Joint Position-Pitch Extraction from Multichannel Audio

Michael Wohlmayr, Marián Képesi

Graz University of Technology, Graz, Austria
SPSC Laboratory
michi_w@sbox.tugraz.at, kepesi@tugraz.at

Abstract

Recently, a method for joint extraction of pitch and location information from two-channel recordings has been introduced. This framework offers a new, natural representation of all acoustic sources in the auditory scene, and has potential to be used as front-end in applications such as advanced tracking of multiple speakers in conference rooms. In this paper, we explore basic properties of this method and propose improvements in performance by using circular arrangements of multiple microphones.

Index Terms: Acoustic arrays, pitch estimation, source localization

1. Introduction

Both pitch and location estimation of acoustic sources are highly active fields of research in the signal processing domain since many decades. The term pitch refers here to the 'fundamental frequency of a quasi-periodic structure in the short-time spectrum of a speech signal' [1]. Tracking the pitch of speech signals is an important building block for applications such as coding or compression. Numerous approaches have been proposed to tackle the problem of pitch estimation, however it seems that a perfect solution suitable for all types of applications remains to be found.

While pitch estimation can usually be accomplished with monaural recordings, source location estimation makes use of multiple microphones, so called microphones arrays. A popular approach for estimation of source location is based on a signals time difference of arrival (TDOA) among microphones, which allows to draw conclusions on the source position using geometric considerations. To simplify matters, it is often assumed that the sources of the acoustic scene have sufficient distance to the microphone array such that their incoming waveforms can be modeled to be planar. This so called far field assumption is valid in many cases [2]. Applications of source localization are speaker tracking and multi-party speech segmentation.

It is remarkable that much effort has been investigated in improving both of these methods, however approaches to *combine* these mechanisms to form a new, intuitive representation of the acoustic scene have been outstanding. Recently, we proposed in [3] a novel approach for *joint* estimation of pitch and location from 2-microphone recordings. It is based on the fact the cross-correlation of the two signals available encodes both the pitch and the TDOA simultaneously. A parameterized sampling (reindexing) of the cross-correlation results in a so-called *Position-Pitch* (PoPi) plane which reveals peaks at locations that correspond to joint position-pitch estimates of all sources in the acoustic scene. As only two microphones are employed, the term 'position' refers here to the direction of arrival (DoA) of a source.

In this paper, we explore basic properties of this representation and seek to improve its performance using multiple microphones with circular arrangement. Throughout this work, we neglect the effects of reverberant room conditions, which are known to detoriate the DoA estimate of TDOA based methods [4]. Since reverberation has a large impact on the method performance, this issue is part of our current research.

The paper is organized as follows: Section 2 gives a quick review of the PoPi-method and defines an alternative formulation in terms of the cross-spectrum. Basic properties of the PoPi-plane are illustrated and will be shown to be directly related with what we will refer to as *phase compensation response* of the microphone array. Section 3 will then explore the improvements gained by using multiple microphones, with the goal to sharpen directional selectivity and remove spatial aliasing. Finally, section 4 gives a conclusion.

2. The Position-Pitch Plane

The Position-Pitch plane, which was recently proposed in [3], is a novel representation of the auditory scene, as it provides a joint estimate of pitch and location of all acoustic sources. Having available a two channel recording of this scene, the PoPi method makes use of the fact that both the pitch and the TDOA are encoded in the cross-correlation of both channels. The cross-correlation R, computed on a short time frame κ, is defined as:

$$R_\kappa(\tau) \quad = \quad \sum_{n=-\frac{N}{2}}^{\frac{N}{2}} x_L(\kappa N + n) \cdot x_R(\kappa N + n + \tau) \quad (1)$$

where x_L and x_R correspond to the left and right microphone recordings, N denotes the length of the time frame, n is the discrete time and τ is the discrete time lag. To evaluate the presence of a periodic signal with fundamental frequency f_0, related to a source at direction φ_0, the cross correlation is then sampled according to:

$$\rho_\kappa(\varphi_0, f_0) = \frac{1}{2K+1} \sum_{k=-K}^{K} R(\lfloor kL(f_0) + n_s(\varphi_0) \rfloor) \quad (2)$$

In this formulation, $L(f_0) = \frac{f_s}{f_0}$ denotes the number of samples of the fundamental period of evaluation, f_s is the sampling frequency, $n_s(\varphi_0)$ denotes the correlation lag corresponding to the DoA of interest, and K is the number of correlation peaks considered in each direction of $R(\tau)$. The PoPi plane $\rho_\kappa(\varphi_0, f_0)$ is computed by steering f_0 and φ_0 over a predefined range, and exhibits large peaks at indices that correspond to pitch and DoA of a source present in the acoustic scene. Equivalently, equation 2 can be expressed in terms of the cross spectrum,

