Design Of Wavelet based Phone Pass Filter (PPF) Using Genetic Algorithm

Amr M. Gody
Assistant professor of signal processing, Cairo University, Fayoum Campus, Faculty of Engineering , Electrical Engineering Department.

Abstract

The paper introduces new direction of handling phone classification problem. The problem is treated by designing a filter like algorithm to separate certain phone or a group of phones from the complete set. The designed filter conveys a good information regarding the phone it extracts. The speech data that is introduced to the filters are prepared by a certain speech model that is based on wavelet transform. This model is introduced into [1] and [2]. A brief illustration of the model is included in the introduction section. The work is applied on the Arabic language phone set. A database of many hours of Arabic dialogs is used. The database is recorded from 30 male speakers. Genetic algorithm is utilized in Phone Pass Filter (PPF) parameters design. The 6 coefficients filters are analyzed to extract all the possible phone properties. The speaker dependency is localized. An average PPF extraction efficiency of 94% is achieved over the trustable dataset.

1. Introduction

In the way of phone recognition , exploring for new feature models is one of the major trends to get new promising domains[3-7]. Many works are done in this direction and many useful results are obtained. In this article a detailed analysis is applied on a proposed feature model introduced in [1] and [2]. The following is a brief description of the work done in [1]. Stable character representation for each phone is achieved. The characters verification by human eye is considered as a measure of success mapping. Database is prepared for a single speaker. Multi-speaker effect was not included. The system used to generate the visual characters is represented in figure 1. A stream of n digitized speech samples is applied to Block number 0 abbreviated as BLK0. A frame of 20 ms is chosen as analysis frame length. No weighting window is applied and no frame overlapping is considered as the study is concerning the central region of the frame (stable region). The output vector F_{1x220} is applied to BLK1 for normalization process. U_{220x1} is a worthy informative vector. It gives an answer of the question “What are the frequency contents at time "t" of this Frame”. The wavelet filter is chosen to be Doubechies filter with 6 analysis levels in depth (6 dyadic frequency bands are chosen to cover the analysis frequency band of the signal under study). The signal is projected on 6 different frequency bands. U_{220x1} is applied to BLK3. This block is considered as a feature generator. The features here are the R.M.S values of the signal in each frequency band. This gives a vector of 6 elements as each element represents the power of the signal in the corresponding frequency band. The vector V_{6x1} is generated from BLK3. Table 1 gives a description of V_{6x1} vector.

Table 1, vectors explanation for the system considering 11025 (Hz) sampling rate and 220 samples/frame ($\approx$20 (ms)) analysis frame length[6]

Frequency range (KHz)	Element index in V_{6x1}
2.7-5.5	1
1.4-2.7	2
0.69-1.4	3
0.344-0.69	4
0.172-0.344	5
0.086-0.172	6

In this work the genetic optimization algorithm is utilized to optimize a certain filters parameters to extract the Arabic phones individually from the composite speech signal. The genetic algorithm is best for finding the global peaks. The following paragraph is best describe the genetic algorithm GA[8].

One of the central challenges of computer science is to get a computer to do what needs to be done, without telling it how to do it. Genetic programming addresses this challenge by providing a method for automatically creating a working computer program from a high-level problem statement of the problem. Genetic programming achieves this goal of *automatic programming* (also sometimes called *program synthesis* or *program induction*) by genetically breeding a population of computer programs using the principles of Darwinian natural selection and biologically inspired operations. The operations include reproduction, crossover (sexual recombination), mutation, and architecture-altering operations patterned after gene duplication and gene deletion in nature.

Genetic programming is a domain-independent method that genetically breeds a population of computer programs to solve a problem. Specifically, genetic programming iteratively transforms a population of computer programs into a new generation of programs by applying analogs of naturally occurring genetic operations. The genetic operations include crossover (sexual recombination), mutation, reproduction, gene duplication, and gene deletion.

