ON FREQUENCY DOMAIN MODELS FOR TDOA ESTIMATION

Jesper Rindom Jensen[1], *Jesper Kjær Nielsen*[23], *Mads Græsbøll Christensen*[1], *Søren Holdt Jensen*[3]

[1]Aalborg University
Audio Analysis Lab, AD:MT
{jrj,mgc}@create.aau.dk

[2]Bang & Olufsen A/S
Struer, Denmark

[3]Aalborg University
Dept. of Electronic Systems
{jkn,shj}@es.aau.dkt

ABSTRACT

Time-difference-of-arrival (TDOA) estimation is an important problem in many microphone signal processing applications. Traditionally, this problem is solved by using a cross-correlation method, but in this paper we show that the cross-correlation method is actually a restricted special case of a much more general method. In this connection, we establish the conditions under which the cross-correlation method is a statistically efficient estimator. One of the conditions is that the source signal is periodic with a known fundamental frequency of $2\pi/N$ radians per sample, where N is the number of data points, and a known number of harmonics. The more general method only relies on that the source signal is periodic and is, therefore, able to outperform the cross-correlation method in terms of estimation accuracy on both synthetic data and artificially delayed speech data. The simulation code is available online.

Index Terms— (Fractional) TDOA Estimation, Fundamental Frequency Estimation, Generalised Cross-correlation

1. INTRODUCTION

The estimation of an angle or a location, from which an unknown source signal originates, is an important problem in many applications. In, e.g., audio applications, such estimates can be used to separate simultaneously talking speakers, to attenuate unwanted background noise, and to estimate the geometry of a room [1–3]. These direction-of-arrival (DOA) and source localisation estimation problems can be boiled down to the problem of estimating the time-difference-of-arrivals (TDOA) between the sensors of an array, and we here consider the problem of estimating the TDOA between two sensors. Such TDOA estimates between sensor pairs are often required input to algorithms (such as the popular SRP-PHAT algorithm [1]) operating on data recorded by more than two microphones [2].

The collection of correlation-based methods referred to as the generalised cross-correlation (GCC) methods [4] is by far the most widely used way to compute TDOA estimates in audio applications. In contrast to radar and sonar applications, the source signal is here typically a wideband signal so statistically efficient algorithms such as MUSIC [5] and ESPRIT [6] developed for narrowband signal models cannot be used directly. Moreover, the broadband version of the MUSIC algorithm has a high computational cost [7] in contrast to the GCC methods, which can be implemented efficiently using an FFT algorithm when these methods are formulated in the frequency domain. Another advantage to formulating the signal model in the frequency domain is that the delay parameter is also separated analytically from the source signal and can be modelled as a continuous parameter. Consequently, most papers on DOA estimation and source localisation for audio applications take their outset in a frequency domain model.

The main points in this paper are most easily demonstrated if we consider the simplest parametric model for TDOA estimation

$$\begin{aligned} x_1(n) &= s(n) + e_1(n) \\ x_2(n) &= \beta s(n-\eta) + e_2(n) \end{aligned} \tag{1}$$

for $n = 0, 1, \ldots, N-1$ where the signals $x_i(n)$, $s(n)$, and $e_i(n)$ are the i'th sensor signal, the source signal, and the noise on sensor i, respectively. The scalars $\beta > 0$ and $\eta \in [-N/2, N/2)$ are the attenuation and the relative delay in samples, respectively, of the source signal from sensor 1 to 2. If the source signal is a periodic signal with the fundamental frequency $2\pi/N$ radians per sample (or an integer multiple thereof), the model in (1) can be written in the frequency domain as

$$\begin{aligned} X_1(k) &= S(k) + E_1(k) \\ X_2(k) &= \beta S(k) \exp(-j2\pi k\eta/N) + E_2(k) \end{aligned} \tag{2}$$

