Peripheral Features for HMM-based Speech Recognition

Takashi FUKUDA, Masashi TAKIGAWA and Tsuneo NITTA
Graduate School of Eng., Toyohashi University of Technology
1-1 Hibariga-oka, Tempaku, Toyohashi JAPAN E-mail:nitta@tutkie.tut.ac.jp

ABSTRACT

This paper describes an attempt to extract peripheral features of a point $\mathbf{c}(t_i,q_j)$ on a time-quefrency (**TQ**) pattern by observing n×n neighborhoods of the point, and then to incorporate these peripheral features into the **MFCC**-based feature extractor of a speech recognition system as a replacement to dynamic features. In the design of the feature extractor, firstly, the orthogonal bases extracted directly from speech data by using **KLT** of 7×3 blocks on a **TQ** pattern are adopted as the peripheral features, then, the upper two primal bases are selected and simplified in the form of Δ_t-operator and Δ_q-operator. The proposed feature-set of **MFCC** and peripheral features shows significant improvements in comparison with the standard feature-set of **MFCC** and dynamic features in experiments with an **HMM**-based ASR system. The reason for the increased performance is discussed in terms of minimal-pair tests.

1. INTRODUCTION

Time-spectrum (**TS**) pattern $\mathbf{x}(t,f)$ has long been used for acoustic features in automatic speech recognition (ASR), and recently, dynamic features such as Δ-cepstrum, Δ-power, etc. have been introduced into ASR[1], [2] and the set of **MFCC** and dynamic features is widely used. Dynamic features represent peripheral features of a point on a **TS** pattern $\mathbf{x}(t_i,f_j)$ along the time axis, however, we can obtain more information from n×n neighborhoods of the point.

In the previous work[3], the 7×7 orthogonal bases on TS patterns were extracted directly from a speech database by using **KLT** (see **Figure 1**), and were incorporated into a feature extractor as mapping operators to extract peripheral features. The feature set of **MFCC** and peripheral features showed significant improvements in comparison with a standard feature-set of **MFCC** and dynamic features. In this method, a set **X** with elements $\mathbf{x}(t,f)$ is mapped onto various peripheral features $\mathbf{Y}_m=\mathbf{y}_m(t,f)$, m=1, 2, …, M by using time-frequency mapping operator $\{\mathbf{\Phi}_m\}$ ($\mathbf{\Phi}_m \in \mathbf{\Phi}$):

$$\mathbf{\Phi}_m : \mathbf{X} \rightarrow \mathbf{Y}_m \qquad (1)$$

Figure 2 illustrates an example of the upper three peripheral features of an utterance [kaden'tsa] (cadence).

In this paper, we propose a design methodology for a peripheral feature extractor in the cepstrum-domain instead of the frequency-domain. In the methodology, we observe peripheral features, or mapping operators, on

