

Discrete Word Speech Recognition Using Hybrid Self-adaptive HMM/SVM Classifier *

Saeid Rahati Quchani ⁽¹⁾

Kambiz Rahbar ⁽²⁾

(1) Assistant professor, Islamic Azad University of Mashhad, IRAN

(2) M.S. Sashiraz electro-optic and laser technology research center, Shiraz, IRAN

Received: 3 June 2006; Reviewed: 10 July 2007; Accepted: 12 October 2007

Abstract This research addresses independent speaker's discrete word speech recognition (DWSR) using hybrid Self-adaptive Hidden Markov Model/Support Vector Machine (SA-HMM/SVM) classifier. Our proposed method includes two main units: preprocessing unit, and classification unit. The first unit tries to frame the speech wave into proper segments and extract time-frequency relevant features in a way to maximize relative entropy of time-frequency energy distribution among segments, and the second unit classifies words within the proper classes. To fulfill this goal, SA-HMM calculates word's likelihood to each existing class correspondently, and finally Support Vector Machine (SVM) classifies it by using all classes' likelihood as an input vector. To validate our proposed method, we test it within our IAUM dataset which contains Persian digits uttered by Persian speakers. Comparing the results with the outcomes of a similar method based on the original HMM shows around 1.2% improvement.

Key Words Discrete Word Speech Recognition, Local Orthogonal Discriminate Bases, Hybrid SVM/Self-adaptive HMM classifier

* Corresponding Author:

Address: Engineering Department, Islamic Azad University of Mashhad, IRAN

Tel: 0511- 6613000 Email: rahati@mshdiau.ac.ir

1. Introduction

Speech recognition is amongst the subjects which has been receiving special attention during recent decades. Generally we can categorize speech recognition methods into two main classes: first, methods employing meaningful linguistic parts such as words, syllables and phonemes, and second, methods that are based on signal processing. Each category has its own advantages and disadvantages. Since there are too many different views in breaking the words into syllables or phonemes depending greatly on linguistics features, as well as being un-identical in different languages, our proposed method employs signal processing techniques, to prepare relevant feature vectors.

In the classification process, usually modern speech recognition systems use statistical approaches which are based on the Bayes' rule. HMM, as one of these statistical approaches is one of the most powerful tools for signal modeling, because it uses the probability distribution associated with each state to model the temporal variability that occurs in speech across speakers or phonetic context via an underlying Markov process [1], [5]. The shortcoming of systems like HMM is that the complexity of the system is typically predefined or chosen through a cross-validation process. In other methods like SVM,

the dataset itself defines how complex the classifier needs to be. SVM is a new approach for pattern recognition problems with clear connection to the underlying statistical learning theory. Also, SVM cannot model the temporal structure of speech efficiently [1]; it always finds a global minimum [2]. Here, our proposed method employs hybrid HMM/SVM for classification process.

However, to have a good classification, extracting relevant features from a signal is very important. It is clear that, the performance of statistical classifier like HMM is improved by using methods which reduce the dimensionality of the problems without losing important information. So, here for the problem at hand, just like Shao et al [12], we try to select the features' bases in the way that maximizes relative entropy of time-frequency energy distribution among classes. We try to achieve this goal by using the basis functions which are well localized in the time-frequency plane as feature extractor called modified best-basis mentioned by Saito et al [6].

This paper is organized as follows: Section 2 presents the methodology of proposed method. Sub-section 2.1 describes local orthogonal discriminate bases. Sub-sections 2.2 and 2.3 present the SA-HMM and SVM respectively. Section 3 reports and

analyzes the experimental results, and finally, section 4 lists the conclusions.

