Syllable intelligibility for temporally filtered LPC cepstral trajectories

Takayuki Arai
Sophia University, 7-1 Kioi-cho, Chiyoda-ku, Tokyo, Japan

Misha Pavel
AT&T Labs West, 75 Willow Road, Menlo Park, California 94025

Hynek Hermansky
Oregon Graduate Institute of Science & Technology, P.O. Box 91000, Portland, Oregon 97291-1000

Carlos Avendano
CIPIC, University of California, Davis, California 95616

(Received 19 December 1996; accepted for publication 29 January 1999)

The intelligibility of syllables whose cepstral trajectories were temporally filtered was measured. The speech signals were transformed to their LPC cepstral coefficients, and these coefficients were passed through different filters. These filtered trajectories were recombined with the residuals and the speech signal reconstructed. The intelligibility of the reconstructed speech segments was then measured in two perceptual experiments for Japanese syllables. The effect of various low-pass, high-pass, and bandpass filtering is reported, and the results summarized using a theoretical approach based on the independence of the contributions in different modulation bands. The overall results suggest that speech intelligibility is not severely impaired as long as the filtered spectral components have a rate of change between 1 and 16 Hz. © *1999 Acoustical Society of America.* [S0001-4966(99)01705-1]

PACS numbers: 43.66.Mk, 43.71.Es, 43.72.Ar [JH]

INTRODUCTION

One of the main objectives of front-end processing in robust automatic speech recognition (ASR) is to preserve critical linguistic information while suppressing such irrelevant information as speaker-specific characteristics, channel characteristics, and additive noise. The information suppressed in the front end of the recognizer is lost for the recognition process. To determine information to be preserved, we need to identify those features of the signal that are necessary for human speech recognition. One way to identify the useful features is to suppress a given feature, reconstruct the speech, and determine its intelligibility through perceptual experiments.

Temporal processing, or filtering the time trajectories, of the logarithmic spectrum or cepstral coefficients is becoming a common procedure in current ASR. One reason for this type of processing is that the convolutional distortion, such as the frequency characteristics of a communications channel, is an additive component in the logarithmic spectrum and the cepstrum domains. These channel characteristics are often fixed or only slowly varying in time. Therefore, cepstral mean subtraction (CMS) is often used to eliminate the channel characteristics by subtracting the mean (or dc component) of the time trajectory of each cepstral coefficient.[1] The delta features are calculated as linear regression coefficients over a short segment of a time trajectory to emphasize the dynamic characteristics of the original features.[2] This delta technique is effectively equivalent to applying a finite impulse response (FIR) bandpass filter which eliminates the dc component of the time trajectory and applies 6 dB/oct emphasis on changes up to approximately 12 Hz.

The RelAtive SpecTrAl (RASTA) technique suppresses the spectral components that change more slowly or quickly than the typical range of change of spectral envelope.[3] (Please note that throughout this paper, we address the rate of change of spectral envelope, i.e., the rate at which the source signal is being modulated, by changes in the vocal tract shape, not the rate of change of the signal itself.) The RASTA technique is implemented by the following steps: (1) compute the spectral amplitude, (2) transform the spectral amplitude through a compressing static nonlinear transformation, (3) filter the time trajectory of each transformed spectral component, (4) transform the filtered speech representation through an expanding static nonlinear transformation, and (5) perform optional processing. The logarithmic function is often used for the nonlinear transformation. RASTA processing also eliminates the dc component but, unlike the delta feature computation, it passes components between 1 and 12 Hz unattenuated. Both delta and RASTA techniques appear to achieve some degree of robustness to channel variations.

Thus, the front end suppresses some information from the speech signal by filtering the time trajectories of the cepstral coefficients. The relatively slow rates of cepstral change, or low-modulation frequencies, include such information as channel characteristics, speaker information, and voice quality, which are assumed not crucial for human speech communication. Similarly, the relatively fast rates of cepstral change, or high-modulation frequencies, might be less important for human speech communication.

To justify this approach, it is essential to identify the

contribution of different modulation frequency bands of cepstral coefficients to human speech recognition. In this paper, we used the LPC-based approach for at least two reasons. First, LPC is the most common technique in speech engineering; therefore, our results are directly applicable to many LPC-based ASR systems. Second, the results can enhance our understanding of the temporal properties of the speech signals. This is due to the fact that, at least in theory, the LPC analysis separates speech information into two components: the sound source and the vocal tract. Hence, the LPC technique allows us to manipulate these components independently and permits us to study the dynamics of each.

The goal of this study is to examine the effect of filtering the time trajectories of the spectral envelope on the intelligibility of the reconstructed speech.

Drullman[4,5] reported the effect of temporal filtering of the spectral envelope on the intelligibility of speech. In his study, the original speech was split into a series of frequency bands. The magnitude envelope of the analytic signal for each band was then low-pass and high-pass filtered. He concluded that low-pass filtering below 16 Hz or high-pass filtering above 4 Hz does not appreciably reduce speech intelligibility.

Drullman's results showed that the low- and high-modulation frequencies of the magnitude spectrum are not essential for the intelligibility of speech. These results are, in principle, consistent with RASTA processing.

In this paper, we will focus on the following question: ''How will speech intelligibility be affected if:

(1) filtering is done in the cepstral trajectories,
(2) the filters are bandpass filters, and
(3) the energy contour is unmodified?''

Prior work does not address these questions for the following reasons. First, Drullman applied the filtering to the magnitude envelope of the analytic signal, which effectively implies filtering of the magnitude spectrum of the speech. It is not obvious that Drullman's results generalize to other features, such as cepstrum, that are typically used for speech recognition. Filtering in a different domain might affect human speech perception differently. This is particularly true if the filtering is performed on a nonlinear transformation of the signal, e.g., the logarithmic function. In contrast to Drullman's study, we examined the effects of temporal filtering of the time trajectories of the LPC cepstrum. Thus, our results have direct implications for cepstrum-based ASR systems.

Second, it is not obvious whether his results for the low-pass and the high-pass experiments can be used to draw any conclusions about bandpass filtering because of a nonlinearity of the human auditory system. Therefore, we investigated empirically the effect of bandpass filtering.

