A Feature Extraction Method for Speech Recognition Based on Temporal Tracking of Clusters in Spectro-Temporal Domain

Nafiseh Esfandian[1], Farbod Razzazi[2], Alireza Behrad[3]
[1] Department of Electrical Engineering, Qaemshahr Branch, Islamic Azad University, Qaemshahr, Iran
[2] Electrical and Computer Engineering, Islamic Azad University, Science and Research Branch, Tehran, Iran
[3]Faculty of Engineering, Shahed University, Tehran, Iran
N.Esfandian@qaemshahriau.ac.ir, razzazi@srbiau.ac.ir, behrad@shahed.ac.ir

Abstract— **In this paper, a novel approach is proposed for secondary feature extraction based on clusters tracking in spectro-temporal domain. Because of high dimensionality of the spectro-temporal features space, this domain is unsuitable for practical speech recognition systems. In order to reduce the dimensions of the feature space, weighted K-means (WKM) clustering technique is applied to spectro-temporal domain. The elements of mean vectors and covariance matrices of clusters are considered as the feature vector of each frame. However the cluster locations change gradually over the time. The main approach is based on the idea that the variations in clusters locations should be temporally tracked frame by frame and the parameters of these variations are considered in the extraction of secondary feature vectors of each speech frame. Several models are used to register the clusters in the new coming frame. In addition, a new architecture is proposed to classify the speech frames by a combining classifier using both tracked and non-tracked secondary features. The assessments were conducted for the proposed feature vectors on classification of several subsets of TIMIT database phonemes. Using tracked secondary feature vectors, the result was improved to 77.4% on voiced plosives classification which was relatively 1.8% higher than the results of non-tracked secondary feature vectors. The results on other subsets showed good improvement in classification rate too.**

Keywords-Speech recognition; Feature extraction; Clustering methods; Image matching; Auditory system; Speech processing

I. INTRODUCTION

The main goal of speech features extraction methods is the extraction of maximum relevant information in the extracted features while reducing the amount of data to minimum. Mel scaled frequency Cepstral coefficients [1] and spectro-temporal features [2-4] are the most frequently used representations of audio signal that are both inspired from physiological audition findings. In recent years, a computational auditory model has been proposed based on psycho-acoustical and neuro-physiological findings in early and central stages of the auditory section of the brain [2]. High dimensionality of spectro-temporal features space makes the system impractical work in this domain and affects the parameter estimation accuracy in the training phase of the speech classifier. Traditional dimensionality reduction methods such as PCA, LDA, MPCA and neural network based features reduction techniques have been used to reduce the number of features in spectro-temporal domain [3, 5]. However, these common methods are not exactly compatible with the speech nature and classification problems.

The proposed method in this paper follows our previous research which clustering methods had been used to reduce the dimensions of spectro-temporal feature space in order to extract secondary features extraction [6, 7]. It had been shown that in the new clustered features space; phonemes are more separable because the classes' information is concentrated in the specific parts of the features space. For this purpose, Gaussian Mixture Model (GMM) and weighted K-means (WKM) clustering methods had been used for features space clustering and the mean vectors and covariance matrices elements of the clusters had been considered in the secondary feature vector of each frame. The results had been shown that better results in phoneme classification had been obtained using WKM clustering in comparison to GMM clustering. In addition, the computational complexity of the secondary feature extraction using WKM clustering was less than GMM clustering. Therefore, in this study, WKM clustering method is used to reduce the spectro-temporal features into a few effective secondary features for each frame. One of the open issues in the previous study was the order of clusters in the feature vector. This sorting order determines the consistency in the meaning of each element of the feature vector and dramatically affects the classification rate, if arranged inappropriately. In previous study, it was assumed that the energies of clusters in each frame are the intrinsic characteristics of the phonemes and can be considered as the measure of clusters sorting in a features vector. Therefore, to overcome this deficiency, the clusters had been sorted based on their energy. Although this presumption is often true in central parts of phonemes, especially in long duration phonemes, however, this cannot be assumed in the gradual interchange of co-articulated phonemes which are frequently occurred in an uttered sentence. The motivation of this study is to improve the classification rate of phonemes by considering the correct order of clusters in a frame. The main approach is based on the idea that the variations in clusters locations should be temporally tracked frame by frame and the order of clusters of each frame should be registered in the whole sequence based on a reference vector before sending them to the classifier. There are various techniques for feature matching and registration over the

