INTELLIGIBILITY OF SPEECH WITH FILTERED TIME TRAJECTORIES OF SPECTRAL ENVELOPES

Takayuki Arai, Misha Pavel, Hynek Hermansky, and Carlos Avendano

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

ABSTRACT

The effect of filtering the time trajectories of spectral envelopes on speech intelligibility was investigated. Since LPC cepstrum forms the basis of many automatic speech recognition systems, we filtered time trajectories of LPC cepstrum of speech sounds, and the modified speech was reconstructed after the filtering. For processing, we applied low-pass, high-pass and band-pass filters. The results of the accuracy from the perceptual experiments for Japanese syllables show that speech intelligibility is not severely impaired as long as the filtered spectral components have 1) a rate of change faster than 1 Hz when high-pass filtered, 2) a rate of change slower than 24 Hz when low-pass filtered, and 3) a rate of change between 1 and 16 Hz when band-pass filtered.

1. INTRODUCTION

One of the main objectives of front-end processing in robust automatic speech recognition (ASR) is to preserve critical linguistic information while suppressing irrelevant information such as speaker-specific characteristics, channel characteristics, and noise. Since the information suppressed in the front end of the recognizer is lost for the recognition process, it is important to identify those features of the signal that are useful for human speech recognition. One way to identify the useful features is to eliminate a given feature, reconstruct the speech, and determine experimentally its intelligibility.

Temporal processing of time trajectories of cepstral coefficients of speech is becoming a common procedure in current ASR. The so-called delta features [1] are calculated as linear regression coefficients over a short segment of a time trajectory to emphasize dynamic characteristics of the original features. This is effectively applying an FIR band-pass filter which eliminates the DC component in the time trajectory and applies 6 dB/oct emphasis on changes up to about 12 Hz. RASTA processing [2] also eliminates the DC but, unlike delta feature computation, passes components between 1 and 12 Hz unattenuated. Both techniques appear to achieve some degree of robustness to channel variations.

The goal of our project was to examine the effect of the temporal filtering of the trajectories of spectral features on the intelligibility of the reconstructed speech.

Drullman [3, 4] 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 and the magnitude envelope of the analytic signal for each band was low-pass and high-pass filtered. Drullman concluded that low-pass filtering below 4 Hz and high-pass filtering above 16 Hz do not appreciably reduce speech intelligibility.

Drullman's results are in principle consistent with RASTA processing. He applied the filtering to the magnitude envelope of the analytic signal, which effectively implies filtering of the magnitude spectrum of the speech. In contrast, both delta features and RASTA are typically applied in the logarithmic spectral (cepstral) domain. Additionally, it is not obvious whether his results for the high-pass and the low-pass experiments imply any conclusions about band-pass RASTA filtering.

In this study, we examined the effects of temporal filtering of the time trajectories of LPC cepstrum. In particular, we measured the intelligibility of speech reconstructed from low-passed, high-passed, and band-passed LPC cepstral coefficients. Thus, our results may have direct implications for the cepstrum-based ASR systems.

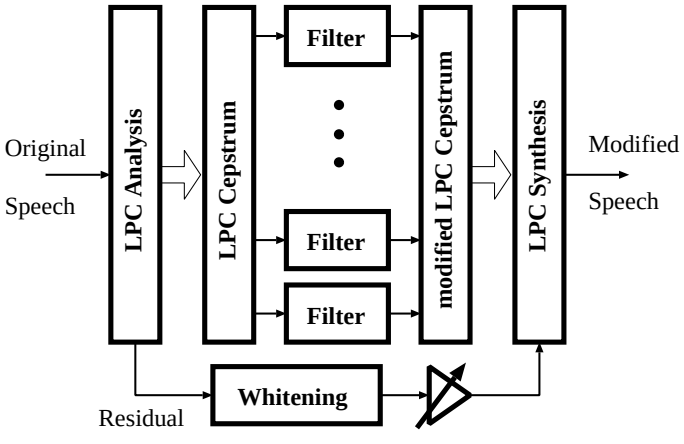

Figure 1: Block diagram of the speech processing system.

