社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.
SP2003-54 (2003-06)

変調スペクトルの貢献度に基づく連続音声認識

金寺　登† 荒井　隆行†† 岡田　賢治†† 浅井　健司††

† 石川高専　〒929-0392 石川県河北郡津幡町北中条
†† 上智大学　〒102-8554 東京都千代田区紀尾井町 7-1
E-mail: †kane@i.ishikawa-nct.ac.jp, ††arai@sophia.ac.jp

あらまし　音声特徴量の時間軌跡をフーリェ変換したものは変調スペクトルと呼ばれ，音声の認識には特定の変調スペクトルが重要であることが知られている．本報告では，音声認識にとって変調スペクトルの各成分がどの程度重要であるかを示す貢献度に応じて変調スペクトルを強調した音声認識特徴量を提案する．自動音声認識実験の結果，提案した特徴量は，雑音環境下において音声認識性能が約 5% 改善されることを確認した．
キーワード　変調周波数，連続音声認識

Continuous Speech Recognitioin Based on the Contribution of Modulation Spectrum

Noboru KANEDERA†, Takayuki ARAI††, Kenji OKADA††, and Kenji ASAI††

† Ishikawa National College of Technology, Tsubata, Ishikawa, 929-0392 Japan
†† Sophia University, 7-1 Kioi-cho, Chiyoda-ku, Tokyo, 102-8554 Japan
E-mail: †kane@i.ishikawa-nct.ac.jp, ††arai@sophia.ac.jp

Abstract　The Fourier transform of the time trajectories of a parameter such as logarithmic spectrum or cepstrum is called the modulation spectrum. In this paper we propose new feature for automatic speech recognition based on the contribution of modulation frequency components. The contribution shows the importance of each modulation frequency component for speech recognition. In proposed method, the time trajectory of each transformed spectral component is filtered by a linear-phase FIR filter with modulation-frequency characteristics based on the contribution as a substitute for RASTA filter. The proposed feature has two important properties: (1) little phase distortion and (2) effective enhancement of important modulation frequency components. The phase distortion of the RASTA filter increases a recognition error. The proposed FIR filter emphasizes important modulation frequency components of speech according to the contribution, while alleviates most modulation frequency components of noise. Testing proposed feature on IPA98 task (Japanese continuous speech recognition task) in noisy environments (SNR 10 dB) gave a relative improvement of 5 % in word accuracy over the MFCC with dynamic feature. The results show proposed modulation filtering based on the contribution of modulation frequency components is effective.
Key words　modulation frequency, continuous speech recognition

1. Introduction

It is known that a particular part of modulation frequency components is important for speech recognition [1]~[5], [12].

Perceptual experiments [2], [3] indicate that some components of the modulation spectrum are more important for the intelligibility of speech than others. Drullman et al. [2], [3] concluded that low-pass filtering below 16 Hz or high-pass filtering above 4 Hz does not appreciably reduce speech intelligibility. Greenberg [6] confirmed the Drullman's results informally. Arai et al. [4], [5] extended Drullman's research [2], [3] 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.

Kanedera et al. [10]~[12] investigated on the effect of filtering of the time trajectories of spectral envelopes on

ASR(automatic speech recognition). The results indicate that most important linguistic information is in modulation frequency components from the range between 1 and 16 Hz, especially between 1 and 10 Hz with the dominant component at around 4 Hz (the average syllabic rate of speech [1]).

We, therefore, carried out digit speech recognition experiments using Modulation FT (Fourier transform) that represents important modulation frequency components. Experimental results on various noise conditions (SNR 10dB) show an 8% reduction in average error rate by using Modulation FT as compared to conventional MFCC with delta features [17]. Furthermore, Momomura et. al. obtained good results of ASR experiments using Modulation Wavelet that represents multi-resolutional features in modulation spectral domain efficiently [19]. Although Modulation FT or Modulation Wavelet is very useful for word recognition, these features could not display these capabilities in tri-phone model, because tri-phone model, which is often used for continuous speech recognition, can not represent long time period corresponding to important low modulation frequency components. Modulation FT or Modulation Wavelet would be useful in case using word model or sub-word model for continuous speech recognition. The alternative to emphasize important low modulation frequency components is pre-filtering the important components.