consecutive frames in image and video processing [8]. In the present study, the Euclidean distance measure is used for temporal tracking of the clusters. In other words, the clusters centers of two frames are matched based on their weighted Euclidean distance. The clusters are sorted using two strategies in feature vectors. In the first strategy, the clusters of each frame are sorted based on energy measure and in the other strategy; the clusters are re-arranged using temporal matching results of the secondary feature vectors sequence. The results of these two strategies are compared to each others. In addition, combining mechanisms of two features sorting models is applied to optimize the classification rate.

The organization of the paper is as follows. Spectro-temporal representation of speech is briefly reviewed in section 2. The proposed phoneme feature extraction and clusters tracking algorithms in spectro-temporal domain are introduced and formulated in sections 3 and 4. Experimental results and performance evaluation of the proposed features on standard datasets for phoneme classification task are presented in section 5. The paper is concluded in section 6.

II. AUDITORY MODEL

Auditory model is a mathematical model for internal ear and the first layer of auditory brain section that is proposed for speech processing applications in recent years [2, 3]. This model is obtained using neuro-physiological, biophysical, and psycho-acoustical investigations at various stages of the auditory system. The block diagram of auditory model is shown in Fig. 1. It consists of two basic stages. In the primary stage of auditory model, the acoustic signal is transformed into the auditory spectrogram. The central stage analyzes the spectrogram to extract the spectro-temporal features.

A. The Primary Stage of the Auditory Model

While the audio signal passes through the ear, the neural sensors of the basilar membrane of the cochlea convert one dimensional audio signal into a two-dimensional auditory spectrogram image which the frequency axis of this 2D image is a tonotopic (nearly logarithmic) axis. Basilar membrane can be considered as a band-pass filter bank. This filter bank includes 128 asymmetric band-pass filters with the frequency responses which are uniformly distributed along the tonotopic axis.

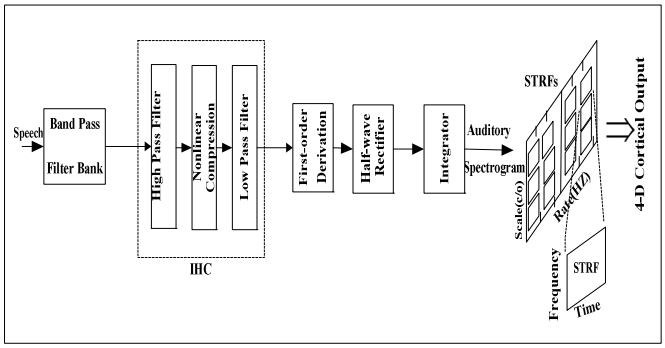

Figure 1. The block diagram of auditory model

The cochlear filters outputs are converted into auditory nerve patterns by an inner hair cell stage (IHC). IHC stage consists of a high-pass filter in time domain, an instantaneous nonlinear compression and a time domain low-pass filter. The last part of this stage is a model of lateral inhibitory network (LIN) activity, which increases the frequency selectivity of the cochlear filters. LIN is approximated by a first order derivative along the tonotopic axis the final output of this stage, is obtained by using a half wave rectifier and an integrator during a short time window [2].