2. Methodology

Fig.1 summarizes our proposed DWSR. As in our previous work [4], [10] in the first unit silence is eliminated from speech signal. To fulfill this goal, two criteria, i.e. energy and zero crossing, should be met. Within this system, mean and standard deviation of domain and passing rate of zero are estimated for calculating the statistical characteristics of background noise. Energy threshold and passing rate of zero are calculated by using statistical characteristics and maximum mean of domain in the distance; threshold mean is used for finding the distance where threshold often passes. It is assumed that starting and terminating points are out of this distance [5], and [11]. In fact, these statistical characteristics are used for finding high and low thresholds (ITL, ITU). Therefore, if the signal energy is higher than ITU or between ITU and ITL, and passing from zero is less than IZCT threshold, it would be speech. In the reverse situation, it would be silence.

After eliminating silence from speech, it becomes framed in the way that each frame could be processed as a speech segment with constant properties. In order to keep the frame's starting and terminating information,

overlapping is used. Frame's length should be as much to keep its static logical. In other words, it should be so small to keep semi-static properties of signal parameters; in case of being within the pace alternate limitations it will lose the energy of semi-static properties. On the other hand, the frame should be long enough to keep pace harmonics. For example, if it is chosen much bigger than the pace alternate, some parameters such as energy changes smoothly and therefore cannot reflect speech signal properties.

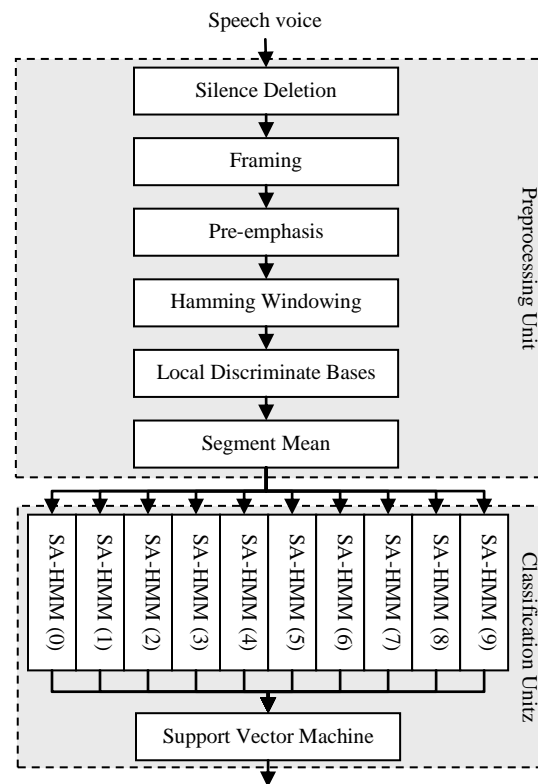


Fig.1 The general diagram of DWSR

After framing, the frames are passed through pre-emphasis filter to eliminate high frequencies of speech signal, drop the properties of spectrum boundaries and make the frames smooth, Then each frame is multiplied by Hamming window to minimize the effect of undesirable speech signal boundaries. The length of window should be as short as possible to accelerate the reaction against the domain variability. But, if the window length is too short, it will not provide a suitable means for producing even energy function.

Now, we can calculate time-frequency features using Local Discriminate Bases (LDB) unit for each frame. This unit is going to be studied in section II. In the Segment Mean unit for decreasing input vector size we represent each frame with its mean to prepare proper feature vector for classification unit.

Finally, in classification unit ten SA-HMMs trained represented for ten classes existing in our dataset. Each SA-HMM calculates the amount of likelihood for input features, and finally, SVM classifies it by using all classes' likelihood as an input vector.

2.1. Local Orthogonal Discriminate Bases (LDB)

The wavelet packet method is a generalization of wavelet decomposition (see Fig.2) which

offers a richer signal analysis. Wavelet packet atoms are waveforms indexed by three naturally interpreted parameters: position and scale as in wavelet decomposition, and frequency. For a given orthogonal wavelet function, a library of wavelet packets bases is generated. Each of these bases offers a particular way of coding signals, preserving global energy and reconstructing exact features. The wavelet packets can then be used for numerous expansions of a given signal. In the orthogonal wavelet decomposition procedure, the generic step splits the approximation coefficients into two parts. After splitting, we obtain a vector of approximation coefficients and a vector of detail coefficients, both at a coarser scale. The information lost between two successive approximations is captured in the detail coefficients. The next step consists of splitting the new approximation coefficient vector; successive details are never re-analyzed.

