Speech Classification Based on Fuzzy Adaptive Resonance Theory

Ching-Tang Hsieh[1] and Chih-Hsu Hsu[2]

[1] Department of Electrical Engineering, Tamkang University, Taipei 251
hsieh@ee.tku.edu.tw

[2] Department of Information Technology, Ching-Kuo Institute of Management & Health, Keelung 203
hsu552@tpts4.seed.net.tw

Abstract

This paper presents a neuro-fuzzy system to speech classification. We propose a multi-resolution feature extraction technique to deal with adaptive frame size. We utilize fuzzy adaptive resonance theory (FART) to cluster each frame. FART was an extension to ART, performs clustering of its inputs via unsupervised learning. ART describes a family of self-organizing neural networks, capable of clustering arbitrary sequences of input patterns into stable recognition codes. In our experiments, the TIMIT database is used and extracts features of each phoneme. The performance of speech classification is 88.66%, demonstrate the effectiveness of the proposed system is encouraging.

Keywords: speech classification, neuro-fuzzy system and fuzzy ART.

1. Introduction

In large vocabulary speech recognition, words are frequently modeled as networks of sub-word units such as phonemes. In other words, a word is modeled acoustically by concatenating phonetic acoustic models according to a pronunciation network stored in a dictionary of phonetic spellings. A benefit of this approach is that it is not necessary for the speaker to train all words in the vocabulary, only the acoustic of phonetic models need to be trained. To identify a phoneme, some of its features in time/frequency or in some other domain must be known. It is a basic requirement of a speech recognition system to extract a set of features for each phoneme. A feature can be defined as a minimal unit, which distinguishes maximally close phonemes. The extracted features are then passed to a network for the recognition of phonemes. The recognized phonemes are combined to give a word utterance.

For simplifying the speech recognition task, speech classification plays an important role in speech recognition system. In speech classification, we extract a set of features of each frame from wavelet transform (WT). The windowed Fouier transform (FT) has uniform resolution over the time frequency plane. It is difficult to detect sudden burst in a slowly varying signal by FT. Recently, WT has been proposed for feature extraction [1-3]. To overcome the problem of fixed resolution extracted from FT, the WT uses adaptive window sizes, which allocate more time to the lower frequency and less time for the higher frequency [4,5].

Several approaches have been proposed for classification problems. Some works focus on conventional probabilistic and deterministic classifiers [6-8]. Another approach uses neural networks to classify patterns. Adaptive resonance theory (ART) describes a family of self-organizing neural networks, capable of clustering arbitrary sequences of input patterns into stable recognition codes [9,10]. Fuzzy-ART (FART) was an extension to ART, performs clustering of its inputs via unsupervised learning [11,12]. The patterns it operates on are assumed to be real-valued vectors with no missing features. A major characteristic of FART is the capability of both incremental (on-line) and batch (off-line) learning.

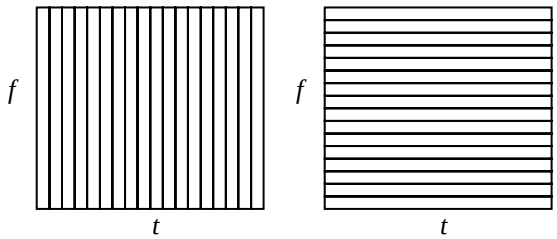
This paper is organized as follows. In section 2 we discuss the features extraction based on wavelet transform. Section 3 briefly describes the class of FART. The experimental results are given in Section 4. Finally, some concluding remarks are presented in Section 5.

2. Features Extraction Based on Wavelet Transform

The multi-resolution formulation of WT is obviously designed to represent signals where a single event is decomposed into finer and finer detail, but it turns out also to be valuable in representing signals where a

time-frequency or time-scale description is desired even if no concept of resolution is needed. In many applications, one studies the decomposition of a signal in terms of basis function. For example, stationary signals are decomposed into the Fourier basis using FT. For nonstationary signals (i.e. signals whose frequency characteristics are time-varying like music, speech, image, etc.) the Fourier basis is ill-suited because of the poor time-localization. The classical solution to this problem is to use the short-time (or windowed) Fourier transform (STFT). However, the STFT has several problems, the most severe being the fixed time-frequency resolution of the basis functions. Wavelet techniques give a new class of bases that have desired time-frequency resolution properties. The "optimal" decomposition depends on the signal studied.