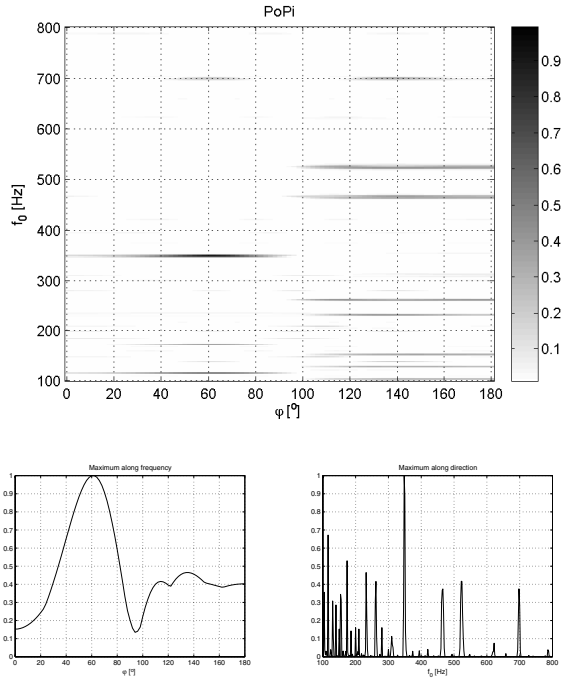

Figure 1: *Top: Normalized PoPi plane for single sound source from direction 60°. The spacing of the two microphone is 0.4 meters. Bottom: Maximum values taken along frequency and DoA, respectively.*

$S_{xy}(e^{j\omega})$, which is the Fourier transform of the cross correlation $R_\kappa(\tau)$. The PoPi plane then evaluates as:

$$\rho_\kappa(\varphi_0, f_0) = \sum_{k=0}^{\lfloor \frac{\pi}{\Omega} \rfloor} |S_{xy}(e^{jk\Omega})| \cdot \cos(\angle S_{xy}(e^{jk\Omega}) + k\Omega n_s) \quad (3)$$

where $\Omega(f_0) = 2\pi f_0 / f_s$. In this formulation, we exploit the fact that the magnitude response, $|S_{xy}(e^{j\omega})|$, holds the harmonic structure of the recorded signals, while the phase response, $\angle S_{xy}(e^{j\omega})$, encodes the delay of signals occurring between two microphones. Again, by steering f_0 and φ_0 over a predefined range, the PoPi plane $\rho_\kappa(\varphi_0, f_0)$ in equation 3 will yield a maximum if

- f_0 is chosen such that $|S_{xy}(e^{j\omega})|$ is sampled at a true pitch frequency $\Omega(f_0)$ and multiples of it. This mechanism is almost identical to the *harmonic sum spectrum* method, a variation of the *harmonic product spectrum* algorithm for pitch estimation, as described in [5].
- φ_0 is chosen such that $n_s(\varphi_0)$ matches the true sample delay n_τ of one source, which is encoded in the phase of the cross spectrum, substantially at frequencies corresponding to high energy harmonics of this source. Only in this case, the cosine will yield a maximum as its argument is zero or any multiple of 2π.

All in all, the magnitude of the cross spectrum is sampled at multiples of the fundamental frequency, and before their summation, each of these samples is weighted by the cosine of the difference between the observed phase $\angle S_{xy}(e^{j\omega})$ and the *steering phase* $\omega\, n_s$. It is remarkable that this mechanism is similar in spirit to [6], a pitch based DoA estimation approach reported to be robust to both environmental noise and multipath propagation. In either case, the algorithm is guided by the fact that all source relevant DoA information will be located at the corresponding harmonics, as these can be assumed to be of considerably higher energy than environmental noise and to be in general well separated from other sources harmonics.