The original best-basis method introduced by Coifman, et. al.[7] extracts relevant features from signal for signal compressing purposes in two stages: first by expanding signal into an orthogonal bases dictionary (i.e., a redundant set of wavelet packet bases or local sine/cosine bases having a binary tree structure) and second by minimizing a certain

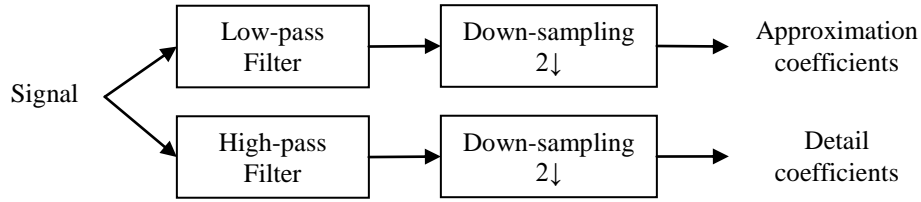


Fig. 2. Single-level discrete 1-D wavelet transform

X ₀₀							
X ₁₀				X ₁₁			
X ₂₀		X ₂₁		X ₂₂		X ₂₃	
X ₃₀	X ₃₁	X ₃₂	X ₃₃	X ₃₄	X ₃₅	X ₃₆	X ₃₇

Fig.3. Wavelet packet decomposition of signal X at depth 3 with best selected tree

information cost function through searching this binary tree. Consider the one-dimensional case, starting with the root node. The best tree is calculated using the following scheme. A node N is split into two nodes N₁ and N₂ if and only if the sum of the entropy of N₁ and N₂ is lower than the entropy of N. This is a local criterion based only on the information available at the node N. For instance Fig.3 shows wavelet packet decomposition of signal X at depth three. The best tree selected based on entropy criteria.

Several entropy type criteria can be used. If the entropy function is an additive function along the wavelet packet coefficients, this algorithm leads to the best tree [9]. For classification, Saito, et. al. [6] substitute certain

information cost function by symmetric relative entropy which is defined for two classes as follows:

$$J(p, q) = \sum_{i=1}^n p_i \log \frac{p_i}{q_i} + \sum_{i=1}^n q_i \log \frac{q_i}{p_i} \quad (1)$$

Where $p = \{p_i\}_{i=1}^n$ and $q = \{q_i\}_{i=1}^n$ are the sequences of normalized energy distributions of signals belonging to each class.

The local discriminate basis algorithm (LDB) as described in [6] is as follow:

1. Selecting an orthogonal bases dictionary which specifies QMFs for a wavelet package dictionary or selecting the local cosine or sine dictionary
2. For each class, construct a time-frequency energy map by:

- a. Normalizing each signal by the total energy of all signals of that class,
 - b. Expanding that signal into the tree-structured subspaces, and accumulating the signal energy in each coordinate,
3. Computing the discriminate measure symmetric relative entropy J among L distributions time-frequency energy maps for each node,
 4. Pruning the binary tree by eliminating children nodes where sum of their discriminate measures is smaller than or equal to the discriminate measure of their parent, and
 5. Ordering and selecting most discriminate basis vector by their discrimination power for constructing classifiers.

2.2. Self-Adaptive Hidden Markov Model (SA-HMM)

There are different models for signals-modeling, such as DFA, Mealy, Poisson, etc. which are classified into two general categories: Statistical and Deterministic models [5]. HMM $\theta(\pi, A, B)$ is one of the statistical signals-modeling schemes which is characterized as follow:

π : The vector of the initial state probabilities, that contains the probability of the (hidden) model being in a particular hidden state at time $t = 1$.

