Concurrent Speaker Localization using Multi-band Position-Pitch (M-PoPi) Algorithm with Spectro-Temporal Pre-Processing

Tania Habib and Harald Romsdorfer

Signal Processing and Speech Communication Lab
Graz University of Technology, Austria
{tania.habib, romsdorfer}@tugraz.at

Abstract

Accurate, microphone-based speaker localization in real-world environments, like office spaces or meeting rooms, must be able to track a single speaker and multiple concurrent speakers in the presence of reverberations and background noise. Our Multi-band Joint Position-Pitch (M-PoPi) algorithm for circular microphone arrays already shows a frame-wise localization estimation score of about 95% for tracking a single speaker in a noisy, reverberant setting. In this paper, we present two extensions of the M-PoPi algorithm to improve the localization estimation accuracy also for multiple concurrent speakers. These extensions are a weighted spectro-temporal fragment analysis as a pre-processing step for the M-PoPi algorithm and a particle filter-based tracking as a post-processing step. Experiments using real-world recordings of two concurrent speakers in a typically reverberant meeting room show an improvement of the frame-wise localization estimation score from 43% using the plain M-PoPi algorithm to 66% using the M-PoPi algorithm with both extensions.

Index Terms: speaker localization, speech processing, array signal processing, direction-of-arrival

1. Introduction

Hands-free communication has gained much attention in recent years. We consider here an acoustic scene, where several human speakers should be captured by a microphone array so that their voices are transmitted or processed by a speech recognition system. In order to track an active speaker in the room, the multiple inputs of the microphone array can be manipulated to enhance or attenuate signals emanating from a particular direction. Moreover, this enhancement is based purely on the knowledge of the source location.

Various approaches exist in literature to localize active sources in an acoustic scene [1]. The most commonly used approach is based on the Time-Difference-of-Arrival (TDoA) method, which is a two-step procedure. In the first step, one or several time delays between different pairs of microphones (i.e., the TDoAs) are estimated. The source position is determined in the second step using the array geometry and estimated TDoAs. Well-known methods in this category are Generalized Cross-Correlation (GCC) and variants [2, 3]. Other methods use frequency-averaged signal power of a Steering Beamformer (SB), where a steered response is generated by steering the beamformer over a predefined spatial region. A method combining both features of SB with the ones used for PHAse Transform (PHAT) weighting of the GCC is known as SRP-PHAT [4]. These methods generally achieve robust performance in real-world acoustic conditions. However, they tend to give poor results for multi-speaker scenarios [5].

A joint Position-Pitch (PoPi) estimation method for speaker localization has been recently proposed [6]. It is based on the observation that the cross-correlation of a stereophonic audio signal encodes both the pitch and the position of the active speaker in an auditory scene. The performance of PoPi algorithm degrades in multi-source environments with strong reverberation. This led to the formulation of the Multi-band Position-Pitch (M-PoPi) algorithm [7], where a pre-processing block inspired from the auditory model [8] is used. The method improved the localization of two concurrent speakers, but it was restricted to vowel utterances [5]. For real concurrent speech signals, summation of cross-correlations across all frequency bands results in inaccurate location estimates for M-PoPi algorithm.

Recent studies on data recorded with binaural mannequin have demonstrated that grouping of location cues over the spectro-temporal regions of speech signal yields robust location estimates [9, 10]. These regions also called as spectro-temporal fragments are generated using pitch information. Further extension of the fragment based system is carried out by weighting the elements in a fragment to improve location estimates. As not all elements are equally affected by reverberation.

In this paper, we show that the improvements made with spectro-temporal fragment approach on data recorded with binaural mannequin can also be made with multi-channel microphone array with omni-directional microphones. Concretely, we have extended our M-PoPi algorithm with an unweighted and weighted spectro-temporal fragment based pre-processing. Section 3 summarizes the details. Furthermore, we have added a particle filter based post-processing stage to the M-PoPi algorithm. In Section 6, we compared all possible combinations of the proposed extension with the M-PoPi algorithm in real-world experiments. The results show that combination of weighted spectro-temporal fragment pre-processing with particle filter based post-processing has the best location estimate performance.

