Proceedings of the Institute of Acoustics

PITCH ANALYSIS OF CONCURRENT SPEECH

Dekun Yang, Georg F. Meyer & William A. Ainsworth

Centre for Human and Machine Perception Research
Department of Communication and Neuroscience
Keele University, Keele, Staffordshire ST5 5BG, UK

1 INTRODUCTION

One of the important properties of a speech signal is pitch which is loosely defined as the perception of a fundamental frequency (F0) of a harmonic component pattern of the speech signal. During the two decades a massive amount of research work has been carried out in pitch analysis with the objectives of both understanding human auditory system and building machines for automatic speech processing [1]. Pitch information has been utilized in many speech analysis tasks ranging from normalization, intonation, through segmentation and segregation, to integration (see e.g. [2] and references therein).

Pitch analysis becomes more complex when the speech signal is a mixture of speech from two different speakers. Psychological studies of human speech perception indicate that people can perceive concurrent speech more accurately when listening two competing voices with different pitches [3, 4]. Pitch analysis of concurrent speech is concerned with the determination of the two F0s of the constituent voices. Applications of pitch analysis of concurrent speech include the F0-based segregation in which the segregation is accomplished through the grouping of the elementary acoustic features of concurrent speech signals by exploiting the F0 difference.

Pitch determination is a difficult task, especially when concurrent speech is involved because two harmonic structures are closely spaced and overlapping in the frequency domain. Existing algorithms [5, 6, 7] fall into two categories: joint estimation approach and harmonic suppression approach. The joint estimation approach is based on the estimation of the first two largest peaks in an autocorrelation function or in a harmonic-sieve pattern. Success of the joint estimation approach depends largely on the reliability of the peak finding process while two harmonic structures are interacting each other. In the harmonic suppression approach the pitch of dominant voice is estimated first and then the second pitch is detected from the remainder signal obtained by suppressing the dominant voice. The harmonic suppression approach also faces the problem of locating the peak reliably especially when two voices have similar powers.

In this paper a method for joint estimation of pitches of concurrent speech is developed to overcome the above-mentioned difficulties. The method adopts the sinusoidal model for speech waveform in which the glottal excitation waveform is assumed to be composed of sinusoidal components of arbitrary amplitude, frequencies and phases [8, 9]. A high order linear prediction model is proposed to parameterize these amplitude and frequencies of the sinusoidal components. The combination of the sinusoidal model and linear prediction technique leads to a high-resolution linear prediction spectrum in which the harmonic components of speech signals can be identified via a harmonic sieve. The proposed method is tested using Keele pitch database.

The rest of the paper is organized as follows. Previous work in pitch estimation using linear prediction models is reviewed in Section 2. A method for estimating pitches of concurrent speech is described in Section 3. Performance of the method is evaluated in Section 4, followed by conclusions in Section 5.

2 LINEAR PREDICTION, FORMANT, AND PITCH

Linear prediction techniques [10] have been used for speech modeling since the early day of speech analysis. The speech signal can be modeled as the response of a linear system to an excitation sequence. The source of the excitation sequence is located at the glottis while the linear system comprises the influence of the vocal tract system, i.e. a linear filter is used to model the frequency property of the vocal tract. An autoregressive (AR) model is usually used to define the linear filter, that is, the current speech sample is predicted from a linear combination of a finite number of past sample, i.e.

$$x(n) = \sum_{k=1}^{M} g_k x(n-k) \qquad n = M+1, M+2, \cdots, N \tag{1}$$

in which M is the order of the linear filter, $x(n)$ is the speech sample, N is the number of samples, and $\{g_i\}$ are the prediction filter coefficients. This set of linear equations can be solved exactly when $N = 2M$ or solved approximately in least-square sense when $N > 2M$. After the prediction filter coefficients $\{g_i\}$ are determined, the formant frequencies can be determined from the zeros of the transfer function of the linear system

$$H(z) = 1 - \sum_{k=1}^{M} g_k z^{-k} \tag{2}$$

or from the peaks in the linear prediction spectrum which is defined as

$$S(f) = \frac{1}{|H(z)|^2_{z=e^{j2\pi f}}} \tag{3}$$

Figure 1 plots the Fourier spectrum and linear prediction spectrum of a vowel /er/ in which the order of

