東北大学機関リポジトリ
TOUR
Tohoku University Repository

Recognition of consonant based on the Perceptron model

著者	牧野　正三
journal or publication title	IEEE International Conference on ICASSP '83. Acoustics, Speech, and Signal Processing
volume	1983
page range	738-741
year	1983
URL	http://hdl.handle.net/10097/46626

Recognition of consonant based on the Perceptron model

Shozo Makino, Takeshi Kawabata and Ken'iti Kido

Research Center for Applied Information Sciences,
Tohoku University

Abstract

This paper proposes a new method for the recognition of consonant based on the Perceptron model. The recognition model is composed of the sensory, feature extraction, response and lateral inhibition layers.

The recognition scores of 90.4% to 98.4% are obtained for unvoiced affricates, unvoiced plosives, unvoiced and voiced fricatives.

1. Introduction

This paper proposes a new method[1] for phoneme recognition, where the segmentation and recognition is carried out simultaneously. The recognition model is composed of the sensory, feature extraction, response and lateral inhibition layers. The model is similar to the Perceptron model[2] composed of the threshold logic unit and the weighted linear sum.

The Perceptron model has two defects; (1) The association layer (feature extraction layer) needs a number of units for extracting features. After the learning, however, most of the units do not contribute to feature extraction.

(2) Physical interpretation for the unit in the association layer selected after the learning is difficult. With the proposed model it is easy to capture the physical meaning of the model for the designer.

The recognition scores are 93.1%, 94.0%, 98.4% and 90.4% for unvoiced affricates, unvoiced plosives, an unvoiced fricative and an voiced fricative, respectively, for the samples in the 212 words uttered by five male and five female speakers.

2. Recognition model

Fig.1 shows the structure of model applied to recognize the four consonant groups as shown in Table 1.

In the sensory layer, input speech is passed through 29-channel BPF's.

In the feature extraction layer, 6 features of the plosive(P), fricative(F), unvoiced(U), non-plosive($\bar{P}$), non-fricative($\bar{F}$) and voiced($\bar{U}$) are extracted from the logarithmic spectrum and the temporal envelope of the logarithmic speech power, using linear discriminant

CH1841-6/83/0000-0738 $1.00 © 1983 IEEE

Authorized licensed use limited to: TOHOKU UNIVERSITY. Downloaded on April 08,2010 at 07:22:07 UTC from IEEE Xplore. Restrictions apply.

function.

In the response layer, the convolutions between the 90 discriminant coefficients and the 90 vectors composed of the above-mentioned 6 features in 15 frames are computed frame by frame, where the Perceptron learning is used for making the coefficients to optimally discriminate a specified consonant group from the others. The frame is regarded a consonant if the computed value exceeds the threshold for the consonant.

In the lateral inhibition layer, the final output is uniquely determined by lateral inhibition even if some consonants are recognized in the previous layer.

The association unit (feature extraction unit) in the Perceptron model is randomly connected with the sensory unit. On the other hand, the feature extraction layer in the present model is systematically connected to extract 6 features with physical meaning. Meanwhile, the response layer in the model is a two dimensional FIR filter in the time and features axes.

3. Sensory layer

The speech is sampled at 24kHz and passed through 29 single tuned digital bandpass filters of Q=6, whose center frequencies are arranged every 1/6 octave between 250 Hz and 6300 Hz. The power of every channel is computed for every frame of 10 ms duration and logarithmically transformed.

4. Feature extraction layer

The logarithmic spectrum of the 29-channel BPF's is defined as vector $\mathbf{x}$. Features of fricative $f_F(t)$ and unvoiced $f_U(t)$ are as follows.

$$f_F(t) = \mathbf{w}_F \cdot \mathbf{x}(t) \quad (1)$$

$$f_U(t) = \mathbf{w}_U \cdot \mathbf{x}(t) \quad (2)$$

The plosive feature is calculated from the temporal envelope $\mathbf{p}(t)$ of the logarithmic speech power.

$$f_P(t) = \mathbf{w}_P \cdot \mathbf{p}(t) \quad (3)$$

where $\mathbf{p}(t) = (p_{t-n}, \ldots, p_{t+n})$

p_t ; logarithmic speech power at the t-th frame

The solution weight vector $\mathbf{w}_F$, $\mathbf{w}_U$ and $\mathbf{w}_P$ for optimally discriminating a specified phoneme group are computed using the discriminant analysis.

