ＴＲ－ＳＬＴ－００８１

Successive Adaptive Harmonic Filtering for Separating Voice Sources Using Single-microphone Input

LEE, Siu Wa　&　SOONG, Frank K.

2004 年 8 月 31 日

概要

Blind source separation has been one of the most popular speech research areas recently. In this work we investigate the feasibility of separating multi-source signals in a single-microphone input scenario. By exploiting the harmonic structure of voiced speech signals, a successive adaptive harmonic filtering system is proposed and successfully tested to separate two mixed voice sources blindly.

（株）国際電気通信基礎技術研究所
音声言語コミュニケーション研究所
〒619－0288「けいはんな学研都市」光台二丁目 2 番地 2 TEL：0774－95－1301

Advanced Telecommunications Research Institute International
Spoken Language Translation Research Laboratories
2-2-2 Hikaridai “Keihanna Science City” 619-0288,Japan
Telephone:+81-774-95-1301
Fax　　:+81-774-95-1308

INTRODUCTION

Signal separation has been one of most popular speech processing research areas recently. Multi-input signals from a microphone array are usually used in most separation systems. However, it may not always be possible to install multi-sensors in every application and it is challenging to solve the blind source separation problem using only single channel input. In this work we investigate the feasibility of multi-source, voiced signal separation in a single-microphone input scenario. By exploiting the harmonic structure of voiced speech, a successive filtering system is designed. The input mixed signal $x(t)$ is related to the source signal $x_1(t)$ and $x_2(t)$ by the following equation,

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

which is called instantaneous mixing. This is shown in Figure 1.

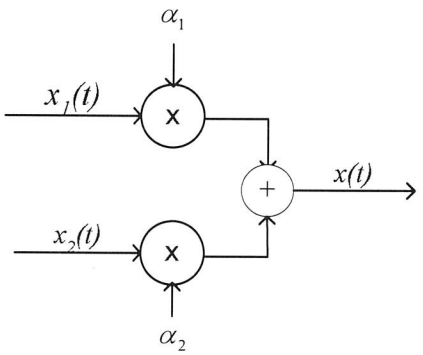

Figure 1 Instantaneous mixing of the two source signal $x_1(t)$ and $x_2(t)$ scaled by α_1 and α_2, respectively, giving the output signal $x(t)$.

A number of assumptions have been made in our problem formulation. They are:

1. Both $x_1(t)$ and $x_2(t)$ are voiced speech.
2. They obey the source-filter model. Figure 2 illustrates the source-filter model, where $u_i(t)$ represents the periodic excitation and $h_i(t)$ is the impulse response of the vocal tract.

3. The fundamental frequencies of the two sources are co-prime to each other, or in other words, the highest common factor (HCF) of f_0^1 and f_0^2 is 1, where f_0^i is the fundamental frequency of source i.

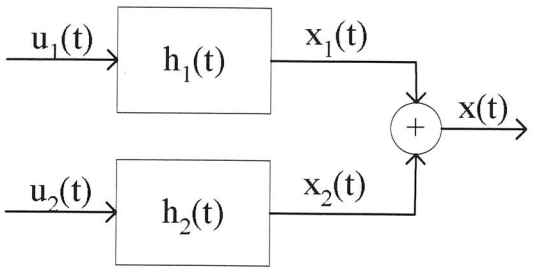

Figure 2 In source-filter model, each source signal is modeled by a vocal tract filter, $h_i(t)$, which is excited by a corresponding excitation source signal, $u_i(t)$.

Spectral Envelope vs Excitation Source for Signal Separation

In the source-filter model, the filter, $h_i(t)$, convolves with the excitation source, $u_i(t)$, and generates the voiced speech signal, $x_i(t)$. This subsection discusses the feasibility of using spectral envelope or excitation source for single-input based separation problem.

