Evaluation of Joint Position-Pitch Estimation Algorithm for Localising Multiple Speakers in Adverse Acoustical Environments

Stephan Gerlach[1], Stefan Goetze[1], Jörg Bitzer[1,2] and Simon Doclo[1,3]

[1] *Fraunhofer Institute for Digital Media Technology (IDMT), Project group Hearing Speech and Audio Technology (HSA), 26129 Oldenburg, Germany, Email: {stephan.gerlach, s.goetze}@idmt.fraunhofer.de*

[2] *Jade University of Applied Sciences, 26129 Oldenburg, Germany, Email: joerg.bitzer@jade-hs.de*

[3] *University of Oldenburg, 26129 Oldenburg, Germany, Email: simon.doclo@uni-oldenburg.de*

Abstract

Automatic speaker localisation, detection and tracking are important challenges in multi-channel hands-free communication systems. In particular, simultaneous localisation of different speakers is of great interest for multi-microphone noise reduction schemes. Besides position, another possible feature to distinguish between different speakers is the fundamental frequency (pitch) of the speakers' voices. The recently proposed Position-Pitch (PoPi) estimation algorithm combines speaker localisation based on well-known cross-correlation approaches with pitch estimation techniques. In this contribution we evaluate the robustness of a modified version of the PoPi algorithm for localising simultaneous speakers in a realistic environment including room reverberation and different signal-to-noise ratios (SNR). In order to improve robustness, we particularly focus on modifications of the frequency-domain phase transformation $\mathrm{T}\{\cdot\}$ used by the original PoPi algorithm.

Joint position-pitch estimation

Methods for speaker localisation as in [5] use a two step approach to combine the estimate of pitch f_0 and localisation, where in a first stage a pitch estimation algorithm is applied and in an second stage the direction of arrival (DoA) φ_0 is determined. The approach used here automatically estimates pitch and position in one step using the so-called Popi plane $\rho(\varphi, f_0)$, i.e.,

$$\rho(\varphi, f_0) = \sum_{n_p} |\phi[n_p]| \cdot \mathrm{T}\left\{\psi[n_p] - \psi_0[n_p]\right\}, \qquad (1)$$

$$\psi_0[n_p] = p \cdot 2\pi \frac{d_{il}\cos(\varphi) f_0}{c}. \qquad (2)$$

To calculate $\rho(\varphi, f_0)$ the cross power spectral density (CPSD) $\phi[n]$ between microphone i and l is used, where the absolute value of the CPSD $|\phi[n]|$ contains pitch information due to harmonic multiples and the phase ψ of the CPSD contains DoA information. The phase $\psi_0[n_p]$ is the expected phase for a considered DoA at harmonic multiples of f_0. The CPSD is resampled via a parametric comb filtering at discrete frequency bins $n_p = p \cdot n(f_0)$ to consider all possible pitch and DoA combinations [3]. Harmonic sources with pitch f_0 arriving from direction φ_0 are represented by high peaks in the PoPi plane. To enhance the basic algorithm in robustness for noise and reverberation as well as for multi-speaker situations preprocessing methods, such as a gammatone filterbank or cepstrum weightning, were added [1, 2].

Phase transformation

To enhance the steering pattern in the Popi plane different phase-transformations can be introduced, whose goal is to emphasize estimations originating from real source measurements. These transformations modify the impact of the phase weighting in the parametric comb filtering towards a focused beam. In this contribution, we evaluate the impact of 3 different phase transformations, which all have in common that they are real valued, even and 2π periodic functions [3], i.e.,

$$\mathrm{T}_1\{x\} = \cos(x), \qquad (3)$$

$$\mathrm{T}_2\{x\} = \frac{1}{1+\beta-\cos(x)}. \qquad (4)$$

The variable x represents the difference between the measured phase ψ and phase ψ_0 derived from the parametric comb filtering representing a DoA of interest. For $x \to 0$, or a multiple of 2π the transformations result in a significant weight, representing a true source. The transformations $\mathrm{T}_1\{\cdot\}$ and $\mathrm{T}_2\{\cdot\}$ have been proposed in [3]. Parameter $\beta > 0$ affects the spread of the preferential direction. Figure 1 shows the effect of transformation $\mathrm{T}_2\{\cdot\}$ for several β values.

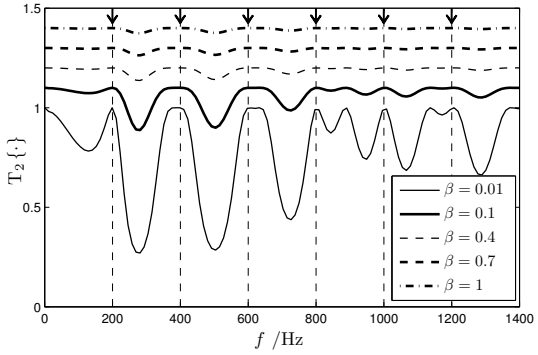

Figure 1: $\mathrm{T}_2\{\cdot\}$ values in the parametric comb-filter for different β, $f_0 = 200\,\mathrm{Hz}$. Results with $\beta > 0.01$ assigend with offset (each step +0.1) for better visual distinction.

In this paper we additionally analyse the transformation

$$\mathrm{T}_3\{x\} = \begin{cases} 1 & \text{, if } (x \bmod 2\pi) < 2\pi\beta \\ 0 & \text{, else} \end{cases} \qquad (5)$$

taking in to account the minimal requirements for a transformation function and achieving a narrower preferential beam.

