Harmonic Filtering for Joint Estimation of Pitch and Voiced Source with Single-microphone Input

S. W. Lee[1], Frank K. Soong[1,2], and P. C. Ching[1]

[1]Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong SAR, China
[2]Microsoft Research Asia, Beijing, China
{yswlee, pcching}@ee.cuhk.edu.hk, frankkps@microsoft.com

Abstract

Standard correlation based methods are not effective in estimating pitch tracks of multiple speech sources from a single-microphone input In this paper, an adaptive harmonic filtering is proposed to jointly estimate the source signals and their corresponding fundamental frequencies. By exploiting the harmonic structure of voiced speech, pitch information of one source is extracted from the pitch prediction filter and the output residual becomes the estimate of the other source. The procedure is iterated successively with a summation constraint. From the evolution of pitch prediction filter, it is shown that the iterative harmonic filtering with the summation constraint is effective to separate multiple pitch tracks into individual ones.[1]

1. Introduction

Pitch estimation has been one of the popular research topics over the past decades and found a wide range of applications, e.g., linear predictive coding (LPC), adaptive comb filtering, speech recognition with tone information and speech source segregation [1 - 3]. Pitch refers to the auditory perception of tone [4]. In human speech production, if sufficiently fast airflow passes and the vocal folds are tensed appropriately, the vocal folds oscillate. This makes the airflow excitation periodic and generates voiced sounds. When the vocal folds are relaxed, the airflow is either (1) a wideband noise-like excitation and produces unvoiced sounds or (2) a plosive excitation generated from an abruptly released pressure [5, 6]. Voiced sounds are periodic and exhibit regular patterns in both waveform and frequency spectrum. The more rapid the oscillation of vocal folds, the higher the pitch. As pitch is related to perception and is not directly measurable from waveform, fundamental frequency (F0) is often estimated instead. F0 is defined as the inverse of the period of vocal fold oscillation.

Correlation-based methods, especially the autocorrelation function (ACF), have been conventionally used to estimate F0 values of a single-source speech signal. ACF searches the lag that provides the greatest similarity between the window and the delayed version. It is known to be relatively robust to noise. However, to cover the necessary F0 range, the window that ACF must be computed is relatively large. This makes the pitch tracking of fast F0 changing speech difficult. The Cross-correlation function (CCF) has been suggested to overcome the above disadvantage, since samples from adjacent windows are involved in the computation.

By looking at secondary cues, such as the second-largest peak in ACF, some works have been reported to use single-period estimation for multiple F0 estimation [7]. Nevertheless, for multi-source signals (with multiple F0s), extending these single-period estimation methods alone is not adequate [8, 9]. Both the first and the second-largest peaks are easily affected by the multiple periodicities.

In this paper, a successive harmonic filtering system is proposed to accurately estimate the corresponding F0s of two individual voiced sources. Individual source signal is alternately removed from the mixed signal and pitch estimation is performed on the residual, rather than using secondary cues from the multi-source signal. At every iteration, one F0 is estimated, while another source signal is estimated as the filter output. Experiments on synthetic speech were carried out. In particular, the evolution of the successive harmonic filtering has been extensively investigated.

2. Pitch predictor

In the proposed successive harmonic filtering system, pitch prediction error filter, which has been widely used for LPC-based speech coding [1, 10], is incorporated. Before moving to the mathematical details of pitch prediction error filter, the speech model used is reviewed first.

2.1. Speech model for voiced sounds

The input multi-source signal $x(n)$ is related to the source signals $x_1(n)$ and $x_2(n)$ by the following equation,

$$x(n) = x_1(n) + x_2(n) \tag{1}$$

Both $x_1(n)$ and $x_2(n)$ are voiced sounds with distinct periodicities. The proposed system is based on the source-filter model. Let $u_i(n)$ and $h_i(n)$ represent the periodic excitation and the impulse response of the vocal tract of source signal $x_i(n)$. Fig. 1 depicts the source-filter model for multi-source signals. It is assumed that the fundamental frequencies of the two sources are co-prime. Mathematically,

$$HCF\left(F0^1, F0^2\right) = 1 \tag{2}$$

where HCF is the highest common factor and $F0^i$ denotes the fundamental frequency of source i.