2.1. Phase Compensation Response

The weighting of spectral magnitude samples is of special significance, as it scores the degree of phase matching at corresponding frequencies and is only maximal if both steering and source phase fully compensate. For this reason, we refer to this cosine term in equation 3 as *phase compensation response* H:

$$H(\omega, \varphi_0) = \cos(\omega(n_\tau - n_s(\varphi_0))) \quad (4)$$

Analysis of this phase compensation response H proves to be useful in understanding the behaviour of the PoPi plane and in designing microphone arrays with multiple microphones and arbitrary geometry. To illustrate this, the top of Fig. 1 shows a PoPi plane that results from a simple scene, where a source from direction $\varphi = 60°$ emits a periodic signal with a pitch of 350Hz. As expected, the PoPi plane exhibits a clear peak at this corresponding position. However, the spread of the peak along φ is considerably large, and there exist several other peaks, although all with smaller magnitude. Both phenomena can be explained in terms of the phase compensation response H, which is shown for two different steering directions in Fig. 2 for the case of two microphones. The x-axis corresponds to the direction of the source, the y-axis to frequency (not pitch). The four superimposed rectangles indicate the values that weight the first four samples of the cross-spectrum magnitude in equation 3 when computing the PoPi plane for pitch 350 Hz at two steering directions according to the figure. Obviously, in case of the upper image all weights are large for steering direction 60°such that the corresponding value in the PoPi plane is large too. However, due to the large spread of the mainlobe, these weights have still a considerable magnitude for steering directions 40°and 80°, which is the reason for the large variance of the peak.

Beyond the high spread of the wanted peak, there appear undesired peaks in the PoPi plane. There are mainly two types of such peaks that arise in this example. In Fig. 1, the first type of undesired peaks is located at direction 60°and at frequencies which are rational multiples of the true pitch. Sampling the cross-spectrum at such frequencies will accidentally include true harmonic peaks whose positive contribution to the sum in equation 3 causes peaks in the PoPi plane. However, the location of such peaks relative to the maximal peak is known, and appropriate suppression of such known patterns is feasible. The second type of undesired peaks is located at direction 140°. The reason for these peaks is spatial aliasing, which again can be explained in terms of the phase compensation response H. As shown at the bottom of Fig. 2, the phase compensation response for steering direction 140°exhibits also a large response for some frequencies of sources from direction 60°. For example, when evaluating the Popi-plane for pitch 700Hz and direction 140°, the first 2 samples of the cross-spectrum magnitude will accidentally hit the 2nd (700Hz) and 4th (1400Hz) harmonic of the source spectrum. Then, these samples are weighted by the phase compensation values indicated by the 2nd and the 4th rectangle, respectively (shown in black), thus also resulting in a peak at [700Hz;140°]. Of course, these two terms also contribute to the overall sum when sampling at frequency 350Hz

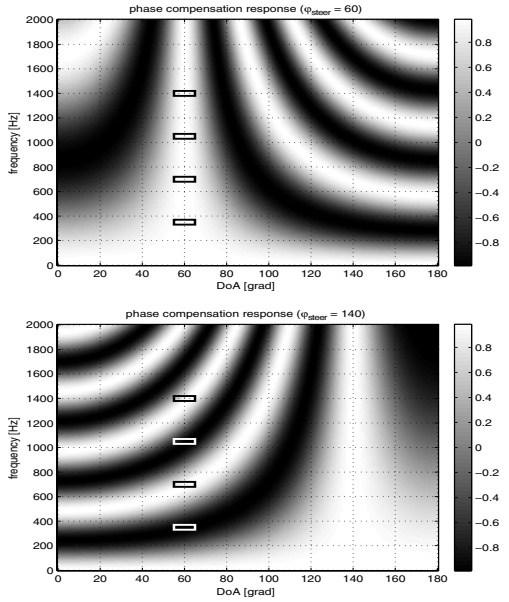

Figure 2: *Visualization of phase compensation response H (equation 4) for two different steering directions (top: 60°, bottom: 140°). The spacing of the two microphones is 0.4 meters. The squares highlight the values used to weight the first four cross-spectrum magnitude samples for the source in direction 60° and the response $\rho(\varphi_0, f_0)$ being computed for pitch 350Hz (equation 3).*

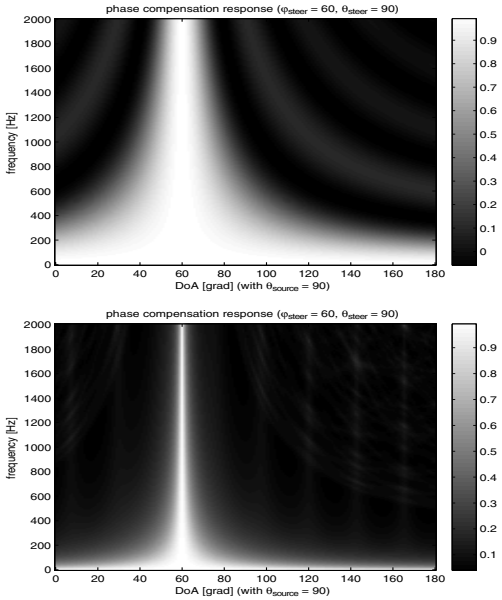

Figure 3: *Top: Phase compensation response H^m. Bottom: Phase compensation response $\tilde{H}^m$ ($\beta = 0.01$). In both cases, the same circular microphone array with 16 microphones and radius 0.2 m is used, and the steering direction is set to 60°.*

