A Method of Extracting Time-Varying Acoustic Features Effective for Speech Recognition

Kazuyo Tanaka and Hiroaki Kojima

Machine Understanding Division, Electrotechnical Laboratory, AIST
1-1-4 Umezono, Tsukuba, Ibaraki 305, Japan
ktanaka@etl.go.jp

ABSTRACT

Feature extraction plays a substantial role in automatic speech recognition systems. In this paper, a method is proposed to extract time-varying acoustic features that are effective for speech recognition. This issue is discussed from two aspects: one is on speech power spectrum enhancement and the other is on discriminative time-varying feature extraction which employs subphonetic units, called demiphonemes, for distinguishing non-steady labels from steady ones. We confirm its potential by applying it to spoken word recognition. The results indicate that recognition scores are improved by using the proposed features, compared with those using ordinary features such as delta-mel-cepstra provided by a well-known software tool.

1. INTRODUCTION

A great deal of efforts have been paid in developing new feature extraction techniques for speech recognition. In the 1980s, several methods were proposed which used time pattern features in HMM (Hidden Markov Model)-based speech recognition[1,2]. But these methods, except those using so-called delta-cepstra(or mel-cepstra)[3], the time patterns showed little effect on recognition scores. As a result, the cepstrum (or mel-cepstrum) and delta-cepstrum of speech wave form became a standard-like feature set for major speech recognition systems, and after that, noticeable progress could not be found in this area.

On the other hand, preceding such works we applied time-varying acoustic patterns and power spectrum enhancement to phoneme classification and proposed a distance measure effective for speech recognition[4], and also proposed a sub-phonetic category set, called demiphoneme, as a descriptive unit of speech recognition[5,6]. A recognition system using this unit achieved a relatively high recognition score compared with those using ordinary phonetic units[7,8].

In the present research, we intend to grasp the potential and limitation of speech power spectra in the automatic speech recognition. At the first stage, the methods are implemented on the bases of the our previous works.

A power spectrum enhancement technique proposed in the next section is basically similar operation to that described in reference[9]. At this time, we confirm how the shape of estimated power spectra(i.e., the criterion in spectrum estimation) affect the result of speech recognition. In the time-varying feature extraction, crucial points of our method are: 1) the time-varying feature is derived by subtracting its average from the original values, as described in reference[4], and 2) estimating new coordinates by applying a simplified discriminant analysis to a speech sample set in which all samples are labeled into demiphoneme sequences.

In the following sections, a concrete procedure is described and the effectiveness of our method is evaluated by speaker-independent spoken word recognition.

2. POWER SPECTRUM ENVELOPE ESTIMATION AND ENHANCEMENT

2.1 Mel-Scaled Filter Bank

The power spectrum envelopes are initially estimated by a pass-band filters, in which the shape of the pass-bands is defined as a Gaussian window on log Fourier power spectra and the center frequency intervals and band widths are an equal

rate in the following mel frequency scale:

$$f^{(mel)} = 700\log_2(1 + f^{(Hz)}/700) \quad (1)$$

The number of channels is presently determined as 64, which cover the frequency band from about 100Hz to 7000Hz. The band width ranges from 25Hz to 220Hz(35 mel in the mel frequency).

2.2 Spectrum Enhancement

The spectrum enhancement plays a role of enhancing phonetically significant features. In our research, it is assumed that the above role is achieved by suppressing gross spectrum variations and enhancing local peaks and valleys. For such purpose we have already proposed a technique, but the following operation is an alternative:

Let us denote the power spectrum by $P(f)$ and its differentiation by $P'(f)$. Then the frequency positions satisfying the following equation are the local peaks or zeros.

$$P'(f) = 0 \quad (2)$$

Therefore, if we modify $P'(f)$ as follows:

$$Q'(f) = Const\sqrt{P'(f)} \quad if\ P'(f) \geqq 0 \quad (3)$$
$$= -Const\sqrt{-P'(f)} \quad if\ P'(f) < 0$$

then $Q(f)$, integral of $Q'(f)$ is a spectrum of which local peak and valley bands are enhanced if *Const* is large enough.