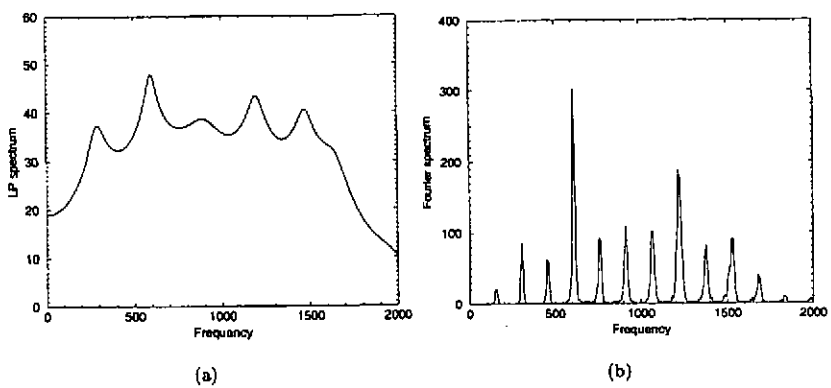


Figure 1. Estimation of formant frequencies from the peaks of linear prediction spectrum. (a) linear prediction spectrum of vowel /er/ obtained using a linear prediction model of 50 order, and (b) its Fourier spectrum for comparison.

the linear prediction model is 50. We can see from Figure 1 that the peaks in linear prediction spectrum correspond to the formant frequencies.

Based on the speech production model pitch can be determined by calculating the autocorrelation function of the excitation signal which can be recovered through inverse filtering. Since the linear prediction model is used to capture the formant characteristics the residual corresponds to the glottal property. Thus, inverse filtering can be used to filter the speech through the inverse of the vocal tract to estimate the excitation signal. After the excitation signal is reconstructed, the pitch is taken to be the maximum peak in the autocorrelation time lag function.

Pitch determination based the linear prediction model is not effective for concurrent speech signals. This is mainly due to the difficulty of finding the two peaks which correspond to the two pitches in the autocorrelation of the linear prediction residual. In addition, the linear prediction equations may not give accurate estimates of the formant frequencies of the mixed speech signals, especially when the formants of the constituent signals are closely spaced in the frequency domain.

3 OVERMODELED LINEAR PREDICTION AND PITCH

Pitch is usually determined by exploiting the periodicity of the speech signal in the time domain or the harmonic structure in the frequency domain. The harmonic structure in the frequency domain is essentially equivalent to the periodicity in the time domain. In what follows we consider a frequency domain based method to estimate the pitches of concurrent speech signals through two steps. The first step is to obtain a spectrum which exhibits the two harmonic structures contained in the mixed speech signals. The second step is to perform a harmonic-sieve operation in the spectrum to find the two pitches.

The motivation of using a linear prediction model comes from the need of a high resolution spectrum which can resolve the closely spaced frequencies in the two harmonics of concurrent speech. Since pitch varies with time pitch analysis must be performed within a short analysis window. In these situations the Fourier spectrum cannot provide the spectrum of sufficient resolution due to to the Gabor-Heisenberg inequality. To alleviate such a resolution problem we use linear prediction spectrum instead of Fourier spectrum for the pitch estimation.