for $k = 0, 1, \ldots, N-1$ where $X_i(k)$, $S(k)$, and $E(k)$ are the discrete Fourier transform coefficients of the signals $x_i(n)$, $s(n)$, and $e_i(n)$, respectively. The frequency-domain model in (2) suffers from a number of problems. First of all, it is often too restrictive. Although the source signal is often approximately periodic on a short time scale in audio applications, the assumption on the fundamental frequency is usually not satisfied in practice. This will lead to artefacts which are commonly referred to as edge effects [8–10]. These edge effects can be avoided by introducing appropriate zero padding, but this will colour the noise spectrum by a rank-deficient correlation matrix [8]. Another problem is that the frequency-domain model in (2) cannot be used for fractional TDOA estimation since a non-integer delay of a real-valued source signal results in a complex-valued sensor signal!

Due to these problems, we instead propose a different model which only assume the source signal to be periodic, but not that the fundamental frequency is $2\pi/N$ radians per sample. Modelling the fundamental frequency as an unknown parameter and estimating it jointly with the TDOA or DOA is not a new idea [11–15]. However, we here show that this model is actually more general than the traditional frequency domain model as the latter is a special case of the former, and we also establish the conditions under which the cross-correlation method is a statistically efficient estimator. In this connection, we propose a new approximate maximum likelihood estimator for joint fundamental frequency and TDOA estimation which outperforms the cross-correlation method on both synthetic data and artificially delayed speech data. In contrast to the traditional cross-correlation method, the proposed estimator also produces fractional delay estimates without resorting to, e.g., interpolation methods.

The work by J. R. Jensen was supported by the Danish Council for Independent Research, grant ID: DFF – 1337-00084. The work by J. K. Nielsen was supported by InnovationsFonden. The work by M. G. Christensen was supported by the Villum Foundation.

2. JOINT FUNDAMENTAL FREQUENCY AND TDOA ESTIMATION

As we alluded to in the introduction, we will here not make any assumption in our signal model on the fundamental frequency. As we detail below, we instead assume that the source signal is periodic with an unknown fundamental frequency and an unknown number of harmonic components. We will also make the assumption that the noise is white and Gaussian. Although this will potentially lead to poor estimation performance in audio applications with significant reverberation, the assumption is sufficient here to demonstrate our main points.

2.1. The Model

Any zero-mean, real-valued, and periodic source signal can be written as

$$s(n) = \sum_{k=1}^{L} A_k \cos(\omega_0 k n + \phi_k) = \sum_{k=-L}^{L} \alpha_k \exp(j\omega_0 k n) \quad (3)$$

where $A_k > 0$, $\phi_k \in [-\pi, \pi)$, $\omega_0 \in (0, \pi/L)$, and $\alpha_k = \alpha^*_{-k} = A_k \exp(j\phi_k)/2$ are the amplitude, phase, fundamental frequency, and complex amplitude, respectively. Note that $A_0 = 0$ which corresponds to the physical fact that the source signal has no DC-component. If we delay the source signal by the delay η, we therefore obtain that

$$s(n-\eta) = \sum_{k=-L}^{L} \alpha_k \exp(j\omega_0 k n) \exp(-j\xi k) \quad (4)$$

where we have defined $\xi \triangleq \omega_0 \eta$. In matrix-vector notation, the signal model in (1) can therefore be written as

$$\boldsymbol{x} = \begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{Z}(\omega_0) \\ \beta \boldsymbol{Z}(\omega_0) \boldsymbol{D}(\xi) \end{bmatrix} \boldsymbol{\alpha} + \boldsymbol{e} = \boldsymbol{H}(\beta, \xi, \omega_0)\boldsymbol{\alpha} + \boldsymbol{e} \quad (5)$$

