New Temporal Features for Robust Speech Recognition with Emphasis on Microphone Variations

Jia-lin Shen and Wen L. Hwang
Institute of Information Science, Academia Sinica
Nankang, Taipei, Taiwan, Republic of China
phone : 886-2-7883799-{2407,1609}, fax : 886-2-7824814
e-mail : {jlshen,whwang}@iis.sinica.edu.tw

ABSTRACT

Although the delta and RASTA methods have been widely used in extracting the temporal properties of stationary features for robust speech recognition, there is still a need to investigate new temporal features for better performance. In this paper, we present two new temporal features for robust processing of speech signals with emphasis on microphone variations. First, the temporal feature is derived from a bank of RASTA-like filters, in where the parameters of each filter in this bank are estimated according to the statistical properties of the speech signals. Secondly, a parameterized temporal filter (called PTF) is proposed. The filter can be described by four parameters, the passband, the beginning transition, the ending transition, and the smoothness of the magnitude of the filter response. These parameters altogether determine the magnitude of the frequency response of the PTF, and a transformation algorithm is then used to derive the temporal coefficients with real and causal characteristics. The discriminative ability of PTF features can be further enhanced using the minimum classification error (MCE) algorithm. Experimental results show that the RASTA features is inferior to the PTF features both in quiet condition and in the presence of microphone variations.

1. Introduction

In recent years, robustness has become a crucial factor for the success of a speech recognition system [1]. A novel representation of speech signals in front-end processing would definitely be helpful for robust speech recognition [2-7]. It is well-known that recognition performance is improved if temporal variations of the stationary cepstrum, such as delta cepstrum and RASTA (RelAtive SpecTrAl) features, are incorporated into the recognition system [5][8]. Both delta cepstrum and RASTA features have been shown to be effective in channel noise reduction. Delta cepstrum, which approximates the first order temporal derivative of the cepstrum, is obtained from the output of a FIR band-pass filter (called delta filter). On the other hand, the RASTA method adds an extra pole to the delta filter based on the finding that the greater sensitivity of human hearing to the magnitude response of the temporal trajectories of frequency bands (also called modulation frequency) is at around 4Hz [9]. RASTA processing has several drawbacks:

1. Its filter has an infinite number of coefficients. This degrades the recognition performance, especially in sub-unit based systems because the short-term characteristics of speech signals could be smeared due to the long processing window.
2. As compared to cepstral features, RASTA processing has better recognition rates in mismatched training and testing environments while its performance is significantly degraded in matched environments.
3. RASTA features have poorer compatibility with stationary features than do the delta features in both matched and mismatched environments (see Table 2).

Therefore, much freedom remains in designing new temporal features for robust speech recognition. After conducting several experiments searching for relatively important modulation frequency bands in recognition performance, we conclude that the following requirements must be met by a temporal filter: (1) the filter should focus on relatively important modulation frequencies of speech signals while suppressing relatively unimportant ones; (2) the filter should be a FIR filter or should have fast decay in the time domain; and (3) the discriminative ability of its temporal features could be enhanced for speech recognition.

With the aim of understanding the relative weights of the modulation frequencies, a statistical method was used in which the peaks and cut-off modulation frequencies from a large set of speech data were identified. Then, a bank of RASTA-like filters, which have formulas similar to that of a RASTA filter but are different in the pole locations and bandwidths, was used jointly in extracting temporal features. The pole of a RASTA-like filter in the bank was adjusted to match as closely as possible one of the peak modulation frequencies estimated by speech signals. According to our experiments, the derived temporal features were able to discriminate more speech phonetic characteristics than could delta or RASTA processing, with more than 40% error rate reduction in both matched and mismatched environments as compared to the results of RASTA processing. However, the improvement of the recognition rate came with an increment in the dimensions of the temporal features, which was proportional to the number of filters used in the processing. Apparently, a larger feature size led to a longer processing time. Furthermore, a RASTA-like filter is still an IIR filter, which takes a long-term window; therefore, there is a risk of the smearing of the short-term characteristic of speech.

Motivated by the need for a filter with (1) a small feature dimension for real time applications, (2) fast decay in the time domain, and (3) covering most of the modulation frequencies that are sensitive to speech signals, we propose a parameterized temporal filter (called PTF hereafter). The filter is characterized by four parameters, the duration of the passband, the beginning transition, the ending transition, and the smoothness of the magnitude of the filter response. These parameters are empirically estimated to improve speech recognition performance. Our PTF has a passband covering most of the modulation frequencies that are sensitive to speech signals. The temporal decay of the PTF is determined by the smoothness at the beginning and the ending transitions. Finally, the filter’s coefficients are further refined to enhance the ability to discriminate models by minimizing recognition errors using the minimum classification error (MCE) algorithm [7]. Experimental results show that the new feature can produce nearly 20% of error rate reduction in comparison with a RASTA feature of the same dimension.

This paper is organized into 5 sections. The speech database and some initial processing are presented in section 2. Section 3 describes temporal feature extraction from a RASTA-like filter bank. In section 4, the parameterized temporal filter (PTF) is described. Also, we show how the discriminative capability of the temporal filter is further enhanced with the minimum classification error (MCE) algorithm. Finally, section 5 gives the concluding remarks.

2. Speech Database and Initial Processing

In this study, recognition of Mandarin base syllables in quiet condition and in the presence of microphone variations was studied. Mandarin Chinese is a monosyllabic structured tonal language. The total number of Mandarin syllables is 1345. Each Mandarin syllable is associated with a tone, and there exist 4 lexical tones and 1 neutral tone. If the differences in tones are disregarded, these 1345 Mandarin syllables are reduced to the 408 different base syllables [10]. Conventionally, the Mandarin syllables are composed of 22 INITIAL's and 41 FINAL's where INITIAL means the initial consonant while FINAL means the vowel part, including possible medial and nasal endings. It is believed that accurate recognition of all Mandarin base syllables is the key problem for large-vocabulary Mandarin speech recognition [10].