The spectral envelope has been commonly modeled by an all-pole filter. By using linear prediction coding (LPC), it can be extracted easily. Equation (2) shows the relationship between the input and output signals.

$$x(t) = u_1(t) * h_1(t) + u_2(t) * h_2(t) \tag{2}$$

In frequency domain,

$$X(\omega) = U_1(\omega)H_1(\omega) + U_2(\omega)H_2(\omega) \tag{3}$$

Or in the Z-transform domain, we obtain

$$H_1(z) = \frac{X_1(z)}{U_1(z)} = \frac{1}{1 - \sum_k a_{k1} z^{-k}} = \frac{1}{A_1(z)} \tag{4}$$

$$H_2(z) = \frac{X_2(z)}{U_2(z)} = \frac{1}{1 - \sum_k a_{k2} z^{-k}} = \frac{1}{A_2(z)} \tag{5}$$

where a_{ki} is the LPC coefficients of $H_i(z)$. Let Q be the number of poles in $A_i(z)$ and p_{ij} be the poles of $H_i(z)$. Equation (4) and (5) becomes,

$$H_1(z) = \frac{1}{1 - \sum_k a_{k1} z^{-k}} = \frac{1}{A_1(z)} = \frac{1}{(z - p_{11})(z - p_{12})\mathrm{K}\ (z - p_{1Q})} \tag{6}$$

$$H_2(z) = \frac{1}{1 - \sum_k a_{k2} z^{-k}} = \frac{1}{A_2(z)} = \frac{1}{(z - p_{21})(z - p_{22})\mathrm{K}\ (z - p_{2Q})} \tag{7}$$

Substituting Equation (6) and (7) in Equation (3), we obtain,

$$X(z) = \frac{U_1(z)}{A_1(z)} + \frac{U_2(z)}{A_2(z)} \tag{8}$$

$$\begin{aligned} X(z) &= \frac{U_1(z)A_2(z) + U_2(z)A_1(z)}{A_1(z)A_2(z)} \\ &= \frac{U_1(z)(z - p_{21})(z - p_{22})\mathrm{K}\ (z - p_{2Q}) + U_2(z)(z - p_{11})(z - p_{12})\mathrm{K}\ (z - p_{1Q})}{(z - p_{11})(z - p_{12})\mathrm{K}\ (z - p_{1Q})(z - p_{21})(z - p_{22})\mathrm{K}\ (z - p_{2Q})} \end{aligned} \tag{9}$$

In Equation (9), it is found that the poles located in the denominator of source signals appear in the nominator with summation of the two excitations.

To extract the spectral envelopes of different sources, the poles should be identified. However, locating the poles and zeros in Equation (9) is difficult and non-linear processing is often involved.

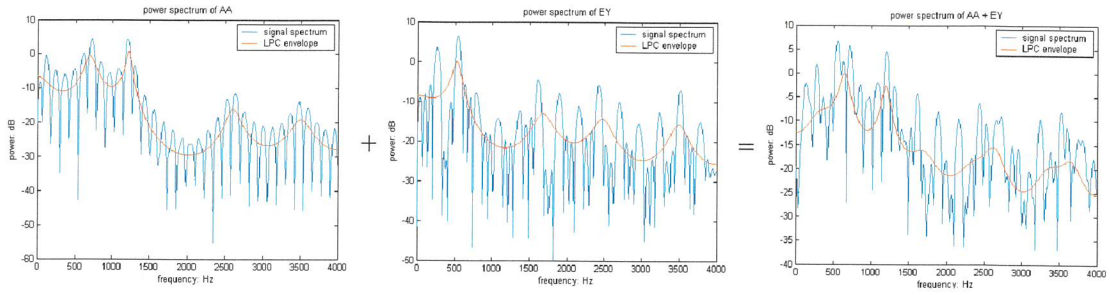

Figure 3 Spectral envelopes of the two source signals and the mixed signal. The envelope is obtained by finding the LPC coefficients of the signal spectrum.

Figure 3 depicts the spectral envelopes extracted from two source signals and the corresponding mixed signal. In order to extract the spectral envelopes of each source signal from the input mixed spectrum, it is necessary to locate the poles for each source. However, comparing the rightmost figure to the left and the centre ones, only some of the poles remain in the spectral envelope, while other poles disappear as a result of the zeros in Equation (9).

