An Accurate Speech Classification based on Fuzzy ARTMAP Neural Networks and Wavelet Packet

M.H. Radfar and K. Faez
Dept. of Electrical Eng., Amir Kabir Univ., Tehran, Iran
hradfar@aut.ac.ir | kfaez@aut.ac.ir

Abstract

This paper presents an accurate voiced/unvoiced/ transition/silence speech classifier that is used as an integral part of any toll quality speech coder. Due to ability of wavelet packet in decomposition of time–frequency plane with high resolution, first five discriminate features based on energy concentration of wavelet packet coefficients for each speech classes in time-frequency plane are extracted. Then a Fuzzy ARTMAP neural network classifier, which has been shown a powerful tool for non-stationary signal classification is employed. Experimental results show the proposed approach gives considerable performance improvements in some aspects with respects to the conventional methods

1. Introduction

In recent years emerging application for toll quality speech coder at low bit rate such as third generation of wireless networks has encouraged a lot of research in this area [1], [2]. Aim is to achieve a toll quality, and low bit rate speech coder simultaneously. Unfortunately, the two objectives are in contrast with each other. To cope with this dilemma, an adaptive hybrid-coding algorithm must be implemented for each speech class. Therefore, speech classifier is an inseparable part of any toll quality speech coder. Whatever a classifier is more accurate, more quality is achievable. Mainly low bit rate speech coders classify speech segments into voice/ unvoice source and encode each class with different methods, resulting in low bit rate vocoder especially below 2.4 kbps [3],[4]. However in order to preserve speech quality a more accurate classification than rough classification into voiced /unvoiced (V/U) is needed. Degradation of quality in low bit rate vocoder is mainly due to assumption of simple model for speech signal. Commonly, speech segments that have strong periodicity are identified as voices and unvoices are represented by noise model. Nevertheless transitional speech [6], [7], such as non periodic glottal pulse, onsets (transition from unvoice to voice) and plosives (b,t,g,k,q) don't follow these characteristics, and usually are misclassified by a speech classifier. Consequently need for an accurate multi speech classifier is manifested. Like other pattern recognition systems, speech classifiers consist of two parts, a feature extractor and a classifier. The most commonly speech classification features are zero crossing rates, first auto correlation coefficient, first LPC coefficient, speech peakiness and signal energy, which has been extensively used and developed by researchers [12].[22]. These parameters work properly when high level voice or unvoice are to be classified, but fail for transitional speech. Especially when auto correlation coefficients are used as classification parameters, non-periodic glottal pulses due to weak periodicity are classified as unvoiced but in fact the vocal cord is involved during constructing these sounds. Beyond these features, recently, multiresolution analysis with wavelet transform has been used widely, owing to its potential for dealing with non-stationary signal. Application of wavelet transform in speech classification and pitch detection was first introduced by Kadambe [18] who was inspired by Mallat [19] in which signal discontinuity are detected by decomposition of signal into wavelet scales and stable local maxima across several scales i.e. 3 through 5 are determined. These maxima identify edges or discontinuity in signals. Recently this approach has been developed by other contributions [8], [20]. However, most effort has been focused on event based-detection property of wavelet in speech signals where instances of glottal closure are determined by wavelet transform. Ability of wavelet transform in analysis of time-frequency plane has not been considered well up to now in speech area in compares with short time Fourier transform (STFT) and Mel frequency cepstral coefficients (MFCC). Among wavelet transform algorithms wavelet packet provides better resolution in time frequency domain while preserve computational complexity in comparison with MFCC and STFT .As will be shown, wavelet packet (WP) coefficients energy for each speech class is dispersed in different parts of time frequency plane. This favorite attribute makes WP coefficients a useful discriminated feature in speech classification.

In the second part of system, generally two strategies have been introduced. First by presenting a proper threshold, speech classes are determined [8], [9]. Second a neural network classifier such as Back propagation [10][14], self-organized map [11], and recurrent neural network [13] are employed to determine the corresponding class. Fuzzy ARTMAP neural network uses the class of adaptive resonance theory architecture designed for supervised learning, first introduced by Carpenter [15] and develops in other contributions [16], [17]. Fast and stable learning of large non-stationary databases has made Fuzzy ARTMAP as an outstanding candidate for classification problems.