The proposed filtering system is not only aimed at extracting $F0^1$ and $F0^2$, but also estimating the sources $x_1(n)$ and $x_2(n)$. For voiced speech sources, energy is dominant at multiples of F0. After mixing with other sources, through a linear process, the speech energy at harmonic frequencies remains there due to the eigenfunction property of sinusoids. This allows the pitch extraction from the mixed input $x(n)$,

[1] This work is partially supported by a grant awarded by the Hong Kong Research Grants Council. Thanks also go to the ATR SLT Labs, where S. W. Lee and Frank K. Soong conducted this work last year.

without removing the envelope of the vocal tract. By removing the speech harmonics of $x_i(n)$, this helps to separate $x_i(n)$ from $x_j(n)$, where $i \neq j$ (For a purely periodic single-source signal, if the harmonic structure is removed, the output response should be zero). Fig. 2 illustrates this simple idea to estimate sources using a single-microphone input with the help of pitch knowledge.

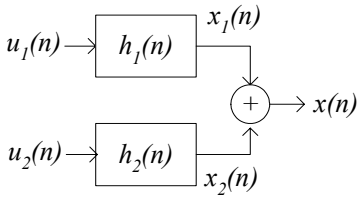

Figure 1: Source-filter model for a two-source signal $x(n)$.

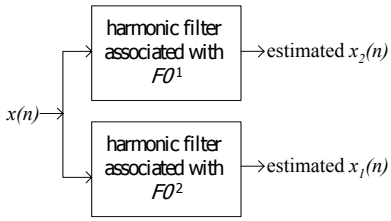

Figure 2: The harmonic structure of $x_i(n)$ is removed by an adaptive filter and the filter output is the estimated signal of the other source.

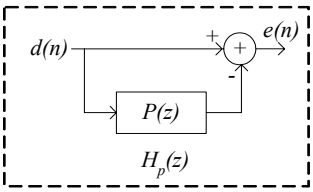

Figure 3: Block diagram of a pitch prediction error filter.

2.2. Pitch prediction error filter

To construct a simple pitch prediction error filter, a single tap delay filter can be used. It is used to reduce distant sample correlation. The block diagram is shown in Fig. 3. Let $H_p(z)$ and $P(z)$ be the pitch predictor error filter and the prediction filter, respectively. The output is the difference between the current input and the prediction filter output. The time-domain filter of lag M is defined by the impulse response,

$$h_p(n) = \delta(n) - p(n) = \delta(n) - \beta\delta(n - M) \tag{3}$$

This filter has zeros at the frequency $1/MT$ and all its multiples, where T is the sampling period. For single-source signals, the lag M can be found as follows. A normalized correlation based function $\tau(m)$ is first computed [10],

$$\tau(m) = \frac{\phi(0,m)}{\sqrt{\phi(0,0)\phi(m,m)}} \tag{4}$$

where

$$\phi(i,j) = \sum_{n=0}^{N-1} d(n-i)d(n-j) \tag{5}$$

and $d(n)$ is the input signal to the pitch predictor and N is the frame length. The lag where maximum $\tau(m)$ occurs is taken as M. $\tau(m)$ is a normalized CCF, which facilitates the pitch extraction, as $\tau(m)$ tends to be close to 1 for lags corresponding to multiples of pitch period.

In the proposed harmonic filtering system, a three-tap pitch prediction error filter is used, rather than single tap. Multi-tap predictor can deal with non-integral samples of pitch periods, by interpolating the fractional sample delay with multi-taps. It also yields a higher prediction gain than a single-tap filter [10]. The resulting $h_p(n)$ is,

$$\begin{aligned} h_p(n) = \delta(n) &- \beta_1\delta(n-M) \\ &- \beta_2\delta(n-(M+1)) - \beta_3\delta(n-(M+2)) \end{aligned} \tag{6}$$

