# Formant frequency estimation of high-pitched vowels using weighted linear prediction[a)]

Paavo Alku[b)] and Jouni Pohjalainen
*Department of Signal Processing and Acoustics, Aalto University, P.O. Box 13000, FI-00076 Aalto, Finland*

Martti Vainio
*Institute of Behavioural Sciences, SigMe Group, University of Helsinki, P.O. Box 9, FI-00014 Helsinki, Finland*

Anne-Maria Laukkanen
*Department of Speech Communication and Voice Research, University of Tampere, Fi-33014 Tampere, Finland*

Brad H. Story
*Speech Acoustics Laboratory, University of Arizona, Tucson, Arizona 85721*

(Received 2 January 2012; revised 12 March 2013; accepted 10 June 2013)

All-pole modeling is a widely used formant estimation method, but its performance is known to deteriorate for high-pitched voices. In order to address this problem, several all-pole modeling methods robust to fundamental frequency have been proposed. This study compares five such previously known methods and introduces a technique, Weighted Linear Prediction with Attenuated Main Excitation (WLP-AME). WLP-AME utilizes temporally weighted linear prediction (LP) in which the square of the prediction error is multiplied by a given parametric weighting function. The weighting downgrades the contribution of the main excitation of the vocal tract in optimizing the filter coefficients. Consequently, the resulting all-pole model is affected more by the characteristics of the vocal tract leading to less biased formant estimates. By using synthetic vowels created with a physical modeling approach, the results showed that WLP-AME yields improved formant frequencies for high-pitched sounds in comparison to the previously known methods (e.g., relative error in the first formant of the vowel [a] decreased from 11% to 3% when conventional LP was replaced with WLP-AME). Experiments conducted on natural vowels indicate that the formants detected by WLP-AME changed in a more regular manner between repetitions of different pitch than those computed by conventional LP. 
[http://dx.doi.org/10.1121/1.4812756]

PACS number(s): 43.70.Jt, 43.72.Ar [CHS] Pages: 1295–1313

## I. INTRODUCTION

Resonances of the vocal tract, the formants, are parameters of fundamental importance in all areas of speech science and technology. Therefore, many methods have been developed during the past decades to estimate formant frequencies from speech signals. In this area, the algorithms exploiting linear prediction (LP) have especially gained momentum. LP refers to a family of parametric spectral estimation techniques that represent the envelope of the power spectrum of discrete-time signals using an all-pole digital filter structure (Makhoul, 1975a; Markel and Gray, 1976). The name LP refers to the optimization of the all-pole filter coefficients: each time domain signal sample is predicted as a linear combination of a given number of previous samples. The optimal coefficients are defined, in the case of conventional LP, by searching for the filter taps that minimize the square of the prediction error, the residual.

LP is well-suited for formant estimation due to its close connection to the source-filter theory of speech production (Fant, 1970; Markel and Gray, 1976). Since formants are spectral regions of large energy, they are weighted most heavily in the error criterion of LP and thus represented most accurately. LP enables computing a smoothed all-pole fit, determined by a small number of parameters, to the power spectrum of speech. Consequently, formants can be easily identified from the all-pole spectra by using, for example, simple peak picking. In addition, all-pole models computed by the most well-known form of linear predictive techniques, conventional LP analysis based on the autocorrelation method, are guaranteed to be stable and the analysis can be performed non-iteratively with high computational efficiency.

Even though conventional LP analysis has several benefits, its formant estimation performance is known to suffer from inaccuracies when analyzing high-pitched speech (Makhoul, 1975a,b; El-Jaroudi and Makhoul, 1991). In particular, the estimates of the lowest formants are biased by the pitch harmonics. This degradation in accuracy of LP is due to the least squares error criterion used in conventional

[a)]Portions of this work were presented in "Improved formant frequency estimation from high-pitched vowels by downgrading the contribution of the glottal source with weighted linear prediction," Proceedings of Interspeech, Portland, OR, September 2012.

[b)]Author to whom correspondence should be addressed. Electronic mail: paavo.alku@aalto.fi

LP. High-pitched voices are characterized by a relatively large concentration of the sound energy at the fundamental frequency (F0) and its few lowest integer multiples. The all-pole LP model optimized by the least squares criterion favors these high-energy regions and, consequently, might not find a good spectral match for formants. In addition to this general frequency domain explanation, the degradation of LP in the formant estimation of high-pitched voices can be explained by aliasing which takes place in the autocorrelation domain (El-Jaroudi and Makhoul, 1991).

Modifications to conventional LP have been proposed in many studies in order to compute all-pole models that are less affected by F0. Hermansky *et al.* (1984) utilized a frequency domain approach, in which interpolation was used to attenuate the contribution of harmonics from a voiced speech spectrum. Miyoshi *et al.* (1987) proposed a sample-selective linear predictive algorithm based on solving a set of over-determined LP equations using two stages. The first stage, conventional LP, discards those prediction equations that yield large residuals and the second stage employs only the prediction equations that give relatively small prediction errors. Lee (1988) studied different cost functions that give more weight to small residual samples while down-weighting the prediction error samples of large amplitude. By utilizing maximum-likelihood-type estimation together with Huber's psi-function, Lee (1988) proposed the iterative Robust Linear Prediction (RBLP) algorithm.

El-Jaroudi and Makhoul (1991) proposed a modification to conventional LP, an algorithm named Discrete All-pole Modeling (DAP). DAP utilizes a discrete version of the Itakura-Saito distortion measure as the error criterion, which is implemented iteratively in order to solve the optimal filter coefficients. The DAP algorithm by El-Jaroudi and Makhoul (1991) has since been utilized in several studies (e.g., Fröhlich *et al.*, 2001; Alku *et al.*, 2006) to model the vocal tract of high-pitched speech. Ma *et al.* (1993) studied an approach which is similar to the work of both Miyoshi *et al.* (1987) and Lee (1988). Their method, Weighted Linear Prediction (WLP), is based on introducing a temporal weighting function by which the square of the prediction residual is multiplied. With this weighting function, certain residual samples can be assigned a larger importance in the least-squares optimization of the filter coefficients. Rahman and Shimamura (2007) suggested a method that employs homomorphic deconvolution in the autocorrelation domain. More specifically, their method, entitled Linear Prediction using Refined Autocorrelation (LPRA), transforms the speech signal into the cepstral domain in which the contributions of the voice source and vocal tract are separated by utilizing a straightforward truncation. In contrast to many other modifications of conventional LP, the method proposed by Rahman and Shimamura (2007) is guaranteed to result in stable all-pole models.

Finally, formant estimation of high-pitched speech by exploiting temporal change of pitch was recently studied by Wang and Quatieri (2010). Their work is motivated by the fact that the traditional source-filter model can be considered a sampling process in which the formant envelope is sampled by harmonics of the periodic source signal. As the pitch increases, the harmonic sampling becomes more sparse resulting in poor formant estimates. However, by assuming that pitch changes, the spectral sampling of the underlying formant envelope can be improved. Based on this motivation, Wang and Quatieri (2010) proposed two realizations of a two-dimensional analysis framework to address the formant estimation from high-pitched speech. A similar approach has also been used by Shiga and King (2003). It is, though, worth emphasizing that neither of these methods can be used if formants of a single speech frame with constant pitch are to be estimated.

In the present investigation, linear predictive analyses of different types are used to study formant frequency estimation of high-pitched speech. Formant tracking is a procedure in which formant contours are computed, either automatically or semi-automatically, over a time span that typically covers one word or sentence and formant estimates from several consecutive frames are combined. In the past four decades, several formant tracking methods have been developed utilizing, for example, the Newton-Raphson iteration (Olive, 1971), amplitude and frequency modulation (Potamianos and Maragos, 1996), and adaptive Kalman filtering (Deng *et al.*, 2007). Rather than studying formant tracking, this investigation focuses on how accurately a set of selected linear predictive methods are able to estimate vowel formants over a single data frame. As shown by the literature review above, this topic has been widely covered by several studies published in the past decades. Despite the wealth of articles on the topic, there are, however, issues in the study area which are not covered in the previous investigations. Therefore, launching the present study is justified and motivated by the following three rationales.

First, it will be shown that the weighted LP algorithm proposed by Ma *et al.* (1993) can be combined with a new straightforward weighting function. With this new weighting, the present study indicates that the performance of WLP in formant frequency estimation improves considerably.

Second, previous articles typically compare the performance of the underlying algorithm with conventional LP. Larger comparisons between different modified LP methods, however, are limited and they typically involve comparison with DAP, as used, for example, by Rahman and Shimamura (2007), or with LPRA as used in the study by Wang and Quatieri (2010). In contrast, the comparison in the present study includes a larger number of potential linear predictive algorithms that have been proposed for formant estimation of high-pitched speech.

Third, performance of a formant estimation method is typically assessed by using synthetic vowels produced by some form of source-filter modeling. This kind of evaluation, however, might not be truly objective because the test material and the methods to be assessed are based on similar models of human voice production. Therefore, the current study takes advantage of a different strategy in performance evaluation of the selected linear predictive methods. The idea is to use *physical modeling* of the vocal folds and the vocal tract in order to synthesize vowels with known formant frequencies. This approach is different from the one where conventional source-filter synthesis is used, because the test signals are generated by a physical law, rather than by a parametric digital model similar to the all-pole model assumed in LP.

Computation of all-pole models by utilizing temporal weighting of the squared residual plays an essential role in the current study. Therefore, the next section is devoted solely to WLP, which is first described as proposed by Ma *et al.* (1993). In the same section, a novel modification is introduced to WLP in order to make formant estimation less sensitive to the biasing effects of F0. Described in Sec. III are the linear predictive methods that were selected for comparison in the present investigation. The speech material and methods of quantifying formant estimation accuracy are described in Secs. IV and V, respectively. The experiments conducted are described in Sec. VI and their results reported in VII, separately for synthetic and natural speech. Finally, discussion and conclusions of the study are given in Secs. VIII and IX, respectively.

## II. WEIGHTED LINEAR PREDICTION

### A. Formulation and optimization

WLP is a linear predictive method for computing all-pole models of speech by temporally weighting the square of the residual in the optimization of the model parameters. The idea in temporal weighting is to emphasize the contribution of certain pre-selected time samples as a part of the minimum squares type of optimization while de-emphasizing those speech samples that are considered to contain data that is less desirable for the modeling task in question. The samples whose contribution is to be de-emphasized might, for example, be corrupted by environmental noise. Another corruption, which combines WLP directly to the topic of the present study, is the excitation of the vocal tract, the glottal volume velocity waveform. This source signal of speech is definitely needed in order to excite the resonances of the vocal tract and thereby to create the acoustic speech pressure signal. However, the rapid fluctuation of the vocal folds in high-pitched speech increases the number of vocal tract excitation events per time unit. This strengthened activity of the source causes all-pole modeling to be too focused on the excitation rather than on the vocal tract filter. Consequently, all-pole models computed by LP become affected by the glottal source resulting in poor accuracy in formant estimation.

Following the notations used by Ma *et al.* (1993), the mathematical derivation of the WLP model can be expressed as follows. The residual energy of the $p$th order WLP model can be written as

$$E = \sum_{n=n_1}^{n_2} e_n^2 \cdot W_n = \sum_{n=n_1}^{n_2} \left( s_n - \sum_{k=1}^{p} a_k s_{n-k} \right)^2 \cdot W_n, \qquad (1)$$

where $e_n$ is the residual, $W_n$ is the temporal weighting function, $s_n$ is the speech signal to be modeled, and $a_k\,(1 \le k \le p)$ are the predictor coefficients. The residual energy is minimized in the time span between $n_1$ and $n_2$. In the case of the autocorrelation method, $n_1 = 1$ and $n_2 = N + p$, and the speech signal is assumed to be zero outside the interval $[1, N]$. The optimal WLP filter coefficients can be determined by setting the partial derivatives of Eq. (1) with respect to each $a_k$ to zero. This results in the WLP normal equations

$$\sum_{k=1}^{p} a_k \sum_{n=n_1}^{n_2} W_n \cdot s_{n-k} s_{n-i} = \sum_{n=n_1}^{n_2} W_n \cdot s_n s_{n-i}, \quad 1 \le i \le p. \qquad (2)$$

Note that conventional LP is obtained as a special case of WLP: if $W_n$ is chosen as a finite nonzero constant for all $n$, it becomes a multiplier of both sides of Eq. (2) and cancels out leaving the LP normal equations (Makhoul, 1975a). Equation (2) can also be expressed in matrix form as

$$\left( \sum_{n=n_1}^{n_2} W_n \cdot \mathbf{s}_n \mathbf{s}_n^T \right) \mathbf{a} = \sum_{n=n_1}^{n_2} W_n \cdot s_n \mathbf{s}_n, \qquad (3)$$

where $\mathbf{a} = [a_1, a_2, ..., a_p]^{\mathrm{T}}$ and $\mathbf{s}_n = [s_{n-1}, s_{n-2}, ..., s_{n-p}]^{\mathrm{T}}$.

### B. Weighting with the short-time energy function

In the study by Ma *et al.* (1993), temporal weighting was computed from the speech signal by using the short-time energy (STE) function

$$W_n = \sum_{i=0}^{M-1} s_{n-1-i}^2, \qquad (4)$$

where the length of the energy window is denoted by $M$. The use of STE is motivated by the fact that it enables a straightforward computation of weighting that, overall, over-weights speech samples that occur after the glottal closure and under-weights those located in the glottal open phase (see Fig. 1 in Magi *et al.*, 2009). It is, however, worth emphasizing that this rationale is general and does not utilize precise extraction of the glottal open and closed phase. Instead, the use of STE is based on the general observation that speech samples have large amplitudes immediately after the instant of glottal closure (e.g., Strube, 1974). Because of these high-amplitude samples, the energy computation in Eq. (4) results in large $W_n$ values for time indices located approximately in the glottal closed phase.