In this paper we propose new feature for continuous speech recognition based on the contribution which shows the importance of each modulation frequency component for speech recognition. The new feature emphasizes important modulation frequency components such as Modulation FT and Modulation Wavelet.

2. Features for ASR based on the contribution of modulation frequency components

2.1 Contribution to recognition performance

The contribution of modulation frequency components shows the importance of each modulation frequency component for speech recognition [12]. By including a modulation frequency band, the probability of error is multiplied by the factor of 1/(the corresponding contribution). That is, if the contribution of a modulation frequency band is greater than 1.0, it represents that the recognition performance of the system including the band will be improved.

2.2 Contribution for continuous speech

To extract the contribution of modulation frequency components for continuous speech, the modulation spectrum at each frame was extracted by 64-point DFT with the hamming window for 64 frames picked up from the time trajectories of 8-th order PLP (perceptual linear predictive coding) [7] and logarithmic power around the frame. As each component of the modulation spectrum is corresponding to output of the modulation band-pass filter, some modulation frequency components combine together to make feature which contains information of the corresponding range in modulation frequency domain. We carried out continuous automatic speech recognition experiments using the features made from various combinations of modulation frequency components. In the recognizer, syllable HMMs were used without a language model to avoid the effect of the language model. Each syllable was modeled by 5 states (including nonemitting initial and final states). The training data was 3000 sentences uttered by 10 male and 10 female speakers selected from ATR Japanese database set C (continuous speech, 6000 sentences uttered by 40 speakers, 150 sentences for each speaker) [24]. The evaluation data was 3000 sentences uttered by other 10 male and 10 female speakers.

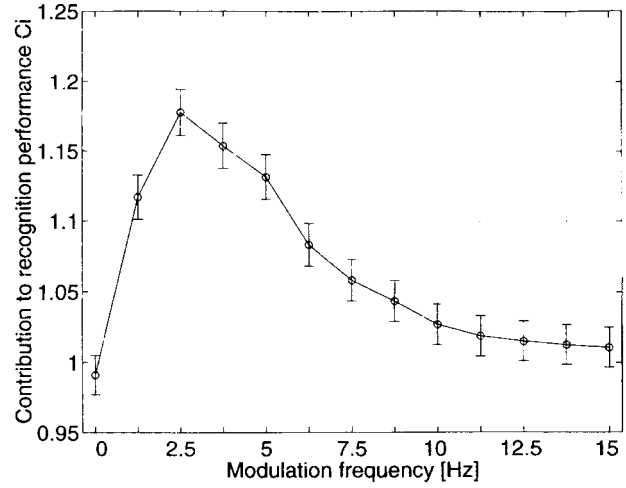

Fig. 1 The contribution of modulation frequency components to recognition performance for continuous speech

Figure 1 shows the contribution to recognition performance derived from these recognition results [16]. In Fig. 1, each bar indicates the contribution with 95% confidence intervals to the overall ASR accuracy by the corresponding modulation frequency band. The results indicate that most important linguistic information is in modulation frequency components from the range between 1 and 16 Hz, especially between 1 and 10 Hz.