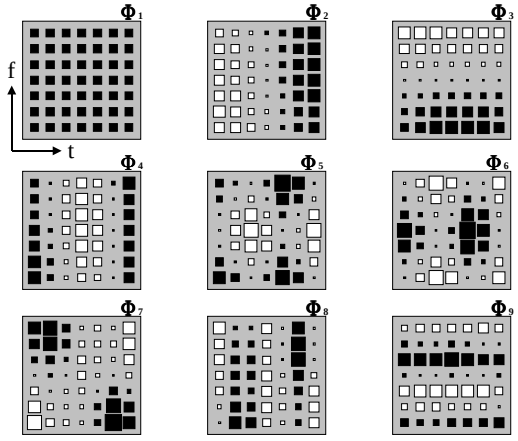

Figure 1 7×7 orthogonal basis on TS pattern

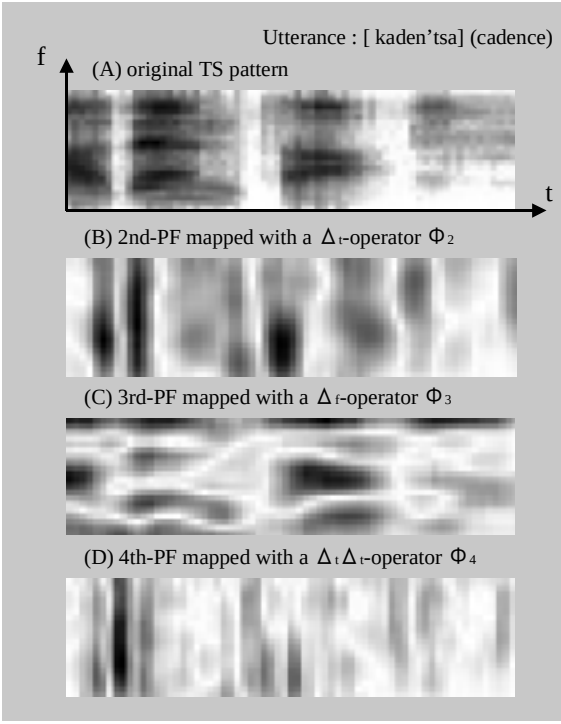

Figure 2 Time-spectrum pattern and peripheral features in frequency domain

time-quefrency (**TQ**) pattern, then incorporate the mapping operators into the feature extractor of a **HMM**-based ASR system after simplifying $\{\mathbf{\Phi}_m\}$and making them symmetrical. The feature-set of **MFCC** and

peripheral features is compared with a standard feature-set of **MFCC** and dynamic features in word recognition experiments with a **HMM**-based **ASR** system.

This paper is organized as follow: Section 2 discusses the geometrical structure of 7×3 blocks in time-quefrency space. Section 3 then outlines a method of implementing the peripheral features in a feature extractor of an ASR system together with **MFCC** parameters. Section 4 describes the experimental setup and results, and section 5 provides the discussion.

2. PERIPHERAL FEATURES ON TIME - QUEFRENCY (TQ) PATTERN

Our previous work[3] showed that many types of geometrical structures are observed on **TS** patterns. In this section, we observe peripheral features on **TQ** patterns instead of **TS** patterns. **Figure 3** shows the upper nine orthogonal bases of 7×3 blocks on **TQ** patterns $\{\mathbf{c}(t_i,q_j)\}$, j = 1, 2, …, 12. Tease are converted by **DCT** of the output of **BPFs** $\{\mathbf{x}(t_i,f_j)\}$, j=1, 2, …, 24. In the figure, black and white squares represent positive and negative values, respectively, and the size of each square represents amplitude. The 7×3 orthogonal bases were extracted by using Karhunen-Loeve transform (**KLT**) from the speech data described in section 4.1.

From a space-operational point of view, $\mathbf{\Phi}_1$ is considered to be a smoothing operator, and as such is neutral and generally has no effect on feature extraction for ASR. $\mathbf{\Phi}_2$ and $\mathbf{\Phi}_3$ are the first and second derivative operators with respect to the quefrency axis (Δ_q-operator and $\Delta_q\Delta_q$-operator), respectively, $\mathbf{\Phi}_4$, $\mathbf{\Phi}_7$ and $\mathbf{\Phi}_9$ are the first, second and third derivative operator with respect to the time axis(Δ_t-operator, $\Delta_t\Delta_t$-operator, $\Delta_t\ \Delta_t\ \Delta_t$ -operator), respectively, and $\mathbf{\Phi}_5$, $\mathbf{\Phi}_6$ and $\mathbf{\Phi}_8$ are subspaces that represent ridges and/or valleys on **TQ** patterns.

TQ space operators, or mapping operators, $\mathbf{\Phi}_m$ map a **TQ** pattern **c**(t,q) onto peripheral features $\mathbf{Y}_m=\mathbf{y}_m(t, q)$, m=1, 2, …, M. An element $\mathbf{y}_m(t, q)$ of peripheral features is calculated with 7×3 neighborhoods of **c**(t,q) and $\mathbf{\Phi}_m = \boldsymbol{\varphi}_\mathbf{m}(t,q)$ by the following equation:

$$y_m(t,q) = \sum_{i=-3}^{3}\sum_{j=-1}^{1} c(t+i, q+j)\Phi_m(i,j) \qquad (2)$$

Figure 4 illustrates an example of the upper three peripheral features of utterance [kaden'tsa] (cadence). Here, the mapping operators $\{\mathbf{\Phi}_m\}$, m=2,3,4 were applied to the **TQ** pattern after simplifying $\{\mathbf{\Phi}_m\}$ and making them symmetrical. In this case, the three mapping operators are correspondent to the Δ_q-operator, $\Delta_q\Delta_q$-operator, and Δ_t-operator, respectively. In the figure, (A) is an original **TS** pattern, (B) is the **MFCC** obtained by converting the **TS** pattern with **DCT**, and (C), (D), and (E) represent the 2nd-**PF** (**PF** : Peripheral Feature) mapped with a Δ_q-operator $\mathbf{\Phi}_2$, the 3rd-**PF** mapped with a $\Delta_q\ \Delta_q$ -operator $\mathbf{\Phi}_3$, and the 4th-**PF** mapped with Δ_t-operator $\mathbf{\Phi}_4$, respectively. In the figure, patterns of peripheral features and **MFCC** are displayed as absolute values.

