PITCH ESTIMATION OF STEREOPHONIC MIXTURES OF DELAY AND AMPLITUDE PANNED SIGNALS

Martin Weiss Hansen, Jesper Rindom Jensen and Mads Græsbøll Christensen

Audio Analysis Lab, AD:MT, Aalborg University, Denmark
{mwh,jrj,mgc}@create.aau.dk

ABSTRACT

In this paper, a novel method for pitch estimation of stereophonic mixtures is presented, and it is investigated how the performance is affected by the pan parameters of the individual signals of the mixture. The method is based on a signal model that takes into account a stereophonic mixture created by mixing multiple individual channels with different pan parameters, and is hence suited for use in automatic music transcription, source separation and classification systems. Panning is done using both amplitude differences and delays. The performance of the estimator is compared to one single-channel, two multi-channel and one multi-pitch estimator using synthetic and real signals. Experiments show that the proposed method is able to correctly estimate the pitches of a mixture of three real signals when they are separated by more than 25 degrees.

Index Terms— Pitch estimation, multi-channel processing, noise reduction, maximum likelihood.

1. INTRODUCTION

Pitch is an important feature of harmonic signals, such as short segments of music and speech. It is related to the fundamental frequency, which is the reciprocal of the period of a harmonic signal. Pitch estimation has applications in problems such as separation [1], enhancement [2], compression [3], modification [4], transcription [5], classification [6], time-delay estimation [7] and source localization [8].

Many pitch estimation methods exist, i.e., non-parametric methods based on autocorrelation [9, 10], the average magnitude difference function (AMDF) [11] and the harmonic product spectrum [12]. A drawback of these methods is that they can not distinguish between the fundamental pitch period and multiples of it, and they exhibit poor performance under noisy conditions. Another significant group of methods consists of statistical parametric methods, such as maximum likelihood (ML) [12]. These methods are based on parametric descriptions of the signals that we wish to analyze. It is worth noting that a lot of material, in particular music, is available in stereo. Therefore, exploiting this multi-channel property, multi-channel pitch estimation is interesting. One such method based on a multi-microphone periodicity function (MPF) is presented in [13], while a multi-microphone maximum a posteriori (MAP) approach is taken in [14]. A multi-channel maximum likelihood (MC ML) pitch estimator, which allows for different conditions in the channels is presented in [15], and a collection of statistical, parametric methods are presented in [16].

Pitch estimation is useful when analyzing musical performances. To the authors' knowledge no parametric method exists that exploit the channel pan parameters of stereophonic mixtures to obtain pitch estimates. A stereophonic mixture is created in recording studios by mixing several stereophonic signals. Each of these signals might have different mixing parameters, such as panning and equalization. In this paper, we take a closer look at mixtures composed of amplitude and delay panned signals. Amplitude panning is a frequently used virtual source positioning technique, where different gains are applied to the individual channels of a signal. The perception of direction is dependent on these gain factors [17]. A time delay can be added to one of the channels of the signal to enhance the spatial quality of the signal and to add depth [18]. If a signal is delayed by more than 1 ms in a stereo setup, the perceived direction of the source is determined mostly by the signal which arrives first [19]. According to [18], the spatial quality of a signal is enhanced by using delays in the 12 to 40 ms range. The effect is called the Haas effect [20]. The idea of separating sources from a multi-channel mixture is used within the source separation [21] and array processing [22] research communities but it has, to the knowledge of the authors, not been applied within the area of pitch estimation and its application in, for example, music transcription.

In this paper, we propose a pitch estimation method for such stereophonic mixtures. In this work, these mixtures are assumed to be created by mixing several stereophonic channels with known pan parameters. The method is based on the ML principle, where each signal is modeled as a sum of delayed and attenuated sinusoids. The aim of the work pre-

This work was supported in part by the Villum Foundation, the Danish Council for Independent Research, grant ID: DFF 1337-00084, and the Danish Council for Strategic Research of the Danish Agency for Science Technology and Innovation under the CoSound project, case number 11-115328. This publication only reflects the authors' views.