Many works are implemented in the area of speech processing that makes use of GA. Tang's article [9] introduces the genetic algorithm (GA) as an emerging optimization algorithm for signal processing. After a discussion of traditional optimization techniques, the article reviews the fundamental operations of a simple GA and discusses procedures to improve its functionality. The properties of the GA that relate to signal processing are summarized, and a number of applications, such as IIR adaptive filtering, time delay estimation, active noise control, and speech processing, that are being successfully implemented are described in this article.

Kwong. S.,1996 [10] introduces the genetic algorithm (GA), which is used to solve the nonlinear time alignment problem of pattern comparisons in ASR. Experimental results show that the GA has a better performance than the Dynamic Time Warping DTW. In addition, two derivatives of GA: the hybrid GA and the parallel GA are also presented.

Juola. P, 1996 [11] describes a method of encoding binary data into a “radio alphabet” using a feature-based distance metric to measure phonetic confusability, then using this metric in a genetic algorithm to select appropriate words from a larger list of candidate words.

Chau, C.W., 1997 [12] uses the GA for HMM training. He concluded that the GA mimics natural evolution and performs global searching within the defined searching space. Experiments showed that using the GA for HMM training (GA-HMM training) results in a better performance than using other heuristic algorithms to optimize the model parameters in order to best describe the trained observation sequences.

Ching-Tang, 1997 [13] stated that most of the speech segmentation works are based on the thresholds of parameters to segment the speech data into phonemic units or syllabic units. In the paper, the threshold decision as a clustering problem is formulated. Feature parameters extracted from the analysis frame are clustered into three types: silence, consonants, and vowels. Distributed fuzzy rules which have been used in clustering the numerical data are used for this task. The distributed fuzzy rules, which do not need many training data, have good performance in clustering problems and are beneficial for clustering the features of speech data. Such a method, however, has many fuzzy if-then rules. Genetic-algorithm-based method is utilized for selecting a small number of significant fuzzy if-then rules to construct a compact fuzzy classification system with high classification power. Effectiveness of this approach has been substantiated by classification experiments for continuous radio news speech samples uttered by two females and two males.

In section 2, the data preparation phase is illustrated. In section 3, the idea after Phone Pass Filter PPF is introduced. If the filtering efficiency is high enough then the filter parameters will contains many speech information regarding the filtered phone. The information analysis of the filter parameters is introduced in section 4. The effect of multi speakers are investigated in section 5. Then the phone raw features will be illustrated in section 6. The application of the PPF to cluster the phones into groups that share in some features are introduced in section 7. Then finally the conclusions from this research paper is highlighted in section 8.

2. Data preparation phase

The target of this research is to measure the proposed speech model[1] . A brief description of the model is given in the introduction section. The database consists of many Arabic sentences read in a normal and in a formal way. The complete set of sentences are read by 30 male speakers. The recordings are noise free. The

context is gathered from many subjects to get as much as possible of phonetic patterns. All the database is transcripted. The Speech Filling System software SFS is used to automatically segment the database. A random verification test is applied on the segmented database. The labeling is TIMIT standard. Not all the segmented sentences are hand verified. All recordings are 32kHz , 16 bit, mono.

The Arabic phones that is fully clustered as 36 classes (Table 2). The speech signal segments of each class are gathered and concatenated from the complete Database using a C++ module as shown in figure 2. In brief, the C++ module organizes the complete speech database into 36 speech signals. Each signal represents a phone class speech data.

Referring to figure 2, the outputs of module 1 are 36 wave files (Table 2). The 36 classes waveform files are introduced to the feature extractor module 2. This module is built in the MatLab environment to extract the features of the 36 waveform files and constructs a 36 feature files. Each features file contains a columns that each represents a features vector of 20 (ms) analysis period.

3. PPF coefficients Design

We have a target of 36 filters each is a PPF of a single class. As illustrated in section 2, speech features representing time frames are gathered in a single file for each class. This implies that the class file contains not only class properties but also multi speakers effect as the database is taken from 30 different male speakers. Genetic algorithm is utilized to estimate the best filter coefficients for each class. As indicated in the introduction section, this algorithm is highly recommended in finding the absolute peak points.

