Hindawi Publishing Corporation
EURASIP Journal on Applied Signal Processing
Volume 2006, Article ID 67960, Pages 1–16
DOI 10.1155/ASP/2006/67960

A Robust Formant Extraction Algorithm Combining Spectral Peak Picking and Root Polishing

Chanwoo Kim,[1] Kwang-deok Seo,[2] and Wonyong Sung[3]

[1] *School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213-3891, USA*
[2] *Computer and Telecommunications Engineering Division, Yonsei University, Wonju, Gangwon 220-710, Korea*
[3] *School of Electrical Engineering and Computer Science, Seoul National University, Gwanak-gu, Seoul 151-744, Korea*

Received 22 September 2004; Revised 27 July 2005; Accepted 22 August 2005

Recommended for Publication by Ulrich Heute

We propose a robust formant extraction algorithm that combines the spectral peak picking, formants location examining for peak merger checking, and the root extraction methods. The spectral peak picking method is employed to locate the formant candidates, and the root extraction is used for solving the peak merger problem. The location and the distance between the extracted formants are also utilized to efficiently find out suspected peak mergers. The proposed algorithm does not require much computation, and is shown to be superior to previous formant extraction algorithms through extensive tests using TIMIT speech database.

1. INTRODUCTION

The formant is one of the most important features in speech signals,and is used for many applications, such as speech recognition, speech characterization, and synthesis. Previous formant extraction methods can largely be classified into spectral peak picking, root extraction, and analysis by synthesis [1–4]. The spectral peak picking methods and their variants have been widely used for a long time because of low computational complexity, but they often seriously suffer from the peak merger problems [1–3], where two adjoining formants are identified into a single one. The root extraction methods try to find out all the locations of roots by solving a prediction-error polynomial obtained from linear prediction coefficients (LPC), which obviously requires much computation [5]. An efficient method for evaluating the pole locations by iteratively computing the number of poles in a sector in the z-plane has been reported in [2]. However, the accuracy of the root extraction methods can hardly be high because it is not always clear to determine whether a root obtained forms a formant or just shapes the spectrum [5].

In this paper, we propose a new formant extraction algorithm that conjoins the spectral peak picking method and the root polishing scheme. In the proposed algorithm, the formant candidates are found by using the spectral peak picking method. Later, the possibility of peak mergers for each peak is examined using the screening condition among the formant frequencies of speech. As for the suspected peaks, the number of poles forming each peak is evaluated using Cauchy's integral formula. If the number of poles constituting a spectral peak is two, then the root polishing is conducted for separating the merged formants.

In this study, we used the TIMIT core test set, a widely known speech database, to compare the performance of different extractors [6]. For this purpose, we used the phone location information from TIMIT label files and compared the extracted formant values for a specific phone with the formant distribution of English vowel phonemes described in [7].

The organization of this paper is as follows: in Section 2, previous works on formant extraction methods are briefly reviewed and discussed. In Section 3, we explain characteristics of merged formants. Section 4 introduces the proposed robust formant extraction algorithm. Section 5 includes several core experimental results to prove the robustness of the proposed algorithm. We end with the concluding remarks in Section 6.

2. REVIEW OF THE PREVIOUS WORKS

In this section, we will briefly explain previous research regarding formant extraction. Basically, the speech production process is often modeled by the concatenation of the vocal tract and the lip radiation filters, while the excitation signal is generated by the glottis. References like [1] or [5] cover the theoretical backgrounds on the derivation of this

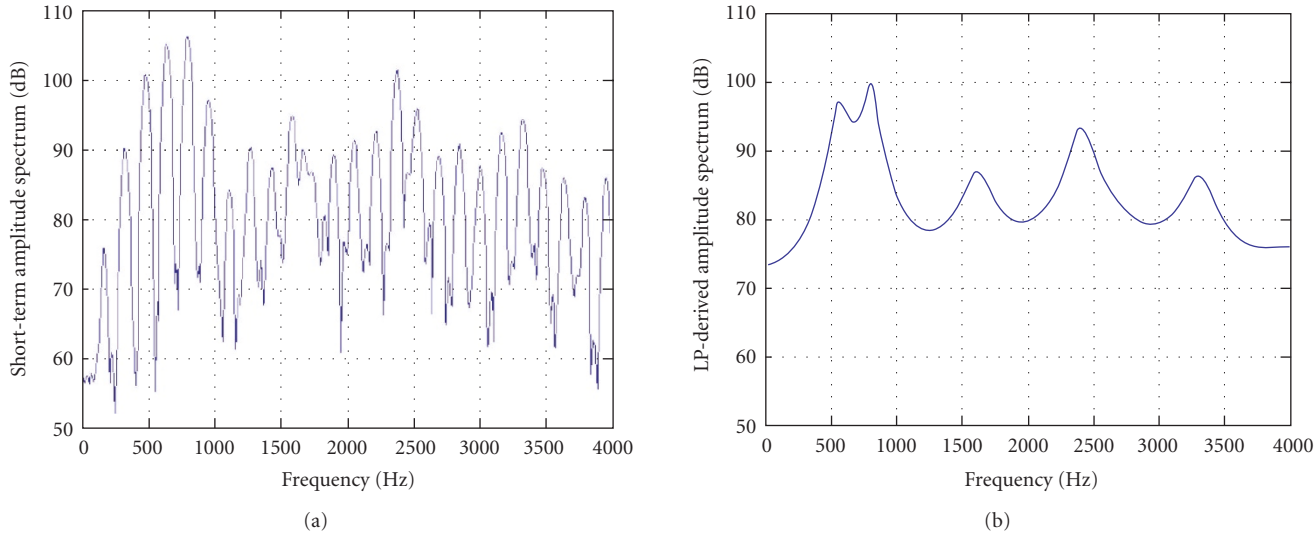


Figure 1: (a) Short-term amplitude spectrum, and (b) LP-derived amplitude spectrum of "ae" sound.

model in detail. Since the vocal tract itself is a tube with a varying cross-sectional area, it has resonant frequencies like any other tubes. These resonances are called formants, and the frequencies at which they occur are often referred to as the formant frequencies. We will explain the spectral peak picking, root extraction, and analysis-by-synthesis methods, which are the three large categories of formant extraction methods as stated in Section 1. It is an established fact that in most cases, the vocal tract system can be modeled as an all-pole system [1, 5]. Thus, the vocal tract system $H_v(z)$ can be appropriately modeled as follows:

$$H_v(z) = \frac{G_v}{\sum_{k=0}^{I} \alpha_k z^{-k}}, \tag{1}$$

where G_v is the gain factor. In this equation, we use the subscript v to denote the vocal tract system.

More importantly, it has been established by previous research that the coefficients α_k, $0 \le k \le I$, are suitably modeled by LP coefficients [1]. Thus, by computing LP coefficients, we can model the vocal tract and obtain information on formants.

2.1. Spectral peak picking method

The spectral peak picking method and its variants have been widely used for formant extraction [1–5, 8–10]. In most cases, instead of the short-term spectrum itself, smoothed spectra, such as linear prediction (LP) spectrum or cepstrally smoothed spectrum are often employed [1, 3, 5]. However, LP spectra are more often used for this purpose, since they show conspicuous peaks. Additionally, it has been verified that the prediction-error polynomial obtained from LP coefficients is closely related to the vocal tract filter, which generates the formants [1, 5]. Figure 1(a) shows the short-term spectrum of the "ae" sound, and Figure 1(b) illustrates the LP spectrum of this signal.

Here, we will briefly explain how the LP spectrum is computed, and how formant frequencies are obtained from this spectrum. Let us denote LP coefficients of a short-term speech signal by a_k, $0 \le k \le N_{\mathrm{LP}}$, where N_{LP} is the prediction order. From these LP coefficients, we can construct the following prediction-error filter:

$$A(z) = \sum_{k=0}^{N_{\mathrm{LP}}} a_k z^{-k}. \tag{2}$$

As mentioned above, previous studies show that the vocal tract filter is modeled as an all-pole system, and the vocal tract filter in (1) can be obtained from the prediction-error filter in (2) which is also known as the inverse filter (IF) [5, 10].

By performing FFT of sufficient order like 256 or 512, on the zero-padded LP coefficients, we can obtain a reasonable amplitude spectrum of the vocal tract system shown in (1).

In this paper, we will call the spectrum, obtained by the above-mentioned procedure, LP spectrum. As the name suggests, this type of formant extractors tries to find resonances on the spectrum. In general, spectral peak picking methods are advantageous in that, they show relatively reliable results, and they do not require much computation. However, as previously mentioned in the introduction, the peak merger problem is the most inherent problem. Several techniques have been proposed so far to resolve the peak merger problem [3, 11]. In [3], LP spectra are computed inside the unit circle to increase the resolving power against the peak merger cases. In [11], poles inside the unit circle have been intentionally moved on the unit circle. However, as discussed in [5], they are not perfect in distinguishing merged peaks and obtaining desired formant frequencies.

2.2. Root extraction method

Formant extraction using the root extraction method is explained in several texts and papers [1, 2, 5]. In this method, like the spectral peak picking method, we first compute linear prediction (LP) coefficients and obtain the prediction-error filter $A(z)$. Comparing with (1), we can easily find that the roots of this polynomial $A(z)$ correspond to the poles of the vocal tract system. Thus, we can obtain candidates for formants by solving $A(z) = 0$, using numerical methods.

When poles are kept sufficiently apart, and one of these poles, $z = r_0 e^{j\phi_0}$, forms a formant, the formant frequency F, and the formant bandwidth B can be represented by the following equations [1]:

$$F = \frac{f_s}{2\pi}\phi_0, \tag{3}$$

$$B = -\frac{f_s}{\pi}\ln(r_0), \tag{4}$$

where r_0 is the magnitude of the pole, ϕ_0 is the phase of the pole, f_s is the sampling frequency, F is the formant frequency, and B is the 3-dB formant bandwidth. Thus, if we find the roots of the prediction-error polynomial, we can obtain the formant frequencies using (3). In addition, we can get the bandwidth information from (4).