$A = \{a_{ij}\}$: The state transition matrix, holding the probability of a hidden state given the previous hidden state.

$B = \{b_j(k)\}$: The confusion matrix, containing the probability of observing a particular observable state given that the hidden model is in a particular hidden state.

An important point to remember is that the number of states in classical HMMs was usually predefined and fixed during training. The basic philosophy of HMM is that the signal can be well modeled, if its parameters are carefully and correctly chosen, i.e., successfully trained. We train HMM with known samples and finally obtain a model that is the nearest to the signal source, in the sense of a predefined criterion; e.g., maximum likelihood (ML) in the Baum–Welch training algorithm. In pattern recognition applications, different signal sources probably have different state numbers, thereby cannot be well modeled by HMMs with a fixed state number. If the numbers of predefined states are greater than the real word, then the training takes more time, so it needs more samples to float additional states. On the other hand, if the number of states is less than the real word, then the signal cannot be well modeled [3], [8]. According to Self-adaptive HMM $\theta(N, \pi, A, B)$ design, an HMM automatically matches its states numbers (N) to the real state number of the signal source which is being modeled. The idea behind this design is that the true states

numbers had less entropy than other false states numbers [3]. The entropy (H) for the model can be calculated approximately by sum of partial entropies, i.e.:

$$H(\theta) = H(\pi) + H(A) + H(B) \quad (2)$$

where:

$$H(\pi) = -\sum_j^N \pi_j \log \pi_j \quad (3)$$

$$H(A) = -\sum_i^N \sum_j^N a_{ij} \log a_{ij} \quad (4)$$

$$H(B) = -\sum_j^N \sum_k^M b_j(v_k) \log b_j(v_k) \quad (5)$$

where M is the number of observable symbols and $V = \{v_1, v_2, \dots, v_M\}$ is the symbol set.

2.3. Support Vector Machine (SVM)

SVM is a learning system that uses a hypothesis space of linear functions in a high dimensional feature space to estimate decision surfaces directly rather than modeling a probability distribution across training data. It uses support vector (SV) kernel to map the data from input space to a high-dimensional feature space which facilitates the problem to be processed in linear form. SVs are samples that have non-zero multipliers at the end of optimization process which is referred to equation (7).

SVM always finds a global minimum because it usually tries to minimize a bound on the

structural risk, rather than the empirical risk [2]. Empirical risk is defined as measured mean error rate on the training set as bellow:

$$R_{\text{emp}}(\alpha) \equiv \frac{1}{2l} \sum_{i=1}^l |y_i - f(x_i, \alpha)| \quad (6)$$

where l is number of observation, y_i is class label and x_i is sample vector. Structural risk is defined as a structure of divided entire class of function into nested subset and finding the subset of function which minimizes the bound on the actual risk.

SVM achieves this goal by minimizing the following Lagrangian formulation:

$$L_P \equiv \frac{1}{2} \|w\|^2 - \sum_{i=1}^l \alpha_i y_i (x_i \cdot w + b) + \sum_{i=1}^l \alpha_i \quad (7)$$

where α_i is positive Lagrange multipliers

3. Experimental Results

The proposed method is validated by using IAUM Persian digits dataset. This dataset contains one thousand digits uttered by hundred Persian speakers. It should be noted that this dataset has ten classes labeled from zero to nine.

Here, we are going to present some outcomes of each two-main unit: preprocessing unit, and classification unit. In the first unit, as described in Section 2, for all items in dataset silence eliminated from speech signal through meeting two criteria: energy and zero crossing. Then, speech signals are framed into proper units. Each frame passed through pre-emphases

filter and multiplied by Hamming window to eliminate high frequencies of speech and minimize the effect of undesirable speech signal boundaries. At the end, LDB time-

frequency features extracted from each frame (Fig. 4) and represented with its mean to feed-forward in the second main unit as a proper feature vector for classification.

