Improved formant frequency estimation from high-pitched vowels by downgrading the contribution of the glottal source with weighted linear prediction

Paavo Alku[1], *Jouni Pohjalainen*[1], *Martti Vainio*[2], *Anne-Maria Laukkanen*[3], *Brad Story*[4]

[1] Department of Signal Processing and Acoustics, Aalto University, Finland
[2]Department of Speech Sciences, University of Helsinki, Finland
[3]Department of Speech Communication and Voice Research, University of Tampere, Finland
[4]Speech Acoustics Laboratory, University of Arizona, USA

paavo.alku@aalto.fi

Abstract

Since performance of conventional linear prediction (LP) deteriorates in formant estimation of high-pitched voices, several all-pole modeling methods robust to F0 have been developed. This study compares five such previously known methods and proposes a new technique, Weighted Linear Prediction with Attenuated Main Excitation (WLP-AME). WLP-AME utilizes weighted linear prediction in which the square of the prediction error is multiplied with a weighting function that downgrades the contribution of the glottal source in the model optimization. Consequently, the resulting all-pole model is affected more by the vocal tract characteristics, which leads to more accurate formant estimates. By using synthetic vowels created with a physical modeling approach, the study shows that WLP-AME yields improved formant frequency estimates for high-pitched vowels in comparison to the previously known methods.

Index Terms: formants, linear prediction

1. Introduction

Linear prediction (LP) is a spectral estimation technique that is widely used in estimation of the vocal tract resonances, the formants. LP is well-suited for formant estimation due to its close connection to the source-filter theory of speech production [1]. However, the performance of conventional LP is known to deteriorate for high-pitched speech [2], [3]. In particular, the estimates of the lowest formants are biased by the spectral components generated by the glottal source, F0 and its harmonics, due to the error criterion used in conventional LP.

Modifications to LP have been proposed in many studies in order to compute all-pole models that are less affected by F0 and its harmonics. Lee [4] studied different cost functions that give more weight to small residual samples while down-weighting the prediction error samples of large amplitude. El-Jaroudi and Makhoul [3] utilized the Itakura-Saito distortion measure in the model optimization. Ma *et al*. [5] developed a method, Weighted Linear Prediction (WLP), in which a temporal weighting function is used in order to reduce the biasing effect of F0. Rahman and Shimamura [6] suggested a cepstral domain method to remove the biasing effect of F0. Finally, Wang and Quatieri [7] proposed recently a method in which the temporal change of pitch is exploited to improve the spectral sampling of the vocal tract resonances.

In this study, several linear predictive methods are compared in formant frequency estimation of high-pitched vowels. In addition, a new weighting function is proposed for WLP that downgrades the contribution of the glottal source and thereby yields formant estimates that are less biased by F0 and its harmonics. The study utilizes physical modelling of voice production in order to synthesize test vowels with known formant frequencies. With this approach, test signals are generated by a physical law, rather than by a parametric digital model similar to the all-pole model assumed in LP.

2. Weighted linear prediction

Weighted Linear Prediction (WLP) is a method for computing all-pole models by temporally weighting the square of the residual in the filter optimization. Hence, the contribution of certain pre-selected samples that are considered to be less desirable can be de-emphasized in the optimization. In formant estimation, an over-active role of the glottal source, which takes place when the vocal folds vibrate rapidly in high-pitched speech, can be regarded as this kind of a less desirable phenomenon. Hence, WLP is a justified choice in searching for linear predictive techniques robust to the effects of high pitch.

The optimization of the WLP model can be expressed according to [5] 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, and $a_k (1 \le k \le p)$ are the predictor coefficients. The residual energy is minimized between indices n_1 and n_2. In the autocorrelation method, $n_1 = 1$ and $n_2 = N + p$, and the signal is assumed to be zero outside [1, *N*]. The optimal WLP filter is determined by setting the partial derivatives of Eq. 1 with respect to each a_k to zero. This results in the WLP normal equations

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

Equation 2 can also be expressed in matrix form as

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

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

In [5], temporal weighting was computed from speech by using the short-time energy (STE) function

$$W_n = \sum_{i=0}^{M-1} s_{n-1-i}^2 \quad , \tag{4}$$

where M denotes the length of the energy window. STE 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 [8].

