ON PROPERTIES OF MODULATION SPECTRUM FOR ROBUST AUTOMATIC SPEECH RECOGNITION

Noboru Kanedera ∗ *Hynek Hermansky* †‡ *Takayuki Arai* ‡

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

ABSTRACT

We report on the effect of band-pass filtering of the time trajectories of spectral envelopes on speech recognition. Several types of filter (linear-phase FIR, DCT, and DFT) are studied. Results indicate the relative importance of different components of the modulation spectrum of speech for ASR. General conclusions are: (1) most of the useful linguistic information is in modulation frequency components from the range between 1 and 16 Hz, with the dominant component at around 4 Hz, (2) it is important to preserve the phase information in modulation frequency domain, (3) The features which include components at around 4 Hz in modulation spectrum outperform the conventional delta features, (4) The features which represent the several modulation frequency bands with appropriate center frequency and band width increase recognition performance.

1. INTRODUCTION

Temporal processing of time trajectories in the logarithmic spectrum domain is becoming a common procedure in automatic speech recognition (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). In delta features [2], components of the modulation spectrum around 10 Hz are typically enhanced while lower and higher components are suppressed. RelAtive SpecTrAl processing (RASTA) [3] passes components of the modulation spectrum between about 1 and 12 Hz. Such processing effectively modifies the so-called modulation spectrum of speech [3].

Perceptual experiments indicate that some components of the modulation spectrum are more important for the intelligibility of speech than others. This fact was also confirmed in Japanese [7] and English [9]. Drullman et al. [5, 6] concluded that low-pass filtering below 16 Hz or high-pass filtering above 4 Hz does not appreciably reduce speech intelligibility. Arai et al. [7] extended Drullman's research [5, 6] to the logarithmic domain and applied not only low-pass or high-pass filters but also band-pass filters. The results of these experiments suggest that most of the information necessary to preserve intelligibility is the range between 1 and 16 Hz.

Nevertheless, the relative importance of various components of the modulation spectrum is not well known. Therefore Kanedera et al. [14] investigated the relative importance of different components of the modulation spectrum of speech. These results indicate that 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. In a noisy environment, the range below 1 Hz is not contributing useful information and the recognition performance can be typically improved by eliminating it from the recognition process. Such trends of the relative importance of modulation frequencies were unchanged despite changes of recognizers and features used.

2. RELATIVE IMPORTANCE OF MODULATION FREQUENCIES

This section describes work on the relative importance of modulation frequencies in ASR. The experimental system is shown in Figure 1. It consists of a module which extracts logarithmic spectrum, one which filters time trajectories of components of the logarithmic spectrum, and of a pattern classification module which yields the final recognition results. The recognition accuracy $p(f_L, f_U)$ of the system is a function of the lower cutoff frequency f_L and the upper cutoff frequency f_U of the band-pass filter.

To derive some notion of the relative importance of various modulation frequencies, we defined the contribution $I(f_L, f_U)$ to recognition performance resulting from inclusion of the range between f_L and f_U by

$$I(f_L, f_U) = \frac{1}{N-1}\Big[\sum_{l<f_L}\{p(l, f_U) - p(l, f_L)\} + \sum_{u>f_U}\{p(f_L, u) - p(f_U, u)\}\Big], \quad (1)$$

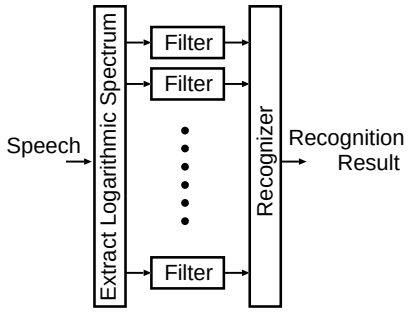


Figure 1: Block diagram of the ASR system with temporal processing.

where N is the number of ranges of modulation frequency [14].