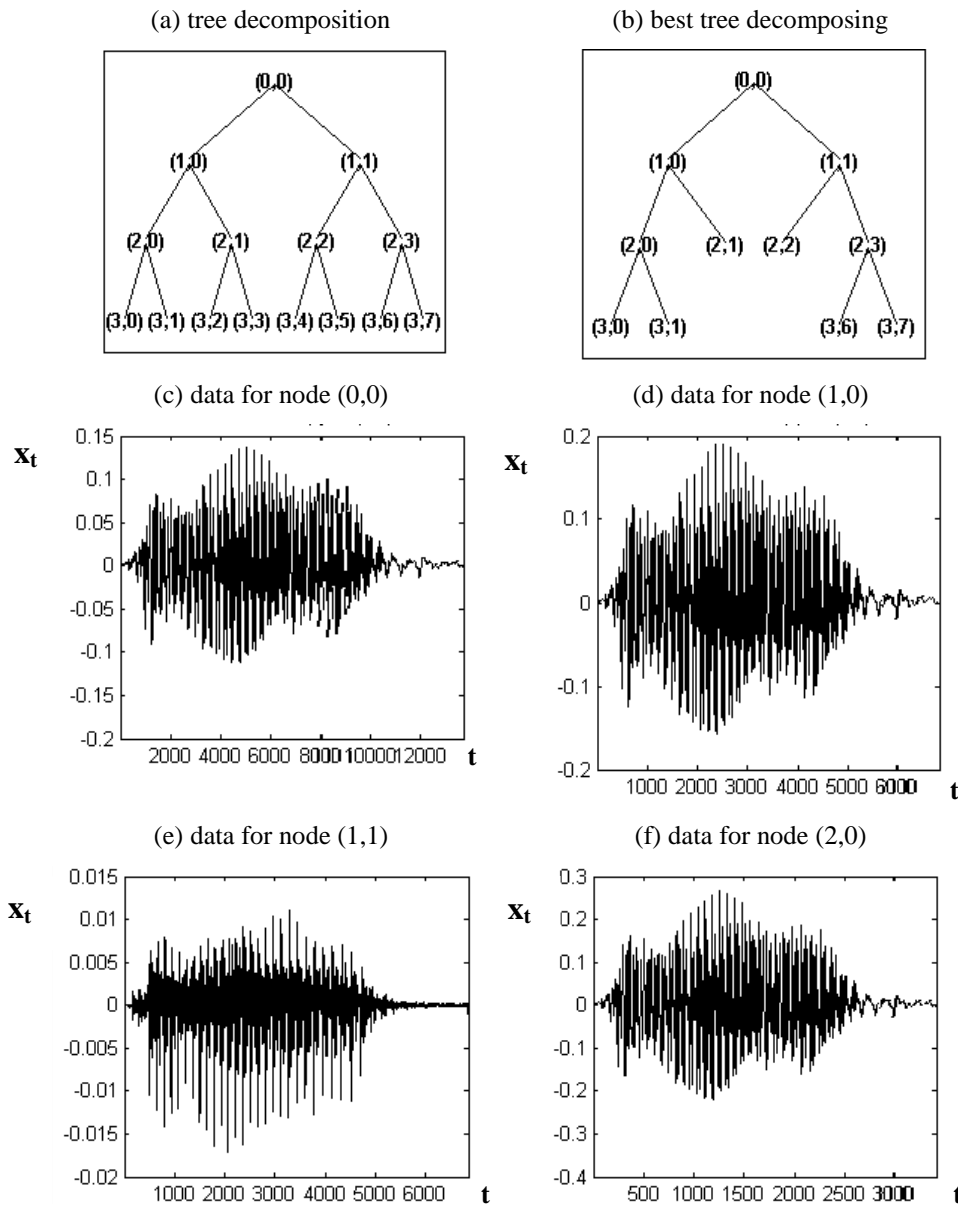


Fig.4. Wavelet packet decomposition and best selected tree for Persian number 2 uttered by a Persian speaker. (a) Wavelet packet decomposition, (b) LDB (best) selected tree, and the rest are wavelet approximation and detail coefficients