Third, because the temporal change of the magnitude envelope was filtered in Drullman's experiment, the energy contour and the temporal change of the spectrum were both affected. Instead, we focus only on the modification of the spectral change; in our experiment, the energy contour of the modified signal is kept the same as that of the original signal.

In this paper, we first describe the signal processing based on a residual-excited LPC vocoder. This signal processing consists of the LPC cepstral analysis of the speech signal, temporal filtering of the cepstrum, and reconstruction of the speech. The stimuli for the perceptual experiments are described in Sec. II. In the same section, the experimental procedure, subjects, and data analysis are described. We then describe two experiments with low-pass and high-pass conditions (experiment 1), and bandpass conditions (experiment 2). Finally, we discuss further issues based on the experimental results, including the information for intelligibility using a density function over the modulation frequency (in Sec. IV).

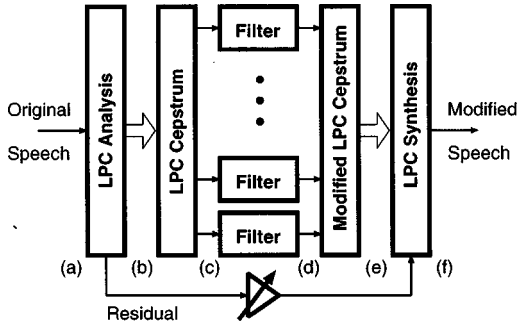

FIG. 1. Block diagram of the speech-processing system.

I. SIGNAL PROCESSING

An overview of our signal-processing method is illustrated in Fig. 1. It consists of applying a frame-by-frame LPC analysis to the original speech, then filtering the time trajectories of the resulting LPC cepstral coefficients. Subsequently, the modified speech signal is reconstructed by an LPC synthesis technique. The filters used in this study were either low-pass, high-pass, or bandpass, with different cutoff frequencies covering the frequency band of interest.

The signal-processing technique was based on a residual-excited LPC vocoder. This approach permits the construction of the entire continuum from the nonfiltered signal to the complete removal of all LPC information. In the range between those two extremes, we were able to examine speech intelligibility as a function of the frequency content of the temporal trajectories of the LPC cepstral coefficients.

Figure 2 shows an example of an utterance: (a) the original speech, and (b) its spectrogram, the time trajectory of the first LPC cepstral coefficient (c) before and (d) after filtering, and (f) the modified speech and (e) its spectrogram. In the original time trajectory, one can see very high modulation frequencies as well as the dc component [Fig. 2(c)]. After bandpass filtering between 1 and 16 Hz, fast and slow modulations are removed, but the major components remain [Fig. 2(d)]. A comparison of the spectrograms of Fig. 2(b) and (e) indicates that the major spectral transitions were preserved.

A. LPC cepstral representation

The speech signals were first analyzed by a 12th-order linear prediction technique, with pre-emphasis. The energy and the 12 LPC coefficients were calculated at each frame using the parameters shown in Table I. Following the LPC

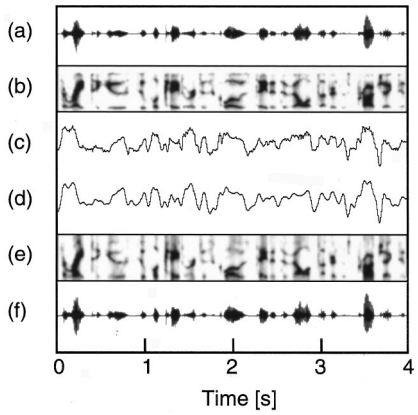

FIG. 2. Signals and spectrograms of an utterance. Each plot corresponds to the letter marked in Fig. 1. (a) Original speech. (b) LPC-based spectrogram of the original speech. (c) Time trajectory of the first LPC cepstral coefficient. (d) Bandpass filtered time trajectory of (c) with bandpass filter between 1 and 16 Hz. (e) LPC-based spectrogram of the modified speech. (f) Modified speech.

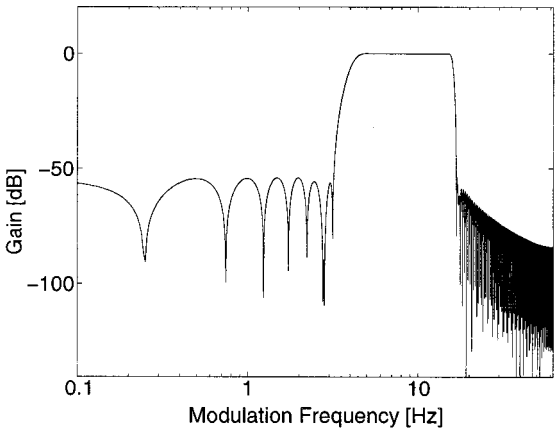

FIG. 3. The frequency characteristics of a bandpass filter designed by the windowing method. For this plot, the cutoff frequencies are 4 and 16 Hz.

analysis, the LPC coefficients were converted to cepstral coefficients. To achieve a logarithmic spectrum with a sufficiently high resolution, we calculated all cepstral coefficients up to quefrency of 16 ms.

B. Filtering of the cepstral coefficients

The time trajectory of each cepstral coefficient was processed by a temporal filter. The filters were identical at all quefrencies except that the coefficient at zero quefrency was discarded. The bandpass filters (BPFs) were implemented as 257-tap finite impulse response (FIR) filters with linear phase. Their coefficients were designed by the windowing method (Hamming window). For each filter, the slope within the transient band is approximately 48 dB/Hz. Figure 3 shows the magnitude frequency characteristics of a sample bandpass filter.

C. Reconstruction of the speech

The filtered LPC cepstral coefficients were used to compute the modified power spectrum at each frame. A 12th-order LPC filter was calculated from the autocorrelation function obtained by applying the inverse Fourier transform of the power spectrum. In the last stage of the signal processing, we reconstructed speech sounds using the modified LPC coefficients together with the residual signal.