The rest of the paper is structured as follows: Section 2 gives an introduction of the M-PoPi algorithm; followed by the description of spectro-temporal fragment generation in Section 3; the particle filter based tracking is presented in Section 4; Section 5 provides experimental framework details followed by results and analysis in Section 6. Finally, Section 7 draws some conclusions and outlines future work.
2. The M-PoPi Algorithm

The M-PoPi algorithm is an extension of the standard PoPi algorithm. It pre-processes the audio input signals by a filterbank of 64 overlapping bandpass gammatone filters, with center frequencies spaced uniformly on the equivalent rectangular bandwidth (ERB) scale between 50 Hz and 8000 Hz. For every considered microphone pair, the cross-correlations of the two input signals are calculated after being filtered by corresponding filters of the filterbank. These cross-correlations are normalized and added across all frequency bands to form a summary cross-correlation function. To evaluate the presence of a periodic signal with unknown fundamental frequency \( F_0 \), related to a source position at \( \varphi_0 \), the summary cross-correlation is sampled as:

\[
\rho(\varphi_0, F_0) = \frac{1}{2K+1} \sum_{k=-K}^{K} R(k \cdot L(F_0) + O(\varphi_0)), \quad (1)
\]

with \( L(F_0) = \sum_{f} \) as the number of samples with respect to the fundamental frequency \( F_0 \), and \( O(\varphi_0) = d \cdot \cos(\varphi_0) \cdot F_0 \) as correlation lag corresponding to the Direction-of-Arrival (DoA) with respect to \( \varphi_0 \). \( \hat{F}_0 \) is the sampling frequency, \( c \) is the speed of sound in air and \( d \) is the distance between a pair of microphones. 

2\( K + 1 \) is the number of correlation peaks considered. The term position will be referred as DoA in rest of the paper.

The resulting position-pitch relations can be represented in a plane known as a PoPi plane, that reveals the peaks at locations that corresponds to joint position-pitch estimates of the active sources in an acoustic scene. Fig. 1 presents the resulting PoPi planes for M-PoPi and standard PoPi algorithm in case of a multi-speaker scenario, where two female speakers are active in a single frame.

In practice, the PoPi plane is evaluated only for predefined values of \( L(F_0) \) and \( O(\varphi_0) \), which are pre-calculated for frequencies \( F_0 = [80, \ldots, 400] \) Hz, and DoA candidates \( \varphi_0 = [0^\circ \ldots 180^\circ] \) with a step-size of 1°. In the case of uniform circular array (UCA), the position-pitch are estimated for pairs of oppositely placed microphones and then summed up. In order to cover a 360° view, the 0°-180° response of each oppositely placed microphone pair was first mirrored around its’ axis before the summation.

In case of speech signals, the M-PoPi algorithm performs well for a single speaker with approx. 95% frame-wise correctness score [11], but deteriorates for multiple concurrent speakers under strong reverberant conditions. This happens because the summation of cross-correlations across all frequency bands enhances the dominant speaker with higher energy in comparison to the weaker speaker with relatively low energy. Selecting and grouping of frequency channels by the dominating source over a certain time improves the location estimates for multiple speakers.

3. Spectro-Temporal Fragment Generation

In [12], a monaural based system is designed to extract spectro-temporal regions dominated by a single speaker in presence of multiple concurrent speakers. The method is based on analysis of auto-correlogram, which exhibits dendritic (or tree-like) structure for periodic sounds. The correlogram is obtained by passing the mixed signal through a gammatone filterbank, with similar settings used for M-PoPi algorithm and running a short-term auto-correlation on the half-wave rectified outputs of each filter. Multiple pitch estimates are computed and a rule-based tracker is used to form the pitch tracks. The best matched channels are recruited for that particular pitch track. If more than one pitch track is simultaneously active, the frequency channels matching best to the track are assigned to each pitch track. Different spectro-temporal regions are formed by matching each pitch track with the correlogram peaks presented in every frequency channel. Hence resulting in spectro-temporal regions dominated by a single source.

In this paper, we undergo a study to investigate the significance of integrating the spectro-temporal regions within M-PoPi algorithm for concurrent speaker localization. Unlike in the previous mention study, we employ a microphone array consisting of 24 omni-directional microphones to record concurrent speech. The current setup lacks any directionality or human like characteristics and hence is more vulnerable to surrounding acoustic conditions. We treat the array as a coherent set of sensors, for which the spectro-temporal regions are generated using one reference microphone pair consisting of microphone 1 and 13. The reason behind using this pair was simply based on the notion that it is also used as a reference to localize speakers with respect to the array.