2.3 Lin-Log RASTA

RASTA(RelAtive SpecTrAl processing) [8] improves recognition performance passing the important modulation frequency band. The time trajectory of each transformed spectral component is filtered by RASTA filter with modulation-frequency characteristics shown in Fig.2. The RASTA filter passes the modulation frequency components from the range between 1 and 12 Hz. The range is consistent with important

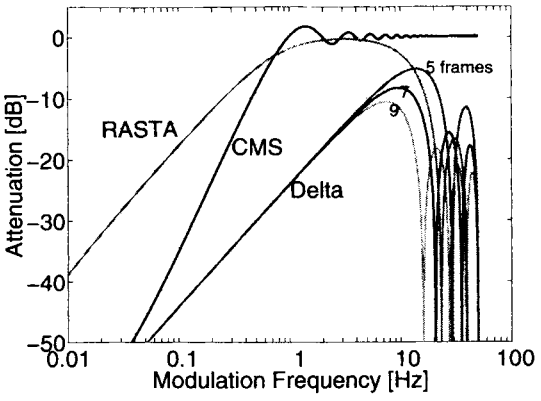

Fig. 2 Modulation-frequency characteristics of CMS, Delta, and RASTA.

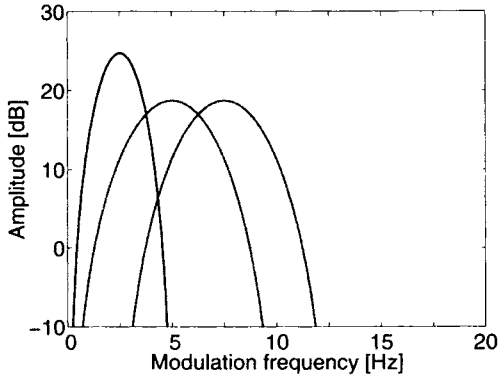

Fig. 3 Modulation-frequency characteristics of Modulation FT.

modulation frequency band shown in Fig.1.

In [9], Hermansky et al. have proposed as a substitute for the logarithmic transform in RASTA processing the function

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

where J is a signal-dependent positive constant, and x is a critical band power spectrum. The amplitude-warping transform (1) is linear-like for $J \ll 1$ and logarithmic-like for $J \gg 1$. This method is robust for both convolutional and additive noise.

2.4 Modulation FT

RASTA described in Sec.**2. 3** extracts important modulation frequency band using a band-pass filter. On the other hand, Modulation FT [17] extracts several important frequency bands efficiently using multi-resolutional band-pass filters with modulation-frequency characteristics shown in Fig. 3. In Fig. 3, the second component of 64-point DFT and the second and third components of 32-point DFT were used.

In digit speech recognition experiments [17] on various noise conditions (SNR 10dB), approximately 8% improvement in average error rate was obtained by using Modulation FT as compared to conventional MFCC with delta features [22].

2.5 Features for ASR based on the contribution of modulation frequency components

We propose new feature for ASR based on the contribution of modulation frequency components. In proposed method, the amplitude-warping transform shown in Eq.(1) is employed, and the time trajectory of each transformed spectral component is filtered by a linear-phase FIR filter with modulation-frequency characteristics shown in Fig.1 as a substitute for RASTA filter.

The proposed feature has two important properties: (1) little phase distortion and (2) effective enhancement of important modulation frequency components. The phase distortion of the RASTA filter causes a recognition error [23]. The proposed FIR filter emphasizes important modulation frequency components of speech according to the contribution, while alleviates most modulation frequency components of noise.

3. Word speech recognition using important modulation spectrum

3.1 Experimental setup

To evaluate features described in Sec.**2.** for isolated-word recognition in noisy environments, ATR Japanese database set C (216 words, 20 male and 20 female speakers) [24] was used. The system was trained on clean speech uttered by 15 male and 15 female speakers, while the test data uttered by other 5 male and 5 female speakers were degraded by additive background noise with SNR 0 dB. The four sets of these training and test data were used in Jack-knife method. The 15 kinds of noise was used from NOISEX-92 database [15] as additive background noise.

3.2 Results using word HMMs

Table 1 shows isolated-word recognition results using word HMMs in various noise environments. The HMM Tool kit (HTK) was used to train a Gaussian mixture HMM. In this case, each of the 216 words were modeled by 8 states (including nonemitting initial and final states), and there were 2 mixtures per states. Covariance matrics for each mixture were assumed to be diagonal. In Table 1, clean shows word error rate without noise, and additive noise shows word error rate for the speech degraded by 15 kinds of noise with SNR 0 dB.