Ideally, the residual signal would contain only the sound source information. In practice, however, the residual signal may also contain some information about the vocal tract shape, so the LPC residual sometimes yields a relatively intelligible signal. In the first half of this study (experiment 1), we further whitened the residual signal to reduce the intelligibility. The whitening was not as helpful as we expected, so we did not whiten the residual signal in the second half of this study (experiment 2).

Finally, to avoid measuring the effect of modifying the energy contour, we matched the total energy in each frame of the reconstructed speech to the energy in the related frame of the original speech. Thus, our reconstructed speech had the same energy contour as the original speech, but its spectral envelope structure was modified.

TABLE I. Conditions for LPC analysis.

Order of LPC analysis	12
Window	Hamming
Frame length	32 ms
Frame period	8 ms
Pre-emphasis	0.98

II. STIMULUS

A. Speech samples

The original speech sounds were obtained from a Japanese syllable database used for articulation tests at NTT Japan. To generate stimuli for this study, we selected the voice of a 24-year-old female. Each sentence contained a target Japanese syllable in the carrier phrase ''Kankonbai ——— oruso.'' The original speech signal was quantized with a 16-bit resolution and sampled at 48 kHz. Our stimuli were processed and presented at an 8-kHz sampling rate.

The original data set contained 100 Japanese syllables. We selected a subset of 31 syllables covering the three corner vowels /a/, /i/, and /u/, and Japanese consonants /p/, /b/, /t/, /d/, /k/, /g/, /s/, /ʃ/, /ts/, /tʃ/, /dz/, /dʒ/, /n/, /m/. The 31 syllables are shown in Table II. Each syllable is a vowel (V) or consonant–vowel (CV) syllable.

TABLE II. Japanese syllables used in this study.

	Unvoiced consonants			Voiced consonants		
Vowels	/a/	/i/	/u/			
Stops +Vowels	/pa/	/pi/	/pu/	/ba/	/bi/	/bu/
	/ta/			/da/		
	/ka/	/ki/	/ku/	/ga/	/gi/	/gu/
Fricatives +Vowels	/sa/	/ʃi/	/su/			
Affricates +Vowels		/tʃi/	/tsu/	/dza/	/dʒi/	/dzu/
Nasals +Vowels				/ma/	/mi/	/mu/
				/na/	/ni/	/nu/

B. Stimulus conditions

Stimuli were divided into conditions according to the amount of low-pass, high-pass, and bandpass filtering of the LPC cepstral coefficients.

For experiment 1, the time trajectories of the LPC cepstral coefficients were filtered with low-pass and high-pass filters with cutoff frequencies f_C, where $f_C = \{0,1,2,3,4,5,6,8,12,24,48,f_N\}$ [Hz], where f_N is equal to half of the frame rate, i.e., $f_N = 62.5$ Hz. A complete set of the 13 conditions (including clean speech) applied to all 31 syllables was presented to our subjects in a session consisting of 403 ($=13\times31$) stimuli.

For experiment 2, the time trajectories of the LPC cepstral coefficients were filtered with bandpass filters. The bandpass filters had lower cutoff frequencies f_L and upper cutoff frequencies f_U, where $f_L = \{0,1,2,4,8,16,32,f_N\}$ [Hz] and $f_U = \{0,1,2,4,8,16,32,f_N\}$ [Hz] ($f_L \leq f_U$). Note that when $f_L = 0$ the filter is a low-pass filter, and when $f_U = f_N$ the filter is a high-pass filter. A complete set of the 30 conditions (including clean speech) applied to all 31 syllables was presented to our subjects in a session consisting of 930 ($=30\times31$) stimuli.

C. Procedure

We used the method of constant stimuli, with stimuli presented in random order. Each subject participated in four sessions. Combinations of syllables and filtering conditions were randomized across sessions and subjects.

The stimuli were generated by the digital-to-analog (D/A) converter of a SPARC-20 workstation at 8-kHz sampling rate and presented using high-quality headphones (Sennheiser HD 250 II) at a comfortable listening level. On each trial, the subject heard an isolated syllable preceded and followed by 1-s intervals of silence. Following each stimulus presentation, subjects indicated their answer and then initiated the next trial. Each stimulus was presented only once.

Subjects interacted with the experimental setup using a graphical user interface and a mouse input device. As shown in Fig. 4, the monitor screen showed icons for all 31 possible stimuli, and subjects were asked to select the icon of the most likely stimulus. In addition to the stimulus icons, there were buttons to allow corrections and to indicate completion of trials.

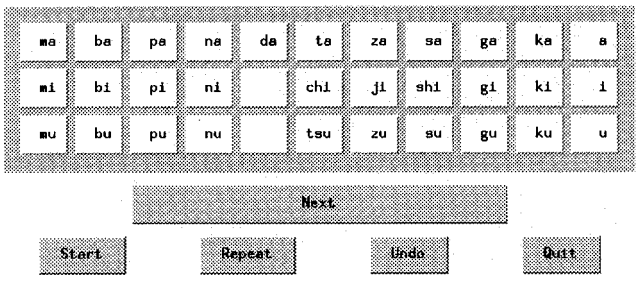

FIG. 4. Graphical user interface for the experiments.

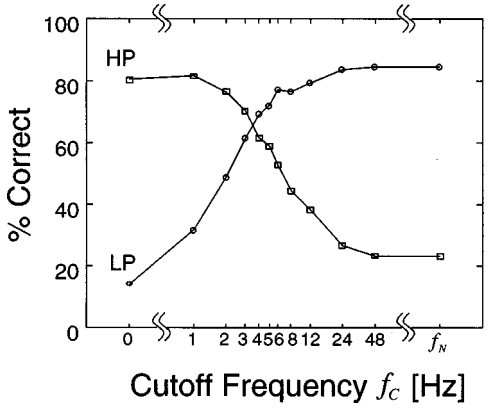

FIG. 5. Results for low-pass (LP) and high-pass (HP) filtering.

D. Subjects

A total of 20 Japanese native speakers participated in the study. The subjects were paid for their participation. Of the 20 subjects, 16 participated in experiment 1 and four in experiment 2. No subject reported having any previous hearing problem.