The remainder of this paper organized as follows. First the procedure for extracting features is explained, and then a brief review of Fuzzy ARTMAP architecture will be presented. Finally experimental results are reported and compared with the conventional method.

2-Feature Extraction based on wavelet Packet

Aim of any feature extraction system is to choose those features, which are most effective for preserving the class separability and to prune non-relevance information.

As discussed before the objective is to classify each sub frame of speech signal (each frame contains 160 samples and is sampled in 8 KHz) as voice, unvoiced, transition, or silence. As it is proved in [21], if the scaling function and wavelets form an orthonormal basis (e.g. Daubechie mother wavelets), the Parsevall theorem relates energy of the signal to energy in each of components and their wavelet coefficients:

$$\int |g(t)|^2 dt = \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{k=\infty} |d_j(k)|^2 \qquad (1)$$

Where the two-dimensional set of coefficients $d_j(k)$, are the discrete wavelet transform (DWT) of $g(t)$. According to this fact, we have a framework for describing the signal by wavelet coefficients in different parts of joint time-frequency domain of time frequency plane. We use dyadic wavelet packet transform, because the flexible tree structure makes it possible to have equivalent sub bands in the whole time frequency plane. In the corresponding wavelet packets situation, each detail coefficient vector is also decomposed into two parts using the same approach as in the approximation vector splitting. This offers the richest analysis of each frame. After examining various speech frames, we found the following partitioning of time frequency appropriate for our purposes. The time frequency plane is divided into different cells as illustrate in Fig 1.

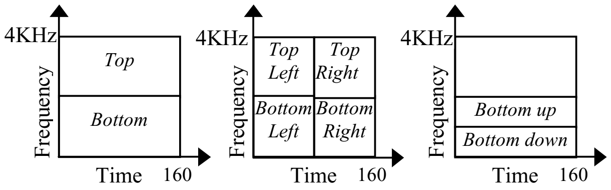


Figure 1-The time frequency plane is divided into different cells

Then the density of energy in each region is obtained by WP coefficients, and finally five sets of features are determined as follows. Decomposition of the wavelet packet tree is obtained in level three. After applying of wavelet packet with Daubechie 8 mother wavelet, the WP coefficient in level three i.e. 8 bands, with 8×160 coefficients, are obtained. We use the wavelet packet because it provides much finer and adjustable resolution. The features are extracted as follows. The term *E()* hereafter means the energy of WP coefficients in one of the indicated region in Fig 1, and is computed as follows

$$\boldsymbol{E(desired\ reigon\ )} = \sum_{j\&k\in desired..reigon} \sum \boldsymbol{d^2{}_j\ (k)} \qquad (2)$$

2-1 Voiced /Unvoiced/Silence Classification

As shown in Fig 2 (a) and (b), most parts of WP coefficients energy for a voiced frame are placed in low frequency part of time frequency plane and evenly distributed in time. In other hand, energy distribution for an unvoiced frame is evenly distributed in all over the time frequency plane. Therefore, we can define feature one by

$$\boldsymbol{feature1} = \frac{\boldsymbol{E(Bottom)}}{\boldsymbol{E(Top)}} \qquad (3)$$

Where for example, *E(Top)* means the WP coefficients energy in band 4through8 (2000Hz to 4000Hz).This parameter is especially useful to discriminate a reliable voiced subframe from unvoiced sub frame. In [8], they have proposed a similar parameter and introduced the threshold to make the relevance decision. However using the threshold solely provides good results, only when a reliable voice is to be classified. The term reliable voiced speech here means a voiced segment with strong periodicity and high energy. So, in order to separate unreliable voiced sub frame from unvoiced one, we consider the second feature that is defined by

$$\boldsymbol{feature2} = \frac{\boldsymbol{E(Bottom\ -\ down)}}{\boldsymbol{E(Bottom\ -\ up)}} \qquad (4)$$

If the value of feature 1 is not too high to make an explicit decision, then if the value of the second feature is high, the sub frame is voiced speech otherwise it is classified as unvoiced.