The excitation of the glottal source is most prominent at the instant of glottal closure. Therefore, one could argue that the biasing effect of F0 could be decreased by using WLP with a weighting function that downgrades the contribution of samples that are located near the instant of closure. One such simple weighting function, denoted as the Attenuated Main Excitation (AME), is shown in Fig. 1 together with a synthetic glottal flow derivative pulse. The AME window has one amplitude parameter, denoted by *d* in Fig. 1, that determines the level of attenuation. In addition, the function uses two relative time-domain parameters: (1) duration quotient $(DQ = (T_1/T)\cdot 100\%)$ and (2) position quotient $(PQ = (T_2/T_1)\cdot 100\%)$. In order to avoid abrupt changes, the function follows a linear ramp which is set to a constant value (of 3 samples). By using the AME window as the weighting function W_n in Eq. 3, a new all-pole method, denoted as WLP-AME, is obtained. The WLP-AME calls for identifying the glottal closure instants, indicated by t_{me} in Fig. 1. This calls for using either electroglottography (EGG) or epoch extraction methods such as that described in [9].

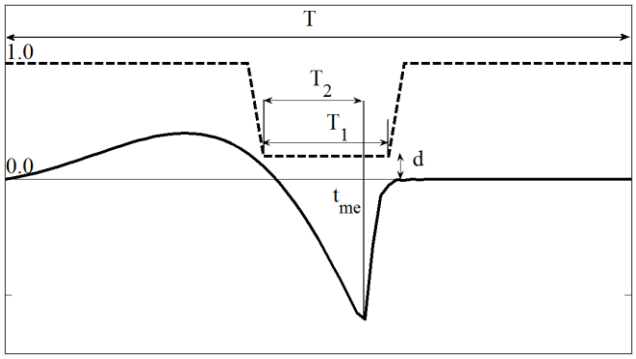

Figure 1: *Waveform of the AME function (dashed) together with a differentiated glottal flow (solid) synthesized with the LF-model. Parameters of the AME function correspond to the fundamental period T, the duration T_1 of the attenuated section, the time T_2 between the beginning of the attenuated section and the position of the main excitation t_{me}, and the amplitude d of the attenuated section.*

In the present study, the parameters of the AME function were optimized as follows. First, a set of synthetic vowels were generated by using the Liljencrants-Fant (LF) [10] waveform as an excitation. LF parameters were varied to create four phonation modes (modal, breathy, whispery, and creaky) according to [11]. F0 values of the excitation signals were varied between 100 Hz and 450 Hz with an increment of 50 Hz. All-pole vocal tract models of ten synthetic English vowels were created according to [12]. Second, WLP-AME analysis was performed for all the sounds by varying the parameters shown in Fig. 1 as follows: (a) six different values were used for *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%). Third, frequencies of the lowest four formants were computed from each WLP-AME filter and the estimated values were compared with the true ones by computing the relative formant error measure [6]. Finally, the parameter vector yielding the minimum formant error was sought for resulting in the following optimal values of the AME function: *d* = 0.01, *DQ* = 40%, and *PQ* = 80%. These settings were then used in all WLP-AME analyses of this study.

3. Experiments

The goal of the present study was to evaluate the performance of the WLP-based techniques described in section 2 in formant frequency estimation of high-pitched speech and compare WLP with several all-pole modeling methods. The selected methods are described next in this section. After this, the synthetic speech material is described.