E. Data analysis

We summarized the data for each condition and each stimulus in terms of the proportion of correct responses to total responses. The proportion of correct CV syllables is defined as P_c(CV), and a response was scored as correct only if both constituents (phonemes)—vowel and consonant—were recognized correctly. The resulting overall proportions of correct responses were averaged over all stimuli for each condition. In addition, we analyzed the results for each phonetic category. The proportion correct for each category is denoted as P_c(category), e.g., P_c(C) for consonants and P_c(V) for vowels.

III. EXPERIMENTAL RESULTS

A. Experiment 1

Experiment 1 consisted of low-pass and high-pass filtering of the LPC cepstral trajectories. The overall summary of the results averaged over stimuli and subjects is shown in Fig. 5. The abscissa of the graph shows cutoff frequencies and the ordinate represents the proportion of correct responses P_c(CV) of each CV. Each point is an average of 31 stimuli, 8 subjects, and 4 sessions for a total of 992 ($=31\times8\times4$) trials. Assuming a binomial distribution of responses, the largest standard error of the estimates is less than 2%. The corresponding error bars were omitted for clarity.

The performance for the original speech averaged over the four subjects was 85.8% and ranged from 75.8% to 99.2%. The average score for the residual signal was 18.5% and ranged from 7.3% to 27.4%. The useful range of the information in the LPC cepstral trajectories was therefore 18.5%–85.8%.

In the low-pass condition, the performance begins to decrease gradually below 24 Hz. In the high-pass condition, the decrease in performance begins above 1 Hz. The low-pass

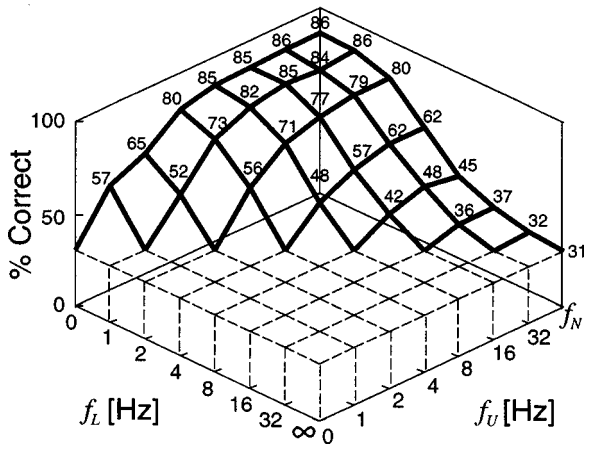

FIG. 6. Results on 31 syllables (full set) for bandpass filtering. The number at each node shows the proportion of correct responses.

and high-pass curves intersect between 3 and 4 Hz. This modulation frequency indicates the center of information, where the information is equally divided into lower and higher modulation bands. This modulation frequency is consistent with the average syllabic rate of speech.[6]

B. Experiment 2

Four native Japanese subjects participated in experiment 2. Each subject participated in four experimental sessions. Each session consisted of 930 stimuli.

The results for the bandpass condition are shown in Fig. 6. The vertical axis is the proportion of correct responses P_c(CV), while the other two axes are the cutoff frequencies, f_L and f_U. Each point is an average of 31 stimuli, 4 subjects, and 4 sessions for a total of 496 ($=31\times4\times4$) trials. The largest standard error of a binomial distribution with the same number of trials is less than 2%.

In experiment 2, the performance for the original speech averaged over four subjects was 86.1% and ranged from 84.7% to 89.5%. The average score for the residual signal was 30.6% and ranged from 17.7% to 41.9%. The useful range of the information in the LPC cepstral trajectories was therefore 30.6%–86.1%.

The data from this experiment are also consistent with those of the low-pass and high-pass conditions in experiment 1.

Table III is the confusion matrix for the original signal in experiment 2. As shown in this table, there are several syllables for which the original signal is hard to understand. This could be attributed to the low sound quality of the 8-kHz sampled stimuli. Misperception of the consonant of the syllables was common, while most of the vowels were perceived correctly. Figure 7 shows the proportion of correct responses for the 21 CV syllables which are perceived perfectly for the original signal (/ka/, /ki/, /ku/, /ga/, /gi/, /gu/, /sa/, /ʃi/, /su/, /dza/, /dzu/, /tʃi/, /tsu/, /da/, /na/, /ni/, /nu/, /pi/, /ba/, /bi/, /ma/).

As can be seen in Fig. 7 (subset) and Fig. 6 (full set), the global trends in both cases are the same. As in Fig. 6, the trend in Fig. 7 is not affected when the time trajectories have components between 1 and 16 Hz.

TABLE III. Reduced confusion matrix of responses for the original signal in the bandpass experiment. Syllables that were perceived perfectly were omitted from the matrix. The total number of responses for each syllable is 16.

	Response								
Stimulus	/u/	/ka/	/ku/	/gi/	/gu/	/ta/	/ni/	/nu/	Total
/dʒi/				3					3
/ta/		12							12
/pa/						7			7
/pu/			10		2				12
/bu/					6			2	8
/mi/							11		11
/mu/	2							12	14

C. Cue trading

To maintain the original phonetic information for human perception, a decrease in one feature can be offset by an increase in another cue; this tradeoff is known as cue trading.[7] To illustrate the phenomenon of cue trading, we projected Fig. 7 onto two different planes. The two planes are shown in Fig. 8: (a) the proportion of correct syllables P_c(CV) versus f_U, and (b) the proportion of correct syllables P_c(CV) vs f_L. In Fig. 8(a), the graphs of $f_L=0$ Hz and $f_L=1$ Hz match when $f_U>4$ Hz, while the graphs of $f_L=0$ Hz and $f_L=1$ Hz do not match when $f_U\leqslant4$ Hz. That is, if we have components at 4 Hz and above, then we can compensate for the lost cues below 1 Hz, but if we lose the components at 4 Hz and above, then we cannot compensate for those lost cues. Similarly, in Fig. 8(b) the graphs of $f_U=16$ Hz and $f_U=f_N$ match when $f_L<4$ Hz, while the graphs of $f_U=16$ Hz and $f_U=f_N$ do not match when $f_L\geqslant4$ Hz. That is, if we have the components at 4 Hz and below, then we can compensate for the lost cues above 16 Hz, but if we lose the components at 4 Hz and below, then we cannot. This suggests that the component at 4 Hz is necessary for cue trading. A modulation frequency of 4 Hz corresponds to the average syllabic rate of speech.[6]

