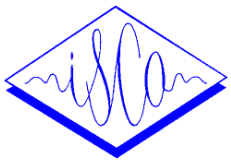
ISCA Archive
http://www.isca-speech.org/archive



3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

NOISE ROBUST SPEECH RECOGNITION USING A DYNAMIC-CEPSTRUM

Kiyoaki Aikawa [1] *and* *Tsuyoshi Saito* [2]

[1] ATR Human Information Processing Research Laboratories
2-2 Hikaridai, Seikacho, Sorakugun, Kyoto 619-02 Japan
[2] Toyohashi Univ. of Technology
1-1 Hibarigaoka, Tempakucho, Toyohashishi, Aichi 441 Japan

ABSTRACT

Noise robust speech recognition is achieved using a dynamic-cepstrum. The dynamic-cepstrum is a new spectral representation incorporating time-frequency forward masking. The time-frequency masking can suppress the spectral components commonly included in the current spectrum and in the preceding spectra. This feature suggests the applicability of the dynamic-cepstrum to noisy speech recognition. Speaker-dependent and speaker-independent phoneme recognition experiments are conducted using hidden Markov models. Experimental results demonstrate that the dynamic-cepstrum outperforms the conventional cepstrum on robustness against stationary noise and amplitude-modulated noise. The dynamic-cepstrum is also found to be superior to the conventional cepstrum combined with a delta-cepstrum.

1. INTRODUCTION

Time-frequency two-dimensional masking characteristics were first investigated through perception experiments [1]. Following this, a new spectral representation incorporating time-frequency masking and called the dynamic-cepstrum was proposed [2]. Excellent recognition performance has been obtained by applying the dynamic cepstrum to continuous speech recognition [3].

The time-frequency masking mechanism produces a masking pattern by accumulating frequency-smoothed preceding spectra. By subtracting the masking pattern from the current spectrum, the spectral components commonly included in the current speech spectrum and the preceding spectra are suppressed. This feature suggests the applicability of the dynamic-cepstrum to noisy speech recognition.

This paper evaluates the robustness of dynamic-cepstrum against additive stationary and amplitude-modulated noises. The dynamic-cepstrum is compared with the conventional cepstrum for speaker-dependent and speaker-independent phoneme recognition. This paper also reports the results when the delta-cepstrum [4] is additionally used.

2. DYNAMIC-CEPSTRUM

The time-frequency two-dimensional masking model is constructed based on the perception experimental results reported in [1]. The forward masking pattern evoked by a pure tone is relatively sharp at the masker frequency immediately after the end of the masker. The masking pattern loses its sharpness as a function of elapsed time after the masker is turned off.

The mechanism producing the masking pattern is modeled by spectral smoothing which depends on the masker-signal time interval [2]. In this model, a perceived spectrum $P(\omega, i)$ at time i is obtained by subtracting a masking pattern $M(\omega, i)$ from a spectrum $S(\omega, i)$. The perceived spectrum is given by

$$P(\omega, i) = S(\omega, i) - M(\omega, i) \tag{1}$$

where ω is the frequency and the intensity of the spectrum is represented in logarithm for simulating perceptual intensity.

Smoothed preceding spectra are accumulated into masking pattern $M(\omega, i)$. Let $h(\lambda, n)$ denote the impulse response of the spectral smoothing lifter as a function of frequency λ and time n. Then, the masking pattern is obtained by

$$M(\omega, i) = \sum_{n=1}^{N} \int_{-\infty}^{\infty} S(\omega - \lambda, i - n) h(\lambda, n) d\lambda \tag{2}$$

where N is the maximum masking duration.

The dynamic-cepstrum is defined by the inverse Fourier transform of perceived spectrum $P(\omega, i)$ and given by

$$b_k(i) = c_k(i) - m_k(i) \tag{3}$$

where $c_k(i)$ is kth cepstrum coefficient, and $m_k(i)$ is the inverse Fourier transform of the masking pattern; $m_k(i)$ is called the masking coefficient.

The masking coefficient is obtained by multiplying lifter gains to cepstral coefficients in the quefrency domain. The kth masking coefficient $m_k(i)$ at time i is given by the sum of the weighted preceding cepstrum coefficients as

$$m_k(i) = \sum_{n=1}^{N} c_k(i - n) l_k(n) \tag{4}$$

where $l_k(n)$ is the kth lifter gain for smoothing the spectrum n frames before and the inverse Fourier transform of the impulse response $h(\lambda, n)$.