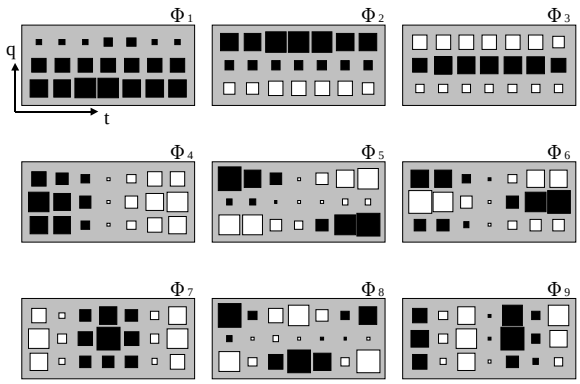

Figure 3 7×3 orthogonal bases on TQ pattern

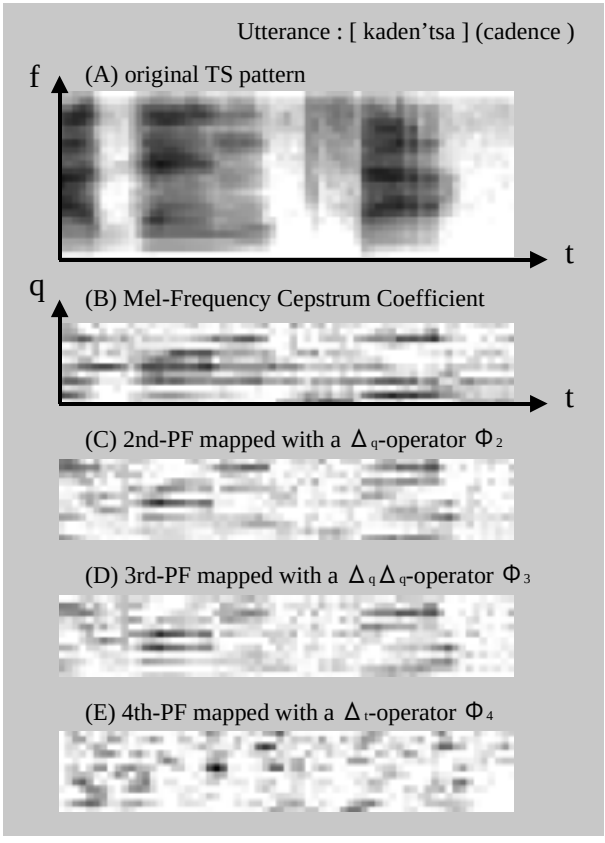

Figure 4 Peripheral features in cepstrum domain

3. COMBINING PERIPHERAL FEATURES WITH MFCC

This chapter describes the methods of extracting peripheral features and combining them with **MFCC** in a feature extractor. **Figure 5-A** shows a standard feature parameters used in current **HMM**-based ASR systems. In the feature extractor, input speech is sampled at 16 kHz and a 512-point **FFT** of 25 ms Hamming-windowed

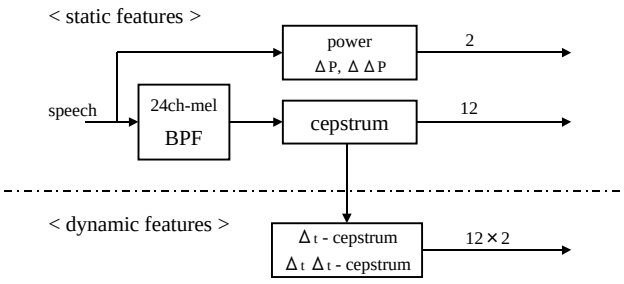

Figure 5-A MFCC with dynamic features

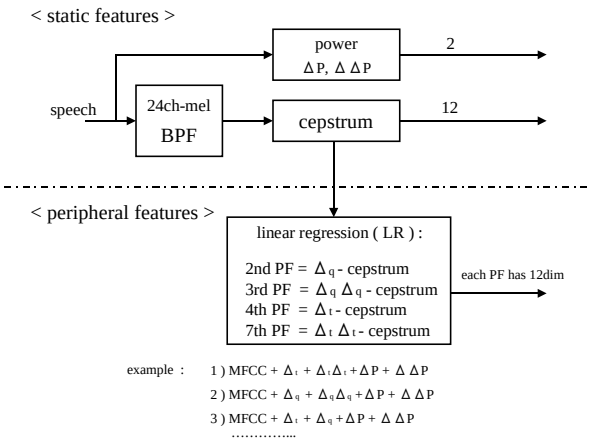

Figure 5-B MFCC with peripheral features

speech segments is applied every 10 ms. The resultant **FFT** power spectrum is then integrated into the output of 24ch-**BPFs** with mel-scaled center-frequencies. Then, 38 feature parameters including 12 static parameters (mel-cepstrum), ΔP (logarithmic power), ΔΔP, and 24 dynamic features (Δ_t, $\Delta_t\Delta_t$) are extracted after converting the output of **BPFs** into cepstrum coefficients (**MFCC**).