Figure 9 shows the results in terms of proportion of

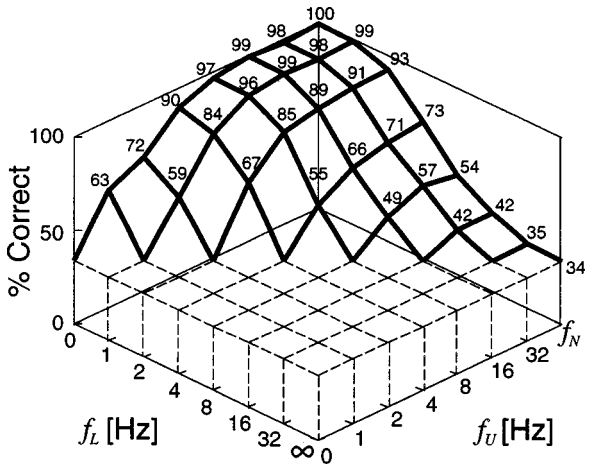

FIG. 7. Results on 21 CV syllables (subset) for bandpass filtering. The number at each node shows the proportion of correct responses.

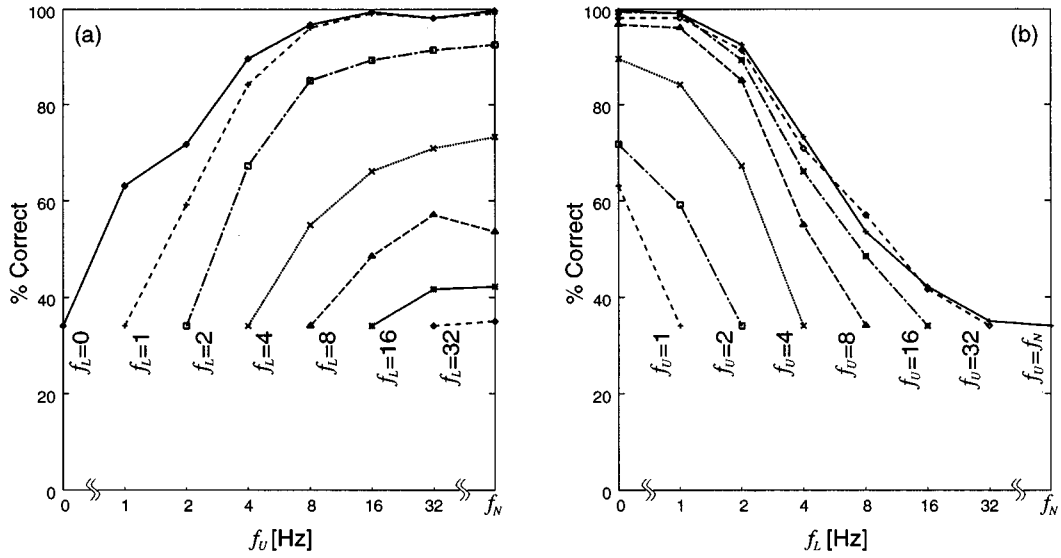

FIG. 8. Projection of Fig. 6 onto two different planes. (a) A plot of the proportion correct versus f_U. (b) A plot of the proportion correct versus f_L.

correct CV syllables P_c(CV), consonants P_c(C), and vowels P_c(V). As can be seen in Fig. 9(a), P_c(C) is sensitive to higher-modulation frequency components and is lower than P_c(V) when $f_U<16$ Hz. On the other hand, as shown in Fig. 9(b), P_c(V) is sensitive to lower-modulation frequency components and is lower than P_c(C) when $f_L>1$ Hz.

As can be seen in Fig. 9, P_c(CV) is larger than the product of P_c(C) and P_c(V). Fletcher showed that the articulation probability of a CV syllable will be the product of the articulation probabilities of the C and the V.[8] The results in our domain, however, suggest that consonants and vowels do not contribute to intelligibility independently.

The results for each consonant category were analyzed as shown in Fig. 10. The proportion of correct responses for each category drops when $f_U<16$ Hz, and P_c(stops) and P_c(nasals) are sensitive to higher-modulation frequency components [Fig. 10(a)]. As shown in Fig. 10(b), the proportion of correct responses for each category drops when $f_L>2$ Hz, and P_c(fricatives) and P_c(affricates) are sensitive to lower-modulation frequency components. We observed that the modulation frequency component at 4 Hz is essential for the sounds having longer duration, such as fricatives, and that the much higher-modulation frequency components are essential for the sounds having shorter duration, such as stops.

IV. DISCUSSION

In Drullman's experiment, the temporal change of the magnitude envelope was filtered. As a result, both the energy