3.1. All-pole modeling methods to be compared

The following all-pole methods were compared: (1) conventional LP [2] , (2) Robust Linear Prediction (RBLP) [4], (3) Discrete All-pole Modeling (DAP) [3], (4) Linear Prediction using Refined Autocorrelation (LPRA) [6], (5) Weighted Linear Prediction with the Short-Time Energy weighting function (WLP-STE) [5], and (6) Weighted Linear Prediction with Attenuated Main Excitation (WLP-AME) developed in this study.

All the analyses were computed with the autocorrelation criterion using a frame length of 25 ms, Hamming windowing, and a first order all-zero pre-emphasis with zero at *z* = 0.97. The prediction order was set to *p* = 10. In addition to the prediction order, the methods to be compared have different parameters whose optimal values were selected according to previous studies. RBLP was computed by using Huber's psi-function with *c* = 1.5 and the Iterative Reweighted Least Squares Algorithm [4]. DAP was implemented by using the α value of 0.6 and 20 iterations [3]. LPRA was computed according to [6] 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. WLP-STE was computed as in [5] by setting *M* = *p*. In WLP-AME, glottal closure was estimated as the instant of the negative peak of the glottal area function that was available for the test vowels (see section 3.2).

3.2. Test vowels synthesized with physical modeling

Accuracy assessment of a formant estimation method calls for using synthetic speech with known formant values. Therefore, previous investigations have exclusively utilized voices generated 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 present study takes advantage of a different approach, the physical modelling of the speech production mechanism, in generation of the test utterances with known formant frequency values.

A computational model of the speech production system was used to generate vowels representative of an adult male, adult female, and approximately a five year-old child. The voice source component used consisted of a kinematic representation of the medial surface of the vocal folds [13, 14] for which the F0, 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 [15]. 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 was specified by an area function representative of an [a], [i], and [æ] vowel. The area functions were taken from [16] for the adult male vowels and from [17] for the adult female vowels. The child-like area functions were not measured directly but rather generated with an acoustic sensitivity function approach [18] 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 [19]. 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 that included energy losses due to yielding walls, viscosity, heat conduction, and radiation at the lips [19].

Each of the vowels for the male, the female, and the child were generated with eight F0 values, ranging from 100 Hz to 450 Hz in 50-Hz increments. 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. Vowel duration was 0.4 seconds and F0 was maintained constant during the utterance. Sampling frequency was finally set to 10 kHz.

4. Results

Formant estimation accuracy was quantified using the following straightforward error measure [6]:

$$d_{\mathrm{err},i} = 100\% \cdot \frac{\left| F_{\mathrm{est},i} - F_{\mathrm{tru},i} \right|}{F_{\mathrm{tru},i}}, \qquad (5)$$

where $F_{\mathrm{est},i}$ is the estimated ith formant frequency extracted by peak-picking the underlying all-pole spectrum and $F_{\mathrm{tru},i}$ denotes the true ith formant used in the physical modeling. The error measure given in Eq. 5 was defined separately for the lowest three formants (*i.e.* $1 \le \mathrm{i} \le 3$). In addition, the number of spectral peaks found in the all-pole spectra was quantified by a relative number, denoted by n_{pks}, defined as the proportion of the number of analyses showing at least three spectral peaks to the total number of analyses conducted.

Results are shown for [a], [i], and [æ] in Tables 1, 2, and 3, respectively, by pooling together the eight F0 categories. The date corroborate previous findings [6, 7] according to which the formant estimation error is largest for the first formant (F1). In addition, the results indicate that RBLP and DAP yield in general better estimation accuracy than conventional LP. This improvement, however, is not consistent over all vowels and formants analyzed and there are cases where conventional LP was better than RBLP or DAP. LPRA yielded the largest estimation error for F1. The clearest result was observed for the WLP-AME analysis which yielded the smallest estimation error for all vowels and for all error measures. Finally, the relative number of those analyses where an all-pole modeling method is able to indicate all the three lowest formants varied between 50% and 100%. WLP-AME found all the three lowest resonances in each analyzed signal for the vowels [a] and [æ]. For [i], however, the close location of the second (F2) and third (F3) formant made it difficult for WLP-AME to distinguish the two resonances and they were quite often smeared into a single peak.