Each function in a basis can be considered schematically as a tile in the time-frequency plane, where most of its energy is concentrated. Nonoverlapping tiles can schematically capture orthonormality of the basis functions. With this assumption, the time-frequency tiles for the standard basis and the Fourier basis are shown in Fig. 1.



(a) (b)

Fig. 1 (a) Standard time domain basis (b) Standard frequency domain basis

The discrete wavelet transform (DWT) is another signal –independent tiling of the time-frequency plane suited for signals where high frequency signal components have shorter duration than low frequency signal components. Fig. 2 shows the corresponding tiling description, which illustrates time-frequency resolution properties of a DWT basis.

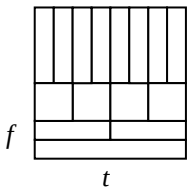


Fig. 2 Three-scale of wavelet basis.

The definition of the scaling function $\phi_{j,k}(t)$ and wavelet function $\psi_{j,k}(t)$ is given by [4,5].

$$\phi_{j,k}(t) = 2^{j/2}\phi(2^j t - k) \qquad j,k \in Z \qquad (1)$$

$$\psi_{j,k}(t) = 2^{j/2}\psi(2^j t - k) \qquad j,k \in Z \qquad (2)$$

This two-variable set of basis function is used in a way similar to the short time Fourier transforms. A set of expansion functions such that any signal can be represented by the series

$$f(t) = \sum_k c_{j+1}(k)\phi_{j+1,k}(t) \qquad (3)$$

or in terms of the next scale as

$$f(t) = \sum_k c_j(k)\phi_{j,k}(t) + \sum_k d_j(k)\psi_{j,k}(t) \qquad (4)$$

A signal space of multi-resolution approximation is decomposed by WT in an approximation (lower resolution) space and a detail (higher resolution) space. In order to generate a basis system that would allow higher resolution decomposition at higher frequencies, we will iterate the WT recursively to divide the approximation space, giving a left binary tree structure. The wavelet packet (WP) was proposed by Ronald Coifman [13] to allow a finer and adjustable to particular signals or signal classes.

3. Fuzzy Adaptive Resonance Theory

ART describes a family of self-organizing neural networks, capable of clustering arbitrary sequences of input patterns into stable recognition codes [9,10]. Grossberg attempted to address the stability-plasticity dilemma: how can a learning system remain plastic (adaptive) in response to new, unseen information, yet remain stable in response to irrelevant information? How can a system preserve its already acquired knowledge and at the same time be flexible enough to accommodate new information to be store? How can the system decide when to alternate from the stable to the plastic state and vice verse? Grossberg's answer to the stability-plasticity dilemma was the ART. In an ART-based network, information reverberates between the network's layers. Learning is possible in the network, when resonance of the neural activity occurs. According to ART, resonance occurs (1) when an already learned pattern is presented and the network recalls / recognizes it and (2) when a novel input pattern is presented, the network realizes that the pattern constitutes new information and then enter resonant state to memorize it.

FART was an extension to ART, performs clustering of its inputs via unsupervised learning [11,12]. The patterns it operates on are assumed to be real-valued vectors with no missing features. Also, FART is an exemplar-based method for clustering,

meaning that instead of memorizing individual patterns it aggregates them into unique category via the use of (in general, overlapping) hyper-rectangles, which define groupings of patterns. Forming groups out of a mass of observations is a form of compression, which FART uses to form abstract rules about the distribution of the data and achieves generalization.

A major characteristic of FART is the capability of both incremental (on-line) and batch (off-line) learning. In the former case, the system is capable of incorporating new evidence about its environment as it becomes available. In the later one, the training set is repeatedly presented until the system learns the presented knowledge either to perfection or to an acceptable degree of accuracy.