Considering the excitation source, as sinusoids are eigenfunctions, the speech harmonics in $x_i(t)$ remains unchanged after a linear operation (i.e., instantaneous or convolutional mixing). At multiples of f_0^i, the speech energy is dominant. By removing the speech harmonics of $x_i(t)$, this can help separating $x_j(t)$ from $x_i(t)$. This is the basic idea of the proposed single-input based separation system.

Figure 4 depicts the conceptual model of how to exploit the harmonic structure of voiced speech for source separation.

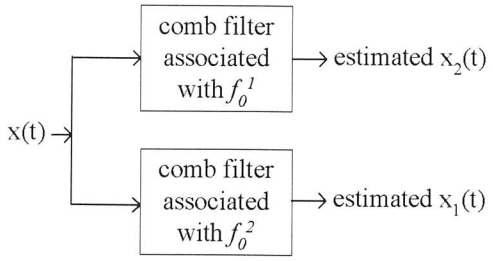

Figure 4 The harmonic structure of $x_i(t)$ is removed by an adaptive comb filter and the filter output is the estimated signal of the other source.

PITCH PREDICTION FILTER

A three-tap, pitch prediction filter is used as the comb filter to remove distant sample correlation by adaptive comb filtering. The block diagram of the pitch prediction filter is shown in Figure 5. The output is the difference between the speech input and the prediction output of filter $P(z)$. The filter $P(z)$ is called the pitch prediction filter; while the overall filter $H_p(z)$ is called the pitch prediction error filter.

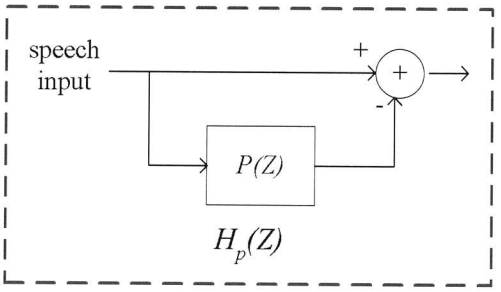

Figure 5 Pitch prediction error filter $H_p(z)$, by adapting the filter coefficients of $P(z)$ to minimize the squares of the output errors.

$$H_p(z) = 1 - P(z) \tag{10}$$

$$P(z) = \beta_1 z^{-M} + \beta_2 z^{-(M+1)} + \beta_3 z^{-(M+2)} \tag{11}$$

The transfer function of the prediction comb filter is given in Equation (10). Multiple (3 in this case) taps around the delay of one pitch period are used to give,

- Higher prediction gain
- To cope with non-integer samples of pitch period

The filter characteristics are governed by the parameter M, β_1, β_2 and β_3. The pitch lag M is first estimated by finding the lag with the maximum correlation coefficient or prediction gain. Then, a set of system equation which uses present values to predict future values distant by M samples is formulated as shown in Equation (12),

$$\begin{bmatrix} x(1) & x(0) & x(-1) \\ x(2) & x(1) & x(0) \\ \mathrm{M} & \mathrm{M} & \mathrm{M} \\ x(N+1) & x(N) & x(N-1) \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} \cong \begin{bmatrix} x(M) \\ x(M+1) \\ \mathrm{M} \\ x(M+N) \end{bmatrix} \qquad (12)$$

$$X\underset{\sim}{\beta} \cong \underset{\sim}{x}$$

By using Least-Squares solution of β, the coefficients β_1, β_2 and β_3 can be found by,

$$\underset{\sim}{\beta} = (X^T X)^{-1} X^T \underset{\sim}{x} \qquad (13)$$

Prediction Gain

Prediction gain is used to assess the extent to which the predictors remove redundancies by measuring the resulting residual energy. It is defined as,

$$prediction\ gain = \frac{average\ input\ energy}{average\ prediction\ residual} \qquad (14)$$

For a single-source signal (only one fundamental frequency exists), maximum prediction gain occurs at correct pitch lag and the spectrum of the resultant prediction filter has valleys at speech harmonic frequencies. Figure 6 shows an example of a pitch prediction filter for a single-source /AA/ signal with fundamental frequency 120 Hz.