However, as mentioned earlier, there are several inherent problems in obtaining formant frequencies using the root extraction algorithm. Firstly, and most importantly, it is very difficult to tell whether an obtained root just shapes the spectrum or actually contributes to forming a formant [5]. If we use an LP order of 14 in obtaining $A(z)$, then there may be up to seven complex conjugate root pairs. Among these seven root pairs, we need to select three root pairs if we want to obtain the first three formant frequencies F_1, F_2, and F_3. Therefore, the root extraction method is not as reliable as the spectral peak picking method. Secondly, obtaining roots of $A(z)$ requires very high computational complexity. So, in most cases, this method is not used in real-time implementation, but for research purposes [5].

When we perform polynomial roots solving, first we can employ numerical algorithms such as Laguerre's method, Muller's method, the Eigenvalue method, and so on. It is computationally burdensome to obtain all the roots using one of these methods. To reduce the computational amount when a single root $z = z_0$ of a polynomial is obtained, we deflate the original polynomial by $(z - z_0)$ and recursively apply the roots solving algorithm. However, when deflating, round-off error often occurs and it can be accumulated. Thus, the obtained roots cannot be quite accurate. To alleviate this problem, after all of the approximate roots of $A(z) = 0$ are identified, we further polish roots which will be described in Section 2.4.

2.3. Analysis-by-synthesis method

In the analysis-by-synthesis method, we construct a synthetic spectrum and try to obtain minimized errors between the synthetic spectrum and the actual spectrum. The synthetic spectrum is obtained using the approximated formant frequencies. Thus, if the differences between the synthetic spectrum and the actual spectrum are very small, the approximated formant frequencies are close to the actual formant frequencies. Analysis-by-synthesis approximations are performed iteratively as follows: firstly, we obtain a rough estimation on formant frequencies. Secondly, using these estimated values, we obtain more accurate values that can reduce the above-mentioned differences between the synthetic and the actual spectra. This process is performed using some systematic procedures, like dynamic programming. After that, if the spectral distance is still larger than a predefined constant, then the second step is repeated. The algorithms introduced in [4, 12] describe variants of the analysis-by-synthesis type of formant extractors.

2.4. Root polishing algorithm

As previously mentioned in Section 2.2, roots obtained from the typical roots solving method and the deflation scheme often suffer from accumulated round-off errors [13, 14]. These errors accumulate when successive deflation steps are applied. So, accompanied with the roots solving procedure, root polishing is generally performed to obtain more accurate values. The root polishing algorithm works as follows [13]:

(1) *Initialization*: obtain an approximate root $z = z_0$, using the roots solving method described in Section 2.2. Set $n = 0$.

(2) *Recursion*: repeat (2-a), (2-b), and (2-c) until $n \le N_0$, where N_0 is the iteration limit.

(2a) obtain z_{n+1} by

$$z_{n+1} = z_n - \frac{A(z_n)}{A'(z_n)}, \tag{5}$$

where $A(z)$ is the prediction-error polynomial shown in (2),

(2b) test whether the following stopping condition (6) is met. If so, terminate.

$$\left| z_{n+1} - z_n \right| < \varepsilon, \tag{6}$$

(2c) set $n = n + 1$.

(3) *Termination*: take z_{n+1} as the polished root.

Unlike most root solving methods, the Newton-Raphson algorithm shows quadratic convergence [14]. Thus, the polishing step requires far less computation compared to the roots solving step. We can obtain polished roots with the required accuracy by adjusting the tolerance in (6). If the application requires more accuracy, then we need to adopt a smaller value for ε. An ε value of 10^{-4} is generally suitable for reliably obtaining formant frequencies.

3. CHARACTERISTICS OF MERGED FORMANTS

In this section, we will develop two conditions related to the poles of the vocal tract system filter. The first one deals with

the magnitude of the poles when these poles form formants. Previous research shows that some of the poles of the vocal tract system filter just shape the spectrum without a direct relation to formants [5]. Using information on the bandwidths of formants, we will derive conditions in which poles form formants. And the other condition is related to the phase difference of two adjacent poles when peak merger occurs. Although the derivation process tells us that these conditions are necessary, there may be rare exceptions to the obtained condition, since these conditions are based on assumptions obtained from experimental results by Dunn [15]. As established by previous research, two peaks that are quite close to each other are sometimes merged and appear to be a single peak. As mentioned previously, this is one of the most difficult problems occurring when we use the spectral peak picking method to extract formants. In the proposed system, the peak merger problem is resolved by inspecting the number of poles around the suspected peak using Cauchy's integral, and subsequently applying the root polishing scheme, which will be described in Section 4. For this purpose, we need to define a region, in the z-domain, where we will employ these procedures. Based on the phase difference information on the merged poles that is derived in this section, we can set an appropriate inspection region. Consequently, we only need to inspect poles inside this inspection region, where two poles may result in a single peak. These two conditions, derived in this section, are incorporated in the proposed system in order to efficiently separate a merged peak into two distinct peaks.

3.1. Magnitude condition for forming a formant

It is obvious that a pole whose magnitude is close to 1 will likely form a formant, while one that is far from 1 will not. A condition on the magnitude of a pole that can form a spectral peak can be derived as follows. From (4), we can establish the following relationship:

$$r_{\min,i} = \exp\left(-\frac{\pi}{f_s} B_{\max,i}\right), \tag{7}$$

where $B_{\max,i}$ is the maximum bandwidth for the ith formant, and $r_{\min,i}$ is the minimum magnitude of a pole that is related to the ith formant.

Previously, Dunn investigated into the range of formant bandwidths [15]. From his research, it is known that the maximum formant bandwidths of F_1, F_2, and F_3 are 160 Hz, 200 Hz, and 300 Hz, respectively. In the case of an 8 kHz sampling rate, we obtain the following results:

$$r_{\min,1} = 0.9391, \qquad r_{\min,2} = 0.9245, \qquad r_{\min,3} = 0.8889. \tag{8}$$

However, previous research shows that there exists significant variability in vowel formant characteristics. Additionally, in deriving (8), the effects of any nearby poles are ignored. Considering these facts, we should allow more tolerance to (8) for guaranteeing a more reliable condition. After repeated experiments, we obtained the following as a new condition:

$$0.8 \le r < 1.0. \tag{9}$$

In the above equation, the inequality of $r < 1.0$ is added due to the stability requirement on poles.

As shown in the following sections, this condition is employed to decide whether a pole obtained by root polishing is related to an actual formant. Note that this condition is not a sufficient condition, but a condition based on experimental results where a pole forms a formant. Thus, it cannot be used as an absolute decision rule. Admittedly, in deriving this condition, we used the experimental results on the formant bandwidths obtained by Dunn [15]. Thus, there may still exist some exceptions to this constraint (9). However, investigation into actual speech signals revealed that there seldom are such exceptions. However, by using constraint (9), we can reduce possible errors of obtaining fallacious formants. The distribution of poles of 726 frames in the z-domain is depicted in Figure 2. While many poles are satisfying (9), some of them are not. From this result, we can conclude that the latter poles are probably not directly related to the actual formants. In this figure, we also find the fact that, poles in the high-frequency region generally have smaller magnitudes, which complies with (8).

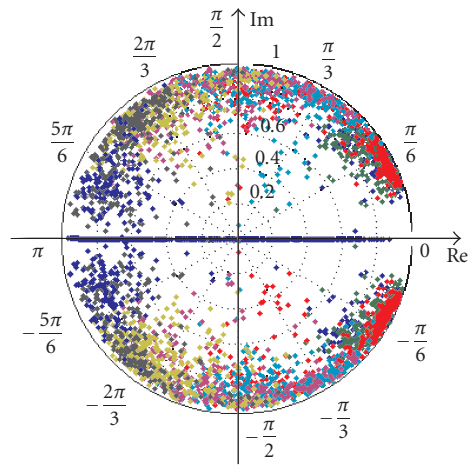


Figure 2: Distribution of poles in speech frames.

3.2. Phase condition for a peak merger

In this section, we will derive a condition on the phase difference between two poles under the following condition: two poles are directly related to two distinct formants and, at the same time, these two formants appear as a single-merged peak in the linear prediction (LP) spectrum.

Generally, the magnitude of the vocal tract system is modeled by the following equation [5]:

$$\left|H_v(e^{j\omega})\right| = \frac{G_v}{\left|\prod_{k=0}^{N}\left(1 - p_k e^{-j\omega}\right)\right|}, \tag{10}$$

where N is the order of the system, and p_k, $0 \le k \le N$, is the

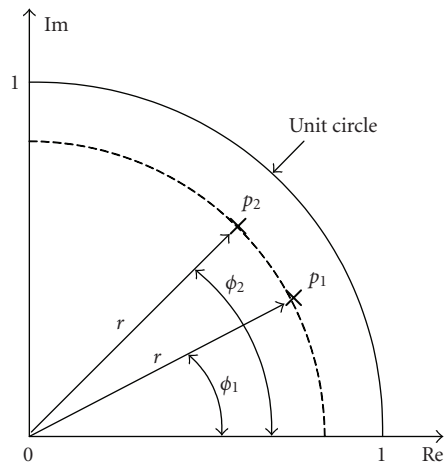


Figure 3: Two poles in the z-domain.

kth pole of the system. In this equation, ω denotes the normalized angular frequency, defined as $\omega = 2\pi(f/F_s)$, where f is the continuous-signal frequency, F_s is the sampling rate.

Without loss of generality, let us consider a case where two poles, $p_1 = r_1 e^{j\phi_1}$ and $p_2 = r_2 e^{j\phi_2}$ in (10), incur a peak merger problem. Figure 3 shows the location of these two poles in z-domain. As stated previously, a peak merger problem occurs when two distinct formants are merged into a single peak. It follows that p_1 and p_2 are the poles that form two distinct formants, even though they may appear as a single peak in the LP spectrum. Since these two poles are directly related to distinct formants, they should satisfy the constraint of (9). As shown by a lot of previous research, the peak merger occurs when these poles are very close to each other, which means that the phase difference between these two poles is small. Accordingly, in the vicinity of these two poles, (10) can be approximated by the following two-pole system:

$$|H_\nu(e^{j\omega})| \approx \frac{G'_\nu}{|1 - r_1 e^{j\phi_1} e^{-j\omega}|\,|1 - r_2 e^{j\phi_2} e^{-j\omega}|}, \quad (11)$$

where G'_ν is the gain of this modified system.

Additionally, some scrutiny on the spectrum shape reveals that the largest phase difference is obtained when each peak has the largest possible bandwidth. From (4), we find that it implies the smallest possible value of r. Thus, we obtain the largest phase difference when both magnitudes of the poles are the same and they have the minimum possible value for r. From this fact, we can substitute r_1 and r_2 in (11) with a common value r.

