AN EXACT SUBSPACE METHOD FOR FUNDAMENTAL FREQUENCY ESTIMATION

Mads Græsbøll Christensen

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

ABSTRACT

In this paper, an exact subspace method for fundamental frequency estimation is presented. The method is based on the principles of the MUSIC algorithm, wherein the orthogonality between the signal and and noise subspace is exploited. Unlike the original MUSIC algorithm, the new method uses an exact measure of the angles between the subspaces. This makes a difference, for example, when the fundamental frequency is low, for real signals, or when the number of samples is low. In Monte Carlo simulations, the performance of the new method is compared to a number of state-of-the-art methods and is demonstrated to lead to improvements in certain, critical cases. Moreover, it is demonstrated on a speech signal that the method can be applied to speech signals and is robust towards noise.

Index Terms— Speech analysis, fundamental frequency estimation, pitch estimation, subspace methods

1. INTRODUCTION

Many signals of interest to mankind are periodic or approximately so. This is, for example, the case for voiced speech. Such signals can be decomposed into sums of sinusoids whose frequencies are integer multiples of a fundamental frequency and the problem of finding this fundamental frequency, sometimes also referred to as pitch estimation, is the topic of the present paper. Pitch estimation is an important topic in speech processing as it has a multitude of different applications, including separation [1], localization [2], dereverberation [3], feedback cancellation [4], diagnosis of illnesses [5], and detection of emotion and stress [6]. As a result, a host of different methods have been proposed over the years for solving this important problem, e.g., [7–15], and we refer the interested reader to [16] for an overview. Many of these methods are, implicitly or explicit, based on asymptotic approximations and this causes trouble in certain situations. This is the case for a low number of observations, for low fundamental frequencies and for real signals.

In this paper, we propose a new method for dealing with these problems in the context of fundamental frequency estimation. It is a subspace method based on the orthogonality between the signal and noise subspaces, a principle known from the classical MUSIC algorithm [17]. Unlike the MUSIC algorithm and its adaptation to real signals [18], the proposed method method is based on an exact measure of the angles between subspaces [19–21]. It is generally not feasible to employ such exact measures in unconstrained frequency estimation with several nonlinear parameters. However, it is well-suited for the fundamental frequency estimation problem as it only involves one nonlinear parameter.

The remaining part of this paper is organized as follows: In Section 2, we introduce the basic signal model, the underlying assumptions and define the problem at hand. Then, in Section 3 the proposed method is presented. We then investigate the performance of the proposed method under various conditions and compare it to a number of state-of-the-art methods in Section 4. Finally, we conclude on our work in Section 5.

2. COVARIANCE MATRIX MODEL

We will now introduce the problem at hand and the signal model. The observed real signal $x(n)$ is comprised of L sinusoidal components having frequencies that are integer multiples of a fundamental frequency ω_0, real amplitude $A_l > 0$, and phases $\phi_l \in [0, 2\pi)$. Moreover, we assume that an additive noise source $e(n)$ is present, which is here assumed to be white with variance σ^2. The signal model can be expressed for $n = 0, \ldots, N-1$ as

$$x(n) = \sum_{l=1}^{L} A_l \cos\left(\omega_0 l n + \phi_l\right) + e(n). \tag{1}$$

The problem at hand is then to estimate ω_0, which, for a given L, can be in the range $\omega_0 \in (0, \frac{\pi}{L})$. For a collection of samples $\{x(n)\}$, the model above can be expressed as

$$\mathbf{x}(n) = \mathbf{Z}\mathbf{a} + \mathbf{e}(n), \tag{2}$$

with the following definitions:

$$\mathbf{x}(n) = \left[\, x(n)\; x(n+1)\; \cdots\; x(n+N-1) \,\right]^T \tag{3}$$

$$\mathbf{Z} = \left[\, \mathbf{z}(\omega_0)\; \mathbf{z}^*(\omega_0)\; \cdots\; \mathbf{z}(\omega_0 L)\; \mathbf{z}^*(\omega_0 L) \,\right], \tag{4}$$

This work was supported by the Villum Foundation.

