SPEAKER-INDEPENDENT SPEECH RECOGNITION USING MICRO SEGMENT SPECTRUM INTEGRATION

Kiyoaki Aikawa

NTT Human Interface Laboratories, Speech and Acoustics Laboratories
1-1 Hikarino-oka, Yokosuka-shi, Kanagawa 239-0847 Japan
aik@nttspch.hil.ntt.co.jp

ABSTRACT

This paper proposes a new spectral estimation method for automatic speech recognition. The spectrum estimated with the conventional data window of around 30 ms shows harmonic structure in the voiced portions of speech data. The harmonic frequency interval is often comparable to the formant frequency interval for female voices with high F0, which results in spectral estimation error. The new idea is to estimate spectrum by taking the Lp norm of the time series of the spectrum obtained from a very short speech segment. The new method, called the micro-segment spectrum integration, provides (1) precise spectral estimation not affected by harmonic structure, and (2) noise-robustness by suppressing noisy speech segments. Phoneme recognition experiments demonstrate that the micro-segment spectrum integration method outperforms conventional spectral estimation methods.

1. INTRODUCTION

Recent automatic speech recognition (ASR) techniques estimate the speech spectrum using a speech frame with length of around 30 ms. This window length was determined so as to stabilize spectral estimation by covering more than two pitch periods. The spectrum obtained in this way shows harmonic structure in voiced regions because of the repeated wave form in the data window. The harmonic frequency-interval is comparable to the formant frequency interval for high F0 (fundamental frequency) voices such as those of females and children. This makes it difficult to correctly locate the real formants. Therefore, spectral envelope estimation accuracy is reduced for such high F0 voices.

The human auditory system has much higher temporal resolution than conventional ASR systems. The temporal resolution depends on the frequency channel that is shown in the gammatone filter model [1]. There is a trade-off between the temporal resolution and frequency resolution. This paper addresses the problem of extracting phonetic information using high temporal resolution without losing the reliability of spectral estimation.

This paper proposes a novel spectral estimation method for automatic speech recognition that is based on integrating small pieces of spectrum using Lp norm. If the data window length is shorter than the pitch period, the estimated spectrum can avoid the effect of the harmonic structure formed by the repeated speech signal. Pitch synchronous spectral analysis is a method which provides a spectrum unaffected by the harmonic structure.

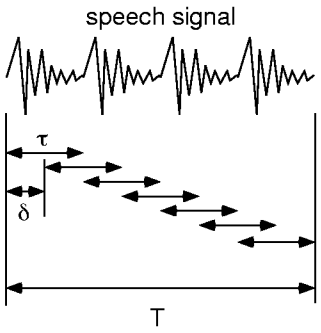


Figure 1. Micro-frame sequence. T is the conventional data window size for speech spectral analysis.

2. MICRO SEGMENT SPECTRUM

Typical speech recognition schemes extract spectra using a data window of around 30 ms wide. Fundamental frequencies mostly distribute between 80 Hz and 350 Hz. Pitch pulse interval is 12.5 ms for F0 = 80 Hz and 2.9 ms for F0 = 350 Hz. The data window size is chosen to cover at least two pitch period even for a possible lowest F0.

If the data window size is shorter than the pitch period, the estimated spectrum does not show harmonic structure. Pitch synchronous analysis is a special case that adaptively employs the same the data window size as the pitch period. This paper uses a data window that is shorter than the pitch period. The speech signal does not show periodicity within such a short data window. This short data window is called the micro frame and the spectrum estimated from the short segment of speech is called the micro-segment spectrum (MSS).