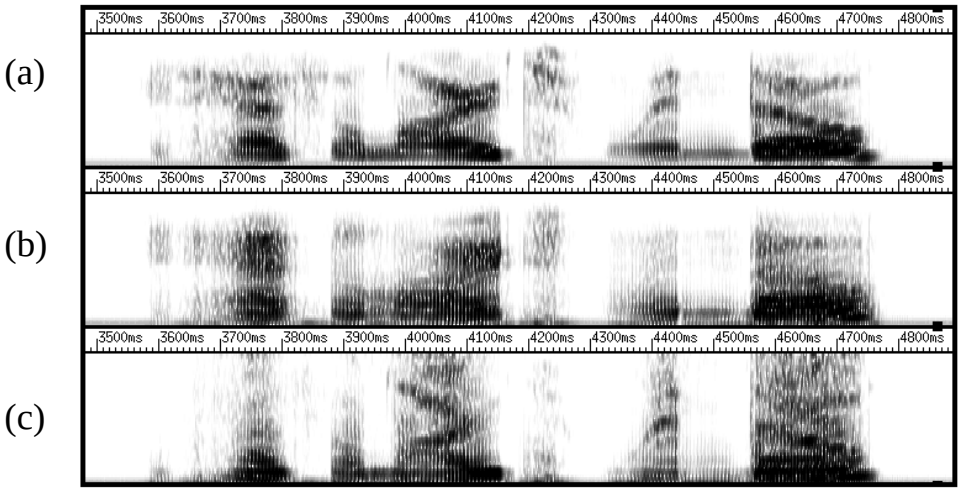

Figure 2: Spectrograms of the utterance "(it) has the large window." (a) Original speech. (b) Speech modified by low-pass filtering at 4 Hz. (c) Speech modified by high-pass filtering at 4 Hz.

2. SIGNAL PROCESSING

The signal-processing method which we used is illustrated in Fig. 1. It consisted of applying frame-by-frame LPC analysis to the original speech and filtering the time trajectories of the resulting LPC cepstral coefficients. Subsequently, the modified speech signal was reconstructed by the LPC synthesis technique. The filters used in this study were either low-pass or high-pass with different cutoff frequencies spanning the entire range of possible frequencies.

The technique was based on a residual-excited LPC vocoder. This approach permits the construction of the entire continuum from the original signal (no filtering) to its LPC residual (complete elimination of spectral envelope variations). 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 spectral envelopes.

2.1. LPC Cepstral Representation

The speech sounds 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 1. Following the LPC analysis, the LPC coefficients were converted to the cepstral coefficients. To keep the original resolution of the logarithmic spectrum, we calculated all cepstral coefficients up to the quefrency of 32 ms (256 points).

2.2. Filtering of the Cepstral Coefficients

The time trajectory for each cepstral coefficient was processed by a temporal filter. The filters were identical at all quefrencies except for zero quefrency. The unit sample function as a filter yields no filtering of the time trajectories. A low-pass filter (LPF) allows slow variations to pass through, while a high-pass filter (HPF) allows fast variations to pass through in the time trajectories. A band-pass filter (BPF) passes through variations in which the rate of change is within a certain range. A filter which has all its coefficients set to zero removes all spectral envelope information. The filters were implemented as a 257-tap FIR filter whose coefficients were calculated by the windowing design method with a Hamming window.

Table 1: Conditions for LPC analysis.

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

Figure 2 shows spectrograms of the utterance "(it) has the large window." A comparison of the spectrograms indicates that the steady-state signal components were preserved in the low-passed signal in Fig. 2(b), whereas the dynamic components (transitions) were preserved in the high-passed signal in Fig. 2(c).

2.3. Reconstruction of the Speech

The filtered LPC cepstral coefficients were used to compute the modified power spectrum at each frame. A 12-order LPC filter was calculated from the auto-correlation function obtained by applying the inverse Fourier transform of the power spectrum.

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 typically yields a relatively intelligible signal. To reduce the intelligibility of the residual, we further whitened the residual signal by applying a second 12th order LPC analysis to the residual from the first LPC analysis.

Table 2: Japanese syllables used in this study.

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

In the last stage of the signal processing we reconstructed speech sounds using the modified LPC coefficients together with the double-whitened residual signal. Care was taken so that the total energy in each frame of the reconstructed speech would match 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.