3. EXTRACTION OF THE TIME-VARYING ACOUSTIC FEATURES

3.1 Demiphoneme Labels

The demiphoneme(called APSeg in some of our papers) is a subphonetic unit for Japanese speech recognition[5-7]. It is defined for a phoneme sequence from an acoustic-phonetic sense, such as:

Phoneme sequence: /yokohama/ (city name)

Demiphoneme sequence:

<Y-YY-YO-OO-OK-QK-KK-KO-OO-OH-
HH-HA-AA-AM-MM-MA-AA-A>

Therefore, we can classify the labels into two classes, as follows:

Acoustically sustained or steady labels: YY, OO, QK, KK, HH, AA, MM, etc.

Acoustically non-steady(transitional) labels: <Y, YO, OK, KO, OH, MA, etc.

Note that time-varying acoustic patterns are essential for the non-steady labels and demiphoneme label set are currently defined for Japanese speech but it will be possible to define a similar category for other languages.

3.2 Time-Varying Acoustic Features

First let us denote an acoustic feature vector series by X_t, $t=0,1,2,\ldots$,(at present, this feature vector is mel-cepstra which is derived by a cosine expansion of power spectrum and is sampled at the 5ms interval). Then let us introduce a matrix representation to represent time-varying features, as follows:

$$R_t = \begin{bmatrix} r_{11}(t) & \cdots & r_{1,2N+1}(t) \\ \cdots & \cdots & \cdots \\ r_{M1}(t) & \cdots & r_{M,2N+1}(t) \end{bmatrix} = [X_{t-N} \quad \cdots \quad X_t \quad \cdots \quad X_{t+N}] \quad (4)$$

Rt generally represents time-space patterns, but as indicated in the previous papers[4], it is not effective to represent time-varying features(this fact is also confirmed in the present experiments). Thus we introduce the following modification to Rt by subtracting the average value of each time axis:

$$\tilde{R}_t = \left[\tilde{r}_{m,n}\right] \quad (5)$$

where $\tilde{r}_{m,n}(t) = r_{m,n}(t) - \frac{1}{(2N+1)}\sum_{n=1}^{2N+1} r_{m,n}(t)$

After here, Rt and $\tilde{R}_t$ are treated as vectors, that is, the order of $M \times (2N+1)$ vectors. (In the following discussions, N=3).

3.3 Estimating Eigen Vectors for the New Feature Vectors

It is known that linear discriminant analysis is equivalent to K-L Expansion of class-centroid vectors of a given sample distribution in its normalized feature space where within-class-variances are normalized. (Note that the derived axes are not orthogonal with each other in the linear discriminant analysis.) Since we intend to keep orthogonality in the transformation for obtaining new feature vectors from original features, we use only K-L Expansion of the centroid vectors without normalizing the within-class-variances.

We think that (a)as for the steady demiphoneme labels, their absolute values in the original feature space are essential for representing the distinction, and (b)as for the non-steady labels, their time-varying patterns are essential.

Thus, we derive two kinds of coordinates: one is derived by applying the K-L Expansion to the centroid vectors of sample distributions for steady demiphoneme labels. The other is derived from non-steady demiphoneme labels by the same operation, except that the time-varying patterns defined in eq.(5) are used in this case. These coordinates(eigen vectors) are represented by E_{ab} and E_{tv}, respectively.

In the present experiments, the coordinates (eigen vectors) are estimated by using a phonetically balanced word set[10] which contained 1542 words uttered by 6 male speakers(9252 samples in all). The word samples are all labeled into demiphonemes.

In carrying out speech recognition, the new feature vector series is calculated from original feature vector series Rt by transforming it using two kinds of the eigen vector sets E_{ab} and E_{tv} obtained by the above K-L Expansions. The first half of the new feature vector $Y_{ab}(t)$ represents the distinction of steady labels and the latter half $Y_{tv}(t)$ represents those of non-steady labels as follows:

$$Y(t) = \begin{bmatrix} Y_{ab}(t) \\ \lambda Y_{tv}(t) \end{bmatrix} \qquad (6)$$

where λ is a normalizing factor.