The speech database used in the following experiments was produced by 3 speakers. For each speaker, 4 collections of all 1345 Mandarin syllables using two types of microphones, C410 and D3700, were produced, respectively, where 3 collections of Mandarin syllables were used for training, and 1 collection was used for testing. Here, C410 is a close-talking and noise-canceling capacitor microphone with a flat frequency response while D3700 is a hand-held dynamic microphone. The averaged signal-to-noise (SNR) of the recording data obtained using these two microphones was 45dB and 35dB, respectively. In this paper, matched condition means that the training and testing data were both produced by microphone C410 while in mismatched condition, the testing data were recorded by microphone D3700. The left-to-right continuous density hidden Markov Model (CHMM) was used for each INITIAL/FINAL unit. All of the speech data were obtained in an office-like laboratory environment. They were low-pass filtered and digitized by an Ariel S-32C DSP board with sampling frequency 16kHz. In the feature extraction process, after end-point detection was performed, a 20 ms hamming window was applied every 8 ms with a pre-emphasis factor of 0.95. The temporal feature extraction process will be discussed in the following.

2.1. Temporal Feature Extraction

A block diagram of the temporal feature extraction process is shown in Fig. 1. The input speech is first processed by short-time Fourier transform (STFT), and then a mel-spectrogram is obtained by filtering the corresponding power spectrum of each frame with a set of 30 triangular band-pass filters, spaced uniformly on a mel-frequency scale. The well-known mel-frequency cepstral coefficient (MFCC) is obtained by processing the logarithm of the derived mel-spectrogram by means of discrete cosine transform (DCT). The temporal features are then obtained by filtering the temporal trajectories of the MFCC features. We use *H(z)* to denote the z-transform of the temporal filter. In fact, the DCT and the temporal filter *H(z),* as shown in the last two blocks of Fig. 1, process the frequency domain and time domain independently. We thus obtain the same temporal features whether we first apply DCT and then *H(Z),* or first apply *H(Z)* and then DCT. In Table 1, we summarize various types of temporal features with respect to *H(z*). One can see that the RASTA filter is derived from the delta filter with an extra pole, i.e., $1/1-0.98z^{-1}$. Fig. 2 shows the overlays of the corresponding frequency and temporal responses of the RASTA and delta filters. A common point for RASTA and delta filters is the spectral zero at the zero modulation frequency, which filters out channel noise. Howerer, it can be noted that the peak frequency in the RASTA filter is located at a lower frequency due to the additional pole, which supports the finding in [6]. Also, the higher modulation frequencies of the RASTA filter are further suppressed as compared to those of the delta filter. Although the RASTA filter provides better performance in both matched and mismatched conditions than can the delta filter, the relatively long-term processing window very often smears the short-term phonetic characteristics of speech signals. Finally, another well-known temporal filter is the cepstral mean subtraction (CMS) process, which subtracts the means of the long term behavior of the mel-frequency band spectra.

3. Temporal Feature Extraction from the RASTA-like Filter Bank

3.1. The Modulation Frequency Spectrum of Speech Signals

As mentioned previously, the RASTA filter was developed to take advantage of the sensitivity of human hearing to a modulation frequency at around *4* Hz [5]. In this section, we describe an alternative approach to temporal filter design. We will not confine ourselves to one temporal filter in feature extraction. Instead, our temporal features are extracted from a bank of filters, with each filter keeping the simplicity of the RASTA formula but adjusting its pole location to a “relatively important” modulation frequency band found by analyzing the statistical properties of the modulation frequency spectrum of speech signals.

We regard modulation frequencies with magnitudes corresponding to local maxima as “relatively important” ones. Since these modulation frequencies capture more dynamic variations of speech signals than can their neighborhoods, we believe that they should be emphasized in speech recognition. On the other hand, because it is impossible to develop methods for the compensation of every types of noise, emphasis on the important parts of speech signals is highly desired for robust speech recognition. In the following, we will describe our method, which finds “relatively important” modulation frequencies:

(1) We compute the averaged magnitude response of the modulation frequency *m* at the mel-frequency *f* for all the Mandarin syllables, S, that is:

$$Mel(f,m)=\frac{\mathbf{1}}{|S|}\sum\nolimits_{v\in S}\mathrm{E}\{Mag(f,m,v)\},$$

where the expectation E is applied to the training set of syllable *v*, and Mag is the magnitude of the modulation frequency *m* of syllable *v* at the mel-frequency *f*. In Fig. *3*, *4* out of all the *30 Mel(f, m)* are plotted. It is noteworthy that similar shapes are obtained for all the mel-frequency bands.

(2) Since the shape of *Mel(f, m)* is similar to those of all the mel-frequency bands, we can sum over the mel-frequency *f,* and obtain:

$$Mel(m) = \sum_{f} Mel(f, m).$$

The “relatively important” modulation frequencies are selected from the local maxima of *Mel(m).* Fig. *3* indicates two spectral peaks at around *0* Hz and *6* Hz, respectively. In fact, if only a limited number of syllables is used for analysis, more than two peaks are obtained. Also, the magnitudes at lower modulation frequencies are much larger than those at higher modulation frequencies. The cut-off speech modulation frequency is about *20* Hz. Although the features obtained by RASTA and delta filters are approximately consistent with our findings as previously described, it is significant that the appearance of multiple peaks in the magnitudes of the histogram of the speech modulation frequency is not fully explored in delta-MFCC or RASTA-MFCC features.

Similar experiments were conducted on noisy speech signals with added white noise up to *20dB* and with a convolution microphone noise, respectively. Let *N(m)* represent the resultant magnitude response of noisy signals and *S(m)* represent the speech signal. Fig. *4* plots the graph of *(|N(m) - S(m)|) / S(m)* versus *m*. It is clear that the largest variations appear at zero modulation frequency in the microphone as well as in added noise variations. Furthermore, except in the region enclosing the zero modulation frequency, the modulation frequencies with strong responses to speech signals approximately corresponds to those frequencies which are more insensitive to environmental variations (enclosed by dashed lines in Fig. *4*). Three regions are indicated in Fig. *4,* and the low modulation frequency region corresponds to the bandwidth at which most of the speech properties are included but also are most likely to be contaminated by environmental variations. The middle modulation frequency region (enclosed by dashed lines) has fewer speech signals properties, but is more robust to environmental variations. Finally, the high modulation frequency region, which contains the fewest speech properties but is easily influenced by noises, is the least desired region for speech recognition. Accordingly, a temporal filter can be developed to emphasize two statistically important components, that is, the low and middle modulation frequency regions, of speech signals to further improve recognition performance. Our approach is to develop a bank of RASTA-like filters which adapt to the multiple peaks of speech signals for temporal feature extraction.