978-0-9928626-3-3/15/$31.00 ©2015 IEEE

sented in this paper is to estimate the pitches of the individual signals that constitute a stereophonic mixture, when the mixing parameters, i.e., the amplitude and delay pan parameters, of the signals are known. It should be noted that in this work we consider finding the pan parameters a separate problem.

The remainder of the paper is organized as follows. In Section 2, the signal model is introduced. The proposed pitch estimator is described in Section 3. The experimental setup and results are presented in Section 4, and the work is concluded in Section 5.

2. SIGNAL MODEL

We now introduce the signal model and assumptions. Consider a K-channel mixture, where the data in channel k at time n can be represented by the snapshot $\mathbf{x}_k(n) \in \mathbb{C}^N$, i.e.,

$$\mathbf{x}_k(n) = [x_k(n) \quad x_k(n+1) \quad \cdots \quad x_k(n+N-1)]^T, \quad (1)$$

for $k = 0, \ldots, K-1$, where $x_k(n)$ is the signal in channel k at time n. We assume that the snapshot (1) is composed of M sources spatially enhanced by amplitude and delay panning. An example of an amplitude pan law that could be applied in a stereophonic mix, i.e., $K = 2$, is [23]

$$g_k = \begin{cases} \cos\theta_m, & \text{for } k = 0. \\ \sin\theta_m, & \text{for } k = 1. \end{cases} \quad (2)$$

where $k = 0$ and $k = 1$ denote the signals at the left and right loudspeaker, respectively, and θ_m is the angle between the pan direction and the left loudspeaker for the mth source. The aperture of the speakers is 90°, resulting in equal amplitudes for $\theta_m = 45°$, while only one channel will be active when $\theta_m = 0°$ or $\theta_m = 90°$. As previously mentioned, delays can be used to enhance the spatial perception [19, 18]. We model the kth channel as a linear superposition of M attenuated and delayed sources, corrupted by noise $\mathbf{e}_{k,m}(n)$, at time n i.e.,

$$\mathbf{x}_k(n) = \sum_{m=0}^{M-1} g_{k,m}\mathbf{s}_m(n - f_s\tau_{k,m}) + \mathbf{e}_{k,m}(n), \quad (3)$$

where $m = 0, \ldots, M-1$, and

$$s_m(n - f_s\tau_m) = \sum_{l_m=1}^{L_m} \alpha_{l,m} e^{jl_m\omega_{0,m}n} e^{-j\omega_{0,m}l_m f_s\tau_m}$$

is a delayed version of the mth source, $l_m = 1, \ldots, L_m$ is the harmonic index, where L_m is the model order, f_s is the sampling frequency, $\omega_{0,m}$ is the fundamental frequency, $\alpha_{l,m} = A_{l,m}e^{\phi_{l,m}}$, where $A_{l,m}$ is the real amplitude of the l_mth harmonic, $\phi_{l,m}$ its phase, and $g_{k,m}$ and $\tau_{k,m}$ denote the gain and delay applied to the signal, respectively. It should be noted that although the signal model is complex, it can be used on real signals by applying the Hilbert transform. We model the kth channel in (3) as a sum of L_m harmonically related complex sinusoids, in Gaussian noise $\mathbf{e}_{k,m}(n)$ with noise covariance $\mathbf{Q}_{k,m}$, i.e.,

$$\mathbf{x}_k(n) = \sum_{m=0}^{M-1} \mathbf{Z}_m(n)\mathbf{G}(k,m)\mathbf{a}_m + \mathbf{e}_{k,m}(n), \quad (4)$$