$$\mathbf{a} = \frac{1}{2}\left[A_1 e^{j\phi_1}\ A_1 e^{-j\phi_l}\ \cdots\ A_L e^{j\phi_L}\ A_L e^{-j\phi_L}\ \right]^T \quad (5)$$

$$\mathbf{z}(\omega) = \left[\ 1\ e^{j\omega 1}\ \cdots\ e^{j\omega l(N-1)}\ \right]^T \quad (6)$$

$$\mathbf{e}(n) = [\ e(n)\ e(n+1)\ \cdots\ e(n+M-1)\ ]^T. \quad (7)$$

The covariance matrix of $\mathbf{x}(n)$ is [18]

$$\mathbf{R} = \mathrm{E}\left\{\mathbf{x}(n)\mathbf{x}^H(n)\right\} = \mathbf{Z}\mathbf{P}\mathbf{Z}^H + \sigma^2\mathbf{I} \quad (8)$$

where $(\cdot)^H$ denotes the Hermitian transpose, and the amplitude covariance matrix $\mathrm{E}\left\{\mathbf{a}\mathbf{a}^H\right\} = \mathbf{P}$ is given by

$$\mathbf{P} = \mathrm{E}\left\{\begin{bmatrix} a_1a_1^* & a_1^*a_1^* & \dots & a_1a_L^* & a_1^*a_L^* \\ a_1a_1 & a_1^*a_1 & \dots & a_1a_L & a_1^*a_L \\ \vdots & \vdots & & \vdots & \vdots \\ a_La_1^* & a_L^*a_1^* & \dots & a_La_L^* & a_L^*a_L^* \\ a_La_1 & a_L^*a_1 & \dots & a_La_L & a_L^*a_L \end{bmatrix}\right\}. \quad (9)$$

Assuming that the phases ϕ_l are uniformly distributed and independent over l we have that $\mathrm{E}\left\{\frac{A_k}{2}e^{j\phi_k}\right\} = 0$ and that $\mathrm{E}\left\{\frac{A_k}{2}e^{j\phi_k}\frac{A_l}{2}e^{-j\phi_l}\right\} = \frac{A_k}{2}\mathrm{E}\left\{e^{j\phi_k}\right\}\frac{A_l}{2}\mathrm{E}\left\{e^{-j\phi_l}\right\} = 0$ for $k \neq l$. Moreover, for $k = l$ we get that $\mathrm{E}\left\{\frac{A_k}{2}e^{j\phi_k}\frac{A_k}{2}e^{-j\phi_k}\right\} = \frac{A_k^2}{4}$. Therefore, the amplitude covariance matrix $\mathbf{P}$ becomes $\mathbf{P} = \frac{1}{4}\mathrm{diag}\left(\left[\ A_1^2\ A_1^2\ \cdots\ A_L^2\ A_L^2\ \right]\right)$, which means that the diagonal structure obtained for complex signals is retained for real signals, and the so-called covariance matrix model, therefore, still holds.

The eigenvalue decomposition (EVD) of the covariance matrix is $\mathbf{R} = \mathbf{U}\mathbf{\Gamma}\mathbf{U}^H$, where $\mathbf{\Gamma}$ is a diagonal matrix containing the positive eigenvalues, γ_k, ordered as $\gamma_1 \geq \gamma_2 \geq \ldots \geq \gamma_M$. Moreover, it can easily be seen that $\gamma_{2L+1} = \ldots = \gamma_M = \sigma^2$. $\mathbf{U}$ contains the M orthonormal eigenvectors of $\mathbf{R}$, i.e., $\mathbf{U} = \left[\ \mathbf{u}_1\ \ \cdots\ \ \mathbf{u}_M\ \right]$. Let $\mathbf{S}$ be formed from a subset of the columns of this matrix as $\mathbf{S} = \left[\ \mathbf{u}_1\ \ \cdots\ \ \mathbf{u}_{2L}\ \right]$. We denote the subspace spanned by the columns of $\mathbf{S}$ as $\mathcal{S} = \mathcal{R}(\mathbf{S})$ and refer to it as the signal subspace. Similarly, let $\mathbf{G}$ be formed from the remaining eigenvectors as $\mathbf{G} = \left[\ \mathbf{u}_{2L+1}\ \ \cdots\ \ \mathbf{u}_M\ \right]$. We refer to the space $\mathcal{G} = \mathcal{R}(\mathbf{G})$ as the noise subspace. Using these definitions, we now obtain $\mathbf{U}\left(\mathbf{\Gamma} - \sigma^2\mathbf{I}\right)\mathbf{U}^H = \mathbf{Z}\mathbf{P}\mathbf{Z}^H$. Introducing $\mathbf{\Psi} = \mathrm{diag}([\ \gamma_1 - \sigma^2\ \cdots\ \gamma_{2L} - \sigma^2\ ])$, this leads to the following partitioning of the EVD:

$$\mathbf{R} = \left[\ \mathbf{S}\ \ \mathbf{G}\ \right]\left(\begin{bmatrix}\mathbf{\Psi} & \mathbf{0} \\ \mathbf{0} & \mathbf{0}\end{bmatrix} + \sigma^2\mathbf{I}\right)\begin{bmatrix}\mathbf{S}^H \\ \mathbf{G}^H\end{bmatrix}, \quad (10)$$

which shows that we may write $\mathbf{S}\mathbf{\Psi}\mathbf{S}^H = \mathbf{Z}\mathbf{P}\mathbf{Z}^H$. As the columns of $\mathbf{S}$ and $\mathbf{G}$ are orthogonal and $\mathcal{R}(\mathbf{Z}) = \mathcal{R}(\mathbf{S})$, it follows that $\mathbf{Z}^H\mathbf{G} = \mathbf{0}$, which is the subspace orthogonality principle used in the MUSIC algorithm [17].

3. PROPOSED METHOD

In practice, the estimated noise subspace eigenvectors will not be perfect due to the observation noise and finite observation lengths, and the above relation is, therefore, only approximate. A measure must then be introduced to determine how close a candidate model $\mathbf{Z}$ is to being orthogonal to $\mathbf{G}$. Traditionally, this has been done using the Frobenius norm [17]. However, this measure is only an accurate measure of the angles between the two spaces for orthogonal vectors in both $\mathbf{Z}$ and $\mathbf{G}$, and, the asymptotic orthogonality of the column of $\mathbf{Z}$ may not always be accurate. Instead, we propose to measure the orthogonality as follows. The principal angles $\{\xi_k\}$ between the two subspaces $\mathcal{Z}$ and $\mathcal{G}$ having projection matrices $\mathbf{\Pi}_Z$ and $\mathbf{\Pi}_G$, are defined recursively for $k = 1, \ldots, K$ as [20]

$$\cos(\xi_k) = \max_{\mathbf{y}}\max_{\mathbf{z}}\frac{\mathbf{y}^H\mathbf{\Pi}_Z\mathbf{\Pi}_G\mathbf{z}}{\|\mathbf{y}\|_2\|\mathbf{z}\|_2} \quad (11)$$

$$\triangleq \mathbf{y}_k^H\mathbf{\Pi}_Z\mathbf{\Pi}_G\mathbf{z}_k = \kappa_k. \quad (12)$$

where K is the minimal dimension of the two subspaces, i.e., $K = \min\{2L, M-2L\}$ and $\mathbf{y}^H\mathbf{y}_i = 0$ and $\mathbf{z}^H\mathbf{z}_i = 0$ for $i = 1, \ldots, k-1$. This results in a set of angles that are bounded and ordered, i.e., $0 \leq \xi_1 \leq \ldots \leq \xi_K \leq \frac{\pi}{2}$. As can be seen, $\{\kappa_k\}$ are the ordered singular values of the matrix product $\mathbf{\Pi}_Z\mathbf{\Pi}_G$, and the two sets of vectors $\{\mathbf{y}\}$ and $\{\mathbf{z}\}$ are the left and right singular vectors of the matrix product, respectively. The singular values are related to the Frobenius norm of $\mathbf{\Pi}_Z\mathbf{\Pi}_G$ and hence its trace, denoted with $\mathrm{Tr}\{\cdot\}$, as $\|\mathbf{\Pi}_Z\mathbf{\Pi}_G\|_F^2 = \sum_{k=1}^K \kappa_k^2$. If this Frobenius norm is zero, then the non-trivial angles are all $\frac{\pi}{2}$, i.e., the two subspaces are orthogonal. We can use this expression to find the fundamental frequency as

$$\hat{\omega}_0 = \arg\min_{\omega_0}\|\mathbf{\Pi}_Z\mathbf{\Pi}_G\|_F^2. \quad (13)$$