where we have defined

$$\begin{aligned}
\boldsymbol{z}(\omega) &\triangleq \begin{bmatrix} 1 & \exp(j\omega) & \cdots & \exp(j\omega(N-1)) \end{bmatrix}^T \\
\boldsymbol{Z}(\omega_0) &\triangleq \begin{bmatrix} \boldsymbol{z}(-L\omega_0) & \cdots & \boldsymbol{z}(-\omega_0) & \boldsymbol{z}(\omega_0) & \boldsymbol{z}(L\omega_0) \end{bmatrix} \\
\boldsymbol{D}(\xi) &\triangleq \operatorname{diag}\Big(\exp(jL\xi), \ldots, \exp(j\xi), \\
&\qquad \exp(-j\xi), \ldots, \exp(-jL\xi) \Big) \\
\boldsymbol{\alpha} &\triangleq \begin{bmatrix} \alpha_{-L} & \cdots & \alpha_{-1} & \alpha_1 & \cdots & \alpha_L \end{bmatrix}^T \\
\boldsymbol{H}(\beta, \xi, \omega_0) &= \begin{bmatrix} \boldsymbol{Z}(\omega_0) \\ \beta \boldsymbol{Z}(\omega_0) \boldsymbol{D}(\xi) \end{bmatrix} .
\end{aligned}$$

Moreover, we assume the noise to be white and Gaussian with variance σ^2. Thus, the observation model is the normal distribution with probability density function (pdf)

$$p(\boldsymbol{x}|\boldsymbol{\alpha}, \beta, \xi, \omega_0, \sigma^2) = \mathcal{N}(\boldsymbol{H}(\beta, \xi, \omega_0)\boldsymbol{\alpha}, \sigma^2 \boldsymbol{I}_{2N}) \quad (6)$$

where $\boldsymbol{I}_{2N}$ is the $2N \times 2N$ identity matrix.

2.2. An Approximate ML Estimator

The observation model in (6) consists of the $2L$ linear parameters in $\boldsymbol{\alpha}$, the noise variance σ^2, and the nonlinear parameters β, ω_0, and ξ. We obtain the maximum likelihood estimates of these parameters if the observation model is maximised w.r.t. to these parameters. The linear parameters and noise variance are easily separated out of the optimisation problem leaving us with the non-convex optimisation problem

$$(\hat{\beta}, \hat{\xi}, \hat{\omega}_0) = \underset{\beta>0, \xi\in[-\pi,\pi), \omega_0 \in (0, \pi/L)}{\arg\max} J(\beta, \xi, \omega_0) \quad (7)$$

where the cost function is given by

$$J(\beta, \xi, \omega_0) = \boldsymbol{x}^H \boldsymbol{H}(\beta, \xi, \omega_0) \left[\boldsymbol{H}^H(\beta, \xi, \omega_0) \boldsymbol{H}(\beta, \xi, \omega_0) \right]^{-1} \times \boldsymbol{H}^H(\beta, \xi, \omega_0) \boldsymbol{x} \, . \quad (8)$$

This cost function is also sometimes referred to as the nonlinear least squares (NLS) cost function. Although possible in principle, it is not computationally feasible to perform the 3D-search for the global maximum over the highly nonlinear cost function in (8). However, an approximate method, which is much faster, can be used instead as described below.

When the fundamental frequency is not close to 0 or π (relative to N), a good approximation to the product $\boldsymbol{Z}^H(\omega_0)\boldsymbol{Z}(\omega_0)$ is a scaled identity matrix. That is,

$$\boldsymbol{Z}^H(\omega_0)\boldsymbol{Z}(\omega_0) \approx N \boldsymbol{I}_{2L} \, . \quad (9)$$

This approximation is exact asymptotically in N or if the fundamental frequency is on the Fourier grid $\{2\pi k/N\}_{k=0}^{N-1}$. Under this approximation, we have that

$$\boldsymbol{H}^H(\beta, \xi, \omega_0)\boldsymbol{H}(\beta, \xi, \omega_0) \approx (1+\beta^2) N \boldsymbol{I}_{2L} \quad (10)$$