where $\mathbf{a}_m = [\alpha_{1,m} \quad \cdots \quad \alpha_{L,M}]^T$ is a vector of complex amplitudes, $\mathbf{Z}_m(n)$ is a Vandermonde matrix, defined as $\mathbf{Z}_m(n) = [\mathbf{z}_{1,m}(n) \quad \cdots \quad \mathbf{z}_{L_M,m}(n)]$, where $\mathbf{z}_{l,m}(n) = [1 \quad e^{j\omega_{0,m}} \quad \ldots \quad e^{j\omega_{0,m}l_m(N-1)}]^T$, and $\mathbf{G}(k,m)$ is a diagonal matrix, i.e.,

$$\mathbf{G}(k,m) = \begin{bmatrix} g_{k,m}e^{-j\omega_{0,m}f_s\tau_{k,m}} \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots g_{k,m}e^{-jL_m\omega_{0,m}f_s\tau_{k,m}} \end{bmatrix}.$$

Assuming that $\mathbf{Q}_{k,m}$ is invertible, the likelihood function of (4) can be written as [16, 15]

$$p(\mathbf{x}_k(n); \boldsymbol{\omega}_0) = \frac{1}{\pi^N \det(\mathbf{Q}_{k,m})} e^{-\mathbf{e}_{k,m}^H(n)\mathbf{Q}_{k,m}^{-1}\mathbf{e}_{k,m}(n)}. \quad (5)$$

If the deterministic part of the signal is stationary, and $\mathbf{e}_{k,m}(n)$ is independent and identically distributed over n and k, the likelihood of the observed set of vectors $\{\mathbf{x}_k(n)\}$ can be written as

$$p(\{\mathbf{x}_k(n)\}; \boldsymbol{\omega}_0) = \prod_{k=0}^{K-1} p(\mathbf{x}_k(n); \boldsymbol{\omega}) =$$
$$\prod_{k=0}^{K-1} \frac{1}{\pi^N \det(\mathbf{Q}_{k,m})} e^{-\mathbf{e}_{k,m}^H(n)\mathbf{Q}_{k,m}^{-1}\mathbf{e}_{k,m}(n)}.$$

If the noise $\mathbf{e}_{k,m}(n)$ is white, but with different variance in each channel, i.e, $\mathbf{Q}_{k,m} = \sigma_{k,m}^2\mathbf{I}$, (5) can be written as

$$p(\mathbf{x}_k(n); \boldsymbol{\omega}_0) = \frac{1}{(\pi\sigma_{k,m}^2)^N} e^{-\frac{1}{\sigma_{k,m}^2}\|\mathbf{e}_{k,m}(n)\|^2},$$

and the log-likelihood is $\ln p(\mathbf{x}_k(n); \boldsymbol{\omega}_0) = -N \ln(\pi\sigma_{k,m}^2) - \frac{1}{\sigma_{k,m}^2}\|\mathbf{e}_{k,m}(n)\|^2$, which for all channels is

$$\ln p(\{\mathbf{x}_k(n)\}; \boldsymbol{\omega}_0) = -N\sum_{k=0}^{K-1} \ln(\pi\sigma_{k,m}^2) - \frac{\|\mathbf{e}_{k,m}(n)\|^2}{\sigma_{k,m}^2}. \quad (6)$$

3. PROPOSED METHOD

We will now derive the proposed pitch estimator. To do this, the log-likelihood (6) is maximized wrt. the parameters that we wish to estimate. The noise variance $\sigma_{k,m}^2$ and the pan matrix $\mathbf{G}_{k,m}$ are specific to channel k of the mth source. The complex amplitudes $\mathbf{a}_m$ and the matrix $\mathbf{Z}_m(n)$ of the mth