MFCC+Δ shows 12-th order MFCC with CMS [21] and their dynamic features Δ with Δ of logarithmic power. PLP+Δ shows 8-th order PLP [7] with CMS and their dynamic features. Modulation FT shows feature described in Sec.**2. 4**. "Real" shows results using only real part of DFT with modulation-frequency characteristics shown in Fig. 3. "Real&Imag." shows results using both real part and imaginary part of the DFT. Modulation FT using only real part

Table 1 Isolated word recognition results using word HMMs (Word error rate[%])

Feature	MFCC +Δ	PLP +Δ	PLP-RI +Δ+Δ^2	Modulation FT Real	Modulation FT Real &Imag.
clean	9.8	8.1	4.4	2.5	1.8
additive noise					
babble	42.6	44.9	41.5	16.8	16.4
buccaneer1	31.8	29.7	29.9	18.7	12.6
buccaneer2	26.2	25.9	27.3	19.5	14.3
destroyerengine	43.6	49.8	36.4	25.3	17.8
destroyerops	34.2	30.9	36.1	11.0	8.4
f16	36.4	30.2	31.3	17.4	12.8
factory1	33.9	30.8	32.8	16.1	11.3
factory2	36.5	36.2	28.2	10.8	7.7
hfchannel	60.1	49.7	26.4	27.3	20.4
leopard	39.8	46.2	42.2	6.6	4.9
m109	30.3	29.9	27.2	8.1	6.3
machinegun	54.9	56.7	64.1	32.8	30.1
pink	26.7	24.5	23.6	15.6	11.4
volvo	24.2	27.4	11.2	3.4	2.2
white	34.1	40.0	21.0	26.4	20.0
all noises (mean)	37.0	36.9	32.0	17.0	13.1
feature size	25	17	26	27	54

Table 2 Isolated word recognition results using tri-phone HMMs (Word error rate[%])

Feature	MFCC +Δ	PLP +Δ	PLP-RI +Δ+Δ^2	Modulation FT Real	Modulation FT Real &Imag.
clean	5.6	5.7	3.5	8.5	9.4
additive noise					
babble	54.9	68.7	31.1	48.2	21.0
buccaneer1	13.2	19.4	6.3	19.9	20.9
buccaneer2	14.7	23.8	7.4	20.1	22.1
destroyerengine	14.9	34.9	8.7	27.0	24.5
destroyerops	37.8	56.8	19.0	21.0	17.1
f16	12.7	21.6	7.4	18.3	20.1
factory1	16.4	28.5	8.2	19.1	19.4
factory2	13.0	23.8	5.4	14.2	16.0
hfchannel	36.3	30.5	7.7	25.5	27.3
leopard	42.9	48.6	11.2	16.6	14.2
m109	18.8	38.0	6.5	14.9	14.7
machinegun	51.8	55.9	60.7	49.3	32.4
pink	8.6	12.5	5.1	16.0	19.2
volvo	8.1	8.9	4.0	9.2	10.8
white	12.2	15.8	6.8	21.4	26.5
all noises (mean)	23.7	32.5	13.0	22.7	20.4
feature size	25	17	26	27	54

gave a relative improvement of 20% in averaged word error rate over MFCC with dynamic feature in noisy environments. This result indicates that Modulation FT is effective not only for English digit recognition but also for Japanese isolated-word recognition using word HMMs.

PLP-RI shows results using feature described in Sec.**2. 5**. Using word HMMs, PLP-RI also gave a relative improvement of 5% in averaged word error rate over MFCC with dynamic feature in noisy environments.