Recall that for single-source speech signals, correlation based pitch extraction can be used to find M. However, for multi-source speech signals, prediction gains for different M is searched and M is set to the lag that provides the maximum value. Prediction gain represents the extent to which predictors remove redundancies by measuring the output energy. It alleviates the problem of using correlation based extraction that the maximum value may happen at multiples of the Least Common Multiple of the source periods. The prediction gain G, is measured by,

$$G = \frac{\sum_n d^2(n)}{\sum_n e^2(n)} \tag{7}$$

where $e(n)$ is the filter output. A set of system equation that uses present values to predict future values distant by M samples is formulated and is given by,

$$\begin{bmatrix} d(1) & d(0) & d(-1) \\ d(2) & d(1) & d(0) \\ \vdots & \vdots & \vdots \\ d(N+1) & d(N) & d(N-1) \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} \cong \begin{bmatrix} d(M) \\ d(M+1) \\ \vdots \\ d(M+N) \end{bmatrix} \tag{8}$$

or in a matrix form,

$$\mathbf{D\beta} \cong \mathbf{d} \tag{9}$$

The least squares solution is,

$$\boldsymbol{\beta} = (\mathbf{D}^{\mathbf{T}}\mathbf{D})^{-1}\mathbf{D}^{\mathbf{T}}\mathbf{d} \tag{10}$$

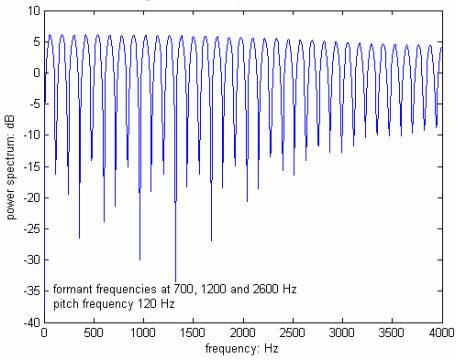

Figure 4: Power spectrum of a pitch prediction error filter.

An example of the pitch prediction error filter is plotted in Fig. 4. It is derived from a synthetic single-source signal /AA/, with F0 equals 120 Hz and formant frequencies at 700, 1220 and 2600 Hz.

3. Successive harmonic filtering system

In this section, the proposed successive harmonic filtering system will be introduced. It is a recursive iterative algorithm for finding an optimal pitch prediction error filter given the input signal *d(n)*. Fig. 5 depicts the architecture of the proposed system (where *x(n)* is a two-source input in this case).

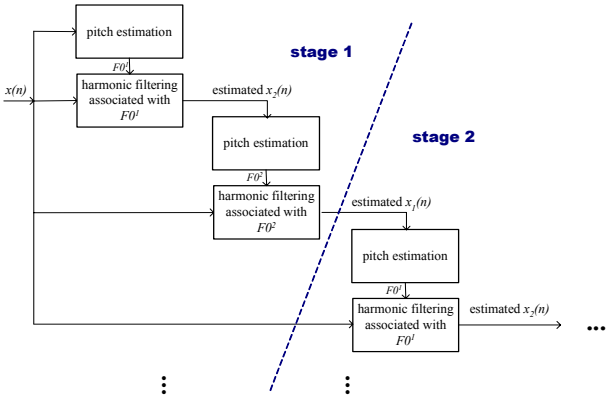

Figure 5: The architecture of the proposed successive harmonic filtering system.

At every iteration, the pitch prediction error filter is used to suppress the harmonic energy of one periodic sound in the mixed input *x(n)*. The filter characteristics are derived from *x(n)* (at the first iteration) or from the newly found error residual from the previous iteration. By filtering the mixed input *x(n)* with this pitch prediction error filter, the other source signal is estimated as the output. Assuming that after passing through the first iteration, $F0^1$ is estimated along with the estimated $x_2(n)$. This error residual $x_2(n)$ enters the filtering system again to jointly estimate $F0^2$ and $x_1(n)$ at the second iteration. The whole process is repeated successively to refine the signal estimates and corresponding pitch values. It is repeated until the prediction gain for each source converges.

A summation constraint,

$$x(n) = x_1(n) + x_2(n) \tag{11}$$

is implicitly adopted during each harmonic filtering, since the filter output is treated as the other source.

This summation constraint is beneficial to pitch extraction and separation of multi-source signals. Pitch extraction is inaccurate during the first iteration, if the input *d(n)* contains two voiced speech signals and the distant sample correlations modeled in Equation (6) and (8) are incorrect. In most cases, only part of the harmonic energy of $x_1(n)$ is removed and the error residual actually contains $x_2(n)$ filtered by the prediction error filter and a fractional trace of $x_1(n)$. At the next iteration, as $x_2(n)$ is much stronger than the trace of $x_1(n)$, the correlations associated with $F0^2$ become more distinctive and better modeled by the two equations. Hence, most of $x_2(n)$ energy is removed. From now on, the input *d(n)* to the pitch prediction error filter becomes more and more single-source like. This self-control of the summation constraint leads to a *F0* estimation with improving accuracy, as the iteration goes. In the meantime, the prediction gain for each source is monotonically increasing and finally reaches an asymptote.