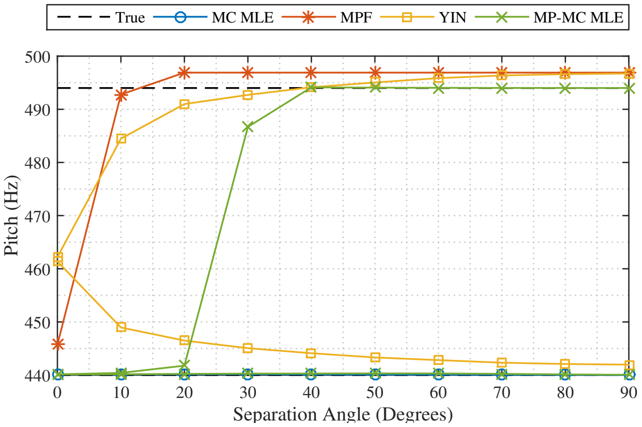

Fig. 1. Pitch estimates for different separation angles. The mixture is composed of two synthetic signals with amplitude panning applied.

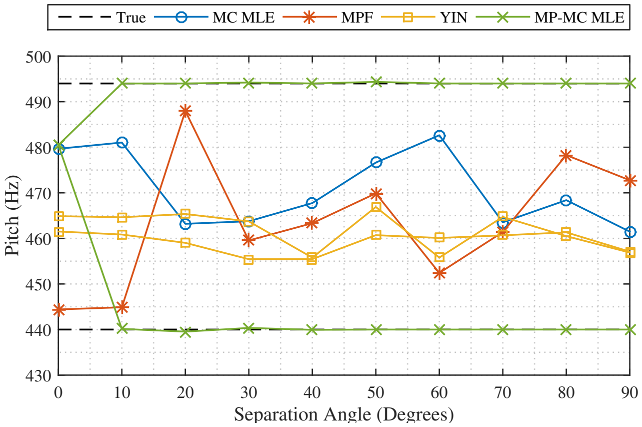

Fig. 2. Pitch estimates for different separation angles. The mixture is composed of two synthetic signals with delay panning applied.

signal are shared among all channels. First the log-likelihood (6) is differentiated wrt. the complex amplitudes $\mathbf{a}_m$, and we equate with zero to obtain the amplitude estimates

$$\hat{\mathbf{a}}_m = \left[\sum_{k=0}^{K-1} \frac{\mathbf{G}^H(k,m)\mathbf{Z}_m^H(n)\mathbf{Z}_m(n)\mathbf{G}(k,m)}{\sigma_{k,m}^2}\right]^{-1} \sum_{k=0}^{K-1} \frac{\mathbf{G}^H(k,m)\mathbf{Z}_m^H(n)\mathbf{x}_k(n)}{\sigma_{k,m}^2}. \tag{7}$$

The amplitude estimates in (7) can be used to form a noise estimate for $n = 0, \dots, N-1$. If (6) is differentiated wrt. the noise variance on sensor k, and equated to zero, we can solve for the variance, with $\hat{\mathbf{e}}_{k,m}(n) = \mathbf{x}_k(n) - \mathbf{Z}_m(n)\mathbf{G}(k,m)\hat{\mathbf{a}}_m$, resulting in the noise variance estimate

$$\hat{\sigma}_{k,m}^2 = \frac{1}{N}\|\hat{\mathbf{e}}_{k,m}(n)\|^2. \tag{8}$$

Combining (6) and (8) results in the concentrated log-likelihood for all n and k

$$\ln p(\{\mathbf{x}_k(n)\}; \boldsymbol{\omega}_0) = -NK \ln(1+\pi) - N \sum_{k=0}^{K-1} \ln \hat{\sigma}_{k,m}^2.$$

The maximum likelihood estimator for the pitch of the mth signal can then be stated as

$$\hat{\omega}_{0,m} = \underset{\{\omega_{0,m}\} \in \Omega_{0,m}}{\arg\min} \sum_{k=0}^{K-1} \ln \|\mathbf{x}_k(n) - \mathbf{Z}_m(n)\mathbf{G}(k,m)\hat{\mathbf{a}}_m\|^2,$$

where $\Omega_{0,m}$ is a set of fundamental frequencies. It should be noted that the pan parameters can be found by adding search dimensions to the above estimator. This is not done here, but it could be exploited that the pan parameters are usually fixed for longer periods of time.