and multiples of it. However, in this case the negative contribution of the first and the third sample (shown by white rectangles in bottom of Fig. 2) is even larger, such that all in all there is no peak at [350Hz;140°].

From this illustration can be seen that major properties of the PoPi plane, namely the spread of the direction estimate and spatial aliasing, are a direct consequence of the properties of the phase compensation response H, as defined in equation 4. Ideally, H would emphasize all frequencies in a small vicinity of the steering direction and zero out all other directions. In the following, we will study methods to modify H and thereby bring it closer to the ideal case.

3. Modified Phase Compensation Response

To reduce the effects of spatial aliasing, we make use of M microphones with circular arrangement. Let m_i denote the i-th microphone and let $(\varphi_i, \theta_i, \rho_i)^T$ denote its spherical coordinates in space. In this notation, φ refers to the azimuth angle, θ to the zenith angle and ρ to the euclidean distance from the origin. Using multiple pairs $\{m_i, m_j\}$ of microphones, we construct a new phase compensation response H^m by summing the single responses H of each pair employed. Therefore, the global source direction (φ_s, θ_s) and steering direction (φ_0, θ_0) of the array must be translated to the relative source and steering delays, $n_\tau^{i,j}(\varphi_s, \theta_s)$ and $n_s^{i,j}(\varphi_0, \theta_0)$, that result for each microphone pair $\{m_i, m_j\}$. Then, the overall phase compensation response H^m is given as:

$$H^m(\omega, \varphi_0, \theta_0) \quad = \quad \sum_{i=1}^{M} \sum_{j=i+1}^{M} \cos(\omega(n_\tau^{i,j} - n_s^{i,j})) \quad (5)$$

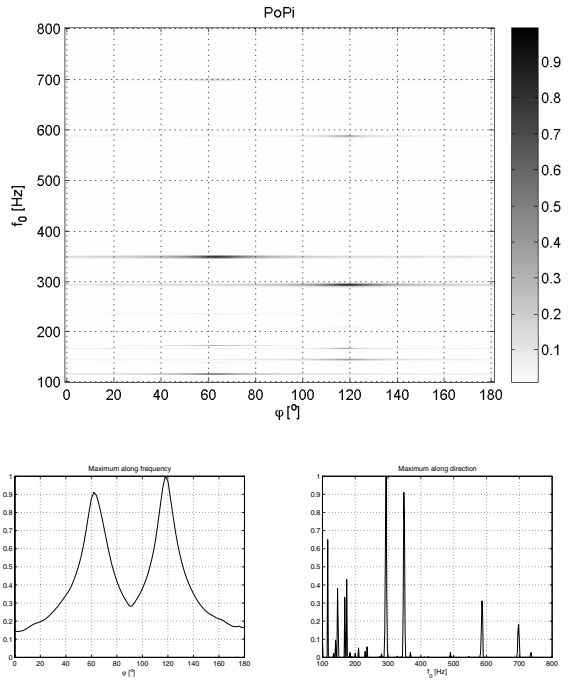

Figure 4: *Top: Modified PoPi plane using a circular array with 16 microphones and radius 0.2 m for two sources from direction 60° and 120°, respectively.*

where, for ease of notation, we skip the dependency of the individual delays on the relative source and steering directions.

The top of Fig. 3 shows H^m that results from a circular array with 16 microphones and radius 0.2 m. All elements of the circular array are arranged on the same height y in space, and for simplicity, we assume here that both the source location and steering direction are restricted to this y = const plane. In contrast to the two-microphone case in Fig. 2, the effects of spatial aliasing are well suppressed due to the usage of 16 microphones. However, the spread of the steering direction still remains unaltered.