A speech frame of length T is divided into N short speech segments with overlap between adjacent segments. T is the typical frame length used in existing speech recognizers. Any spectral estimation method can be used for obtain-

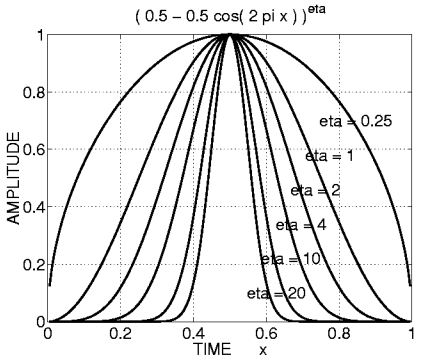


Figure 2. Sinusoidally changing frequency channel output enhanced by power of η.

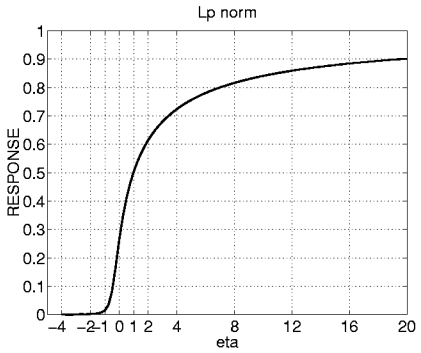


Figure 3. Lp norm of the sinusoidal signal.

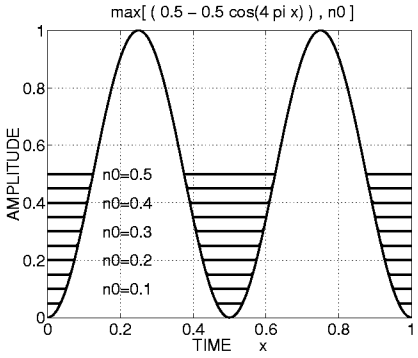


Figure 4. Sinusoidal frequency channel output partially buried in noise.

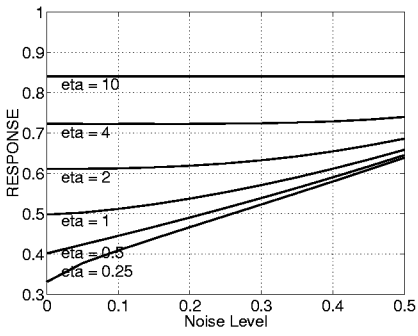


Figure 5. Micro Segment Spectral Integration value for the frequency channel output corrupted by a stationary noise.

ing MSS. Figure 1 illustrates how MSS is obtained from a speech frame T. τ, δ, and T denote micro frame width, micro frame shift, and frame length, respectively. Number of micro frames in a speech frame is

$$N = \frac{T-\tau}{\delta} + 1. \quad (1)$$

If MSS are simply summed, a conventional spectrum would be produced. If the squared MSS are summed, however, a peak-weighted sum of spectra is produced. A method for integrating the MSS using Lp norm is proposed to formulate a generalized sum of the MSS. The method is called the Micro Segment Spectrum Integration (MSSI). The MSSI spectrum is given by

$$S(\omega,t) = \{\frac{1}{N}\sum_{i=0}^{N-1} M(\omega,t+\delta i)^{\eta}\}^{1/\eta}. \quad (2)$$

where $M(\omega,t+\delta i)$ denotes the i-th MSS in the speech frame beginning at time t and must be positive. ω is the angular frequency, and η denotes the power factor of the Lp norm. The GPD (Generalized Probabilistic Descent) method uses the Lp norm for MCE (Minimum Classification Error) training [2].

Above formulation integrates the frequency channel output sequences separately for each center frequency. A simplified way to integrate MSS is to accumulate the MSS with weighting by the powered log-energy. Let the energy and the energy-normalized spectrum of a micro frame be $u(t)$ and $Q(\omega,t)$, the MSSI spectrum can be computed by

$$u(t+\delta i) = \int_{-\pi}^{\pi} M(\omega,t+\delta i)d\omega \quad (3)$$

$$Q(\omega,t+\delta i) = M(\omega,t+\delta i)/u(t+\delta i) \quad (4)$$

$$S(\omega,t) = \frac{\sum_{i=0}^{N-1} u(t+\delta i)^{\eta} Q(\omega,t+\delta i)}{\sum_{i=0}^{N-1} u(t+\delta i)^{\eta}} \quad (5)$$

This equation sums the MSS with the weight of $u(t+\delta i)^{\eta}$.