4. EXPERIMENTS

We now present the experimental evaluation of the proposed pitch estimator, which has been compared to a single-channel auto-correlation-based method, namely YIN [10], the multi-channel MPF method in [13] and finally the multi-channel ML pitch estimator in [15]. In the evaluation of the proposed method the pan parameters are assumed to be known, and the objective is to see how these pan parameters influence the performance of the pitch estimator. A stereophonic mixture, i.e. $K = 2$, consisting of $M = 2$ synthetic signals, s_0 and s_1, with fundamental frequencies $f_{0,0} = 440$ Hz and $f_{0,1} = 494$ Hz have been used for the evaluation.

Three experiments were conducted using synthetic signals, to assess the performance of the proposed method. In the experiments the pitches of the signals are estimated for 10 different pan settings. 200 Monte-Carlo simulations were performed for each setting. In the first setting two synthetic signals are positioned in the middle of the scene. For each of the following settings, the signals are panned away form the center. The amplitude pan law (2) [23] is used. For all three experiments the mixture was analyzed using non-overlapping frames of length $N = 200$ samples, which corresponds to 25 ms at a sampling frequency of 8 kHz, and the results are generated by estimating the pitch in all frames for each setting, and averaging the resulting estimates. The true values are plotted for comparison.

In the first experiment only amplitude panning was applied to the signals, i.e., $\tau_{k,m} = 0$ for all k and m. The single-channel YIN method estimates the pitch for each of the K channels of the mixture, while the MPF and MC MLE methods operate on the multichannel mixture. The results show convergence towards the true pitches at smaller separation angles for the proposed method, compared to the other methods. The results are shown in Figure 1. In the second experiment delay panning was used, i.e. $\theta_m = 45°$ for all m,

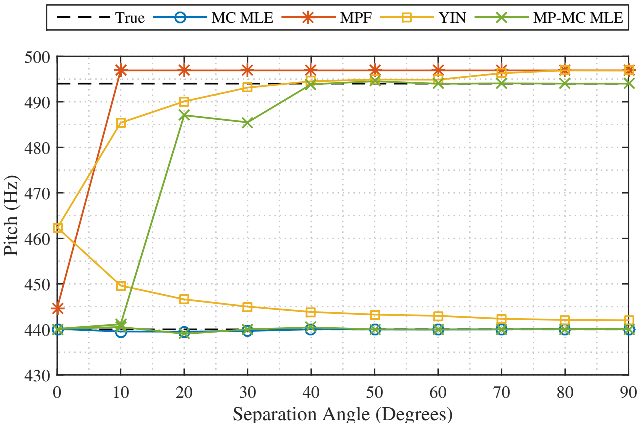

Fig. 3. Pitch estimates for different separation angles. The mixture is composed of two synthetic signals with amplitude and delay panning applied.

which in turn means that g_k are all equal. Delays were added to the attenuated channel of each signal, varying from 0 ms to 40 ms. In this experiment, none of the methods to which the proposed method is compared give the true values on average. The signal model allows for different delays $\tau_{k,m}$, which is why this result is expected. The results are shown in Figure 2. In the third experiment a combination of amplitude and delay panning were used. The gains g_k for each signal were varied as in the first experiment, and the delays $\tau_{m,k}$ were varied as in the second experiment. In this experiment, the results are similar to the results of the first experiment, only more pronounced. The results are shown in Figure 3.

