A Novel Feature-Extraction for Speech Recognition Based on Multiple Acoustic-Feature Planes

Tsuneo NITTA
Multimedia Engineering Laboratory, TOSHIBA CORPORATION
70 Yanagi-cho, Saiwai-ku, Kawasaki 210 JAPAN
E-mail: nitta@sp.mmlab.toshiba.co.jp

Abstract

This paper describes an attempt to incorporate the functions of the auditory nerve system into the feature extractor of speech recognition. The functions include four types of well-known responses to sound stimuli: local peaks of spectrum in steady sound, ascending FM sound, descending FM sound, and sharply rising and falling sound. Each function is realized in the form of a three-level derivative operator and is applied to a time-spectrum (TS) pattern X(t,f) of the output of BPF with 26-channels. The resultant acoustic cue of an input speech represented by multiple acoustic-feature planes (MAFP) is compressed by using KLT, then classified. In the experiments performed on a Japanese E-set (12 consonantal parts of /Ci/) extracted from continuous speech, the MAFP significantly improved the error rate from 34.5% and 29.6% obtained by X(t,f) and X(t,f)+Δ_t X(t,f) to 17.0% for unknown speakers (dimension=64).

1. INTRODUCTION

Speech recognition systems have long incorporated the time-spectrum pattern as acoustic features, but in recent years have introduced dynamic features: Δ-cepstrum, Δ-power, etc. [1],[2]. On the other hand, the auditory nerve responding to the stimuli of ascending FM sound, descending FM sound, sharply rising sound (on-type), sharply falling sound (off-type), etc. are well known in addition to spectral peaks in steady sound [3],[4]. This paper describes an attempt to incorporate such functions of the auditory nerve system into the feature extractor in order to establish the precise speech recognition.

The functions incorporated into a feature extractor are realized by four types of derivative operators that correspond to sharply rising and falling sound (RF), sharply ascending FM sound (AF), sharply descending FM sound (DF), and spectral peaks in steady sound or sound changing slowly in time-spectrum space (SP). The derivative operators are applied to a time-spectrum pattern X(t,f) of the output of BPFs and the resultant acoustic cue of an input speech is represented by multiple acoustic-feature planes (MAFP).

This paper is organized as follows: Section 2 outlines the concept of MAFP and its implementation; section 3 explains the experimental setup; and section 4 gives the results and discussion

2. Multiple Acoustic-Feature Planes (MAFP)

2.1 Concept of MAFP

Figure 1 shows a conceptual schema of MAFP. In the figure, the time-spectrum pattern of an input speech /bja/ is mapped onto four acoustic-feature planes that correspond to the four types of acoustic evidence: (1) RF - sharply rising (+) and falling (-) sound, (2) AF – sharply ascending FM sound, (3) DF – sharply descending FM sound, and (4) SP - spectral peaks in steady sound or sound changing slowly in time-spectrum space. Because these acoustic evidences on the MAFP are considered to play a major role in phoneme discrimination [5], if they can be captured by using some feature extraction mechanism, we can achieve precise speech recognition.