A Gaussian lifter gain function given by

$$l_k(n) = \alpha \beta^{n-1} \exp\left(-\frac{k^2}{2(q_0 - \nu(n-1))^2}\right) \tag{5}$$

provides a good speech recognition performance [3]. q_0 denotes the standard deviation of the Gaussian lifter gain, and ν its decreasing rate per frame. For simulating the masking pattern decay reported in [1], the standard deviation of the Gaussian lifter gain decreases as a function of the time delay n.

3. MODULATION FREQUENCY RESPONSE

The modulation frequency is defined by the frequency of the temporal change of a cepstral coefficient. The dynamic-cepstrum coefficients show different response characteristics to the modulation frequency depending on the order "k". Figure 1 shows the difference of transfer functions for the 1st, 8th and 16th dynamic-cepstrum coefficient. The values at the origin of the horizontal axis indicate the responses to stationary input. The maximum modulation frequency is the Nyquist frequency determined by the frame sampling rate.

This figure indicates that slow cepstrum change is more suppressed than rapid change, and low quefrency components of the spectrum are more suppressed than high quefrency components. This feature implies that the dynamic cepstrum is capable of attenuating stationary or slowly changing wide band noises.

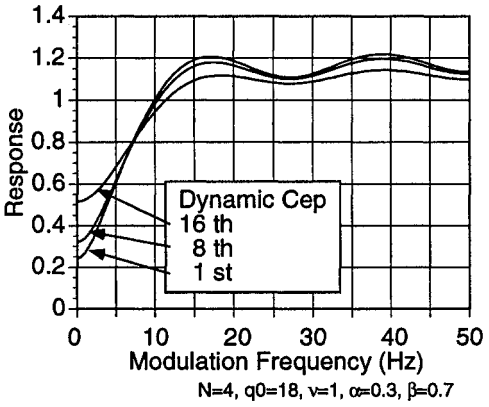


Figure 1. Order-dependent transfer function of each dynamic-cepstral coefficient.

4. NOISE

4.1. Noisy Speech Database

A noisy speech database can be produced by adding white or colored noise to clean speech waves at a specified S/N ratio. The noise amplitude is determined depending on the average power of each speech portion. White Gaussian noise is obtained by converting computer-generated uniform random noise.

As an example of colored noise, pink noise is used. Pink noise is characterized by a -3dB/oct power spectrum decay and can be produced by filtering white Gaussian noise. The power spectrum of pink noise diverges at zero frequency. In this paper, the transfer function in the frequency range less than the threshold ϵ is fixed to a finite value.

The transfer function of the pink noise filter is given by

$$H(\omega) = \begin{cases} \sqrt{\frac{1}{|\omega|}} & \text{if } |\omega| > \epsilon \\ \sqrt{\frac{1}{\epsilon}} & \text{otherwise} \end{cases} \quad (-\pi < \omega \leq \pi). \qquad (6)$$

Pink noise $y(t)$ is obtained from white noise $x(t)$ by

$$y(t) = \sum_{\tau=-L}^{L} x(t-\tau)g(\tau) \qquad (7)$$

where $g(\tau)$ denotes a two-sided symmetric impulse response of the pink noise filter and L denotes its half length. In this paper L is 256 and the threshold ϵ is $\pi/256$.

4.2. Modulated Noise

The dynamic-cepstrum emphasizes spectral change. Then, the speech recognition performance is tested against non-stationary noises. Amplitude-modulated white and pink noises are used as the non-stationary noises. A noisy speech signal is produced by

$$r(t) = s(t) + G\left(1 + \frac{d}{100}\sin(2\pi f t)\right)x(t). \qquad (8)$$

where f and d denote the modulation frequency and modulation depth, respectively. The constant G controls the S/N ratio, $x(t)$ is the stationary noise signal, and $s(t)$ is the speech signal.

5. EXPERIMENT

5.1. Database

The speech database included 5240 common Japanese words (word utterances) and 115 sentences spoken by 10 male and 10 female speakers. A pause was inserted between phrases in the sentence database (phrase utterances). The sampling frequency was 12 kHz, the frame rate was 10 ms, and the frame window size was 30 ms. The 16th order cepstrum was calculated through linear prediction analysis. The masking parameters were fixed to the following values: masking duration $N = 4$ frames, initial standard deviation $q_0 = 18$, standard deviation decreasing rate $\nu = 1$, initial masking decay $\alpha = 0.3$, and medial masking decay $\beta = 0.7$.