An example, computed for the male vowel [æ], demonstrating the performance of WLP-AME is shown in Fig. 2. The figure shows spectra computed by conventional LP (thin curves), and WLP-AME (thick curves) together with the true formant frequencies (vertical lines). Differences between the two methods can be seen when the spectra computed from the low-pitch vowels (lower curves), are compared to the spectra obtained from the high-pitched sounds (upper curves). For the three lowest F0 values, the frequencies of the four lowest formants are almost equal in the LP and WLP-AME spectra. When the F0 increases, however, drifting of the F1 estimate from its true value due to the biasing effect of F0 is clearly seen in the LP spectra. The WLP-AME spectra, however, are able to show formants whose positions remain almost unchanged even though F0 varies from 100 Hz to 450 Hz.

Table 1. Formant estimation results (in %) for the [a] vowel. Each row corresponds to an all-pole method evaluated. Relative errors, defined by Eq. 5, are given for F1-F3 in columns 1–3, respectively. The last column includes the relative number of all-pole analyses indicating at least three formant peaks.

	$d_{\mathrm{err},1}$	$d_{\mathrm{err},2}$	$d_{\mathrm{err},3}$	n_{pks}
LP	11.1	5.9	1.8	79
RBLP	10.5	6.4	1.9	75
DAP	10.5	4.9	1.6	75
LPRA	15.4	4.2	2.3	67
WLP-STE	11.7	5.2	1.6	79
WLP-AME	3.0	1.7	0.5	100

Table 2. Formant estimation results (in %) for the [i] vowel. Each row corresponds to an all-pole method evaluated. Relative errors, defined by Eq. 5, are given for F1-F3 in columns 1–3, respectively. The last column includes the relative number of all-pole analyses indicating at least three formant peaks.

	$d_{\mathrm{err},1}$	$d_{\mathrm{err},2}$	$d_{\mathrm{err},3}$	n_{pks}
LP	12.6	2.8	2.4	75
RBLP	16.3	3.2	2.3	67
DAP	12.4	3.0	1.9	67
LPRA	18.6	3.0	2.6	63
WLP-STE	11.4	2.8	2.1	83
WLP-AME	7.6	1.3	1.2	50

Table 3. *Formant estimation results (in %) for the [ae] vowel. Each row corresponds to an all-pole method evaluated. Relative errors, defined by Eq. 5, are given for F1-F3 in columns 1–3, respectively. The last column includes the relative number of all-pole analyses indicating at least three formant peaks.*

	$d_{\mathrm{err},1}$	$d_{\mathrm{err},2}$	$d_{\mathrm{err},3}$	n_{pks}
LP	9.8	4.3	4.1	100
RBLP	9.3	4.3	3.8	100
DAP	9.4	4.4	3.9	96
LPRA	14.4	3.4	4.1	100
WLP-STE	12.2	4.2	2.9	92
WLP-AME	2.9	1.2	0.8	100

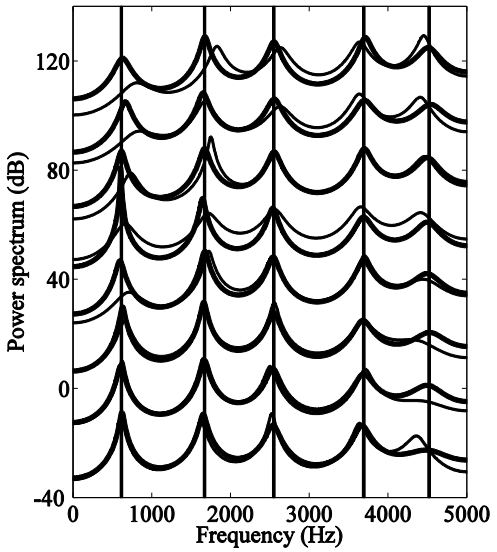

Figure 2: *All-pole spectra computed by conventional LP (thin curve) and WLP-AME (thick curve) from synthetic [æ] vowels of different F0 values. F0 rises from 100 Hz (bottom pair of spectra) to 450 Hz (top pair of spectra) in steps of 50 Hz. True formant values are shown by vertical lines.*