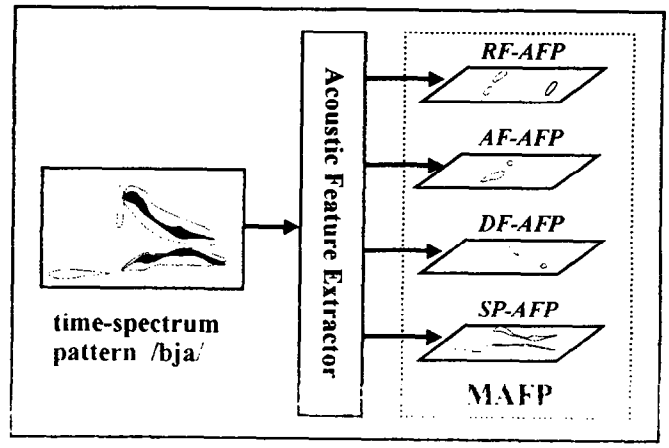


Figure 1 Conceptual schema of MAFP
MAFP: Multiple Acoustic-Feature Planes

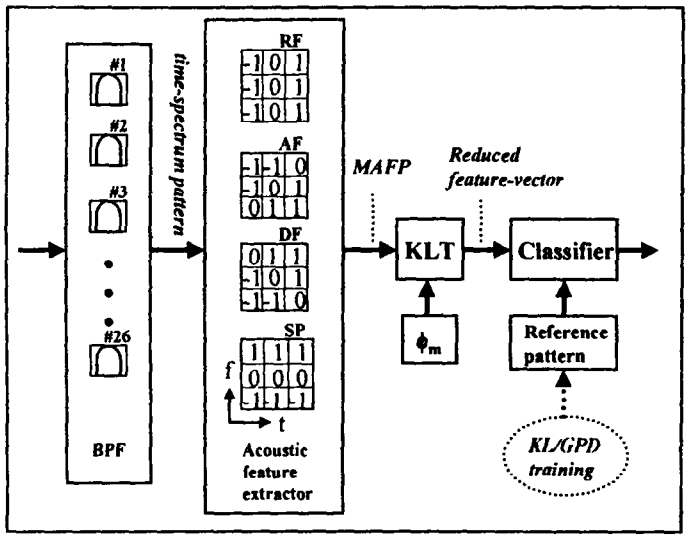


Figure 2 Speech recognition system incorporating MAFP.

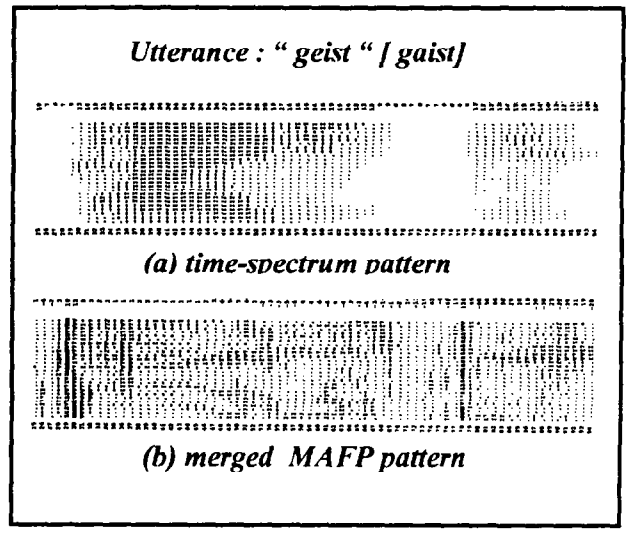


Figure 3 Time-spectrum pattern and merged MAFP.

2.2 Incorporating the MAFP into a Speech Recognition System

One of the early speech recognizer used $\partial X(t,f)/\partial t$ and $\partial X(t,f)/\partial f$ as parameters in the phoneme recognition stage [6]. In recent years, the concept of orientation pattern has been proposed [7]. In this concept, the absolute value $v_1(t,f)$ (= $[(\partial X(t,f)/\partial f)^2 + (\partial X(t,f)/\partial t)^2]^{1/2}$) and the orientation $v_2(t,f)$ (= arctan $\partial X(t,f)/\partial f$ / $\partial X(t,f)/\partial t$) form a vector, so the orientation is devided, or quantized into N directions and only the i-th direction to which the vector belongs has the value of $v_1(t,f)$, while the value for the other directions is 0. In the case of N=8, for example, the i-th element of the orientation pattern $\theta i(t,f)$ (i=1,2,..,8) is given by:

$$\theta i(t,f) = \begin{cases} v_1(t,f) & (i-1)\,\pi/4 - \pi/8 \le v_2(t,f) \\ & \le (i-1)\,\pi/4 + \pi/8 \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