The above features perform well at distinguishing between voiced and unvoiced speech. However they do not enable to separate silence sub frames from voiced or unvoiced speech. A simple investigation shows the energy value of silence frame is too low in comparison with voiced/unvoiced speech. Thus the total energy of a frame is an appropriate feature to distingue silence segments as is expressed in equation 5.

$$\boldsymbol{feature3} = \boldsymbol{E(Total)} \tag{5}$$

2-2 Transitional Classifications

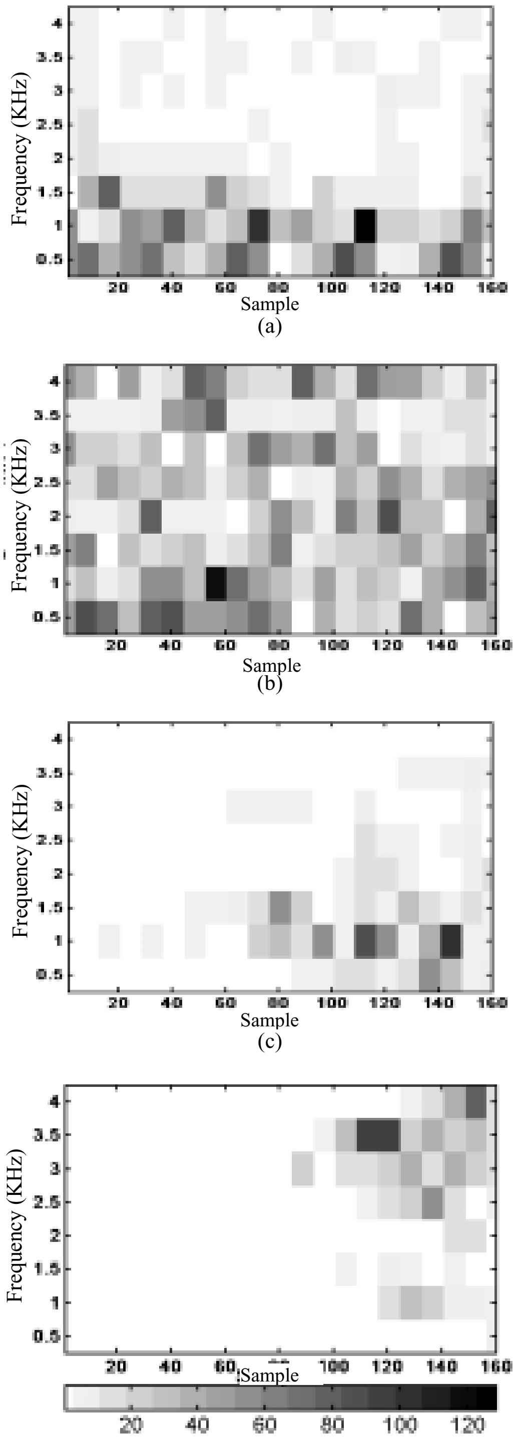
Mainly transitional speech segments occur in the following situations:

First, in onsets where an unvoiced speech is ended and a voiced speech is started. Second, when a plosive sound is uttered such as p, t, b, q. The analysis of transitional speech, especially for onsets shows a transition occurs when a more rapidly waveform with low amplitude (i.e. with weak low frequency components) is followed by a slow varying waveform with high amplitude, (i.e. with dominant low frequency components). The above explanation makes clear that a transitional segment can be distinguished from a voiced or unvoiced speech by comparing the differences between the energy concentrations in bottom-left of time frequency plane to bottom-right of time frequency plane as has been illustrated in Figure 2 (c). Therefore we introduce the forth feature as:

$$\boldsymbol{feature4} = \frac{\boldsymbol{E(Bottom - Right) - E(Bottom - Left)}}{\boldsymbol{E(Total)}} \tag{6}$$

However, this criterion does not work well for all plosive sounds in which a transition happens during passing from a rapidly varying waveform to other rapidly varying waveform with a sudden change in amplitude. In these situations the differences between energy concentrations in top-left region to the top-right of the time frequency plane is a discriminated feature to classify these kinds of transitional sounds, as it has been illustrated in Figure 2 (d) and is expressed in equation 7. Table 1 illustrates the typical value of selected features for each class.

$$\boldsymbol{feature5} = \frac{\boldsymbol{E(Top - Right) - E(Top - Left)}}{\boldsymbol{E(Total)}} \tag{7}$$



(a)

(b)

(c)

(d)

Figure 2: Energy distribution of WP coefficients for (a) voice (b) unvoice (c), (d) transitional segment;