The proposed method is also evaluated using a mixture of three trumpet signals with vibrato, played fortissimo (very loud)[1]. The tones played are A4 ($\approx$ 440 Hz), B4 ($\approx$ 494 Hz) and Db5 ($\approx$ 554 Hz). The fundamental frequencies of the signals are estimated jointly together with the model order using the ANLS method in [16] for comparison, since no ground truth pitches values are available. White Gaussian noise is added to result in an SNR of 20 dB, and the mixture is downsampled from 44.1 kHz to 8 kHz, and converted to a complex signal using the Hilbert transform. A spectrogram of the mixture and the pitch tracks of each signal are shown in Figure 4. The mixture is processed in frames of length $N = 200$ samples, and two of the signals are panned to the sides with a separation angle of 50°, while the third signal is in the center. The proposed method is compared to the MIRtoolbox [24] implementation of the enhanced summary autocorrelation function (ESACF) presented in [25]. The pitch estimates are shown in Figure 5. As the figure shows, the pitch estimates of the proposed estimator are closer to the ANLS estimates than the ESACF estimator. It is worth noting that the proposed method seems to work well, even though the signal model of the proposed method does not model the vibrato of the trumpet.

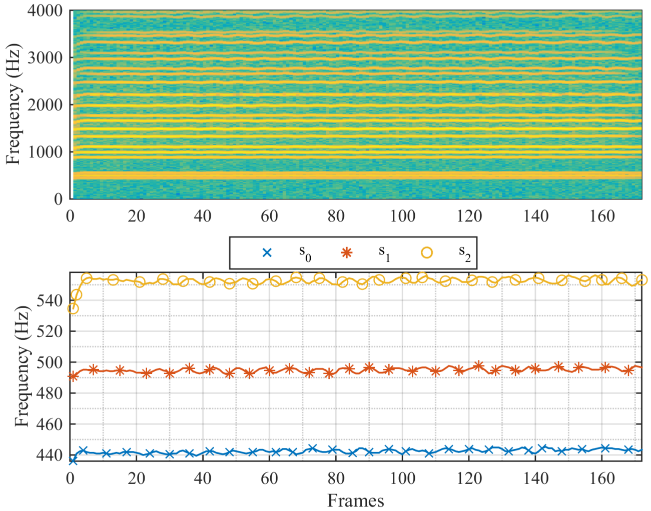

Fig. 4. Spectrogram of trumpet mixture (top), and pitch tracks (bottom).

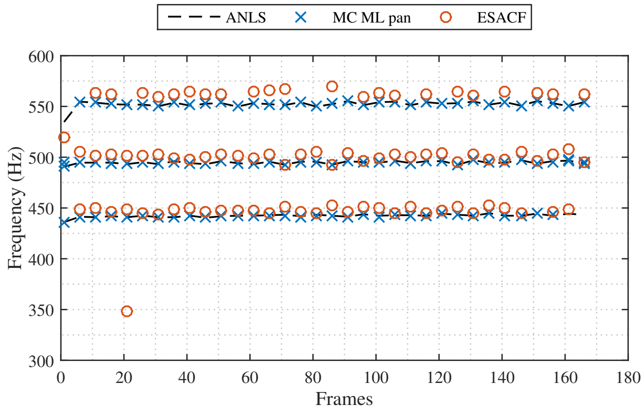

Fig. 5. Pitch estimates of the individual signals of a mixture of three trumpet signals with amplitude and delay panning applied.

5. DISCUSSION

In this paper, a novel method for pitch estimation of stereophonic mixtures has been proposed. The method is based on a maximum-likelihood approach, where a mixture is described using a parametric model, taking amplitude and delay pan parameters into account. Simulations show that the proposed method outperforms the single-channel, multi-channel and multi-pitch methods to which it is compared. An application of the proposed method could be to investigate pan method and settings in recorded mixtures. The method could also be used in transcription and separation systems. As future work it would be interesting to look at joint estimation of the pan parameters and the pitch, since it could be exploited that the pan parameters are stationary for longer periods of time. It would also be interesting to investigate the current noise assumptions, and to extend the current method to allow multiple pitches in the signals that consitute a mixture.

[1] Can be downloaded at `http://theremin.music.uiowa.edu`.

6. REFERENCES

[1] M. I. Mandel, R. J. Weiss, and D. P. W. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 18, no. 2, pp. 382–394, Feb. 2010.

