ON THE IMPORTANCE OF VARIOUS MODULATION FREQUENCIES FOR SPEECH RECOGNITION

Noboru Kanedera †∗, *Takayuki Arai* †‡, *Hynek Hermansky* †‡, *and Misha Pavel* †

† Oregon Graduate Institute of Science & Technology, Portland, Oregon, U.S.A.
‡ International Computer Science Institute, Berkeley, California, U.S.A.
∗ Ishikawa National College of Technology, Japan

ABSTRACT

Temporal processing of the time trajectories in the logarithmic spectrum domain, performed in cepstral mean subtraction, in computation of dynamic features in speech, or in RASTA processing, is becoming a common procedure in current ASR. Such temporal processing effectively enhances some components of the modulation spectrum of speech while suppressing others. It is therefore important to know the relative importance of various components of the modulation spectrum of speech. In this study we report on the effect of band-pass filtering of the time trajectories of spectral envelopes on speech recognition. Results indicate the relative importance of different components of the modulation spectrum of speech for ASR.

1. INTRODUCTION

In current automatic speech recognition (ASR), temporal processing of time trajectories in the logarithmic spectrum domain is becoming a common procedure. Such processing effectively modifies the so-called modulation spectrum of speech [3]. It is therefore important to know the relative importance of various components of the modulation spectrum of speech for both human speech communication and for ASR. Cepstral mean subtraction (CMS) [1] suppresses the DC components of the time trajectories of the cepstrum to alleviate the effects of the convolutional noise introduced, e.g., by the frequency characteristics of the communications channel (additive in logarithmic spectrum or cepstrum). Such temporal processing effectively enhances some components of the modulation spectrum while suppressing others. Thus, in dynamic features [2], components of the modulation spectrum around 10 Hz are typically enhanced while lower and higher components are suppressed [3]. RelAtive SpecTrAl processing (RASTA) [3] passes components of the modulation spectrum between about 1 and 12 Hz.

Perceptual experiments for Japanese syllables, conducted using band-pass filtered time trajectories of LPC cepstrum in the residual-exited LPC system, indicate the relative importance of various components of the modulation spectrum of speech for speech intelligibility [4]. The results of these experiments suggest that most of the information necessary to preserve intelligibility is the range between 1 and 16 Hz.

In this study we report the effect of similar filtering of the time trajectories of spectral envelopes on speech recognition.

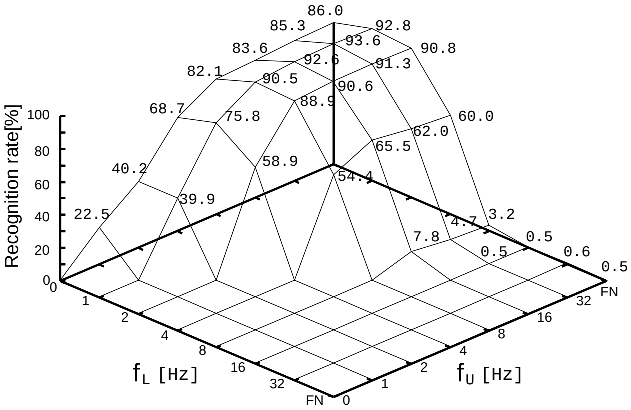

Figure 1: Recognition results for the band-passed time trajectories.

2. SPEECH RECOGNITION EXPERIMENTS AND PERCEPTUAL EXPERIMENTS

To investigate the relative importance of different modulation frequencies for ASR, we performed recognition experiments using 216 phonemically-balanced Japanese words. The system was trained on clean speech while the test data were recorded using a hand-held microphone in a computer room (approximately 55 dB background noise). The training data were drawn from set C of ATR Japanese database and consisted of samples from 10 male native speakers. These tokens, sampled at 20 kHz and quantized to 16 bits, were downsampled to 10 kHz. The test data, uttered by 5 male native speakers, were sampled at 11.025 kHz (also 16 bit quantization). These tokens were passed through a FFT-based filter bank and separated logarithmically into 16 equal parts every 12.5 ms. These were converted to logarithmic values by the Lin-Log function[5],

$$y = \log(1 + Jx), \tag{1}$$

where J is a signal-dependent positive constant. The amplitude-warping transform (1) is linear-like for $J \ll 1$ and logarithmic-like for $J \gg$