B. The Cortical Stage of Auditory Model

The primary auditory stage of the brain analyzes the auditory spectrogram as an image. At this stage, a two-dimensional wavelet transform of auditory spectrogram is calculated. This transform is performed using a spectro-temporal mother wavelet, similar to a two-dimensional Gabor function. In other words, the spectral and temporal modulation contents of the auditory spectrogram are estimated via a bank of modulation-selective 2-D filters. Each filter is tuned to a spectral-temporal modulation index pair. Spectro-temporal impulse responses of these filters are called spectro-temporal response fields (STRFs). Each of STRFs in the bank of directional selective filters can be generated by multiplying two uncoupled complex functions of time and frequency. The resulted STRF is the real part of this multiplication. There are two primitive 2-D STRF types which are named upward (+) and downward (-) respectively which are demonstrated as positive and negative rates respectively. The output of each branch of filter-bank is computed by a convolution of its STRF with the input auditory spectrogram. The output of each branch of filter-bank is computed by a convolution of its STRF with the input auditory spectrogram. Therefore, the cortical representation of speech has four dimensions, scale (Ω in cycles / octave) which is the STRF scaling factor along frequency axis, rate or velocity (ω in Hz) which is the temporal scaling factor of STRF, frequency (f, the number of the band-pass filter) and time (t, the frame number). The two latter parameters show the position of the point in the filtered spectrogram.

The dimensions of this feature space are very large which may bring the system to the curse of dimensionality limitation in the training phase of a speech recognition system. Therefore, the reduction of features space dimensions is a crucial task to train the parameters of artificial speech classifiers efficiently.

III. PREPARE YOUR PAPER BEFORE STYLING PHONEME FEATURE EXTRACTION USING WKM CLUSTERING IN THE SPECTRO-TEMPORAL SPACE

In the first stage of the proposed feature extraction method using WKM clustering, the auditory spectrogram of a speech frame was calculated. Then, the auditory spectrogram is analyzed by a bank of spectro-temporal modulation selective filters that each filter was tuned to a different rate-scale pair. The outputs of the upward and downward cortex STRFs (two amplitudes of the complex outputs) were calculated for each point of four-dimensional space of scale (Ω in cycles / octave), rate (ω in Hz), frequency and time as the coordinates of the auditory output.

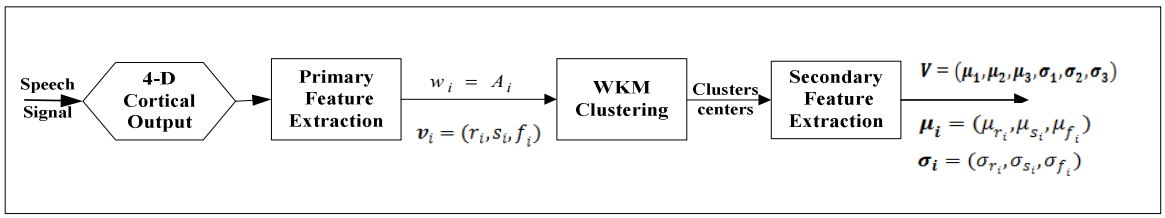

Figure 2. The block diagram of the proposed feature extraction mechanism

As it is common in spectro-temporal space literature, the downward parameters may be represented as negative rates to make a unified 3D space for each frame. In other words, to combine both upward and downward STRF responses, the output of both sets of filters were concatenated in the rate axis. Therefore, in each frame, two 3D cubes with rate axis, scale axis and frequency axis were combined to have one new 3D cube with twice lengthened in the rate axis. Scale, rate and frequency are the coordinates of the point in the new space in each frame which should be considered as the primary feature vectors. In this paper, the auditory spectrogram was obtained using an infinite impulse response filter bank with 128 frequency channels between 180 and 7246 Hz at the resolution of 24 channels per octave. In addition, a time constant of 8ms was used for the leaky time integration and filter-bank outputs were sampled every 4 ms to compute the auditory spectrogram. Temporal parameter of the filters (rate), ranging from 2 to 128 Hz and spectral parameter of the filters (scale), ranging from 0.25 to 8 cycle/octave, were considered to represent the spectro-temporal modulations of the speech signal. Thus, the dimensions of the resulted spectro-temporal feature space were very large (11 (scale filters) ×26 (rate filters) ×128 (frequency channels) =36608 attributes). Therefore, in the proposed cluster-based feature extraction method, WKM clustering is used to extract the features with informative discriminative attributes. It means that the primary features space was segmented into clusters using WKM clustering algorithms. As a result, the main clusters in each speech frame were determined and new feature vectors were extracted with reduced dimensions.