Finally, (13) can be expressed as

$$\hat{\omega}_0 = \arg\min_{\omega_0}\mathrm{Tr}\left\{\mathbf{Z}\left(\mathbf{Z}^H\mathbf{Z}\right)^{-1}\mathbf{Z}^H\mathbf{G}\mathbf{G}^H\right\}. \quad (14)$$

We henceforth refer to this estimator as the angles between subspaces (ABS) method. Interestingly, it is asymptotically equivalent to the estimator proposed in [22] but is different for finite M and N in that it takes the possible non-orthogonality of the sinusoids into account. The estimator requires that a number of quantities are computed as initialization, i.e., only once, namely the EVD of $\mathbf{R}$ and the projection matrix for the noise subspace, which results in a complexity of $\mathcal{O}((M-L)M^2 + M^3)$ with $L < M$. For each candidate fundamental frequency, operations having complexity $\mathcal{O}(L^2M + M^2L + L^3)$ are computed. Regarding the covariance matrix, we use the sample covariance matrix and note that it is not required for this method that its estimate has full rank. It must, however, allow for estimation of a basis for the signal subspace, which requires that $M \leq N - 2L + 1$. Furthermore, we requre that $M \geq 2L + 1$ for the orthogonal complement to the signal subspace to be non-empty, which means that we obtain that $2L+1 \leq M \leq N-2L+1$. Moreover, M should be chosen proportionally to N for the method to be consistent.

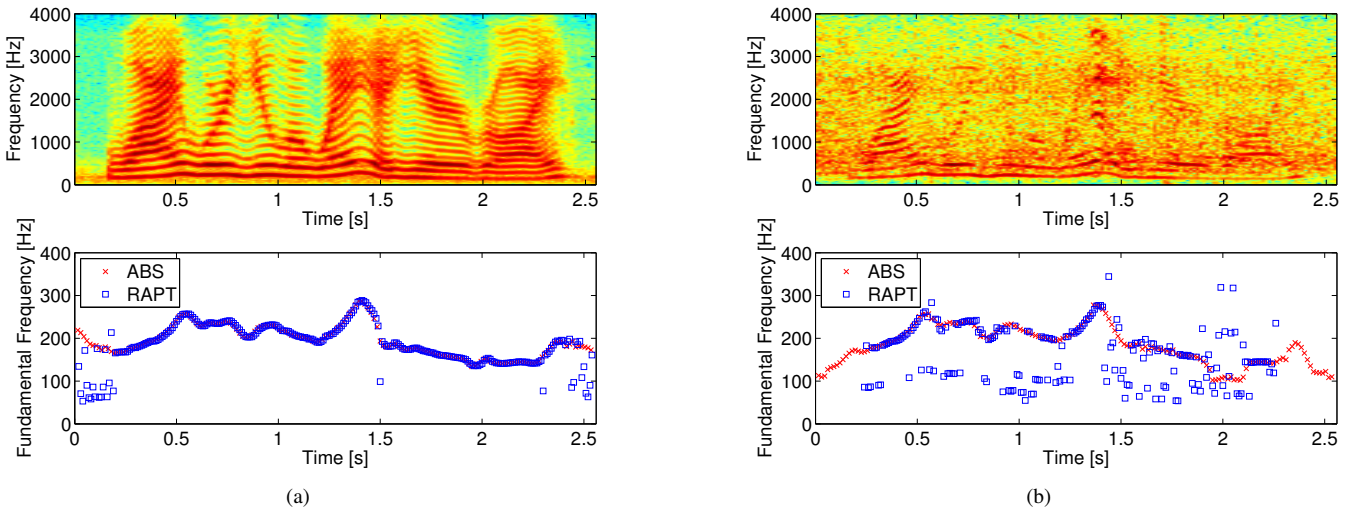

Fig. 1. Real speech signal example. Shown are the spectrograms of the involved signals and the estimates obtained using the proposed method ABS and RAPT [11] for the clean signal (a) and with exhibition hall noise added at 0 dB SNR (b).