A number of new ideas are included in the computation and, as a result, the extracted features yield a clearer physical interpretation than those

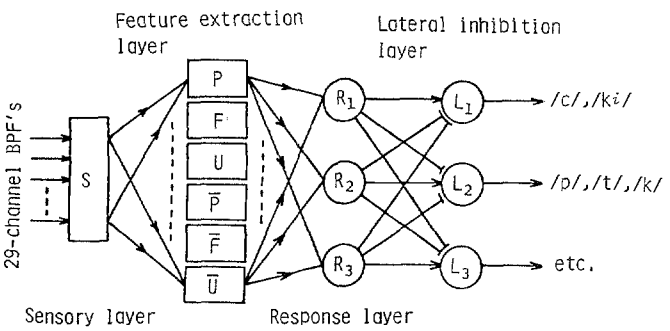

Fig.1 Recognition model

Table 1 Phoneme class

/c/,/ki/	unvoiced affricates
/p/,/t/,/k/	unvoiced plosives
/s/	unvoiced fricative
/z/	voiced fricative

Authorized licensed use limited to: TOHOKU UNIVERSITY. Downloaded on April 08,2010 at 07:22:07 UTC from IEEE Xplore. Restrictions apply.

obtained with the conventional discriminant analysis. Three complementary features $f_{\bar{F}}(t)$, $f_{\bar{U}}(t)$ and $f_{\bar{P}}(t)$ are also computed in the same way.

5. Response layer

Each phoneme is recognized by the response unit independent of the other units. Input features $f^{(i)}(t)$ pass through a FIR filter as follows

$$r(t) = \sum_{i}^{M} \sum_{j=-N}^{N'} a_j^{(i)} f^{(i)}(t+j) \quad (4)$$

$a_j^{(i)}$; Filter coefficients for the i-th fearure

$f^{(i)}(t)$; The i-th feature at the t-th frame.

The unit produces output indicating the existance of the specified phoneme if r(t) exceeds the threshold.

Filter coefficient $a_j^{(i)}$ is computed using the Perceptron learning (fixed-increment error correction training procedure[3]).

Using speech samples labeled by visual inspection, training is carried out. Hereafter this filter is called discriminant filter.

6. Lateral inhibition layer

Because each unit of response layer is constructed independently of each other, output of the response layer is often overlapped. The lateral inhibition layer determines a unique final output from the multi-output in the previous layer. Final output is determined using the priority relation between the phoneme groups, which is computed using the learning method the authors have developed[4].

7. Experiments

The trainings for 4 phoneme and phoneme groups as shown in Table 1 were carried out for the speech samples in the 212 words uttered by 5 male and 5 female speakers.

Fig.2 shows the coefficeints of the discriminant filter for the unvoiced affriates. Unvoiced affricates are characterized by the plosive feature (P^+) and the succeeding fricative feature(F^+).

Fig.3 shows the examples of temporal feature patterns and the output of the discriminant filters for unvoiced affricates and unvoiced plosives, where the specified phonemes are detected correctly.

Table 2 shows the detection score and error scores for the training and test samples, where Set I and Set II include the training and test samples, respectively. The test samples are the 212 words uttered by 5 male and 5 female speakers different from those for the training.

8. Conclusion

A new method for the recognition of consonant is presented and the effectiveness of the method is shown with the experimental results. The new recognition model overcomes the defects in the Perceptron model. The model is shown to have a clear physical meaning and is expected to be useful for the phoneme

Authorized licensed use limited to: TOHOKU UNIVERSITY. Downloaded on April 08,2010 at 07:22:07 UTC from IEEE Xplore. Restrictions apply.

recognition in the speech recognition system for unspecified speakers.