4. SPOKEN WORD RECOGNITION EXPERIMENTS

4.1 Recognition Procedure

To evaluate effectiveness of the proposed features, phoneme-HMM-based word recognition experiments are implemented. **Figure 1** shows a procedure to recognize the word label of the input speech sample. Recognition scores are compared with those obtained by using the mel-cepstra and delta-mel-cepstra provided by HTK(software tool)[11].

Experimental conditions and optional parameters are determined into familiar values. The number of the mel-cepstra used is 12, so that when adding delta-mel-cepstra, the order of feature vector is 24, and the order of the proposed feature vector is the same, that is, 12 derived from the steady label set and 12 from the non-steady label set. All experimental conditions are the same between these two feature sets.

HMM training, recognition and scoring are carried out using the HTK. The topology of the HMMs is a left-to-right model of 4 states and the continuous density distribution type is adopted. 43 phoneme-HMMs are used(this is ordinary for Japanese). The phoneme HMMs are trained using the phonetically balanced word set.

4.2 Experiments and Results

Speaker-independent isolated word recognition test is made using part of the word set(the vocabulary size= 492) uttered by 4 speakers. As described above, two factors, i.e., power spectrum estimation techniques and time-varying feature extraction techniques are examined, so that hereinafter the following abbreviations are used:

a) *Triangular-Window*: Filter bank characteristics for mel-cepstrum calculation provided by HTK.

b) *Gaussian-W+Peak Enhancement*: Power spectrum estimation by the proposed method.

c) *seg(ab)*: Feature set corresponding to Y_{ab} in eq.(6).

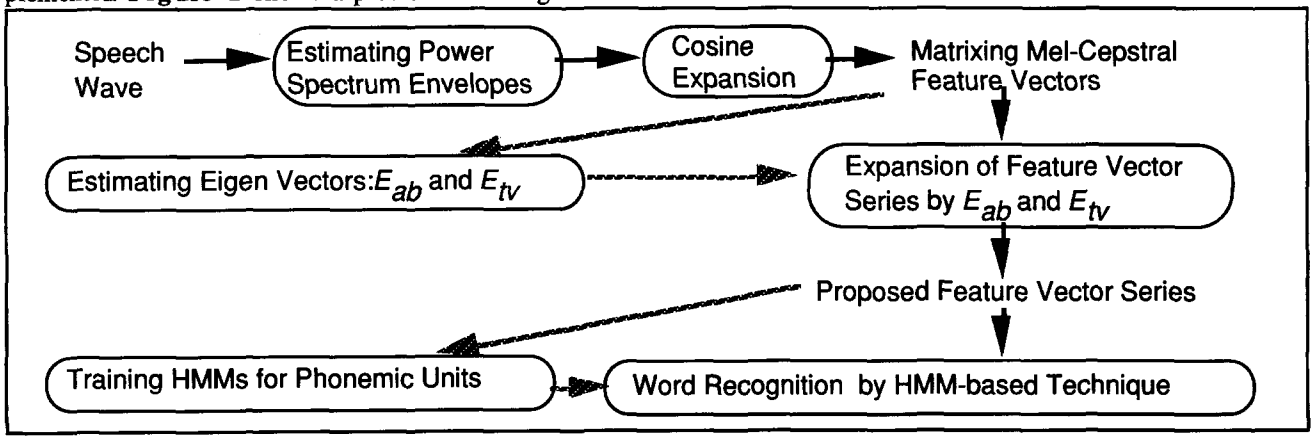

Figure 1 Flow diagram of the isolated word recognition system.

d) *seg(tv)*: Feature set corresponding to Y_{tv}.

Table 1 and **2** show recognition results for this word set. From the results, we found that:

1) for the spectrum estimation, the scores obtained by the proposed technique(*Gaussian*) are improved as, e.g., 82.8% to 88.9%, and 88.7% to 90.8%, in the 2 mixtures case.