3. AUTOMATIC SPEECH RECOGNITION EXPERIMENTS

3.1. Effect of filter

In the previous work [14], long linear phase FIR filters were used to obtain the sharp characteristics in modulation frequency domain. However long non-causal filter causes long time delay. In practical ASR, short filters are desirable to reduce the time delay introduced by the filtering. Here we investigated the relative importance of modulation spectrum using DFT.

Figure 2 shows normalized contribution to recognition performance for DFT filtering with Hamming window. The horizontal axis shows center modulation frequency for each DFT filter. In these experiments, the English words database shown in Table 1 were used. The system was trained on clean speech, while the test data were degraded by convolutional noise and additive background noise. To compare the different cases, contributions are normalized by the maximum value. The features were 8th order PLP and logarithmic energy. The HMM tool kit (HTK) was used to train a Gaussian mixture HMM. In this case, each of the 13 words were modeled by 8 states (including a nonemitting initial and final state), and there were 2 mixtures per states. Covariance matrices for each mixture were assumed to be diagonal.

Although modulation frequency characteristics of the filters are quite different, trends in the relative importance of modulation spectrum components are similar as in the earlier experiments. The range around 4 Hz is useful both in clean environment and in noisy environment. In noisy environment, the range below 2 Hz or above 10 Hz can be less important. In particular, the range below 1 Hz degrade the recognition accuracy.

Table 1: Conditions of word recognition experiment.

Task	13 words Bellcore digit database (0–9, zero, oh, yes, no)
Training	150 speakers (75 males and 75 females)
Test	50 speakers (25 males and 25 females)
Sampling frequency	8 kHz
Window	Hamming (25 ms)
Frame period	12.5 ms
Features	8th order PLP and logarithmic energy

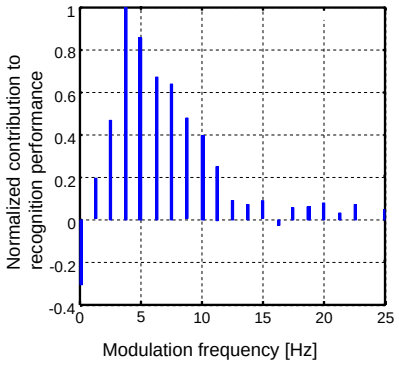
3.2. 2-D cepstrum using selected part of the modulation spectrum

The Fourier transform of the time trajectory of cepstrum has become known as 2-D cepstrum [11, 12]. Imai and Kitamura [11] used a 2-D cepstrum derived for the whole word. Milner [13] used a 2-D cepstrum derived form a relatively short segments of speech on frame-by-frame basis. We used much longer temporal window compared to [13] and aimed at only some selected parts of the modulation spectrum. Thus, similarly to [13] we employ harmonic base functions on frame-by-frame basis but unlike [13] we extract much lower frequency components of the modulation spectrum, and use only some selected ones. In some respects, our temporal processing is more reminiscent to computation of so called dynamic features of speech [2] where Hamming-window weighted harmonic functions are used in place of polynomials used in deriving the dynamic features.

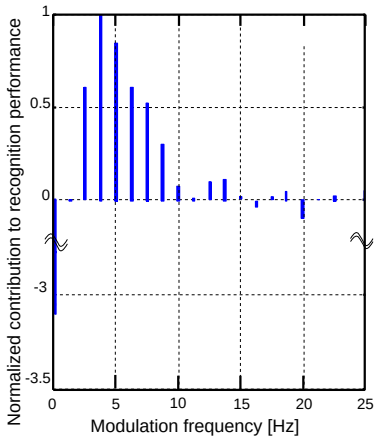
Table 2 shows a comparison between conventional delta features and the 2-D cepstrum using selected part of the modulation spectrum. In clean environment, the test environment was the same as the training environment, while the test data were degraded by additive noise (10 dB) and convolutional noise (HPF, 6 dB/oct), i.e., mismatched environment, in noisy environment.