[2] J. R. Jensen, J. Benesty, M. G. Christensen, and S. H. Jensen, "Joint filtering scheme for nonstationary noise reduction," in *Proc. European Signal Processing Conf.*, 2012, pp. 2323–2327.

[3] E. B. George and M. J. T. Smith, "Analysis-by-synthesis/overlap-add sinusoidal modeling applied to the analysis and synthesis of musical tones," *J. Audio Eng. Soc.*, vol. 40, no. 6, pp. 497–516, 1992.

[4] E. Moulines and J. Laroche, "Non-parametric techniques for pitch-scale and time-scale modification of speech," *Speech Commun.*, vol. 16, no. 2, pp. 175–205, Feb. 1995.

[5] A. Klapuri and M. Davy, Eds., *Signal Processing Methods for Music Transcription*, Springer, New York, 2006.

[6] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 5, pp. 293–302, Jul 2002.

[7] M. S. Brandstein, "A pitch-based approach to time-delay estimation of reverberant speech," in *Proc. IEEE Workshop Appl. of Signal Process. to Aud. and Acoust.*, Oct 1997.

[8] J. R. Jensen, M. G. Christensen, and S. H. Jensen, "Nonlinear least squares methods for joint DOA and pitch estimation," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 21, no. 5, pp. 923–933, 2013.

[9] L. Rabiner, "On the use of autocorrelation analysis for pitch detection," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 25, no. 1, pp. 24–33, Feb 1977.

[10] A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Am.*, vol. 111, no. 4, pp. 1917–1930, 2002.

[11] M. Ross, H. Shaffer, A Cohen, R. Freudberg, and H. Manley, "Average magnitude difference function pitch extractor," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 22, no. 5, pp. 353–362, Oct 1974.

[12] M. Noll, "Pitch determination of human speech by the harmonic product spectrum, the harmonic sum spectrum and a maximum likelihood estimate," in *Proc. Symp. Comput. Process. Commun.* 1969, vol. XIX, pp. pp. 779–797, Polytechnic Press: Brooklyn, New York.

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

[14] T. Gerkmann, R. Martin, and D. Dalga, "Multi-microphone maximum a posteriori fundamental frequency estimation in the cepstral domain," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2009, pp. 4505–4508.

[15] M. G. Christensen, "Multi-channel maximum likelihood pitch estimation," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 409–412, 2012.

[16] M. G. Christensen and A. Jakobsson, *Multi-Pitch Estimation*, Synthesis lectures on speech and audio processing. Morgan & Claypool Publishers, 2009.

[17] V. Pulkki, *Spatial sound generation and perception by amplitude panning techniques*, Helsinki University of Technology, 2001.

[18] B. Katz, *Mastering Audio - The Art and the Science*, Focal Press, 2007.

[19] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*, MIT Press, 1997.

[20] H. Haas, "The influence of a single echo on the audibility of speech," *J. Audio Eng. Soc.*, vol. 20, no. 2, pp. 146–159, 1972.

[21] B. Gold, N. Morgan, and D. Ellis, *Speech and Audio Signal Processing - Processing and Perception of Speech and Music, Second Edition.*, Wiley, 2011.

[22] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*, Springer Topics in Signal Processing. Springer, 2008.

[23] J. C. Bennett, K. Barker, and F. O. Edeko, "A new approach to the assessment of stereophonic sound system performance," *J. Audio Eng. Soc.*, vol. 33, no. 5, pp. 314–321, 1985.

[24] O. Lartillot and P. Toiviainen, "A MATLAB toolbox for musical feature extraction from audio," in *Proc. of the 10th Int. Conference on Digital Audio Effects (DAFx-07)*, 2007.

[25] T. Tolonen and M. Karjalainen, "A computationally efficient multipitch analysis model," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 6, pp. 708–716, Nov 2000.