and this results in that the cost function in (8) can be written as

$$\begin{aligned}
J(\beta, \xi, \omega_0) = \frac{1}{N(1+\beta^2)} \Big[& \boldsymbol{x}_1^H \boldsymbol{Z}(\omega_0) \boldsymbol{Z}^H(\omega_0) \boldsymbol{x}_1 \\
& + \beta^2 \boldsymbol{x}_2^H \boldsymbol{Z}(\omega_0) \boldsymbol{Z}^H(\omega_0) \boldsymbol{x}_2 \\
& + 2\beta \boldsymbol{x}_1^H \boldsymbol{Z}(\omega_0) \boldsymbol{D}^*(\xi) \boldsymbol{Z}^H(\omega_0) \boldsymbol{x}_2 \Big] \, . \quad (11)
\end{aligned}$$

We suggest that this cost function is optimised in the following steps.

1. If all the nonlinear parameters are unknown, an initial estimate of the fundamental frequency can be obtained by using a multi-channel pitch estimator such as the one suggested in [16]. If also the number L of harmonics are unknown, the joint fundamental frequency and model order estimator in [17] can easily be extended to cope with multi-channel data by using the model comparison framework suggested in [18]. If the attenuation β and ξ have been estimated (see the next two steps), the fundamental frequency can be re-estimated by maximising the cost function in (11).
2. When the fundamental frequency is known or has been estimated, the cost function for ξ does not depend on β and reduces to
$$J(\xi) = \boldsymbol{x}_2^H \boldsymbol{Z}(\omega_0) \boldsymbol{D}(\xi) \boldsymbol{Z}^H(\omega_0) \boldsymbol{x}_1 \, . \quad (12)$$
This can be optimised efficiently using an FFT algorithm followed by a 1D line search such as a Fibonacci search.
3. When the fundamental frequency is known or has been estimated and an estimate for ξ has been computed, an estimate for the attenuation parameter is obtained by solving the second order equation
$$\begin{aligned}
0 &= \frac{\partial J(\beta, \xi, \omega_0)}{\partial \beta} \\
&= \beta \left[\boldsymbol{x}_2^H \boldsymbol{Z}(\omega_0) \boldsymbol{Z}^H(\omega_0) \boldsymbol{x}_2 - \boldsymbol{x}_1^H \boldsymbol{Z}(\omega_0) \boldsymbol{Z}^H(\omega_0) \boldsymbol{x}_1 \right] \\
&\quad + (1-\beta^2) \boldsymbol{x}_1^H \boldsymbol{Z}(\omega_0) \boldsymbol{D}^*(\xi) \boldsymbol{Z}^H(\omega_0) \boldsymbol{x}_2 \quad (13)
\end{aligned}$$
for β.

By iterating between the three steps above, an approximate ML estimate can be found. However, we have found that just one iteration gives an acceptable performance, and we have used this setting in the simulation section below.

2.3. An Important Special Case

When the fundamental frequency is set to $\omega_0 = 2\pi/N$ and the number of harmonics is set to $L = \lceil N/2 \rceil - 1$, the approximation in (9) is exact, and the cost function for $\eta = \xi/\omega_0$ does not depend on β and can be written as

$$J(\eta) = \boldsymbol{x}_1^H \boldsymbol{Z}(2\pi/N)\boldsymbol{D}^*(2\pi\eta/N)\boldsymbol{Z}^H(2\pi/N)\boldsymbol{x}_2 = \sum_{k=-\lceil N/2 \rceil+1}^{\lceil N/2 \rceil-1} X_1^*(k)X_2(k)\exp(j2\pi k\eta/N) \quad (14)$$

where $X_1(0) = X_2(0) = 0$. If η is an integer and $X_1(N/2) = X_2(N/2) = 0$ if N is even, the cost function can be written as

$$J(\eta) = \sum_{k=0}^{N-1} X_1^*(k)X_2(k)\exp(j2\pi k\eta/N) \quad (15)$$

which is the cost function of the cross-correlation TDOA estimator [2]. Thus the cross-correlation estimator in (15) is the ML estimator and is therefore an efficient estimator asymptotically if the following conditions are satisfied.

