# Feature Extraction for Speech Recognition Based on Orthogonal Acoustic-Feature Planes and LDA

*Tsuneo NITTA**
Multimedia Engineering Laboratory, TOSHIBA CORPORATION
* Currently, he belongs to Graduate School of Eng. at 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 multiple topological structures, hidden in time-spectrum (TS) patterns, by using multiple mapping operators, and to incorporate the operators into the feature extractor of a speech recognition system. In the previous work, the author proposed a novel feature-extraction method based on MAFP/KLT (MAFP: multiple acoustic-feature planes), in which 3×3 derivative filters were used for mapping operators, and showed that the method achieved significant improvement in preliminary experiments. In this paper, firstly, the mapping operators are directly extracted in the form of a 3×3 orthogonal basis from a speech database. Next, the operators are evaluated, together with 3×3 simplified operators modeled on the orthogonal basis. Finally, after comparing the experimental results, the author proposes an effective feature-extraction method based on MAFP/LDA, in which a Sobel filter is used for mapping operators.

## 1. INTRODUCTION

Although the time-spectrum (**TS**) pattern x(t,f) has long been used for acoustic features in speech recognition systems, dynamic features such as Δ-cepstrum, Δ-power, etc. have been introduced in recent years [1],[2]. The author has proposed a novel feature-extraction method based on multiple acoustic-feature planes (**MAFP**) and showed that the method significantly improved the error rate from 34.5% and 29.6% obtained by x(t,f) and x(t,f)+$\Delta_t$x(t,f) to 17.0% for unknown speakers in preliminary experiments performed on a Japanese E-set (12 consonantal parts of /Ci/) extracted from continuous speech [3].

In the feature-extraction method based on MAFP, a TS pattern x(t, f) is mapped onto multiple **AFPs** (acoustic-feature planes) $y_m(t, f)$, m=1, 2, ..., M by using mapping operator $\mathbf{G_m}$ ($G_m \in G$):

$$G_m: x(t, f) \rightarrow y_m(t, f) \quad (1)$$

In the previous work [3], four types of 3×3 derivative filters used for edge enhancement in image processing were applied as mapping operators $G_m$. The four mapping operators are expected to capture the four types of local acoustic evidence observed in TS patterns of speech: sharply rising and falling sound (**RF-AFP**), spectral peaks in steady sound or sound changing slowly in time-spectrum space (**SP-AFP**), sharply ascending FM sound (AF-AFP), and sharply descending FM sound (DF-AFP). The preliminary experiments showed that RF-AFP and SP-AFP were dominant. In this paper, firstly, the 3×3 orthogonal basis $\{\Phi_1, \Phi_2, ..., \Phi_9\}$ on TS patterns is observed by extracting it directly from a speech database. Next, the author proposes a model that simplified $\{\Phi_m\}$ and compares these two types of mapping operators in phonetic-segment classification tests.

MAFP represents topological structures of TS patterns, however, the reason for the high performance shown in the previous experiments is considered to be that the feature compression, or the feature selection by using the Karhunen-Loeve transform (**KLT**) following the linear mapping, has an important role. In this paper, the feature selection by using linear discriminant analysis (**LDA**) is also investigated to achieve more accurate and effective feature extraction.

This paper is organized as follows: Section 2 discusses the topological structure of 3×3 blocks on TS patterns. Section 3 then outlines an experimental speech recognition system and how MAFP and a feature selector are implemented in the system. Finally, Section 4 gives the experimental setup, the results and discussion.

## 2. 3×3 ORTHOGONAL BASIS EXTRACTED FROM SPEECH DATA

We can observe many types of local variations on TS patterns, however, the orthogonal basis of n×n blocks extracted from speech data is not as complicated as we expected. Figure 1 shows an example of the 3×3 orthogonal basis calculated with the total of 160 million 3×3 blocks of TS patterns by using KLT. Speech data includes all the Japanese phonetic segments with Cv and V structures and is analyzed by a BPF bank with 26 channels described in Section 3.