The genetic algorithm tries to optimize a certain criteria. Genetic Algorithm (GA) is applied to maximize the trained phone recognition rate. Figure 3 illustrates a snapshot of one complete optimization cycle. The features representing one class are applied to the recognizer. The scoring will be calculated. Success rate will be introduced to the genetic optimizer to be maximized by adjusting filter coefficients. As shown in the figure a single class features database of n vectors are applied to a filtering process. This process is simply a one by one multiplier of the vector elements (6 elements). Each coefficient in the filter is a chromosome. The target of the optimizer is to find the best chromosomes that maximize the success rate. In mathematical words, $P(Q|X)$ is to be maximized as Q is the observation and X is the class that the observations are belonging to.

The optimization dynamics is to minimize the interclass distance and maximize the other classes distance for a single class under training. Actually this is not a sufficient condition to achieve a good general phone recognizer which is not our focus in this research. As no constraints are applied to minimize the correlation of different class chromosomes, the PPF(s) could not be used directly in a universal phone recognition system. To justify the filter coefficient design, the other classes are not introduced during the optimization process. The interphones interference will contribute very much in the design process and may cause the optimizer to try to do more effort to eliminate the new disturbance source. The direction of this research is to analyze the new phone model and get indication how far it is promising in ASR. The only disturbance included is the effect of multi speakers. The problem of ASR phone recognizer may be handled after this phase by including the other sets. It may be a multi phase recognizer. Even a grammar may be included to optimize the practical recognizer as the in the market recognizers. Figure 4 gives an illustration of the ideal phone recognition system. The design criteria of PPF(s) needs to be modified to include the interphones interference in order to make it possible achieving such ideal classification patterns.

Figure 5 illustrates the process layout. Each filter is designed to just extract the associated phone. It does not trained to do anything with other phones. A filtering example case is introduced in figure 5. The unknown class (of type class 2) is wrong recognized due to PPF of class 1. The PPF of class 1 makes the unknown class of type class 2 to be diverged from class 1 and to be converged to class 3. The recognition process may classify the unknown class as type 3 not type 2.

4. PPF(s) in work

In this phase the PPF(s) introduced in section 4 will be analyzed. We have 36 class filters and 36 success rates. Some success rates was very poor and some are moderate and some are perfect. Table 2 indicates the different success rates obtained in this experiment.

Table 2: Arabic Phone set sorted by the obtained SR. "C" abbreviates a consonant and "V" abbreviates Vowel

#	Class Symbol	Phone type	SR(%)	Arabic Symbol
1	sp	-----	100	Speech Pause
2	~h	C	100	ح
3	S_h	C	100	ش
4	S_c	C	100	ص
5	~@	C	100	ع
6	D	C	99.991302	د
7	Y	C	99.971199	ى
8	U2	V	99.365799	و
9	W	C	98.921501	و
10	I2	V	97.195602	ى
11	T_h	C	96.882797	ظ
12	sil	C	96.243896	Silence
13	S	C	95.9505	س
14	a2_C	V	93.012703	ا
15	I2^	V	90.008102	Stressed i2
16	H	C	82.462601	هـ
17	Z	C	71.500504	ز
18	F	C	70.800697	ف
19	R_c	C	55.3643	ر
20	D_j	C	44.693001	Stressed ج
21	M	C	42.1119	م
22	x	C	30.408501	خ
23	Q	C	21.0334	ق
24	@	C	13.9481	ء
25	L	C	11.3627	ل
26	B	C	9.8794	ب
27	J	C	9.63177	ج
28	K	C	9.00461	ك
29	G	C	8.49563	non formal character
30	d_C	C	5.62249	ض
31	D$	C	3.86721	non formal character
32	N	C	3.17136	ن
33	T_c	C	2.44499	ط
34	G_h	C	2.17527	غ
35	T	C	1.77982	ت
36	~z	C	1.73214	ذ

As indicated in table 2, there are 18 classes with SR >70%. The rest of classes give low SR. By reviewing the phonetic properties of the low rate classes we noticed that they all have a plosive or short duration behaviors. This kind of phones may be damaged in the automatic segmentation process as it is not manually verified using the spectrogram and hearing aid.