4. EXPERIMENTAL RESULTS

We will start out by demonstrating the applicability of the proposed estimator to speech signals. To do so, we run the estimator on a clean speech signal sampled at 8 kHz and compare the result to those obtained with the Robust Algorithm for Pitch Tracking (RAPT) [11]. In this experiment, the implementation of RAPT from Voicebox [23] is used with standard settings. The proposed method is used with segments of size $N = 240$ and $M = 120$, and pitches are estimated in the range 60 Hz to 400 Hz. The model order is estimated using the principle of [21], which can be integrated into the proposed method, and, as in RAPT, the estimates are smoothed using [24]. The spectrogram of the clean speech signal is shown in the top panel of Figure 1(a) while the obtained estimates are shown in the bottom panel. As can be seen, both estimators estimate the pitch well with only a few errors. In Figure 1(b), the same is shown, only noise, here the exhibition hall noise from the NOIZEUS corpus [25], has now been added to the speech signal at 0 dB signal-to-noise ratio (SNR). It can be seen that RAPT now performs very poorly while the proposed estimator still works well, despite the poor SNR.

Next, the proposed method is compared to a number of other estimators using Monte Carlo simulations by generating signals according to (1) and then applying various estimators to those signals. The so-obtained parameter estimates are then compared to the true parameters and the error is measured in terms of the mean square error (MSE). We compare the proposed method (which, as mentioned, is referred to as ABS) to the weighted least-squares (WLS) method of [9], the approximate nonlinear least-squares (ANLS) method [7, 8, 16], and the optimal filtering method (OPTFILT) [16] and the MUSIC-based method of [22]. Regarding the MUSIC-based method, the proposed method should outperform it under adverse conditions and they should perform the same for high N and M. For each set of experimental conditions, 100 realizations are used and the Cramér-Rao Lower Bound (CRLB) shown is the average over the exact CRLB. The signals were generated with the following parameters, except for the parameters that are varied: a fundamental frequency with $\omega_0 = 0.3129$ is used with five harmonics each having unit amplitude and phases uniformly distributed between $-\pi$ and π. Segments of $N = 100$ samples were used with $M = 50$ and white Gaussian noise added at a 40 dB SNR. Note that this is the SNR for the fundamental frequency estimation problem as defined in [16]. The high SNR is used so that the noise will not be the limiting factor but rather the asymptotic approximations. The results are shown in Figures 2(a), 2(b), 2(c), and 2(d) in terms of the MSE as a function of N, ω_0, the SNR and M. From the figures, a number of interesting observations can be made. Firstly, it can be seen from Figure 2(a) that all methods perform well for a high number of observations, N, except the ANLS method which does not perform well at all. It can also be observed that the methods exhibit different threshold behavior, but the ABS and MUSIC methods perform similarly here. This is, however, not the case when the MSE is observed as a function of the fundamental frequency, as shown in Figure 2(b). From this figure, it can be seen that the MUSIC method is indeed improved by avoiding the approximate measure of orthogonality as is done in the ABS method. In fact, the ABS method now performs as well as any of the other methods. This clearly shows that, as claimed, the exact measure is preferable when dealing with non-orthogonal sinusoids. In Figure 2(c), the MSE is depicted as a function of the SNR. This figure shows that the subspace methods appear to hold an advantage over the WLS and OPTFILT methods