3.2. The Design of the RASTA-like Filter Bank

We require that the RASTA-like filter preserve the simplicity of the RASTA formula but with free parameters in selecting the pole location as well as the cut-off modulation frequency. We can then adjust the pole location to a local maximum, if possible, in the modulation frequency region to be emphasized, and adjust the cut-off modulation frequency to de-emphasize the high modulation frequency region.

The RASTA-like filter is expressed as follows:

$$H(z)=\frac{\sum_{k=-N}^{N} kz^{N+k}}{\sum_{k=-N}^{N} k^{2}(1-\alpha z^{-1})}, \qquad (1)$$

where the parameter N determines the cut-off frequency while the parameter α determines the pole location. The pole location is adjusted to match as closely as possible a local maximum as shown in Fig. 3. At our setting, two RASTA-like filters with the parameter pair (N,α), equivalent to (3, 0.98) and (2, 0.8), are employed, respectively, for temporal feature extraction. For simplicity, we will call the derived feature RASTA-FB-MFCC hereafter. The dimension of the temporal feature for each filter was chosen to be 14. Thus, the RASTA-FB-MFCC has a total dimension of 28. Table 2 gives the recognition performances for Mandarin syllables for different types of features and their combinations. One can see from the table that all the temporal features (delta-MFCC, RASTA-MFCC, and RASTA-FB-MFCC) outperform the stationary feature (MFCC) in mismatched conditions. The recognition rate of the temporal features in both matched and mismatched conditions could be further improved by adding the stationary feature. It is worth mentioning that the combination of MFCC with delta-MFCC produces better performance in all conditions than did MFCC plus RASTA-MFCC. Finally, as shown in the table, RASTA-FB-MFCC had a better performance than did all the other temporal features. In fact, it produced the best performance in mismatched conditions. As compared to the results of MFCC, the averaged error rate was reduced by 59.01% in mismatched conditions, and the recognition rate was increased by 2.9% in matched conditions. In addition, in comparison with the results of MFCC plus delta-MFCC, the recognition rates were reduced by 1.85% in matched conditions but increased by 3.13% in mismatched conditions.

Although the best performance in mismatched conditions was achieved by the proposed RASTA-FB-MFCC, it has double the feature dimension compared to the other features. This increased the processing time and should be prevented. In the next section, we will introduce the other new temporal feature.

4. Feature Extraction from Parameterized Temporal Filter (PTF)

Several issues are involved in designing a temporal feature. First, the dimension of the feature has to be small for real time applications. In other words, the time response of the corresponding temporal filter should have fast decay. Secondly, the temporal filter has to be parameterized in order for it to be practically useful in implementation. One of the most difficult tasks is to determine the right parametric form to achieve optimal performance. Finally, the unknown parameters which defines the filter need to be estimated. In this section, we will describe an approach that uses parameters to develope the magnitude response of a temporal filter, the performance of which is then determined by these parameters. We call our temporal filter, for the sake of simplicity, a parameterized temporal filter (PTF).

This section is organized in two parts. In the first part, we will discuss the parameters used by PTF. These parameters determine the magnitude of PTF in the modulation frequency domain. In the second part, we will construct a real causal filter with knowledge of only the magnitude of modulation frequency responses. Then, the performance with respect to the newly derived PTF features will also be evaluated in the second section.

4.1 The Parameters of PTF

PTF is composed of 4 parameters, the beginning transition band (A), the passband (B), the ending transition band (C), and a smooth parameter (*n*). Although lower and middle modulation frequencies are important in speech recognition, it is very important to understand the contribution of recognition accuracy versus that of the coverage of modulation frequency bands. In PTF, three of the four parameters are used for this purpose. Section B in Fig. 5 denotes the passband while sections A and C denote the beginning and the ending transitions of the modulation frequency response of the filter, respectively. Apparently, section A must remove the zero modulation frequency due to its sensitivity to noises, and the region has to be as short as possible in order to allow section B to cover most of the low modulation frequencies. Section B represents the passband of the modulation frequency. This section should cover at least the region enclosed by the dashed line in Fig. 4. Finally, section C controls the cut-off modulation frequency. We also require that PTF be a real causal filter with fast decay in the temporal domain. A relatively long temporal filter tends to degrade performance in sub-unit based systems and smears the two adjacent units in the continuous speech signals. In order to have sufficient decay in the time domain, we require a certain degree of smoothness of the filter in the modulation frequency domain.

The smoothness of a PTF in the modulation frequency domain can be achieved as follows. Let *r(m)* and *s(m)* be the transition functions from region A to B and B to C, respectively. The function *r(m)* is an increasing function while *s(m)* is a decreasing function. The smoothness at regions A and C is determined entirely by the regularity of the functions *r(m)* and *s(m)*, respectively. According to [14][15], the smoothness of *r(m)* and *s(m)* is decided by real-valued continuously differentiable functions:

$$r_0(m) = s_0(m) = \sin(\frac{\pi}{2} m);$$

$$\text{region A} \quad r_{n+1}(\text{m}) = r_n(0.5(\sin[\pi(\frac{m}{|A|}\text{-}0.5)]+1)), \tag{3}$$

$$\text{region C} \quad s_{n+1}(\text{m}) = s_n(0.5(\cos[\pi(\frac{m-\beta}{|C|})]+1),$$

where β denotes the beginning point of region C while *n* determines the smoothness of the two functions. One can show that $\text{r(m)} = r_{\text{n}}(m)$, $\text{s(m)} = s_n(m) \in C^{2^n-1}$.

4.2 The Design of PTF

In the previous section, we gave the parameters of a PTF. After selection of the parameters, we have only the magnitude responses of the PTF. The phase of the filter has not yet been defined. For speech signal processing, we like our temporal filter to be a real and causal filter. Some constraints must be imposed on the phase of the PTF to obtain real causal coefficients. It was shown that a real and causal filter can be designed based on knowledge of only the magnitude of the filter response [13]. The theorem of Benedetto and Teolist is rephrased as follows :

Theorem. Let M(*m*) be non-zero positive values, and let $\Im$ denote the Hilbert transform. If

$$\int \frac{|\log M(m)|}{1+m^2}\,dm < \infty ,$$

then $H(m) = M(m)e^{-i\Im[\log(M(m))]}$ is the response of a real and causal filter.