The proposed feature extractor providing MAFP is designed by using four types of 3×3 derivative operators, not by using two types of one-directional (1×2) derivative operators. Figure 2 shows an example of the speech recognition system incorporating the MAFP. In the figure, the feature extractor is composed of four three-level derivative operators used for the edge enhancement in image processing [8]. The four operators are expected to capture the four types of acoustic evidence. X(t,f) on the time-spectrum pattern has the following eight neighborhoods

$$\begin{pmatrix} X(t-1,f+1) & X(t,f+1) & X(t+1,f+1) \\ X(t-1,f) & X(t,f) & X(t+1,f) \\ X(t-1,f-1) & X(t,f-1) & X(t+1,f-1) \end{pmatrix} \quad (2)$$

An element X'(t,f) of an acoustic-feature plane (AFP) is given by:

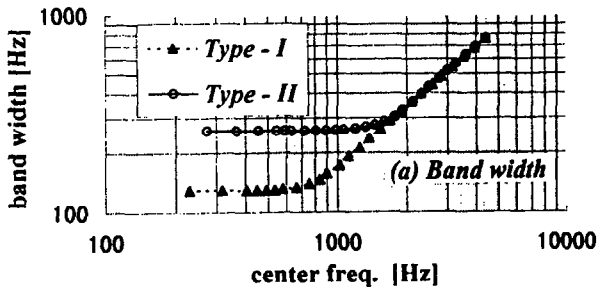
$$X'(t,f) = \sum_{i=-1}^{1} \sum_{j=-1}^{1} X(t+i, f+j)\, W(t+i, f+j) \quad (3)$$

where, W(t,f) is the three-level operator corresponding to each AFP X'(t,f) as shown in Figure 2. A positive sign of X'(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. Figure 3 shows an example of MAFP that represents the utterance /geist/ ([gaist]). In the figure, (a) is an original time-spectrum pattern and (b) represents a pattern into which the four AFPs (RF-, AF-, DF-, and SP-AFP) are merged.

Because MAFP has too many dimensions of 1248 (=26ch.*12frame*4AFP) to directly execute computation at the classification stage, the number of dimensions needs to be reduced. Before executing the reduction by using Karhunen-Loeve transform (KLT), recombination suitable for each time-frequency resolution of AFPs is applied (Table 1). SP-AFP needs a high resolution on the frequency axis, while RF-AFP requires a high resolution on the time axis. The number of dimensions after the recombination becomes 488. After KLT, the reduced feature-vector with the dimensions of 32 – 80 is given. In this paper, the evaluation test for phonetic segments is performed with a classifier based on KL/GPD competitive training [9].

Table 1 Recombination of the dimension of time-axis and frequency-axis.

	number of channels	number of frames
original TS pattern	26	12
RF-AFP	6	12
AF-AFP	13	12
DF-AFP	13	12
S P-AFP	26	4



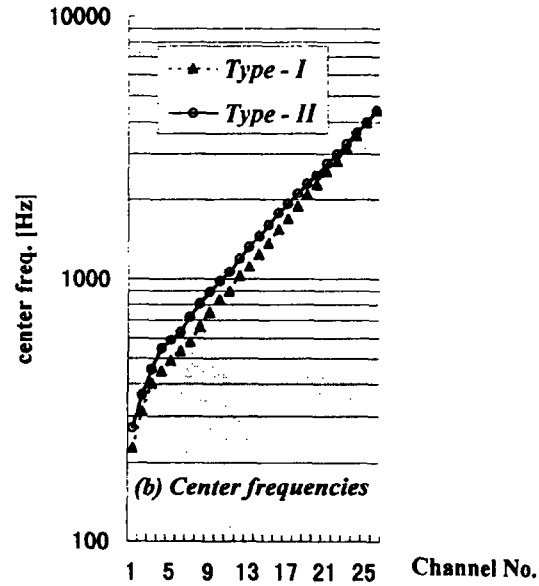


Figure 4 BPF specification

3. Experiments

3.1 Speech Database

The experiments were carried out with two sets of database extracted from continuous speech manually.