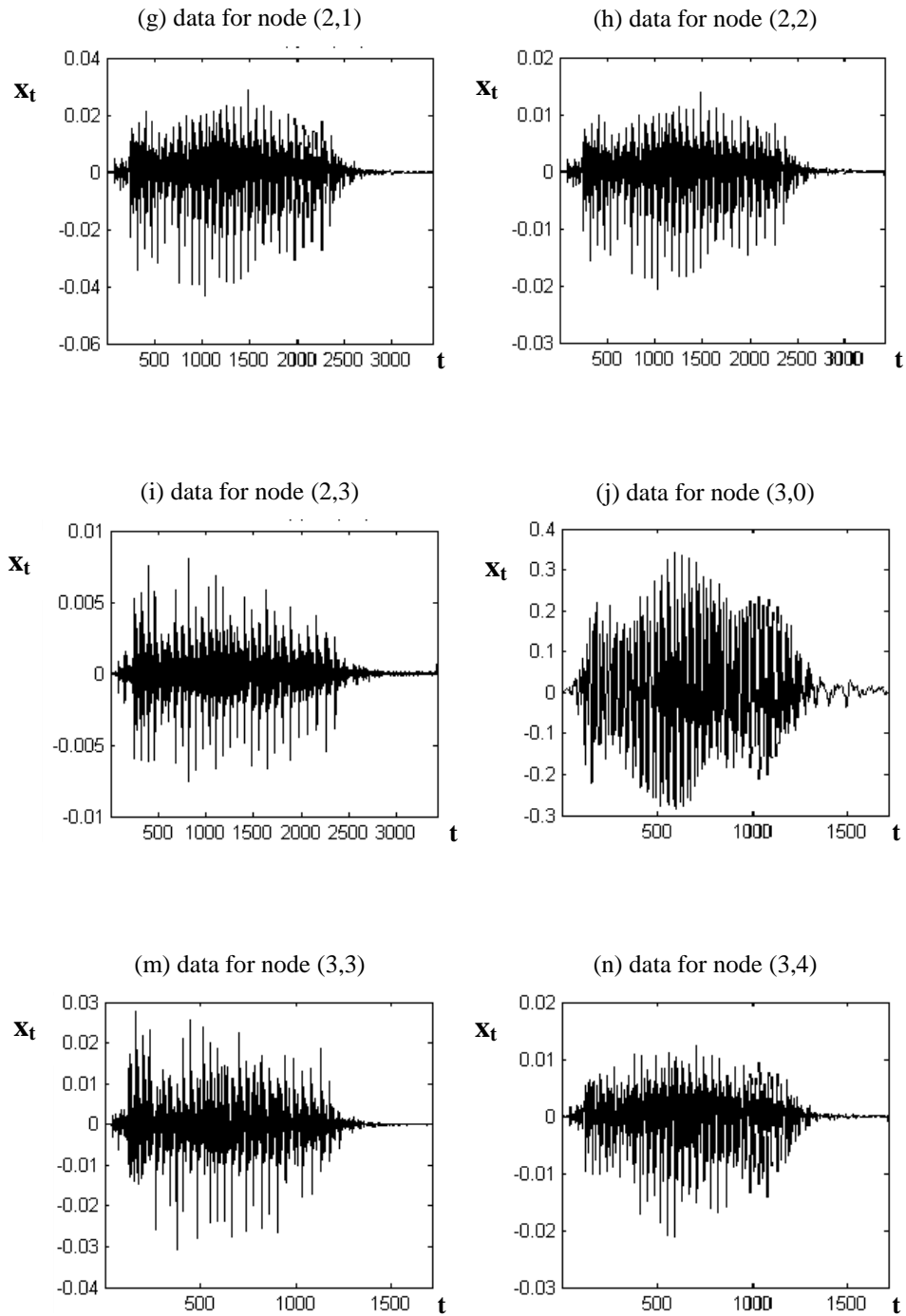


Fig.4. Wavelet packet decomposition and best selected tree for Persian number 2 uttered by a Persian speaker. (a) Wavelet packet decomposition, (b) LDB (best) selected tree, and the rest are wavelet approximation and detail coefficients