3. PERCEPTUAL STUDY

3.1. 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 contains a target Japanese syllable in the carrier sentence "Kankonbai _____ oruso." The original speech signal was quantized with a 16 bit resolution and sampled at 48 kHz. Our stimuli were downsampled to 8 kHz.

3.2. Japanese Syllables

The original data set contains 100 Japanese syllables. We selected a subset of 31 syllables covering 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 2.

3.3. Stimulus Preparation and Presentation

For the low-pass and high-pass experiment, the time trajectories of the LPC cepstral coefficients were filtered with cutoff frequencies f_C [Hz], where $f_C = \{0, 1, 2, 3, 4, 5, 6, 8, 12, 24, 48, \infty\}$. A complete set of 13 conditions, including the clean speech, applied to all 31 syllables was presented to subjects in sessions consisting of 403 stimuli. For the band-pass experiment, the time trajectories of the LPC cepstral coefficients were filtered with lower cutoff frequencies f_L [Hz] and upper cutoff frequencies f_U [Hz], where $f_L = \{0, 1, 2, 4, 8, 16, 32, \infty\}$ and $f_U = \{0, 1, 2, 4, 8, 16, 32, \infty\}$. Each subject participated in four sessions. Combinations of syllables and filtering conditions were randomized across sessions and subjects.

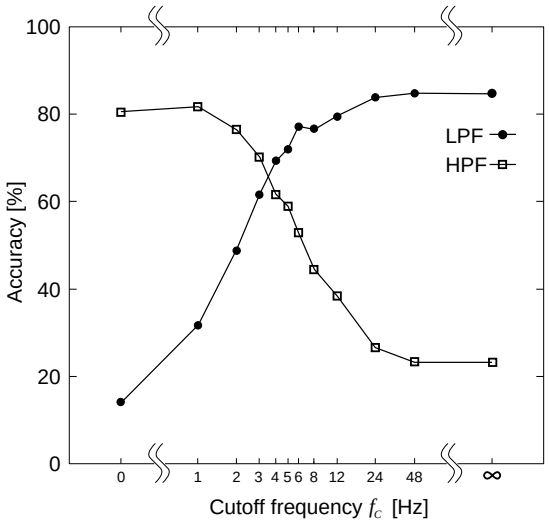

Figure 3: Results for low-pass and high-pass filtering. The number shows the accuracy.

3.4. Procedure

We used the method of constant stimuli with stimuli presented in random order. 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 headphones at a comfortable listening level. On each trial, the subject heard an isolated syllable preceded and followed by a one-second silent interval. Following each stimulus presentation subjects indicated their answer and then initiated the next trial.

Subjects interacted with the experimental setup using a graphical user interface and a mouse input device. The monitor screen showed icons for all 31 possible stimuli and subjects were asked to click on 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.

3.5. Subjects

A total of 16 native speakers of Japanese participated in the study, with eight subjects in each experiment for low-pass and high-pass filtering. All subjects had normal hearing.

4. EXPERIMENTAL RESULTS

Intelligibility was scored by the proportion of correct syllables. A syllable was scored as correct only if both constituents were recognized correctly.

The clean (original) speech had average intelligibility of 85.8% over 16 subjects, ranging from 75.8 to 99.2%. The average residual signal intelligibility was 18.5% over 16 subjects ranging from 7.3 to 27.4%. The useful intelligibility range of the information in LPC cepstral coefficients was therefore 18.5–85.8%.

4.1. Low-Pass and High-Pass Filtering

The resulting intelligibility for the high-pass and low-pass conditions averaged over subjects and syllables is shown in Fig. 3. Consequently, each point is based on 992 trials. The largest standard error of a binomial distribution with the same number of trials is less than 2%. The error bars were omitted for clarity. The graph shows the intelligibility of the low-pass and high-pass filtered LPC trajectories. As seen, the intelligibility starts decreasing below 24 Hz in the low-pass filtered condition. The intelligibility for the high-pass filtered condition starts decreasing above 1 Hz. The intersection of the two curves is located between 3 and 4 Hz. Interestingly, this rate of change approximately coincides with the syllabic rate of speech [5].