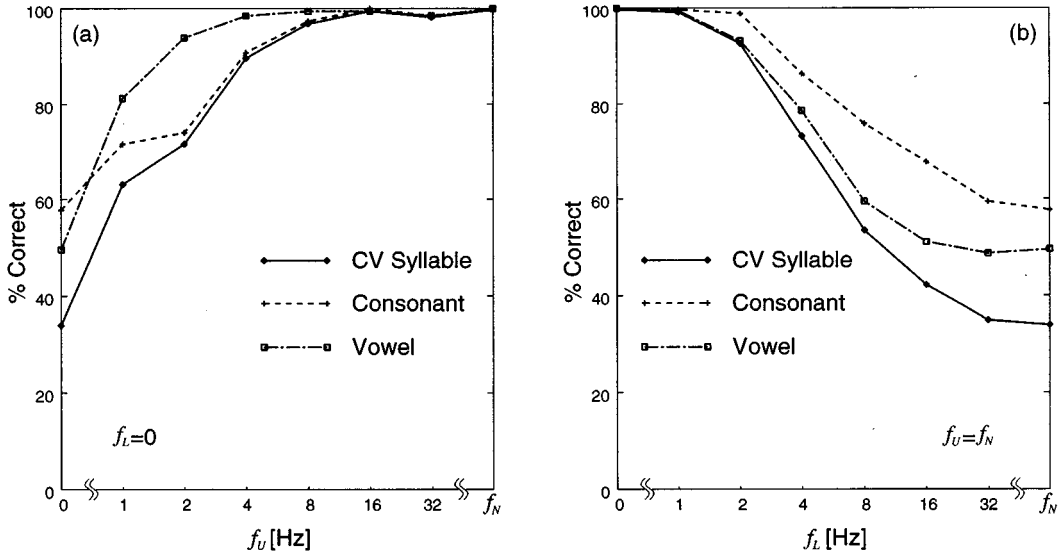

FIG. 9. The result in terms of CV (syllable), C (consonant), and V (vowel). (a) The proportion correct versus f_U when $f_L=0$. (b) The proportion correct versus f_L when $f_U=f_N$.

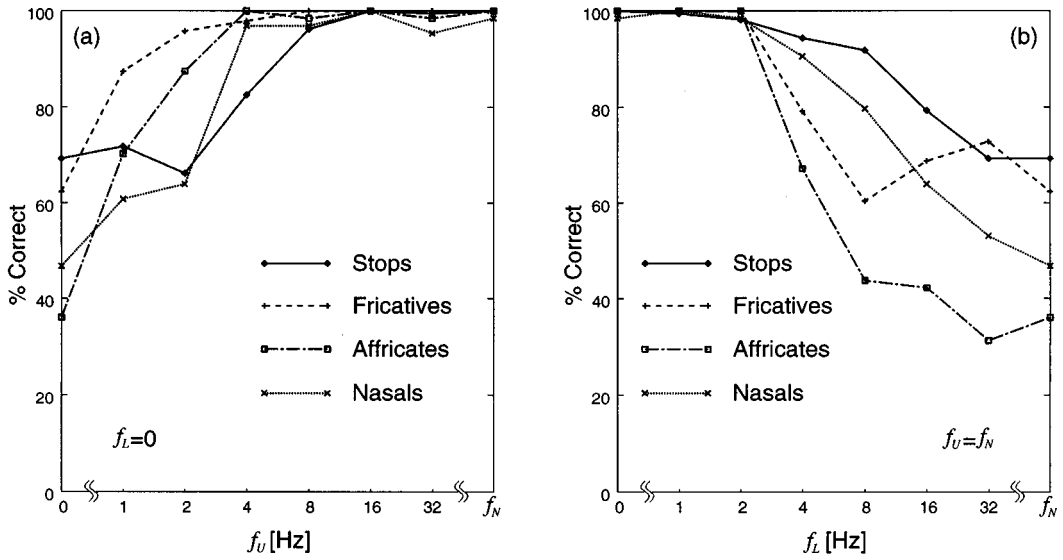

FIG. 10. The result for each consonant category. (a) The proportion correct versus f_U when $f_L=0$. (b) The proportion correct versus f_L when $f_U=f_N$.

contour and the temporal change of the spectrum were affected. We primarily modified the rate of change of the spectral components but kept the LPC residual and the energy contour unchanged.

The LPC residual contains timing information such as plosive bursts, fricative onset, and vowel onset. Therefore, the modified speech signal also contains such timing information as seen in the spectrogram, Fig. 2(e). This technique, however, effectively smears the formant structure of speech as well as the formant transitions. Thus, both vowel and consonant intelligibilities were affected, as seen in Fig. 9. For the residual signal itself, the proportion of correct CV syllables was 34%, whereas for the original signal the proportion of correct CV syllables was 100%. Therefore, our results are valid in the range between those two extremes, and we focus on the relative importance of the modulation frequency.

Figure 6 shows that the modified speech is more intelligible when $f_L \leqslant 1$ Hz and $f_U \geqslant 16$ Hz. The lower limit of 1 Hz suggests that the slowly varying and static components, such as channel characteristics, do not contribute significantly to human speech communication. Similarly, the very fast-changing components above 16 Hz seem to have little effect on intelligibility. In fact, the upper limit of the modulation frequency has important implications for parametric speech coding, particularly for defining how fast we can sample the speech envelope for efficient transmission of speech signals.[9]

In this study, only a target syllable was presented during the experiments. Those syllables were extracted from a longer carrier phrase after temporal filtering. We also conducted a small separate experiment in which we included the carrier phrase as well as the target syllable to see the effect of environmental cues on speech intelligibility. Two native Japanese subjects participated in one experimental session

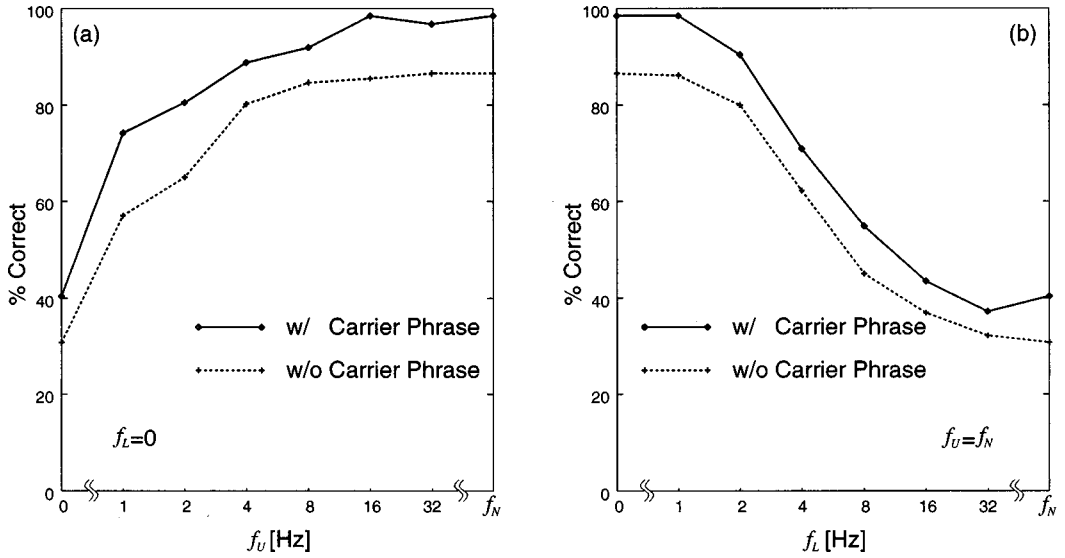

FIG. 11. With and without the carrier phrase. (a) The proportion correct versus f_U when $f_L=0$. (b) The proportion correct versus f_L when $f_U=f_N$.