4. Experimental results and discussions

Stationary synthetic sounds are used to study the performance of the single-input harmonic system. Table 1 lists the details of speech samples for the experiments reported below. The evolution of the harmonic filter is investigated to check if the summation constraint and recursive estimation are effective or not.

synthetic /AA/	formant frequencies: 700, 1220 and 2600 Hz
	F0: 130 Hz
	power: 60 dB
	sampling frequency: 8 kHz
synthetic /EY/	formant frequencies: 480, 1720 and 2520 Hz
	F0: 77 Hz
	power: 60 dB
	sampling frequency: 8 kHz

Table 1: Details of the speech samples

Summing these two sources to form *x(n)*, the proposed filtering system went through 10 iterations and stopped. Fig. 6 plots the input and output waveforms. The estimated pitch lags *M* at iteration 1 to iteration 10 are 60, 103, 61, 104, 62, 104, 62, 104, 62 and 104 respectively. This demonstrates the self-control property. This pitch lag *M* estimate can be refined by doing a parabolic interpolation to the three tap values in the pitch prediction error filter. With such an interpolation, the estimated non-integral pitch periods are 61.5 and 103.9 (in lag). They correspond to *F0* of 130.08 Hz and 77 Hz respectively.

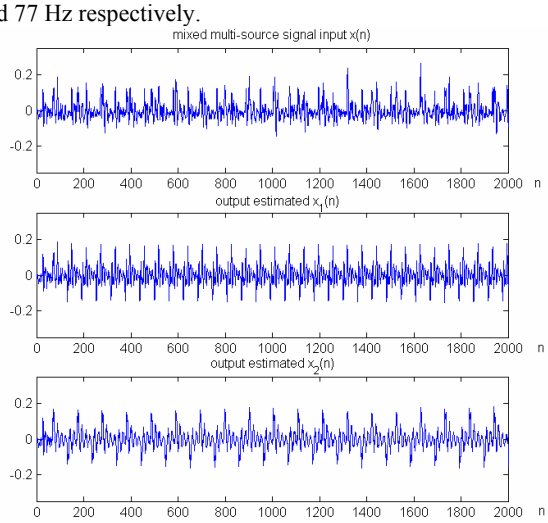

Figure 6: Input and output waveforms.

When the iteration proceeds, the periodicity found in the error residual *e(n)* becomes more single-source like. The notches in the pitch prediction error filter $H_p(z)$ become narrower and sharper. Fig. 7 plots the power spectra of the filters at several iterations.

The prediction gain for a designated source is plotted against stage number in Fig. 8. Stage number is the actual iteration number associated for a designated source. For example, iteration 1, 3 and 5 are stage 1, 2 and 3 for $x_1(n)$ and iteration 2, 4 and 6 are stage 1, 2 and 3 for $x_2(n)$. The final prediction gains are 19.71 and 19.41 dB for $x_1(n)$ and $x_2(n)$ respectively.

Referring to Equation (6), the transfer function of the pitch prediction error filter consists of zeros only. Fig. 9

shows the pole-zero plots at different iterations. When the number of iterations increases, the zeros are pushed closer to the unit circle, making the attenuation at harmonic frequencies as large as possible. The evolution of zeros toward the unit circle is even more distinctive at lower speech harmonics. This phenomenon is also observed in the power spectra.