In section 2, we investigated the 7×3 orthogonal bases on **TQ** patterns and found that the upper primal bases include derivative operators ($\mathbf{\Phi}_2$, $\mathbf{\Phi}_3$) along quefrency-axis, while the standard **MFCC**-based extractor did not include them. **Figure 5-B** shows the procedure for extracting peripheral features, including dynamic features, and **MFCC**.

In figure 5-B, four space operators, giving the four peripheral features Δ_q-, $\Delta_q\Delta_q$-, Δ_t-, and $\Delta_t\Delta_t$- cepstrum, are simplified in the form of 1×3-block operators (Δ_q, $\Delta_q\Delta_q$) and 7×1-block operators (Δ_t, $\Delta_t\Delta_t$), and the derivative operation is replaced by a linear regression calculation. Various peripheral features are combined with MFCC static features (12 MFCC + ΔP + ΔΔP).

4. EXPERIMENTS

4.1 Speech Database

The following three data sets were used :

D1. Acoustic model design set: A subset of "ASJ(Acoustic Society of Japan) Continuous Speech Database", consisting of 4,503 sentences uttered by 30 male speakers (16KHz, 16bit)

D2. Test data set: A subset of "Tohoku University and Matsushita Spoken Word Database", consisting of 100 words uttered by 10 unknown male speakers. The sampling rate was converted from 24KHz to 16KHz.

D3. 7×3 orthogonal basis design data set: A subset of "ASJ News Corpus(ASJ-JNAS)", consisting of 2,662 sentences uttered by 53 male speakers.

4.2 Experimental Setup

The **D1** data set was used to design 43 Japanese monophone-**HMM**s with five states and three loops. In the **HMM**, output probabilities are represented in the form of Gaussian mixtures, and covariance matrices are diagonalized (mixture = 1, 2, 4). Speaker-independent word-recognition tests were then carried out with the **D2** data set. Feature parameters were evaluated by combining Δ_t, $\Delta_t\Delta_t$, ΔP, ΔΔP, Δ_q, and $\Delta_q\Delta_q$ with **MFCC**.

4.3 Result

Figure 6 shows experimental results when various types of features are added to **MFCC**. The results show that:

- The peripheral feature Δ_q-cepstrum provides improvement equal to that of Δ_t-cepstrum, and
- the highest score is obtained when both Δ_t-cepstrum and Δ_q-cepstrum are added.

In the experiment, the feature parameter set "**MFCC**+Δ_t +Δ_q+$\Delta_q\Delta_q$+ΔP+ΔΔP" with a feature dimension of 50 did not score higher than the "**MFCC**+ Δ_t + Δ_q +ΔP+ΔΔP" set because of the small size of the data set. The experimental results suggest that the addition of dynamics along the quefrency axis to the standard **MFCC** parameter set is important.

5. DISCUSSION

Why are the recognition scores of the proposed feature extractor higher than those of the baseline extractor? The proposed extractor had the facility to grasp the variation along quefrency-axis that contributes to recognize vowels and consonants, while the variation along time-axis has strong influence to consonant recognition.

To compare the contribution of two features Δ_t ($\Delta_t\Delta_t$) and Δ_q we prepare two word list, or minimal-pair lists, that are composed from an original 100-word list of D2. In the first list that includes minimal pairs of vowels /a, i, u, e, o, N (independent nasal sound)/, nonsense words were made by replacing a vowel in the original word with a different one. In the second list that contains minimal pairs of consonants /p, t, k, b, d, g, s, c, h, z, m,