The 2-D real cepstrum was derived form each time trajectory of 8th order PLP cepstral coefficients (including logarithmic energy), using DCT on 16 points of the time trajectory (i.e. using 32-points DFT after taking 16-points to create even symmetric sequence weighted by Hamming window). The 2-D complex cepstrum was derived from 32-points of time trajectory weighted by Hamming window. Only second and third components of both real and complex cepstra were used. These components correspond to 5 and 7.5 Hz center frequency, and cover the range between about 3 and 9.5 Hz in modulation frequency.

The number of features was 18 (2 components $\times$ 1 (only



(a) Clean



(b) Noisy

Figure 2: Normalized contribution to recognition performance for DFT filtering.

real part) $\times$ 9 original static features) for using DCT, and 36 (2 components $\times$ 2 (real and imaginary parts) $\times$ 9 original static features) for using DFT. When combining with 9 static features, the number of feature was 27 for the real (DCT-based) cepstrum and 45 for the DFT-based complex cepstrum.

These results show that 2-D cepstrum can significantly outperform conventional delta features, especially in mismatched environment. The results using DFT which uses 32-points temporal window (400 ms) is better than those using DCT which has 16-points temporal window (200 ms).

3.3. Effect of phase

Table 3 shows the effect of including the phase. The conditions were the same as those in Section 3.1. The results of experiments using only absolute values are worse than results from experiments which used both real or imaginary parts. From these results we conclude that it appears to be important to use the phase information of the modulation spectrum.

Table 2: Recognition results of conventional delta features and 2-D cepstrum using selected part of the modulation spectrum.

Feature	Static feature	Feature size	WER [%]	
			Clean	Noisy
Δ, Δ^2	yes	27	1.7	21.7
2-D cepstrum (DCT)	yes	27	0.9	12.8
2-D cepstrum (DFT)	yes	45	0.9	4.6
Δ, Δ^2	no	18	2.8	28.8
2-D cepstrum (DCT)	no	18	4.2	4.8
2-D cepstrum (DFT)	no	36	3.4	3.5

Table 3: Recognition results in various phase conditions.

Feature	Static feature	Feature size	WER [%]	
			Clean	Noisy
2-D cepstrum (DFT)	yes	45	0.9	4.6
Real part	yes	27	1.5	14.2
Imaginary part	yes	27	1.1	10.3
Absolute values	yes	27	5.2	24.2
2-D cepstrum (DFT)	no	36	3.4	3.5
Real part	no	18	7.7	10.5
Imaginary part	no	18	8.0	11.8
Absolute values	no	18	22.5	38.8

3.4. Multi-resolution ASR

The results in Section 3.1 show that most of the useful linguistic information is in modulation frequency components in the range between 1 and 16 Hz, with dominant contributions coming from the range between 2 and 10 Hz.

Table 4 shows the results of several combination using DFT with 16-points, 32-points, and 64-points. The conditions were the same as those in Section 3.1. For each resolution, we selected the components in the range between 2 and 10 Hz. In the case (g), second components using 64-points DFT and second and third components using 32-points DFT were used. These components correspond to 2.5, 5, and 7.5 Hz in center modulation frequency respectively. These modulation frequencies would correspond to roughly word rate, syllabic rate and demi-syllabic rate respectively. Resulting improvements indicate the potential of multi-resolution approach in which several information streams derived from the speech signal using various lengths of the temporal windows for deriving the modulation spectrum based features are combined.

Table 4: Recognition results using multi-resolution.

DFT size	16	32	64	Feature size	WER [%]		Case
					Clean	Noisy	
Order of DFT components	1	—	—	18	2.4	24.9	(a)
	—	2, 3	—	36	3.4	3.5	(b)
	—	—	2–6	90	2.8	2.3	(c)
	1	2, 3	—	54	1.7	6.2	(d)
	—	2, 3	2–6	126	2.0	2.5	(e)
	1	2, 3	2–6	144	1.5	2.2	(f)
	—	2, 3	2	54	1.4	1.9	(g)

4. CONCLUSIONS

Results indicate that most of the useful linguistic information is in modulation frequency components in the range between 1 and 16 Hz, with dominant contributions coming from the range between 2 and 8 Hz. These observations are quite independent of the particular character of filters applied in the modulation spectrum domain.