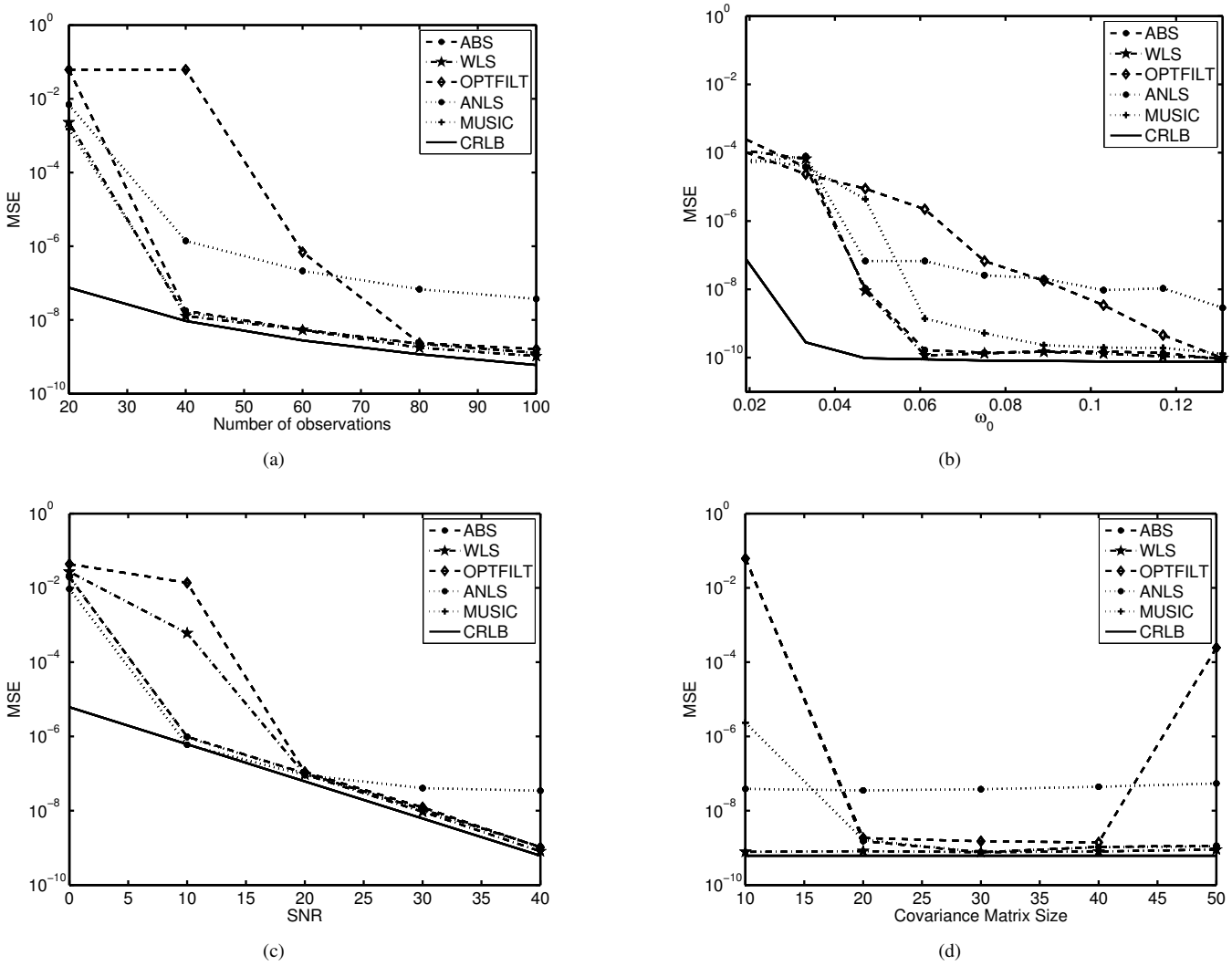

Fig. 2. Performance measured in terms of the mean square estimation error (MSE) as a function of (a) the number of samples, N, and (b) the fundamental frequency, ω_0, (c) the SNR, and (d) the covariance matrix size M.

in terms of being robust towards noise. In this case, it does, though, not appear to matter whether the exact measure of the ABS method or the approximate one of MUSIC is used. Finally, the performance is assessed as a function of M, the covariance matrix size, with the results being shown in Figure 2(d). It can be seen that as long as M is chosen not to low or too high, its value does appear to be all that critical, although this may be different for different fundamental frequencies.