From a space-operational point of view, $\Phi_1$ is considered to be an averaging filter, $\Phi_2$ and $\Phi_3$ are

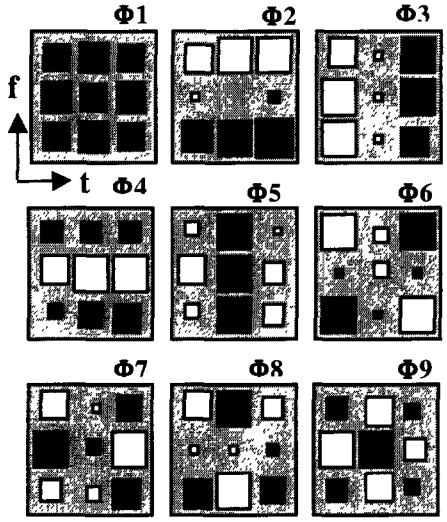


**Figure 1 3×3 orthogonal basis extracted from speech data**

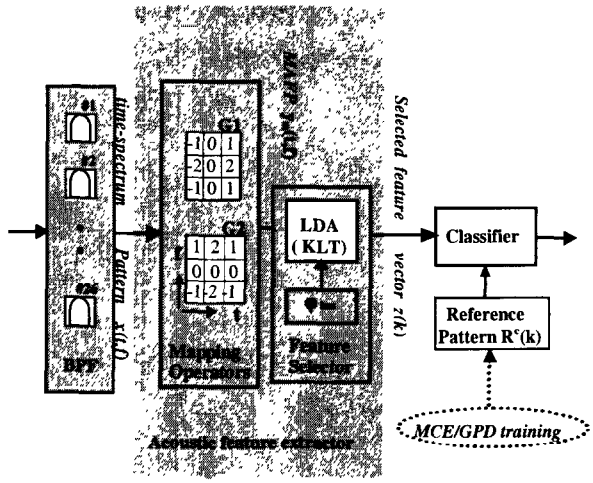


**Figure 2 Experimental ASR system**

the first-order derivative filters with respect to the f-axis and t-axis, respectively, $\Phi_4$ and $\Phi_5$ are the second-order derivative operators with respect to the f-axis and t-axis, respectively, and $\Phi_6$, $\Phi_7$, $\Phi_8$, $\Phi_9$ are subspaces that represent ridges and/or valleys on TS patterns. $\Phi_9$ is similar to Laplacian. $\{\Phi_m\}$ will be implemented in the acoustic feature extractor as mapping operators $G_m$.

## 3. EXPERIMENTAL ASR SYSTEM

Figure 2 shows an experimental ASR system incorporating mapping operators $G_m$, or space filters, and a feature selector. An input speech is sampled at 11 kHz and a 256-point FFT of the 24 ms Hamming-windowed speech segments is applied every 8 ms. The resultant FFT power spectrum is then integrated into a BPF output with 26 dimensions [3].

At the acoustic-feature extraction stage, an output of BPF bank, or a TS pattern x(t, f), is mapped onto MAFP $y_m(t,f)$ by using mapping operators $G_m$. An element $y_m(t, f)$ of MAFP is calculated with 3×3 neighborhoods of x(t, f) and $G_m = g_m(t, f)$ by the following equation:

$$y_m(t, f) = \sum_{i=-1}^{1} \sum_{j=-1}^{1} x(t+i, f+j)\, g_m(i, j) \qquad (2)$$

In Figure 2, mapping operators $\{G_1, G_2\}$ are represented by two types of five-level derivative filters called a Sobel filter in image processing [4]. $G_1$ and $G_2$ in Figure 2 correspond to $\Phi_3$ and $\Phi_2$ in Figure 1, respectively, and are expected to capture two types of dominant acoustic evidence, (1) RF-AFP: sharply rising (+) and falling (-) sound and (2) SP-AFP: spectral peaks in steady sound or sound changing slowly in time-spectrum space. Figure 3 shows an example of 3×3 simplified operators modeled on the orthogonal basis $\{\Phi_m\}$, i= 1, 2, ... , 5. $\{G_1, G_2\}$ is a Sobel filter and $G_1$, $G_2$, $G_3$, $G_4$, $G_5$ are modeled on $\Phi_3$, $\Phi_2$ , $\Phi_1$, $\Phi_5$ , $\Phi_4$, respectively.

Figure 4 shows an example of MAFP that represents the utterance /geist/ ([gaist]). In the figure, (A) is an original time-spectrum pattern and (B) and (C) represent a RF-AFP mapped with $G_1$ and a SP-AFP mapped with $G_2$, respectively. A positive sign of $y_m$ (t, f) means a positive slope, negative sign a negative slope. For example, a clear spectral peak in steady sound is represented by a pair of positive and negative values in SP-AFP. In the figure, AFP patterns are displayed with absolute values.