3. 3 Results using tri-phone HMMs

A tri-phone model needs less training data than word model, because the unit size of tri-phone model is smaller than that of word model. Consequently, tri-phone model is often used for continuous speech recognition. Table 2 shows isolated-word recognition results using tri-phone HMMs in various noise environments. Each tri-phone HMM was modeled by 5 states (including nonemitting initial and final states), and there were 16 mixtures per states. Covariance matrics for each mixture were assumed to be diagonal.

Results of MFCC, PLP and PLP-RI in Tab. 2 were better than those in Tab. 1, while Results of Modulation FT in Tab. 2 were worse than those in Tab. 1. Dynamic features used in MFCC, PLP and PLP-RI emphasize modulation-frequency components at around 10 Hz as shown in Fig. 2. Modulation FT emphasizes modulation-frequency components at around 2.5 Hz, 5 Hz and 7.5 Hz as shown in Fig. 3. The emphasized range of modulation frequency by Modulation FT is lower than that by dynamic features. Generally, it needs longer recognition unit to model the lower range of modulation frequency. Consequently, results of Modulation FT using tri-phone model, whose unit size is shorter than that of word model, got worse. These results suggest that the capability of Modulation FT would not be displayed using short recognition unit model such as tri-phone model. Modulation FT would be useful in case using word model or sub-word model for continuous speech recognition.

The effect of short recognition unit is small in the case of dynamic features used in MFCC, PLP and PLP-RI, because the dynamic features use the higher range of modulation frequency. As the amount of training data per model is increase using short recognition unit, results of MFCC, PLP and PLP-RI using tri-phone model would be improved.

4. Continuous speech recognition based on the contribution of modulation frequency components

4. 1 Experimental setup

To evaluate features described in Sec.**2.**, Japanese dictation toolkit (1998 version) [20] was used. The conditions for training and evaluation conformed to [20]. The acoustic model was a tri-phone model with 16 mixture components per state.

The system was trained on clean speech, while the test data were degraded by additive background noise. The 15 kinds of noise was used from NOISEX-92 database [15].

Table 4 Recognition results (Word recognition accuracies [%] at SNR 10 dB)

Feature	(1) PLP+Δ	(2) Lin-Log RASTA+Δ	(3) PLP-RI +Δ	(4) MFCC +Δ	(5) MFCC -RI+Δ	(6) PLP +Δ+Δ^2	(7) PLP-RI +Δ+Δ^2	(8) MFCC +Δ+Δ^2
clean	90.2	85.8	90.3	92.1	91.4	92.6	91.4	92.6
additive noise								
babble	75.8	79.0	81.6	81.8	81.3	84.0	86.6	87.3
buccaneer1	77.0	77.4	82.5	80.1	80.9	82.6	85.9	83.8
buccaneer2	77.6	77.7	85.0	79.1	79.0	82.1	86.6	83.8
destroyerengine	68.7	73.4	75.0	77.9	80.8	77.5	84.4	85.0
destroyerops	79.5	80.6	79.8	84.1	85.9	85.8	87.2	89.2
f16	76.0	78.0	82.6	80.5	80.9	81.9	86.4	85.9
factory1	77.2	78.3	86.1	80.3	81.3	81.8	87.0	85.3
factory2	82.1	81.2	86.3	86.4	86.7	88.7	90.4	89.3
hfchannel	66.8	75.9	78.4	69.3	71.2	77.5	85.7	79.6
leopard	76.6	82.2	82.0	84.1	85.1	86.1	87.9	88.7
m109	85.3	82.0	86.3	87.5	88.8	88.6	91.0	88.8
machinegun	47.1	42.3	42.8	56.3	57.0	51.7	44.7	58.1
pink	78.9	78.4	84.0	81.5	81.3	82.4	86.3	85.1
volvo	87.2	85.2	89.0	91.6	88.9	92.8	90.8	90.9
white	71.9	75.5	81.3	73.7	74.9	78.7	86.7	80.7
all noises (mean)	75.2	76.5	80.2	79.6	80.3	81.5	84.5	84.1
feature size	17	17	17	25	25	26	26	38