Consequently, the magnitude function of the system function can be represented as shown in (12) by some arithmetic

$$|\hat{H}_\nu(e^{j\omega})| = \frac{G'_\nu}{\sqrt{(1 + r^2 - 2r\cos(\omega - \phi_1))(1 + r^2 - 2r\cos(\omega - \phi_2))}}, \quad (12)$$

where ω is a normalized frequency of the sampled discrete-time signal. Real poles cannot constitute the actual formants, as can be seen in (3). Thus, poles that form formants should exist in complex conjugate pairs. Without loss of generality, we will consider two poles with positive phases in (12) since, as mentioned previously, we consider the range of $-\pi \le \omega \le \pi$ in the following derivation.

In deriving (12) from (11), we used the property that $|H_\nu(e^{j\omega})| = \sqrt{H_\nu(e^{j\omega}) H^*_\nu(e^{j\omega})}$.

If the peak merger occurs, (12) should have a single maximum value. The condition for this can be derived by differentiating the square of the reciprocal of (12) with respect to ω and, examining whether the number of roots of this derivative is one. The derivative of the squared value of (12) is as follows:

$$\begin{aligned}
&\frac{d}{d\omega}\left(\frac{G'^2_\nu}{|\hat{H}_\nu(e^{j\omega})|^2}\right) \\
&\quad = \frac{d}{d\omega}\big((1 + r^2 - 2r\cos(\omega - \phi_1)) \\
&\qquad \times (1 + r^2 - 2r\cos(\omega - \phi_2))\big) \\
&\quad = 2r\sin(\omega - \phi_1)\big((1 + r^2 - 2r\cos(\omega - \phi_2))\big) \\
&\qquad + 2r\sin(\omega - \phi_2)(1 + r^2 - 2r\cos(\omega - \phi_1)) \\
&\quad = 2r\big[\big((1 + r^2)(\sin(\omega - \phi_1)) + \sin(\omega - \phi_2)\big) \\
&\qquad - 2r(\sin(\omega - \phi_1)\cos(\omega - \phi_2) \\
&\qquad + \cos(\omega - \phi_1)\sin(\omega - \phi_2))\big].
\end{aligned} \quad (13)$$

We can further simplify (13) by the addition and the multiplication properties of trigonometric functions into:

$$\begin{aligned}
&\frac{d}{d\omega}\left(\frac{G'^2_\nu}{|\hat{H}_\nu(e^{j\omega})|^2}\right) \\
&\quad = 4r^2\left[\frac{(1 + r^2)}{r}\sin\left(\omega - \frac{\phi_1 + \phi_2}{2}\right)\cos\left(\frac{\phi_2 - \phi_1}{2}\right)\right. \\
&\qquad \left. - \sin\left(2\left(\omega - \frac{\phi_1 + \phi_2}{2}\right)\right)\right] \\
&\quad = 8r^2\sin\left(\omega - \frac{\phi_1 + \phi_2}{2}\right)\left(\frac{1 + r^2}{2r}\cos\left(\frac{\phi_2 - \phi_1}{2}\right)\right. \\
&\qquad \left. - \cos\left(\omega - \frac{\phi_1 + \phi_2}{2}\right)\right).
\end{aligned} \quad (14)$$

Close scrutiny shows that (14) has one to three roots in the range of $0 \le \omega \le \pi$, because $0 \le (\phi_1 + \phi_2)/2 \le \pi$ as assumed previously. Specifically, from the equation of $\sin(\omega - (\phi_1 + \phi_2)/2) = 0$, we can always obtain one root in the range of $0 \le \omega \le \pi$. If $((1 + r^2)/2r)\cos((\phi_2 - \phi_1)/2) < 1$, then we can find out that $|H_\nu(e^{j\omega})^2|$ has two maximum values at $(\phi_1 + \phi_2)/2 \pm \cos^{-1}(((1 + r^2)/2r)\cos((\phi_1 - \phi_2)/2))$ and a single minimum value at $\omega = (\phi_1 + \phi_2)/2$. This case corresponds to two peaks that are distinct in spectrum. However,

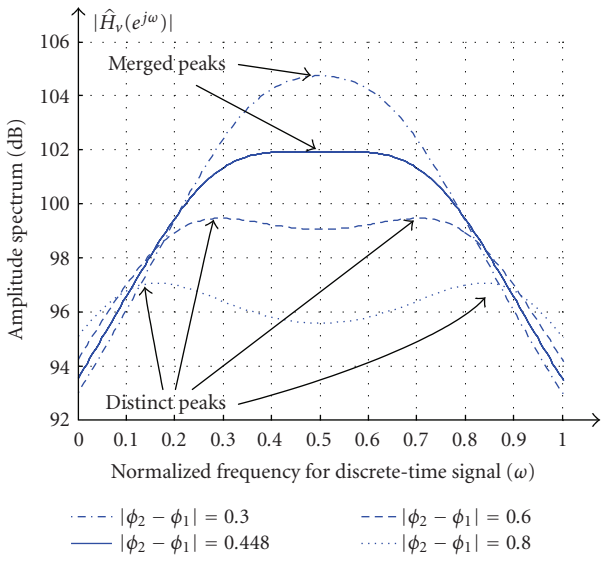


Figure 4: Magnitude plots for different values of $|\phi_2 - \phi_1|$, when $r = 0.8$.

if $((1 + r^2)/2r)\cos((\phi_2 - \phi_1)/2) \geq 1$, then we can easily find that $|H_v(e^{j\omega})^2|$ has a single maximum at $\omega = (\phi_1 + \phi_2)/2$.

Thus, the obtained condition for a peak merger is as follows:

$$|\phi_1 - \phi_2| < 2\cos^{-1}\left(\frac{2r}{1 + r^2}\right). \tag{15}$$

It is evident that as r approaches the unity, the maximum value of $|\phi_2 - \phi_1|$ satisfying (15) becomes smaller. Thus, in order to obtain a condition for a peak merger, r should take the minimum possible value which is in accordance with the previous discussion. From (9) and (15), a condition of $|\phi_1 - \phi_2| < 0.442$ rad is obtained by letting $r = 0.8$ in (15). Figure 4 shows the magnitude response of (12) for several different values of $|\phi_2 - \phi_1|$ when $r = 0.8$. From this figure, we can see that peak mergers actually occur when $|\phi_1 - \phi_2| < 0.442$, which exactly complies with our derived condition.

However, in the actual experiments, directly using (15) sometimes results in miss detections, which are largely due to the approximation involved in deriving (15) and interaction with other poles. Furthermore, an excessively large angle might lead to an increased false alarm probability, by including poles related to another peak. In this context, missed detection means that we do not detect a peak merger, which is actually present, by simply looking into the number of poles in the vicinity of the suspected peak with a central angle specified by (15). Likewise, a false alarm means that we erroneously decide that a peak merger occurs by inspecting the number of poles in the same vicinity around the suspected peak. The region used for testing the number of poles will be described in Section 4.3 in greater detail. After repeated experiments, we found a sector of the central angle 0.5498 rad to be appropriate for reducing error rates. Assuming an 8 kHz sampling rate, this value corresponds to 700 Hz. Therefore, a condition for a peak merger employed in the proposed system is that, the difference between two adjacent formant frequencies should be less than 700 Hz as follows:

$$\left|\frac{F_s}{2\pi}\phi_1 - \frac{F_s}{2\pi}\phi_2\right| < 700\,\text{Hz}, \quad \text{for 8 kHz sampling rate}, \tag{16}$$

where $F_s = 8000$ Hz is the sampling frequency. Note that $(F_s/2\pi)\phi_i$, $i = 1, 2$, is the frequency in Hz that corresponds to the phase of a pole as indicated by (3).

This result is exploited in deriving other conditions in Sections 4.2 and 4.3.

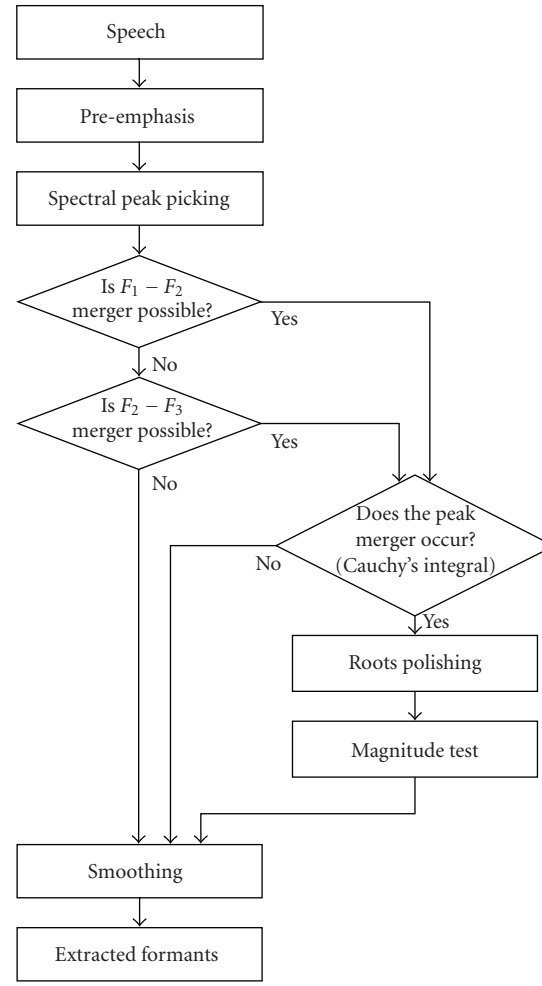


Figure 5: Block diagram of the proposed system.

4. PROPOSED METHOD

The following steps are taken to obtain the formant frequencies in each frame: finding the peaks, examining the formants locations for peak merger checking, computing the number of poles for a suspected peak, and polishing the roots. The block diagram of the proposed system is shown in Figure 5. This figure shows that we employ both the spectral peak picking method and root polishing procedure followed by a test using Cauchy's integral formula.