This theorem shows that by carefully selecting the phase term, a real and causal filter can be constructed from its magnitude. Then, the PTF *h(t)* can be obtained by taking the inverse Fourier transform (*IFFT*) of *H*(*m*):

$$h(t) = IFFT(M(w)e^{-i\Im[\log(M(w))]}) . \tag{4}$$

There is one implementation issue of concern: Since a PTF is a band-pass filter, its response at zero modulation frequency is zero. This causes a problem in computing log (M(0)). We thus assign a small value for zero, say 10^{-5}.

Table 3 shows the recognition rates versus the different combinations of regions A, B and C. To understand the importance of the duration of each region in recognition performance, one region was changed each time while the other two were fixed. In all the experiments, we kept the smoothness parameter *n* at 2. In the first set of experiments, we tried to discover the low modulation frequency bands for the recognition rate. We set the duration of A, |A|, to 0 point, and fixed |C|, the duration of C, to 40 points. It can be found that the best performance was obtained when the duration of B, |B|, equaled 20 points. In the second set of experiments, we varied the duration of C while preserving that of the other two parts with |A|=0 point and |B|=20 points. Our results indicated that the best recognition rates occurred when |C| range from 50 to 60 points. In the last set of experiments, we varied the duration of A while fixing the durations of B and C to 20 and 60 points, respectively. The best performance in mismatched conditions occurred when A contained only 1 point. That is, only the DC component of the filter response was not needed. As the duration of the A increased, the recognition rates decreased in mismatched conditions. In matched conditions, the best performance occurred when A had 2 points. Based on the three sets of experiments, we thus chose the parameters of PTF with |A|=1, |B|=20, |C|=30, and *n* = 2. In Fig. 7 , we compare the temporal responses of the RASTA filter with those of our PTF, with the parameter set (|A|,|B|,|C|,*n*) = (1, 20, 30, 2). It is obvious from the bottom of the Fig. that the PTF had faster decay than did the RASTA filter. This makes it feasible to adopt our PTF features in a sub-unit system.

4.2. Discriminative Temporal Feature Based on MCE Algorithm

Speech recognition can be further improved with the aid of a minimum classification error (MCE) algorithm [11][16][17]. The MCE algorithm allows us to adjust the coefficients of the temporal filter such that the discriminative capabilities of the models can be enhanced. As a consequence, the recognition rate will be improved.

Let y represent the temporal features $\{y_1 ,y_2 ,\ldots ,y_T\}$, where y_t is the t-th temporal feature, and M is the collection of models, given by $\{\lambda_1,\ \lambda_2,\ \cdots\ \}$. Following conventional methods of speech processing, λ_i is modeled as a multi-valued Gaussian distribution with mean and variance, given by $[\mu_{i,\mathbf{1}},\mu_{i,\mathbf{2}},\cdots,\mu_{i,T}]$ and $[\Sigma_{i,\mathbf{1}},\Sigma_{i,\mathbf{2}},\cdots,\Sigma_{i,T}]$, respectively. The loss function in MCE algorithm is usually defined as a sigmoid function, namely:

$$l(d(y,M)) = \frac{1}{1+e^{-\beta d(y,M)}}, \qquad (6)$$

where $d(y,M)$ denotes the misclassification measure of the observed features *y* in the acoustic model M, and β is a pre-determined constant. Let λ_c be the “correct” model used in recognizing *y*; the misclassification which measures the distance from *y* to the acoustic model M can be defined as:

$$d(y,M) = -\log p(y,\lambda_c) + \log\{\tfrac{1}{|M|-1}\sum_{j,j\neq c} p(y,\lambda_j)^{\eta}\}^{1/\eta}, \qquad (7)$$

where λ_j is the model different from λ_c, and η is a positive number. When $d(y,M)$ is positive, this implies misclassification; otherwise means correct decision. For simplicity, we usually assume that $\eta \rightarrow \infty$; then, $d(y,M)$ has the following simpler form :

$$d(y,M) = -\log(p(y,\lambda_c)) + \log(p(y,\lambda_p)), \qquad (8)$$

where λ_p designates the most misclassified model of *y*. From now on, we will assume that the probability function given an observed y in model λ_i is a disjoint normal distribution, namely:

$$P(y,\lambda_i) = \prod_{j=1}^{T} A_i e^{-(y_i-\mu_{i,j})^T \Sigma_{i,j}^{-1}(y_i-\mu_{i,j})} = \prod_{j=1}^{T} N(\mu_{\mathrm{i,j}},\Sigma_{i,j}), \qquad (9)$$

where $N(\mu_{i,j},\Sigma_{i,j})$ represents the Gaussian normal distribution with mean $\mu_{i,j}$ and variance $\Sigma_{i,j}$ for model λ_i. The *t*-th temporal feature y_t is obtained by convolving the stationary features $\{x_t \mid t = 0,1,2\ldots\}$ with the temporal filter *H*, whose coefficients are designated as $\{\alpha_k \mid k = 0,1,,,L\}$, where *L* is the window size of temporal filter *H* :

$$y_t = \sum_{k=0}^{L} \alpha_k x_{t-k} \ . \qquad (10)$$

The coefficients of filter *H* are then adjusted according to the following gradient descent scheme:

$$\alpha_k^{n+1} = \alpha_k^n - \varepsilon(n)\frac{\partial l(d(y,M^n))}{\partial \alpha_k^n} \text{ with k = 0,1, ..,L,} \qquad (11)$$

where ε(n) is the step size satisfying the requirements given in [17] for the above equation to converge, and *n* is the iteration index. It is not difficult to take the partial derivative of $l(d(y,M^n))$:

$$\frac{\partial \ell\left(d\left(y,M^{n}\right)\right)}{\partial \alpha_{\mathrm{k}}^{n}}=\frac{\partial \ell\left(d\left(y,M^{n}\right)\right)}{\partial d\left(y,M^{n}\right)}\cdot\frac{\partial d\left(y,M^{n}\right)}{\partial \alpha_{\mathrm{k}}^{n}}=\ell\left(d\left(y,M^{n}\right)\right)\left(\mathbf{1}-\ell\left(d\left(y,M^{n}\right)\right)\right)\frac{\partial d\left(y,M^{n}\right)}{\partial \alpha_{\mathrm{k}}^{\mathrm{n}}}$$

with

$$\frac{\partial\ d(y,M^{n})}{\partial \alpha_{\mathrm{k}}^{\mathrm{n}}}=\sum_{t}[\gamma_{p,t}^{n}(y_{t}-\mu_{p,t}^{n})\left(\Sigma_{p,t}^{n}\right)^{-\mathbf{1}}-\gamma_{c,t}^{n}\left(y_{t}-\mu_{c,t}^{n}\right)\left(\Sigma_{c,t}^{n}\right)^{-\mathbf{1}}]x_{t-k}\ .$$

$\gamma_{c,t}^{n}$ and $\gamma_{p,t}^{n}$ denote the weighting factors at iteration *n* for the *t*-th feature y_t in models λ_c and $\lambda_{p,}$ respectively. Furthermore, not only are the filter's coefficients adjusted, but the model's parameters, $[\mu_{i,\mathbf{1}},\mu_{i,\mathbf{2}},\cdots,\mu_{i,T}]$ and $[\Sigma_{i,\mathbf{1}},\Sigma_{i,\mathbf{2}},\cdots,\Sigma_{i,T}]$, can also be adjusted accordingly.