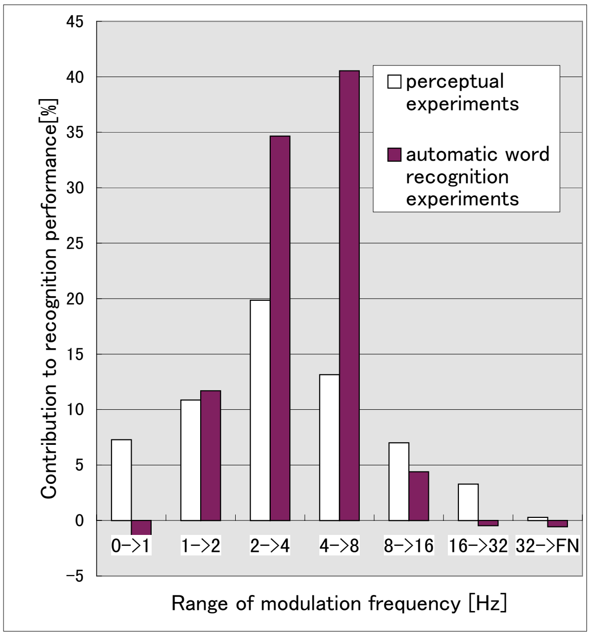

Figure 2: Comparison with the perceptual experiments.

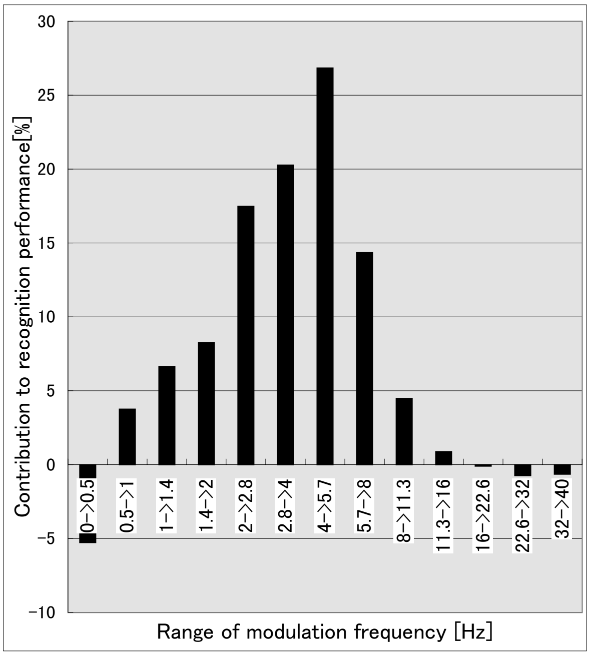

Figure 3: Improvement of recognition accuracy by including each modulation frequency

After this process, the time trajectories were filtered with a 511-tap FIR filter. As we were mainly interested in the relative changes in ASR performance with changes of FIR filter shape rather than in highest obtainable scores, the ASR was done using a simple-but-efficient DTW-based recognizer.

Figure 1 shows recognition results for different band-pass filters applied in the modulation domain. The vertical axis shows the recognition accuracy, while the other axes indicate the lower cutoff frequency f_L and the upper cutoff frequency f_U of the band-pass filters. For some band-pass conditions (indicated by bold letters) there is a noticeable improvement in the recognition accuracy comparing to the accuracy (86%) of the baseline (i.e., no filtering) system.

Figure 2 shows the relative importance of different modulation frequencies. These data were derived by differentiating the surface of the function shown in Figure 1 with a subsequent averaging of all differentials in a given frequency range. [1]

For example, the average improvement resulting from inclusion of the range between 2 and 4 Hz is defined by

$$\frac{1}{7}[\sum_{f_L \leq 2}\{p(f_L,4)-p(f_L,2)\}+\sum_{f_U>4}\{p(2,f_U)-p(4,f_U)\}],$$

where $p(f_1, f_2)$ is the recognition accuracy at $f_L = f_1$, $f_U = f_2$. If $f_L = f_U = f$ then $p(f, f)$ is the prior probability, which is 1/(the number of words) in this ASR experiment.

Average differences in the recognition accuracy are shown in Figure 2. Here each bar indicates average improvement in ASR accuracy resulting from inclusion of a given modulation frequency band. For comparison, results of a similar evaluation using data from perceptual experiments [4] are also shown in Figure 2.