Note that we employed root polishing instead of direct roots solving method. Polishing two roots around the spectral peaks requires far less computation, compared to directly solving all the roots of the linear prediction-error polynomial. Also, as shown in the figure, we perform a test using Cauchy's integral formula, before root polishing, to find out whether the peak comprises two poles or a single pole. Additionally, before the test, we examine whether the peak merger is possible or not, using the data on formants distribution [7]. This procedure is shown in detail in Section 4.2. We apply Cauchy's integral only if the extracted formant frequencies satisfy this screening condition. So, the additional computation required for the entire process of peak resolving, in the proposed system, is far less burdensome than that of direct roots solving method.

4.1. Step I: finding the spectral peaks

First, if needed, the original speech signal is down sampled to 8 kHz since the first three formant frequencies are less than 4 kHz. Then, this signal is preemphasized with a preemphasis coefficient of $\mu = 0.95$, and the spectral peaks are found using LPC spectrum, as in the ordinary spectral peak picking methods [5]. A 14th-order LPC analysis is used. Previous studies show that just increasing the LP-order cannot be the solution to the peak merger problem [3]. Thus, in our cases, Step III and IV are employed to resolve the peak merger problem.

4.2. Step II: the application of screening conditions

Simple formulas for the location of the extracted formants are used to identify, whether or not, they are necessary to resolve the suspected merged peaks. This separation test is based on conditions for peak mergers, which will be explained shortly.

The advantages of this test are two folds. First of all, the amount of computation is reduced significantly, since only a small fraction, about 5% of the peaks, needs to be examined via the subsequent Cauchy's integral and the root polishing method. Secondly, this screening prevents the unnecessary resolving of poles. Note that inadequate resolving of poles often leads to accuracy degradation. This is due to the fact that there may be some poles that are not directly related with the formants. As a result, some of them may exist inside the sector that we intend to examine. Detailed explanation on this sector is given in the following subsection. As mentioned previously, the conditions (9) and (16) are not mathematically strict conditions, but based on mathematical inference from experimental results. Thus, it is still possible that a small number of the roots that are not directly related to formants may exist in this sector. In this case, erroneous resolving may occur. The following conditions are based on the distribution of formant frequencies and give us information on the possibility of peak mergers. In sum, the following conditions reduce both the computational requirement and some erroneous resolving cases.

The screening conditions employed are as follows. Let F_1, F_2, and F_3 be the extracted formant frequencies from the spectral peak picking, and F_1', F_2', and F_3' be their actual frequencies, respectively.

Condition 1

$F_2 - F_1$ (or $F_3 - F_2$) > 700 Hz in the peak merger case.

Justification for this condition: as shown in Figure 6, we can easily see that the difference between F_2 and F_1 would be large when F_1 is formed by merged formants because F_2 actually corresponds to F_3'. This figure shows the case where the peak in the lower frequency is a merged one. To justify the above condition, let us assume that F_1 is a merged formant, and $F_2 - F_1 < 700$ Hz contrary to the above condition. In this case, F_1 needs to be resolved into F_1' and F_2'. As mentioned above, F_2 corresponds to F_3'. Accordingly, from the above-mentioned assumption, we can obtain $F_3' - F_1 < 700$ Hz. It can be roughly assumed that the resolved formant frequencies are located symmetrically centered to F_1, which means $(F_1' + F_2')/2 = F_1$. From the condition for a peak merger (14), it can be derived that $F_3' - F_1' < 1050$ Hz. However, according to the possible formants distribution in [5], $F_3' - F_1' > 1050$ Hz. Thus, the assumption is wrong, and it can be stated that the difference between $F_2 - F_1$(or $F_3 - F_2$) > 700 Hz in the peak merger case.

Condition 2

$F_2 > 1800$ Hz for the peak merger between F_1' and F_2' to occur.

Justification for this condition: if the first peak is formed owing to the peak merger, then the originally extracted F_2 becomes F_3'. As can be seen in the formants distribution in [7], F_3' is larger than 2000 Hz except for "ER" sound. But in the case of "ER" sound, peak merger cannot happen since F_1 and F_2 are widely separated. Thus, if F_2 is less than 1800 Hz, this needs not be resolved.

4.3. Step III: examining peak merger

We will now describe how we can examine the peak merger around a suspected peak that satisfies the screening condition in the previous subsection. Originally, the idea of obtaining the number of poles in a given sector was presented in [2]. We employ Cauchy's integral formula introduced in their work to find out whether the peak is a merged one. When testing peak merger using Cauchy's integral formula, we employed LP prediction in the order of 10. If we adopt an LP polynomial of a much higher order, then there will be many poles that are not related to the actual formant, so it will become difficult to separate merged peaks using the pole information.

Although they perform the integration repeatedly to find out the actual phase of the pole in Snell's algorithm [2], we apply this integration for the purpose of peak merger checking. The advantages of this system can be described in two ways. First, the number of integrations is reduced significantly. Specifically, much iteration is necessary to obtain the phases of poles with sufficient accuracy in Snell's algorithm. However, in the proposed system, this integration is

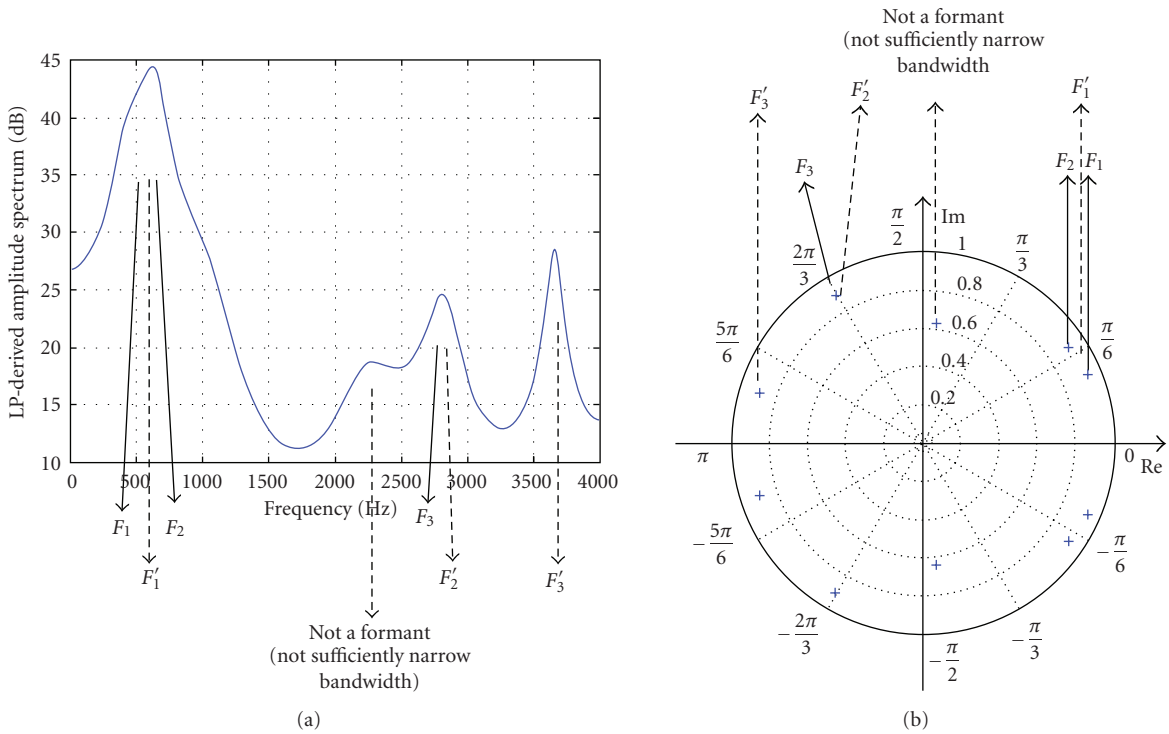


Figure 6: Actual formant frequencies and formant frequencies obtained from spectral peaks when peak merger occurs. (a) LP-derived spectrum, actual formant frequencies (F_1, F_2, and F_3), and formant frequencies obtained from spectral peaks (F_1', F_2', and F_3'), (b) pole locations, actual formant frequencies (F_1, F_2, and F_3), and formant frequencies obtained from spectral peaks (F_1', F_2', and F_3').

performed just once for each peak satisfying the condition in Step II. Secondly, it is very difficult to find out which poles are actually related to formants with Snell's algorithm, since not all of the poles are related to actual formants, as mentioned previously. Consequently, Snell's algorithm shows the performance of a typical formant extractor based on the root extraction algorithm. In contrary, we exploit information on the spectral peak and utilize this integral to resolve the peak merger problems. Thus, we do not suffer from the above-mentioned problem inherent in extractors based on roots solving.

This integration is performed in the vicinity of the peak. Let's assume that the angle related to the spectral peak is ϕ_{PEAK}. The area that we want to examine is shown in Figure 7(a). In this figure, ϕ_3 and ϕ_4 are derived by the following equations:

$$|\phi_3 - \phi_4| = \frac{700\pi}{4000}, \tag{17}$$

$$\frac{|\phi_3 + \phi_4|}{2} = \phi_{\text{PEAK}}. \tag{18}$$

In (17), the reason why we use the central angle of $(700/4000)\pi$ can be found in (16). More specifically, this is due to the fact that we want to find whether two poles satisfying the condition of (9) and (16) exist in the vicinity of a single suspected peak. Additionally, the radii of $r = 0.8$ and $r = 1.0$ are given by (9) as a condition. In the $F_1 - F_2$ resolving case, if $\phi_3 \leq 200\pi/8000$, we take $\phi_3 = 200\pi/8000$, because the lowest possible formant frequency is 200 Hz [7].

Along with this, the contour of Cauchy's integral is shown in Figure 7(b), which is the same as shown in [2]. The reason why we adopt this contour lies in the fact that we can reduce the computational burden significantly compared to the integration along the one in Figure 7(a). When performing the integration along the contour in Figure 7(b), it is possible that poles not meeting the constraint $0.8 < r < 1.0$ are selected. These poles are filtered through the subsequent root polishing algorithm. Note that the root polishing algorithm described in the next subsection gives us the magnitude of the pole as well as its phase.