5. Conclusions

Formant frequency estimation based on LP is known to deteriorate 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. This study analyzed five previously known methods and proposed a new technique, Weighted Linear Prediction with Attenuated Main Excitation (WLP-AME).WLP-AME is based on weighted linear prediction by utilizing a time-domain function which downgrades the effects of the main excitation peak of the glottal flow derivative. 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 characteristics of the vocal tract, which leads to less biased formant estimates.

Results with synthetic vowels produced by a physical modeling approach 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. 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.

6. Acknowledgements

This study was supported by the Academy of Finland (project 127345).

7. References

[1] Markel, J. and Gray, A., Jr., *Linear Prediction of Speech*, Springer-Verlag, 1976.

[2] Makhoul, J., "Linear prediction: A tutorial review", Proc. IEEE, 63:561–580, 1975.

[3] El-Jaroudi, A. and Makhoul, J., "Discrete all-pole modeling", IEEE Trans. Signal Process., 39:411–423, 1991.

[4] Lee, C-H., "On robust linear prediction of speech", IEEE Trans. Acoust., Speech, Signal Process., 36:642–650, 1988.

[5] Ma, C., Kamp, Y. and Willems, L., "Robust signal selection for linear prediction analysis of voice speech", Speech Commun., 12:69–81, 1993.

[6] Rahman, S. and Shimamura, T., "Linear prediction using refined autocorrelation function", EURASIP J. on Audio, Speech and Music Process., Article ID 45962:1–9, 2007.

[7] Wang, T. and Quatieri, T., "High-pitch formant estimation by exploiting temporal change of pitch", IEEE Trans. Audio, Speech, Lang. Process., 18:171–186, 2010.

[8] Magi, C., Pohjalainen, J., Bäckström, T. and Alku, P., "Stabilised weighted linear prediction", Speech Commun., 51:401–411, 2009.

[9] Naylor, P., Kounoudes, A., Gudnason, J. and Brookes, M., "Estimation of glottal closure instants in voiced speech using the DYPSA algorithm", IEEE Trans. Speech Audio Process., 15:34–43, 2007.

[10] Fant, G., Liljencrants, J. and Lin, Q., "A four-parameter model of glottal flow", STL-QPSR, Speech, Music and Hearing, Royal Institute of Technology, Stockholm, Sweden, 4:1–13, 1985.

[11] Gobl, C., "A preliminary study of acoustic voice quality correlates", STL-QPSR, Speech, Music and Hearing, Royal Institute of Technology, Stockholm, Sweden, 4:9–22, 1989.

[12] Gold, B. and Rabiner, L., "Analysis of digital and analog formant synthesizers", IEEE Trans. Audio and Electroacoustics, 16:81–94, 1968.

[13] Titze, I., "Parameterization of the glottal area, glottal flow, and vocal fold contact area", J. Acoust. Soc. Am., 75:570–580, 1984.

[14] Titze, I., *The Myoelastic Aerodynamic Theory of Phonation*. Iowa City: National Center for Voice and Speech, 2006.

[15] Titze, I., "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, 2002.

[16] Story, B., "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, 2008.

[17] Story, B., "Synergistic modes of vocal tract articulation for American English vowels", J. Acoust. Soc. Am., 118:3834–3859, 2005.

[18] Story, B., "A technique for "tuning" vocal tract area functions based on acoustic sensitivity functions", J. Acoust. Soc. Am., 119:715–718, 2006.

[19] Story, B., Speech Simulation with an Enhanced Wave-Reflection Model of the Vocal Tract, Ph.D. Thesis, University of Iowa, Iowa City, 1995.