STE weighting was motivated in Ma *et al.* (1993) with a general approach based on the function of the human voice production mechanism. In addition, STE weighting can be justified, even though not included in the study by Ma *et al.* (1993), from a mathematical perspective as follows. The WLP computation on the left-hand side of Eq. (3) forms the autocorrelation function of the normal equations by summing (from $n_1$ to $n_2$) temporal, "snap-shot" autocorrelation matrices $\mathbf{s}_n \mathbf{s}_n^T$ and multiplying these matrices by $W_n$. At time instant $n = j$, this "snap-shot" matrix can be written as

$$\mathbf{s}_j \mathbf{s}_j^T = \begin{bmatrix} s_{j-1}^2 & s_{j-1}s_{j-2} & \cdots & s_{j-1}s_{j-p} \\ s_{j-2}s_{j-1} & s_{j-2}^2 & \cdots & s_{j-2}s_{j-p} \\ \vdots & \vdots & \cdots & \vdots \\ s_{j-p}s_{j-1} & s_{j-p}s_{j-2} & & s_{j-p}^2 \end{bmatrix}. \qquad (5)$$

A widely used matrix norm in numerical linear algebra is the Frobenius norm (Golub and Van Loan, 1983), which can be written for the matrix of Eq. (5) as

$$\|A\|_F = \sqrt{\sum_{i=1}^{p}\sum_{k=1}^{p}|c_{ik}|^2}, \tag{6}$$

where $c_{ik}$ denotes an element on the $i$th row and $k$th column of the matrix given in Eq. (5). The Frobenius norm is a scalar value which reflects the averaged amplitude level of the elements of the underlying matrix. By taking into account the matrix symmetry in Eq. (5), the square of the Frobenius norm can be written as

$$\begin{aligned}\|A\|_F^2 = {} & (s_{j-1}^4 + s_{j-2}^4 + \cdots + s_{j-p}^4) \\ & + 2\cdot\{s_{j-1}^2(s_{j-2}^2 + s_{j-3}^2 + \cdots + s_{j-p}^2)\} \\ & + 2\cdot\{s_{j-2}^2(s_{j-3}^2 + s_{j-4}^2 + \cdots + s_{j-p}^2)\} \\ & + \cdots + 2\cdot\{s_{j-p}^2 s_{j-p-1}^2\}.\end{aligned} \tag{7}$$

The sum given in Eq. (7) can be simplified to

$$\|A\|_F^2 = (s_{j-1}^2 + s_{j-2}^2 + \cdots + s_{j-p}^2)^2. \tag{8}$$

By selecting a STE window with its duration parameter $M = p$, the following equation can now be written by using Eqs. (4) and (8):

$$\begin{aligned}W_j &= \sum_{i=0}^{M-1} s_{j-1-i}^2 \\ &= \sum_{i=0}^{p-1} s_{j-1-i}^2 = s_{j-1}^2 + s_{j-2}^2 + \cdots + s_{j-p}^2 = \|A\|_F.\end{aligned} \tag{9}$$

In other words, by selecting a STE function with its duration parameter $M$ equal to the order of the linear prediction $p$, temporal weighting in WLP corresponds to multiplying in Eq. (3) each "snap-shot" autocorrelation matrix $\mathbf{s}_n\mathbf{s}_n^T$ by a scalar coefficient $W_n$, whose value equals the Frobenius norm of the corresponding matrix. Hence, the contribution of those "snap-shot" matrices that have large Frobenius norm values is emphasized when individual matrices are summed together while matrices of small Frobenius values are de-emphasized. This implies that the role of speech data of large amplitude values is emphasized in the final normal equations, from which the WLP filter parameter vector $\mathbf{a}$ is solved. From the point of view of noise-robust computation of spectral models, this feature is justified because large speech samples are most likely less vulnerable to (additive) environmental noise. Indeed, STE weighting has shown successful behavior in comparisons of noise-robust feature extraction methods in automatic speech (Magi *et al.*, 2009) and speaker (Saeidi *et al.*, 2010) recognition. From the point of view of the present study, the above derivation also justifies the use of WLP with STE weighting, because emphasizing "snap-shot" matrices of large Frobenius norms can be interpreted as putting more weight on speech samples located after the glottal closure. These speech samples occur in the glottal closed phase during which the excitation from the vocal folds is weak. Hence, by emphasizing the contribution of these samples in the filter optimization, the biasing effect of F0 on formant estimates is expected to become smaller.

### C. Weighting with windows focusing on impulse-like excitations

By using STE weighting, WLP has been shown to yield spectral models that are less vulnerable to the effects of F0 than conventional LP (Ma *et al.*, 1993). It is worth emphasizing, though, that the STE function is unable to down-weight properly the contribution of the glottal excitation and, consequently, formant estimates computed by WLP are still biased by F0. However, it is possible to develop new weighting functions, which down-weight the contribution of the excitation more effectively thereby resulting in more accurate WLP models of the vocal tract transfer function. In order to develop such weighting, an example based on a simple digital source-filter model of speech production is studied. In this example, the glottal excitation and the vocal tract, respectively, are modeled with an impulse train of constant pitch (F0 = $1/T$, where $T$ denotes the period length) and a $p$th-order recursive IIR (infinite impulse response) filter of an all-pole structure. The excitation, shown in Fig. 1(a), is denoted by $x_n$. The transfer function of the vocal tract model is expressed as

$$H(z) = \frac{1}{1 + \sum_{k=1}^{p} b_k z^{-k}}. \tag{10}$$

Assume now that the vocal tract model is stable, its delay line is initialized to zero, and that $x_n$ is a causal signal starting at time index $n = 0$. The output of the filter, the signal denoted $y_n$, can be computed recursively using the well-known difference equations for linear digital systems (Oppenheim and Schafer, 1989) by utilizing the input $x_n$, previous values of the output $y_{n-k}$, $1 \le k \le p$, and the filter coefficients $b_k$, $1 \le k \le p$

$$y_n = x_n - \sum_{k=1}^{p} b_k y_{n-k}. \tag{11}$$

This recursive computation yields an output whose value at $n = 0$ equals the input, i.e., $y_0 = x_0$. For time indices $n > 0$, the output starts to fluctuate and its values gradually attenuate towards zero due to the filter stability, as seen in Fig. 1(b). However, the occurrence of a new non-zero sample of the input at time index $n = T$ causes a rapid increase in the value of $y_n$. This value, evidently, is almost equal to the value of the input at the same time index (i.e., $y_T \approx x_T = 1$) if the recursive contribution caused by the filter delay line is assumed to have been grossly attenuated. By repeating the same analysis over consecutive periods, the value of the synthesized speech signal $y_n$ of this simplified speech production model is observed to almost equal the corresponding value of the excitation $x_n$ at time indices $n = T, 2T,$

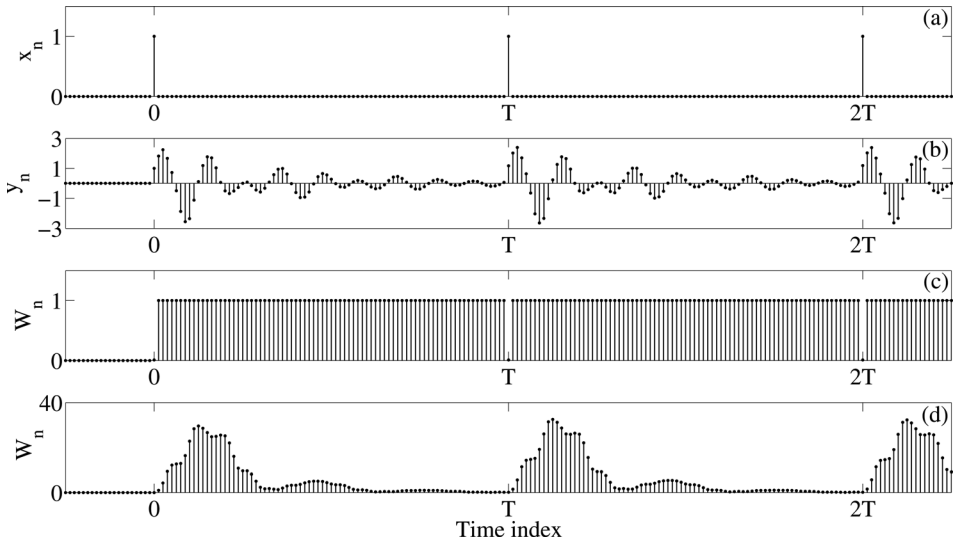


FIG. 1. Time-domain waveforms of a sound synthesis example in which an all-pole model of the vocal tract is excited by an impulse train. The (a) input and (b) output are shown. Weighting functions of the WLP analysis computed according to (c) Eq. (12) and (d) Eq. (4) are depicted.

3 $T$, etc. It is worth emphasizing that the input signal $x_n$ conveys no information about the resonances of the underlying IIR filter. Therefore, impulse-like peaks that occur regularly in $y_n$ at index $n = T$ and its integers multiples serve as distortion originating at the excitation that degrades the performance of conventional LP in the estimation of the resonances of the IIR filter from $y_n$. The higher the pitch, the more severe the distortion due to the larger number of excitation-originating peaks per LP frame in the computation of autocorrelation from $y_n$.

The problem addressed above can, however, be tackled by utilizing the idea of temporally weighted prediction together with a window that de-emphasizes the strong peaks caused by the excitation. One such solution can be obtained by using the following window function

$$W_n = \begin{cases} 1, & n \neq k \cdot T \\ c, & n = k \cdot T, k = 1, 2, 3..., \end{cases} \quad (12)$$

where parameter $c$ is chosen to be a positive real value close to zero (e.g., $c = 0.01$). This window is depicted in Fig. 1(c) for the simple speech production model described above. For comparison, the STE weighting function computed for $y_n$ according to Eq. (4) by using $M = p$ is shown in Fig. 1(d). The all-pole spectra obtained by using these two weighting functions in the WLP computation are shown in Fig. 2 together with the spectrum determined by conventional LP and with the true all-pole spectrum of the IIR filter. In order to demonstrate the effect of pitch, two examples with F0 equal to 100 Hz and 400 Hz are shown in Figs. 2(a) and 2(b), respectively. In the case of lower F0, conventional LP and both of the WLP methods can be seen to be able to estimate the spectral envelope of the IIR filter accurately. However, in the case of the higher pitch, all-pole filters defined by both conventional LP and WLP with STE weighting are greatly distorted due to the occurrence of prominent peaks in $y_n$ originating at the excitation. De-emphasizing the prediction error at these peaks by utilizing WLP with the window given in Eq. (12) results in an all-pole model that can be clearly seen in the figure to be closer to the original IIR filter.

In the example above, the distortion caused by the high-pitched excitation is attenuated by designing a weighting function that de-emphasizes large prediction error samples that occur at the same time instants where the impulse train input is non-zero. While an ideal impulse train excites the

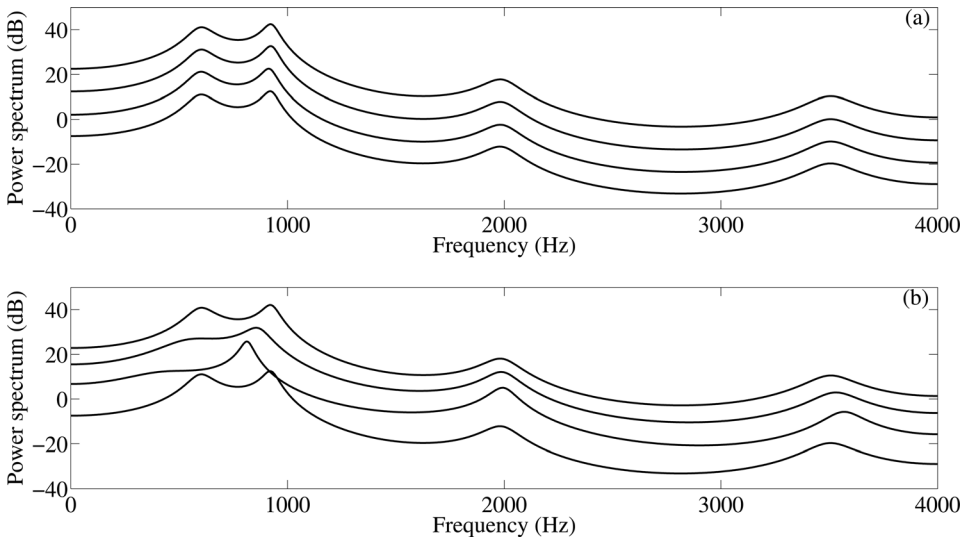


FIG. 2. All-pole spectra computed from the signal shown in Fig. 1(b) with (a) F0 = 100 Hz and (b) F0 = 400 Hz. The spectra shown correspond to, from bottom to top, (1) the original all-pole filter used in the sound synthesis, (2) conventional LP, (3) WLP computed with the STE weighting function [Eq. (4), $M = p$], and (4) WLP computed with the weighting function defined in Eq. (12).

vocal tract by non-zero samples only once per fundamental cycle, the same does not hold true in the production of natural speech. Instead, by assuming that the glottal flow, vocal tract, and lip radiation can be modeled as linear processes and that the lip radiation effect can be combined with the glottal flow via differentiation (Fant, 1970), the excitation is the derivative of the glottal flow having many non-zero samples during one cycle. For this kind of an excitation waveform, the WLP can be implemented with a simple weighting function shown in Fig. 3. This weighting, denoted as the Attenuated Main Excitation (AME) function, de-emphasizes the square of the prediction residual in several time samples that are located in the vicinity of the main excitation of the vocal tract, the instant denoted by $t_{me}$ in Fig. 3. The AME window has one amplitude parameter, denoted by $d$ in Fig. 3, that determines the level of attenuation. In addition, the function uses two relative parameters defined in the time domain as follows. The first, duration quotient (DQ), is a measure of the duration of the attenuated section of the window with respect to the length of the fundamental period: $DQ = (T_1/T) \times 100\%$. The second, position quotient (PQ), is a measure of the position of the main excitation of the vocal tract with respect to the duration of the attenuated section: $PQ = (T_2/T_1) \times 100\%$. In order to avoid abrupt changes, the weighting function follows a linear ramp when changing between its maximum 1.0 and minimum $d$. In order to reduce the number of parameters, the duration of the ramp is set to a constant value (0.4 ms was used which corresponds to three samples with a sampling frequency of 8 kHz). With the above definition, the AME function aims to focus on the vicinity of the main peak of the speech excitation waveform. Therefore, the computation of WLP naturally calls for estimating the locations of the main excitations which can be computed using either electroglottography (EGG) (e.g., Krishnamurthy and Childers, 1986) or epoch extraction techniques developed to estimate glottal closure instants directly from acoustic speech signals (e.g., Murthy and Yegnanarayana, 1999; Naylor *et al.*, 2007).