A feature selector selects effective dimensions among many dimensions of MAFP. This reduces the computation time and memory at the classification stage. Feature selection in the feature selector has two stages. At the first stage, recombination suitable for each time-frequency resolution of AFPs is applied [3]. RF-AFP, for example, needs a high resolution on the time axis (6 ch. × 12 frames), while SP-AFP requires a high resolution on the frequency axis (26 ch. × 3 frames). At the second stage, statistical feature selection based on KLT or LDA is applied. A selected feature vector z(k), k=1,2,..., K is given by the following equation:

$$z(k) = \sum_{m=1}^{M} \sum_{f=1}^{F} \sum_{t=1}^{T} y_m'(t, f)\varphi_{km}(t, f) \quad k=1, 2,..., K \quad (3)$$

where, $y_m'(t, f)$, t=1,2,..., T, f=1,2,..., F is the m-th AFP after recombination, $\varphi_{km}$ (t, f), m=1,2,..., M is the k-th eigen vector set of KLT ($\varphi^{KL}_{km}$ (t, f)) or LDA ($\varphi^{LD}_{km}$ (t,f)). The evaluation test of phonetic segments is

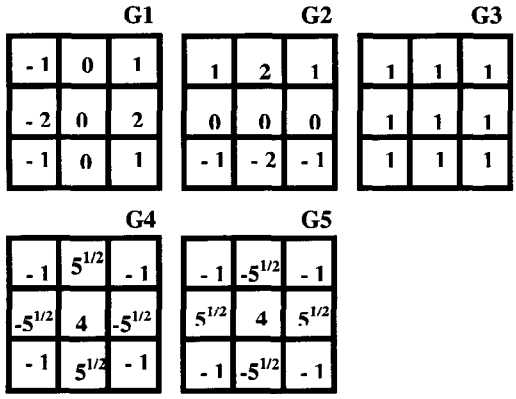


**Figure 3 Mapping operators modeled on orthogonal basis.**

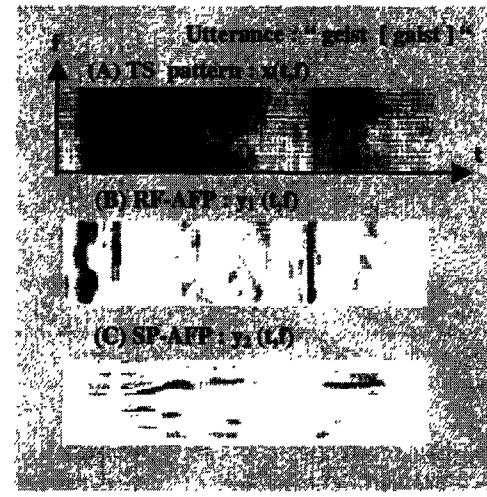


**Figure 4 Time-spectrum pattern and acoustic feature planes (AFPs).**

performed with a classifier based on MCE/GPD competitive training [5]. The author presented the KL/GPD competitive training method, in which both $\varphi^{KL}_{km}(t, f)$ in a feature selector and reference patterns of a classifier were trained and the resultant eigen vectors of $\varphi^{KL}_{km}(t, f)$ put at an angle to one another to minimize classification errors [6]. On the other hand, in the case of $\varphi^{LD}_{km}(t, f)$, because eigen vectors at an angle are extracted to maximize the F-ratio, improvement in recognition accuracy is small when the competitive training for a feature extractor is added. In the following discussion, MCE/GPD training is applied only to a classifier.

## 4. EXPERIMENTS

### 4.1 Speech Database

The experiments were carried out with a Japanese Cv-set extracted from continuous speech manually. The set consists of 68 consonantal parts (v=/a, i, u, e, o/) and includes all the Japanese Cv structures. The number of speakers was 8 (4 males and 4 females) and the total number of samples was 4,523. The evaluation experiments were controlled with the deleted interpolation technique and were performed for unknown speakers (open test).

### 4.2 Comparison for Mapping Operators

Table 1 compares the error rates between the mapping operator $\{\Phi_m\}$ extracted from speech data and the modeled operator $\{G_m\}$. The results were combined into a total in two ways by grouping the consonants (C), as well as Cv. The number of categories are 68 (Cv group) and 13 (C group). LDA was applied for feature selection.

The following five operators were evaluated.

- first-order derivative operators:
  orthogonal basis extracted speech data $\{\Phi_2, \Phi_3\}$
  modeled but symmetrized operators $\{G_1, G_2\}$
- first-order and second-order derivative operators:
  orthogonal basis extracted speech data $\{\Phi_2,\Phi_3,\Phi_4,\Phi_5\}$
  modeled but symmetrized operators $\{G_1,G_2,G_4,G_5\}$
- averaging filter and first-order derivative operators:
  modeled but symmetrized operators $\{G_1,G_2,G_3\}$

The results show:

- The first-order derivative operator is dominant.
- Additional AFP ($m \geq 3$) mapped with the second-order derivative operator or the averaging filter gives no improvement.
- The modeled but symmetrized operator gives enough performance.