We can denote the above-mentioned sector in Figure 7(b) by (19):

$$\begin{aligned} &\Gamma_1 : 0 \leq r \leq 2, \quad \phi = \phi_3, \\ &\Gamma_2 : r = 2, \quad \phi_3 \leq \phi \leq \phi_4, \\ &\Gamma_3 : 0 \leq r \leq 2, \quad \phi = \phi_4. \end{aligned} \tag{19}$$

As shown in [2], we can obtain the number of poles inside this sector by

$$n(\Gamma) = \frac{1}{2\pi j} \int_{\Gamma} \frac{A'(z)}{A(z)} dz, \tag{20}$$

where polynomial $A(z)$ is the prediction-error polynomial, and Γ is the sector composed of three curves Γ_1, Γ_2, and Γ_3 in (19). For the integration on the curves Γ_1 and Γ_3, the composite Simpson's rule [14] is employed. The curves are partitioned into short segments, having an equal length to perform the numerical integration. For the integral on the curve

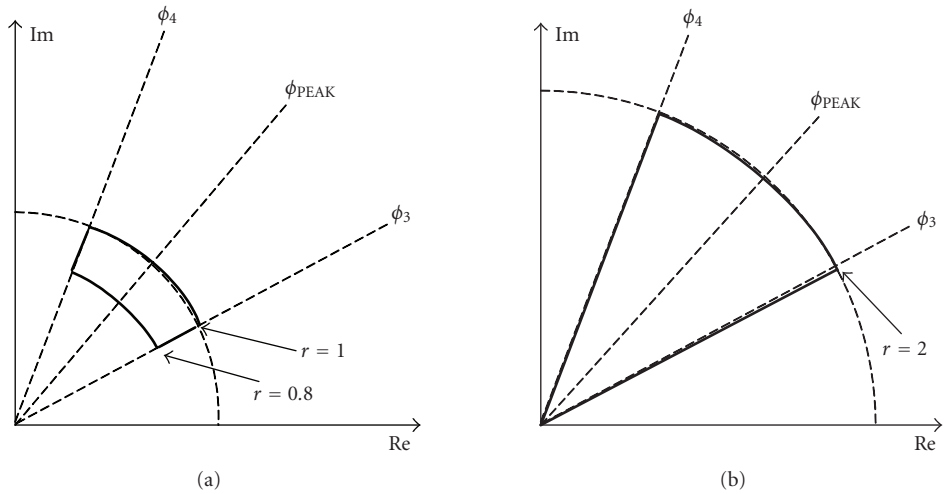


Figure 7: (a) Test area for a peak merger, and (b) contour for Cauchy's integral.

Γ_2, the approximate value of $N|\phi_4 - \phi_3|$ was used to reduce computation as in [2]. In this approximation, N denotes the LPC order. For more details on this approximation value, you are referred to [2].

4.4. Step IV: resolving poles by polishing the roots

If the result of Cauchy's integration in Step III is two, then the two poles that constitute the merged peak are obtained in the following manner. To begin with, it is quite natural that (3) can be applied to these poles because these two poles are directly related to the spectral peak. Thus, the initial approximate phase values of these two values can be given by

$$\phi_0^{(0)} = \phi_1^{(0)} = \frac{2\pi F}{f_s}, \tag{21}$$

where $\phi_0^{(0)}$ and $\phi_1^{(0)}$ are the approximate values of the phases of these two poles, respectively. In the notations of $\phi_0^{(0)}$ and $\phi_1^{(0)}$, the subscript 0 and 1 denote each pole, and the superscript (i) denote the iteration number which will be described subsequently. In (21), F is the frequency of the spectral peak in Hz to which these poles are directly related, and f_s is the sampling frequency of the speech signal. Along with estimating the phase value, we also need to estimate the approximate magnitudes of these two poles. Also note that (3) is derived under the assumption that poles are kept sufficiently apart. When two poles form a single peak, they are quite close to each other. Thus, (21) does not yield quite accurate values in the merged peak case. However, the obtained values from (21) should be in the neighborhood of the actual roots, so we can obtain more accurate values by the root polishing algorithm, which will be explained in detail. As previously mentioned in (9), the typical range of magnitudes of poles that constitute formants is given by $0.8 \le r < 1.0$. Thus, we adopt the initial approximate value of magnitude $r_0^{(0)}$ and $r_1^{(0)}$ as follows:

$$r_0^{(0)} = r_1^{(0)} = 0.9. \tag{22}$$

Thus, from (21) and (22), we obtain the approximate values of these two roots $z_0^{(0)}$ and $z_1^{(0)}$ by

$$z_0^{(0)} = z_1^{(0)} = 0.9e^{j(2\pi F/f_s)}. \tag{23}$$

After obtaining the initial approximation of (23), Bairstow's algorithm [13], that is, a variation of Newton-Raphson method, is used to obtain the roots by polishing this approximate value into the exact value. In Bairstow's algorithm, we try to seek the quadratic factors. Since the coefficients of the prediction-error polynomial $A(z)$ in (2) are all real, then the complex conjugates of $z_0^{(0)}$ and $z_1^{(0)}$ are also roots of $A(z)$.

Specifically, the quadratic factor that has a root of $z_0^{(0)}$ should be the following form:

$$\left(z^2 + B_0^{(0)} z + C_0^{(0)}\right) = 0, \tag{24}$$

where

$$B_0^{(0)} = -z_0^{(0)} - \left(z_0^{(0)}\right)^* = -1.8\cos\left(\frac{2\pi F}{f_s}\right), \tag{25}$$

$$C_0^{(0)} = \left|z_0^{(0)}\right|^2 = 0.81. \tag{26}$$

If we divide the prediction polynomial $A(z)$ by $z^2 + B_0^{(0)} z + C_0^{(0)}$, then we obtain the following relationship:

$$A(z) = \left(z^2 + B_0^{(0)} z + C_0^{(0)}\right) Q(z) + Rz + S, \tag{27}$$

where $Q(z)$ is the quotient, and $Rz + S$ is the linear remainder. In essence, Bairstow's algorithm numerically finds the quadratic factor, which makes both R and S in (25) converge

to 0. Now, Bairstow's algorithm works in the following manner:

(1) *Initialization*: obtain $B_0^{(0)}$ and $C_0^{(0)}$ from (24) and (25). Set $n = 0$,

(2) *Recursion*: repeat (2a), (2b), (and 2c) until $n \leq N_0$, where N_0 is the iteration limit.

(2a) from $B_n^{(0)}$and $C_n^{(0)}$, obtain $B_{n+1}^{(0)}$ and $C_{n+1}^{(0)}$ by employing two-dimensional Newton-Raphson method,

(2b) test whether the coefficient has been converged by applying the following stopping condition. If both of (28) and (29) are met, go to step (3). Otherwise, continue the recursion step.

$$\left|B_{n+1}^{(0)} - B_n^{(0)}\right| \leq \varepsilon_1 \left|B_{n+1}^{(0)}\right| \quad \text{or} \quad \left|B_{n+1}^{(0)}\right| \leq \varepsilon_2, \tag{28}$$

$$\left|C_{n+1}^{(0)} - C_n^{(0)}\right| \leq \varepsilon_1 \left|C_{n+1}^{(0)}\right| \quad \text{or} \quad \left|C_{n+1}^{(0)}\right| \leq \varepsilon_2. \tag{29}$$

In (28) and (29), ε_1 and ε_2 are constants for convergence checking. In our system, we adopt the values of $\varepsilon_1 = 0.001$ and $\varepsilon_2 = 0.0001$,

(2c) set $n = n + 1$.

(3) *Termination*: obtain $z_0^{(n+1)}$ by solving the quadratic equation:

$$z^2 + B_0^{(n+1)} z + C_0^{(n+1)} = 0. \tag{30}$$

Because this equation is quadratic, we generally obtain the roots in the complex conjugate form. Among them, the one with the positive phase value is our desired root $z_0^{(n+1)}$.

After obtaining the desired value of $z_0^{(n+1)}$, we divide the prediction-error polynomial $A(z)$ by $(z^2 + B_0^{(n+1)} z + C_0^{(n+1)})$. And we apply the above-mentioned Bairstow's algorithm once gain to obtain $z_1^{(n+1)}$.

This method has the advantage of not requiring complex arithmetic, while the standard Newton-Raphson method resorts to complex arithmetic for polishing complex roots. Although this method cannot be used broadly, because of the stability problem, in the proposed system, we do not encounter this problem since the initial approximation (23) is sufficiently close to the accurate roots. We can find that the roots converge with sufficient accuracy, satisfying the stopping condition in (28) and (29) after three or four iterations.

Sometimes roots with $r < 0.8$ or outside, this sector may be selected. In this case, the obtained roots should be discarded due to the constraint (9). After obtaining the roots, the formant frequencies can be obtained by (3). This is a clear advantage compared to the bisection method described in [2] or the conventional roots-extraction-type formant extractor [5, 9, 10], which directly solves $A(z) = 0$.

5. RESULTS

Previous research of formants shows that there are high correlations between a specific vowel and its formant frequencies [5, 7]. The following Table 1 shows the typical values of formant frequencies that we used for accuracy checking [5, 7]. These values are used as the decision criterion whether a peak merger occurred or not in the testing phase.

Table 1: Typical values of formant frequencies.

Vowel	F_1	F_2	F_3
iy	270	2290	3010
ih	390	1990	2550
eh	530	1840	2480
ae	660	1720	2410
aa	730	1090	2440
ao	570	840	2410
uh	440	1020	2240
uw	300	870	2240
ah	640	1190	2390
er	490	1350	1690

Figure 8 shows a sample speech frame where a peak merger in the formant frequencies occurred. In this frame, the formant frequencies obtained from the peaks with sufficient bandwidth are $F_1 = 593.8$ Hz, $F_2 = 2712.1$ Hz, and $F_3 = 3514.4$ Hz, respectively. The LP spectrum with LP order 10 in Figure 8(a) confirms this result. However, when tested for peak mergers with this system, the peak in the lower frequency is found to be made of two poles as shown in Figure 8(b), and the subsequent roots testing and polishing procedures modify the formant frequencies in this frame to $F_1 = 569.5$ Hz, $F_2 = 854.3$ Hz, and $F_3 = 2712.1$ Hz. In this case, the pronounced vowel is "AO," and you can find that the corrected formant frequencies are in accordance with the typical frequencies shown in Table 1.

Figure 9 shows the spectrogram of the word "pineapple" and the extracted formant frequencies using the conventional spectral peak picking method and the proposed algorithm. At the onset of speech, the first and the second formants are very close, so they form a single peak. In this part of speech, the pronounced phone is /AA/, thus, as shown in Table 1, the F_1 and F_2 are very close to each other. The region in ellipsis in Figure 9(a) denotes the merged peak. And, in this case, the duration of speech where the peak merge occurs is rather long, so it is very difficult to correct the result using conventional formant tracking or smoothing methods. But, as shown in Figure 9, the proposed algorithm yields desirable results even for this part of the speech.