Continuous density 3-state hidden Markov models (HMMs) were used for the phoneme models[5]. The probability distribution was represented by a Gaussian mixture, and the number of mixtures was eight.

The dynamic-cepstrum and the conventional cepstrum were compared in phoneme recognition for four categories: 23 phonemes, 18 consonants, 6 consonants, and 5 vowels. The 18-consonant set included /b, d, g, m, n, N, p, t, k, s, h, z, r, y, w, ch, ts, sh/, the 6-consonant set included /b, d, g, m, n, N/, and the 5-vowel set included /a, i, u, e, o/. The 23-phoneme set included the 18 consonants and 5 vowels. The HMMs were trained with clean speech and tested for noisy speech.

5.2. Speaker-Dependent Recognition

Speaker-dependent phoneme recognition experiments were carried out for a male speaker. Training samples were collected from 2620 common words. Testing samples were collected from noise-added and different 2620 words and phrases.

5.2.1. Stationary Noise

Noisy speech was produced at three S/N ratios: 10 dB, 20 dB and 30 dB. Figures 2 and 3 compare 23-phoneme recognition results for the three S/N ratios. "clean" means the case without noise addition. Figure 2 gives the results for word utterance and Figure 3 for phrase utterance. "DyC" denotes the dynamic-cepstrum and "Cep" the conventional cepstrum. These figures show a clear improvement of the recognition rates by the dynamic-cepstrum.

Figure 4 shows 23-phoneme recognition results when pink noise was added to the phrase utterance data. The results are similar to those with white noise.

Figure 5 shows the results when delta-cepstrum is additionally used. Even with the addition, the dynamic-cepstrum is superior to the conventional cepstrum.

5.2.2. Modulated Noise

The performance of the dynamic-cepstrum was examined for various amplitude-modulated noises. Table 1 shows the effect of modulation depth. The S/N ratio was 20 dB and the modulation frequency was 10 Hz. The effect

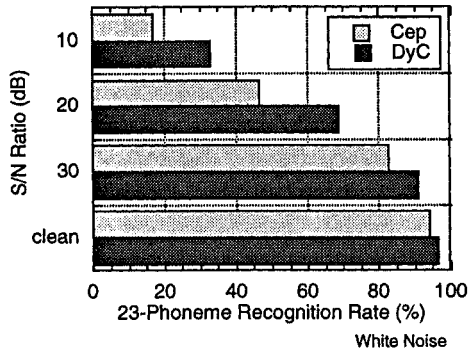


Figure 2. 23-phoneme recognition results for word utterance (white noise added).

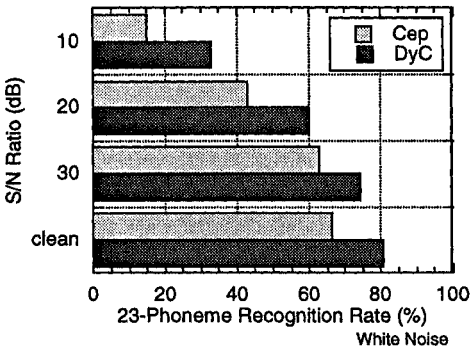


Figure 3. 23-phoneme recognition results for phrase utterance (white noise added).

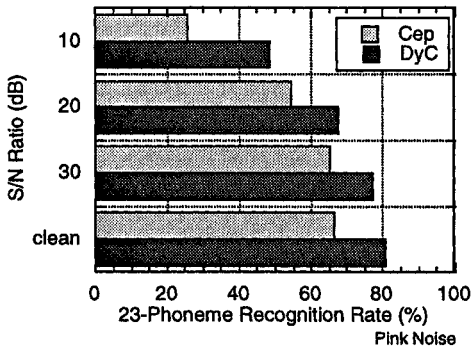


Figure 4. 23-phoneme recognition results for phrase utterance (pink noise added).

of the modulation depth is very small for the recognition results obtained using the dynamic-cepstrum.