In WKM clustering method, K clusters are distributed in the space to model the space. Each sample in the feature space is assigned to the nearest cluster. In addition, a weight is assigned to each feature vector, which is determined according to the importance of each sample in the quality of the clustering overall fitness function [9]. In this clustering algorithm, the weight of each point is considered in the clusters estimation. The weight of each point may be interpreted as a soft repetition of the vector in the cluster center calculation. The magnitude component of each point was considered as the weight in WKM algorithm to emphasis on high energy points of the space in the clustering procedure. The weight of a cluster was defined as the mean magnitude of all recruited samples for each cluster.

A. Secondary Feature Extraction Using WKM Clustering

The secondary feature extraction mechanisms in each frame are shown in Fig. 2 using WKM clustering methods. In the proposed feature extraction method, the samples in the primary feature space (denoted as the vector $\mathbf{v}_i$) were applied to the clustering algorithm to extract secondary features. Each point in the input space was defined as a three dimensional vector $\mathbf{v}_i = (r_i, s_i, f_i)$. In this vector, r denotes the rate, s is the scale, f is the frequency of the output of upward or downward STRFs at each points of the spectro-temporal space. The magnitude components of points $w_i = A_i$ were considered as the weighting factor of input vectors.

These primary feature vectors $\mathbf{v}_i$ were clustered using WKM algorithm assuming diagonal covariance matrix and the centers of clusters were considered as secondary feature vectors. The mean vector and covariance matrix elements of the clusters were considered in the secondary feature vector of each frame as $\mathbf{V} = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\mu}_3, \boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2, \boldsymbol{\sigma}_3)$ assuming three clusters for each frame. $\boldsymbol{\mu}_i$ and $\boldsymbol{\sigma}_i$ are mean and variance vectors of i^{th} cluster. Each mean vector consists of three components as $\boldsymbol{\mu}_i = (\mu_{r_i}, \mu_{s_i}, \mu_{f_i})$ and variance vector consists of three components as $\boldsymbol{\sigma}_i = (\sigma_{r_i}, \sigma_{s_i}, \sigma_{f_i})$. Thus, the secondary feature vector had 18 elements.

IV. TWO MODELS FOR CLUSTERS SORTING IN THE SECONDARY FEATURE VECTORS

A. Clusters sorting in feature vectors using energy measure

In the first strategy, it is assumed that the center of cluster with larger magnitude bears more information. Therefore, the features are sorted in accordance to the clusters amplitude. Time variations of clusters are not considered in this method and the clusters centers sorting are performed based on energy measure. It means that the magnitude component of clusters centers are sorted descending. In addition, variance components of clusters centers are sorted according to the value of their mean components as $\mathbf{V} = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\mu}_3, \boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2, \boldsymbol{\sigma}_3)$.

B. Clusters Sorting in Feature Vectors Using Temporal Tracking Results

The former cluster ordering strategy may cause the system sensitive to noisy conditions; because the cluster locations are changed in the scale, rate and frequency axes during time. Therefore, the results of temporal tracking of the clusters over the time are used for features sorting in the second strategy. In this method, the clusters centers are matched to a reference vector using a distance measure. The result of this matching determines the order of clusters.

While the locations and shapes of the cluster centers change gradually over the time, the clustered should be tracked during consequent frames. In this study, Euclidian distance measure was employed to track the clusters. Two mechanisms in various conditions were used for temporal clusters tracking.