The speech signals from the microphone pair are averaged and processed through fragments generation system, which outputs the fragments formed over both frequency and time. The cross-correlations are computed for every filter output. These cross-correlation profiles are added across the spectro-temporal fragments following the systems proposed in [10] such as:

\[
R(\tau)_{\text{unweighted}} = \frac{1}{L} \cdot \sum_{(f,t) \cdot \ast F_p} CC_{(f,t)}(\tau), \quad (2a)
\]

\[
R(\tau)_{\text{weighted}} = \frac{1}{L} \cdot \sum_{(f,t) \cdot \ast F_p} \psi(\tau) \cdot CC_{(f,t)}(\tau), \quad (2b)
\]

where \( F_p \) is a speech fragment containing cells \( (f,t) \), and \( L \) is the total number of spectro-temporal cells in a fragment. \( CC_{(f,t)}(\tau) \) is the cross-correlation profile for a frequency channel \( f \) and frame \( t \) over a range of time-lags, \( \tau \in \{ \tau_{\text{min}}, \ldots, \tau_{\text{max}} \} \). In (2a), an averaged cross-correlation is computed for a speech fragment, where every cell in the fragment is given equal weight. Whereas in (2b), the summation is carried out by assigning different weights \( \psi(\tau) \) based on the interaural coherence (IC) weighting criterion [13] given as:

\[
\psi(\tau) = \max CC_{(f,t)}(\tau). \quad (3)
\]

The unweighted mean cross-correlation from (2a) is decomposed according to (1), which is termed as M-PoPi_{unweighted}.
algorithm. Whereas the PoPi decomposition using the weighted mean cross-correlation from (2b) is termed as M-PoPi\textsubscript{unweighted} algorithm. The process is repeated for both algorithms for every microphone pair. The resulting planes are later summed for all microphone pairs to generate the final location estimates.

4. Tracking using Particle Filters

Tracking multiple concurrent sources using particle filters is widely used in practice, because they can deal with non-linear, non-Gaussian multimodal problems [14]. The task at hand is to track active sources with source-state defined as \( \alpha_t = [\hat{\phi}_1, \hat{\phi}_2, \cdots, \hat{\phi}_K, T_K] \), where \( \hat{\phi}_k \) is the DoA for source \( k \) and \( T_K \) is the total number of sources active at the current time-step \( t \). We combine both the standard and spectro-temporal integration based M-PoPi algorithms with particle filters. The parameter selection such as; the dynamic model, the localization function and the likelihood function with the main steps of particle filters are summarized as follows:

1. **Initialization of Particle Filters**: The particle filters are randomly distributed in the state-space, \( \alpha_0 \) with associated uniform weights \( w_0 = 1/P, i = 1 : P \). In our experiments, we have chosen \( P = 100 \).

2. **Dynamical Model**: The new set of particles are predicted according to Langevin dynamics model with similar settings used in [14].

3. **Localization Function**: To transform the raw data received at the sensors into localization measurements, we use the output of every M-PoPi algorithm maximized along the DoA dimension denoted by \( y_t \), the measurements at time-step \( t \).

4. **Likelihood Function**: For each M-PoPi algorithm, a pseudo-likelihood function \( F(y_t, \alpha) \) based on the formulation of [14] is derived as:

\[
F(y_t, \alpha) = \max \{ y_t(\hat{\phi}_\alpha), \xi_0 \}^\star,
\]

where \( \hat{\phi}_\alpha \) is the localization parameter corresponding to the state, \( \xi_0 \geq 0 \), and \( r \in \mathbb{R}^+ \). The use of \( r \) and \( \xi_0 \) as explained in [14] is to help shape the localization function and to make it more amenable to recursive estimation. The likelihood function used to assign new weights to the particle filters is then calculated as, \( p(y_t|\alpha^{(i)}) = F(y_t, \alpha^{(i)}) \). In our experiments, we used the values of \( \xi_0 = 0 \) and \( r = 2 \). The new weights corresponding to particles are assigned as:

\[
w_{t}^{(i)} = p(y_t|\alpha^{(i)}),
\]

and normalized to obtain \( \sum_{i=1}^{P} w_{t}^{(i)} = 1 \).

5. **Resampling**: Resample the particles by multiplying the particles with higher weights and deleting the ones with smaller weights to avoid the degeneracy problem using a suitable resampling method. Systematic resampling is used in our scheme and weights are reset to uniform values.

6. **Location Estimation**: The final estimates of the source locations are calculated by clustering the particles set, or using a histogram measure with a carefully selected threshold.

<table>
<thead>
<tr>
<th>( \psi_0 )</th>
<th>( 335^\circ - 38^\circ )</th>
<th>( 221^\circ - 155^\circ )</th>
</tr>
</thead>
</table>

Table 1: Azimuths of 2 concurrent speaker setups.

5. Experimental Framework

The performance of the algorithms is evaluated on data recorded using Yamaha MSP5A loudspeakers and a 24 channel UCA with an inner diameter of 55 cm in the SPSC meeting room measuring 6.02 × 5.32 × 3 m, with reverberation time of \( RT_{60} = 650 \) ms. One of the walls of the room has a large window partly covered by blinds that were set open during the recordings. The floor is covered with standard carpet. No particular effort was made to reduce the reverberations in the room.