A block diagram of a FART module is displayed in Fig. 3, and is comprised by two major subsystems. The attentional subsystem itself consists of three layers of neurons. If the dimensionality of input patterns is M, the module's F0 layer has M nodes and is a pre-processing stage that complete encodes the input patterns. It requires that its input vectors have their entire feature values normalized between 0 and 1. In other words, FART's input domain is M-dimension. F0 transforms an input vector to 2M-dimension, which serves as an input vector to the F1 layer. Although complement coding doubles the size of input patterns, it turns out to be essential for FART to perform clustering. Layer F1 has 2M nodes, while F2 has a large enough number of nodes that will allow the FART module to perform its learning task. All nodes in F1 are interconnected with all nodes in F2 via bottom-up and top-down weights.

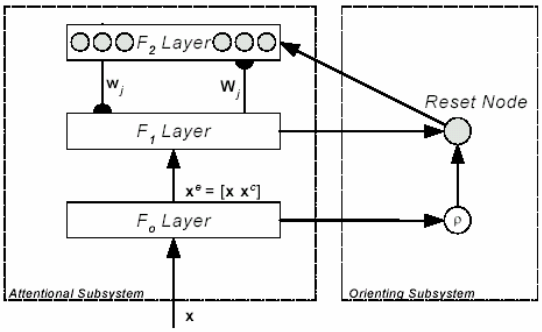


Fig. 3 Block diagram of Fuzzy ART.

FART employs localized rather than distributed learning. The later one applies to multi-layer perceptrons (MLP), in order to learn a single pattern; many weights have to be updated. However, learning a particular pattern in FART only involves the template modification of a single node, whether a category updates or a category creation takes place. Thus, updating weights in FART's learning lies primarily in the search (the combination of repetitive node competition and performing the vigilance test) for a suitable category.

When a single pattern x is being presented, the basic steps of FART training and performance phase are outlined in the below.

For each category j, the category choice function (CCF) $T(w_j \mid \underline{x})$ is defined as

$$T(w_j \mid \underline{x}) = \frac{\left|\underline{x} \wedge w_j\right|}{\alpha + \left|w_j\right|} \tag{5}$$

The CCF reflect the degree to which the weight vector wj is a fuzzy subset of the input vector x. The category match function (CMF) $\rho(w_j \mid \underline{x})$ is defined as

$$\rho(w_j \mid \underline{x}) = \frac{\left|\underline{x} \wedge w_j\right|}{\left|\underline{x}\right|} \geq \rho \tag{6}$$

Resonance occurs if the CMF of the chosen node meets the vigilance criterion. The weight vector wJ is updated according to the equation

$$w_J^{(new)} = \beta(\underline{x} \wedge w_J^{(old)}) + (1-\beta) w_J^{(old)} \tag{7}$$

Fast learning corresponds to setting $\beta = 1$.

1. Present pattern x.
2. Calculate the category choice function (CCF) values $T(w_j \mid \underline{x})$ for all nodes in the F_2 layer according to Eq. (5).
3. Find the smallest node index J, such as
$$J = \arg\max_j \{T(w_j \mid \underline{x})\} \tag{8}$$
4. Perform the vigilance test (VT) for node J as in Eq. (6).
5. If node J passes the VT, proceed to step 6. Otherwise, reset node J and return back to step 3.
6. If FART operates in performance phase, report that $\underline{x}$ is associated to category J, then update J according to Eq. (7) or, if J is uncommitted, report that $\underline{x}$ is a novel pattern.
7. Reinstate the active status of all nodes that have been reset during the node selection search.

In general, FART training may produce overlapping categories, that is, the intersection of some hyper-rectangles corresponding to FART categories might be non-empty. A pattern located inside the intersection of some categories will choose the categories of smallest size.

4. Experimental Results

In speech classification experiments, the database is the Texas Instrument / Massachusetts Institute of Technology (TIMIT) acoustic-phonetic corpus of read

speech [14]. TIMIT database is widely used as a reference for comparison of speech recognition performance [15]. The speech signal was sampled at a rate of 16KHz with 16 bits resolution. A set of features of each frame is extracted from DWT. In this work, Daubechies wavelet is used. Each frame of 16ms was taken and three-scale of DWT decomposition was applied. As the sampling frequency of the speech signal is 16KHz, the frequency bands obtained after decomposition are 0-1, 1-2, 2-4 and 4-8KHz. The normalized energy of each frequency bands is calculated.