## III. ALL-POLE MODELING METHODS TO BE COMPARED

The goal of the present study was to evaluate the performance of the WLP-based techniques described in Sec. II in formant frequency estimation of high-pitched speech and compare WLP with several all-pole modeling methods developed during the past three decades. In order to accomplish this, a set of earlier techniques with the most potential was first selected. The selection criteria and parameter settings of the selected methods are described in this section. In addition, the identification technique utilized in detecting formant frequencies from all-pole models is explained.

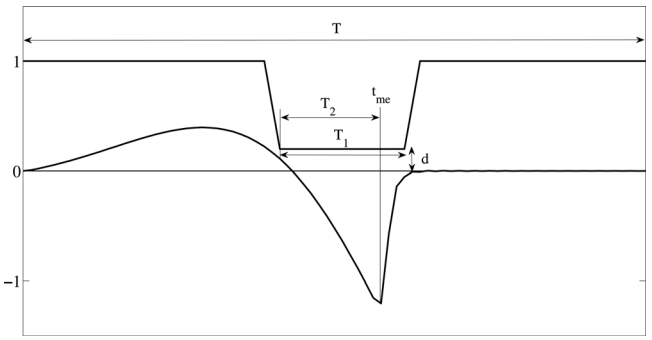


FIG. 3. (Top) Waveform of the Attenuated Main Excitation (AME) function together with (bottom) a differentiated glottal flow synthesized with the Liljencrants-Fant model. Parameters of the AME function correspond to the fundamental period $T$, the duration $T_1$ of the attenuated section, the time $T_2$ between the beginning of the attenuated section and the position of the main excitation $t_{me}$, and the amplitude $d$ of the attenuated section.

### A. Selected methods and their parameter settings

Methods were selected for comparison based on the following criteria. First, only parametric spectral modeling methods based on all-pole filter structures were included, while algorithms utilizing the autoregressive moving average (ARMA) model were excluded. Second, all methods to be included were to be based on an algorithm different from WLP but still specifically developed to cope with the problems caused by high pitch. Third, the method was to be based on a similar set of basic parameters (prediction order, frame length, window type in the computation of the autocorrelation function) that is used in conventional LP. Similarly to conventional LP, none of the methods selected required temporal change of pitch. Based on the criteria above, the following all-pole modeling methods were included in the comparison: (1) conventional LP, (2) Robust Linear Prediction (RBLP) proposed by Lee (1988), (3) Discrete All-pole Modeling (DAP) developed by El-Jaroudi and Makhoul (1991), (4) Linear Prediction using Refined Autocorrelation (LPRA) by Rahman and Shimamura (2007), (5) Weighted Linear Prediction with the Short-Time Energy weighting function (WLP-STE) proposed by Ma *et al.* (1993), and (6) Weighted Linear Prediction with Attenuated Main Excitation (WLP-AME) developed in this study.

All linear predictive methods selected were computed with the autocorrelation criterion using a frame length of 25 ms together with Hamming windowing and a first order all-zero pre-emphasis with zero at $z = 0.97$. Two different orders of prediction were used due to the different sampling frequencies of the synthetic and natural speech: for synthetic vowels (sampled with 10 kHz), the $p$ value was set to 10, while for natural speech (sampled with 8 kHz), the value $p = 8$ was used. RBLP was computed in a similar manner as in Lee (1988) by using Huber's $\psi$-function with $c = 1.5$ and the Iterative Reweighted Least Squares Algorithm [see Eqs. (3.9)–(3.12) in Lee, 1988]. The DAP analysis was implemented by using the $\alpha$ value of 0.6 and 20 iterations [see Sec. IV B and Eq. (48) in El-Jaroudi and Makhoul, 1991]. LPRA was computed according to Rahman and Shimamura (2007) by using a cepstral window, whose length was 3.6 ms and 2.4 ms for vowels with F0 smaller and larger, respectively, than 200 Hz. Homomorphic deconvolution was implemented by using FFT-based conversions between time and cepstral domains [see Eqs. (5)–(9) in Rahman and Shimamura, 2007]. WLP-STE was computed as in Ma *et al.* (1993), that is, by utilizing Eqs. (3) and (4) of the present study and by setting $M = p$.

Finally, the parameters of the AME function were optimized by the following approach. First, a set of synthetic

vowels was generated by using a linear source-filter model with an artificial glottal source waveform and all-pole modeling of the vocal tract. The first time derivative of the glottal flow pulse was simulated with the Liljencrants-Fant (LF) waveform (Fant *et al.*, 1985) by using four different phonation modes (modal, breathy, whispery, and creaky). LF parameters of the four phonation modes were taken from Gobl (1989). F0 values of the glottal source signals were varied between 100 and 450 Hz with an increment of 50 Hz. Ten different synthetic American-English vowels were created by modeling their vocal tract transfer functions with eighth-order all-pole filters whose formant frequencies and bandwidth were taken from Tables I and II of Gold and Rabiner (1968). In total, this synthesis procedure yielded 320 signals (eight F0 values, four phonation modes, ten vowels).

Second, WLP-AME analysis was performed for all the synthetic sounds by varying the parameters of the weighting function shown in Fig. 3 as follows: (a) six different values were used for the parameter $d$ (0.01, 0.03, 0.05, 0.10, 0.15, and 0.20), (b) four values were used for $DQ$ (20%, 40%, 60%, and 80%), and (c) six values were used for $PQ$ (0%, 20%, 40%, 60%, 80%, and 100%). Hence, 144 ($6 \times 4 \times 6$) WLP-AME analyses were performed for each of the 320 signals resulting in a total of 46 080 analyses.

Third, frequencies of the lowest four formants were computed from each WLP-AME filter by peak-picking the corresponding all-pole spectrum. The estimated values were compared with the true ones by computing the relative formant error measure [see Eq. (10) in Rahman and Shimamura, 2007]. Finally, the AME window parameter vector yielding the minimum relative formant error was sought. This procedure resulted in the following optimal values of the AME function: $d = 0.01$, $DQ = 40\%$, and $PQ = 80\%$. These settings were then used in all further WLP-AME analyses of this study.

### B. Formant identification from all-pole models

The identification of formant frequencies from all-pole filters can be computed by searching for the roots of the filter's denominator as used, for example, by Rahman and Shimamura (2007) and by Wang and Quatieri (2010). Alternatively, formants can be identified by looking for the peaks of the corresponding all-pole spectrum, an approach that has been utilized, for example, by Hillenbrand *et al.* (1995) and Hagiwara (1997). In the present study, the latter method was adopted because it has been shown to be more reliable especially when dealing with formants that are located at low frequencies or close to each other (Vallabha and Tuller, 2002). An all-pole spectrum can become overly smooth not showing a sufficient number of local peaks. Alternatively, the spectrum may show spurious peaks whose number, depending on the choice on the order of prediction, might be larger than the number of formants sought. In order to take these phenomena into account, each all-pole spectrum was quantified not only by determining the locations of its spectral peaks (in Hz) but also by counting the number of these peaks within the spectral region from 250 Hz to half the sampling frequency where the lowest three formants are expected to be located.

## IV. SPEECH MATERIAL

Both synthetic and natural vowels were used to evaluate the performance of the different all-pole modeling methods in formant estimation. In the following, the generation and properties of the test vowels are described first for the synthetic vowels, and then for the natural utterances.

### A. Synthetic vowels

A computational model of the speech production system was used to generate synthetic vowel samples representative of an adult male, adult female, and a child aged approximately 5 years. The voice source component of the model used for all vowel samples consisted of a kinematic representation of the medial surface of the vocal folds (Titze, 1984, 2006) for which the fundamental frequency, surface bulging, adduction, length, and thickness are control parameters. The vocal fold length was set to 1.6 cm for the male, 1 cm for the female, and 0.8 cm for the child model. Similarly, the thickness of the vocal folds was set to 0.3 cm for the male, 0.2 cm for the female, and 0.15 cm for the child. As the vocal fold surfaces are set into vibration the model produces a glottal area signal that is coupled to the acoustic pressures and air flows in the trachea and vocal tract through aerodynamic and acoustic considerations as discussed by Titze (2002). The resulting glottal flow was determined by the interaction of the glottal area with the time-varying pressures present just inferior and superior to the glottis.

The vocal tract shape, which extended from glottis to lips, was specified by an area function representative of an [a], [æ], [i], or the neutral vowel. The MRI-based area functions were taken from Story (2008) for the adult male vowels and from Story (2005) for the adult female vowels. The child-like area functions were not measured directly but rather generated with an acoustic sensitivity function approach (Story, 2006) that mapped desired formant frequencies to a plausible vocal tract shape. For each vowel area function the vocal tract length was set to 17.5 cm for the male, 14.1 cm for the female, and 11.4 cm for the child. The tracheal shape was also specified by an area function that extended from the glottis to the bronchi and is based on data reported in Story (1995). Although the length of trachea was scaled for the male, the female, and the child by the same ratios as the vocal tract lengths, the cross-sectional area of the trachea (i.e., shape) was maintained constant for all syntheses. The acoustic wave propagation in the subglottal and supraglottal airspaces was computed with a wave-reflection model (Liljencrants, 1985; Story, 1995) that included energy losses due to yielding walls, viscosity, heat conduction, and radiation at the lips (Story, 1995).

The complete model for the adult male is illustrated in Fig. 4. The three-dimensional representation of the vocal fold medial surfaces is shown bounded on the upstream (subglottal) side by the tracheal section and bronchial termination and on the downstream (supraglottal) side by the tubular representation of the vocal tract (in this case the neutral vowel). A sample glottal flow signal is shown near the junction of the trachea and lower vocal tract, which results from the interaction of the subglottal and supraglottal pressures

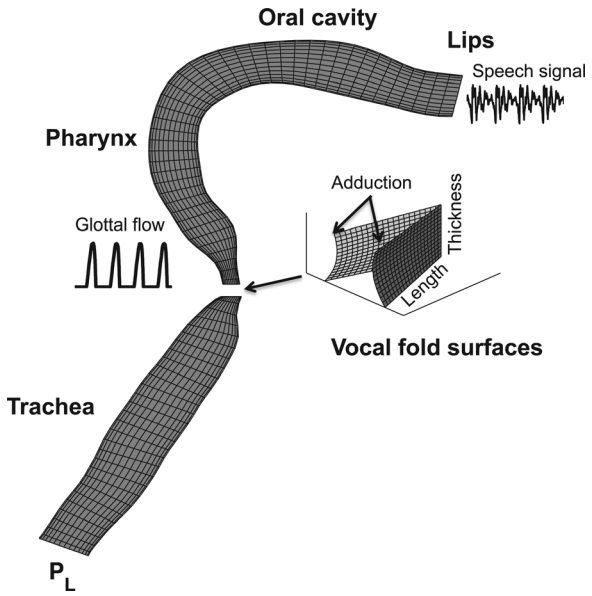


FIG. 4. Illustration of the speech production system as a tubular airway model. The trachea is shown in the lower part of the figure where the bronchial termination is indicated by $P_L$, the lung pressure that drives the system. On the right side is a schematic representation of the vocal fold medial surfaces; for illustrative purposes they are shown magnified and displaced from their appropriate location. In the upper part is shown the vocal tract shape for an adult male neutral vowel. A sample glottal flow signal is indicated near the junction of the trachea and lower vocal tract, and the radiated speech signal is shown near the lips.

with the glottal area. The radiated speech signal is shown near the lips and is analogous to a microphone signal recorded from a real talker. Additional information and examples are available in Story (2013).

Each of the four vowels for the male, the female, and the child were generated with eight fundamental frequencies, ranging from 100 Hz to 450 Hz in 50-Hz increments. The signals were first generated with the sampling frequency of 44.1 kHz from which they were down-sampled to 10 kHz. Although the full range of these F0s is unlikely to be produced by either the male, female, or the child, conducting the experiment with the entire range was desirable for ease in comparison. Each vowel sample was generated with a total duration of 0.4 s and F0 was maintained constant during the sample.

### B. Natural vowels

Natural vowels were recorded from 12 talkers (six males, six females, mean age 39 year, min. age 21 year, max. age 50 year), all native speakers of Finnish. They were instructed to produce sustained speech phonation (not singing) on vowels [a], [æ], and [i] at different pitches. The subjects produced their first sample by using their habitual speaking pitch and increased the pitch diatonically up to a level where they still could comfortably sustain a speech-like sound with a clear vowel quality. Each vowel was phonated for at least 2 s and repeated three times at one pitch. The whole pitch series was accomplished with one vowel first and then again with another vowel. As an aid, reference sounds were given with an electric keyboard instrument.

Speech pressure signals were recorded by a condenser microphone (Brüel & Kjær 4188) that was attached to a sound-level meter (Brüel & Kjær Mediator 2238) serving also as a microphone amplifier. EGG was recorded simultaneously (Glottal Enterprise MC2-1). In order to avoid inconsistency in the synchronization of speech and EGG, the mouth-to-microphone distance (40 cm) was carefully monitored in the recordings. Speech and EGG waveforms were recorded into a computer at a sampling rate of 44.1 kHz, with 16-bit resolution.

The middle utterance in each series of three samples produced using the same pitch was selected on the computer for further analyses. The sampling frequency of speech and EGG was down-sampled to 8 kHz and the propagation delay of the acoustic signal from the glottis to the microphone was compensated for. This propagation delay was estimated by using the vocal tract length of 15 cm and 17 cm for females and males, respectively, the mouth-to-microphone distance of 40 cm, and the sound speed of 359 m/s. These values yielded the propagation delay of 1.53 ms and 1.59 ms for female and male speakers, respectively. In total, the data collection procedure yielded 330 natural vowel utterances (152 produced by females, 178 produced by males) whose F0 varied between 167 and 500 Hz for females and between 84 and 296 Hz for males.

## V. QUANTIFICATION OF FORMANT ESTIMATION ACCURACY FROM ALL-POLE MODELS

Quantifying the formant estimation accuracy of all-pole modeling methods is feasible for synthetic speech because the true values of formants are known. For natural vowels, however, the evaluation is problematic because the correct formant frequencies are unknown. Therefore, different quantification methods were used in the present study for comparing the synthetic and natural vowels. The techniques used are described separately in this section.