The effect of modulation frequency was examined under the following conditions: S/N ratio of 20 dB and modulation depth of 50%. Figure 6 shows the 23-phoneme recognition results for noise-added phrase utterances. The recognition rate decreases as the modulation frequency increases until 40 Hz. The decrease in the recognition rate recovers a little bit at the modulation frequency of 50 Hz. This is because the modulation cycle has become less than the frame window size.

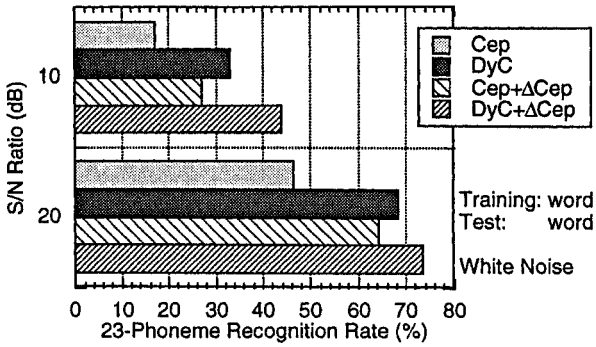


Figure 5. 23-phoneme recognition results for phrase utterance in the case that a delta-cepstrum is additionally used (white noise added).

Table 1. Effect of modulation depth on phoneme recognition. (S/N 20dB, Modulation Frequency 10Hz, White Noise)

Category		word		phrase	
		50%	100%	50%	100%
23 Phonemes	Cep	49.6	57.7	43.9	47.3
	DyC	70.3	70.5	60.6	58.0
18 Consonants	Cep	44.3	53.1	41.0	45.4
	DyC	66.1	66.7	54.9	52.4
6 Consonants	Cep	57.3	65.2	57.0	60.3
	DyC	71.6	74.8	63.6	62.7
5 Vowels	Cep	92.9	94.3	78.6	79.0
	DyC	96.0	96.8	83.6	83.8

5.3. Speaker-Independent Phoneme Recognition

Speaker-independent phoneme recognition was carried out for 10 male and 10 female speakers. Speaker-independent phoneme HMMs were trained with the voices of nine speakers of the same gender excluding the testing speaker. This experiment was repeated 10 times, changing the testing speaker each time. Training and testing phoneme samples were collected from a phonetically balanced 216-word database. Stationary white noise was added to the testing samples at the S/N ratio of 20 dB.

Figure 7 shows the speaker-independent 23-phoneme recognition results for the male speakers. The term "Ave" indicates the average of the 10 speakers. Figure 8 shows the results for the female speakers. These figures

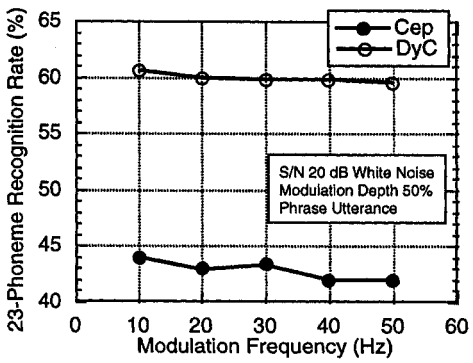


Figure 6. Relationship between modulation frequency and 23-phoneme recognition rate for phrase utterance.

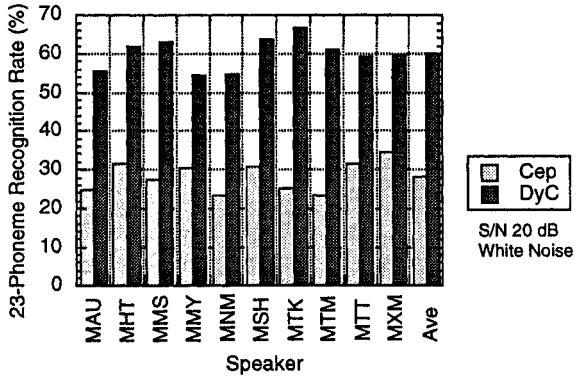


Figure 7. Speaker-independent phoneme recognition results for male speakers.