1. The source signal is a periodic signal with zero-mean.
2. The fundamental frequency of the source signal is $2\pi/N$.
3. The number of harmonics of the source signal is $L = \lceil N/2 \rceil - 1$.
4. The delay is an integer value.

For the special case where N is even, $X_1(N/2) \neq 0$, and $X_2(N/2) \neq 0$, the cross-correlation method can be shown to be a suboptimal estimator. In practice, however, anti-aliasing filters will nearly always ensure that $X_1(N/2) \approx X_2(N/2) \approx 0$ so this special case is only of academic interest.

2.4. Fractional TDOA Estimation

The cross-correlation function in (15) can also be derived from the frequency domain model in (2). However, as we alluded to in the introduction, the cost function should not be used for fractional TDOA estimation even though η appears to be a continuous parameter. To demonstrate this, assume that no noise is present so that

$$X_2(k) = X_1(k)\exp(-j2\pi k\eta_0/N), \quad \text{for } k = 0, 1, \ldots, N-1$$

where η_0 is the true delay. If we insert this into (15) and exploit that $X_1(k) = X_1^*(N-k)$, the cost function becomes complex-valued unless $\eta - \eta_0$ is an integer. This is a well-known problem and has typically been addressed by using various interpolation methods [19–21], fractional delay filters [22, 23], and the fractional Fourier transform [24]. However, these heuristic methods can be completely avoided if the cost function in (14) is used instead, since it is real-valued for any delay η.

3. SIMULATIONS

The proposed TDOA estimator (denoted as AML in the rest of this section) was evaluated and compared with other estimators on synthetic data as well as on artificially delayed speech data. This served to experimentally show the differences between the traditional and proposed models for TDOA estimation. The other methods considered in the evaluations were the NLS method proposed in [25][1], and the generalized cross-correlation (GCC) method with unit and phase transform (PHAT) weighting, respectively [4]. These GCC methods have been modified so that their cost functions are written with symmetric indices as in (14) to allow for fractional TDOA estimates, and we refer to them as GCC and GCCP in the rest of this paper. The differences between the AML method and the NLS method are that the NLS method does not make the asymptotic approximation in (9), but assumes that the source is in the far field (i.e., that $\beta = 1$). For both of these methods, the pitch and harmonic model order were estimated using the method [17]. The for generating the results