3. MSSI FEATURES

The MSSI can represent a variety of functional features by controlling the factor η. When $\eta = 1$, Eq. (2) yields the arithmetic average of the MSS sequence. When $\eta = 1$, Eq. (2) yields the root mean square. When $\eta = -\infty$ and $\eta = \infty$, Eq. (2) outputs the minimum and maximum values of the MSS sequence, respectively.

High energy portions in the sequence of a MSS frequency channel output are emphasized when $\eta > 1$, and low energy portions are emphasized when $\eta < 1$. Therefore, η should be greater than 1 to suppress the low energy MSS which may be corrupted by noise. On the other hand, η should be less than 1 for suppressing high energy burst noise.

Figure 2 shows a frequency channel output raised to the power of η where the output temporally changes as

$$M(i) = 0.5 - 0.5\cos(2\pi i/N) + \varepsilon \quad (6)$$
$$1 < i < N$$

where ε is a small constant that prevents divergence of the Lp norm. Figure 3 shows the summed MSS using Lp norm as

$$F(\eta) = \{\frac{1}{N}\sum_{i=1}^{N} M(i)^{\eta}\}^{1/\eta}. \quad (7)$$

When $\eta = 1$, $F(\eta)$ outputs the value 0.5, the average of $M(i)$. In this paper, the range of η is limited to $0 < \eta$ because speech information is concentrated at the high energy portions of the spectrogram.

Figure 4 illustrates the change in the MSS frequency channel output whose valleys are buried by a stationary noise with ten noise levels. The noise-corrupted MSS sequence is given by

$$M'(i) = \max[M(i), n_0] \quad (8)$$
$$n_0 = (0.05, 0.1, \ldots, 0.5) \quad (9)$$

where n_0 is a stationary noise. Figure 5 shows the MSSI of noise-buried spectral sequences. η is varied from 0.25 to 10. The larger the η is, the more the effect of additive noise is suppressed. When $\eta > 4$, Eq. (7) output is almost independent of the noise level n_0.

4. MSS ESTIMATION

4.1. Power Spectrum method

MSS can be obtained by any spectral estimation method such as DFT or LPC (linear predictive coefficient). DFT is a common spectral estimation method. If the power spectrum of a micro-frame is denoted as $P(k,i)$, MSS is given by

$$M(k,i) = \log[1 + P(k,i)] \quad (10)$$

where k and i are the frequency channel number and MSS serial number in a frame, respectively. Equation (10) is close to a log power spectrum when the amplitude is large, and is close to a linear spectrum when the amplitude is small. The spectral value is always positive.

4.2. Autocorrelation method

LPC-based MSS can be obtained by integrating the autocorrelation coefficients obtained from a micro frame. This method is an extension of Eq. (5). The micro segment autocorrelation coefficients are summed with the weighting of log-energy raised to the power of η. Given the kth autocorrelation coefficient $r(k, t+\delta i)$ at micro frame i, the weighted sum of the micro autocorrelation coefficient is given by

$$R_k(t) = \frac{\sum_{i=0}^{N-1} v(t+\delta i)^{\eta} r_k(t+\delta i)}{\sum_{i=0}^{N-1} v(t+\delta i)^{\eta}} \quad (11)$$

$$v(t+\delta i) = \log[1 + u(t+\delta i)]. \quad (12)$$

Given the autocorrelation coefficient $R_k(t)$, standard LPC analysis can be used to obtain the LPC spectrum and LPC cepstrum. Equation (11) calculates the weighted sum of autocorrelation coefficients and emphasizes high energy MSS.