### A. Quantification of formant estimation accuracy in synthetic vowels

Formant estimation accuracy was quantified for the synthetic vowels using the following straightforward error measure (e.g., Wang and Quatieri, 2010):

$$d_{\mathrm{err},i} = 100\% \times \frac{|F_{\mathrm{est},i} - F_{\mathrm{tru},i}|}{F_{\mathrm{tru},i}}, \tag{13}$$

where $F_{\mathrm{est},i}$ is the estimated $i$th formant frequency extracted by peak-picking the spectrum resulting from the underlying all-pole modeling technique and $F_{\mathrm{tru},i}$ denotes the true $i$th formant used in the physical modeling. The error measure given in Eq. (13) was defined separately for the lowest three formants (i.e., $1 \leq i \leq 3$). In addition, an error measure that takes all three formants into account was used by computing the Euclidean distance (in Hz) between the true and the estimated formant vector

$$d_{\mathrm{euc}} = \sqrt{\sum_{i=1}^{3} (F_{\mathrm{est},i} - F_{\mathrm{tru},i})^2}, \tag{14}$$

where $F_{\mathrm{est},i}$ and $F_{\mathrm{tru},i}$ are defined as in Eq. (13).

Additionally, the number of spectral peaks found in the all-pole spectra was quantified by a relative number, denoted by $n_{pks}$, which is defined as the proportion of the number of analyses showing at least three spectral peaks to the total number of analyses conducted.

### B. Quantification of formant estimation accuracy in natural vowels

For natural vowels, only the first (F1) and second (F2) formant were identified from the computed all-pole spectra. The third formant (F3) was excluded from the analysis because previous studies (e.g., Rahman and Shimamura, 2007) indicate that it is not affected by F0 as much as F1 and F2. In quantifying all-pole models in formant estimation of natural vowels, *a priori* knowledge about the locations of [a], [æ], and [i] in the F1, F2 space was first used to set frequency limits to identify which peaks of the all-pole spectrum were to be regarded as F1 and which F2. For the vowel [a], F1 and F2 were identified as the lowest and second lowest peak, respectively, of the all-pole spectrum in the frequency range between 200 and 1700 Hz. For the vowel [æ], the corresponding frequency limits were 200 and 2500 Hz. In contrast to [a] and [æ], the vowel [i] is characterized by a low F1 and a high F2. Therefore, F1 of the vowel [i] was determined by searching for the all-pole spectrum peak between 200 and 600 Hz and F2 was identified as the lowest peak in a separate frequency range between 1800 and 3500 Hz.

Once both F1 and F2 were identified, their contours as a function of F0 were characterized by first forming a difference which describes how a formant frequency changes between consecutive samples in the vowel series of increasing pitch. The obtained difference series was then quantified with a single numerical value by computing its standard deviation as follows:

$$\delta_i = \mathrm{SD}\{F_{k,i} - F_{k-1,i}\}, \quad 2 \le k \le R, \tag{15}$$

where SD denotes the standard deviation, $F_{k,i}$ denotes the $i$th formant frequency ($i = 1, 2$) computed from the $k$th sample in the vowel series of increasing pitch and $R$ denotes the number of the samples in the series. In principle, if a speaker is capable of keeping his or her formant frequencies constant while raising the pitch and if an all-pole modeling method is able to find formants correctly, Eq. (15) will result in the value zero. Furthermore, if a speaker is unable to preserve the vocal tract resonances in constant positions while raising the pitch but creates a linear change in the formant frequencies, Eq. (15) yields zero if the formants are correctly identified by all-pole modeling. However, if the formant frequencies estimated by the all-pole modeling fluctuate between consecutive samples in the vowel series, a phenomenon whose likely cause is the biasing effect of F0, the difference will become amplified resulting in a positive, nonzero value of $\delta_i$. Hence, Eq. (15) enables a quantitative comparison of how formants estimated from natural speech vary for different all-pole modeling methods when fundamental frequency is increased.

## VI. EXPERIMENTS

Two main experiments were conducted for the synthetic vowels and one for the natural utterances. The first experiment on synthetic speech, the results of which are reported in Sec. VII A 1, corresponds to the ideal case in which the main excitation of the vocal tract, $t_{me}$ shown in Fig. 3, is always correctly obtained. In each fundamental cycle, this epoch was computed by searching for the time instant of the minimum peak of the differentiated glottal area signal given by the computational model (see Fig. 5, for an example).

The second experiment on the synthetic vowels, the results of which are reported in Sec. VII A 2, consisted of analyses in which the correct value of $t_{me}$ was deliberately distorted by a random number. The goal of this experiment was to validate the robustness of WLP-AME with respect to the extraction of the instant of the main excitation. The correct discrete time index of $t_{me}$ was distorted as follows:

$$t_{me}^{d} = t_{me} + t_d, \tag{16}$$

where $t_d$ denotes a random discrete time variable of uniform distribution. In order to vary the amount of distortion, the

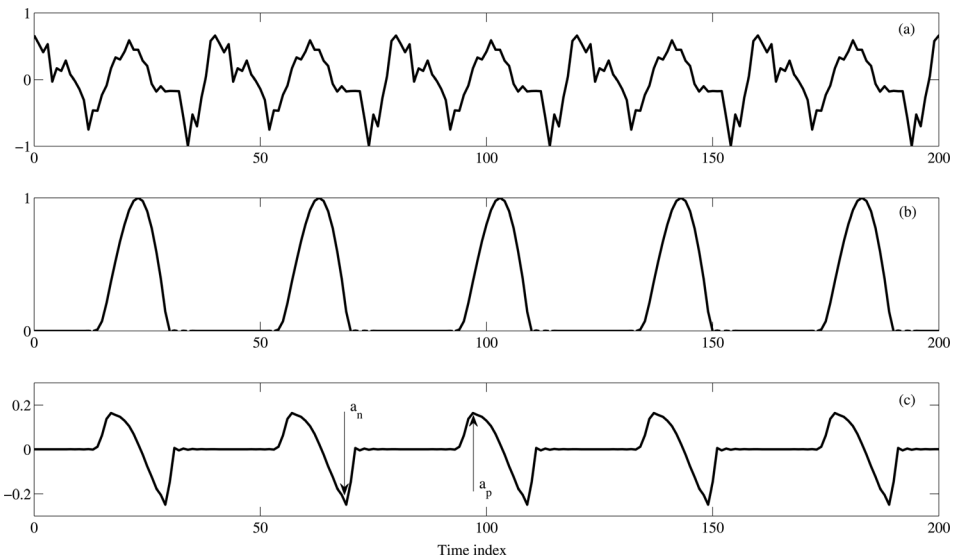


FIG. 5. Signals synthesized with the physical modeling approach by using parameters that correspond to the production of the vowel [a] by a male speaker using F0 of 250 Hz. (a) Example of a speech signal, (b) glottal area function, and (c) differentiated glottal area function. The negative and positive peak of the differentiated area function is shown for one fundamental period by $a_n$ and $a_p$, respectively.

minimum and maximum values of $t_d$ were constrained by a discrete time-domain parameter, denoted by $r$, as follows:

$$-r \leq t_d \leq r. \quad (17)$$

The randomization of $t_{me}$ was conducted separately for each fundamental period in the WLP-AME frame. The value of $r$ was varied from $r=0$ (i.e., correct epochs) to $r=10$ with an increment of two samples. For each value of $r$, the formant estimation performance of WLP-AME was parameterized using the five quantification methods described in Sec. V A.

Combined with the second experiment, the formant estimation accuracy of WLP-AME was also compared with the closed phase covariance (COV) method (Strube, 1974; Wong *et al.*, 1979). The COV analysis is similar to WLP-AME in the sense that it is based on removing the effect of the glottal source in the estimation of the vocal tract resonances in the computation of LP analysis. Differently from WLP-AME, the COV analysis does not use a temporal weighting in the optimization of the filter coefficients, but rather assumes that there exists a closed phase of the glottal cycle during which there is no excitation and the filter coefficients can be computed free from the contribution of the voice source. Since the duration of the closed phase is typically short, the COV analysis, however, uses the covariance criterion (Rabiner and Schafer, 1978) in the computation of the optimal LP filter instead of the autocorrelation criterion that is used in all methods described in Sec. III A. Since both WLP-AME and COV require extracting the instant of the glottal closure, their comparison was conducted jointly by evaluating the effect of random variation in the instant of the main excitation.

Several previous studies (e.g., Yegnanarayana and Veldhuis, 1998; Plumpe *et al.*, 1999; Alku *et al.*, 2009) have indicated that formant estimates computed by COV are greatly affected by both the position of the beginning of the covariance window (with respect to the instant of the main excitation) and the length of the window (i.e., the number of speech samples in the covariance window that are expected to occur in the closed phase). Therefore, the optimal values for both the beginning and length of the covariance window were first determined for each individual speech sound (i.e., separately for each vowel, F0 value, and speaker used in the physical modeling). The beginning of the covariance window was varied between the instants of the negative and positive peak of the differentiated glottal area function. These instants could easily be detected, as shown by an example in Fig. 5, by peak picking the derivative of the area function. Since the time distance between the negative and positive peak depends on the length of the fundamental cycle, the number of the beginning indices to be tested varied between 16 (in the highest F0 category) and 66 (in the lowest F0 category) for each individual speech signal.

For the beginning index of each covariance window, a number of COV analyses were computed by varying the length of the window as in the study by Plumpe *et al.* (1999). [Similarly to Plumpe *et al.* (1999), the length of the covariance window is denoted by $N_w$ in the following.] First, a two-window covariance analysis was computed by utilizing speech data from two consecutive periods. [For further details, see Eq. (7) in Plumpe *et al.* (1999).] The length of this covariance window in both periods was recommended by Plumpe *et al.* (1999) to be set to the value that is "slightly larger than half the desired LP order." Since $p=10$ was used as the LP order in the present study for the synthetic vowels, the window length of the two-window analysis was set to seven samples in both periods. Second, conventional one-window COV analyses were computed by varying $N_w$ between a lower bound and an upper bound, both of which were set to the LP order dependent values suggested by Plumpe *et al.* (1999). With $p=10$, this resulted in using 13 and 20 samples as the lower and upper bound, respectively, in varying $N_w$. Hence, there were altogether nine covariance window lengths to be varied (one for the two-window analysis, eight for the one-window analysis). The optimal covariance window setting was finally determined by searching for the combination of beginning index and length which yielded a window during which the energy of the glottal area function per time instant was smallest.

The experiment conducted for natural vowels, the results of which are reported in Sec. VII B, was limited to involve only two linear predictive methods, conventional LP and WLP-AME. This choice was made because RBLP, DAP, LPRA, and WLP-STE have been previously compared with conventional LP in the respective publications. Instants of $t_{me}$ were delineated in the computation of WLP-AME by searching for the maximum peak amplitude of the differentiated EGG for each glottal cycle inside the analysis frame.

## VII. RESULTS

### A. Synthetic vowels

#### 1. Experiments with correct values of $t_{me}$

Due to the large number of different parameters involved, the data needed to be compressed before it could be expressed. Therefore, a procedure similar to that used by Wang and Quatieri (2010) was utilized: the performance of each all-pole modeling method as a function of F0 is shown graphically for one vowel only ([a]) after which the data in different pitch categories is combined into tables and shown separately for each of the four vowels synthesized. In addition, data originating from child, female, and male vocal tracts are always combined.

Error measures $d_{\mathrm{err},1}$, $d_{\mathrm{err},2}$, $d_{\mathrm{err},3}$, and $d_{\mathrm{euc}}$ are shown as a function of F0 for the synthetic vowel [a] in Figs. 6–9, respectively. In order to enable comparison of six all-pole modeling methods, each error measure is illustrated in three panels. Conventional LP, marked by circles, is included in all panels, while the other five methods are shown only once in the corresponding panel. The data shown agree with the results of several previous studies (e.g., Rahman and Shimamura, 2007; Wang and Quatieri, 2010) in that the formant estimation error increases as F0 rises and that this degradation is most severe for F1. In all, however, the estimation error obtained in the present study is larger than that reported by Wang and Quatieri (2010). This is most likely

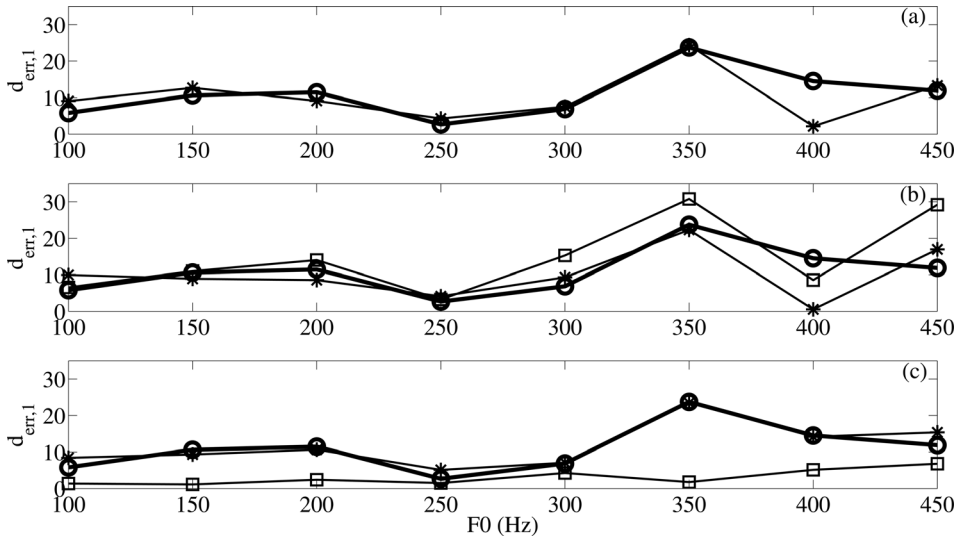


FIG. 6. Estimation error for the first formant as a function of F0. Error $d_{\mathrm{err},1}$ is computed using Eq. (13) with $i = 1$. Six all-pole modeling methods analyzed are shown in the three panels as follows: LP (circles) is shown in all panels, RBLP (asterisks) is shown in panel (a), DAP (asterisks), and LPRA (squares) are shown in panel (b), and WLP-STE (asterisks) and WLP-AME (squares) are shown in panel (c). Analysis was computed from synthetic speech by averaging results in each F0 and method over four vowels ([a], [æ], [i], the neutral vowel) and three speakers (male, female, child) used in the physical modeling approach.