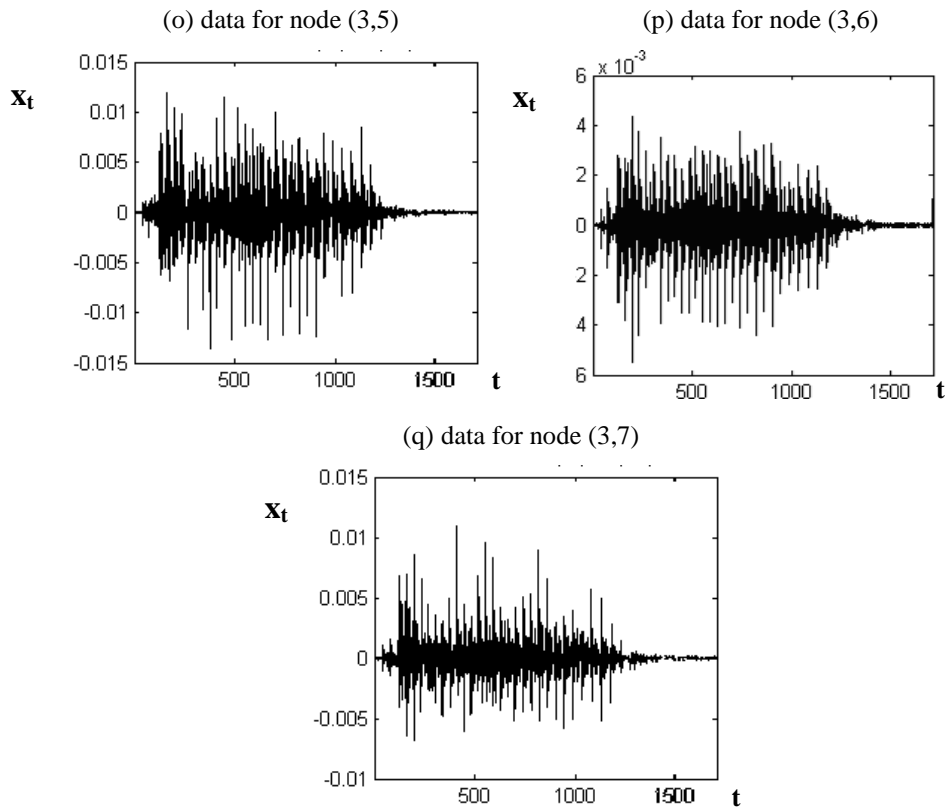


Fig.4. Wavelet packet decomposition and best selected tree for Persian number 2 uttered by a Persian speaker. (a) Wavelet packet decomposition, (b) LDB (best) selected tree, and the rest are wavelet approximation and detail coefficients

In classification unit, ten SA-HMMs trained represented for ten existing classes in our dataset. Each SA-HMM calculated the amount of likelihood for input features and finally SVM classified it by using all classes' likelihood as an input vector. We used part of this dataset for training HMMs, SA-HMMs and SVM separately. After each HMM and SA-HMM were trained, we trained SVM with RBF kernel using particular dataset achieved from main dataset by applying HMMs and SA-HMMs.

Tables 1, 2 and Table 3 show the classification rate when fix and adaptive states number were identified in advance while LDB tree-structured deep varies from three to five.

