Multi-Channel Maximum Likelihood Pitch Estimation

Christensen, Mads Græsbøll

Published in:
2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)

DOI (link to publication from Publisher):
10.1109/ICASSP.2012.6287903

Publication date:
2012

Document Version
Accepted author manuscript, peer reviewed version

MULTI-CHANNEL MAXIMUM LIKELIHOOD PITCH ESTIMATION

Mads Græsbøll Christensen

Dept. of Architecture, Design & Media Technology
Aalborg University, Denmark
mgc@create.aau.dk

ABSTRACT

In this paper, a method for multi-channel pitch estimation is proposed. The method is a maximum likelihood estimator and is based on a parametric model where the signals in the various channels share the same fundamental frequency but can have different amplitudes, phases, and noise characteristics. This essentially means that the model allows for different conditions in the various channels, like different signal-to-noise ratios, microphone characteristics and reverberation. Moreover, the method does not assume that a certain array structure is used but rather relies on a more general model and is hence suited for a large class of problems. Simulations with real signals shows that the method outperforms a state-of-the-art multi-channel method in terms of gross error rate.

Index Terms— Pitch estimation, microphone arrays, multi-channel audio

1. INTRODUCTION

An important property of audio and speech signals is the pitch, and the pitch is one of the most frequently used features in the processing and analysis of acoustic signals. In many cases, the pitch of a signal is simply related to the fundamental frequency, which describes the number of times a periodic signal repeats per time interval. Bandlimited periodic signals can be expressed as a finite weighted sum of harmonically related sinusoids having frequencies that are integer multiples of this fundamental frequency. The problem of finding the fundamental frequency from such periodic signals buried in noise is called fundamental frequency or pitch estimation. Over the years, many different methods for pitch estimation have been devised from classical approaches such as harmonic summation and product methods [1] to more recent methods such as harmonic fitting [2], maximum likelihood [3], optimal filtering [4], subspace methods [5] and Bayesian methods [6, 7]. For an overview, we refer the reader to [5] and the references therein. Despite the host of methods devoted to pitch estimation, it appears that very few methods have been devised for estimating the pitch when multiple channels are available, as would be the case in microphone array processing or in studio recordings of music. Some methods do exist, however, including those of [8, 9] and the joint pitch and localization methods of [10, 11]. There are several reasons why multi-channel pitch estimation should be pursued when multiple channels are available. Firstly, the presence of more data is always beneficial in frequency estimation problems. Secondly, the conditions under which the signals have been recorded may differ from channel to channel and it may be difficult to pick one channel a priori as having the best conditions. Hence, a method using all channels is preferable. Also, spatial localization may make it fairly easy to attenuate noise from certain angles, although we shall not seek to exploit this here.

In this paper, a novel method for multi-channel pitch estimation is presented. It is a maximum likelihood estimator based on a Gaussian assumption and a parametric model of signal of interest. In this model, the fundamental frequency is shared across channels while amplitudes, phases and noise characteristics are allowed to be different for each channel. The model thus takes into account that, e.g., the signal in each channel may have been filtered and that the noise level may be different. Hence, the presented method is based on a quite general model that can be assumed to work in many different situations. An important aspect of this work is that the integration across channels is done in a mathematically tractable manner.

The rest of the paper is organized as follows: In Section 2, the underlying parametric model and statistical assumptions are presented after which the proposed method is derived in Section 3. Then, in Section 4 some experimental results demonstrating the advantages of the proposed method are presented. Finally, we conclude on the work in Section 5.

2. FUNDAMENTALS

We will now present the signal model and associated assumptions. The proposed method operates on a signal vector $\mathbf{x}_k(n) \in \mathbb{C}^M$ at time n (termed a snapshot) for the kth channel, defined as $\mathbf{x}_k(n) = [\, x_k(n) \; \cdots \; x_k(n+M-1) \,]^T$ which is constructed from the observed signal from the kth channel $x_k(n)$, for $n = 0, \ldots, N-1$. We model this vector as a sum of L harmonically related complex sinusoids in