Table 3 Word recognition accuracies [%]

Feature	Feature size	clean	SNR 20dB	SNR 10dB	SNR 0dB
(1) PLP+Δ	17	90.2	85.5	75.2	51.8
(2) Lin-Log RASTA+Δ	17	85.8	84.1	76.5	52.9
(3) PLP-RI+Δ	17	90.3	87.9	80.2	56.0
(4) MFCC+Δ	25	92.1	88.3	79.6	57.8
(5) MFCC-RI+Δ	25	91.4	88.4	80.3	57.9
(6) PLP+Δ+Δ^2	26	92.6	89.8	81.5	58.0
(7) PLP-RI+Δ+Δ^2	26	91.4	90.1	84.5	63.9
(8) MFCC+Δ+Δ^2	38	92.6	90.1	84.1	63.0

4. 2 Results

Table 3 and 4 show recognition results under various noise conditions for following features with CMS [21].

(1) PLP+Δ ··· 8th-order PLP cepstrum and their dynamic features Δs (with Δ of logarithmic power).

(2) Lin-Log RASTA+Δ ··· feature described in Sec.**2. 3** and its dynamic feature.

(3) PLP-RI+Δ ··· features filtered with modulation frequency characteristic shown in Fig.1 for 8th-order PLP cepstrum, and their dynamic features (proposed feature based on contribution of modulation frequency components described in Sec.**2. 5**).

(4) MFCC+Δ ··· 12th-order MFCC and their dynamic features Δs (with Δ of logarithmic power).

(5) MFCC-RI+Δ ··· features filtered with modulation frequency characteristic shown in Fig.1 for 12th-order MFCC, and their dynamic feature.

(6) PLP+Δ+Δ^2 ··· feature (1), and its Δ^2.

(7) PLP-RI+Δ+Δ^2 ··· proposed feature (3), and its Δ^2.

(8) MFCC+Δ+Δ^2 ··· feature (4), and its Δ^2.

In Table 3, clean shows word accuracy without noise, and SNR (20dB, 10dB, 0dB) shows mean of word accuracies for the speech degraded by 15 kinds of noise with corresponding SNR. Table 4 shows word accuracies for the speech degraded by 15 kinds of noise at SNR 10dB.

4. 3 Discussion

Lin-Log RASTA extracts the modulation frequency components between 1 and 12 Hz from PLP. Comparing PLP and Lin-Log RASTA in Table 3, Lin-Log RASTA is better than PLP in noisy environments below SNR 10 dB, while PLP is better than Lin-Log RASTA in clean environments. The results suggest modulation filtering methods such as Lin-Log RASTA is effective in noisy environments. This is consistent with Arai's research [4], [5] which suggests that important modulation frequency components for speech perception exist in the range between 1 and 16 Hz, while the modulation frequency components of noise are distributed in broad range. The modulation filtering methods take advantage in noisy environments by keeping the important modulation frequency components for speech recognition while alleviating most of modulation frequency components of noise.

In case feature size is 17, proposed feature (3) obtained best performance. The results show modulation filtering based on the contribution of modulation frequency components is effective.

Feature (5) filtered with modulation frequency characteristics shown in Fig.1 for MFCC outperforms MFCC(6) in noisy environments. Proposed feature (7) in noisy environments (SNR 10 dB) gave a relative improvement of 5 % in word accuracy over the MFCC(6) with dynamic feature. These results show the proposed feature based on contribution of modulation frequency components take advantage for continuous speech recognition in noisy environments.

5. Conclusions

we proposed new feature for automatic speech recognition based on the contribution of modulation frequency components. The modulation filtering methods take advantage in noisy environments by keeping the important modulation frequency components for speech recognition while alleviating most of modulation frequency components of noise. The results show modulation filtering based on the contribution of modulation frequency components is effective for continuous speech recognition.