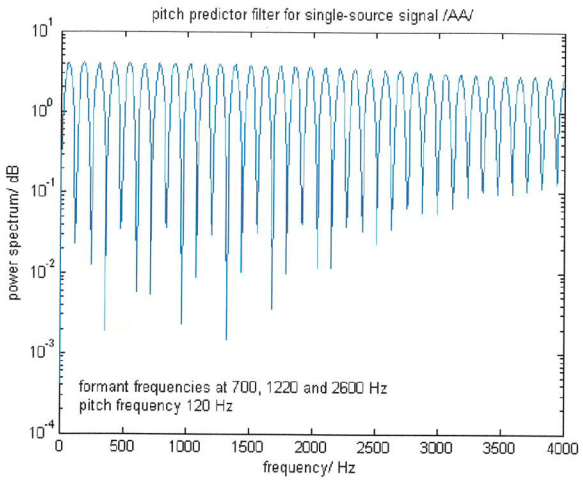

Figure 6 Power spectrum of the pitch prediction error filter for a single-source signal /AA/, with fundamental frequency 120 Hz and formant frequencies at 700, 1220 and 2600 Hz.

SEPARATION BY COMB FILTERING

Two separation systems, namely system 1 and system 2 were designed with successive use of adaptive comb filtering.

Successive Source Separation System 1

Comb filtering is successively applied until the prediction gain reaches an asymptote, as illustrated in Figure 7. After knowing the filter coefficients in all iterations, comb filters which come from the same channel are grouped together. Finally, the input mixed signal $x(t)$ is convolved with each group to obtain an estimated source signal.

For the first comb filter, it is apparent that, due to the interference of the second speech source, the pitch lag estimate wil not be very accurate and as a result, the notches of the prediction error filter do not fall at the correct fundamental frequencies. This deviation from the correct pitch harmonics, together with other estimation errors, make the source separation inadequate.

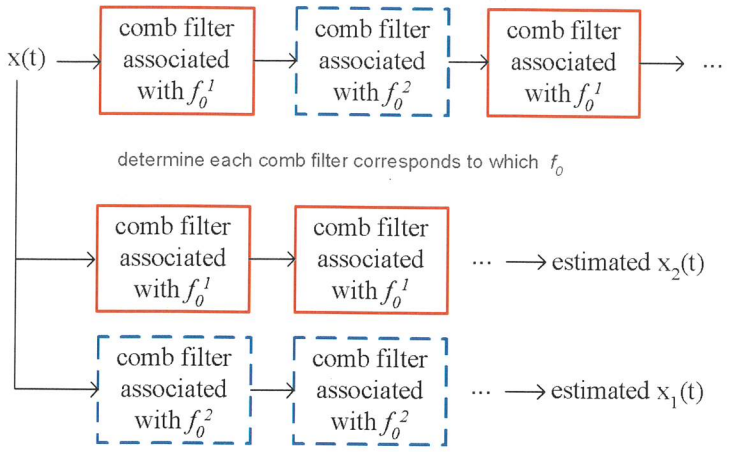

Figure 7 The architecture of the successive separation system 1. Comb filtering is applied in successive stages. After deriving the filter coefficients, all filters correspond to the same channel $x_i(t)$ are grouped together and the mixed input $x(t)$ is convolved with the cascaded filters to obtain an estimated signal for $x_j(t)$.

Successive Source Separation System 2

To prevent accumulated error propagation along the separation process in the first system, a summation constraint is adopted and the architecture of the separation system is modified and a new block diagram, we shall call it system 2, is depicted in Figure 8.