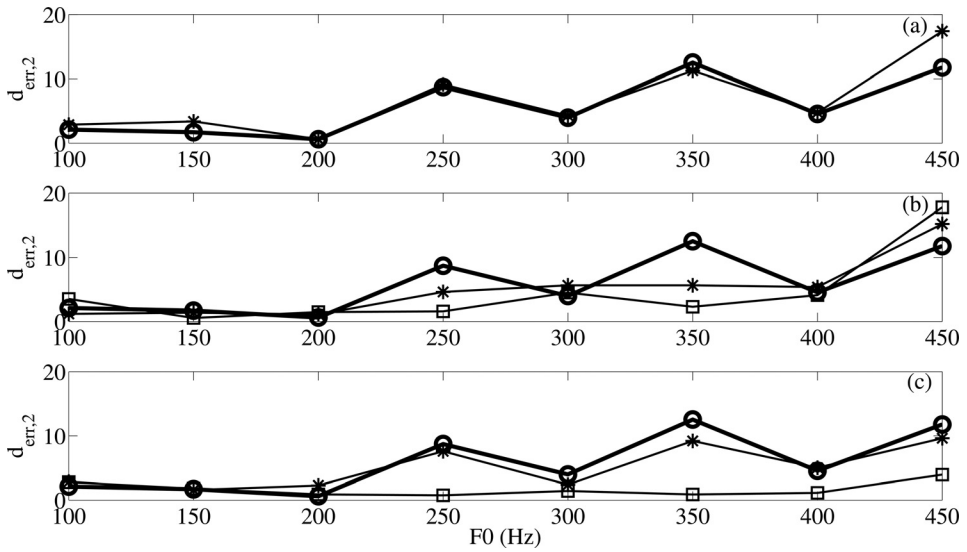


FIG. 7. Estimation error for the second formant as a function of F0. Error $d_{\mathrm{err},2}$ is computed using Eq. (13) with $i = 2$. Six all-pole modeling methods analyzed are shown in three panels as follows: LP (circles) is shown in all panels, RBLP (asterisks) is shown in panel (a), DAP (asterisks) and LPRA (squares) are shown in panel (b), and WLP-STE (asterisks) and WLP-AME (squares) are shown in panel (c). Analysis was computed from synthetic speech by averaging results in each F0 and method over four vowels ([a], [æ], [i], the neutral vowel) and three speakers (male, female, child) used in the physical modeling approach.

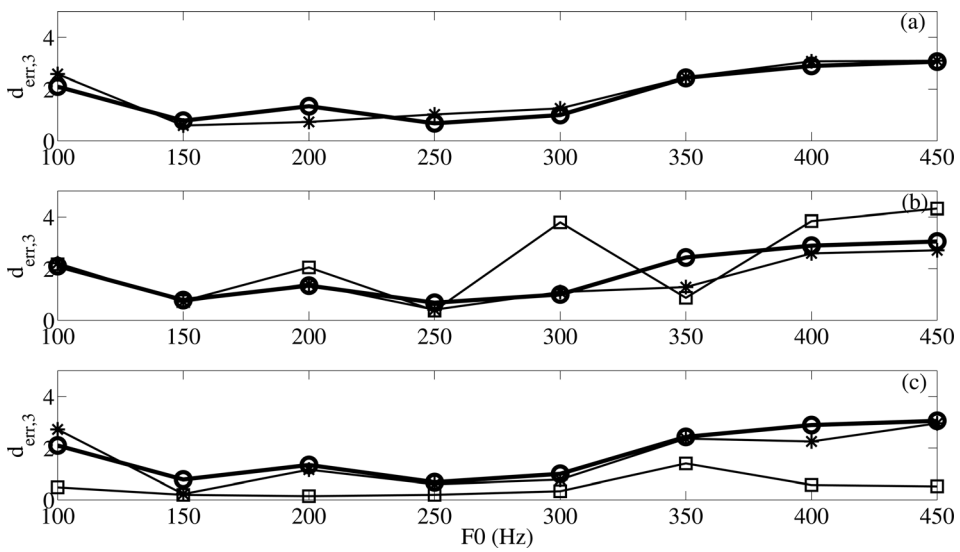


FIG. 8. Estimation error for the third formant as a function of F0. Error $d_{\mathrm{err},3}$ is computed using Eq. (13) with $i = 3$. Six all-pole modeling methods analyzed are shown in three panels as follows: LP (circles) is shown in all panels, RBLP (asterisks) is shown in panel (a), DAP (asterisks) and LPRA (squares) are shown in panel (b), and WLP-STE (asterisks) and WLP-AME (squares) are shown in panel (c). Analysis was computed from synthetic speech by averaging results in each F0 and method over four vowels ([a], [æ], [i], the neutral vowel) and three speakers (male, female, child) used in the physical modeling approach.

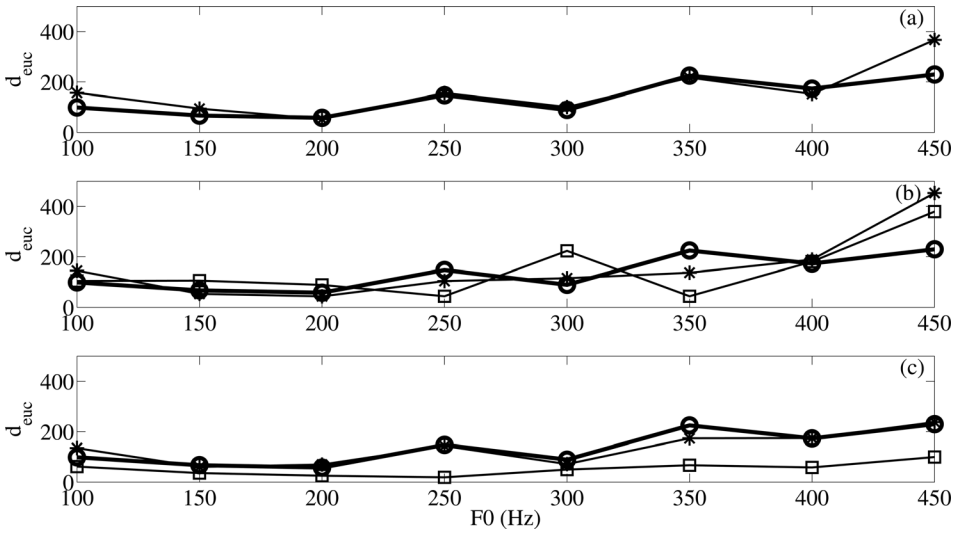


FIG. 9. Euclidean formant estimation error as a function of F0. Error $d_{euc}$ is computed using Eq. (14). Six all-pole modeling methods analyzed are shown in the three panels as follows: LP (circles) is shown in all panels, RBLP (asterisks) is shown in panel (a), DAP (asterisks) and LPRA (squares) are shown in panel (b), and WLP-STE (asterisks) and WLP-AME (squares) are shown in panel (c). Analysis was computed from synthetic speech by averaging results in each F0 and method over four vowels ([a], [æ], [i], the neutral vowel) and three speakers (male, female, child) used in the physical modeling approach.

explained by the use of different excitation waveforms in the vowel synthesis: Simple impulse trains were used by Wang and Quatieri (2010) while the present study employed the physical modeling approach, complicating the separation of the vocal tract and the voice source. The magnitude of the estimation error observed in the present study is also larger than that reported by Rahman and Shimamura (2007). However, the difference between the present study and the investigation of Rahman and Shimamura (2007) is smaller than the difference between the present study and the investigation by Wang and Quatieri (2010). This finding appears logical because Rahman and Shimamura (2007) utilized more realistic LF-modeled excitation waveforms in their vowel synthesis, and thus, their synthesis methods were more similar to the present study. The estimation error, shown in Figs. 6–9, is not a simple monotonically increasing function of F0, and in this follows the data in Wang and Quatieri (2010). Furthermore, the accuracy of RBLP and DAP is in general better than that of conventional LP, even though there are individual cases where conventional LP outperforms either of these two robust methods. In contrast, the performance of LPRA was surprisingly inferior to that of conventional LP especially when F0 was larger than 250 Hz. This might be explained by the simple truncation operation utilized in LPRA in the cepstral domain to reduce the biasing effect of F0. Although this approach yielded good results in the study by Rahman and Shimamura (2007), it does not seem to work with the present test material. The clearest result, however, is seen in Figs. 6–9 when WLP-AME is compared with the rest of the all-pole modeling techniques. For the great majority of the cases, the formant estimation error of WLP-AME is seen to be smaller than that of any other technique. The difference between the two WLP-based methods is also remarkably large indicating that the AME function is able to reduce the biasing effect of the glottal source better than the STE weighting.

Data for [a], [æ], [i], and the neutral vowel are shown in Tables I–IV, respectively, by pooling together the eight F0 categories. These data indicate again that the estimation error is largest for the first formant. Comparison of the different vowels also shows that for the vowel [i], which has low F1, the estimation error is in general larger than for the other vowels analyzed. The results also corroborate previous findings reported by Lee (1988) and El-Jaroudi and Makhoul (1991) in showing that RBLP and DAP yield better estimation accuracy than conventional LP (except for the vowel [i]). Estimation accuracy for the second and third formant is, however, less consistent, including cases where conventional LP was better than RBLP or DAP. LPRA yielded the largest estimation error for F1, except for the neutral vowel for

TABLE I. Formant estimation results for the synthetic [a] vowels. Each row corresponds to an all-pole modeling method evaluated (notation is described in Sec. III A). Relative errors in F1, F2, and F3, defined by Eq. (13), are given in columns 1–3, respectively. The Euclidean distance between the true and the estimated formant, defined by Eq. (14), is given in column 4. The last column includes the relative number of all-pole analyses indicating at least three formant peaks.

| | $d_{err,1}$ (%) | $d_{err,2}$ (%) | $d_{err,3}$ (%) | $d_{euc}$ (Hz) | $n_{pks}$ (%) |
|---|---|---|---|---|---|
| LP | 11.13 | 5.86 | 1.79 | 136.15 | 79.17 |
| RBLP | 10.53 | 6.42 | 1.85 | 146.81 | 75.00 |
| DAP | 10.49 | 4.87 | 1.55 | 152.36 | 75.00 |
| LPRA | 15.43 | 4.18 | 2.33 | 140.42 | 66.67 |
| WLP-STE | 11.68 | 5.17 | 1.63 | 132.67 | 79.17 |
| WLP-AME | 3.04 | 1.70 | 0.48 | 52.30 | 100.00 |

TABLE II. Formant estimation results for the synthetic [æ] vowels. Each row corresponds to an all-pole modeling method evaluated (notation described in Sec. III A). Relative errors in F1, F2, and F3, defined by Eq. (13), are given in columns 1–3, respectively. The Euclidean distance between the true and the estimated formant, defined by to Eq. (14), is given in column 4. The last column includes the relative number of all-pole analyses indicating at least three formant peaks.

| | $d_{err,1}$ (%) | $d_{err,2}$ (%) | $d_{err,3}$ (%) | $d_{euc}$ (Hz) | $n_{pks}$ (%) |
|---|---|---|---|---|---|
| LP | 9.83 | 4.27 | 4.10 | 217.29 | 100.00 |
| RBLP | 9.30 | 4.30 | 3.76 | 199.89 | 100.00 |
| DAP | 9.39 | 4.37 | 3.85 | 211.13 | 95.83 |
| LPRA | 14.44 | 3.40 | 4.08 | 236.10 | 100.00 |
| WLP-STE | 12.21 | 4.22 | 2.87 | 186.65 | 91.67 |
| WLP-AME | 2.89 | 1.24 | 0.79 | 56.49 | 100.00 |

TABLE III. Formant estimation results for the synthetic [i] vowels. Each row corresponds to an all-pole modeling method evaluated (notation is described in Sec. III A). Relative errors in F1, F2, and F3, defined by Eq. (13), are given in columns 1–3, respectively. The Euclidean distance between the true and the estimated formant, defined by Eq. (14), is given in column 4. The last column includes the relative number of all-pole analyses indicating at least three formant peaks.

| | $d_{\mathrm{err},1}$ (%) | $d_{\mathrm{err},2}$ (%) | $d_{\mathrm{err},3}$ (%) | $d_{\mathrm{euc}}$ (Hz) | $n_{\mathrm{pks}}$ (%) |
|---|---|---|---|---|---|
| LP | 12.59 | 2.84 | 2.37 | 150.46 | 75.00 |
| RBLP | 16.27 | 3.18 | 2.25 | 155.95 | 66.67 |
| DAP | 12.35 | 3.01 | 1.93 | 145.95 | 66.67 |
| LPRA | 18.59 | 2.99 | 2.63 | 163.47 | 62.50 |
| WLP-STE | 11.39 | 2.83 | 2.12 | 134.36 | 83.33 |
| WLP-AME | 7.59 | 1.30 | 1.15 | 78.52 | 50.00 |

which it showed the second worst performance among the methods compared. The clearest result was, again, observed for the WLP-AME analysis which yielded the smallest estimation error for all four vowels and for all four error measures.

Finally, the relative number of those analyses, where an all-pole modeling method is able to predict all three lowest formants, i.e., $n_{\mathrm{pks}}$, varied between 50% and 100%. When the four vowels were combined, this value was between 80% and 90% for all six methods. It is worth emphasizing that WLP-AME found all three lowest resonances in each analyzed signal for three of the vowels ([a], [æ], and the neutral vowel). For the vowel [i], however, the close location of the second and third formant made it difficult for WLP-AME to distinguish the two resonances and they were quite often smeared into a single peak.

A representative example demonstrating the performance of WLP-AME in formant estimation is shown in Fig. 10. The figure, computed for the vowel [æ] involves two all-pole methods, conventional LP (thin curves), and WLP-AME (thick curves). Characteristic differences between the two methods can be seen when spectra computed from low-pitch vowels, shown by the lower curves, are compared to the spectra obtained from high-pitched sounds, depicted by the upper curves. When the three lowest F0 values are compared, the four lowest peaks occur at nearly the same frequency in both methods. For vowels of higher pitch, however, the F1 estimate drifts from its true value due to the biasing effect of F0 in the spectra computed by conventional LP. The WLP-AME spectra, however, are able to show formant peaks whose positions remain almost unchanged even though F0 varies from 100 to 450 Hz.