Acknowledgement

This work was partially supported by the Ministry of Education, Culture, Sports, Science and Technology in Japan, Grant-in Aid for Scientific Research (C). 14580246.

References

[1] 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. Amer., Vol. 77, pp. 1069 – 1077, 1985.

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

[3] R. Drullman, J. M. Festen, and R. Plomp, "Effect of reducing slow temporal modulations on speech reception," J. Acoust. Soc. Amer., Vol. 95, pp. 2670 – 2680, 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, Philadelphia, pp. 2490 – 2493, 1996.

[5] T. Arai, M. Pavel, H. Hermansky and C. Avendano, "Syllable intelligibility for temporally filtered LPC cepstral trajectories." J. Acoust. Soc. Amer.. Vol. 105, No. 5, pp. 2783 – 2791, 1999.

[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.

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

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

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

[10] N. Kanedera, T. Arai, H. Hermansky, and M. Pavel, "On the importance of various modulation frequencies for speech recognition," Proc. Eurospeech, Rhodes, Greece, pp. 1079 – 1082, 1997.

[11] N. Kanedera, H. Hermansky and T. Arai, "On properties of modulation spectrum for robust automatic speech recognition," Proc. IEEE ICASSP, Seattle, WA, pp. II-613 – II-616, 1998.

[12] N. Kanedera, T. Arai, H. Hermansky, and M. Pavel, "On the relative importance of various components of the modulation spectrum for automatic speech recognition," Speech Communication, Vol.28, pp.43 – 55, 1999.

[13] N. Kanedera, T. Arai, H. Hermansky, and T. Funada, "On properties of the modulation spectrum for robust automatic speech recognition (in Japanese)," Technical report of IEICE, SP97-70, pp.15 – 22, 1997.

[14] N. Kanedera, T. Arai, and T. Funada, "On the robustness of automatic speech recognition using multi-resolution modulation spectrum (in Japanese)," Technical report of IEICE, SP98-51, pp.45 – 52, 1998.

[15] A. Varga and H. J.M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," Speech Communication, Vol. 12, No. 3, pp. 247 – 251, 1993.

[16] N. Kanedera, T. Arai, and T. Funada, "Investigations on information of speech recognition and speaker identification in modulation spectrum (in Japanese)," Technical report of IEICE, SP2000-34, Vol.100, No.240, pp.15 – 22, 2000.

[17] N. Kanedera, T. Arai, and T. Funada, "Robust automatic speech recognition emphasizing important modulation spectrum (in Japanese)," IEICE Trans. D-II, Vol.J84-D-II, No.7, pp.1261 – 1269, 2001.

[18] N. Kanedera, T. Arai, K. Okada, and Y. Momomura, "Continuous speech recognition based on the contribution of modulation frequency components (in Japanese)," Technical report of IEICE, SP2002-64, pp.41 – 46, 2002.

[19] Y. Momomura, K. Okada, T. Arai, N. Kanedera, Y. Murahara, "Using the modulation complex wavelet transform for feature extraction in automatic speech recognition," Proc. Eurospeech, Vol.4, pp.2639 – 2642, 2001.

[20] T. Kawahara, A. Lee, T. Kobayashi, K. Takeda, N. Minematsu, K. Itou, M. Yamamoto, A. Yamada, T. Utsuro, and K. Shikano, "Japanese Dictation Toolkit – 1998 version – (in Japanese)," J. Acoust. Soc. Jpn., **56**, 4, pp. 255 – 259, 2000.

[21] B. S. Atal, "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, 1974.

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

[23] V. Johan, B. Louis, "Channel normalization techniques for automatic speech recognition over the telephone," Speech Communication, Vol.25, pp.149 – 164, 1998.

[24] H. Kuwabara, Y. Sagisaka, K. Takeda, M. Abe, "Construction of ATR Japanese speech database as a research tool (in Japanese)," ATR Technical Report, TR-I-0086, 1989.

ミツバ綜合印刷株式会社