In Fig. 7, we show two implementations of MCE algorithms. In the first implementation (indicated by the solid lines in the figure), the filter's coefficients are updated whenever a new observation *y* is obtained. Then, after several observations, the model's parameter is retrained with the maximal likelihood method [18] from the newly derived filter coefficients. After that, the model is used to re-estimate the filter's coefficients. These two phases are iteratively run until convergence is reached. In the second implementation, the model and filter parameters are adjusted jointly at each iteration. As in most of the numerical methods, the performance of the MCE algorithm depends on the initial parameters; in our case, this is the coefficients of the temporal filter. In the following, we will present the performance of RASTA and PTF features with/without use of MCE algorithms.

The recognition rates in both matched and mismatched conditions versus various temporal features with/without enhancement by means of MCE algorithm are listed in Table 4. The first set of data, corresponding to the top three rows, is the results for the PTF features while the second set, containing the bottom three rows, is for the RASTA features. The first row in each set is the result obtained without using the MCE algorithm. The numbers enclosed in parentheses in the last two rows in each set correspond to the method of implementation of the MCE algorithm, given in the previous paragraph.

It is obvious that the best performance occurred when the filter and model parameters were estimated jointly using the PTF features as the initial parameters, denoted as (PTF(2)). Compared to the results of the PTF features without applying the MCE, the error rates with the MCE were reduced by more than 7% and 10% in matched and mismatched conditions, respectively. Similarly, use of the MCE as post-processing in the RASTA filter also improved recognition performance. Comparing the performance of the PTF(2) and RASTA(2) features, the error rates of PTF(2) were reduced by 3.42% and 12.77% in matched and mismatched conditions, respectively. Comparing the results of PTF(2) to those of MFCC, approximately the same recognition accuracy was achieved in matched conditions while PFT(2) had an error rate reduction of 42.80% in mismatched conditions.

5. Conclusion

We have presented two new temporal features, RASTA-FB-MFCC and PTF, for robust processing of speech signals with emphasis on environments with microphone variations. Although our experimental data were obtained in the presence of the microphone variations, these two features can be extended appropriately to noisy environments with added noises. Experimental results show that these two features have better performance than do the well-known delta and RASTA features for convolution noises in both matched and mismatched conditions.

RASTA-FB-MFCC, which uses multiple filters to capture the dynamic speech characteristics, had the best performance among all the temporal features but at the expense of doubling the dimensions compared to the others. The PTF features outperformed the RASTA features of the same dimension and have several advantages: (1) the filter is designed based on a small set of parameters; (2) the PTF filter has faster decay in the time domain than does the RASTA filter. This makes the PTF features feasible for use in sub-unit systems. Furthermore, both the RASTA features and PTF features can be enhanced, and resulting in improvement of the recognition rate using the MCE algorithm.

Acknowledge: We thank Dr. Lee, Lin-Shan for many valuable discussions and suggestions.

References :

1. S. Furui, “Toward Robust Speech Recognition Under Adverse Conditions”, *Proceedings of the ESCA Workshop in Speech Processing*, pp. 31-42, 1992.
2. O. Ghitza, "Auditory Nerve Representation as a Front-end for Speech Recognition in a Noisy Environment", *Computer Speech and language*, 1 (2); 109-130, Dec. 1986.
3. D. Mansour and B.H. Juang, "The Short-time Modified Coherence Representation and Noisy Speech Recognition", IEEE Trans. Acoust., Speech, Signal Processing, pp. 795-804, June 1989.
4. S. Seneff, “A Computational Model for the Peripheral Auditory System : Application to Speech Recognition Research”, *IEEE Int. Conf. Acoust., Speech, Signal Proc.*, pp. 1983-1986, 1986.
5. H. Hermansky and N. Morgan, "RASTA Processing of Speech", *IEEE. Trans. Speech and Audio Processing*, Vol.2, No.4, Oct. 1994.
6. S.B. Davis, P. Mermelstein, “Comparison of Parametric Representations for Mononsyllabic Word Recognition in Continuously Spoken Sentences”, *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-28, No. 4, pp. 357-366, 1980.
7. Jia-lin Shen, Wen L. Hwang and Lin-shan Lee, "Robust Speech Recognition Features Based on Temporal Trajectory Filtering of Frequency Band Spectrum", *Int. Conf. Spoken Lang. Proc.*, pp. 881-884, 1996.
8. S. Furui, "Cepstral Analysis Technique for Automatic Speaker Verification", *IEEE Trans. Acoust. Speech, Signal Processing*, ASSP-29, pp.254-272, Apr. 1981.
9. G. Green, "Temporal Aspects of Audition", Ph.D. Thesis, Oxford, 1976.
10. L.S. Lee, *et. al*, “Golden Mandarin (I) - A Real-time Mandarin Speech Dictation Machine for Chinese Language with Very Large Vocabulary”, *IEEE Trans. On Speech and Audio Processing*, Vol. 1, No. 2, pp. 158-179, Apr. 1993.
11. W. Chou, B. H. Juang, C. H. Lee, "Segmental GPD Training Of HMM Based Speech Recognizer", *IEEE Int. Conf. Acoust., Speech, Signal Proc.*, pp. 473-476, 1992.
12. B.A. Dautrich, L.R. Rabiner, "On the Effects of Varying Filter Bank Parameters on Isolated Word Recognition", *IEEE Trans. Acoustics, Speech, Signal Proc.*, ASSP-31 (4): 793-807, Aug. 1983.