Gaussian noise $\mathbf{e}_k$ having covariance matrix $\mathbf{Q}_k$, i.e.,

$$\mathbf{x}_k(n) = \mathbf{Z}(n)\mathbf{a}_k + \mathbf{e}_k(n), \tag{1}$$

with $\mathbf{a}_k = [\, A_{k,1}e^{j\phi_{k,1}} \;\cdots\; A_{k,L}e^{j\phi_{k,L}} \,]^T$ being a vector containing the complex amplitudes of the signal in the kth channel. Moreover, the matrix $\mathbf{Z}(n)$ is a Vandermonde matrix at time n, defined as $\mathbf{Z}(n) = \left[\begin{array}{ccc} \mathbf{z}_1(n) & \cdots & \mathbf{z}_L(n) \end{array} \right]$, where the mth entry of the column vector $\mathbf{z}_l(n) \in \mathbb{C}^M$ is defined as $[\mathbf{z}_l(n)]_m = e^{j\omega_0 l(n+m-1)}$ with $\omega_0 \in \Omega_0$ being the fundamental frequency, i.e., the parameter we seek to find in the set $\Omega_0 = (0, 2\pi/L)$. We assume that G vectors $\mathbf{x}_k(n)$ have been observed for each channel. We define the signal and noise parameter vector $\boldsymbol{\theta}_k$ for the kth channel containing the fundamental frequency ω_0, the complex amplitudes $\{A_{k,l}e^{j\phi_{k,l}}\}$ and the noise covariance matrix $\mathbf{Q}_k$. Regarding the model order L, we remark that it is possible to extend the proposed method to joint fundamental frequency and order estimation using the MAP principle [5]. However, for simplicity, we do not describe that here. Assuming that $\mathbf{Q}_k$ is invertible, the likelihood function (for complex signals) of $\mathbf{x}_k(n)$ can then be written as

$$p(\mathbf{x}_k(n); \boldsymbol{\theta}_k) = \frac{1}{\pi^M \det(\mathbf{Q}_k)} e^{-\mathbf{e}_k^H(n)\mathbf{Q}_k^{-1}\mathbf{e}_k(n)}, \tag{2}$$

with $\det(\cdot)$ denoting the matrix determinant. Now, assuming that the deterministic part is stationary and $\mathbf{e}_k(n)$ is independent and identically distributed over n as well as independent over k, the likelihood of the observed set of vectors $\{\{\mathbf{x}_k(n)\}_{n=0}^{G-1}\}_{k=1}^{K}$ (or $\{\mathbf{x}_k(n)\}$ for short) across channels can be written as

$$\begin{aligned} p(\{\mathbf{x}_k(n)\}; \{\boldsymbol{\theta}_k\}) &= \prod_{k=1}^{K}\prod_{n=0}^{G-1} p(\mathbf{x}_k(n); \boldsymbol{\theta}_k) \\ &= \prod_{k=1}^{K} \frac{1}{\pi^{MG}\det(\mathbf{Q}_k)^G} e^{-\sum_{n=0}^{G-1}\mathbf{e}_k^H(n)\mathbf{Q}_k^{-1}\mathbf{e}_k(n)}. \end{aligned} \tag{3}$$

There are several ways in which the noise covariance matrix can be estimated, but they are, however, all fairly involved and they are hence best avoided. Moreover, it may be difficult to say anything about the noise covariance matrix $\mathbf{Q}_k$ a priori. In that case, the best solution is to assume that the noise is white in each channel[1] but that the noise has different variance σ_k^2, i.e.,$\mathbf{Q}_k = \sigma_k^2\mathbf{I}$. Similar arguments hold regarding the assumption of the noise being independent across channels. With the above assumptions, the likelihood function for a single snapshot for channel k reduces to

$$p(\mathbf{x}_k(n); \boldsymbol{\theta}_k) = \frac{1}{(\pi\sigma_k^2)^{GM}} e^{-\frac{1}{\sigma_k^2}\|\mathbf{e}_k(n)\|^2}, \tag{4}$$

and the log-likelihood function is then $\ln p(\mathbf{x}_k(n); \boldsymbol{\theta}_k) = -M\ln(\pi\sigma_k^2) - \frac{1}{\sigma_k^2}\|\mathbf{e}_k(n)\|^2$, which across all channels and snapshots under the aforementioned conditions yields

$$\begin{aligned} &\ln p(\{\mathbf{x}_k(n)\}; \{\boldsymbol{\theta}_k\}) = \\ &-GM\sum_{k=1}^{K}\ln(\pi\sigma_k^2) - \sum_{k=1}^{K}\sum_{n=0}^{G-1}\frac{\|\mathbf{e}_k(n)\|^2}{\sigma_k^2}. \end{aligned} \tag{5}$$

[1]The white Gaussian distribution can be shown to be the one maximizes the entropy of the noise [12].