The results indicate that most of the useful linguistic information is in the modulation frequency components between 1 Hz and 16 Hz, with the dominant component at around 4 Hz. In perceptual experiments, which used noise-free speech, some benefit was derived from including the modulation frequency range between 0 and 1 Hz. This range is clearly not useful from the ASR point of view, however. With speech from realistic environment, the modulation frequency components below 1 Hz or above 16 Hz do not contain any useful information and hence may be excluded from the ASR process. The outcome of this experiment is consistent with the current RASTA processing of speech.

Figure 3 shows the details in the improvement of recognition accuracy for different modulation frequencies. The trends shown in Figure 2 and Figure 3 are similar, although the extent of improvement is different because the ranges of the modulation frequencies are different.

[1] For example, consider the improvement of the recognition accuracy resulting from inclusion of a modulation frequency range between 2 and 4 Hz. In Figure 1, the recognition accuracy is 90.6% at f_L = 2 Hz, f_U = 16 Hz, and it is 65.5% at f_L = 4 Hz, f_U = 16 Hz. Accordingly, the improvement resulting from inclusion of the range between 2 and 4 Hz under f_U = 16 Hz is 25.1%. We can also calculate the improvement resulting from inclusion of the range between 2 and 4 Hz under f_L = 1 Hz by subtracting the accuracy at f_L = 1Hz, f_U = 2Hz from the accuracy at f_L = 1 Hz, f_U = 4 Hz.

Table 1: Conditions of word recognition experiment (1) – (7)

Task	13 words Bellcore digit database (0–9, zero, oh, yes, no)
Recognizer	DTW
Training	20 speakers (10 males and 10 females)
Test	50 speakers (25 males and 25 females)
Sampling frequency	8 kHz
Window length	25 ms
Frame period	12.5 ms

Table 2: Conditions of word recognition experiment (8) – (13)

Task	216 words ATR Japanese database set C
Recognizer	DTW
Training	10 male speakers
Test	10 male speakers
Sampling frequency	10 kHz
Window length	25 ms
Frame period	12.5 ms

3. RELATIVE IMPORTANCE OF MODULATION FREQUENCIES IN VARIOUS ENVIRONMENTS

The relative importance of different modulation frequency components may change with environment. We hence investigated the relative importance of different modulation frequency components in various noise environments.

In these experiments, the English words database shown in Table 1 and the Japanese database shown in Table 2 were used. The system was trained on clean speech, while the test data were degraded by convolutional noise and additive background noise.

Figure 4 shows normalized improvements of recognition accuracy in various environments. Environments (14) and (15) in Figure 4 are the same as those in Figure 2. To compare the different cases, improvements are normalized by the maximum value. Black rectangles indicate a decrease in the recognition performance caused by including a given modulation frequency band.

The range below 2 Hz or above 8 Hz becomes less important in noisy environments. Especially, the range below 1 Hz noticeably degrades the recognition accuracy in noisy environments.

Greenberg [6] suggests that speech can withstand channel decorrelations as long as about 120 ms without significant loss in intelligibility. This would be consistent with our results which show that the range of modulation frequency between 2 Hz and 8 Hz (125 ms) is always important in any environment.

4. CONCLUSION

The effect of filtering the time trajectories of spectral envelopes on speech recognition was investigated. These results indicate that: (1) most of the useful linguistic information is in modulation frequency components from the range between 1 Hz and 16 Hz, with the dominant component at around 4 Hz, (2) the range of modulation frequency between 2 Hz and 8 Hz is useful in all environments, (3) in a clean environment, all ranges are useful, (4) in a noisy environment, the range below 2 Hz or above 16 Hz sometimes degrades the recognition accuracy. The outcomes of these experiments are consistent with the current RASTA processing of speech.

Acknowledgements

We acknowledge the collaboration of Sangita Tibrewala and Carlos Avendano of Oregon Graduate Institute of Science and Technology on this project.

5. REFERENCES

[1] B. S. Atal, "Effectiveness of Linear Prediction Characteristics of Speech Wave for Automatic Speaker Identification and Verification," J. Acoustic. Soc. Amer., **55**, pp. 1304–1312, 1974.

[2] S. Furui, "Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum," IEEE Trans. Acoust., Speech, Signal Proc., **ASSP–34**, pp. 52–29, 1986.