2) for the time-varying feature extraction, the scores obtained by the proposed method(*seg(ab)+ seg(tv)*) are improved as, e.g., 82.8% to 88.7%, and 88.9% to 90.8% in the 2 mixtures case.

Table 1 Word Recognition Scores(%) in Several Conditions.

Spectrum Estimation	Feature Set	Number of Mixtures		
		1	2	4
Triangular-Window	melcep+Δmelcep	80.6%	82.8%	*
	seg(ab)+seg(tv)	83.8	88.7	91.5
Gaussian-W+ Enhancement	melcep	-	76.8	78.6
	seg(ab)	-	84.4	89.4
	melcep+Δmelcep	-	88.9	91.1
	seg(ab)+seg(tv)	-	90.8	92.8

Table 2 Individual Speaker's Recognition Scores(%) in Case of the Bold Line Box Shown in Table1.

Feature Set	Spk1	Spk2	Spk3	Spk4	ave.
melcep+Δmelcep	93.5	83.9	92.1	94.9	91.1
seg(ab)+seg(tv)	91.9	88.4	95.3	95.7	92.8

Deference of the scores decreases when the scores become high. This may be due to saturation of the recognition score which depends on characteristics of individual utterance data sets. Observing individual speaker's scores, relatively worse speakers results tend to be more improved.

5. CONCLUDING REMARKS

Further investigations are needed for fixing a base line of this feature extraction method. For example, effect in adverse environment is still ambiguous.

From the discriminant analysis viewpoint, the proposed feature extraction method should be primarily applied to a demiphoneme-based speech recognition. Therefore, we are planning more large vocabulary speech recognition in which the demiphonemes will be adopted as the recognition unit.

References:

[1]B.A. Hanson, B.H. Applebaum, "Robust speaker-independent word recognition using static, dynamic and acceleration features," ICASSP-90 pp.857-860 (1990).

[2] E. Tsuboka,"Hidden Markov Model embedded dynamic features of speech spectrum,"(in Japanese) IEICE Technical Report SP87-97 (1987).

[3] S. Frui,"Speaker independent isolated word reconition using dynamic features speech spectrum," IEEE Trans. ASSP-34, No.1, pp.52-59 (1986).

[4] K. Tanaka, "A parametric representation and clustering method for phoneme recognition," IEEE Trans. ASSP-29, No.6 pp.1117-1127(1981).

[5] K. Tanaka, S.Hayamizu, K.Ohta, "A demiphoneme network representation of speech and automatic labeling techniques for speech database construction," ICASSP-86, Paper 7.1 (1986).

[6] K. Tanaka, S.Hayamizu, K.Ohta, "Automatic labeling of known speech using a demiphoneme network representation and a parameter series segmentation," (in Japanese) J. Acoust. Soc. Jp Vol.42, No.11, pp.860-868(1986).

[7] S.Hayamizu, K.Tanaka, K.Ohta, "A large vocabulary word recognition system using rule-based network representation of acoustic characteristic variation," ICASSP-88, Paper S5.8 (1988).

[8] K. Tanaka, S.Hayamizu, K.Ohta, "Sorting and clustering of acoustic phonetic variations based on a fine-labeled speech database with applications for automatic word recognition," (in Japanese) Trans. IEICE Vol.J73-D-II, No.10, pp.1619-1629(1990).

[9] K.Tanaka, "A dynamic processing approach to phoneme recognition(part I): Feature extraction," IEEE Trans. ASSP-27, 6, pp.596-608 (1979).

[10] S.Hayamizu, K.Tanaka, S.Yokoyama, K.Ohta, "Generation of VCV/CVC balanced word sets for speech data base,"(in Japanese) Bull. ETL, Vol.49, No.10, pp.803-834(1985).

[11] S.J. Young, et.al, HTK V1.5 & V2.0 User Manual, Entropic Research Laboratory, Inc., 1993 & 1995.