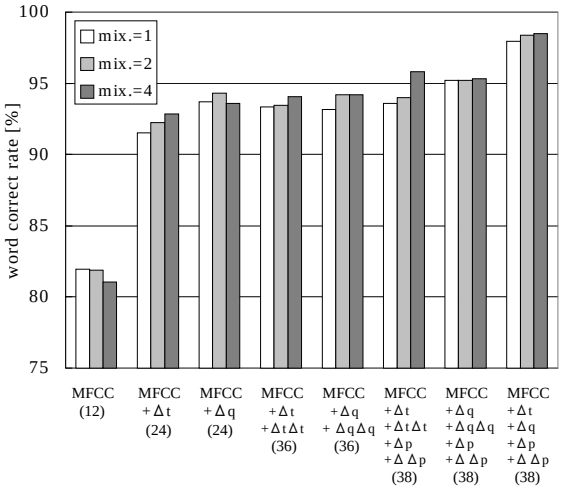

Figure 6 Comparison between MFCC parameter sets

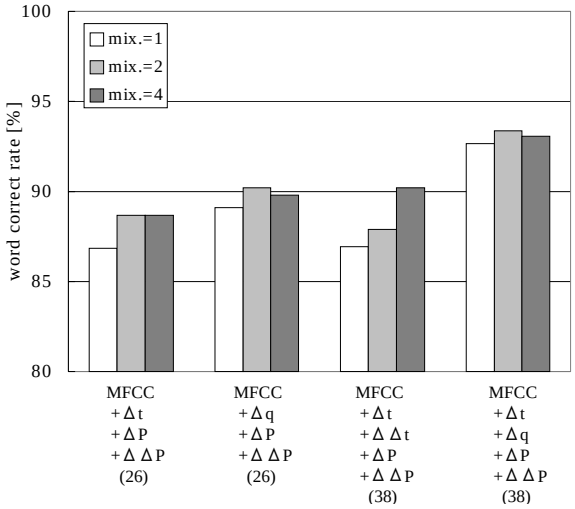

Figure 7-A minimal-pair test of vowels

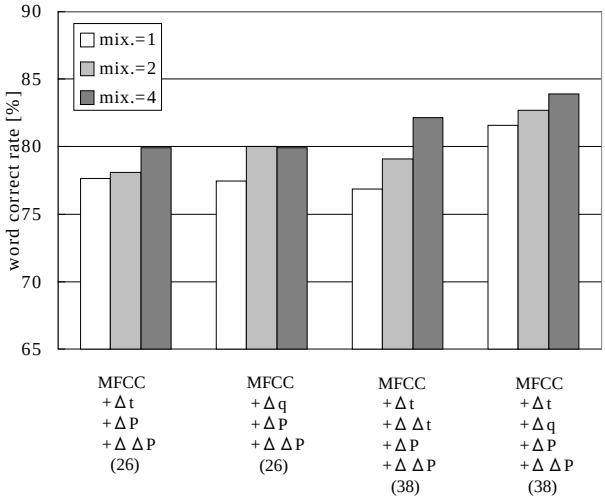

Figure 7-B minimal-pair test of consonants

n, r, j, w/, nonsense words were rebuilt by changing a consonant with the other one to form such a confusable pair as [wi:kude:] (weekday) ↔ [wi:kube:].

After combining each word list with the original 100-word list, 200-word recognition experiments for "Minimal pairs of vowels" and "Minimal pairs of consonants" were carried out. Figure 7-A and -B are the experimental results of "Minimal pairs of vowels" and "Minimal pairs of consonants", respectively. The result in Figure 7-A shows that Δ_q parameter is superior to Δ_t parameter for the discrimination of vowels in spoken words, while the result in Figure 7-B indicates that Δ_q -parameters and Δ_t -parameters equally contribute to performance in the discrimination of consonants.

6. CONCLUSION

A framework for incorporating multiple geometric structures into the feature extractor of ASR systems was proposed. The design methodology of mapping operators for extracting peripheral features was given by observing the orthogonal bases of speech and by incorporating primal components into a feature extractor in a simplified form. The proposed feature extractor that combines peripheral features with **MFCC** showed significant improvements in comparison with the standard **MFCC**-based feature extractor in the **HMM**-based word recognition experiments. Dynamics along the quefrency axis plays an important role both in vowel and consonant recognition.

REFERENCES

[1] K. Elenius and M. Blomberg, "Effect of emphasizing transitional or stationary parts of the speech signal in a discrete utterance recognition system", IEEE Proc. ICASSP'82, pp.535-538 (1982).

[2] S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum", IEEE Trans. Acoust. Speech Signal Process. ASSP-34, pp.522-59 (1986).

[3] T. Nitta, M. Takigawa, and T. Fukuda, "A Novel Feature Extraction Using Multiple Acoustic Feature Planes for HMM-based Speech Recognition " Proc. ICSLP'00, Vol.1, pp.385-388 (2000).