We evaluated the proposed method on a TIMIT core test set, which comprises 240 speech samples spoken by 10 speakers. In the test phase, we performed the accuracy decision in the Mel scale. If the extracted ith formant frequency in the Mel scale is closest to the jth formant frequency in this table, in Mel scale and $i \neq j$, then we conclude the extraction result to be inaccurate. Otherwise, we decide this result to be accurate. This decision criterion is employed in the following accuracy evaluation. Since there are some variations in actual formant frequencies, this test criterion cannot be used for checking the accuracy of extracted formant frequencies with very high reliability. However, this criterion is very

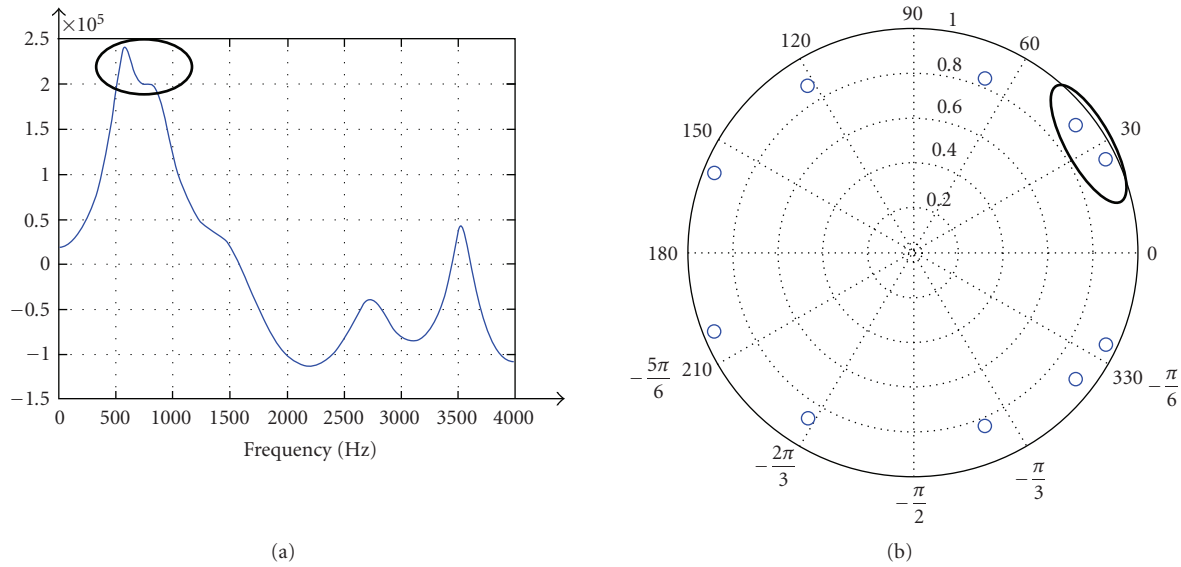


Figure 8: (a) LP spectrum, and (b) the pole locations of a frame. The ellipses indicate poles forming a peak merger.

useful for detecting errors due to the peak merger for large speech DBs like TIMIT, since a computer program for testing this criterion can be easily implemented. If this criterion tells us that a peak merger or extraction error occurs, we also check whether this test result is correct by investigating the extracted formant frequencies and comparing them with the spectrogram of the speech, and identifying the phone label of the speech.

Table 2 shows phone label information in a speech sample in a TIMIT DB. This sample can be found in TEST/DR1/MJSW0/SI1640.PHN. Since the original TIMIT phone label information is given in units of the sampling index, we changed the base unit from the sampling index to time in this table.

Figure 10 shows extracted formant results on this part of speech. In Figure 10(a), the formant extraction result was obtained using the standard ESPS formant extraction algorithm incorporated in WaveSurfer [16]. As widely known, the ESPS formant extractor shows good performance in most cases. In obtaining this figure using WaveSurfer, we first down sampled this 16 kHz TIMIT speech sample into 8 kHz one. According to our experiments, errors occur more frequently, when we use the ESPS formant extractor for 16 kHz speech samples, rather than 8 kHz speech samples. Figures 10(b) and 10(d) show the formant extraction result using the conventional spectral peak picking method and root extraction algorithm without additional smoothing. Compared to these results, Figure 10(c) illustrates the formant extraction result obtained, using the proposed method. As shown in Figure 10(a), the ESPS formant extractor appears more robust against the peak merger problem. This is because the ESPS formant extractor is not based on the spectral peak picking method, but on the root extraction method. As stated before, most of the formant extractors based on the root extraction algorithm have difficulty in selecting roots that are directly related to actual formants. However, in the case of the ESPS formant extractor, a modified Viterbi algorithm is employed to find the most probable poles related to actual formants. By adopting this scheme, the ESPS formant extractor shows sufficiently good performances in most cases. However, even the ESPS formant extractor sometimes misses in selecting some resonances. As shown in this figure, for the /W/ phone, the extractor incorrectly selects the third formant frequency. By looking into the spectrogram in detail and following the movement of the spectral peaks, we can find that the fourth formant frequency obtained for the /W/ sound should be the third formant frequency. The proposed algorithm shown in Figure 10(c) shows a better result, even without sophisticated smoothing algorithms. Another advantageous aspect of our proposed algorithm is that it requires far less computation compared to the formant extractors based on roots solving, as previously described in Section 2. Figure 10(b) shows the formant extraction result obtained using the conventional peak picking algorithm. As you can see in this figure, there are many errors in the extracted formant frequency due to the peak merger problems. Compared to this result, our proposed algorithm in Figure 10(c) shows good performance in resolving the peak merger. When a smoothing algorithm is not employed, the extraction result obtained using the root extraction algorithm shows the poorest result as shown in Figure 10(d). In the time between 0.75s and 0.78s, it seems that there are errors in the proposed method and the spectral peak picking method as shown in Figures 10(b) and 10(c). During this part of the speech signal, the ESPS formant extractor shows somewhat better results. This is due to the fact that, for nasalized sounds, we need an additional zero to model the vocal tract system [1]. If the zero is located in the vicinity of the pole that forms a formant, it is very difficult to extract that formant from the LP spectrum. In these particular cases, formant extractors based on root solving may show better results.

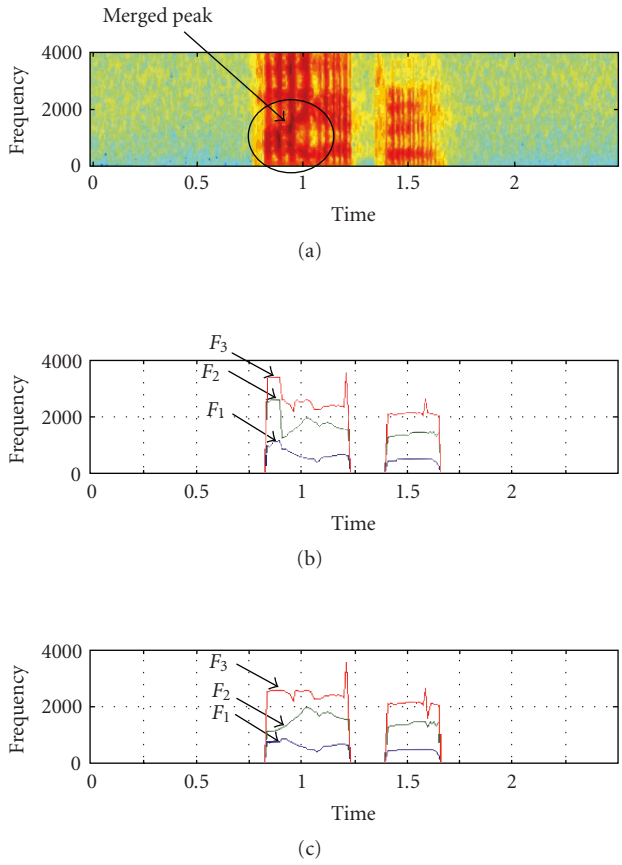


Figure 9: (a) Speech spectrogram (ellipsis in this figure denotes the merged peak), (b) formants tracking result with the conventional spectral peak picking method, and (c) formants tracking result with the proposed method.

Table 2: Phone location of a portion of a sample speech in TIMIT DB (speech file: TEST/DR1/MJSW0/SI1640.WAV).

Beginning time	Ending time	Phone name
0.455	0.500	vcl
0.500	0.617	w
0.617	0.702	ah
0.702	0.785	n
0.785	0.805	vcl

Table 3 shows the formant extraction results using the conventional spectral peak picking method for a speech sample in the well-known TIMIT DB. In obtaining these formant values, we used an LP order of 14 and an FFT order of 512. The window size is 30ms, and the frame rate is 10ms. Figure 10(b) illustrates the plot of the formants obtained from this speech sample, in the range of $0.5 \le t \le 0.65$, using the conventional spectral peak picking method. As shown in Table 3 and Figure 10(b), peak mergers occurred many times in the /W/ phone. This merger occurred since the first and second formant frequencies are very close, as shown in Figures 10(a) and 10(c).

In contrast, Table 4 shows the extraction results when the proposed algorithm is employed. As you can see in this table, the peak merger problems have been successfully figured out. Figure 10(c) also shows that we obtain the desired formant frequency at this region, and continuity in the extracted formant frequency can be maintained.

More detailed information on pole locations and LP spectra can be found in Figure 11 at different time t. Figures 11(c), 11(d), and 11(e) show the cases of peak merger. By comparing with the pole locations in z-plane, you can find that the peak in the lowest frequency around 600 Hz is actually composed of two spectral peaks. As previously mentioned, the /W/ sound is pronounced in this part of speech. However, by employing Cauchy's integral and root polishing scheme, we can distinguish two resonances and obtain correct values as shown in Figure 10(c) and Table 3.

After testing our algorithm on this test set, we can conclude that most of the $F_1 - F_2$ merger problems occured in the "AA" and "AO" sounds. Note that the difference between F_1 and F_2 is very small in these sounds as shown in Table 1. The "AA" and "AO" vowels constitute about 5.2% of the 10 vowels we tested. The TIMIT core test set had 250 samples. During the test, using these sounds, the proposed system yielded a performance accuracy of 87%, which is significantly higher than that of the spectral peak picking method's 81.6%. This result proves that the proposed method is robust enough for the peak merger problem. On the other hand, the performance of the root extraction method is the worst (less than 50%). This is partly because there is no clear way to relate the solved roots to formants. Note too that no smoothing technique was employed for any of the extractors during the evaluation process. For the speech frames, where peak mergers do not happen, the proposed algorithm and the spectral peak picking method showed almost the same performance.