We are concerned with the pitch estimation from the linear prediction spectrum which exhibits the harmonics rather than the formants of speech signals. We first model the concurrent speech signal $x(t)$ as a sum of sinusoidal signals such that

$$x(n) = \sum_k [a_k \cos(j2\pi nkf_1) + b_k \cos(j2\pi nkf_2)] + w(n) \tag{4}$$

in which f_1 and f_2 are the two pitches, a_k and b_k are the amplitudes of the kth harmonic frequencies of the two competing voiced speech, and $w(n)$ is white noise. We then use a high order linear prediction model for modeling the multiple sinusoidal signals. The idea of estimating the frequencies of multiple sinusoidal signals based on a high order prediction model is not new. Indeed, during last decade a substantial amount of work has been devoted to the linear prediction based spectral estimation techniques [11, 12, 13, 14, 15, 16, 17, 18]. In our work the linear prediction technique is applied for attaining a high resolution spectrum of concurrent speech signals. Given N speech signal samples $x(n), n = 1, 2, \ldots, x(N)$, we choose the linear prediction model order to be $L(L < N/2)$. The linear prediction equation can be written in matrix form

$$\begin{bmatrix} x(L) & x(L-1) & \cdots & x(1) \\ x(L+1) & x(L) & \cdots & x(2) \\ \vdots & \vdots & \vdots & \vdots \\ x(N-1) & x(N-2) & \vdots & x(N-L) \\ x(2) & x(3) & \vdots & x(L+1) \\ \vdots & \vdots & \vdots & \vdots \\ x(N-L) & x(N-L+1) & \vdots & x(N) \end{bmatrix} \begin{bmatrix} g_1 \\ g_2 \\ \vdots \\ g_L \end{bmatrix} = \begin{bmatrix} x(L+1) \\ x(L+2) \\ \vdots \\ x(N) \\ x(1) \\ x(2) \\ x(N-L) \end{bmatrix} \tag{5}$$

or

$$\mathbf{Ag} = \mathbf{b} \tag{6}$$

The prediction coefficient vector can be obtained via the singular value decomposition (SVD) of the data matrix $\mathbf{A}$, i.e.

$$\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^* \tag{7}$$

where "$*$" denotes complex conjugate transpose, Σ is a diagonal matrix whose diagonal elements σ_i are known as singular values, the column vectors of $\mathbf{U}$ and $\mathbf{V}$ are the eigenvectors corresponding to the signal subspace and the noise subspace. The prediction coefficient vector can be obtained by

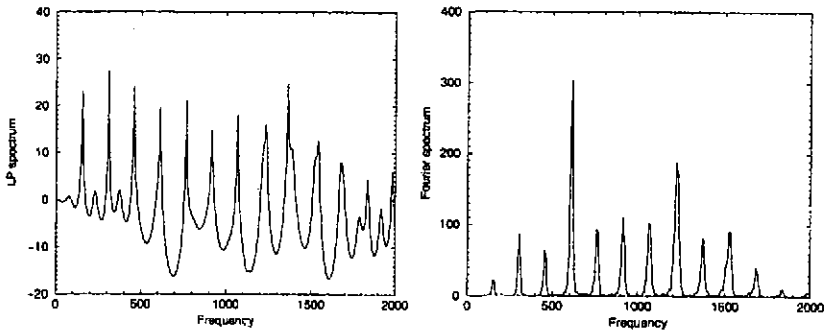
$$\mathbf{g} = \sum_k \frac{(\mathbf{u}_k^* \mathbf{r})}{\sigma_k} \mathbf{u}_k \tag{8}$$

The harmonic frequencies reflect the peaks in the linear prediction spectrum. The main advantage of using SVD to determine the prediction coefficients is that the robustness can be improved by the truncation of the set of singular values [12, 13]. Suppose that there are a total M_0 harmonic frequencies contained in concurrent speech signals. Instead of using all eigenvectors we only use the $M(2M_0 < M < L)$ eigenvectors corresponding to the M largest singular values for the prediction coefficient estimation, i.e. the prediction coefficient vector is given by

$$\mathbf{g} = \sum_{k=1}^{M} \frac{(\mathbf{u}_k^* \mathbf{r})}{\sigma_k} \mathbf{u}_k \tag{9}$$

It is noted that the predominant harmonic frequencies correspond to the M largest singular values. The predominant harmonic frequencies remain in the spectrum because the estimate retains the M largest singular values and associated eigenvectors. Figure 2 (a) shows the linear prediction spectrum of a vowel /er/ obtained using a linear prediction model in which the model order is 250 and the polynomial coefficients are computed using the 50 largest eigenvectors. For comparison, the Fourier spectrum of the vowel is also given in Figure 2 (b). We can see from Figure 2 that the peaks of the linear prediction spectrum are clearly the harmonic frequencies of the vowel.