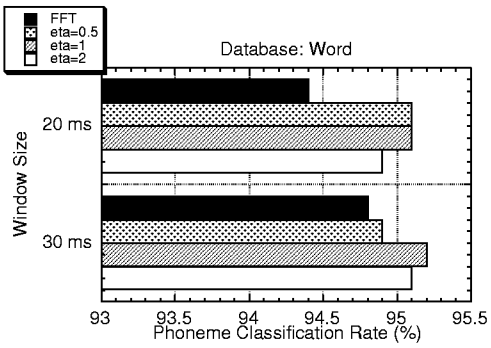


Figure 6. Comparison between a DFT and three types of micro-segment spectrum for word database.

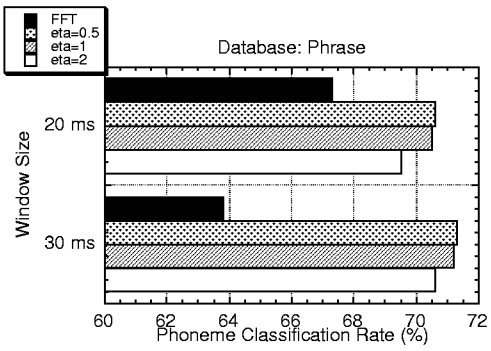


Figure 7. Comparison between a DFT and three types of micro-segment spectrum for phrase database.

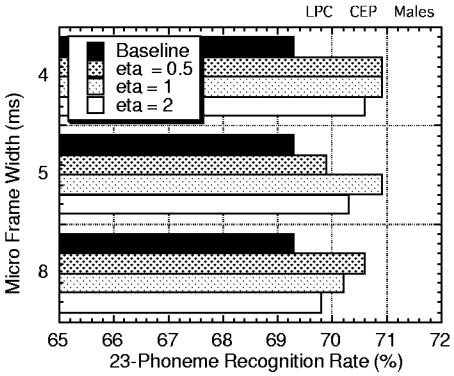
5. EXPERIMENT

Speaker-dependent and speaker-independent phoneme recognition experiments were carried out for 23 Japanese phonemes including 18 consonants and 5 vowels. The sampling frequency was 12 kHz. A three-state left to right continuous distribution HMM was used as the phone model. The output probability was represented with an 8 Gaussian mixture. Common frame width was 30 ms, and frame shift was 10 ms. Each MSSI spectrum was represented in cepstrum form in the speech recognizer. The Hamming window was used for obtaining MSS.

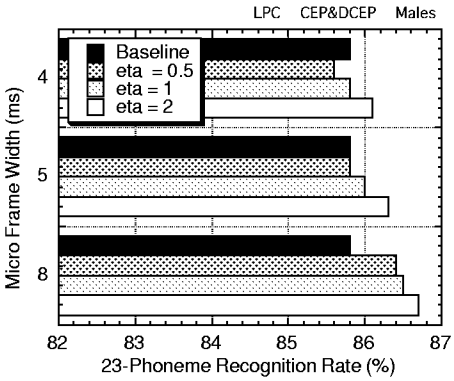
5.1. Speaker-Dependent recognition

A speaker-Dependent phoneme recognition experiment was conducted using a database that included 5240 Japanese common words and 274 phrases spoken by a male speaker. HMMs were trained with a 2620-word set and tested with the other 2620-word set and a 274-phrase set. MSSI spectrum was estimated with DFT-based method.

Performance was compared for various combinations of η and frame size T. Figures 6 and 7 show speaker-dependent 23-phoneme recognition results using DFT for MMS estimation. Figure 6 shows the result for the word database and Figure 7 shows the result for the phrase database. MSSI method shows better performance than conventional DFT based spectral estimation. The recognition performance is stable for all combinations of η and frame size.