**Table 1 Comparison for Mapping Operators ( error rate [%] )**

| Type | Cv / C | Feature dimension 12 | 32 | 48 |
|---|---|---|---|---|
| $\{\Phi_2,\Phi_3\}$ | Cv | 33.0 | 25.5 | 24.8 |
| | C | 24.0 | 18.7 | 18.6 |
| $\{\Phi_2,\Phi_3,\Phi_4,\Phi_5\}$ | Cv | 38.0 | 32.3 | 31.8 |
| | C | 30 5 | 24.0 | 24.9 |
| $\{G_1,G_2\}$ | Cv | 31.6 | 25.6 | 23.9 |
| | C | 25.5 | 19.5 | 17.4 |
| $\{G_1,G_2,G_3\}$ | Cv | 32.7 | 26.9 | 25.1 |
| | C | 24.6 | 19.3 | 18.6 |
| $\{G_1,G_2,G_4,G_5\}$ | Cv | 32.9 | 26.1 | 24.6 |
| | C | 25.5 | 18.8 | 17.8 |

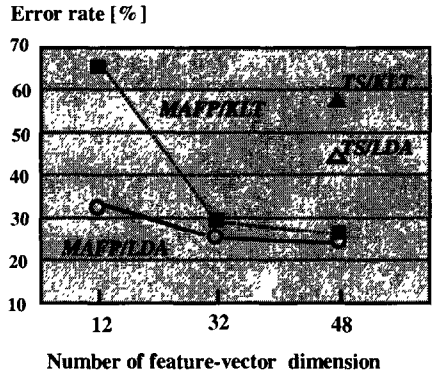


**Figure 5-A MAFP/KLT vs. MAFP/LDA : Classification of Cv group.**

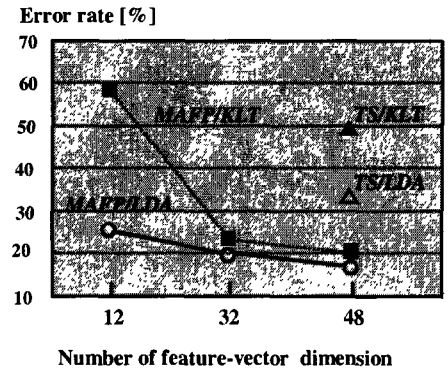


**Figure 5-B MAFP/KLT vs. MAFP/LDA : Classification of C group**

In the following experiments, the mapping operator is fixed with the Sobel operator $\{G_1, G_2\}$.

### 4.3 Comparison for Feature Selectors

The following four acoustic features were evaluated.
- original TS pattern + KLT (TS/KLT)
- original TS pattern + LDA (TS/LDA)
- MAFP + KLT (MAFP/KLT)
- MAFP + LDA (MAFP/LDA)

Figure 5-A and 5-B show the error rate of Cv group and C group, respectively. The results show:

- MAFP/LDA has comparatively higher performance than MAFP/KLT.
- MAFP/LDA can maintain accuracy in the range of small feature dimensions.
- In the case of the original TS pattern, the difference in performance between KLT and LDA at the dimension 48 is wide. This shows that MAFP has the advantage that comes with its explicit topological structure.

## 5. CONCLUSION

A framework for incorporating multiple topological structures into the feature extractor of a speech recognition system was proposed. Moreover, the design methodology of mapping operators in the feature extractor was given by observing the orthogonal basis of speech and modeling it. The proposed method based on MAFP/LDA showed significant improvements in comparison with the convenient method of TS/KLT and TS/LDA in the experiments with a Japanese Cv-set sp eech database and can maintain accuracy in the range o f small feature dimensions.

## ACKNOWLEDGEMENTS

The author gratefully thanks Takeshi Inoue for his assistance with experiments.

## 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, " A novel feature-extraction for speech recognition based on multiple acoustic-feature planes", IEEE Proc. ICASSP'98, pp. 29-32 (1998).

[4] L.S. Davis, " A survey of edge detection techniques", CGIP, 4, 3, pp.248-270 (1975).

[5] B.H. Juang and S. Katagiri," Discriminative learning for minimum error classification", IEEE Trans. Signal Processing, 40, 12, pp.3043-3054 (1992).

[6] T. Nitta and A. Kawamura, "Designing a reduced feature-vector set for speech recognition by using KL/GPD competitive training", Eurospeech'97,pp.2107-2110 (1997).