(a) Japanese V-set: 5 vowels and an independent nasal sound /N/. The number of speakers was 20 (10 males and 10 females) and the total number of samples was 1161.

(b) Japanese E-set: 12 consonantal parts of Ci: #i, ki shi, chi, ni, hi, mi, ri, gi, ji, bi, pi. The number of speakers was 8 (4 males and 4 females) and the total number of samples was 425.

3.2 Experimental Setup

Speech was sampled at 11 kHz and a 256-point FFT of the 24 ms Hamming-windowed speech segments was applied. The frame-rate was 8 ms.

BPF design: Two types of BPF banks were tested. Both were 26 critical-band filters as shown in Figure 4 - (a). Type-II has rather wide band-widths in the low frequencies to give stability in high pitch speech [10]. Because it is important for the application of 3×3 derivative-operators not to include more than one event within a 3×3-window, the center frequencies in Figure 4 - (b) as well as the frame rate were decided by investigating many time-spectrum patterns.

Acoustic features evaluated: The Following four acoustic features were evaluated with V-set and E-set databases.

- Time-spectrum pattern (TS)
- TS + Δ_t-parameter (TS+Δ)
- 4-AFP-merged pattern (merged MAFP; (b) in Fig. 3)
- MAFP (Type-I and Type-II BPF bank)
- MAFP + TS

Δ_t-parameters were computed by $\Delta_t = X(t-2, f) - X(t+2, f)$.

4. Results and Discussion

The evaluation experiments were controlled with the deleted interpolation technique and were performed for unknown speakers (open test).

• **BPF :** Table 2 shows the results of BPF comparison for the acoustic features of MAFP. The table shows that Type- II with wider band-widths in the low frequencies has a more stable performance than Type- I. The BPF type was therefore fixed to Type- II in the following experiments and discussions.

Table 2 BPF Comparison for MAFP (error rate)

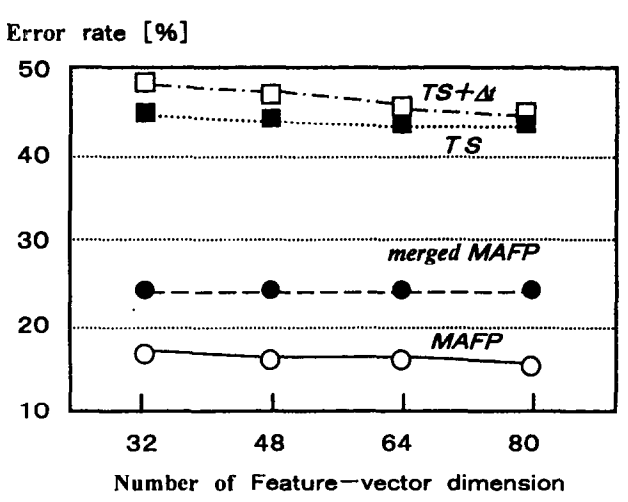
Data	BPF	Feature dimension 32	48	64	80
V-set :	Type- I	17.9	17.4	17.5	16.8
	Type-II	16.7	15.9	15.8	15.7
E-set :	Type- I	21.9	19.2	17.5	17.5
	Type-II	18.8	18.2	17.0	15.8

• **TS :** Figure 5 shows error rates for different acoustic features. Because speech samples in the V-set include weak utterances and nasalized vowels as well as vowels in various contexts, the error rate was comparatively high.