To illustrate the effectiveness of the high order linear prediction model for estimating the pitches of concurrent speech, we give an example in which the concurrent speech signal contains the voice from a male



(a) (b)

Figure 2. Estimation of harmonic frequencies from the peaks of linear prediction spectrum. (a) the linear prediction spectrum of a vowel /er/ obtained using a linear prediction model in which the model order is 250 and the polynomial coefficients are computed using the 50 largest eigenvectors. (b) its Fourier spectrum for comparison.

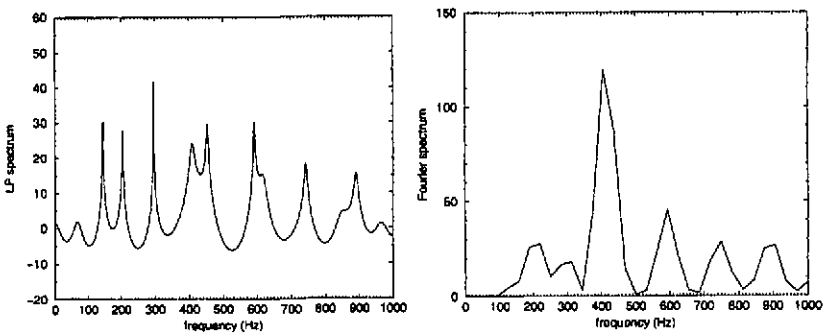
speaker and a female speaker (details of the construction of concurrent speech for our study will be discussed in next section). Figure 3 (a) shows the linear prediction spectrum of the mixed signal with 51.2ms in duration. The order of the linear prediction model is 250, and the prediction coefficients were computed using the using the 50 largest eigenvectors of the sample data matrix. It can been seen in Figure 3 (a) that the two harmonic structures are clearly visible from the peaks in the spectrum. For comparison we plot the Fourier spectrum of the mixed signal in Figure 3 (b), in which it can be seen that the two harmonic structures cannot be resolved by Fourier transform due to the short data record.

4 EXPERIMENTAL RESULTS

In this section we provide some experiments to evaluate the performance of the proposed method using pitch database [19] developed by Keele University. The Keele pitch database contains simultaneous recorded speech and laryngograph data. The propagation delay between the glottis and the microphone was found to be a negligible amount of two samples, and thus we assumed that he recorded speech and laryngograph data are aligned. The laryngograph signal provides a reliable measure of vocal fold contact; a rapid rise in the laryngograph signal corresponds to the rapid closure of the vocal cords and thus the peaks in the differentiated laryngograph signal indicate the instants of glottal closure. The derivative of the laryngograph signal is used as the reference pitch value in our experiments. Voicing decisions can also obtained from the laryngograph signal. Regions of unvoiced speech were detected by an absence of peaks in the differentiated laryngograph signal.

Concurrent speech signals were constructed by mixing two voices obtained from the Keele pitch database. We considered the scenario in which two constituent voices have the same powers in the mixed speech signal. Two reference pitches obtained from the two constituent voices are used as the reference pitches of the mixed speech signal. Figure 4 (a) shows an example of the reference pitches in which one speaker is male while the other is female. The sampling frequency is 10kHz. The speech signal was divided into a

sequence of overlapping frames at a frame rate of 20ms. Each frame is 51.2ms in duration.



(a) (b)

Figure 3. Comparison of the linear prediction spectrum and Fourier spectrum for estimating pitches of concurrent speech. (a) the linear prediction spectrum obtained using a linear prediction model in which the model order is 250 and the polynomial coefficients are computed using the 50 largest eigenvectors. (b) the Fourier spectrum.