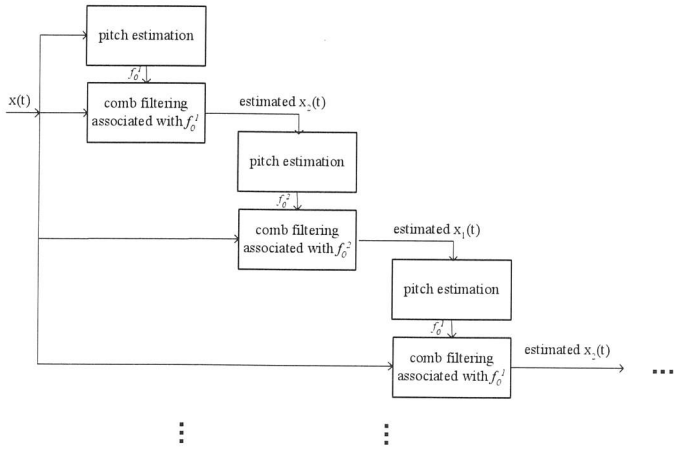

Figure 8 The architecture of the successive source separation system 2. Comb filtering is applied in successive stages. At each iteration, a comb filter is derived according to the input. The estimated source signal is then found by passing the mixed signal $x(t)$ through the comb filter. At the next iteration, a comb filter is derived according to the newly found estimation from previous stage and the whole process repeats itself till the prediction gain of each channel reaches its asymptote.

A summation constraint $x(t) = x_1(t) + x_2(t)$ is implicitly applied during each comb filtering. Ideally, if estimated $x_i(t)$ is free from $x_j(t)$, the comb filtering and prediction gain should be exactly like the case of single channel. Hence, prediction gains in successive iterations should be monotonically increasing and finally reach an asymptote.

EXPERIMENTAL SETUP

Short stationary speech signals were used to study the performance of the single-input based source separation system. Both synthetic and real speech is used in our experiments.

synthetic /AA/	formant frequencies: 700, 1220 and 2600 Hz
	pitch: 130 Hz
	output power: 60 dB
Synthetic /EY/	formant frequencies: 480, 1720 and 2520 Hz
	pitch: 77 Hz
	output power: 60 dB
Recorded male voiced speech /AA/ and /IY/	

Table 1 Experimental setup. Both synthetic and real speech is used to evaluate the performance of the proposed separation system.

RESULTS

When passing through more and more iterations, the estimated source signal becomes more and more single-source like and the notches of the comb filters become narrower and sharper.