6. CONCLUDING REMARKS

In this study, a robust formant extraction algorithm, which sequentially applies the spectral peak picking, formants location examining, and the root polishing, is developed. One of the most notable advantages of the proposed system lies in its robustness against the peak merger problem that was extremely difficult to be solved using conventional spectral peak picking methods. Although roots solving method in themselves show poor accuracy, we successfully exploited information on poles around the merged peaks in tackling the above-mentioned pole merger problem. We also propose the root polishing scheme for obtaining two distinct formant frequencies from a single merged peak, requiring significantly less computation when compared to the direct roots solving method. As well, several conditions on poles and formants are devised in order to enhance the accuracy result and/or reduce the computational burden. Consequently, this method not only shows better results at the intensive test using TIMIT

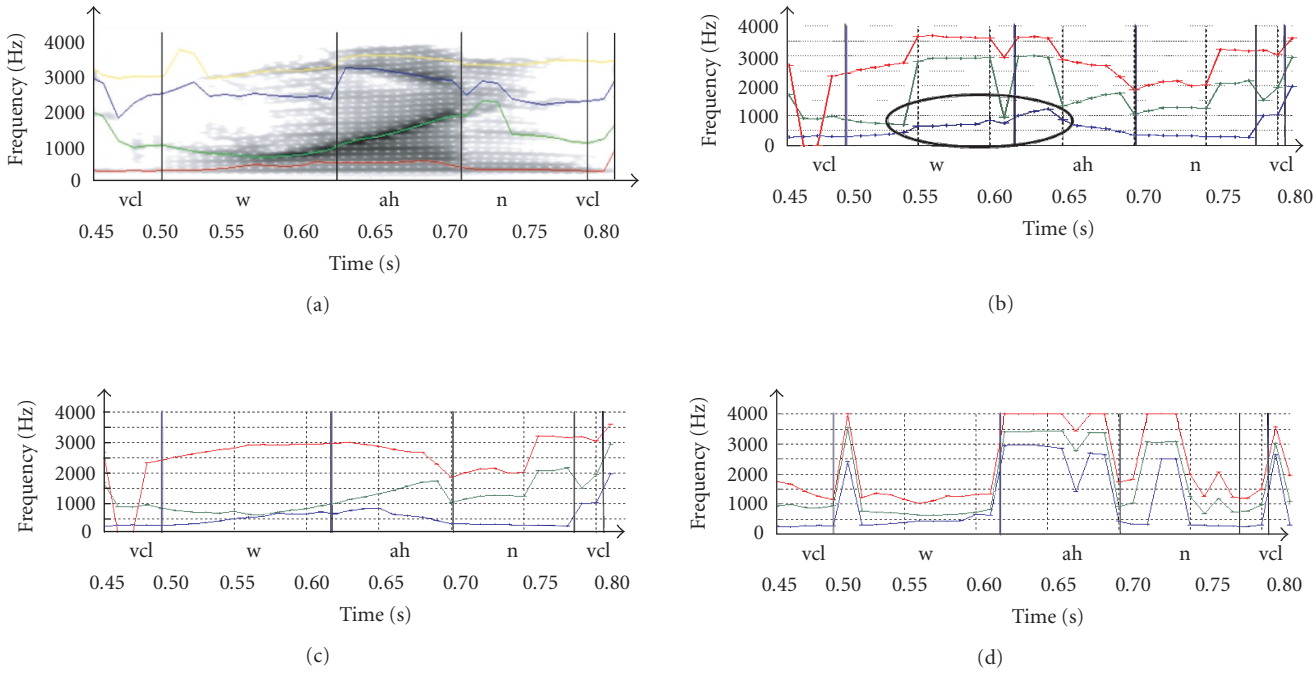


Figure 10: Spectral shape and the location of poles for a speech sample in TIMIT DB(TEST/DR1/MJSW0/SI1640.WAV) (ellipsis in this figure denotes the merged peak). (a) Formant frequencies obtained using WaveSurfer, (b) formant frequency obtained using the spectral peak picking method, (c) formant frequency obtained using the proposed algorithm, and (d) formant frequency obtained using root extraction algorithm.

Table 3: Formant extraction results for a speech sample in TIMIT DB using the spectral peak picking method (speech file: TEST/DR1/MJSW0/SI1640.WAV).

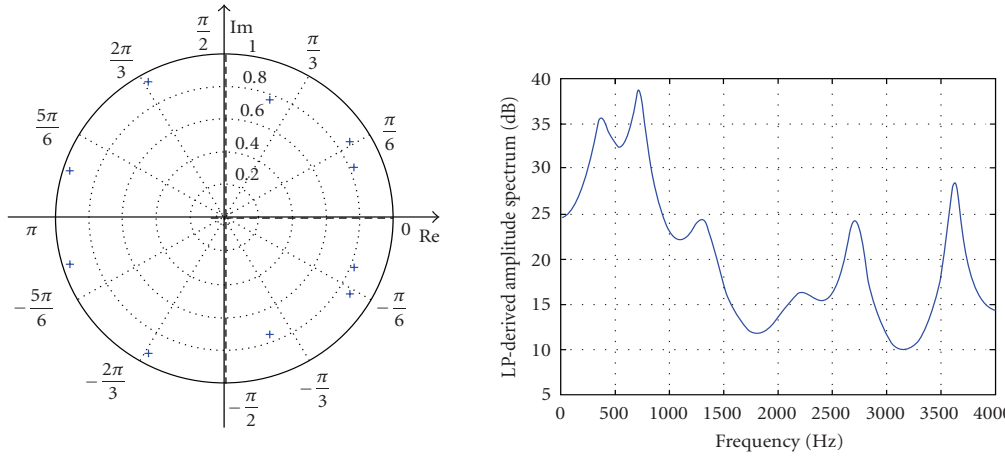
Time (t)	F_1	F_2	F_3	Merger
0.50	296.9	859.4	2421.9	Not merged
0.51	312.5	781.3	2531.3	Not merged
0.52	343.8	734.4	2612.4	Not merged
0.53	375.0	718.8	2693.2	Not merged
0.54	437.5	687.5	2765.6	Not merged
0.55	640.6	2812.5	3656.3	Merged
0.56	640.6	2921.9	3687.5	Merged
0.57	671.9	2921.9	3642.1	Merged
0.58	687.5	2921.9	3624.6	Merged
0.59	703.1	2934.1	3613.4	Merged
0.60	843.8	2947.8	3587.2	Merged
0.61	734.5	937.5	2957.2	Not merged
0.62	1000.0	2979.8	3627.4	Merged
0.63	1140.6	3000.0	3650	Merged
0.64	1218.7	2942.1	3592.2	Merged
0.65	859.3	1328.1	2875.0	Not merged

Table 4: Formant extraction results for a speech sample in TIMIT DB using the proposed algorithm (speech file: TEST/DR1/MJSW0/SI1640.WAV).

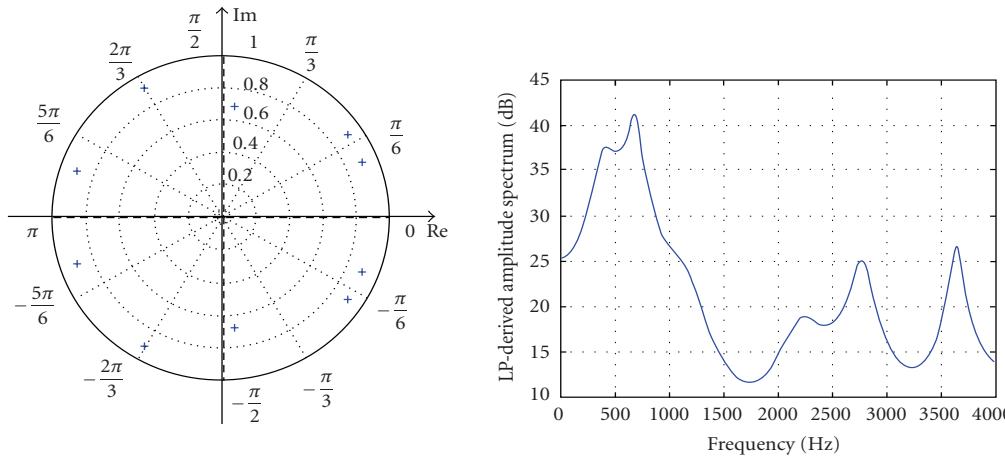
Time (t)	F_1	F_2	F_3	Merger
0.50	296.9	859.4	2421.9	Not merged
0.51	312.5	781.3	2531.3	Not merged
0.52	343.8	734.4	2612.4	Not merged
0.53	375	718.8	2693.2	Not merged
0.54	437.5	687.5	2765.6	Not merged
0.55	509.6	749.6	2812.5	Resolved
0.56	556.8	651.3	2921.9	Resolved
0.57	579.4	636.3	2932.1	Resolved
0.58	687.5	721.4	2921.9	Resolved
0.59	665.5	785.3	2934.1	Resolved
0.60	666.8	841.7	2947.8	Resolved
0.61	734.4	937.5	2957.2	Not merged
0.62	672.7	1007.0	2979.8	Resolved
0.63	782.1	1140.6	3000	Resolved
0.64	841.2	1218.8	2942.1	Resolved
0.65	841.2	1328.1	2875.0	Not merged

speech database, but also requires very little additional computation. The reason for this is, because root polishing needs to be applied only to a small portion of the speech frames.