4.2. Band-Pass Filtering

One native Japanese subject (the first author of this paper) participated in the preliminary perceptual experiment with band-pass filtering cepstral coefficient trajectories. This particular experiment subject also participated in both the high-pass and the low-pass experiments and is a good representative of typical trends observed in these experiments. There were four experimental sessions, Each session consisted of 29 conditions applied to all 31 syllables, totaling 899 stimuli.

The resulting intelligibility is shown in Fig. 4. This graph shows that the modified speech is highly intelligible when $f_L \leq 1$[Hz] and $f_U \geq 16$[Hz]. The data points show the average over 124 trials. The largest standard error of a binomial distribution with the same number of trials is less than 5%. Because this subject participated in the low-pass and high-pass experiment and generated results consistent with the other subjects, we feel that these data are representative. The data from this preliminary band-pass experiment are also consistent with those of the low-pass and high-pass experiments.

5. CONCLUSION

The results of these experiments suggest that speech intelligibility is not severely impaired as long as the filtered LPC cepstral coefficients have

1) a rate of change faster than 1 Hz as observed in the high-pass filtering experiment,
2) a rate of change slower than 24 Hz as observed in the low-pass filtering experiment, and
3) a rate of change between 1 and 16 Hz as observed in the band-pass filtering experiment.

Thus, it appears that feature rates-of-change outside of these limits are due to various nonlinguistic aspects of the speech signal and can be suppressed. The results provide additional support for RASTA-like processing of cepstral features in ASR.

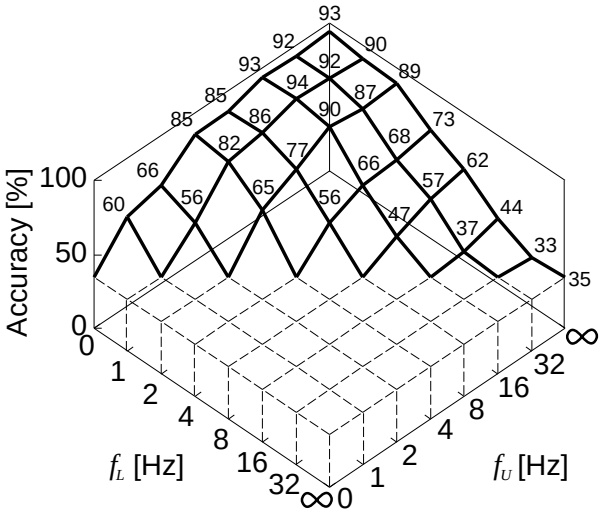

Figure 4: Results for band-pass filtering. The number shows the accuracy.

6. ACKNOWLEDGEMENTS

We acknowledge the assistance of Yonghong Yan, Troy Bailey and Brian Mak, who helped with the setup of the initial perceptual experiment. Thanks to Sadaoki Furui and the members of his laboratory at NTT for the use of their speech database and for their helpful comments. Finally, we would like to thank the subjects who participated in the experiments. This research was supported in part by grants from the DoD under MDA-904-94-C-6169 , the NSF/ARPA under IRI-9314959 with the additional funding provided by the member companies of the Center for Spoken Language Understanding (CSLU).

7. REFERENCES

1. Furui, S. "Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum," *IEEE Trans. Acoust., Speech, Signal Proc.*, ASSP-34: 52–59, 1986.
2. Hermansky, H., Morgan, N. "RASTA Processing of Speech," *IEEE Trans. Speech and Audio Proc.*, 2: 578–589, 1994.
3. Drullman, R., Festen, J. M., and Plomp, R. "Effect of Temporal Envelope Smearing on Speech Reception," *J. Acoustic. Soc. Amer.*, 95: 1053–1064, 1994.
4. Drullman, R., Festen, J. M., and Plomp, R. "Effect of Reducing Slow Temporal Modulations on Speech Reception," *J. Acoustic. Soc. Amer.*, 95: 2670–2680, 1994.
5. Houtgast, T., Steeneken, H. J. M. "A Review of the MTF Concept in Room Acoustics and its Use for Estimating Speech Intelligibility in Auditoria," *J. Acoustic. Soc. Amer.*, 77: 1069–1077, 1985.