TABLE IV. Formant estimation results for the synthetic neutral vowels. Each row corresponds to an all-pole modeling method evaluated (notation is described in Sec. III A). Relative errors in F1, F2, and F3, defined by Eq. (13), are given in columns 1–3, respectively. The Euclidean distance between the true and the estimated formant, defined by Eq. (14), is given in column 4. The last column includes the relative number of all-pole analyses indicating at least three formant peaks.

| | $d_{\mathrm{err},1}$ (%) | $d_{\mathrm{err},2}$ (%) | $d_{\mathrm{err},3}$ (%) | $d_{\mathrm{euc}}$ (Hz) | $n_{\mathrm{pks}}$ (%) |
|---|---|---|---|---|---|
| LP | 10.26 | 4.26 | 1.74 | 145.92 | 100.00 |
| RBLP | 10.16 | 3.78 | 1.72 | 139.29 | 100.00 |
| DAP | 10.18 | 4.35 | 1.62 | 146.74 | 100.00 |
| LPRA | 10.99 | 4.35 | 2.84 | 178.96 | 100.00 |
| WLP-STE | 12.56 | 3.40 | 1.65 | 146.84 | 100.00 |
| WLP-AME | 3.00 | 0.63 | 0.39 | 33.09 | 100.00 |

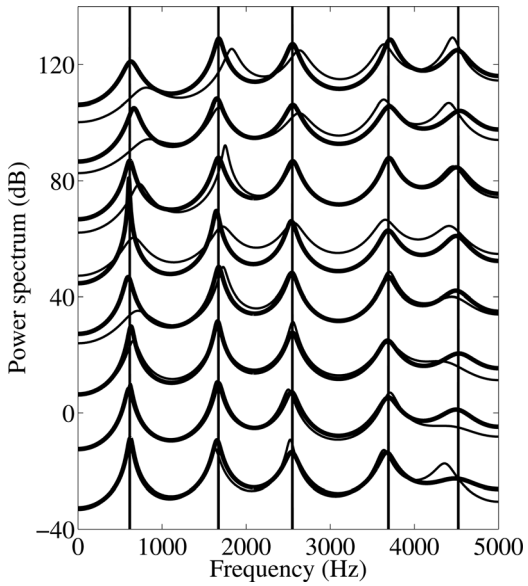


FIG. 10. All-pole spectra computed by conventional LP (thin curve) and WLP-AME (thick curve) from synthetic vowel sounds of different F0 values. F0 rises from 100 Hz (bottom pair of spectra) to 450 Hz (top pair of spectra) in steps of 50 Hz. Speech signals were synthesized with the physical modeling approach by using parameters that correspond to the production of the vowel [æ] by a male speaker. True formant values are shown by vertical lines.

### 2. Experiments with randomized values of $t_{me}$

In comparing WLP-AME to the selected all-pole modeling methods in case of incorrect positioning of the main excitation, results were combined in the same manner as in Sec. VII A 1 (i.e., F0 categories and data from the male, female, and child speaker were pooled together). Table V shows the five quantification methods as a function of parameter $r$ for the synthetic vowel [a]. To avoid expanding this section too much, the results of the other three vowels are not included as separate tables as in Sec. VII A 1 but the main observations are discussed.

The data indicates, as expected, that the estimation error increases when the value of $r$ is raised. In addition, the estimation error is again largest for F1. By comparing the formant estimation errors of Table V to the corresponding values in Table I, it can, however, be seen that the error in WLP-AME is smaller than that of the other all-pole modeling techniques even though correct values of $t_{me}$ were not used. In particular, the relative error in F1 varied between 3.04% and 6.60% when computed with WLP-AME with varying values of $r$, whereas the F1 error was larger than 10.49% when any of the other all-pole modeling methods was used. The same finding was observed for all the other vowels as well: The relative error in F1 was always smaller in WLP-AME than in the other all-pole methods. For WLP-AME, the maximum relative error in F1 was 9.04% which was obtained for the vowel [ae] with $r = 10$. For F2 and F3, the relative error in

TABLE V. Formant estimation results of WLP-AME as a function of the randomization parameter $r$ for the synthetic [a] vowels. The top row corresponds to using correct epochs, after which the maximum random error in the extraction of the main excitation, defined in Eq. (16), increases row by row. Relative errors in F1, F2, and F3, defined by Eq. (13), are given in columns 2–4, respectively. The Euclidean distance between the true and the estimated formant, defined by Eq. (14), is given in column 5. The last column includes the relative number of all-pole analyses indicating at least three formant peaks.

| $r$ (samples) | $d_{\mathrm{err},1}$ (%) | $d_{\mathrm{err},2}$ (%) | $d_{\mathrm{err},3}$ (%) | $d_{\mathrm{euc}}$ (Hz) | $n_{\mathrm{pks}}$ (%) |
|---|---|---|---|---|---|
| 0 | 3.04 | 1.70 | 0.48 | 52.30 | 100.00 |
| 2 | 4.01 | 1.71 | 0.88 | 67.82 | 100.00 |
| 4 | 5.66 | 2.77 | 0.82 | 79.84 | 100.00 |
| 6 | 6.44 | 3.67 | 1.25 | 100.58 | 100.00 |
| 8 | 6.41 | 3.49 | 1.35 | 99.41 | 96.00 |
| 10 | 6.60 | 5.01 | 1.20 | 110.32 | 88.00 |

WLP-AME was also smaller than that of any other method when $r < 10$. With $r = 10$, the relative error in F2 and F3 was approximately of the same magnitude as in the other selected all-pole methods.

Formant frequency estimation results are shown for COV and WLP-AME in Tables VI and VII, respectively, as a function of the randomization parameter $r$ by pooling together the four vowels, three speakers and eight F0 categories. Data in the tables indicate that the formant frequency estimates of both methods deteriorate when the random error in the extraction of $t_{me}$ increases. It can be observed, however, that the averaged formant estimation errors are smaller for WLP-AME in all three formants in each value of $r$. The performance of the COV analysis was better than that of LP, RBLP, DAP, LPRA, and WLP-STE reported separately for each vowel in Tables I–IV, but it failed to reach the performance of WLP-AME. The lower estimation accuracy of the COV analysis is explained by the absence of a sufficiently long closed phase particularly in the glottal area functions of the female and child speaker used in the physical modeling.

### B. Natural vowels

Formant frequencies estimated from the natural vowels by LP and WLP-AME are illustrated for F1 and F2 in Figs. 11 and 12, respectively. Formant frequency values are shown separately for each of the three vowels and for the male and female speakers. Furthermore, in order to demonstrate how the formants estimated by the two all-pole modeling methods behave when F0 is raised, F1 of the vowel [a] is shown as an example. The behavior of F1 of the vowel [a] is depicted in Fig. 13 separately for male (upper graph) and female (lower graph) talkers as a function of the repetition order in the vowel series of increasing pitch.

TABLE VI. Formant estimation results of COV as a function of the randomization parameter $r$, data averaged over all four synthetic vowels. The top row corresponds to using correct epochs, after which the maximum random error in the extraction of the main excitation, defined by Eq. (16), increases row by row. Relative errors in F1, F2, and F3, defined by Eq. (13), are given in columns 2–4, respectively. The Euclidean distance between the true and the estimated formant, defined by Eq. (14), is given in column 5. The last column includes the relative number of all-pole analyses indicating at least three formant peaks.

| $r$ (samples) | $d_{\mathrm{err},1}$ (%) | $d_{\mathrm{err},2}$ (%) | $d_{\mathrm{err},3}$ (%) | $d_{\mathrm{euc}}$ (Hz) | $n_{\mathrm{pks}}$ (%) |
|---|---|---|---|---|---|
| 0 | 6.70 | 4.17 | 1.88 | 102.85 | 80.00 |
| 2 | 6.44 | 4.18 | 2.20 | 135.54 | 82.00 |
| 4 | 8.37 | 4.46 | 2.09 | 134.97 | 79.00 |
| 6 | 10.24 | 5.13 | 2.76 | 131.79 | 71.00 |
| 8 | 11.21 | 4.94 | 2.10 | 152.14 | 81.00 |
| 10 | 10.46 | 4.73 | 2.60 | 167.90 | 76.00 |

TABLE VII. Formant estimation results of WLP-AME as a function of the randomization parameter $r$, data averaged over all four synthetic vowels. The top row corresponds to using correct epochs, after which the maximum random error in the extraction of the main excitation, defined by Eq. (16), increases row by row. Relative errors in F1, F2, and F3, defined by Eq. (13), are given in columns 2–4, respectively. The Euclidean distance between the true and the estimated formant, defined by Eq. (14), is given in column 5. The last column includes the relative number of all-pole analyses indicating at least three formant peaks.

| $r$ (samples) | $d_{\mathrm{err},1}$ (%) | $d_{\mathrm{err},2}$ (%) | $d_{\mathrm{err},3}$ (%) | $d_{\mathrm{euc}}$ (Hz) | $n_{\mathrm{pks}}$ (%) |
|---|---|---|---|---|---|
| 0 | 4.13 | 1.22 | 0.70 | 55.10 | 88.00 |
| 2 | 4.71 | 1.41 | 0.85 | 66.34 | 90.00 |
| 4 | 6.14 | 2.16 | 1.18 | 88.07 | 91.00 |
| 6 | 6.17 | 2.63 | 1.24 | 95.19 | 93.00 |
| 8 | 7.25 | 2.84 | 1.53 | 111.49 | 90.00 |
| 10 | 7.84 | 3.59 | 1.79 | 125.91 | 93.00 |

Statistical analyses were performed for the natural vowels with the linear mixed-effects models (Baayen *et al.*, 2008; R Development Core Team, 2009), where effects with absolute $t$-values larger than 2 were interpreted to be significant. The analysis was computed in two parts. In the first one, the main goal was to test for the possible effect of the all-pole modeling method (LP, WLP-AME) on the value of F1 and F2. In the second test, the primary goal was to find out whether the two all-pole methods show a statistically significant difference in the value of $\delta_i$ defined in Eq. (15).

In the first statistical analysis, effects of the *method* (LP, WLP-AME), *gender* (male and female), *vowel* ([a], [æ], [i]), and *F*0 (repetition number) were computed separately for F1 and F2. In both analyses, the vowel [a], the female gender, and the LP method was used as an intercept against which the significance of the formant value was tested. The results for F1 and F2 are summarized, respectively, in Tables VIII and IX. As expected, these data indicate that both F1 and F2 were significantly related to the vowel and gender, a finding that is also demonstrated in Fig. 11 and Fig. 12. However, the method did not show a significant effect on either F1 or F2. In addition, there was a strong effect of F0 ($t = 7.61$ and $t = 3.30$ in F1 and F2, respectively) indicating that the formant frequencies were higher in high-pitched vowels. This finding is demonstrated for F1 of the vowel [a] in Fig. 13. The effect of F0 further interacts with vowel ($t = -4.04$ for [i]:repetition in F1; $t = -3.08$ for [æ]:repetition in F2; $t = -7.38$ for [i]:repetition in F2). With both formants, there is also a strong male:[i]:repetition interaction depicting different effect of F0 between the genders.

In the second statistical test, $\delta_i$ was used as the dependent measure while the predictors were *method* (LP, WLP-AME), *vowel* ([a], [æ], [i]), speaker's *gender* (male, female),

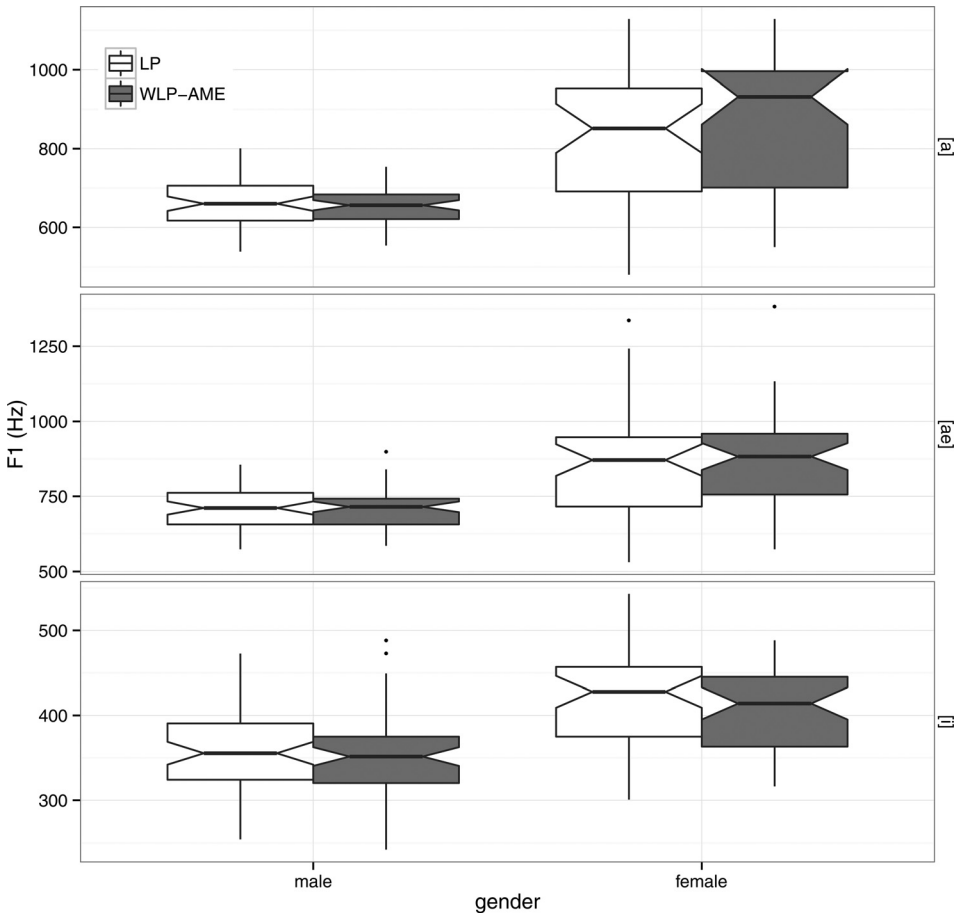


FIG. 11. Frequency of F1 estimated by LP (white boxplot) and WLP-AME (gray boxplot) from natural vowels. Data computed from the vowel [a], [æ], and [i] are illustrated, respectively, in the upper, middle, and lower panel. F1 values are shown separately for the male and female talkers.