5. DISCUSSION

In this paper, a new method for fundamental frequency estimation has been presented. The method, which is a subspace method, avoids the commonly used asymptotic approximations of other methods, including also the classical MUSIC algorithm [17], its real extension [18] and the harmonic summation method [7], which forms the basis of many state-of-the-art methods. Instead, the method is based on an exact measure of the angles between subspaces. In simulations, the method was demonstrated to outperform its approximate counterpart for low fundamental frequencies, a situation where the aforementioned asymptotic approximations become inaccurate. The experiment involving a noisy speech signal also clearly demonstrated how the method is robust to noise compared to the RAPT method [11]. An added benefit of the proposed method is that the principle used herein can be extended using [21] to account also for an unknown number of harmonics. While the proposed method is computationally more intensive than its predecessors, its improved performance under adverse conditions and generally more accurate estimates may prove beneficial in certain, critical applications, like diagnosis of illnesses or speech separation.

6. REFERENCES

[1] M. Stark, M. Wohlmayr, and F. Pernkopf, "Source-filter based single channel speech separation using pitch information," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 19(2), pp. 242–255, 2011.

[2] M. S. Brandstein, "Time-delay estimation of reverberated speech exploiting harmonic structure," *J. Acoust. Soc. Am.*, vol. 105(5), pp. 2914–2919, 1999.

[3] T. Nakatani, K. Kinoshita, and M. Miyoshi, "Harmonicity-based blind dereverberation for single-channel speech signals," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15(1), pp. 80–95, 2007.

[4] K. Ngo, T. van Waterschoot, M. G. Christensen, M. Moonen, S. H. Jensen, and J. Wouters, "Adaptive feedback cancellation in hearing aids using a sinusoidal near-end model," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, 2010, pp. 181–184.

[5] A. Tsanas, M. A. Little, P. E. McSharry, J. Spielman, and L.O. Ramig, "Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease," *IEEE Trans. Biomed. Eng.*, vol. 59(5), pp. 1264–1271, 2012.

[6] G. Zhou, J. H. L. Hansen, and J. F. Kaiser, "Nonlinear feature based classification of speech under stress," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 9(3), pp. 201–216, 2001.

[7] 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.

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

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

[10] 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, 2002.

[11] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds., chapter 5, pp. 495–518. Elsevier Science B.V., 1995.

[12] A. Nehorai and B. Porat, "Adaptive comb filtering for harmonic signal enhancement," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 34(5), pp. 1124–1138, Oct. 1986.

[13] Z. Jin and D. Wang, "HMM-based multipitch tracking for noisy and reverberant speech," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 19(5), pp. 1091–1102, 2011.

[14] J. Tabrikian, S. Dubnov, and Y. Dickalov, "Maximum a posteriori probability pitch tracking in noisy environments using harmonic model," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 12(1), pp. 76–87, 2004.

[15] M. Goto, "A real-time music-scene-description system: predominant-f0 estimation for detecting melody and bass lines in real-world audio signals," *Speech Communication*, vol. 43, pp. 311–329, 2004.

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

[17] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34(3), pp. 276–280, Mar. 1986.

[18] P. Stoica and A. Eriksson, "MUSIC estimation of real-valued sine-wave frequencies," *Signal Processing*, vol. 42, pp. 139–146, 1995.

[19] R. T. Behrens and L. L. Scharf, "Signal processing applications of oblique projection operators," *IEEE Trans. Signal Process.*, vol. 42(6), pp. 1413–1424, 1994.

[20] G. H. Golub and C. F. V. Loan, *Matrix Computations*, The Johns Hopkins University Press, 3rd edition, 1996.

[21] M. G. Christensen, A. Jakobsson, and S. H. Jensen, "Sinusoidal order estimation using angles between subspaces," *EURASIP J. on Advances in Signal Processing*, pp. 1–11, 2009, Article ID 948756.

[22] M. G. Christensen, A. Jakobsson and S. H. Jensen, "Joint high-resolution fundamental frequency and order estimation," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15(5), pp. 1635–1644, 2007.

[23] M. Brookes, "VOICEBOX: A speech processing toolbox for MATLAB," MATLAB toolbox, 2006, Imperial College, London, UK.

[24] H. Ney, "Dynamic programming algorithm for optimal estimation of speech parameter contours," *IEEE Trans. Syst., Man, Cybern.*, vol. 13(3), pp. 208–214, 1983.

[25] Y. Hu and P. Loizou, "Subjective evaluation and comparison of speech enhancement algorithms," *Speech Communication*, vol. 49, pp. 588–601, 2007.