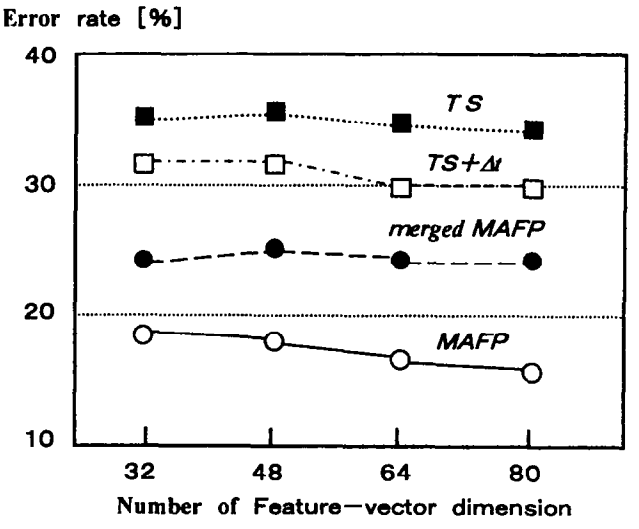
• **TS+Δ_t:** Δ_t-parameter improves the error rate in the E-set test (Figure 5 - (b)). On the other hand, Δ_t-parameter is not effective for steady sound and could not improve the performance of the V-set (Figure 5 – (a)).

• **merged MAFP:** The merged MAFP shown in Figure 3 - (b) is significantly improved the error rate both of the V-set and E-set.

• **MAFP:** The MAFP improved the performance even further. The improvement was about one half in the error rate compared with the TS pattern for both of the V-set



(a) **Japanese V-set** (a, I, u, e, o, N)



(b) **Japanese E-set** (#i, ki, shi, chi, ni, hi, mi, ri gi, ji, bi, pi)

Figure 5 Error rates for different acoustic features

and E-set. In the experiments on the E-set, for example, the MAFP significantly improved the error rate from 34.5% and 29.6% obtained by TS and TS+Δ_t to 17.0% (dimension=64).

The reason for this high performance is considered to be that each AFP constructs a topological subspace and the MAFP represents good-natured acoustic cues.

• **Compactness of MAFP :** When TS pattern is added to other acoustic features (Δ_t-parameter, merged MAFP, etc.), because such acoustic features do not include all the acoustic cues, the error rate is decreased in general. Table 3 compares the results between MAFP and MAFP+TS. The results show that the MAFP holds phonetic information without original TS patterns.

Table 3 MAFP vs. MAFP+TS (error rate %)

(BPF: Type-II)		Feature dimension			
Data	Feature	32	48	64	80
V-set :	MAFP	16.7	15.9	15.8	15.7
	MAFP+TS	18.4	16.9	16.5	16.1
E-set :	MAFP	18.8	18.2	17.0	15.8
	MAFP+TS	20.4	18.4	17.0	15.8

5. Conclusion

A new framework for incorporating the functions of the auditory nerve system into the feature extractor of speech recognition was proposed. The proposed method of multiple acoustic-feature planes (MAFP) showed significant improvements in the experiments with Japanese V-set and E-set speech databases.

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. Hashimoto, Y. Katayama, K. Murata, and I. Taniguchi, "Pitch-synchronous response of cat cochlear nerve fibers to speech sounds", Jpn. J. Physiol., 25, pp.633-644 (1975).

[4] T. Watanabe, Jpn. J. Physiol., 22, pp.569-583 (1972).
T. Watanabe, Jpn. J. Physiol., 22, pp.569-583 (1972).

[5] P. Ladefoged, "A course in phonetics", 2nd Edit., New York: Harcourt, Brace, Jovanovich (1982).

[6] T.B. Martin, "Practical application of voice input to machine", Proc. IEEE, 64-4 (1976).

[7] R. Oka and H. Matsumura, "Speaker independent word speech recognition using the blurred orientation pattern obtained from the vector field of spectrum", Proc. IJCPR (1988).

[8] J. Prewitt, "Object enhancement and extraction", in Picture Processing and Pcychopictorics, B.S. Lipkin and A.Rosenfeld Eds., New York, Academic Press (1970).

[9] 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).

[10] D.H. Klatt, "A digital filter bank for spectral matching", IEEE Proc. ICASSP'76, pp.573-576 (1976).