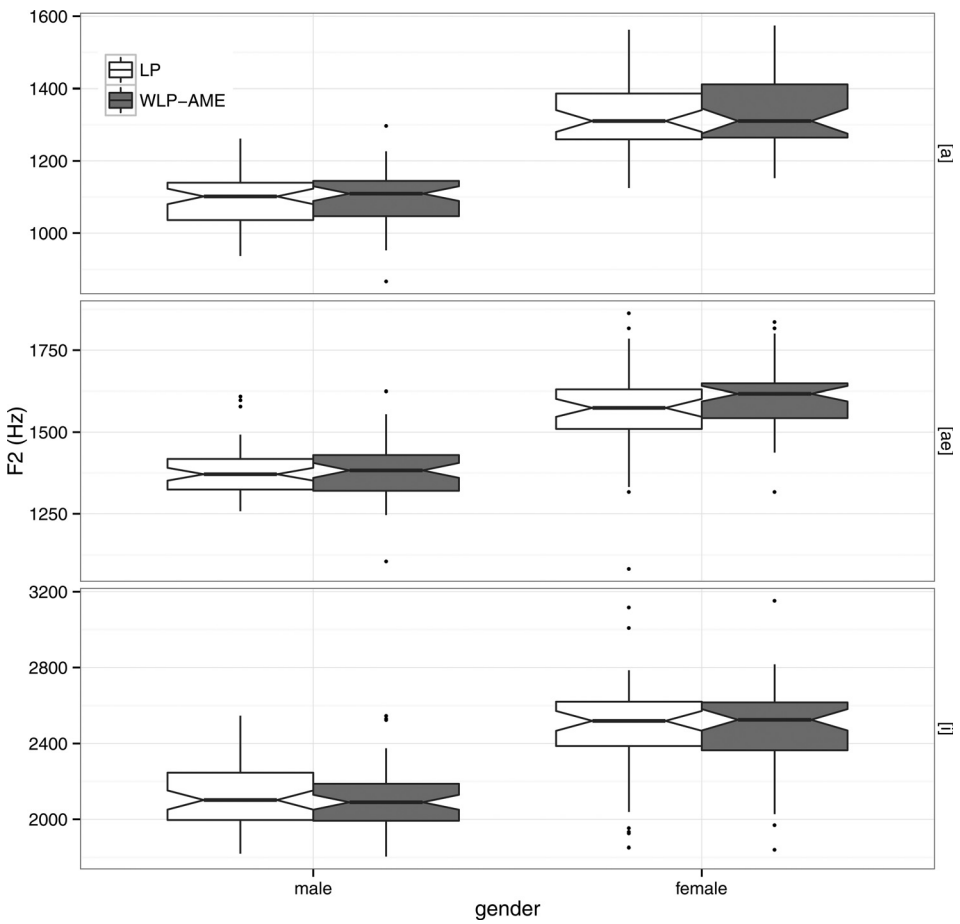


FIG. 12. Frequency of F2 estimated by LP (white boxplot) and WLP-AME (gray boxplot) from natural vowels. Data computed from the vowel [a], [æ], and [i] are illustrated, respectively, in the upper, middle, and lower panel. F2 values are shown separately for the male and female talkers.

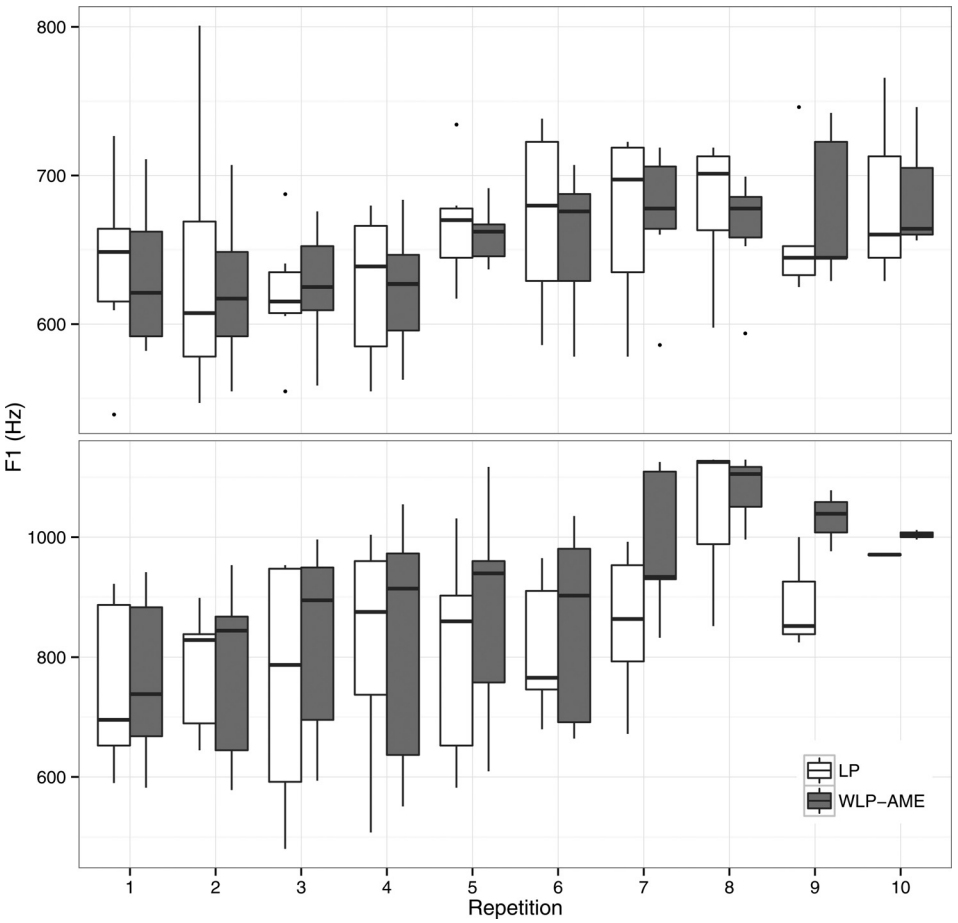


FIG. 13. Frequency of F1 of the vowel [a] estimated by LP (white boxplot) and WLP-AME (gray boxplot) as a function of the repetition order in producing vowels of increasing pitch. Data are shown separately for the (top) male and (bottom) female talkers.

and *formant* (F1, F2). The linear mixed-effects models were estimated using the vowel [a], the first formant, the LP method, and the female gender as an intercept against which the significances were tested. A model with all values (i.e., all vowels and both formants) was fitted to see the global effect of the method. Likelihood ratio tests (analysis of variance function in R) showed that by-subject random slopes for the fixed-effect predictors (vowel, formant, gender) did not improve the fit significantly. Results from the best-fitting model are summarized in Table X.

As can be seen from Table X, the WLP-AME method differs significantly from LP ($t = -2.04$). There is also a clear effect of gender with males showing significantly lower values ($t = -5.74$). In addition, the vowel [i]:F2 interaction is significant ($t = 3.79$), but the [æ]:F2 interaction is not significant.

Finally, two examples are shown demonstrating the behavior of all-pole spectra computed by LP and WLP-AME from natural vowels of increasing F0. The examples were computed from the repetition series of the vowel [æ]

TABLE VIII. Results from statistical analyses using the mixed-effects modeling on the frequency of F1 of natural vowels. The intercept stands for the female gender, the vowel [a], and the LP method.

| | Estimate | Standard error | *t* value |
|---|---|---|---|
| Intercept | 748.27 | 23.14 | 32.34 |
| male | −130.57 | 20.63 | −6.33 |
| [æ] | −5.36 | 21.28 | −0.25 |
| [i] | −362.00 | 21.51 | −16.83 |
| repetition | 25.31 | 3.33 | 7.61 |
| WLP-AME | 9.90 | 6.45 | 1.54 |
| male:[æ] | 48.84 | 28.70 | 1.70 |
| male:[i] | 56.74 | 28.77 | 1.97 |
| male: repetition | −18.70 | 4.18 | −4.47 |
| [æ]:repetition | 1.58 | 4.51 | 0.35 |
| [i]:repetition | −19.55 | 4.84 | −4.04 |
| male:[æ]:repetition | 0.66 | 5.79 | 0.11 |
| male:[i]:repetition | 20.59 | 5.97 | 3.45 |

TABLE IX. Results from statistical analyses using the mixed-effects modeling on the frequency of F2 of natural vowels. The intercept stands for the female gender, the vowel [a], and the LP method.

| | Estimate | Standard error | *t* value |
|---|---|---|---|
| Intercept | 1258.28 | 31.02 | 40.56 |
| male | −212.59 | 32.94 | −6.45 |
| [æ] | 331.26 | 33.98 | 9.75 |
| [i] | 1352.61 | 34.05 | 39.72 |
| repetition | 17.52 | 5.31 | 3.30 |
| WLP-AME | 9.21 | 10.26 | 0.90 |
| male:[æ] | 12.95 | 45.84 | 0.28 |
| male:[i] | −336.18 | 45.73 | −7.35 |
| male: repetition | −9.09 | 6.68 | −1.36 |
| [æ]:repetition | −22.22 | 7.21 | −3.08 |
| [i]:repetition | −54.93 | 7.44 | −7.38 |
| male:[æ]:repetition | 9.56 | 9.25 | 1.03 |
| male:[i]:repetition | 57.95 | 9.31 | 6.23 |

TABLE X. Results from statistical analyses using the mixed-effects modeling on the value of $\delta_i$ defined by Eq. (15) for natural vowels. The intercept stands for the female gender, the vowel [a], the LP method, and the first formant.

| | Estimate | Standard error | *t* value |
|---|---|---|---|
| Intercept | 90.77 | 21.22 | 4.28 |
| [æ] | 14.44 | 28.31 | 0.51 |
| [i] | −28.26 | 28.31 | −1.00 |
| F2 | 3.43 | 28.31 | 0.12 |
| WLP-AME | −17.42 | 8.54 | −2.04 |
| male | −49.03 | 8.54 | −5.74 |
| [æ]:F2 | −0.44 | 40.04 | −0.01 |
| [i]:F2 | 151.54 | 40.04 | 3.79 |

produced by a male and a female talker and these data are shown, respectively, in Fig. 14 and Fig. 15. For visual clarity, spectra of both all-pole modeling methods are shown only for four consecutive samples in the repetition series of increasing F0. The figures demonstrate that F1 and F2 can be identified in both LP and WLP-AME spectra as clear local peaks. In addition, these resonances are located approximately in the same positions in the spectra computed by the two methods, a finding that was also observed in the statistical tests summarized in Tables VIII and IX. It can be seen, however, that the formant frequencies computed by LP change inconsistently when F0 is raised: Formants either move up or down in frequency when pitch increases. One possible explanation for this behavior is the sensitivity of LP to the biasing effect of F0. The corresponding peaks indicated by WLP-AME, however, follow a more regular trend in which the formant peak in general rises when F0 increases.

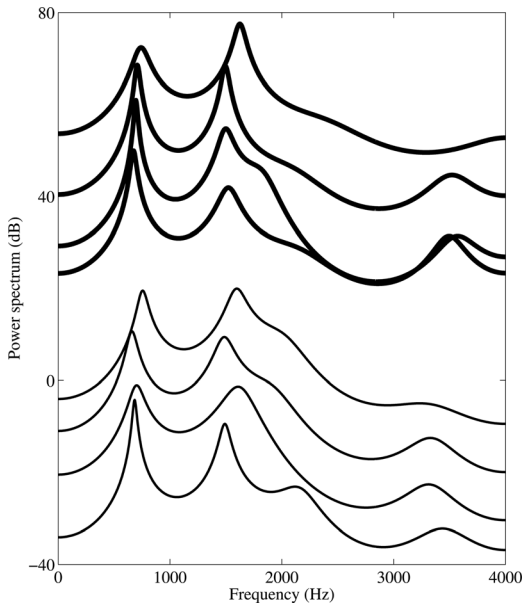


FIG. 14. Examples of all-pole spectra computed by conventional LP (thin curve) and WLP-AME (thick curve) for four natural vowels of increasing F0. F0 rises from 175 Hz to 260 Hz corresponding, respectively, to the bottom and top spectrum in both groups of four spectra. Spectra were computed from the natural [æ] vowels produced by a male speaker.

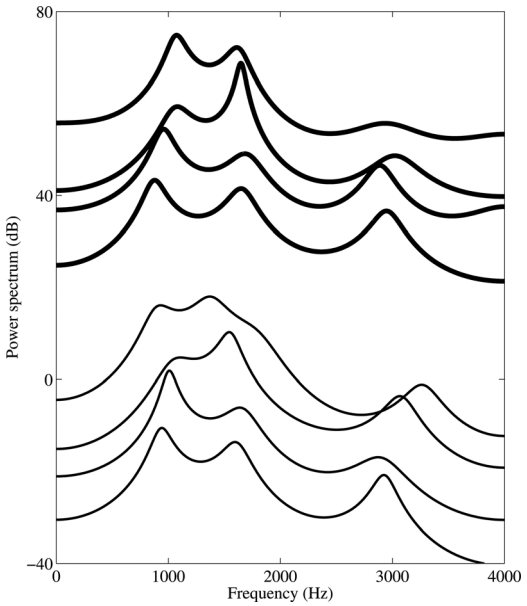


FIG. 15. Examples of all-pole spectra computed by conventional LP (thin curve) and WLP-AME (thick curve) for four natural vowels of increasing F0. F0 rises from 335 Hz to 445 Hz corresponding, respectively, to the bottom and top spectrum in both groups of four spectra. Spectra were computed from the natural [æ] vowels produced by a female speaker.

## VIII. DISCUSSION

All-pole modeling is a widely used parametric spectrum estimation method, but its performance in formant estimation is known to deteriorate for high-pitched sounds. In particular, accurate estimation of the lowest formants is known to be difficult due to the biasing effect caused by the sparse harmonic structure of high-pitched sounds. In order to tackle this problem, several all-pole modeling methods which are robust with respect to F0 have been proposed during the past decades. This study analyzed five such previously known methods and proposed a new technique, WLP-AME.

WLP-AME is based on temporally weighted linear prediction in which the square of the prediction error is multiplied by a given parametric time-domain weighting function. This weighting function is constructed using a waveform whose value is one at all time instants during the fundamental period except in the vicinity of the time instant of the main acoustical excitation of the vocal tract when it drops close to zero. With this weighting function, the contribution of those speech samples that are greatly affected by the glottal excitation can be diminished in the computation of the optimal filter coefficients. Consequently, the resulting all-pole model will be affected more by the spectral characteristics of the vocal tract, which leads to less biased formant estimates in the analysis of high-pitch speech.