Table 1: Typical value of proposed features for voiced/unvoiced/transition/silence segments

Feature	Silence	Voice	Unvoice	Transition
1	2.2983	9.2805	1.59	3.8321
2	1.757	1.5729	1.1686	2.2777
3	24.392	8055.94	623.903	4061.16
4	.1045	.07277	.027082	.82063
5	.1019	.029257	.078936	.075245

3. The Fuzzy ARTMAP Neural Network

A detailed description of the fuzzy ARTMAP neural network can be found in [15]. Here just a brief review is presented. The fuzzy ARTMAP neural network consists of two fuzzy ART modules, (ARTa, ARTb) as well as an extra inter ART module, shown in Figure 4. Each fuzzy Art is an extension of ART1 system to enable the network to handle the continuous inputs through the use of fuzzy AND operator ($\wedge$), instead of the logical intersection ($\bigcap$). In order to prevent category proliferation [15], input vectors are normalized by complementary coding where if $a \in [0,1]^M$ denotes original input, then the new F_0 layer input vector I ($I = (a, a^c) \in [0,1]^{2M}$, where $a^c = \{a_i^c\}$ and $a_i^c = 1 - a_i$) is fed into network.

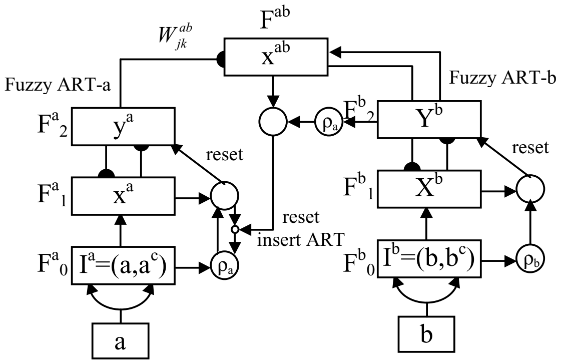


Figure 4: Fuzzy ARTMAP architecture

Thus, the complement coded input and W_j are of dimension 2M. During learning, the adaptive weights or the long term memory (LTM) are adjusted according to the winner take all rule i.e. at most one F2a node can become active at a given time. Map field realizes the match-tracking rule, whereby the vigilance parameter for ARTa increase in response to a miss match at ARTb. After each input in ARTa find a proper output in ARTb, the Map field weights and LTM in ARTa are trained. Thus the number of outputs nodes in ARTa, i.e. F2a, can be increased to some extends, while the number of the Map filed nodes are constant and are equal to the number of output classes. This property of Fuzzy ARTMAP is especially interesting when two input segments with different features belonging to one class are classified. Thus while we preserve generality, the error rate decreases too.

4. Experimental Results

In order to obtain a proper set of training data, twenty sentences from 10 male and 10 female of popular TIMIT database are selected. The sentences are divided into 10 msec (80 sample) sub frames and then are manually labeled as voice, unvoice, transition, and silence. The result of manual class decision is shown in table 2. Each entry shows the percentage of manual class decision.