For this reason, we further employ a direct modification of the phase compensation response. In its original formulation, the cosine in equation 4 acts as a similarity measure of two phases, namely the cross-spectrum phase introduced by the DoA of one source and the phase it should have for this source coming from a certain steering direction. In principle, any real valued, even and 2π periodic function could be used for this purpose. We suggest to use an alternative measure ψ, defined as

$$\psi(x) \quad = \quad \frac{1}{1+\beta-\cos(x)} \tag{6}$$

where $\beta > 0$ is used to control the spread of ψ, see Fig. 5. We then generalize H in equation 4 by replacing the cosine with ψ:

$$\tilde{H}(\omega, \varphi_0) = \psi(\omega(n_\tau - n_s(\varphi_0))) \tag{7}$$

Clearly, this generalized response $\tilde{H}$ can again be extended by using multiple microphones, just as described above, thus yielding a response denoted by $\tilde{H}^m$. For comparison, the bottom of Fig. 3 shows the resulting modified phase compensation response $\tilde{H}^m$, again for the same 16 microphone circular array as above. Not only the spatial aliasing is suppressed, but also the spread of the steering direction, directly controlled by parameter β, is reduced.

Using this extended response, the PoPi plane then evaluates as:

$$\tilde{\rho}^m_\kappa(\varphi_0, f_0) \quad = \quad \sum_{k=0}^{\lfloor \frac{\pi}{\Omega} \rfloor} \sum_{i=1}^{M} \sum_{j=i+1}^{M} |S_{ij}(e^{jk\Omega})| \psi(\angle S_{ij}(e^{jk\Omega}) + k\Omega n_s^{i,j}) \tag{8}$$

where $S_{ij}(e^{j\omega})$ denotes the cross-spectrum between m_i and m_j.

Fig. 4 shows the PoPi plane $\tilde{\rho}^m_\kappa$ that results for two sources from directions 60°and 120°, respectively. As expected, there are no more peaks due to spatial aliasing, and the spread of both DoA estimates is clearly reduced.

4. Conclusion

In this paper, we investigated approaches to extend the original formulation of the PoPi method. First, by reformulating the intuitive time-domain approach in terms of the cross-spectrum, we see that the method is a natural combination of mechanisms which are similar to algorithms that have earlier been designed separately for pitch and DoA estimation. A quantity has been introduced which we refer to as the *phase compensation response* H. Based on this descriptor, basic properties of the PoPi plane such as DoA estimation spread and spatial aliasing can be explained. Direct modification of H and further extension by using multiple pairs of microphones can increase the performance of the method. In general, the resulting modified phase compensation response helps in exploring PoPi-relevant properties of diverse geometric arrangements of microphones in space.

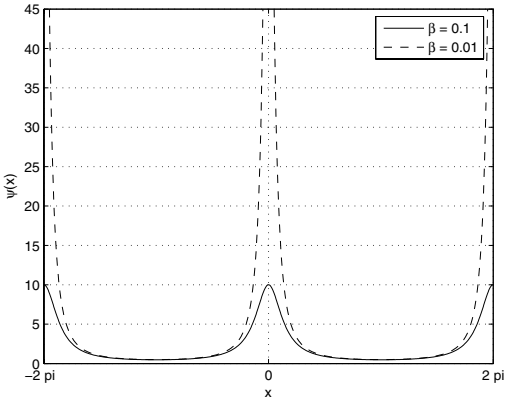

Figure 5: *An alternative function to measure the grade of phase compensation. The variance can directly by controlled by parameter* $\beta > 0$.

Future work will investigate methods for robust DoA estimates under conditions of large reverberation. As reverberation can be interpreted as injection of noise into the phase of the cross spectrum, one of the options is to average the phase from multiple pairs of microphones with specific geometric arrangements, thus yielding a robust estimate of the direct path DoA.

5. References

[1] Vary, P., Heute, U. and Hess, W., Digitale Sprachsignalverarbeitung, Teubner, Stuttgart, 1998.

[2] Brandstein, M. and Ward, D., Microphone Arrays, Springer, Berlin, 2001.

[3] Képesi, M., Pernkopf, F. and Wohlmayr, M., "Joint Position-Pitch Tracking for 2-Channel Audio", International Workshop on Content-Based Multimedia Indexing, 2007.

[4] Gustafsson, T., Rao, B.D. and Trivedi, M., "Source Localization in Reverberant Environments: Modeling and Statistical Analysis", IEEE Transactions on Speech and Audio Processing, Vol.11, 2003.

[5] Schroeder, M., "Parameter Estimation in Speech: a Lesson in Unorthodoxy", Proceedings of the IEEE, Vol.58, 1970.

[6] Brandstein, M.S., "A Pitch-based Approach to Time-Delay Estimation of Reverberant Speech", Applications of Signal Processing to Audio and Acoustics, 1997.