,m_l}(\tau_{kl}(x)), \quad (4) \]

where \( \tau_{kl}(x) \) is the inter-microphone time-delay function (IMTDF). This function is very important, since it represents the theoretical direct path delay for the microphone pair \((k,l)\) resulting from a point source located at \( x \). The IMTDF is mathematically expressed as
\[ \tau_{kl}(x) = \frac{|x - x_k| - |x - x_l|}{c}, \quad (5) \]

where \( c \) is the speed of sound, and \( x_k \) and \( x_l \) are the microphone locations.

The SRP-PHAT algorithm consists in evaluating the functional \( P'_n(x) \) on a fine grid \( G \) with the aim of finding the point-source location \( x_s \) that provides the maximum value:
\[ \hat{x}_s = \arg \max_{x \in G} P'_n(x). \quad (6) \]

### 2.1. Modified SRP-PHAT Functional

Recently, the authors proposed a new strategy where, instead of evaluating the SRP functional at discrete positions of a spatial grid, it is accumulated over the GCC lag space corresponding to the volume surrounding each point of the grid as follows:
\[ P''_n(x) = \sum_{k=1}^{M} \sum_{m_k=1}^{M} \int_{\tau=L_{k11}(x)}^{(n+1)\tau} R_{m_k,m_l}(\tau). \quad (7) \]

The GCC accumulation limits \( L_{k11}(x) \) and \( L_{k12}(x) \) are determined by the gradient of the IMTDF corresponding to each microphone pair, thus, taking into account the spatial distribution of possible TDOAs resulting from a given array geometry, as explained in [6].

### 3. Speaker Detection

In the next subsections, we describe how active speakers are detected in our system, which requires a previous discrimination between speech and non-speech frames based on the distribution of location estimates.

#### 3.1. Distribution of Location Estimates

Our first step to speaker detection is to analyze the distribution of the location estimates \( \hat{x}_s \), when there is an active speaker talking inside the room from a static position. In this context, six microphones were placed on the walls of the video-conferencing room and a set of 12 recordings from different speaker positions were analyzed to obtain the resulting location estimates. Figure 1(a) shows an example of three two-dimensional histograms obtained from different speaker locations. It can be observed that, since the localization algorithm is very robust, the resulting distributions when speakers are active are significantly peaky. Also, notice that the shape of the distribution is very similar in all cases but centered in the actual speaker location. As a result, we model the distribution of estimates as a bivariate Laplacian as follows:
\[ p(\hat{x}_s|H_s(x_s)) = \frac{1}{2\sigma_x\sigma_y} \exp\left(-\sqrt{2\left(\frac{|x-x_s|}{\sigma_x} + \frac{|y-y_s|}{\sigma_y}\right)}\right), \quad (8) \]

where \( p(\hat{x}_s|H_s(x_s)) \) is the conditional probability density function (pdf) of the location estimates under the hypothesis \( H_s(x_s) \) that there is an active speaker located at \( x_s = [x_s, y_s] \). Note that the variances \( \sigma^2_x \) and \( \sigma^2_y \) may depend on the specific microphone set-up and the selected processing parameters. This dependence will be addressed in future works. On the other hand, a similar analysis was performed to study how the distribution changes when there are not active speakers, i.e. only noise frames are being processed. The resulting histogram can be observed in Figure 1(b), where it becomes apparent that the peakedness of this distribution is not as significant as the one obtained when there is an active source. Taking this into account, the distribution of non-speech frames is modeled as a bivariate Gaussian:
\[ p(\hat{x}_s|H_n) = \frac{1}{2\pi\sigma_x\sigma_y} \exp\left(-\frac{x^2}{2\sigma_x^2} + \frac{y^2}{2\sigma_y^2}\right), \quad (9) \]

where \( p(\hat{x}_s|H_n) \) is the conditional pdf of the location estimates under the hypothesis \( H_n \) that there are not active speakers, and the variances \( \sigma^2_x \) and \( \sigma^2_y \) are those obtained with noise-only frames.
The suitability of the proposed models has been tested by a fitting procedure based on trust region optimization, having a R-square parameter above 0.95 in both cases.