C. Clusters Matching Over the Consecutive Frames

In the first mechanism, the cluster centers are matched over the consecutive frames. Therefore, the distances between each cluster center of the current frame and all cluster centers of the previous frame are computed. Then, the best match of the current frame cluster centers with the cluster centers of the previous frame is determined by minimizing the Euclidean distance between their corresponding features. Three clusters were assumed for each frame. Therefore, a 3×3 distance matrix is obtained for each frame. Each element of distance matrix, $dis(i,j)$ is defined as

$$dis(i,j)=\sum_{k=1}^{n}(\mathbf{C}_{iC}(k)-\mathbf{C}_{jp}(k))^2 \qquad (1)$$

Where $\mathbf{C}_{iC}$ and $\mathbf{C}_{jp}$ are the ith cluster center of the current cluster and the jth cluster center of previous frame respectively and n is the numbers of features in each cluster center. Each cluster center vector have six components which is defined as

$$\mathbf{C}_i=(\mu_{r_i},\mu_{s_i},\mu_{f_i},\delta_{r_i},\delta_{s_i},\delta_{f_i}) \qquad (2)$$

In this matching strategy, the clusters centers of the first frame of each phoneme are sorted in descend according to their amplitudes. Then, the cluster centers of the next frames are rearranged by the matching results using the status matrix that defines possible permutations of the clusters in features vectors. Status matrix is defined as

$$\mathbf{S}=\begin{pmatrix}1&1&2&2&3&3\\2&3&1&3&2&1\\3&2&3&1&1&2\end{pmatrix} \qquad (3)$$

Each column of the status matrix shows the position of clusters centers in each speech frame after cluster matching over the time. In fact, the column number of the status matrix shows the matching status number. Although in some frames, the positions of clusters centers remain unchanged, (e.g. the first column of the status matrix); however, the positions of most clusters centers change over the time. To determine the clusters matching result of each frame, the cost vector is defined using the status matrix. Cost(k) is the cost function of the kth matching status according to the kth columns of the status matrix. The cost function is calculated for all columns of status matrix in each frame as

$$\text{Cost(k)}=\sum_{i=1}^{3}w_i.dis(i,S(i,k)) \qquad (4)$$

where $\text{w}_\text{i}\in\{0,1\}$ is the death/birth factor of the ith cluster. This factor is zero, if the minimum distance between a cluster in current frame and the clusters of previous frame is more than the empirical threshold value T.

$$w_i=\begin{cases}0 & if \quad \min(dis(i,j))\geq T\\ 1 & Otherwise\end{cases} \qquad (5)$$

If this factor is zero, it means that a cluster is dead and a new cluster is born in the new frame. In this case, the ith cluster with $\text{w}_\text{i}=0$ is not considered in the cluster matching procedure. Finally, the best matching is determined regarding to the minimum value of the cost functions of the status matrix columns as

$$\text{BM}=\underset{\text{k}}{\text{Argmin}}(\text{Cost}) \qquad (6)$$

$$\text{B}est\ \ Matching(i)=S(i,BM)\ \ ,i=1,2,3 \qquad (7)$$

The clusters locations of the current frame (except the first frame) are permuted according to the best matching that is obtained from Equation (7).

D. Clusters Matching Using the Reference Vector

In another matching mechanism, the clusters of the primary feature vectors which are sorted using the energy measure are matched with the clusters of a reference vector. Two reference vectors are assessed in this clusters matching strategy.

In GM strategy, the frames of all phonemes are matched according to a global references frame. This global reference vector is calculated by averaging all of the training feature vectors of all phonemes. The calculated global reference vector is also used for clusters matching in the test phase of phoneme classification. Then, the distances between each cluster center of current frame and the cluster center of the global reference vector is computed using Equation (1). In this case, $\mathbf{C}_{jp}$ is the jth cluster center of the global reference vector. Finally, the cluster centers are rearranged using matching results that are obtained between each frame and the reference vector according to Equation (7).

In CBM mechanism, the reference vectors are determined by averaging between all feature vectors of each class in the training phase. In this mechanism, the number of reference vectors is depended on the numbers of the phonemes classes that should be classified. Assuming a class for each unknown utterance is a prerequisite for this reference vector selection mechanism, which contradicts with the main goal of a classification application. Therefore, the classifier architecture should be adapted to this tracking mechanism. The block diagram of this architecture is proposed in Fig. 3. In this architecture, the distances between each cluster center in the current frame and the cluster centers of the reference vector are calculated using the Equation (1). In this case, $\mathbf{C}_{jp}$ is the jth cluster center of a reference vector. In the training phase, the features vectors of each class are matched to the reference vector of the same class. In contrast, in the test phase, the features vectors of an unknown phoneme are matched to the reference vectors of all classes.