consisting of 465 stimuli. Figure 11 shows the results in terms of proportion of correct responses with and without the carrier phrase. From Fig. 11 there is approximately 10% gain on average when the target is presented with the carrier phrase. It appears that this gain is due to formant transitions into the consonants of the target syllables. However, the general trend remains the same.

A. Theoretical analysis

The regularity of the relationship between the cutoff frequencies and intelligibility suggests that the information relevant for speech communication is distributed over the temporal modulation frequency range. In this section, we develop a quantitative model that relates the intelligibility and information density as a function of the energy at different modulation frequencies. This model was motivated by the model of the articulation index developed by Fletcher.[8]

The general notion is that each modulation-frequency component contributes independently to the collection of features that are necessary for recognition. In particular, if the information in two nonoverlapping bands A and B is combined, the resulting intelligibility, i.e., the probability of a correct response, P_c, is given by

$$P_c(A+B)=1-[1-P_c(A)][I-P_c(B)], \tag{1}$$

where the intelligibility of each band separately is given by $P_c(A)$ and $P_c(B)$. Moreover, if $P_c(A+B)$ is an additive function of information I_A and I_A, then P_c has the form

$$P_c=1-e^{-(I_A+I_B)}. \tag{2}$$

Given this formulation, we need to determine the relationship between the information measure I and the intervals of the modulation frequency. We assume that the amount of information in a small neighborhood of frequency f, $I(f,f+\Delta f)$, is proportional to a continuous density function D so that $I(f+\Delta f)=D(f)\Delta f$. The amount of information in a band (interval) of frequencies $f_1<f_2$ is then given by the integral

$$I(f_1,f_2)=\int_{f_1}^{f_2} D(f)df. \tag{3}$$

The information density function $D(\cdot)$ must be determined empirically.

Given this model, we can compute the probability of correct identification of the syllables in experiment 2 (shown in Fig. 7) by integrating information between low- and high-frequency limits f_L and f_U. Thus,

$$P_c(f_L,f_U)=1-e^{-I(f_L,f_U)}, \tag{4}$$

$$=1-e^{-\int D(f)df}. \tag{5}$$

We found empirically that the following function:

$$D(f)=1\Big/\left[1+\left(\frac{f-f_{\max}}{\alpha}\right)^2\right], \tag{6}$$

where $f_{\max}$ and α are constants to be determined. Then, by integrating Eq. (6)

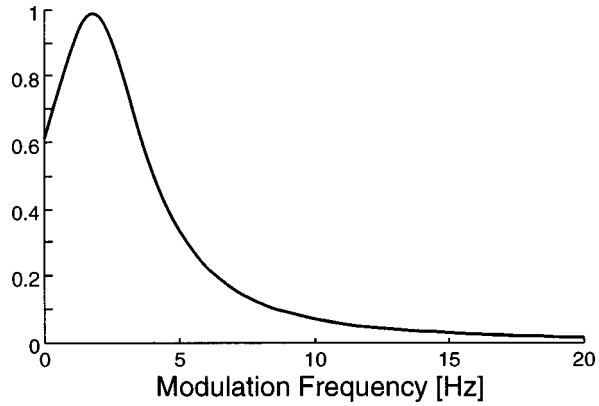

FIG. 12. Density function $D(f)$.

$$I(f_L,f_U)=\alpha\left[\tan^{-1}\left(\frac{f_U-f_{\max}}{\alpha}\right)-\tan^{-1}\left(\frac{f_L-f_{\max}}{\alpha}\right)\right]. \tag{7}$$

The estimated P_c, or $\hat{P}_c$, is obtained by minimizing χ^2, where $f_{\max}$ and α are parameters. Because P_c in Fig. 7 is ranging over the interval $P_0\leqslant P_c<1$, we use the following definition instead of (4):

$$\hat{P}_c(f_L,f_U)=\gamma(1-e^{-\beta I(f_L,f_U)})+P_0. \tag{8}$$

Then, the optimal fit gives us:

$$f_{\max}=1.789,\quad \alpha=2.255,\quad \beta=0.487,$$

$$\gamma=0.726,\quad P_0=0.333,$$

and $D(f)$ is shown in Fig. 12. In this case, χ^2 is 61. Figure 13 shows the fit between P_c and $\hat{P}_c$.

The information distribution function with normalization is defined as the accumulative curve of the density function $D(f)$ as follows:

$$\overline{I}(f)=\int_0^f D(f)df\Big/\int_0^\infty D(f)df. \tag{9}$$

Figure 14 shows $\overline{I}(f)$ as a function of the modulation frequency f in the optimal case. As shown in this figure, 10%,

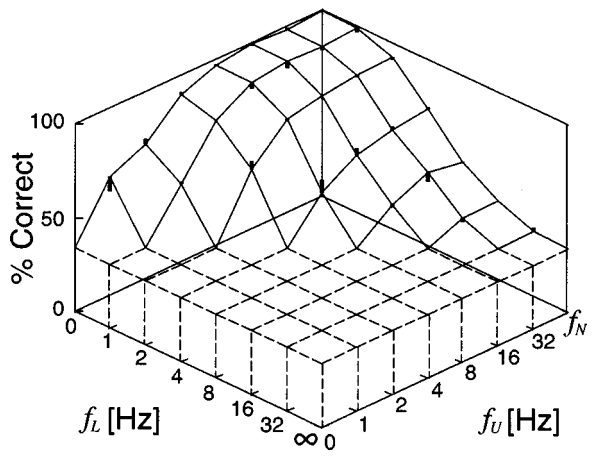

FIG. 13. Fit between P_c and $\hat{P}_c$. The vertical bars show the difference $\hat{P}_c(f_L,f_U)-P_c(f_L,f_U)$ from each $P_c(f_L,f_U)$.