In all experiments the parameter configuration was as follows: the linear prediction model order was 200; the polynomial coefficients were computed using the 50 largest eigenvectors. The linear prediction spectrum obtained by the polynomial coefficients was used to determine the pitches of the mixed speech signals. A simple harmonic-sieve mechanism was used to find the two F0s from the peaks in the linear prediction spectrum. The proposed method has been tested using voiced frames. A frame of mixed signals is said to be voiced if at least one constituent speech signal is voiced (recall that the voiced decision was made for each frame of unmixed speech before the mixing procedure). Pitch estimates of the mixed signal are said to be correct if both estimated F0s are within 20% of the reference values. Figure 4 (b) shows some estimated pitches of the concurrent speech signal in which one speaker is male and the other is female. For all mixed speech signals the average rate of correct pitch estimation was found to be 77% for voiced frames. By a close inspection, we found the most errors occur in the following situations:

1. one of the constituent speech signals is in voiced-unvoiced transitions or unvoiced-voiced transitions.
2. one of the constituent speech signals is unvoiced.
3. two pitches of the constituent speech signals are harmonically related.

In the first case the error is due to the shortage of voiced signal samples in transitional frames. Note that the voiced decision is made based on the averaging effect of the differentiated laryngograph signal over the frame. The number of voiced signal samples may be smaller than the frame length in transitional frames. However, if the number of voiced signal samples is less than twice the model order the high-order linear prediction equations cannot be solved. One way to alleviate such a problem is to use a lower order of the linear prediction model at the cost of decreasing the resolution of the linear prediction spectrum. The

problem of how to choose the optimal linear prediction model order for transitional frames is currently being investigated.

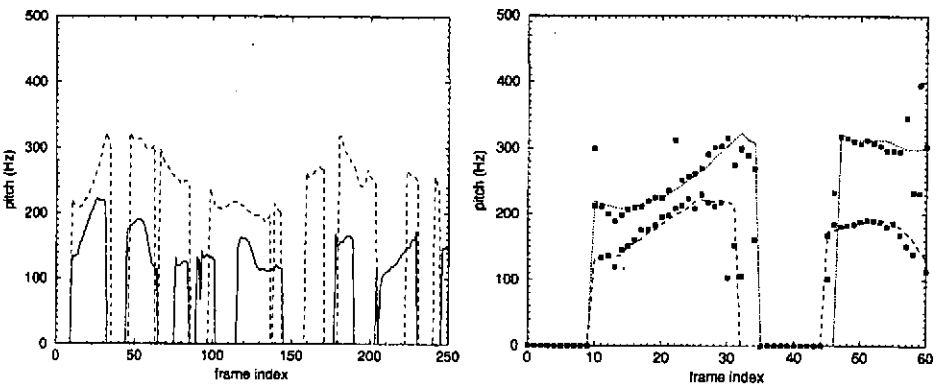


Figure 4. (a) Reference pitch contours of a mixed speech, in which solid contour is male pitch and dashed contour is female pitch. (b) the estimated pitches shown in black squares.

In the second case where only one harmonic is present the false pitch estimates are the source of errors. This is because the algorithm is designed to find two pitches by assuming the existence of two harmonic structures in the mixed signal, and one of the pitch estimates must be wrong when only one voice is present. The solution to this problem may be detecting the number of speakers before the pitch analysis or removing the false pitch estimates by some heuristic strategies.

In the third case it is impossible to find the two pitches correctly using the frame based approach because the two harmonics coincide with each other. The only way to overcome this difficulty is to employ a multi-frame strategy to incorporate the pitch estimates over successive frames. Pitch is a time-varying feature contained in speech signals and it is rare that two pitches of concurrent speech are harmonically related over a long period (e.g. see Figure 4). Based on the assumption that pitches vary slowly within an observation interval, pitch contours can be constructed via multi-frame integration. Our on-going work is to develop a multi-frame based method in which candidate peak estimates obtained from each frame are combined to determine meaningful pitch contours.