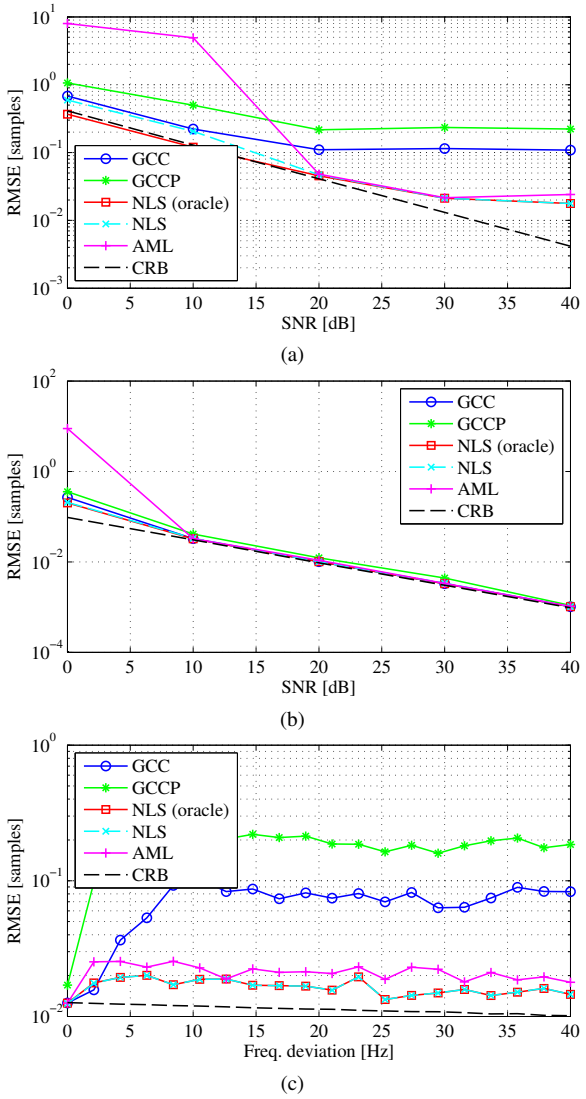

Fig. 1. Performance of the (A)NLS and GCC(P) methods in scenarios with a stereo harmonic signal (a) in different SNRs as well as (c) with different fundamental frequency deviations, and (b) a stereo white Gaussian noise signal at different SNRs.

[1]This method was derived for joint DOA and pitch estimation, but we modified it for TDOA estimation when the pitch is known or estimated.

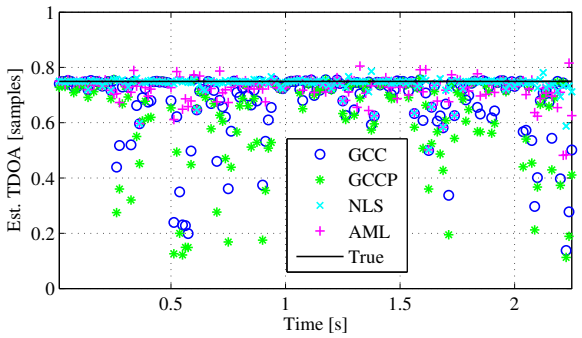

Fig. 2. TDOA estimates of a real speech, stereo signal obtained using the (A)NLS and GCC(P) methods.

presented in this section can be downloaded from `http://kom.aau.dk/~jkn/publications/publications.php`.

Firstly, we describe the evaluation of the methods on synthetic data. In one experiment, the methods were evaluated in a scenario where the signal of interest was 100 samples of a real-valued and periodic source consisting of five harmonics with unit amplitudes and random phases. The fundamental frequency, in radians per sample, was sampled from $\mathcal{U}(0.1, 0.15)$. A synthetic stereo recording was then obtained by generating an additional signal by delaying the signal of interest with approximately 0.6 samples and multiplying it with $\beta = 0.75$. Both signals were observed in white Gaussian noise with a variance corresponding to a certain SNR for the first channel. With this setup, 100 Monte-Carlo simulations were conducted for different SNRs for the first channel, yielding the results in Figure 1a. The labels 'NLS (oracle)' and 'CRB' denotes the NLS method applied with oracle pitch information and the Cramér-Rao bound, respectively. To summarize these results, the NLS yields similar performance (except for low SNRs), no matter if the true or estimated pitch was used. Moreover, the NLS slightly outperformed the ANLS method. The main result, however, is that all of these methods outperformed the GCC(P) methods, relying on the traditional frequency domain model. This clearly illustrates the benefit of using the proposed model. The reason to the floor on the (A)NLS performances for SNRs greater than 30 dB are: the large sample approximation in (9) used in the ANLS method, and the far field assumption used in the NLS method (i.e., that $\beta = 1$).