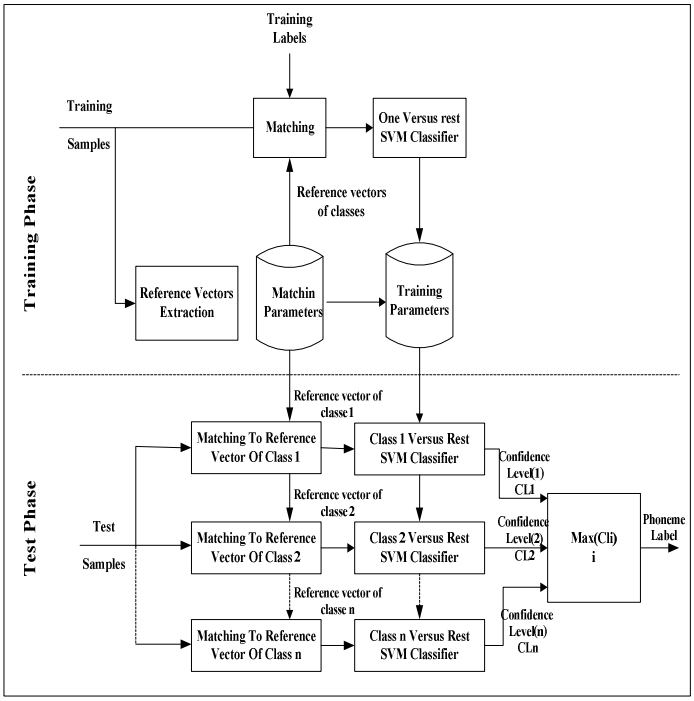

Figure 3. The block diagram of clusters matching mechanism using a class based reference vectors

The feature vectors that are matched with reference vector of i^{th} class are classified using the i^{th} class versus the rest binary SVM classifier. The class of an unknown phoneme is determined with respect to the maximum value of decision levels of the classifiers outputs. In other words, each frame of an unknown phoneme is matched with the best reference vector of the classes by using the confidence levels of the classifiers outputs.

E. Combining the Classification Results of Tracked and Non-Tracked Features

After a subtle error analysis of different proposed classification methods, it was observed that there is a considerable mismatch between error samples sets in tracked and non-tracked classifiers. This led us to design a combining classifier mechanism to tune the result.

In this mechanism, parallel classifiers are trained using two types of tracked and non-tracked secondary features vectors. To have consistent architectures for classifiers, the proposed one-versus-the-rest SVM classifier architecture was used for the phoneme classification in each branch. Finally, phoneme classification is performed using the combination of confidence values of the classifiers that are obtained in each branch. In this fusion strategy, $CL_1 = (CL_{11}, CL_{21}, ..., CL_{n1},)$ and $CL_2 = (CL_{12}, CL_{22}, ..., CL_{n2},)$ are the confidence levels vectors for each unknown phoneme that are obtained using the outputs of n classifiers of the first and second branches respectively. In this paper, the overall confidence of the i^{th} class CL_i is empirically evaluated by a few fusing rules:

$$Maximum\ rules\ \ CL_i = \max_i(CL_{i1}, CL_{i2}) \quad (8)$$

Finally, the maximum confidence value will indicate the winner class:

$$C = \max_i(CL_i) \quad (9)$$

V. EXPERIMENTAL RESULTS

A. Exprimental setup

The overall evaluation of the tracked features is tested on classification of main categories of phonemes. Although, /b/, /d/, /g/ as one of hard to discriminate set of phonemes was used as the benchmark of many studies in this field [10, 11]. Therefore, in this study, most of the experiments are conducted on /b, /d/, /g/ phonemes to evaluate and tune the tracking performance of proposed cluster tracking strategies. The evaluation of proposed feature extraction method is performed on clean speech and the phonemes are selected from TIMIT acoustic-phonetic continuous speech corpus which contains short sentences spoken by male and female speakers from 8 major dialect regions of the United States [12]. TIMIT contains 6300 sentences, where 10 sentences have been spoken by each of 630 male and female speakers.