As the target of this research is to test the properties of the proposed features, so that the low SR classes are excluded. The effect of excluding the low SR classes may be included in the overall phone recognition system as Phone Uncertainty given in equation 1. We dived the 18 which represents the number of reliable phone classes by the 36 which represents the total number of the database classes. This ratio may gives an indication of the recognizer trained using the segmented database.

$$PU\,\% = \frac{18}{36} \times 100 = 50\% \qquad \textbf{(1)}$$

5. Multi Speakers Effect

Now we have 18 PPF(s) with high SR. The PPF(s) need to be analyzed to get the common and discriminative components in the filter vectors. Two points may be considered.

Filters are trained to optimize single class extraction. So any speaker related feature may be given a low weight.

The database that is used during the training process is taken from 30 different speakers. So the multi speakers effect is considered during the training process.

The above two points make it clear that we have to seek about the common low weight features in the different class filters. The common low weight feature components across the different filters give indication about which features are best to be used in a speaker identification using the proposed features model.

As the optimization process used here is not a weighting process (0.0 to 1.0), so we could not detect what is the common low weight features by direct method. The genetic optimizer is given the flexibility in the weighting process to get the best recognition rate that may be obtained from the available database. The optimizer gives weights from (-10 to 10) for each feature component. This may lead us to another direction to find if any common actions are made on the similar feature components across the different classes.

$$X = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{61} & \cdots & a_{6n} \end{pmatrix} \quad \textbf{(2)}$$

In (2), the X matrix contains the different class filters of the 18 classes under test. Each vector represents a filter of a certain class. Each vector is a 6 element components that represent the 6 features model weights as indicated in section 3.

Autocorrelation is applied to each row to get how far the row elements are correlated.

$$R_x(\tau) = \frac{1}{\|X\|^2} \sum_{n=0}^{N-\tau-1} X(n+\tau) X(n) \quad \textbf{(3)}$$

$$Y = \begin{pmatrix} R_{x1}(\tau) \\ \vdots \\ R_{x6}(\tau) \end{pmatrix} = \begin{pmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{61} & \cdots & a_{6m} \end{pmatrix}_{m=35} \quad \textbf{(4)}$$

In (4), the matrix Y will contain 6 rows. Each row is the Auto Correlation Function of the corresponding row elements. The function will produce 17 lag components on both values of τ (negative and positive) plus one zero-lag value. We may exclude the negative lags as it is symmetric function.

$$C = Max\{Y\} = \begin{pmatrix} c_{11} \\ \vdots \\ c_{61} \end{pmatrix} \quad \textbf{(5)}$$

The C vector in (5) represents the maximum correlation that may be obtained from the matrix X. A typical result obtained is given in table 3.

Table 3: Max correlation values that may be obtained from the filters matrix.

Element	Value	Frequency range (KHz)
C11	0.309421	2.7-5.5
C21	0.404181	1.4-2.7
C31	0.250128	0.69-1.4
C41	0.231227	0.344-0.69
C51	0.169341	0.172-0.344
C61	0.578382	0.086-0.172

As shown in table 3, C_{21} and C_{61} indicate the maximum correlating components. So that we may conclude that the most speaker effects will be in the corresponding bands information.