3. PROPOSED METHOD

We will now proceed to derive the proposed estimator. To do this, we first observe that the noise variance σ_k^2 and the complex amplitude vector $\mathbf{a}_k$ are specific to channel k while the fundamental frequency in $\mathbf{Z}$ is shared among all channels. Hence, the two former parameters can be estimated directly from the individual channels (for a particular fundamental frequency candidate). The maximum likelihood estimate of the amplitudes for channel k can readily be shown to be

$$\hat{\mathbf{a}}_k = \left(\sum_{n=0}^{G-1}\mathbf{Z}^H(n)\mathbf{Z}(n)\right)^{-1}\sum_{n=0}^{G-1}\mathbf{Z}^H(n)\mathbf{x}_k(n). \tag{6}$$

This, in turn, can be used to form a noise estimate for $n = 0, \ldots, G-1$ as $\hat{\mathbf{e}}_k(n) = \mathbf{x}_k(n) - \mathbf{Z}(n)\hat{\mathbf{a}}_k$ and, from this, a maximum likelihood noise variance estimate for channel k as

$$\hat{\sigma}_k^2 = \frac{1}{GM}\sum_{n=0}^{G-1}\|\hat{\mathbf{e}}_k(n)\|^2. \tag{7}$$

Inserting these quantities into (5) then yields the concentrated log-likelihood for channel k at time n $\ln p(\mathbf{x}_k(n); \omega_0) = -M\ln\pi - M\ln\hat{\sigma}_k^2$, which depends only on the fundamental frequency ω_0, and the maximization of this function over the fundamental frequency would then leads to the maximum likelihood estimate for channel k. For all n and k, this yields

$$\ln p(\{\mathbf{x}_k(n)\}; \omega_0) = -GMK\ln\pi - GM\sum_{k=1}^{K}\ln\hat{\sigma}_k^2. \tag{8}$$

The maximum likelihood estimator (MLE) can finally be stated as

$$\hat{\omega}_0 = \arg\min_{\omega_0\in\Omega_0}\sum_{k=1}^{K}\ln\hat{\sigma}_k^2. \tag{9}$$

To summarize how the estimator works for each candidate fundamental frequency $\omega_0 \in \Omega_0$, the amplitudes are first found using (6) where after the noise variance is estimated for each channel k using (7). Then, the variances are integrated across channels as in (8) and the fundamental frequency can then be determined using (9). An interesting special case can be obtained as follows. For $M = N$ only one signal vector will be available for each channel. We denote this as $\mathbf{x}_k = \mathbf{x}_k(0)$ and similarly for the other quantities. Then the channel k estimators reduce to

$$\hat{\mathbf{a}}_k = \left(\mathbf{Z}^H\mathbf{Z}\right)^{-1}\mathbf{Z}^H\mathbf{x}_k \quad \text{and} \quad \hat{\sigma}_k^2 = \frac{\mathbf{x}_k^H\boldsymbol{\Pi}_Z^{\perp}\mathbf{x}_k}{N} \tag{10}$$

Parameter	Value	Parameter	Value
Sound vel.	340 m/s	Room Dim.	[5 4 6] m
Source pos.	[2 3.5 2] m	Samples	4096
Reverb time	0.4 s	Mic.	Hypercard.
Ref. order	-1	Number of Mic.	4
Mic. pos.	Random	Mic. orient.	Random

Table 1. Experimental settings for the mirror method.

where $\mathbf{\Pi}_Z = \mathbf{Z}\left(\mathbf{Z}^H\mathbf{Z}\right)^{-1}\mathbf{Z}^H$ and $\mathbf{\Pi}_Z^\perp = \mathbf{I} - \mathbf{\Pi}_Z$. Noting then that the columns of $\mathbf{Z}$ are asymptotically orthogonal, i.e., $\lim_{M\to\infty} M\mathbf{\Pi}_Z = \mathbf{Z}\mathbf{Z}^H$, the resulting estimator can be written as follows:

$$\hat{\omega}_0 = \arg\min_{\omega_0\in\Omega_0} \sum_{k=1}^{K} \ln\left(\|\mathbf{x}_k\|^2 - \frac{1}{N}\|\mathbf{Z}^H\mathbf{x}_k\|^2\right). \quad (11)$$