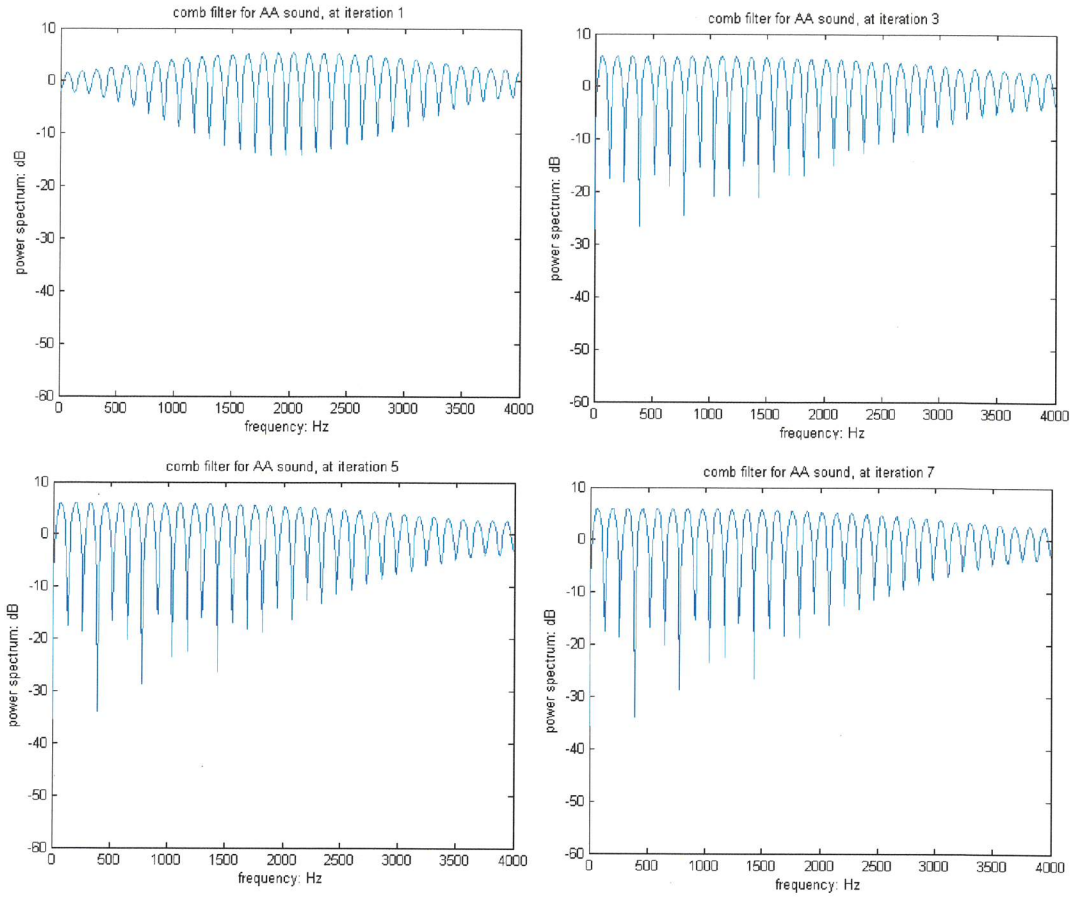

Figure 9 Power spectrum of the pitch prediction filter for the /AA/ channel. The spectral valleys become sharper in successive stages.

The two synthetic speech is mixed together at a mixing power ratio of 1 : 1. The mixed signal is fed into to the source separation system 2. In Figure 10, plots the prediction gains and the zeros of the prediction error filter, are depicted, respectively.

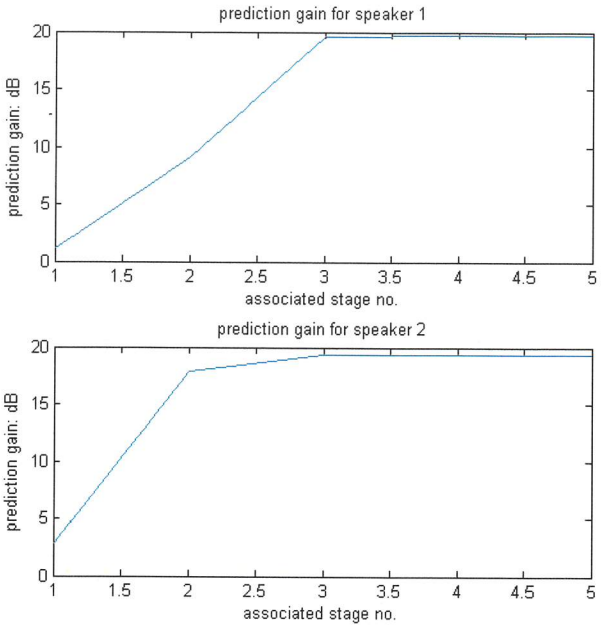

Figure 10 Prediction gain versus processing stage number for a designated speaker.