Reference

[1] Kawabata,T., Makino,S. and Kido,K.: "Consonant recognition using the Perceptron learning",(in Japanese), Paper of TG on speech, ASJ., S82-24(June,1982)

[2] Rosenblatt,F.: "Principle of Neurodynamics", Spartan Books(1961)

[3] Nilsson,N.J.:"Learning Machines.", McGraw-Hill, New York(1965)

[4] Kawabata,T., Makino,S. and Kido,K.: "Organization of lateral inhibition.", (in Japanese), Paper of autumn meeting of ASJ., pp 175-176(Oct.,1982)

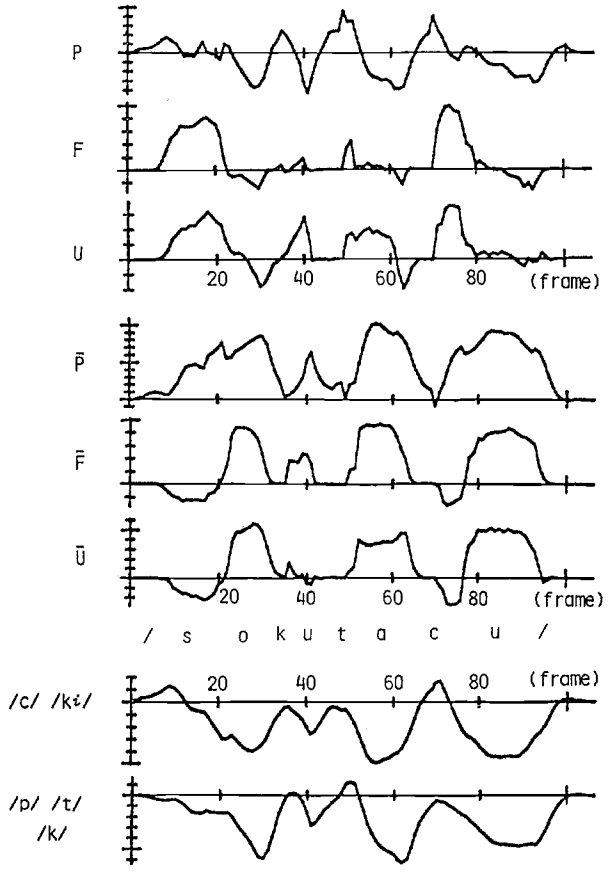

Fig.3 The examples of the temporal feature patterns and the outputs of the discriminant filters.

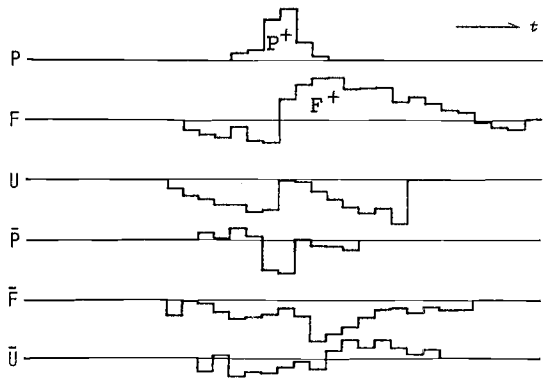

Fig.2 The examples of the temporal feature patterns and the outputs of the discriminant filters.

Table 2 The detection score and error scores

IN \ OUT	/c/ /ki/	/p/ /t/ /k/	/s/	/z/	Number of samples
/c/,/ki/	93.1%	1.6%	1.6%	2.8%	320
/p,t,k/	3.2	94.0	0.0	0.5	793
/s/	0.0	0.0	98.4	0.8	494
/z/	1.9	0.0	0.5	90.4	209
Vowel	0.02	0.3	1.1	1.6	5718
Others	0.3	1.9	1.6	3.3	2962

(Set I)

IN \ OUT	/c/ /ki/	/p/ /t/ /k/	/s/	/z/	Number of samples
/c/,/ki/	90.6%	4.7%	1.6%	2.2%	318
/p,t,k/	3.7	92.7	0.1	0.3	816
/s/	0.2	0.2	94.8	4.2	498
/z/	2.9	0.0	2.4	89.1	210
Vowel	0.03	0.05	0.5	0.4	5786
others	0.5	2.6	0.6	2.2	3015

(Set II)

Authorized licensed use limited to: TOHOKU UNIVERSITY. Downloaded on April 08,2010 at 07:22:07 UTC from IEEE Xplore. Restrictions apply.