Table 2: Manual Class decision

Voice	Unvoice	Transition	Silence
45	35	16.25	3.75

Then Daubechie 8 wavelet packet transform was performed on each frame. For restricting the effect of boundary at the frame edge, we extend each frame from left and right to 8 msec, however this amendment introduces 8 msec delays into system. In next step, five features are computed as discussed in pervious section. The features are normalized according to complementary coding as was discussed in section 3. Thus 10 inputs are fed into network. Fuzzy ARTMAP parameters are set as follows: choice parameter is set to zero, the ARTa based lined vigilance parameter, $\overline{\rho}_a$ is set to zero, the learning rate, β is set to one i.e. fast learning. Epsilon is set to 0.001, this value is added to ARTa vigilance parameter when a mismatch happens in F_{ab} , so a reset does not occur in ARTa. And finally the MAP field vigilance parameter ρ_{ab} is set to 0.95. The number of F^a_2 output nodes is initially set to 20. However due to representing the discriminating features just 12 nodes are used at the F_2^a output layer. In order to examine the reliability of features, a fixed set of 4000 chosen exemplars was normalized and presented to the Fuzzy ARTMAP system. After training the system, we test validation of the approach by presenting 4000 test samples. The neural network classification confusion matrix on the test database is presented in table 2 In order to evaluate the accuracy of method, five features (the signal energy, the first reflection coefficient, the rate of zero crossing, the first coefficient of the 10th order LPC and the peak amplitude of signal) that are extensively used in other contributions [12], [14], [5] are extracted from the training data set (4000 exemplars as mentioned later) and fed into a back propagation neural network with 10 hidden layers and 4 outputs. After training, the confusion matrix is obtained by introducing of the test samples to the back propagation classifier. The method that is denoted by (a) in Table 3 demonstrates the results taken from WP-Fuzzy ArtMap classifier and the method (b) indicates the results from the Back propagation classifier. Each entry in the confusion matrix shows the percentage of the automatically detected class in comparison with manual class decision. As it is clear from the table 3, the error rate in the method b is somehow higher than those of the method a, especially for transitional speech. It must be mentioned here, in classifying the speech segments manually, the non-periodic glottal pulses were considered as an voiced speech rather than transitional speech, because naturally vocal cord is exited during the construction of non-periodic speech, although their residual signal exhibits weak periodicity, but the experiments show [6] when the non-periodic sounds are classified as voiced sounds, there is no or only slight degradation of perceived quality as long as the estimated pitch is not extremely high or low.

Table 3:Matrix confusion for proposed (a) WP-FuzzyArtMap classifier and (b) Back propagation

Manual	Methods	Unvoice	Voice	Transition	Silence
Unvoice	(a)	33	0.6	1.125	0.275
	(b)	31.8	0.5	1.979	0.721
Voice	(a)	0.35	43.7	0.75	0.2
	(b)	0.55	42.8	1.45	0.2
Transition	(a)	1.175	0.95	14.125	0
	(b)	2.575	1.55	11.9	0.225
Silence	(a)	0.45	0	0	3.3
	(b)	0.45	0	0	3.3

5. Conclusion

In this research, a more accurate speech classifier was presented. In comparison with other proposed methods, this approach uses a new set of features that appropriately characterize speech classes. Moreover, instead of using the thersholding to make the decision, like what was done in [9], a fast and stable neural network was employed to classify the proposed features. Flexibility of partitioning time-frequency plane into finer cells in addition to low computational complexity in comparison with short time Fourier transform and MFC coefficients, makes wavelet packet a powerful tool for the classification purposes. Furthermore reported results show that the accuracy of this model is comparable with other conventional model, and this method also performs a finer classification more than the voice/unvoice classification, that has been extensively used in other approaches. In the next step of this research we are going to evaluate the model accuracy in noisy environment.

References

[1] E. Shlomot, V. Cuperman, and A. Gersho, “ Hybrid coding: combined harmonic and waveform coding of speech at 4 kb/s” Speech and Audio Processing, IEEE Transactions on, Volume: 9 Issue: 6, pp. 632 –646, Sept. 2001

[2] O. Gottesman, A. Gersho, “Enhanced waveform interpolative coding at low bit-rate “Speech and Audio Processing, IEEE Transactions on, Volume: 9 Issue: 8, pp. 786 –798, Nov. 2001

[3] A. McCree, Kwan Truong, E.B. George, T.P. Barnwell and V. Viswanathan, “A 2.4 kbit/s MELP coder candidate for the new U.S. Federal Standard “Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings, 1996 IEEE International Conference on, Volume: 1, pp. 200 -203, May 1996

[4] Tian Wang, K. Koishida, V. Cuperman, A. Gersho and J.S. Collura, “A 1200 bps speech coder based on MELP” Acoustics, Speech, and Signal Processing, 2000. ICASSP '00. Proceedings. 2000 IEEE International Conference on, Volume: 3, pp. 1375 -1378, June 2000

[5] *J.A Marks,”* Real time speech classification and pitch detection”, COMSIG 88. Southern African Conference on , 24 June 1988

[6] Jongseo Sohn, and Wonyong Sung, “A low resolution pulse position coding method for improved excitation modeling of speech transition” Acoustics, Speech, and Signal Processing, 1999. ICASSP '99. Proceedings, 1999 IEEE International Conference on, Volume: 1, pp. 265 -268, March 1999