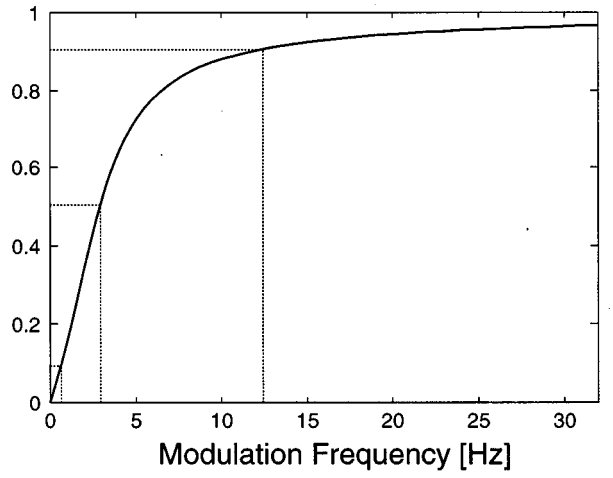

FIG. 14. Distribution function $\overline{I}(f)$.

50%, and 90% percentiles are at 0.7, 2.9, and 11.7 Hz, respectively. The information function shows that most of the information is involved within the range between 1 and 12 Hz. These low-modulation frequencies are, therefore, essential for the human auditory system to extract and recognize speech information.[10]

V. CONCLUSION

The intelligibility of speech with filtered time trajectories of the spectral envelope was investigated. We extended previous research[4,5] to the logarithmic domain and applied bandpass filters in addition to low-pass and high-pass filters. For the result of the perceptual experiments, the information for intelligibility was described by a density function over modulation frequency. The results of these experiments suggest that speech intelligibility is not severely impaired as long as the filtered LPC cepstral coefficients have a rate of change between 1 and 16 Hz.

In contrast to Drullman's study, the effects of temporal filtering of the time trajectories of the cepstrum were examined in this study. In particular, we determined the effect of bandpass filtering. Thus, our results may have direct implications for cepstrum-based ASR systems, and they are important as a benchmark of how the spectral representations used in ASR relate to human speech recognition.

Recently, Kanedera showed that the performance of the speech recognition for the 13-word Bellcore digit database task and the 216 Japanese word-recognition task had the highest recognition rate when the bandpass filter between 1 and 16 Hz was used.[11] This result is consistent with that of our perceptual experiment. The results provide additional support for RASTA-like processing of cepstral features in ASR.

ACKNOWLEDGMENTS

We acknowledge the assistance of Yonghong Yan, Troy Bailey, Brian Mak, and Ronald Cole of the Oregon Graduate Institute of Science & Technology (OGI), who helped with the setup of the initial perceptual experiment; Steven Greenberg of the International Computer Science Institute (Berkeley, California), Robert Damper of the University of Southampton and Pieter Vermeulen of OGI, who gave us useful comments; and Karen Ward of OGI, who helped us to proofread this manuscript. Thanks to Sadaoki Furui of the Tokyo Institute of Technology and the former members of his laboratory at NTT for lending their speech database and for their helpful comments. We would also like to thank the subjects who participated in the experiments. Finally, we would like to thank the two anonymous reviewers, who gave us many valuable suggestions. This research was supported in part by grants from the DoD under Grant No. MDA-904-94-C-6169 and the NSF/ARPA under Grant No. IRI-9314959, with additional funding provided by the member companies of the Center for Spoken Language Understanding (CSLU).

[1] B. S. Atal, ''Effectiveness of Linear Prediction Characteristics of the Speech Wave for Automatic Speaker Identification and Verification,'' J. Acoust. Soc. Am. **55**, 1304–1312 (1974).
[2] S. Furui, ''Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum,'' IEEE Trans. Acoust., Speech, Signal Process. **ASSP34**, 52–59 (1986).
[3] H. Hermansky and N. Morgan, ''RASTA Processing of Speech,'' IEEE Trans. Speech Audio Process. **2**, 578–589 (1999).
[4] R. Drullman, J. M. Festen, and R. Plomp, ''Effect of Temporal Envelope Smearing on Speech Reception,'' J. Acoust. Soc. Am. **95**, 1053–1064 (1994).
[5] R. Drullman, J. M. Festen, and R. Plomp, ''Effect of Reducing Slow Temporal Modulations on Speech Reception,'' J. Acoust. Soc. Am. **95**, 2670–2680 (1994).
[6] T. Houtgast and H. J. M. Steeneken, ''A Review of the MTF Concept in Room Acoustics and its Use for Estimating Speech Intelligibility in Auditoria,'' J. Acoust. Soc. Am. **77**, 1069–1077 (1985).
[7] B. C. J. Moore, *An Introduction to the Psychology of Hearing* (Academic, New York, 1989).
[8] H. Fletcher, *Speech and Hearing in Communication* (Krieger, Huntington, NY, 1953).
[9] J. L. Flanagan, ''Parametric Coding of Speech Spectra,'' J. Acoust. Soc. Am. **68**, 412–419 (1980).
[10] S. Greenberg, ''Understanding Speech Understanding: Towards a Unified Theory of Speech Perception,'' in *Proceedings of the ESCA Tutorial and Advanced Research Workshop on the Auditory Basis of Speech Perception*, edited by W. A. Ainsworth and S. Greenberg (Keele University, Staffordshire, UK, 1996), pp. 1–8.
[11] N. Kanedera, T. Arai, H. Hermansky, and M. Pavel, ''On the Importance of Various Modulation Frequencies for Speech Recognition,'' in *Proceedings of the 5th European Conference on Speech Communication and Technology*, Greece, Vol. 3, pp. 1079–1082 (1997).