The study first compared the performance of five previous methods and that of WLP-AME in formant estimation of synthetic vowels. Synthetic vowels were created with a physical modeling approach in order to get test material whose production mechanisms are different from those assumed in the all-pole models evaluated. Results indicated that for the great majority of the cases WLP-AME yielded formant estimation errors that were smaller than any of those

computed by the five previously known methods. WLP-AME was also shown to be robust with respect to the extraction of the glottal closure instant: The mean formant error was better than that of the other five all-pole modeling methods utilizing the autocorrelation criterion and also better than that of COV analysis even though the extracted instant of glottal closure was deliberately distorted.

In addition, WLP-AME was compared with the most widely used all-pole modeling technique, conventional LP, by estimating formants from natural vowels with increasing F0. The analysis was limited to the lowest two formants that could be reliably obtained from the all-pole models. Results indicated, as expected, that both F1 and F2 increased significantly when pitch was raised. However, there was no significant difference between LP and WLP-AME in the frequency of either F1 or F2 when the corresponding formant frequencies were pooled together from all repetitions of different F0 values. In other words, the two methods indicated F1 and F2 in the frequency positions that were, on average, not different. Despite this, there was, however, a statistically significant difference between the two methods when the formant frequency contour was analyzed between consecutive repetitions of vowels with increasing pitch. This analysis indicated, interestingly, that the jitter in the frequency of F1 and F2, estimated from series of vowels with increasing F0, was significantly smaller in WLP-AME than in LP. A plausible explanation why F1 and F2 estimated by WLP-AME increased in a more regular manner as a function of F0 in comparison to resonances given by LP is the reduced biasing effect of F0 to the WLP-AME models.

Attenuating the effect of the glottal excitation was computed in WLP-AME using a conceptually simple temporal weighting function. In the present study, the parameters of the AME function were first optimized with synthetic vowel data and their values were kept unchanged in the rest of the analyses. Since the shape of the glottal flow in natural speech is known to change remarkably due to, for example, adjustments in vocal intensity (e.g., Holmberg *et al.*, 1988) or the type of phonation (e.g., Alku and Vilkman, 1996), a justified topic of future work would be to search for methods to adapt the AME function. In addition, it is noted that all the linear predictive analyses involved in the present study are classical in the sense that they use time-invariant filter coefficients that are updated once per frame. A more flexible paradigm is to utilize time-varying AR-modeling (e.g., Schnell and Lacroix, 2008; Rudoy *et al.*, 2011) in which linear predictive filter coefficients evolve in time. Combining the proposed WLP-AME method with the time-varying AR modeling approach is another topic of future studies which would maybe help in detecting vocal tract variation in continuous high-pitched speech.

In conclusion, this study shows that the performance of all-pole modeling in formant estimation of high-pitched speech is improved when WLP-AME is used instead of conventional LP or the selected group of previously developed robust all-pole modeling techniques. WLP-AME has several features that are similar to those used in the conventional LP analysis: Both methods use a similar kind of a frame structure and they can be implemented either with the autocorrelation or covariance criterion. In addition, the computational complexity of WLP-AME is reasonably low provided that the instants of the glottal closure are known. Moreover, WLP-AME is similar to LP in the sense that there is no need to assume that the fundamental frequency is changing during the analysis frame, a pre-requisite that is required, for example, in the techniques proposed by Wang and Quatieri (2010) and Shiga and King (2003). Hence, in principle, WLP-AME could be easily used as an alternative to conventional LP as a method to compute all-pole spectra in formant estimation. There are, however, two differences that might limit the use of WLP-AME. The method, like several of its counterparts such as RBLP and DAP, does not guarantee the stability of the all-pole model. In addition, WLP-AME calls for identifying the instants of glottal closures in order to build the weighting function. If EGG is available, as in the present study, the detection of closure instants can be easily computed from the differentiated EGG. If, however, the speech pressure signal alone is available, the WLP-AME analysis needs to be combined with an epoch extraction method, such as the DYPSA algorithm (Naylor *et al.*, 2007) or the group delay-based method (Murthy and Yegnanarayana, 1999), capable of estimating the glottal closure instants from the acoustic speech signal.

## IX. CONCLUSIONS

WLP-AME was proposed as a linear predictive method to estimate formants from high-pitched speech. The method downgrades the contribution of the main excitation of the glottal source with a straightforward temporal weighting function, which leads to less biased formant estimates in the analysis of high-pitched speech. WLP-AME was shown to yield improved formant frequency estimates for high-pitched vowels in comparison to the previously known methods and the method was shown to be robust with respect to the extraction of the glottal closure instant.

## ACKNOWLEDGMENTS

This work was supported by the Academy of Finland (projects 256961 and 135003, the LASTU Research Programme) and the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 287678 (the Simple4all project).

Alku, P., Airas, M., Björkner, E., and Sundberg, J. (**2006**). "An amplitude quotient based method to analyze changes in the shape of the glottal pulse in the regulation of vocal intensity," J. Acoust. Soc. Am. **120**, 1052–1062.

Alku, P., Magi, C., Yrttiaho, S., Bäckström, T., and Story, B. (**2009**). "Closed phase covariance analysis based on constrained linear prediction for glottal inverse filtering," J. Acoust. Soc. Am. **120**, 3289–3305.

Alku, P., and Vilkman, E. (**1996**). "A comparison of glottal voice source quantification parameters in breathy, normal, and pressed phonation of female and male speakers," Folia Phoniatr. Logop. **48**, 240–254.

Baayen, R., Davidson, D., and Bates, D. (**2008**). "Mixed-effects modeling with crossed random effects for subjects and items," J. Mem. Lang. **59**, 390–412.

Deng, L., Lee, L., Attias, H., and Acero, A. (**2007**). "Adaptive Kalman filtering and smoothing for tracking vocal tract resonances using a continuous-

valued hidden dynamic model," IEEE Trans. Audio Speech Lang. Process. **15**, 13–23.
El-Jaroudi, A., and Makhoul, J. (**1991**). "Discrete all-pole modeling," IEEE Trans. Signal Process. **39**, 411–423.
Fant, G. (**1970**). *Acoustic Theory of Speech Production* (Mouton, The Hague), pp. 15–26.
Fant, G., Liljencrants, J., and Lin, Q. (**1985**). "A four-parameter model of glottal flow," *STL-QPSR 4* (Speech, Music and Hearing, Royal Institute of Technology, Stockholm, Sweden), pp. 1–13.
Fröhlich, M., Michaelis, D., and Strube, H. (**2001**). "SIM—Simultaneous inverse filtering and matching of a glottal flow model for acoustic speech signals," J. Acoust. Soc. Am. **110**, 479–488.
Gobl, C. (**1989**). "A preliminary study of acoustic voice quality correlates," *STL-QPSR 4* (Speech, Music and Hearing, Royal Institute of Technology, Stockholm, Sweden), pp. 9–22.
Gold, B., and Rabiner, L. (**1968**). "Analysis of digital and analog formant synthesizers," IEEE Trans. Audio Electroacoust. **16**, 81–94.
Golub, G., and Van Loan, C. (**1983**). *Matrix Computation* (Johns Hopkins University Press, Baltimore, MD), p. 55.
Hagiwara, R. (**1997**). "Dialect variation and formant frequency: The American English vowels revisited," J. Acoust. Soc. Am. **102**, 655–658.
Hermansky, H., Fujisaki, H., and Sato, Y. (**1984**). "Spectral envelope sampling and interpolation in linear predictive analysis of speech," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, San Diego, CA, pp. 2.2.1–2.2.4.
Hillenbrand, J., Getty, L., Clark, M., and Wheeler, K. (**1995**). "Acoustic characteristics of American English vowels," J. Acoust. Soc. Am. **97**, 3099–3111.
Holmberg, E., Hillman, R., and Perkell, J. (**1988**). "Glottal airflow and transglottal air pressure measurements for male and female speakers in soft, normal, and loud voice," J. Acoust. Soc. Am. **84**, 511–529.
Krishnamurthy, A., and Childers, D. (**1986**). "Two-channel speech analysis," IEEE Trans. Acoust. Speech Signal Process. **34**, 730–743.
Lee, C.-H. (**1988**). "On robust linear prediction of speech," IEEE Trans. Acoust. Speech Signal Process. **36**, 642–650.
Liljencrants, J. (**1985**). "Speech synthesis with a reflection-type line analog," DS dissertation, Dep. of Speech Comm. and Music Acoustics, Royal Inst. of Technol., Stockholm, Sweden, pp. 1–125.
Ma, C., Kamp, Y., and Willems, L. (**1993**). "Robust signal selection for linear prediction analysis of voice speech," Speech Commun. **12**, 69–81.
Magi, C., Pohjalainen, J., Bäckström, T., and Alku, P. (**2009**). "Stabilised weighted linear prediction," Speech Commun. **51**, 401–411.
Makhoul, J. (**1975a**). "Linear prediction: A tutorial review," Proc. IEEE **63**, 561–580.
Makhoul, J. (**1975b**). "Spectral linear prediction: Properties and applications," IEEE Trans. Acoust. Speech Signal Process. **23**, 283–296.
Markel, J., and Gray, A., Jr. (**1976**). *Linear Prediction of Speech* (Springer-Verlag, Berlin), pp. 1–288.
Miyoshi, Y., Yamato, K., Mizoguchi, R., Yanagida, M., and Kakusho, O. (**1987**). "Analysis of speech signals of short pitch period by a sample-selective linear prediction," IEEE Trans. Acoust. Speech Signal Process. **35**, 1233–1240.
Murthy, P. S., and Yegnanarayana, B. (**1999**). "Robustness of group-delay-based method for extraction of significant instants of excitation from speech signals," IEEE Trans. Speech Audio Process. **7**, 609–619.
Naylor, P., Kounoudes, A., Gudnason, J., and Brookes, M. (**2007**). "Estimation of glottal closure instants in voiced speech using the DYPSA algorithm," IEEE Trans. Speech Audio Process. **15**, 34–43.
Olive, J. (**1971**). "Automatic formant tracking by a Newton-Raphson technique," J. Acoust. Soc. Am. **50**, 661–670.
Oppenheim, A., and Schafer, R. (**1989**). *Discrete-Time Signal Processing* (Prentice-Hall, Englewood Cliffs, NJ), pp. 33–39.
Plumpe, M., Quatieri, T., and Reynolds, D. (**1999**). "Modeling of the glottal flow derivative waveform with application to speaker identification," IEEE Trans. Speech Audio Process. **7**, 569–586.
Potamianos, A., and Maragos, P. (**1996**). "Speech formant frequency and bandwidth tracking using multiband energy demodulation," J. Acoust. Soc. Am. **99**, 3795–3806.
Rabiner, L., and Schafer, R. (**1978**). *Digital Processing of Speech Signals* (Prentice-Hall, Englewood Cliffs, NJ), Chap. 8, pp. 403–404.
Rahman, S., and Shimamura, T. (**2007**). "Linear prediction using refined autocorrelation function," EURASIP J. Audio Speech Music Process. **45962**, 1–9.
R Development Core Team (**2009**). *A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria), pp. 1–409.
Rudoy, D., Quatieri, T., and Wolfe, P. (**2011**). "Time-varying autoregressions in speech: Detection theory and applications," IEEE Trans. Audio Speech Lang. Process. **19**, 977–989.
Saeidi, R., Pohjalainen, J., Kinnunen, T., and Alku, P. (**2010**). "Temporally weighted linear prediction features for tackling additive noise in speaker verification," IEEE Signal Process. Lett. **17**, 599–602.
Schnell, K., and Lacroix, A. (**2008**). "Time-varying linear prediction for speech analysis and synthesis," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Las Vegas, Nevada, USA, pp. 3941–3944.
Shiga, Y., and King, S. (**2003**). "Estimating the spectral envelope of voiced speech using multi-frame analysis," in *Proceedings of Interspeech*, Geneva, Switzerland, pp. 1737–1740.
Story, B. (**1995**). "Speech simulation with an enhanced wave-reflection model of the vocal tract," Ph.D. thesis, University of Iowa, Iowa City, pp. 1–352.
Story, B. (**2005**). "Synergistic modes of vocal tract articulation for American English vowels," J. Acoust. Soc. Am. **118**, 3834–3859.
Story, B. (**2006**). "A technique for 'tuning' vocal tract area functions based on acoustic sensitivity functions," J. Acoust. Soc. Am. **119**, 715–718.
Story, B. (**2008**). "Comparison of magnetic resonance imaging-based vocal tract area functions obtained from the same speaker in 1994 and 2002," J. Acoust. Soc. Am. **123**, 327–335.
Story, B. (**2013**). "Phrase-level speech simulation with an airway modulation model of speech production," Comp. Speech Lang. **27**, 989–1010.
Strube, H. (**1974**). "Determination of the instant of glottal closure from the speech wave," J. Acoust. Soc. Am. **56**, 1625–1629.
Titze, I. (**1984**). "Parameterization of the glottal area, glottal flow, and vocal fold contact area," J. Acoust. Soc. Am. **75**, 570–580.
Titze, I. (**2002**). "Regulating glottal airflow in phonation: Application of the maximum power transfer theorem to a low dimensional phonation model," J. Acoust. Soc. Am. **111**, 367–376.
Titze, I. (**2006**). *The Myoelastic Aerodynamic Theory of Phonation* (National Center for Voice and Speech, Iowa City, IA), pp. 1–430.
Vallabha, G., and Tuller, B. (**2002**). "Systematic errors in the formant analysis of steady-state vowels," Speech Commun. **38**, 141–160.
Wang, T., and Quatieri, T. (**2010**). "High-pitch formant estimation by exploiting temporal change of pitch," IEEE Trans. Audio, Speech, Lang. Process. **18**, 171–186.
Wong, D., Markel, J., and Gray, A., Jr. (**1979**). "Least squares glottal inverse filtering from the acoustic speech waveform," IEEE Trans. Acoust. Speech Signal Process. **27**, 350–355.
Yegnanarayana, B., and Veldhuis, N. (**1998**). "Extraction of vocal-tract system characteristics from speech signals," IEEE Trans. Speech Audio Process. **6**, 313–327.