The array has been designed with 24 Behringer ECM8000 omni-directional microphones connected to an M-Audio Firewire Audiophile Mobile Recording Interface under control of a laptop computer. We used a subset of Keele [15] database for recordings. A set of speech files containing male and female utterances were mixed into longer segments modeling different speaker interaction behaviors in a spatialized multi-speaker scenarios. The array was placed in the center of the room; the loudspeakers were positioned at a height of 1.39 m maintaining a constant distance of approx. 2 m from the array.

Different loudspeaker positions were used for male-female and female-female setups, which are presented in Tab. 1. The playback and recording process was controlled by software on a single laptop and the captured audio was saved directly to the hard disk of the laptop with 16 bit resolution and sampling rate of 48 kHz. We used a frame level metric, denoted as Acc [10], to evaluate the algorithms given as:

\[
Acc = \frac{1}{N} \sum_{n=1}^{N} \delta^*(\hat{\phi}_0, \hat{\phi}_n) \times 100\% ,
\]

where \( N \) is the number of frames, \( \delta^* \) is defined as

\[
\delta^*(a, b) = \begin{cases} 1 & \text{if } |a - b| \leq \Delta \\ 0 & \text{otherwise} \end{cases}
\]

where \( \Delta \) is a grace boundary around the true angle. For the results shown in Fig. 2 and in Fig. 3, we varied \( \Delta \) from 1° to 360°. For the discussion, we consider a \( \Delta \) of 5°, as this corresponds to a minimal 35 cm inter-speaker distance.

6. Results and Analysis

To compare the M-PoPi algorithm with the proposed methods, a frame-length of 42.7 msec with a frame-shift of 20.8 msec is used on 10 sec long recordings. The results of all algorithms for the male-female and the female-female case are outlined as frame correctness scores averaged over both speakers in Fig. 2 and Fig. 3, respectively.

The average frame correctness of the M-PoPi algorithm in the male-female scenario is just around 35%. In this scenario, the strong dominance of the male voice over the female voice results in a poor frame correctness score. The inclusion of unweighted spectro-temporal fragment generation with the M-PoPi algorithm (M-PoPi\textsubscript{unweighted}) improves localization accuracy in this scenario by absolute 18.2%. The additional weighting of the fragment cells (M-PoPi\textsubscript{weighted}) further increases localization accuracy to 66.1%.

In the female-female scenario, the M-PoPi algorithm has a frame correctness score of about 51%. M-PoPi\textsubscript{unweighted} results in an absolute improvement of 8.1%, M-PoPi\textsubscript{weighted} further improves the results to about 66%. Unlike the male-female
scenario, a relative difference of 10 dB Signal-to-Signal-Ratio (SSR) exists between the two speakers. This SSR makes the localization of both speakers more challenging and a relatively smaller but still significant improvement in localization accuracy is observed.

The localization accuracy of all three algorithms are improved by including particle filters, as shown in Fig. 2 and in Fig. 3. However, these improvements are more visible for the standard M-PoPi algorithm. To correctly compare the improvements of the particle filter, the parameters of the particle filtering method are not tuned to any particular algorithm, but are kept constant for all methods. It can be observed from the correctness scores that the fragment based systems would require more tuning and adjustment of parameters. On average, by using a particle filter based tracking, an absolute improvement of around 7% is obtained.

From these significant improvements in both scenarios, we observe that the process of generating spectro-temporal representations of speech mixtures into sets of coherent fragments carry more information regarding spatial cues for strong as well as for weak sources than summing the cross-correlation across all frequency channels. We also observe that in both scenarios the frame correctness score does not total to 100% due to some missing frame estimates during speech-silence transitions.

7. Concluding Remarks

In this paper, we extend the M-PoPi algorithm with spectro-temporal fragment-based analysis to improve localization accuracy for concurrent speakers. The proposed methods have been incorporated into a particle filtering framework, and tested on multi-channel concurrent speaker recordings using a circular microphone array in a meeting room with reverberation and background noise. The experiments show that the addition of the spectro-temporal fragment-based analysis eliminates deteriorations due to dominant speakers independent of an addition of a particle filter-based post-processing stage. The particle filter-based post-processing itself improves speaker localization correctness by additional 7% on average. These improvements results partly from assigning pitch value to the position of the sources and partly from introducing a spectro-temporal integration stage with application of multiple sensor pairs. Future work will focus on detailed analysis of particle filtering framework for both the M-PoPiunweighted and M-PoPiweighted algorithms using mobile concurrent sources.

8. References