In the first stage of the evaluation system, the phonemes were selected from TIMIT database. Then primary features were extracted from spectro-temporal space for each phoneme and secondary features vectors are obtained using new cluster-based feature extraction method. After features sorting procedure, the new features vectors were classified using the proposed classifier architecture. Radial basis function (RBF) was used as SVM kernel.

B. Phoneme Classification Results of one-versus-the-rest SVM

One of the determinant parameters of efficiency of matching strategies is death/birth factor. In Fig. 4, classification results of (/b/, /d/, /g/) phonemes in each matching strategy using death/birth threshold values are shown. The best phoneme classification rates are obtained with smaller death/birth threshold value in matching strategy 1. In fact, temporal tracking are often not performed for small death/birth threshold values and the clusters are sorted using energy measure; because the death/birth factor is zero ($w_i = 0$) for most of the frames in this case. It is clear from the results that first matching strategy was not successful for temporal tracking of the clusters. The best results of temporal tracking are obtained using T=1.7 for this strategy and T=1.5 was the optimum threshold for strategies 2 (GM) and 3 (CBM).

C. Combining the classification Results

The results of /b/, /d/, /g/ phonemes classification using energy-ordered and temporally tracked features and combining mechanism for the best death/birth threshold values are tabulated in table 1. The results show that the temporally tracked features gave better results in comparison to energy-ordered features. In addition, it can be observed that the best phoneme classification results are obtained by fusing two classifiers. This is due to the fact that the errors of two

classifiers are not the same. The classification rates on different categories of phonemes using energy-ordered and temporally tracked features and combining mechanism were evaluated and the results are tabulated in table 2. As it can be observed, the classification results using temporally tracked features was improved in comparison to energy-ordered features in all categories of phonemes. In addition, phoneme classification results were fine tuned using combining mechanism.

VI. CONCLUSION

In this paper, cluster tracking methods were employed for enhancing the discriminative behavior of spectro-temporal secondary features. Secondary features vectors were extracted using WKM clustering in spectro-temporal domain and the mean vectors and covariance matrices elements of the clusters were considered in the secondary features vector of each frame. Two strategies were used for clusters sorting in features vectors. In the first strategy, the clusters were sorted with respect to their energy in spectro-temporal space. In the second strategy, the cluster centers were sorted in feature vectors based on temporal tracking results. Various matching strategy were used for temporal tracking of the clusters. In overall, GM and CBM strategies were successful in comparison to the first matching strategy. In addition, fusing the classifiers showed good performance to cover the errors in tracked and non-tracked approaches.

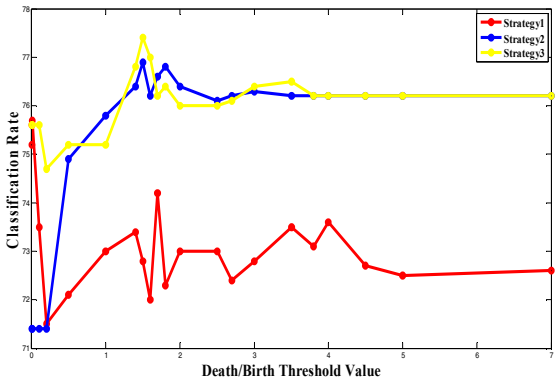

Figure 4. Classification results of (/b/, /d/, /g/) phonemes versus death/birth threshold

TABLE I. FRAME-WISE CLASSIFICATION RATES USING ENERGY-ORDERED FEATURE AND TEMPORALLY TRACKED FEATURES AND COMBINING MECHANISM ON /B/, /D/, /G/ PHONEMES