Table 1. Comparison of HMM base classification error (LDB tree-structured deep = 3)

Method	Error Rate on Train Dataset	Error Rate on Test Dataset
Proposed method	12.5%	13.6%
Proposed method using HMM instead of SA-HMM	13.1%	16.2%

Table 2. Comparison of HMM base classification error (LDB tree-structured deep = 4)

Method	Error Rate on Train Dataset	Error Rate on Test Dataset
Proposed method	9.3%	10%
Proposed method using HMM instead of SA-HMM	8.9%	11%

Table 3. Comparison of HMM base classification error (LDB tree-structured deep = 5)

Method	Error Rate on Train Dataset	Error Rate on Test Dataset
Proposed method	8.5%	9.1%
Proposed method using HMM instead of SA-HMM	8.3%	11.3%

These tables show that when the model entropy decreases, the recognition power increases because of better modeling. Additionally, using LDB with deeper tree-structure decreases error rate.

References

1. A. Ganapathiraju, J.E. Hamaker and J. Picone, “Application of Support Vector Machines to Speech Recognition,” *IEEE Trans. Signal Processing*, vol. 52, no. 8, Aug. (2004).
2. C.J.C. Burges, “A Tutorial on Support Vector Machines for Pattern Recognition,” *Knowledge Discovery Data Mining*, vol. 2, no.2, pp. 121-167, (1998).
3. J. Li, J. Wang, Y. Zhao, and Z. Yang, “Self-Adaptive Design of Hidden Markov Models,” *Pattern Recognition Letters* 25, pp. 197-210, (2004).
4. K. Rahbar and M. Rahbar, “Discrete Words Speech Recognition (DWR) Using Self-adaptive Hidden Markov Model (SAHMM),” in *Proc. Int. Conf. GSPx 2005 Pervasive Signal Processing, USA*, (2005).

4. Conclusions

In this paper we studied independent speakers DWSR based on hybrid SA-HMM/SVM classifier. Our method includes two main units: a) Preprocessing unit that tries to frame the speech into proper segments and extract time-frequency relevant features through maximizing relative entropy of time-frequency energy distribution among segments, and b) Classification unit which classifies words into proper classes by calculating degree of words likelihood with SA-HMM and classifying it through SVM classifier by using all classes’ likelihood as an input vector.

We validated this method within the IAUM dataset and found that LDB with deeper tree-structure provides better features vector for classification. Additionally, by decreasing the model entropy, the recognition power increases.

5. L.R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," in Proc. IEEE 77, No. 2, Feb (1989).
6. N. Saito and R.R. Coifman, "On Local Orthogonal Bases for Classification and Regression," in Proc. IEEE Int. Conf. ICASSP-95 Acoustics, Speech, and Signal Processing, vol. 3, pp. 1529-1532, May (1995).
7. R.R. Coifman and M.V. Wickerhauser, "Entropy-Based Algorithms for Best Basis Selection," IEEE Trans. Information Theory, vol. 38, Issue 2, Part 2, pp. 713-718, Mar. (1992).
8. S. Kwong, Q.H. He, K.W. Ku, T.M. Chan, K.F. Man and K.S. Tang, "A genetic Classification Error Method for Speech Recognition," Signal Processing, no. 82, pp.737-748, (2002).
9. S. Mallat, "A Wavelet Tour of Signal Processing, Academic Press," (1998).
10. S. Rahati, K. Rahbar, "Local Orthogonal Discriminate Bases to Hybrid SVM/Self-Adaptive HMM Classifier for Discrete Word Speech Recognition," IEEE Int. Symp. ISSPIT 2006, Vancouver, Canada., August (2006).
11. V. Digalakis, S. Tsakalidis, C. Harizakis and L. Neumeyer, "Efficient Speech Recognition using Subvector Quantization and Discrete-Mixture HMMS," Computer Speech and Language, no.14, pp. 33-46, (2000).
12. Y. Shao and C.H. Chang, "Wavelet Transform to Hybrid Support Vector Machine and Hidden Markov Model for Speech Recognition," Circuits and Systems, 2005, in Proc. IEEE Int. Symp. ISCAS 2005, vol. 4, pp. 3833-3836, May (2005).