Results also indicate that it is important to preserve the phase information of the modulation spectrum. The features which include components of the modulation spectrum from around 4 Hz outperform the conventional delta features, especially in mismatched training and test environments. Using several modulation frequency bands which correspond to word rate, syllabic rate, and demi-syllabic rate, appears to be beneficial.

Finally, the results show that it may be useful to use some selected parts of the modulation spectrum in multi-stream ASR.

Acknowledgments

We acknowledge the collaboration of Sangita Tibrewala, Misha Pavel, Carlos Avendano, Narendranath Malayath, and Sarel van Vuuren of Oregon Graduate Institute of Science and Technology on this project. We also thank B. Yegnanarayana of the Indian Institute of Technology, for useful comments.

5. REFERENCES

[1] B. S. Atal (1974), "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," J. Acoust. Soc. Amer., Vol. 55, No. 6, pp. 1304 – 1312.

[2] S. Furui (1986), "Speaker-independent isolated word recognition using dynamic features of speech spectrum," IEEE Trans. Acoust. Speech Signal Process., Vol. ASSP-34, No. 1, pp. 52 – 59.

[3] H. Hermansky and N. Morgan (1994), "RASTA processing of speech," IEEE Trans. Speech and Audio Process., Vol. 2, No. 4, pp. 578 – 589.

[4] T. Houtgast and H. J. M. Steeneken (1985), "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria," J. Acoust. Soc. Amer., Vol. 77, pp. 1069 – 1077.

[5] R. Drullman, J. M. Festen, and R. Plomp (1994), "Effect of temporal envelope smearing on speech perception," J. Acoust. Soc. Amer., Vol. 95, pp. 1053 – 1064.

[6] R. Drullman, J. M. Festen, and R. Plomp (1994), "Effect of reducing slow temporal modulations on speech perception," J. Acoust. Soc. Amer., Vol. 95, pp. 2670 – 2680.

[7] T. Arai, M. Pavel, H. Hermansky and C. Avendano (1996), "Intelligibility of speech with filtered time trajectories of spectral envelopes," In Proc. of the ICSLP, Philadelphia, pp. 2490 – 2493.

[8] H. Hermansky, N. Morgan and H. Hirsch (1993), "Recognition of speech in additive and convolutional noise based on RASTA spectral processing," Proc. IEEE ICASSP, Minneapolis, MN, pp. II-83 – II-86.

[9] S. Greenberg (1996), "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.

[10] H. Hermansky (1990), "Perceptual linear predictive (PLP) analysis for speech," J. Acoust. Soc. Amer., Vol. 87, No. 4, pp. 1738 – 1752.

[11] S. Imai and T. Kitamura (1976), "Speech analysis using two dimensional cepstrum," Transactions of the Institute of Electronics and Communication Engineers of Japan, Vol. 59-A, No. 12, pp. 1096 – 1103. In Japanese.

[12] T. Kitamura and K. Katayanagi (1989), "Digit recognition using static and dynamic features of two dimensional mel-cepstrum," Transactions of the Institute of Electronics, Information and Communication Engineers of Japan, Vol. J72-A, No. 4, pp. 640 – 647. In Japanese.

[13] B. Milner (1996), "Inclusion of temporal information into features for speech recognition," In Proc. of the ICSLP, Philadelphia, pp. 256 – 259.

[14] N. Kanedera, T. Arai, H. Hermansky, and M. Pavel (1997), "On the importance of various modulation frequencies for speech recognition," Proc. Eurospeech '97, Rhodes, Greece, pp. 1079 – 1082.