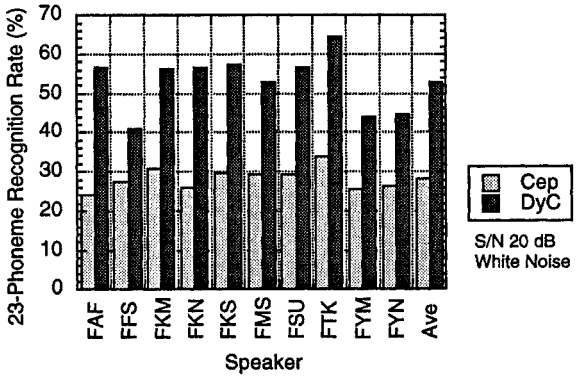


Figure 8. Speaker-independent phoneme recognition results for female speakers.

show the superiority of the dynamic-cepstrum on speaker-independent noisy speech recognition. The recognition error was reduced to 3/5 of that by the conventional cepstrum.

Figure 9 shows the speaker independent 23-phoneme recognition results for male speakers when the dynamic-cepstrum or the conventional cepstrum was used in combination with the delta-cepstrum. Figure 10 shows the results for female speakers under the same conditions. The dynamic-cepstrum provides better performance over the conventional cepstrum even when used in combination with the delta-cepstrum.

6. CONCLUSIONS

The dynamic-cepstrum outperformed the conventional cepstrum on noisy speech recognition. The speaker-dependent and speaker-independent phoneme recognition results demonstrated that the dynamic-cepstrum is robust against additive white and pink noises, even if the noise is amplitude-modulated. The dynamic-cepstrum is superior to the cepstrum even when a delta-cepstrum is additionally used.

ACKNOWLEDGMENT

The authors would like to thank Dr. Yoh'ichi Tohkura of ATR and Prof. Tatsuo Yoshida of Toyohashi University of Technology for providing us the opportunity to do this cooperative research, and Dr. Hideki Kawahara for his valuable suggestions on noise generation.

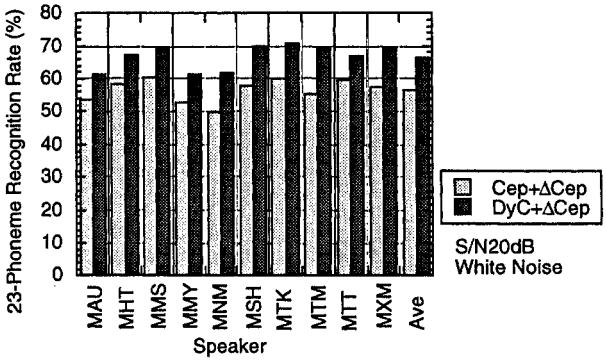


Figure 9. Speaker-independent phoneme recognition results for Cep/DyC in combination with ΔCep for male speakers.

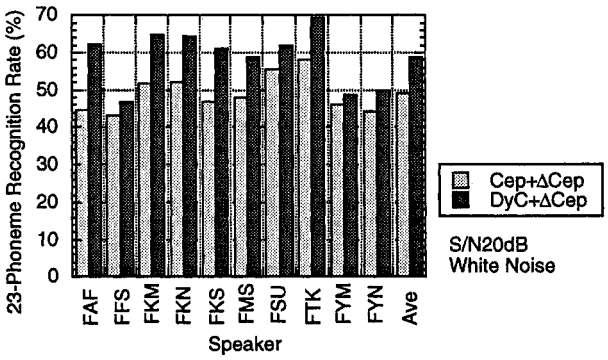


Figure 10. Speaker-independent phoneme recognition results for Cep/DyC in combination with ΔCep for female speakers.

REFERENCES

[1] E. Miyasaka, "Spatio-temporal characteristics of masking of brief test-tone pulses by a tone-burst with abrupt switching transients", *J. Acoust. Soc. Jpn*, vol. 39, no. 9, pp. 614–623 (in Japanese) (1983).

[2] K. Aikawa, H. Kawahara and Y. Tohkura, "Dynamic cepstral parameter incorporating time-frequency masking and its application to speech recognition", *J. Acoust. Soc. Am.*, vol. 92, no. 4, Pt.2, pp. 2476 (Oct.1992).

[3] K. Aikawa, H. Singer, H. Kawahara and Y. Tohkura, "A dynamic cepstrum incorporating time-frequency masking and its application to continuous speech recognition", *Proc. ICASSP'93*, vol. II, pp. 668–671 (Apr.1993).

[4] S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum", *IEEE Trans.*, vol. ASSP-34, no. 1, pp. 52–59 (1986-02).

[5] P. F. Brown: "The acoustic-modeling problem in automatic speech recognition", PhD thesis, Carnegie-Mellon University (1987).