We utilize FART to search their clusters. The clustering of phonemes is shown in Table 1. The results of speech classification with Daubechies wavelet were in Table 2.

Table 1 The clustering of phonemes

Cluster	Phoneme
1	silence
2	b, d, g, p, t, k
3	jh, ch, s
4	sh, z, zh, f, th, v, dh, m, n, ng, hh
5	l, r, w, y, iy, ih, eh, ey, ae, aa, aw
6	ay, ah, ao, oy, ow
7	uh, uw, er, ax, axr

Table 2 The results of speech classification.

Cluster	Classification
1	95.2%
2	83.1%
3	84.2%
4	86.5%
5	91.2%
6	89.8%
7	90.6%
Average	88.66%

5. Concluding remarks

In this paper, a neuro-fuzzy model for speech classification is presented. We propose a multi-resolution feature extraction technique to deal with adaptive frame size. We utilize fuzzy adaptive resonance theory to cluster each frame. The fuzzy rules with variable fuzzy regions were defined by activation regions, which show the existence region of data for a class. In the near future, we will try to apply wavelet transform to adjust features to speech recognition system.

6. References

[1] J. W. SEOK and K. S. BAE, "A Novel Endpoint Detection Using Discrete Wavelet Transform," IEICE Trans. Inf. & Syst., Vol. E82-D, No. 11, Nov. 1999, pp. 1489-1491.

[2] S. H. Chen and J. F. Wang, "Application of Wavelet Transform for C/V segmentation on Mandarin Speech Signals," IEE Proc. Vision, Image and Signal Processing, Vol. 148, No. 2, April 2001, pp. 133-139.

[3] O. Farooq and S. Datta, "Phoneme recognition using wavelet based features," Information Sciences 150, 2003, pp. 5-15.

[4] C. S. Burrus, R. A. Gopinath, and H. Guo, "Introduction to Wavelets and Wavelet Transforms," Prentice-Hall, 1998.

[5] G. Strang and T. Nguyen, "Wavelets and Filter Banks," Wellesley Cambridge, 1997.

[6] L. R. Rabiner and B. H. Juang, "An introduction to hidden Marcov models," IEEE ASSP Magazine, Jan. 1986, pp. 4-16.

[7] L. R. Rabiner and B. H. Juang, "Fundamentals of speech recognition," Prentice-Hall, 1993.

[8] X. Huang, A. Acero and H. W. Hon, "Spoken language processing: A guide to theory, algorithm, and system development," Prentice-Hall, 2001.

[9] G. A. Carpenter and S. Grossberg, "The ART of adaptive pattern recognition by a self-organizing neural network," IEEE Computer, Vol. 21, March 1988, pp.77-88.

[10] G. A. Carpenter and S. Grossberg, "Search mechanisms for adaptive resonance theory (ART) architectures," International Joint Conference on Neural Networks, June 1989, pp. 201-205.

[11] G. A. Carpenter, S. Grossberg, and D. B. Rosen, "Fuzzy ART: an adaptive resonance algorithm for rapid, stable classification of analog patterns," IJCNN-91, Vol. 2, July 1991, pp. 411 – 416.

[12] G. C. Anagnostopoulos, "Novel approaches in adaptive resonance theory for machine learning," Ph. D. thesis, University of Central Florida, 2001.

[13] R. R. Coifman and M. V. Wickerhauser, "Entropy-based algorithms for best basis selection," IEEE Trans. on Information Theory, Vol. 38, March 1992, pp. 713-718.

[14] TIMIT Acoustic-Phonetic Continuous Speech Corpus, NIST Speech Disc 1-1.1, Oct. 1990.

[15] R. Chengalvarayan and D. Li, "Use of generalized dynamic feature parameters for speech recognition," IEEE Trans. on Speech and Audio Processing, Vol. 5, Issue: 3, May 1997, pp. 232-242.