The next experiment was on a scenario where the signal of interest was a broadband, white Gaussian noise signal (N-periodic). This corresponds to $\omega_0 = 2\pi/N$ radians per sample and a harmonic model order of $N/2-1$. Again, a stereo recording was generated by delaying and attenuating this signal as before. White Gaussian noise was added to each channel at different SNRs for the first channel. With this setup, the results depicted in Figure 1b were produced. In this scenario, all methods yield similar performance and attain the CRB for SNRs above 10 dB. This supports our claim that the widely used frequency domain model in (2) is just a special case of our proposed model for TDOA estimation. It should be noted, however, that real signals are never perfectly N-periodic, and the GCC(P) methods will therefore generally not show this optimum performance in practice. We also note that in scenarios with broadband signals like this, the fundamental frequency is very low and difficult to estimate [26]. Nonetheless, the AML and NLS methods show optimum performance when the fundamental frequency is estimated. The last experiment was again with a harmonic signal as in the first experiment on synthetic data. However, in this experiment the SNR on the first channel was varied while the fundamental frequency was $\frac{2 \cdot f_s}{N}$ Hz plus a varying frequency deviation, where the sampling frequency, f_s, was 8 kHz. Monte Carlo simulations were run for different frequency deviations, producing the results in Figure 1c. When the fundamental frequency is on the N-point frequency grid (i.e., no frequency deviation), the GCC, AML and NLS all yield the same performance and attain the CRB. In the more realistic events of frequency deviations, the performance of all methods decreases for an increasing deviation. Above 5 Hz, the AML and NLS methods clearly outperforms the GCC(P) methods. With no attenuation ($\beta = 1$), the NLS method attains the CRB even with frequency deviations[2], so extending the NLS method with estimation of the attenuation factor, will clearly result in a method outperforming GCC(P), in all cases.

	GCC	**GCCP**	**AML**	**NLS**
RMSE [samples]	0.148	0.201	0.056	0.036

Table 1. RMSEs corresponding to the estimates shown in Figure 2.

We also evaluated the methods on artificially delayed speech data. The data set used here was a female speech signal of the sentence "Why were you away a year, Roy?". To be able to evaluate the accuracy of the obtained TDOA estimates, the stereo recording was generated by delaying this speech signal using a RIR generator [27], such that the true TDOA is approximately 0.75 samples. No reverberation or additional noise was added in this process. The TDOA was estimated using the aforementioned methods over time from blocks of 100 samples, which corresponds to 12.5 ms at a sampling frequency of 8 kHz, from the two channels. This resulted in the estimates shown in Figure 2, corresponding to the RMSEs in Table 1. These result show that the AML and NLS methods clearly outperformed the GCC(P) methods, which produced much more spurious TDOA estimates in this more realistic evaluation scenario. This indicates that the proposed model for TDOA estimation is indeed useful in practice.

4. CONCLUSION

In this paper, we have established the connection between the traditional cross-correlation method and a more general maximum likelihood method in which the fundamental frequency of the periodic signal is modelled as an unknown and continuous parameter. In this connection, we established the four conditions under which the cross-correlation method is a statistically efficient estimator and demonstrated experimentally that significant improvement can be achieved by using the maximum likelihood method instead. The conditions are actually quite restrictive as they require the fundamental frequency of the unknown source signal to be $2\pi/N$, where N is the number of data points, the number of harmonics to be $\lceil N/2 \rceil - 1$, and the delay to be an integer value. We demonstrated how the latter assumption can easily be lifted by using symmetric frequency indices around zero. This followed automatically from the proposed model where the fundamental frequency and the number of harmonics are modelled as unknown parameters. Moreover, the derived approximate likelihood estimator for this model was demonstrated to outperform the cross-correlation method on both synthetic and real-world data.

[2]This was evident from results not presented in this paper.

5. REFERENCES

[1] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays - Signal Processing Techniques and Applications*, M. S. Brandstein and D. B. Ward, Eds., chapter 8, pp. 157–180. Springer-Verlag, 2001.

[2] J. Chen, J. Benesty, and Y. A. Huang, "Time delay estimation in room acoustic environments: An overview," *EURASIP J. on Advances in Signal Process.*, vol. 2006, pp. 1–19, May 2006.

[3] J. Benesty, J. Chen, and Y. A. Huang, *Microphone array signal processing*, Berlin, Germany: Springer-Verlag, 2008.

[4] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 4, pp. 320–327, Aug. 1976.

[5] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, Mar. 1986.

[6] A. Paulraj, R. Roy, and T. Kailath, "Estimation of signal parameters via rotational invariance techniques- ESPRIT," *Rec. Asilomar Conf. Signals, Systems, and Computers*, pp. 83–89, Nov. 1985.