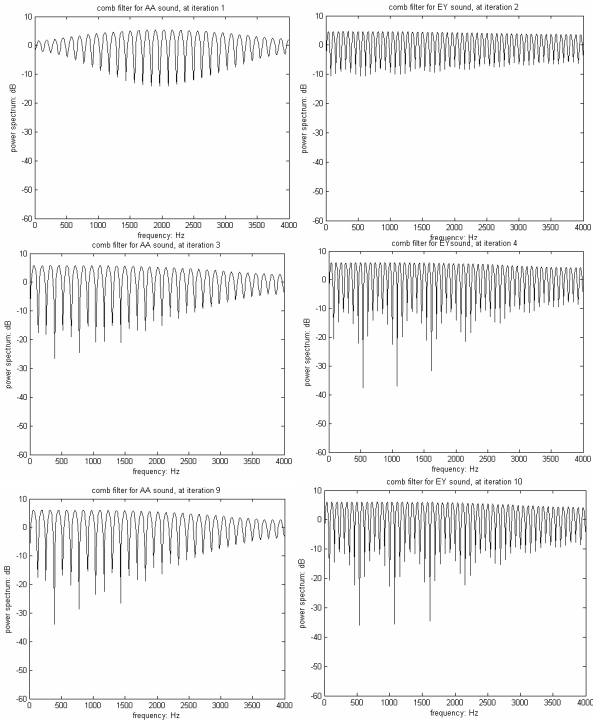

Figure 7: Several snapshots of the power spectra of pitch prediction error filters (comb filters) in iteration. Plots on left hand side associate with one source; plots on right hand side associate with another.

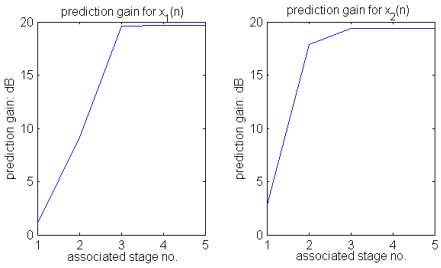

Figure 8: Prediction gain versus stage number for a designated source.

The proposed system is effective for source separation and gives high prediction gains; however, speech from different sources may congregate together at frame boundaries. This problem is related to the tracking of pitch contour in the system design. It is believed that pitch contour information is necessary and the source estimate should be constituted by those output frames from the same source. This issue is highly important for natural speech inputs, as shorter frame length is needed for dealing with the non-stationary nature of speech signals.

Only voiced speech is considered in the current system design. Voiced speech is not the one and only speech type; unvoiced speech and short pauses also present in human speech. For practical deployment, these mixed sourced signals need to be tested.

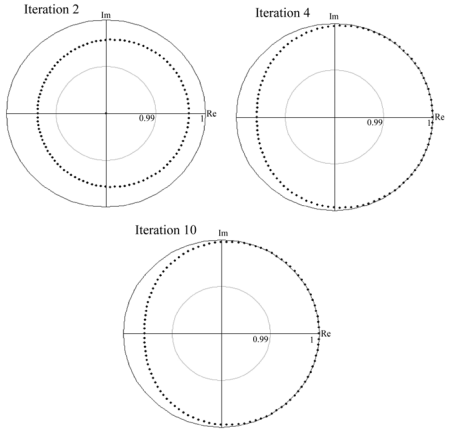

Figure 9: The zeros of the pitch prediction error filter.

5. Conclusions

A joint estimation algorithm for estimating pitch tracks and the corresponding voiced signals in a single channel is proposed. It is a recursive filtering process which exploits the harmonic structure in voiced speech. From the experimental results, we show that the proposed algorithm is effective in estimating F0's and corresponding voiced source signals in a single-microphone input. As voiced sources are successively separated from one iteration to the next, the output residual helps to refine the F0 and source signal estimates. The summation constraint is useful to filter out one source and to improve pitch estimation and it is confirmed by checking the evolution of the pitch filter in successive iterations.

6. References

[1] J. R. Deller, J. H. L. Hansen, and J. G. Proakis, *Discrete-Time Processing of Speech Signals*. New York: IEEE Press, 2000.

[2] N. M. Veilleux and M. Ostendorf, "Probabilistic parse scoring with prosodic information," in *Proc. ICASSP*, 1993, pp. 51-54.

[3] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*. London: The MIT Press, 1990.

[4] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. New Jersey: Prentice Hall, 2001.

[5] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Prentice Hall, 1993.

[6] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. New Jersey: Prentice Hall, 1978.

[7] M. Weintraub, "A theory and computational model of auditory monaural sound separation," Ph.D. dissertation, Stanford University, 1985.

[8] A. de Cheveigné and H. Kawahara, "Multiple period estimation and pitch perception model," *Speech Communication*, vol. 27, pp. 175-185, Apr. 1999.

[9] T. W. Parsons, "Separation of speech from interfering speech by means of harmonic selection," *J. Acoust. Soc. Amer.*, vol. 60, pp. 911-918, Oct. 1976.

[10] R. P. Ramachandran and P. Kabal, "Pitch prediction filters in speech coding," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, pp. 467-478, Apr. 1989.