6. Phone related features

Starting from the array X in (2), we need to get the cross correlation of the different row vectors at $\tau = 0$.

$$R_{xy}(\tau) = \frac{1}{\|X\|\|Y\|} \sum_{n=0}^{N-\tau-1} X(n+\tau) Y(n) \quad \textbf{(6)}$$

Using (6), we may put $\tau = 0$ to get the zero lag cross correlation between two rows. This correlation function reflects how far the information hold by a feature vector is correlated. The zero lag is chosen to get the maximum correlation possible between the two vectors as each feature component is multiplied by another feature component in the same filter vector. The sum in the cross correlation gives an overall metric value over all available filters. Figure 6, illustrates this choice. Each row representing one feature component along the 18 available filters. So each row will be correlated with the other 5 rows and itself. Each row is a pivot when it is correlated with itself. The other correlations will be a relative to that pivot. For example, correlating row X and Row Y with variable $\tau = 0$, cause the correlation function in (6) use elements of the same mask vector in the multiplication as it does not make sense to multiply elements from different masks (filters). The summation of the multiplications in (6) gives an overall estimation of the correlation between the two features that are represented by rows X and Y. The division by $\|X\|\|Y\|$ makes a normalization of the summation value to make it possible to use it in further comparisons. Figure 7 indicates the obtained results. The vertical arrow points to a separation point. As equation (6) is applied 6 times for each row to measure the relative feature correlation relatively to the pivot row(X row in equation 6 where $X \in \{1,2,3,4,5,6\}$). So, we have 6 groups of correlations for each has a 6 relative correlations. All the 6 group results are gathered into a single graph as shown in figure 7. A separation between group results are indicated by a vertical arrow at the beginning of each group and the first element in each group is highlighted by a dark color. For example, component 1 is highly correlated with component 2 as shown in the first group results in figure 7. Component 2 is highly correlated with both components 1 and 3 as indicated in the second group results. So we may drop component 2 as it is highly correlated with both 1 and 3. Also it is better to drop component 2 as it is speaker dependent component as indicated in section 5.

7. Phones Grouping

In this section the filters are treated in such way to get how far they are correlated. The cross correlation is made between the columns instead of rows in figure 6. Each column is treated as a pivot when it is correlated to itself then it is correlated with the other 17 filter vectors.

$$R_{xy}(\tau) = \frac{1}{\|X\|\|Y\|} \sum_{n=0}^{6-\tau-1} X(n+\tau) Y(n) \quad \textbf{(7)}$$

Equation (7) is the cross correlation relation that will be applied on the columns (see figure 6). Equation (8) represents the correlation matrix of filter number 1 with all other filters. Each element in R_1 is a vector of 11 elements that represent the cross correlation values with respect to the lag variable τ.

$$R_1 = \begin{pmatrix} R_{11}(\tau) & R_{12}(\tau) & \cdots & R_{1n}(\tau) \end{pmatrix}_{11x\,18} \quad \textbf{(8)}$$

The next step is to find out how far the vector filter under test is correlated with the other 17 filters. The maximum correlation is chosen in this research. $\overline{R}_1$ is constructed as its elements are the maximum correlation value in each column. The first element will be the autocorrelation of the filter with itself. This value will be evaluated to 1. Equation (9) represents the above words.

$$\overline{R}_1 = Max\,(R_1) = \left(\hat{R}_{11}(\tau) \quad \hat{R}_{12}(\tau) \quad \hat{R}_{13}(\tau) \quad \cdots \quad \hat{R}_{1n}(\tau) \right)_{1\times 18} \tag{9}$$

The whole R matrix is constructed by repeating the above steps to all other filter vectors. Equation (10) illustrates the compound R matrix.

$$R = \begin{pmatrix} \overline{R}_1 \\ \overline{R}_2 \\ \vdots \\ \overline{R}_{18} \end{pmatrix}_{18\times 18} = \begin{pmatrix} 1 & \overline{R}_{12}(\tau) & \cdots & \overline{R}_{1n}(\tau) \\ \overline{R}_{21}(\tau) & 1 & \cdots & \overline{R}_{2n}(\tau) \\ \vdots & \vdots & \cdots & \vdots \\ \overline{R}_{n1}(\tau) & \overline{R}_{n2}(\tau) & \cdots & 1 \end{pmatrix}_{18\times 18} \tag{10}$$

Each row in (10) gives a measure of how far the other filter vectors are correlated with the corresponding pivot filter vector. So we can use this matrix to construct phone groups. The ideal case is to get 18 groups as this is the situation of optimal discrimination. The threshold of 0.7 is chosen to construct the groups. If any filter is more than or equal 0.7 correlation with the pivot filter so it is belonging to the same group as the pivot. The results are tabulated in table 4.