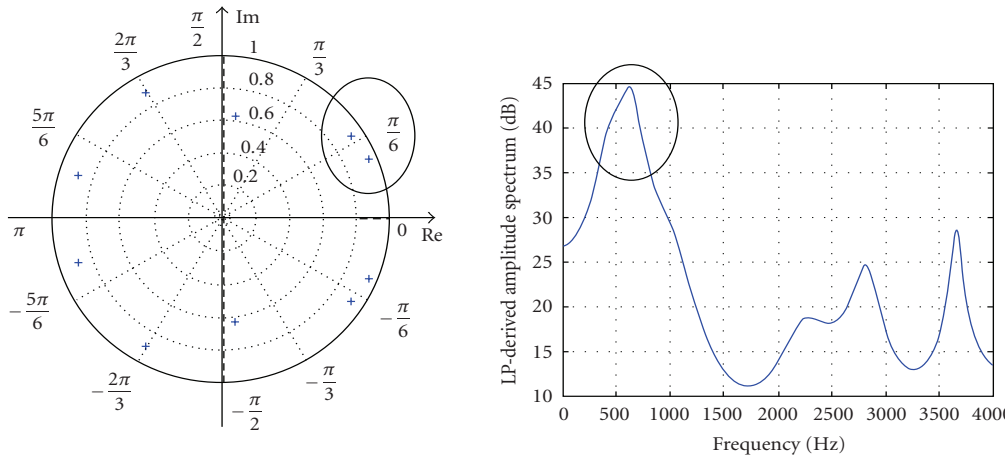
The proposed method is now being incorporated into our previously developed vowel-pronunciation checking system for foreign language learning [8] to obtain improved



(a)



(b)



(c)

FIGURE 11: Continued.

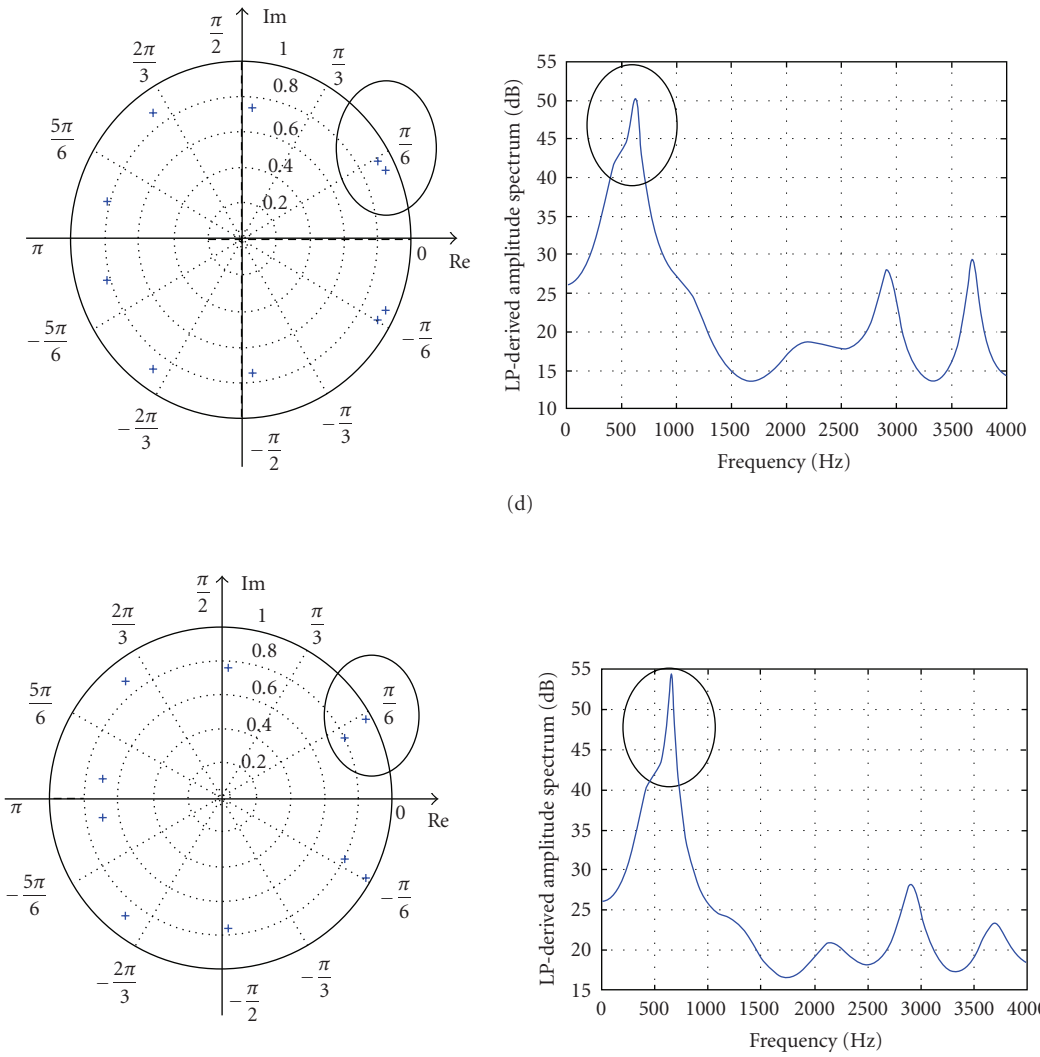


Figure 11: Pole locations and LP spectra for a speech sample in TIMIT DB (TEST/DR1/MJSW0/SI1640.WAV) (ellipsis in this figure denotes the merged peak). (a) Pole location and LP spectrum at time 0.53s, (b) pole location and LP spectrum at time 0.54s, (c) pole location and LP spectrum at time 0.55s, (d) pole location and LP spectrum at time 0.56s, and (e) pole location and LP spectrum at time 0.57s.

performance compared in [8]. We also expect that applications such as speech recognition, formant vocoder, or text-to-speech system (TTS) will benefit from this robust extractor.

ACKNOWLEDGMENT

This study was supported by the National Research Laboratory program (2000–X–7155), Brain Korea 21 Project (0019-19990027) in Seoul National University, and Yonsei University Research Fund of 2005.

REFERENCES

[1] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, Englewood Cliffs, NJ, USA, 1978.

[2] R. C. Snell and F. Milinazzo, "Formant location from LPC analysis data," *IEEE Transactions on Speech and Audio Processing*, vol. 1, no. 2, pp. 129–134, 1993.

[3] S. S. McCandless, "An algorithm for automatic formant extraction using linear prediction spectra," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 22, no. 2, pp. 135–141, 1974.

[4] L. Welling and H. Ney, "Formant estimation for speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 1, pp. 36–48, 1998.

[5] J. R. Dellar Jr., J. G. Proakis, and J. H. L. Hansen, *Discrete-Time Processing of Speech Signals*, Macmillan, New York, NY, USA, 1993.

[6] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "Darpa TIMIT acoustic-phonetic continuous speech corpus," Tech. Rep. NISTIR 4930, U.S.

Department of Commerce, National Institute of Standards and Technology, Gaithersburg, Md, USA, 1993.
[7] G. E. Peterson and H. L. Barney, "Control methods used in a study of the vowels," *Journal of the Acoustical Society of America*, vol. 24, no. 2, pp. 175–194, 1952.
[8] C. Kim and W. Sung, "Vowel pronunciation accuracy checking system based on phoneme segmentation and formants extraction," in *Proceedings of International Conference on Speech Processing*, pp. 447–452, Daejeon, Korea, August 2001.
[9] J. D. Markel, "Digital inverse filtering: a new tool for formant trajectory estimation," *IEEE Transactions on Audio and Electroacoustics*, vol. 20, no. 2, pp. 129–137, 1972.
[10] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *Journal of the Acoustical Society of America*, vol. 50, no. 2B, pp. 637–655, 1971.
[11] G. S. Kang and D. C. Coulter, "600 bits per second voice digitizer (linear predictive formant vocoder)," Naval Research Laboratory Report 8043, Washington, DC, USA, November 1976.
[12] C. G. Bell, H. Fujisaki, J. M. Heinz, K. N. Stevens, and A. S. House, "Reduction of speech spectra by analysis-by-synthesis techniques," *Journal of the Acoustical Society of America*, vol. 33, no. 12, pp. 1725–1736, 1961.
[13] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C*, Cambridge University Press, Cambridge, UK, 1992, pp. 376.
[14] R. L. Burden and J. D. Faires, *Numerical Analysis*, Brooks/Cole, Pacific Grove, Calif, USA, 1997.
[15] H. K. Dunn, "Methods of measuring vowel formant bandwidths," *Journal of the Acoustical Society of America*, vol. 33, no. 12, pp. 1737–1746, 1961.
[16] WaveSurfer, Center for Speech Technology (CTT) at KTH, Stockholm, Sweden, available at http://www.speech.kth.se/wavesurfer/.

Chanwoo Kim received the B.S. and M.S. degrees in electrical engineering from Seoul National University, Seoul, Korea, in 1998 and 2001, respectively, and is currently working toward the Ph.D. degree at the School of Computer Science, Carnegie Mellon University. From 2000 to 2002, he worked on speech recognizers and embedded signal processing systems for Edumedia Technologies. From 2003 to 2005, he was with LG Electronics. His research interests include multimedia systems, speech recognition system, speech analysis, and embedded systems for signal processing.

Kwang-deok Seo received the B.S., M.S., and Ph.D. degrees in electrical engineering from Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea, in 1996, 1998, and 2002, respectively. From August 2002 to February 2005, he was with LG Electronics. Since March 2005, he has been a Faculty Member in the Computer and Telecommunications Engineering Division, Yonsei University, Gangwon, Korea, where he is an Assistant Professor. He has over 30 pending or issued patents and has published over 30 papers in the areas of multimedia coding, multimedia signal processing, and multimedia communication systems. He is a Member of KICS, IEEE, and IEICE.

Wonyong Sung received the B.S. degree in electronic engineering from the Seoul National University in 1978, the M.S. degree in electrical engineering from the Korea Advanced Institute of Science and Technology (KAIST) in 1980, and the Ph.D. degree in electrical and computer engineering from the University of California, Santa Barbara, USA in 1987. From 1980 to 1983, he worked at the Central Research Laboratory of the Gold Star (currently LG Electronics) in Korea. He has been a Member of the Faculty of the Seoul National University since 1989. From January of 1998 to December of 1999, he worked as a Chief of the SEED (System Engineering and Design Center) in Seoul National University. He was an Associate Editor of the IEEE Transaction Circuits and Systems II from 2000 to 2001, and is a design and implementation technical committee Member of the IEEE Signal Processing Society. He founded a venture company, Edumedia Technologies, in 2000, and has developed a handheld educational device for kids, Speaking Partner. His major research interests are the development of fixed-point optimization tools, implementation of VLSI for digital signal processing, parallel implementation of multimedia programs, and development of multimedia software for handheld devices.