13. John H. Benedetto and Anthony Teolist, "A Wavelet Model and Data Compression", *Applied and Computational Harmonic Analysis*, pp. 3-28, 1993.

14. H. Malvar, “Lapped Transforms for Efficient Transform/Subband Coding”, *IEEE Trans. Acoustics, Speech, Signal Proc*., ASSP-38: 969-978, 1990.

15. M. V. Wickerhauser, “Adapted Wavelet Analysis from Theory to Software”, A.K. Peters, 1994.

16. R. Chengalvarayan, L. Deng, “ Use of Generalized Dynamic Feature Parameters for Speech Recognition”, *IEEE Trans. On Speech and Audio Processing*, Vol. 5, No. 3, pp. 232-242, May 1997.

17. B.H. Juang, W. Chou, C.H. Lee, “Minimum Classification Error Rate Methods for Speech Recognition”, *IEEE Trans. On Speech and Audio Processing*, Vol. 5, No. 3, pp. 257-265, May 1997.

18. L. Rabiner, and B. H. Juang, ”Fundamentals of Speech Recognition”, Prentice-Hall Intermational, Inc., 1993.

Feature	*H(z)*
MFCC	1
RASTA-MFCC	$\frac{0.1z^{4}(2+z^{-1}-z^{-3}-2z^{-4})}{1-0.98z^{-1}}$
delta-MFCC	$0.1z^{4}(2+z^{-1}-z^{-3}-2z^{-4})$
MFCC with CMS	high-pass filter

Table1. Summary of various types of features with respect to the temporal filter *H(z)*.

Feature (Dimension)	Matched (%)	Mismatched (%) (microphone variations)
MFCC (14)	87.29	70.63
delta-MFCC (14)	81.71	76.51
MFCC plus delta-MFCC (28)	92.04	84.83
RASTA-MFCC (28)	82.97	79.18
MFCC plus RASTA-MFCC (28)	90.41	83.42
RASTA-FB-MFCC (28)	90.19	87.96

Table 2. The recognition rates of different features in matched and mismatched conditions.

Feature		matched	mismatched (microphone variations)

A=0,
C=20
	B=0	83.72	71.97

	B=10	84.76	72.27
	B=20	**85.13**	**73.90**
	B=40	83.20	68.70

A=0,
B=20
	C=15	84.98	72.21

	C=25	86.39	73.68
	C=30	**87.43**	**72.57**
	C=35	87.58	72.04
	C=40	87.81	71.90

B=20,
C=30
	A=1	**86.32**	**81.26**

	A=2	86.47	80.45
	A=3	85.20	80.07
	A=4	83.05	75.84

Table 3. The recognition rates of PTF features from different values of A, B and C, with *n* fixed at *2*.

Feature	Matched	Mismatched (Microphone Variations)
PTF(A=1,B=20,C=30,n=2)	86.32	81.26
PTF(1)	87.14	80.15
PTF(2)	87.29	83.20
RASTA-MFCC	82.97	79.18
RASTA(1)	84.98	79.26
RASTA(2)	86.84	80.74

Table 4. The recognition rates for RASTA and PTF features with/without enhancement by means of the MCE algorithm.

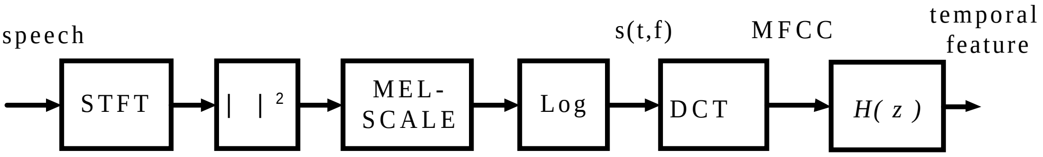

Figure1. A block diagram of the temporal feature extraction process. (DCT denotes discrete cosine transform, and *H(z)* denotes the filter used to process the temporal trajectories of MFCC.)

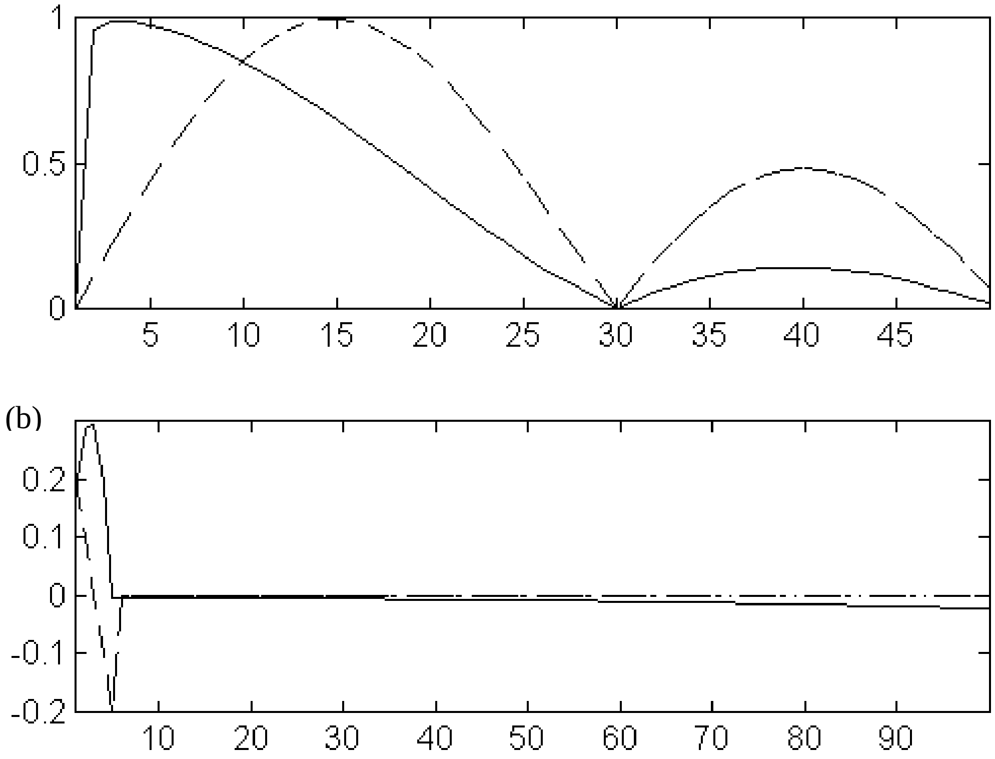

Figure 2. (a) The frequency responses. (b) The impulse responses for delta-MDCC (dashed line) and RASTA-MFCC (solid line).