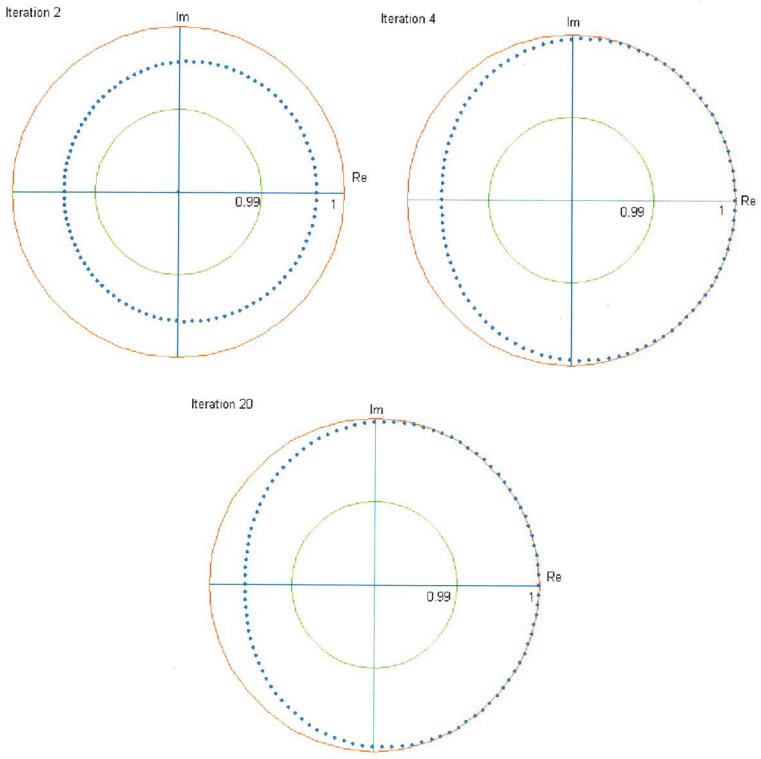

Figure 11 The zeros of the pitch prediction error filter. When the number of iteration increases, the zeros are pushed closer to the unit circle. The evolution of zeros toward the unit circle is even more distinctive for lower speech harmonics.

The proposed separation system was further evaluated by using sources with different power ratios. It is believed that if one of the source signals is much stronger than the other one, the prediction gain of the dominant channel should be higher than the case when the two sources are with the same power; the prediction gain for the other channel of lower power should be lower. Figure 12 verifies this conjecture by plotting the prediction gains for three cases of power ratios at 1:1, 7:3 and 9:1.

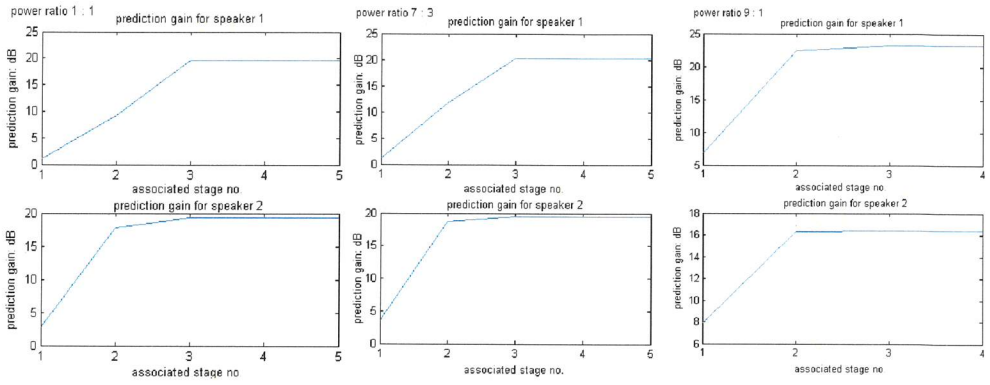

Figure 12 Prediction gains for source inputs with various power ratios. The leftmost is the case where the two sources are of equal power. The centre one is the case where the power ratio is 7 to 3, while the rightmost is the case where the power ratio is 9 to 1.

Besides synthetic speech, real voiced speech signals were recorded to test the performance of the separation system for real speech inputs. The experimental results have shown that the proposed alternate and successive harmonic filtering system is effective for separating voiced sources; however, even at ~20dB prediction gain, there is still some residual crosstalk which is subjectively noticeable. This requires further investigations.

DISCUSSIONS

From the experimental results, we have demonstrate the feasiblity of separating two voiced sources using single microphone input with the proposed successive adaptive comb filtering. Alternate, 2-channel successive, adaptive harmonic filtering with a summation constraint is the key to separate the mixed two voice sources. The proposed system was first evaluated by both source signals with the same and different power levels. It is found that the proposed system works at different signal to interference ratios (SIR's).

In the current implementation, only voiced speech signals were tested. Mixed sourced signals (voiced, unvoiced and silent signals) need to be tested for practical deployment. Besides, information-theoretic criterion (e.g. AIC or MDL ...) is necessary for determining the number of sources.