Matching Strategy	1	2	3
Classification Results *Energy-ordered features*	75.7	75.7	75.7
Death/Birth Threshold Value	1.7	1.5	1.5
Classification Results *Temporally tracked features*	74.2	76.9	77.4
Combining Mechanism *Classification Results*	76.3	77.6	77.8
Fusion Rule	Maximum	Maximum	Maximum

TABLE II. ME-WISE CLASSIFICATION RATES USING ENERGY-ORDERED FEATURE AND TEMPORALLY TRACKED FEATURES AND COMBINING MECHANISM IN MAIN CATEGORIES OF PHONEMES

	Class Phonemes	**Energy-ordered Feature Vectors**	**Temporally Tracked Feature Vectors**	**Matching Strategy**	**Combining Mechanism**
Consonant	***Voiced Plosives (b,d,g)***	75.6	77.4	CBM	77.8
	Unvoiced Plosives (p,t,k)	70.6	71.4	GM	71.9
	Voiced Fricatives (v,dh,z)	83.6	84.2	GM	84.7
	Unvoiced Fricatives (f,s,sh)	89.6	90.1	CBM	90.6
	Nasals (m,n,ng)	52.5	53.9	CBM	54.4
Vowel	***Front Vowel (ih,ey,eh,ae)***	64.4	65.2	CBM	65.9
	Back Vowel (uw,uh,ow,aa)	74.9	75.7	GM	76.2

REFERENCES

[1] S. Davis, P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition inmcontinuously spoken sentences," IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 28(4), pp. 357-366, 1980.

[2] T. Chi, P. Ru, S. A. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," Acoustical Society of America Journal, Vol. 118(2), pp. 887-906, 2005.

[3] N. Mesgarani, M. Slaney, S. A. Shamma, "Discrimination of speech from non-speech based on multiscale spectro-temporal modulations," IEEE Transactions on audio, speech, and language processing, Vol. 14, pp. 920–930, may 2006.

[4] N. Mesgarani, S. V. David, J. B. Fritz, S. A. Shamma, "Phoneme representation and classification in primary auditory cortex," Journal of the Acoustical Society of America, Vol. 123, pp. 899 –909, 2008.

[5] B. T. Meyer, B. Kollmeier, "Robustness of spectro-temporal features against intrinsic and extrinsic variations in automatic speech recognition," Speech Communication, Vol. 53 (5), pp. 753-767, 2011.

[6] N. Esfandian, F. Razzazi, A. Behrad, S. Valipour, "A Feature selection method in spectro-temporal domain based on Gaussian mixture models," Proceedings of International Conference on Signal Processing, Bejing, China, pp. 522–525, October 2010.

[7] N. Esfandian, F. Razzazi, A. Behrad, "A clustering based feature selection method in spectro-temporal domain for speech recognition," Engineering Applications of Artificial Intelligence, Accepted to publish, 2011.

[8] B. Zitová, J. Flusser, "Image registration methods: a survey," Image Vision Computing, Vol. 21(11), pp. 977-1000, 2003.

[9] K. Kerdprasop, N. Kerdprasop, P. Sattayatham, "Weighted K-means for density-based clustering," Lecture notes in Computer Science, Vol. 3589, pp. 488-497, 2005.

[10] B. Gas, J. L. Zarader, C. Chavy, M. Chetouani, "Discriminant neural predictive coding applied to phoneme recognition," NeuroComputing, Vol. 56, pp. 141–166, 2004.

[11] M. Yousefi Azar, F. Razzazi, "A neural predictive coding feature extraction scheme in DCT domain for phoneme recognition," Neural Computing & Applications, Accepted to publish, DOI: 10.1007/s00521-010-0450-0, 2010.

[12] W. M. Fisher, G. R. Doddington, K. M. Goudie–Marshall, "The DARPA speech recognition research database: Specifications and status," Proceedings of DARPA Workshop on Speech Recognition, pp. 93–99, February 1986.