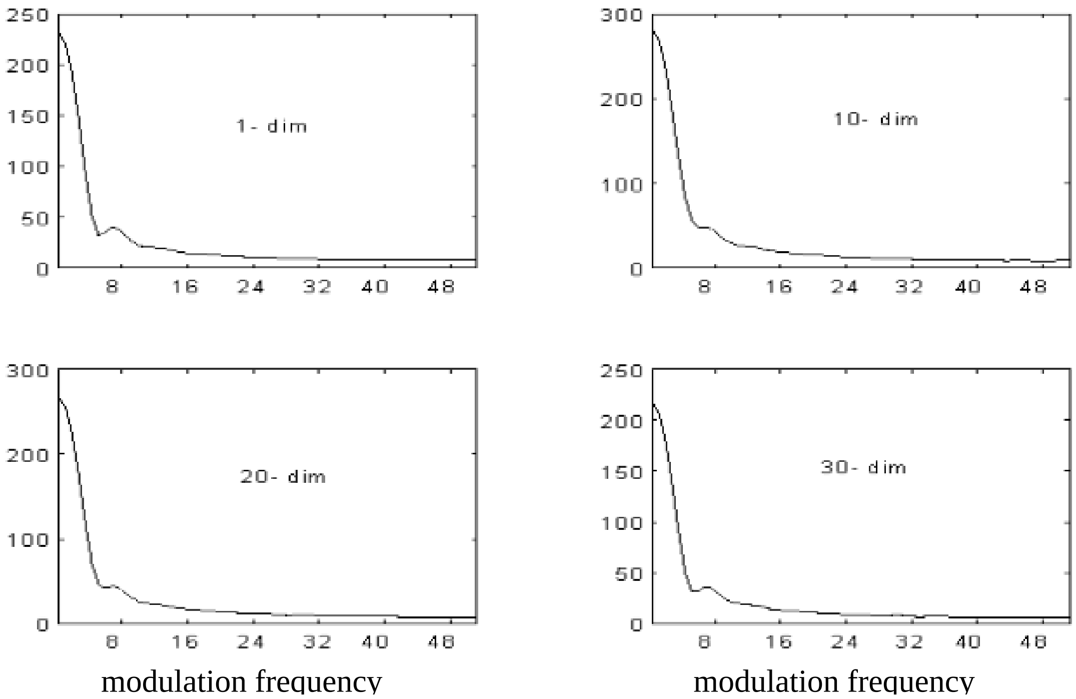

Figure 3. The averaged magnitude versus the modulation frequency for Mandarin syllables.

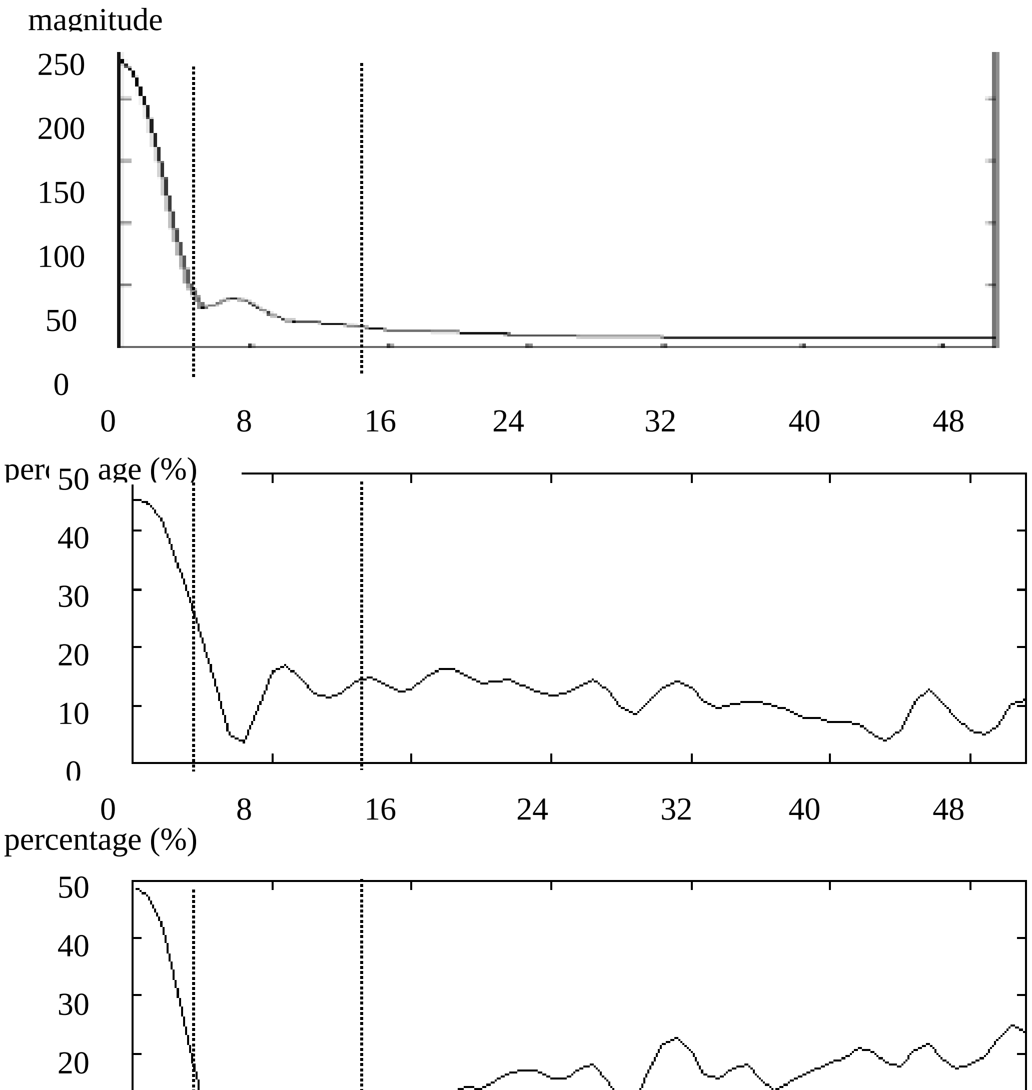

Figure 4. A plot of (|magnitude of noise - magnitude of clean | / magnitude of clean) versus the modulation frequency for Mandarin syllables with (a) 20 dB of white noise and (b) microphone variations.