3.2. Speech/Non-Speech Discrimination

In the last subsection, it has been shown that speech frames are characterized by a bivariate Laplacian probability density function. A similar analysis of location estimates when there are not active speakers results in a more Gaussian-like distribution, which is characterized by a shape less peaky than a Laplacian distribution. This property is used in our system to discriminate between speech and non-speech frames by observing the peakedness of a set of accumulated estimates:

$$C = \begin{bmatrix}
\hat{x}_s(n) & \hat{y}_s(n) \\
\hat{x}_s(n-1) & \hat{y}_s(n-1) \\
\vdots & \vdots \\
\hat{x}_s(n-L-1) & \hat{y}_s(n-L-1)
\end{bmatrix} = [c_x, c_y],$$  \hspace{1cm} (10)

where $L$ is the number of the accumulated estimates in matrix $C$. A peakedness criterion based on high-order statistics was evaluated. Since the kurtosis of a normal distribution equals 3, we propose the following discrimination rules for active speech frames:

$$\text{Kurt}(c_x) \begin{cases} 
\geq 3 & \text{speech} \\
< 3 & \text{non-speech}
\end{cases},$$  \hspace{1cm} (11)

$$\text{Kurt}(c_y) \begin{cases} 
\geq 3 & \text{speech} \\
< 3 & \text{non-speech}
\end{cases},$$  \hspace{1cm} (12)

where a frame is selected as speech if any of the above conditions is fulfilled.

3.3. Camera Steering

To provide a suitable camera stability, a set of target positions were pre-defined coinciding with the actual seats in the videoconferencing room. The localization system will be responsible for communicating the camera which of the target positions is currently active. This process involves two main steps. First, it is necessary to discriminate between speech and non-speech frames as explained in Section 3.2. If a burst of speech frames is detected, then the estimated target position is forwarded to the camera when it does not match the current target seat. Since all the target positions are assumed to have the same prior probability, a maximum-likelihood criterion is followed:

$$\hat{x}_t = \arg \max_{x_t} p(\hat{x}_t | H(x_t)), \hspace{0.5cm} t = 1 \ldots N_t,$$  \hspace{1cm} (13)

where $x_t$ is one of the $N_t$ pre-defined target positions. Given that the likelihoods have the same distribution centered at different locations, the estimated target position $\hat{x}_t$ is the one which is closest to the estimated location $\hat{x}_t$.

4. EXPERIMENTS

To evaluate the performance of our proposed approach a set of recordings was carried out in a videoconferencing test room with dimensions 6.67 m x 5.76 m x 2.10 m. A set of 6 omnidirectional microphones were placed on the walls of the room. To be precise, 4 of the microphones were situated at the 4 corners of the ceiling of the room and the other two microphones were placed at the same height but in the middle of the longest walls. Figure 2 shows the microphone set-up, the camera location and the different seats occupied by the participants. Black dots represent the 12 pre-defined target locations used to select the active speaker seat.

The experiment consisted in recording speakers talking from the different target positions (only one speaker at each time) with the corresponding space of silence between two talking interventions. The recordings were processed with the aim of evaluating the performance of our system in discriminating speech from non-speech frames and determining the active speaker so that the camera can point at the correct seat. With this aim, the original recordings were manually labeled as speech and non-speech fragments. The processing used a sampling rate of 44.1 kHz, with time windows of 2048 samples and 50% overlap. The location estimates were calculated using the modified SRP-PHAT functional, as explained.

Fig. 1. Two-dimensional histograms showing the distribution of location estimates. (a) Distribution obtained for three different speaker locations. (b) Distribution for non-speech frames.
in Section 2. The discrimination between speech and non-speech frames was carried out by calculating the kurtosis of the last \( L \) estimated positions, as explained in Section 3.2.

4.1. Results

Table 1 shows the percentage of correctly detected speech (% SP) and non-speech (% N-SP) frames with different number of accumulated positions \( L = 5, 10, 15, 20 \). Moreover, the processing was performed considering two different spatial grid sizes (0.3 m and 0.5 m). The percentage of speech frames with correct target positions (% T) is also shown in the table. It can be observed that, generally, the performance increases with a finer grid and with the number of accumulated estimates \( L \). These results were expectable, since the involved statistics are better estimated with a higher number of location samples. Although it may seem that there are a significant number of speech frames that are not correctly discriminated, it should be noticed that this is not a problem for the correct driving of the camera, since most of them are isolated frames inside speech fragments that do not make the camera change its pointing target.

5. CONCLUSION

This paper presented a microphone array system for camera-steering to be used in a multiparticipant videoconferencing environment based on the well-known SRP-PHAT algorithm. The distribution of location estimates obtained with a modified SRP-PHAT functional was analyzed, showing that location estimates follow different distributions when speakers are active and allowing to discriminate between speech and non-speech frames under a common localization framework. The results of experiments conducted in a real room suggest that, using a moderately high number of accumulated location estimates, it is possible to discriminate with significant accuracy between speech and non-speech frames, which is sufficient to correctly detect an active speaker and make the camera point at his/her pre-defined location.

6. REFERENCES