[7] J. P. Dmochowski, J. Benesty, and S. Affes, "Broadband MUSIC: opportunities and challenges for multiple source localization," in *Proc. IEEE Workshop on Appl. of Signal Process. to Aud. and Acoust.* IEEE, 2007, pp. 18–21.

[8] J. C. Chen, R. E. Hudson, and K. Yao, "Maximum-likelihood source localization and unknown sensor location estimation for wideband signals in the near-field," vol. 50, no. 8, pp. 1843–1854, 2002.

[9] Y. Isbi and A. J. Weiss, "DFT model errors for finite length observations with spatially distributed sensors," in *IEEE Conv. Electrical and Electronics Engineers in Isreal*. IEEE, 2008, pp. 080–084.

[10] A. Yeredor, "Analysis of the edge-effects in frequency-domain TDOA estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2012, pp. 3521–3524.

[11] X. Qian and R. Kumaresan, "Joint estimation of time delay and pitch of voiced speech signals," *Rec. Asilomar Conf. Signals, Systems, and Computers*, vol. 1, pp. 735–739, Oct. 1995.

[12] G. Liao, H. C. So, and P. C. Ching, "Joint time delay and frequency estimation of multiple sinusoids," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2001, vol. 5, pp. 3121–3124.

[13] L. Y. Ngan, Y. Wu, H. C. So, P. C. Ching, and S. W. Lee, "Joint time delay and pitch estimation for speaker localization," in *Proc. IEEE Int. Symp. Circuits and Systems*, May 2003, vol. 3, pp. 722–725.

[14] J. X. Zhang, M. G. Christensen, S. H. Jensen, and M. Moonen, "Joint DOA and multi-pitch estimation based on subspace techniques," *EURASIP J. on Advances in Signal Process.*, vol. 2012, no. 1, pp. 1–11, Jan. 2012.

[15] M. Wohlmayr and M. Képesi, "Joint position-pitch extraction from multichannel audio," in *Proc. Interspeech*, Aug. 2007, pp. 1629–1632.

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

[17] J. K. Nielsen, M. G. Christensen, and S. H. Jensen, "Default Bayesian estimation of the fundamental frequency," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 3, pp. 598–610, Mar. 2013.

[18] J. K. Nielsen, M. G. Christensen, A. T. Cemgil, and S. H. Jensen, "Bayesian model comparison with the g-prior," *IEEE Trans. Signal Process.*, vol. 62, no. 1, pp. 225–238, 2014.

[19] G. Jacovitti and G. Scarano, "Discrete time techniques for time delay estimation," *IEEE Trans. Signal Process.*, vol. 41, no. 2, pp. 525–533, Feb. 1993.

[20] M. M. McCormick and T. Varghese, "An approach to unbiased subsample interpolation for motion tracking," *Ultrasonic Imaging*, vol. 35, no. 2, pp. 76–89, 2013.

[21] F. Viola and W. F. Walker, "A spline-based algorithm for continuous time-delay estimation using sampled data," *IEEE Trans. Ultrason., Ferroelectr., Freq. Control*, vol. 52, no. 1, pp. 80–93, 2005.

[22] M. Olsson, *Contributions to delay, gain, and offset estimation*, Ph.D. thesis, Linköping University, 2008.

[23] D. L. Maskell and G. S. Woods, "The estimation of subsample time delay of arrival in the discrete-time measurement of phase delay," *IEEE Trans. Instrum. Meas.*, vol. 48, no. 6, pp. 1227–1231, 1999.

[24] K. K. Sharma and S. D. Joshi, "Time delay estimation using fractional Fourier transform," *Elsevier Signal Processing*, vol. 87, no. 5, pp. 853–865, 2007.

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

[26] M. G. Christensen, "Accurate estimation of low fundamental frequencies from real-valued measurements," vol. 21, no. 10, pp. 2042–2056, Oct. 2013.

[27] E. A. P. Habets, "Room impulse response generator," Tech. Rep., Technische Universiteit Eindhoven, 2010, Ver. 2.0.20100920.