Experimental setup

In order to evaluate the effects of the different phase transformations and the relevance of parameter β on the performance of the algorithm, several simulations have been made. In the utilized test scenario a line array with six microphones (intermicrophone distance $d_{il} = 22\,\text{cm}$) was used. Recordings of two speakers with voiced vocals (male $f_0 = 127\,\text{Hz}$ and female $f_0 = 175\,\text{Hz}$) served as simultaneous sources. They were convolved with measured room impulse responses [4] from a room of approximate size $l = 4.6\,\text{m} \times w = 5.1\,\text{m} \times h = 2.5\,\text{m}$. The distance between the microphone array and speaker positions was 3.3 m and the distance between the speakers was 4 m, which results in an angular difference of 63°. Reverberation time was $\tau_{60} \approx 550\,\text{ms}$. To achieve the desired SNR, diffuse speech-shaped noise was added to the simulations. Eight SNR values from ∞ to $-10\,\text{dB}$ as well as eight different β values reaching from 0.001 to 1 were evaluated. A block-based processing was implemented with a blocksize of 85ms (50% overlap) at a sampling rate of $f_s = 48\,\text{kHz}$. To track peaks in the PoPi plane evoked by sources, a particle-filter [1] was used and its output served as source estimation. Performance was measured in terms of the accuracy rate (*Acc*) for all blocks with a fault tolerance of $\Delta = \pm 10°$ in angle and $\Delta = \pm 10\,\text{Hz}$ in pitch compared to the real source characteristics.

Results

All three variants showed a very good accuracy up to 10dB SNR. The phase transformation $\mathrm{T}_1\{\cdot\}$ does not include a β parameter, hence the results only depend on the SNR and are shown in Table 1. $\mathrm{T}_1\{\cdot\}$ shows good results up to 5 dB SNR. Results for $\mathrm{T}_2\{\cdot\}$ are depicted

Table 1: DoA performance rate with phase transformation $\mathrm{T}_1\{\cdot\}$ for different SNR values and $\tau_{60} \approx 550\,\text{ms}$. Two simultanous speakers (male/female).

SNR/dB	$\to\infty$	20	15	10	5	0	-5	-10
Acc/% $\mathrm{T}_1\{\cdot\}$	99.5	99.2	97.9	93.8	96.7	56.7	65	35.8

in Figure 2 and show even slightly better results for low SNR (0 dB, −5 dB), if a high β value (≈ 0.8) is choosen. The phase transformation $\mathrm{T}_3\{\cdot\}$ achieves results comparable with the other two approaches (at appropriately high β values) as can be seen in Figure 3.

Conclusion

All three phase transformations were able to achieve high accuracy rates ($> 90\%$) up to 10 dB SNR with the used scenario. The modification of phase transformation is not that essential for the performance of the PoPi algorithm if parameter β is carefully chosen. Transformation $\mathrm{T}_2\{\cdot\}$ and $\mathrm{T}_3\{\cdot\}$ showed good results even for lower SNR values. For SNR below 0 dB the results obtained by $\mathrm{T}_3\{\cdot\}$ were slightly better than for the other transformations.

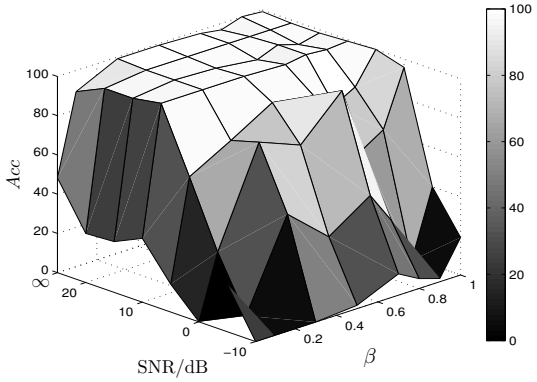

Figure 2: DoA rate with phase transformation $\mathrm{T}_2\{\cdot\}$ for diffrent SNR and β values and ($\tau_{60} \approx 550\,\text{ms}$). Two simultanous speakers (male/female).

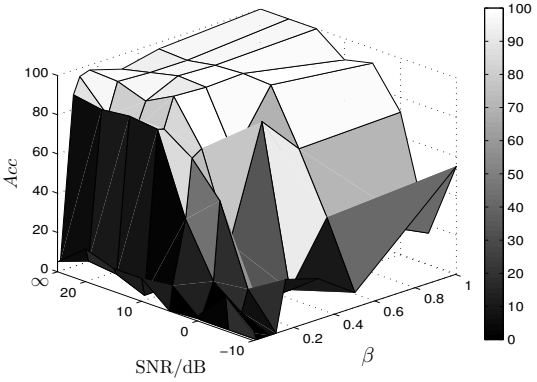

Figure 3: DoA rate with phase transformation $\mathrm{T}_3\{\cdot\}$ for diffrent SNR and β values and ($\tau_{60} \approx 550\,\text{ms}$). Two simultanous speakers (male/female).

Literature

[1] Habib, T. and Romsdorfer, H.: Comparison of SRP-Phat and multiband-PoPi algorithms for speaker localisation using particle filters. In: 13th International Conference on Digital Audio Effects, DAFX, Graz, Austria, Sep. 2010.

[2] Kepesi, M., Ottowitz, L. and Habib, T.: Joint position-pitch estimation for multiple speaker scenarios. In: Hands-Free Speech Communication and Microphone Arrays, HSCMA, pp. 85-88, May 2008.

[3] Wohlmayr, M. and Képesi, M.: Joint Position-Pitch Extraction from Multichannel Audio. In: Interspeech, Antwerp, Belgium, pp. 1629-1632, Aug. 2007.

[4] Habets, E.: Room impulse response generator, June 2010. `http://home.tiscali.nl/ehabets/rir_generator.html`

[5] Christensen, H.; Ma, N.; Wrigley, S. N. and Barker, J.: Integrating pitch and localisation cues at a speech fragment level. In: Interspeech, Antwerp, Belgium, pp. 2769-2772, Aug. 2007.