Finally, we observe that $\|\mathbf{Z}^H\mathbf{x}_k\|^2$ is just the sum over the squared magnitude of the Fourier transform of $x_k(n)$, denoted $X_k(\omega)$ evaluated in a set of frequencies (in this case those of the candidate harmonics), i.e., it can be evaluated efficiently using an FFT as $\|\mathbf{Z}^H\mathbf{x}_k\|^2 = \sum_{l=1}^{L}|X_k(\omega_0 l)|^2$. This can be seen as a simple extension of the classical harmonic summation method [1]. It should be noted that the amplitude estimate in (6) also can be computed efficiently for large M this way, although a phase shift must also be introduced to compensate for $\mathbf{Z}(n)$ being time-varying. It should also be noted that had we assumed that the noise variance was known, the result would have been different. In that case, the estimator would reduce to the maximizer of the weighted sum over the spectra across all channels, i.e., $\sum_{k=1}^{K}\sum_{l=1}^{L}|X_k(\omega_0 l)|^2/\sigma_k^2$. Moreover, if we also assume that the noise variance is the same for all channels, the resulting estimator would simply be the maximizer of $\sum_{k=1}^{K}\sum_{l=1}^{L}|X_k(\omega_0 l)|^2$. Both are to be contrasted with the sum over the logarithm in (11), which shows that the way in which the cost function should be integrated across channels depends on whether the noise variance is known and the same.

4. EXPERIMENTAL RESULTS

To investigate the performance of the proposed method we proceed as follows. We will follow a procedure similar to the test methodology of [9]. A set of single-pitch audio mono signals from the EBU SQAM discs is used, namely the trumpet, violin and horn signals. The signals are down-sampled by a factor of four from the original 44.1 kHz sampling frequency and converted to complex signals using the Hilbert transform, and the signals are processed (by all algorithms) in segments of 40 ms with 50 % overlap and an FFT size of 8192. From these signals, four different channels are generated using E. Habets' implementation[2] of the mirror method [13] with settings as shown in Table 1. Note that microphone positions and orientations were picked randomly. To demonstrate the merits of the proposed method, we add white Gaussian noise to each channel and test two different scenarios: one in which the noise level is the same in all channels, a scenario we will refer to as symmetrical, and one where the noise level is different in each channel, which we will refer to as asymmetrical. For the asymmetrical scenario, the SNRs in the individual channels were -15, -10, -5, and 0 dB, respectively, offset by an overall SNR value. A ground truth pitch is estimated from the clean multi-channel signal using the MPF method [9]. We will compare the proposed method to the MPF method, which, in [9], was demonstrated to outperform a multi-channel version of YIN [14] and the method of [8]. We will evaluate the fast, approximate version in (11) of the proposed method (denoted MLE), which relies on FFTs to compute the log-likelihood. Finally, we will, for reference, also compare to the performance obtained with the simple multi-channel extension of the classical harmonic summation method (denoted HS) mentioned in Section 3, which is also an approximate maximum likelihood estimator when the noise variance is the same in all channels. For the methods that require that the model order is estimated, we used the MAP criterion [5]. As in [9], we will measure the gross error rate (GER), defined as a relative error of more than 20 % relative to the ground-truth, under different conditions. It should be noted that this methodology favors the MPF method as any consistent error in estimates obtained from the clean and noisy multi-channel signals will not be punished for this method. The results are shown in Figures 1(a) and 1(b) for the two scenarios, respectively. A number of observations can be made from the figures. The proposed method performs the best for both scenarios having the lowest GER. The HS method can be seen to perform well for the symmetrical case, as predicted by the theory. However, it can also be seen to to break down when the noise level differs between channels. The MPF method is capable of handling this scenario due to the normalization procedure in [9]. However, it can generally be seen to be more sensitive to low SNRs than the proposed method.

5. CONCLUSION

In this paper, a novel multi-channel pitch estimator has been proposed. The method is based on a maximum likelihood approach, and it is based on a parametric model where the signal in each channel is modeled as a sum of harmonically related sinusoids in noise. The amplitudes and phases are allowed to vary across channels to account for different acoustic propagation paths and the signal-to-noise level is allowed to vary as well. The model is hence quite general and can be used in many different scenarios. Simulations demonstrate that the method generally performs well and outperforms a state-of-the-art method, especially under adverse conditions.