[7] Li Chunyan and V. Cuperman, “Enhanced harmonic coding of speech with frequency domain transition modeling” Acoustics, Speech, and Signal Processing, 1998. ICASSP '98. Proceedings of the 1998 IEEE International Conference on, Volume: 2, pp. 581 –584, May 1998

[8] F.C.A. Brooks, L. Hanzo, “A multiband excited waveform-interpolated 2.35-kbps speech codec for bandlimited channels” Vehicular Technology, IEEE Transactions on, Volume: 49 Issue: 3, pp. 766 –777, May 2000

[9] J. Stegmann, G. Schroder, and Fischer, K.A., ”Robust classification of speech based on the dyadic wavelet transform with application to CELP coding” Acoustics, Speech, and Signal Processing,. ICASSP-96. Conference Proceedings, 1996 IEEE International Conference on, Volume: 1, pp.546 -549 1,May

[10] Jiang Minghu, Yuan Baozong, and Lin Biqin, “The consonant/vowel (C/V) speech classification using high-rank function neural network (HRFNN)” Signal Processing, 1996, 3rd International Conference on, Volume: 2, 14-18 Oct. 1996 pp.1469 -1472

[11] P.P. Boda, “Robust voiced/unvoiced speech classification with self-organizing maps” Circuits and Systems, 1995. ISCAS '95, 1995 IEEE International Symposium on, Volume: 2, pp.1516 -1519, 28 April-3 May 1995

[12] A. Bendiksen, and K. Steiglitz, “ Neural networks for voiced/unvoiced speech classification” Acoustics, Speech, and Signal Processing, 1990. ICASSP-90, 1990 International Conference on, pp.521 -524, April 1990

[13] Wei-Chen Chang, and A.W.Y. Su,” A novel recurrent network based pitch detection technique for quasi-periodic/pitch-varying”, Neural Networks, IJCNN '02, Proceedings of the 2002 International Joint Conference on, Volume: 1, pp. 816 –821,

[14] T. Ghiselli-Crippa, A. El-Jaroudi, “Voiced-unvoiced-silence classification of speech using neural nets”, Neural Networks, 1991, IJCNN-91-Seattle International Joint Conference on, Volume: ii, 8-14 July 1991 pp. 851 -856 vol.2

[15] G.A. Carpenter, S. Grossberg, N. Markuzon, J.H. Reynolds and D.B. Rosen, “ Fuzzy ARTMAP: A neural network architecture for incremental supervised learning of analog multidimensional maps” Neural Networks, IEEE Transactions on, Volume: 3 Issue: 5, pp. 698 –713, Sept. 1992

[16] I. Dagher, M. Georgiopoulos, G.L. Heileman, and G. Bebis, “An ordering algorithm for pattern presentation in fuzzy ARTMAP that tends to improve generalization performance” Neural Networks, IEEE Transactions on, Volume: 10 Issue: 4, pp. 768 –778, July 1999

[17] E. Gomez-Sanchez, Y.A. Dimitriadis, Cano-Izquierdo, J.M., and Lopez-Coronado, J., “ μARTMAP: use of mutual information for category reduction in Fuzzy ARTMAP” Neural Networks, IEEE Transactions on, Volume: 13 Issue: 1, pp. 58 –69, Jan. 2002

[18] S. Kadambe, and G.F. Boudreaux-Bartels, “Application of the wavelet transform for pitch detection of spe ech signals”, Information Theory, IEEE Transactions on, Volume: 38 Issue: 2, pp. 917 –924,March 1992

[19] S.J. Mallat and W.L. Hwang, ”Singularity Detection and Processing with Wavelets”, Information Theory, IEEE Transactions on, Volume: 38 Issue: 2, pp. 617 –642,March 1992

[20] S.H. Chen, and J.F. Wang, “Noise-robust pitch detection method using wavelet transform with aliasing compensation” Vision, Image and Signal Processing, IEE Proceedings-, Volume: 149 Issue: 6, pp. 327 -334, Dec. 2002

[21] S.G. Mallat, “A Wavelet Tour of Signal Processing”, Academic Press, 2nd edition (September 15, 1999)

[22] R.P. Cohn, “Robust voiced/unvoiced speech classification using a neural net” Acoustics, Speech, and Signal Processing, 1991. ICASSP-91, 1991 International Conference on, vol.1, pp. 437 -440,April 1991