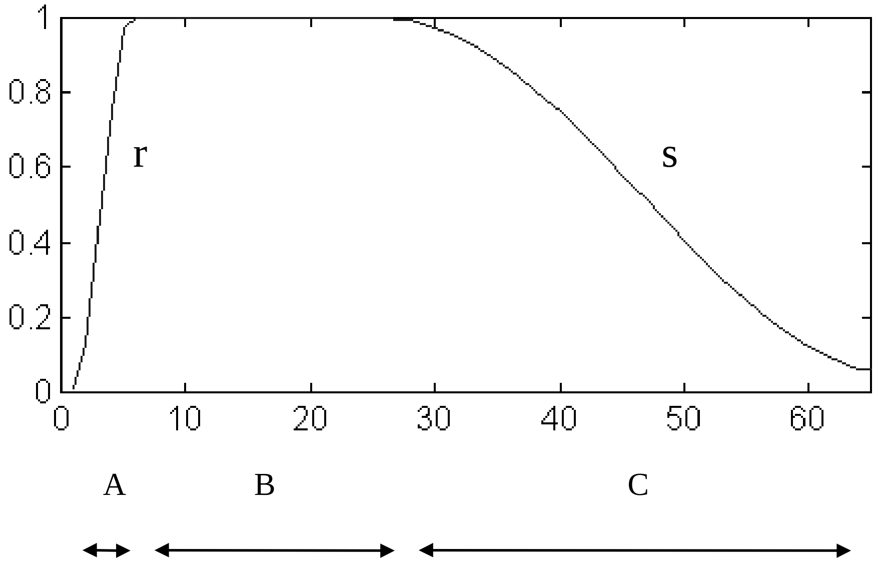

Figure 5. The designed envelope of the PTF filter.

(a)

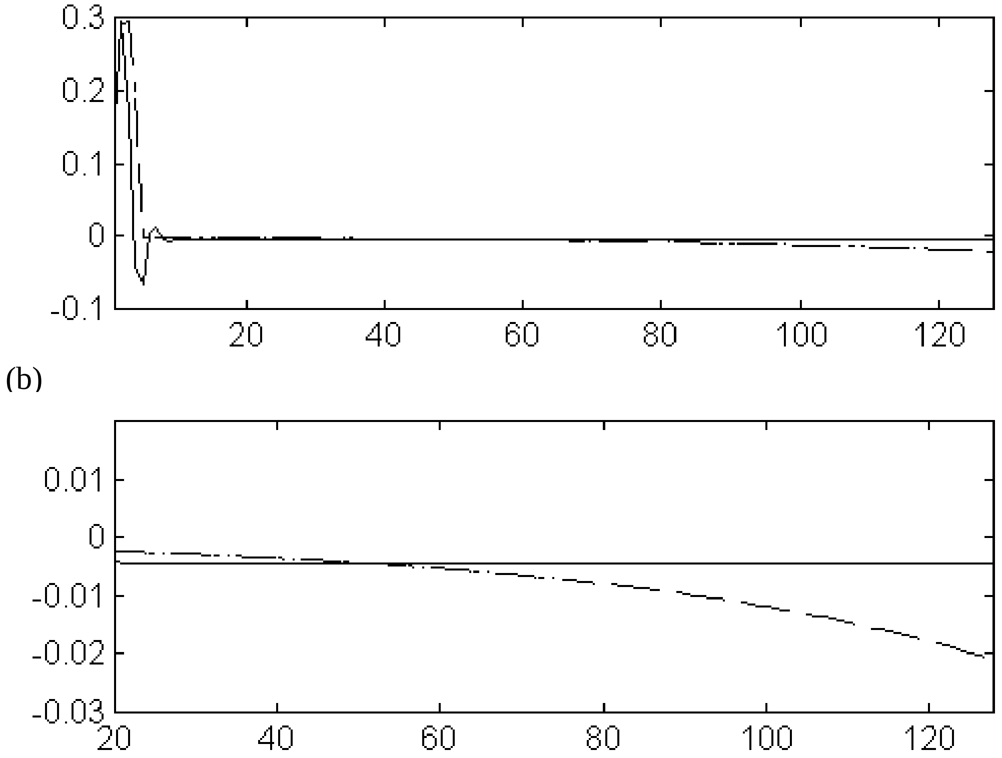

Figure 6. The frequency responses for the RASTA filter(dashed line) and the PTF filter(solid line). In the bottom plot, the tails of the two filters are highlighted.

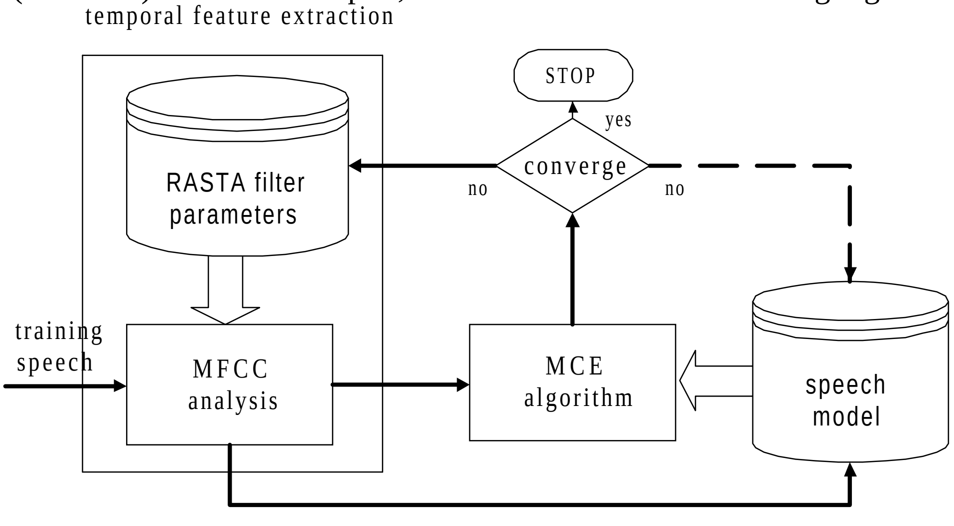

Figure 7. A block diagram of the discriminative temporal feature extraction process. (The dotted line is added for the second estimation method.)