[2] http://home.tiscali.nl/ehabets/rir_generator.html

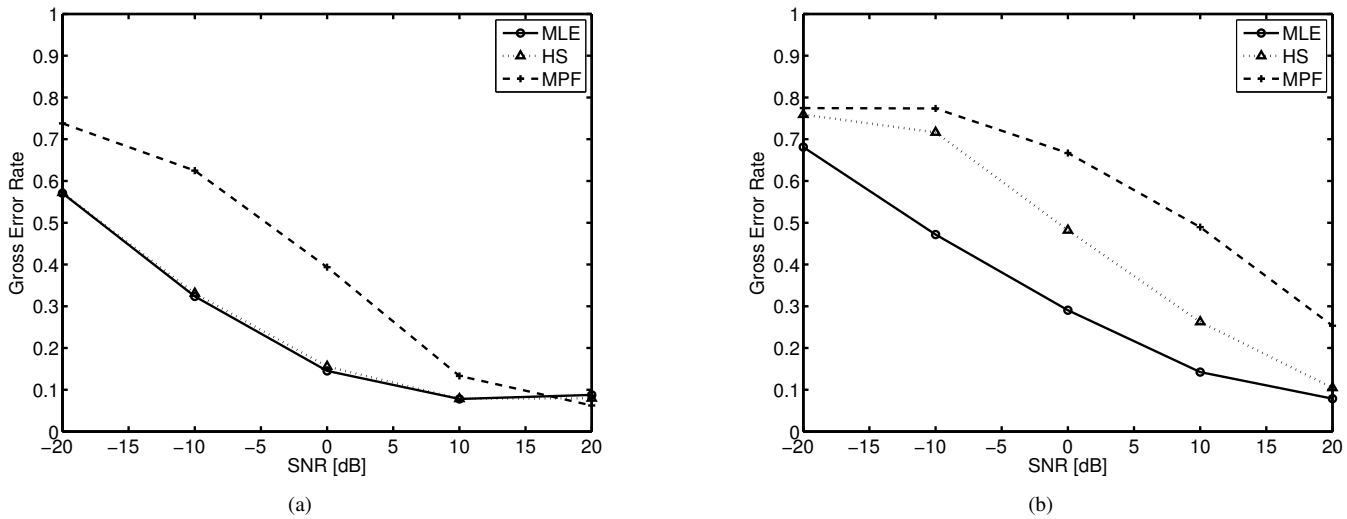

Fig. 1. Performance measured in terms of gross error rate as a function of the SNR in dB for (a) symmetrical noise level and (b) asymmetrical noise level. The SNR for the latter case is the overall SNR.

6. REFERENCES

[1] M. Noll, "Pitch determination of human speech by harmonic product spectrum, the harmonic sum, and a maximum likelihood estimate," in *Proc. Symposium on Computer Processing Communications*, 1969, pp. 779–797.

[2] H. Li, P. Stoica, and J. Li, "Computationally efficient parameter estimation for harmonic sinusoidal signals," *Signal Processing*, vol. 80, pp. 1937–1944, 2000.

[3] B. G. Quinn and P. J. Thomson, "Estimating the frequency of a periodic function," *Biometrika*, vol. 78(1), pp. 65–74, 1991.

[4] M. G. Christensen, J. L. Højvang, A. Jakobsson, and S. H. Jensen, "Joint fundamental frequency and order estimation using optimal filtering," *EURASIP J. on Advances in Signal Processing*, vol. 2011(1), pp. 13, 2011.

[5] M. G. Christensen and A. Jakobsson, *Multi-Pitch Estimation*, vol. 5 of *Synthesis Lectures on Speech & Audio Processing*, Morgan & Claypool Publishers, 2009.

[6] S. Godsill and M. Davy, "Bayesian harmonic models for musical pitch estimation and analysis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2002, vol. 2, pp. 1769–1772.

[7] A. T. Cemgil, H. J. Kappen, and D. Barber, "A generative model for music transcription," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 2, pp. 679–694, Mar. 2006.

[8] L. Armani and M. Omologo, "Weighted autocorrelation-based f0 estimation for distant-talking interaction with a distributed microphone network," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2004, vol. 1, pp. 113–116.

[9] F. Flego and M. Omologo, "Fobust f0 estimation based on a multi-microphone periodicity function for distant-talking speech," in *Proc. European Signal Processing Conf.*, 2006.

[10] M. Kepesi, F. Pernkopf, and M. Wohlmayr, "Joint position-pitch tracking for 2-channel audio," in *Proc. Int. Workshop Content-Based Multimedia Indexing*, 2007, pp. 303–306.

[11] J. R. Jensen, M. G. Christensen, and S. H. Jensen, "Joint DOA and Fundamental Frequency Estimation Methods based on 2-D Filtering," in *Proc. European Signal Processing Conference*, 2010.

[12] G. L. Bretthorst, "An introduction to parameter estimation using Bayesian probability theory," in *Max. Entropy and Bayesian Methods*, P. Fougere, Ed., pp. 53–79. 1990.

[13] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65(4), pp. 943–950, 1979.

[14] A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Am.*, vol. 111(4), pp. 1917–1930, Apr. 2002.