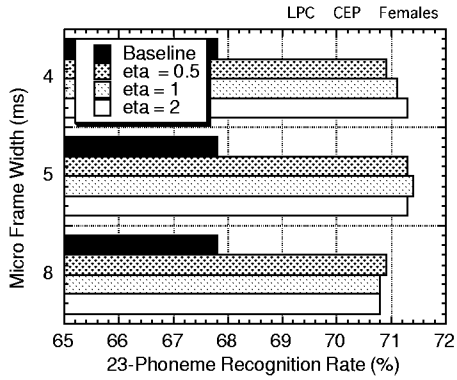


(a) Cepstrum only

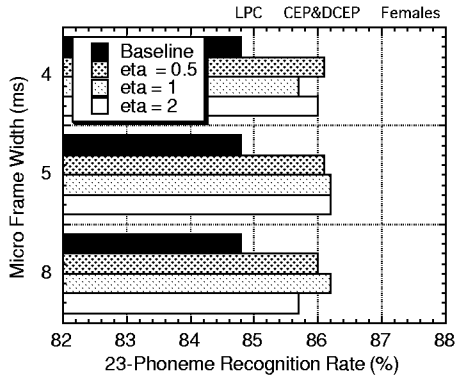


(b) Cepstrum and Delta-Cepstrum.

Figure 8. Performance comparison of three types of micro-segment spectrum integration and a conventional LPC-based spectrum (baseline) for male voice.



(a) Cepstrum only



(b) Cepstrum and Delta-Cepstrum

Figure 9. Performance comparison of three types of micro-segment spectrum integration and a conventional LPC-based spectrum (baseline) for female voice.

5.2. Speaker-Independent recognition

A speaker-independent experiment was conducted using a database that included phonetically balanced 216 Japanese word sets spoken by 10 male and 10 female speakers. Phone HMMs were created as gender models. HMMs were trained with nine speakers and tested for the speaker who was not included in the training set. The test speaker was changed over the 10 speakers of the same gender.

MSS was obtained by LPC-based method. Phoneme recognition rates were compared with those achieved by the baseline method using log LPC spectra. MSSI analysis reduced the phoneme recognition error by 10 % compared with conventional spectral estimation methods.

Various combinations of micro frame width τ, micro frame shift δ, and η were tested. The micro frame widthes were 4, 5, and 8ms. The micro frame shift was 2ms. η was set at 0.5, 1, and 2. LPC analysis order and cepstrum order was 16. The Delta-cepstrum window size was 70 ms. Phoneme recognition rates were compared with that of the conventional LPC cepstrum method.

Figures 8 and 9 show 23-phoneme recognition results for males and females, respectively. In each figure, (a) is the result with cepstrum only and (b) is the result with cepstrum and delta-cepstrum. These figures show that MSSI spectrum offers better phoneme recognition performance than the conventional LPC spectrum. Typical recommended micro frame width was 5ms and η should be 1 or 2.

6. CONCLUSIONS

The micro-segment spectrum integration (MSSI) method has been proposed for precise spectrum estimation. The MSSI spectrum is obtained by the Lp norm of a sequence of short segment spectra. The proposed method is stable even when estimating the speech spectrum with high fundamental frequency. Since each MSS was obtained from a short period of the spectrum, it can reduce the spectral estimation error caused by harmonic structure. Speaker-dependent and speaker-independent phoneme recognition experiments demonstrated that the MSSI method achieves higher performance than conventional spectral estimation based on LPC or DFT.

REFERENCES

[1] R. D. Patterson, K. Robinson, J. Holdsworth, D. Mckeown, C. Zhang and M. Allerhand, "Complex sounds and auditory images", *Auditory Physiology and Perception (Ed. Y. Cazals et al., Pergamon Press, Oxford)*, pp. 429–446 (1992).

[2] S. Katagiri, C.-H. Lee and B.-H. Juang, "New discriminative training algorithms based on the generalized probabilistic descent method", *Proc. IEEE NN-SP Workshop*, pp. 299–308 (1991).