[3] H. Hermansky and N. Morgan, "RASTA Processing of Speech," IEEE Trans. Speech and Audio Proc., **2**, pp. 578–589, 1994.

[4] T. Arai, M. Pavel, H. Hermansky, and C. Avendano, "Intelligibility of Speech with Filtered Time Trajectories of Spectral Envelopes," In Proc. of the ICSLP, pp. 2490-2493, Philadelphia, October, 1996.

[5] H. Hermansky, N. Morgan, and H. Hirsch, "Recognition of Speech in Additive and Convolutional Noise Based on RASTA Spectral Processing," IEEE Int. Conf. Acoust., Speech, Signal Processing, pp. II-83 – II-86, 1993.

[6] S. Greenberg, "Understanding Speech Understanding — Towards a Unified Theory of Speech Perception," In *Proc. of the ESCA Tutorial and Advanced Research Workshop on the Auditory Basis of Speech Perception*, Keele, England, pp. 1–8, 1996.

	Task	Test environment	
		Convolutional noise	Additive noise
(1)	Bellcore digit database (13 words)	nothing	nothing
(2)			20 dB
(3)			10 dB
(4)			0dB
(5)		HPF (6 dB/oct)	nothing
(6)			20 dB
(7)			10 dB
(8)	ATR Japanese database set C (216 words)	nothing	nothing
(9)			20 dB
(10)			10 dB
(11)		BPF (300 Hz -- 3 kHz)	nothing
(12)			20 dB
(13)			10 dB
(14)		computer room	
(15)	perceptual experiments		

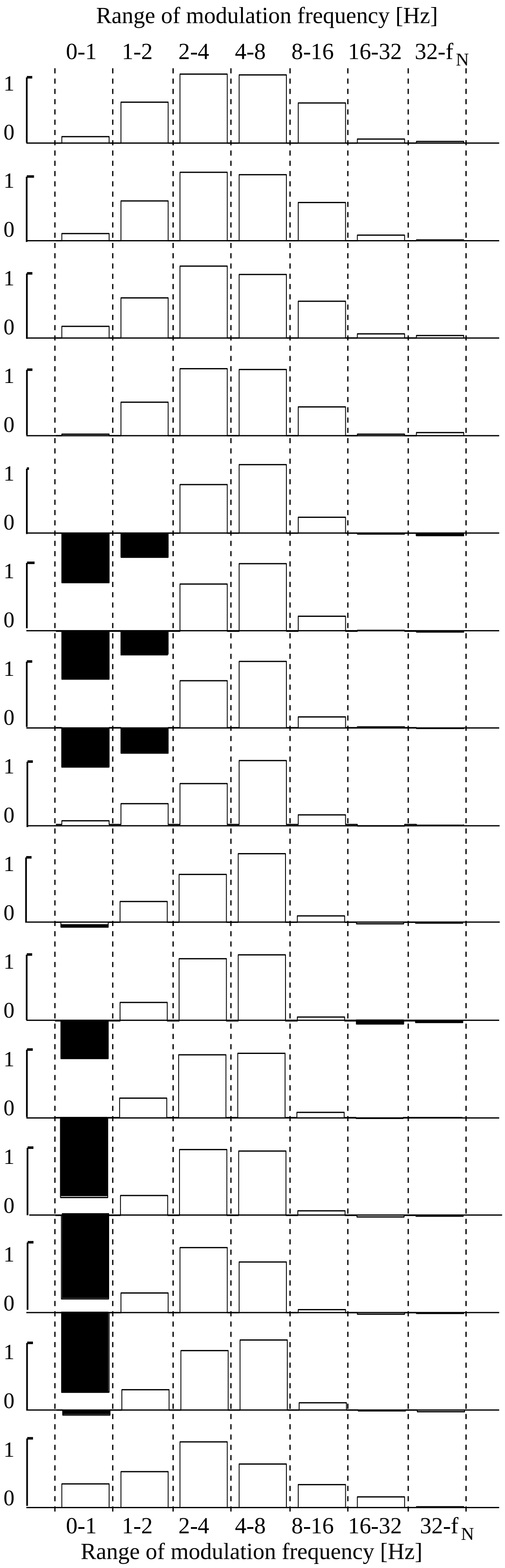

Figure 4: Normalized improvements of recognition accuracy in various environments