5 CONCLUSIONS

A method for estimating the pitches of concurrent speech signals has been proposed in this paper. The method is based the modeling of speech signals as a sum of sinusoidal waves with frequencies of multiple

times of fundamental frequency. To account for close-spaced overlapping harmonic structures contained in concurrent speech signals, a set of linear prediction equations are used to construct a high resolution linear prediction spectrum. The pitches of concurrent speech are determined from the peaks in the linear prediction spectrum. Experimental results were provided to evaluate the performance of the method. The sources causing estimation errors were pointed out and how to improve the estimation performance was discussed.

ACKNOWLEDGMENT

This work is supported by EPSRC Grant GR/L05655 and EPSRC Grant GR/K77754.

6 REFERENCES

[1] D. J. Hermes. Pitch analysis. In M. Cooke, S. Beet, and M. Crawford, editors, *Visual Representations of Speech Signals*, pages 3–25. John Wiley and Sons Ltd, 1993.

[2] W. A. Ainsworth. Pitch and the perception of speech sounds. In *Proceeding of the Institute of Acoustics*, 1998.

[3] P. F. Assmann and Q. Summerfield. Modeling the perception of concurrent vowels - vowels with different fundamental frequencies. *Journal of the Acoustical Society of America*, 88(2):680–697, 1990.

[4] P. F. Assmann and D. D. Paschall. Pitches of concurrent vowels. *Journal of the Acoustical Society of America*, 103(2):1150–1160, 1998.

[5] T. F. Quatieri and R. G. Danisewicz. An approach to co-channel talker interference suppression using a sinusoidal model for speech. *IEEE Transactions on Acoustics Speech and Signal Processing*, 38(1):56–69, 1990.

[6] D. P. Morgan, E. B. George, L. T. Lee, and S. M. Kay. Cochannel speaker separation by harmonic enhancement and suppression. *IEEE Transactions on Speech and Audio Processing*, 5(5):407–424, 1997.

[7] A deCheveigne. Modeling the perception of multiple pitches. In *Proc. IJCAI workshop on Computational Auditory Scene Analysis*, 1997.

[8] R. J. McAulay and T. F. Quatieri. Speech analysis/synthesis based on sinusoidal representation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34(4):744–754, 1986.

[9] T. F. Quatieri and R. J. McAulay. Speech transformations based on a sinusoidal representation. *IEEE Transactions on Acoustics Speech and Signal Processing*, 34(6):1449–1464, 1986.

[10] J. Makhoul. Linear prediction: A tutorial review. *Proceedings of the IEEE*, 63(4):561–580, 1975.

[11] S. M. Kay and S. T. Marple Jr. Spectrum analysis - A modern perspective. *Proceedings of the IEEE*, 69:1390–1418, 1981.

[12] R. Kumaresan and D. W. Tufts. Estimating the parameters of exponentially damped sinusoids and pole-zero modeling in noise. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 30(6):833–840, 1982.

[13] D. W. Tufts and R. Kumaresan. Estimation of frequencies of multiple sinusoids: Making linear prediction perform like maximum likelihood. *Proceedings of the IEEE*, 70(9):975–989, 1982.

[14] S. M. Kay and A. K. Shaw. Frequency estimation by principal component AR spectral estimation method without eigendecomposition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 36(1):95–101, 1988.

[15] S. Haykin. *Array Signal Processing*. Prentice Hall, Englewood Cliffs, NJ, 1985.

[16] S. M. Kay. *Modern Spectral Estimation: Theory and Application*. Prentice Hall, Englewood Cliffs, NJ, 1988.

[17] M. A. Rahman and K. Yu. Total least squares approach for frequency estimation using linear prediction. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 35(10):1440–1454, 1987.

[18] J. A. Cadzow. Spectral estimation: An overdetermined rational model equation approach. *Proceedings of the IEEE*, 70:907–939, 1982.

[19] F. Plante, G. Meyer, and W. A. Ainsworth. A pitch extraction reference database. In *EUROSPEECH'95*, volume 1, pages 837–840, 1995.