Table 4: Phone Groups

Group Number	
1	Z S_c
2	Y
3	W U2 D
4	T_h S
5	sil F
6	S_h
7	~@
8	~h
9	a2_C
10	H
11	I2
12	I2^
13	sp

Table 4 indicates that the 18 phones are belonging to 13 groups. This indicates that the features must not used as it is directly in the recognition system. Instead it may be better to make a hierarchal recognition by classifying the phone into its group. Then internally we may find out what features may be discriminative. We may construct an uncertainty measure by dividing the 13 actual group count by the 18 optimal group count that leads to 72%.

$$GU\,\% = \frac{13}{18} \times 100 = 72\% \tag{11}$$

The estimated overall recognition system performance may be calculated by multiplying the average phone success rate by the grouping uncertainty level by phone uncertainty level.

$$\eta = \overline{S}R \times GU\,\% \times PU\,\% = 0.94 \times 0.72 \times 0.5 \times 100 \approx 34\% \tag{12}$$

Equation (12) gives a prior estimation of the recognition rate that may be obtainable from such system that will implement the PPF(s) directly in a whole recognition system. This efficiency may be elevated if the constraint of multi speaker effect is included in the real recognizer.

8. Conclusion

A 94% of success rate is achieved for phone identification. This rate actually does not imply a phone recognition system with 94% success rate. The high rate in a single phone extraction gives the PPF a great importance as it encapsulates the phone properties. Analyzing the different PPF(s) gives very interesting results that may be very useful in future phone recognition system.

9. References

[1] Amr M. Gody, "Graphical Phone Representation", The Fourth conference on language Engineering CLE'2003, PP: 132-142,Oct. 2003

[2] Amr M. Gody, "Natural Hearing Model Based On Dyadic Wavelet", The third conference on language Engineering CLE'2002, PP: 37-43,Oct. 2002.

[3] Ljolje, A.; Riley, M.D.,"Automatic segmentation and labeling of speech", International Conference on Acoustics, Speech, and Signal Processing, ICASSP-91., volume:1, PP: 473 -476, Apr. 1991.

[4] Lamel, L.F.; Gauvain, J.-L.,"Experiments on speaker-independent phone recognition using BREF", IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP-92., , volume:1, PP: 557 -560, Mar. 1992.

[5] Robinson, T.; Hochberg, M.; Renals, S.,"IPA: improved phone modeling with recurrent neural networks", IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP-94., volume:1, PP: I/37 -I/40, Apr. 1994.

[6] Ruxin Chen; Jamieson L., "Experiments on the implementation of recurrent neural networks for speech phone recognition", Thirtieth Asilomar Conference on Signals, Systems and Computers, volume:1, PP: 779 -782, Nov. 1996.

[7] R. Chen, L. H. Jamieson, "Explicit modeling of coarticulation in a statistical speech recognizer", Proc. Int. Conf. Acoustic, Speech, Signal Processing, Atlanta, GA, PP: 463-466, May 1996.

[8] http://www.genetic-programming.com/gpanimatedtutorial.html

[9] Tang, K.S., "Genetic algorithms and their applications" , Signal Processing Magazine, IEEE, Volume: 13 , Issue: 6, PP: 22 - 37 , Nov. 1996

[10] Kwong, S., "Genetic algorithm for optimizing the nonlinear time alignment of automatic speech recognition systems" , IEEE Transactions on Industrial Electronics, Volume: 43 , Issue: 5, PP: 559 – 566, Oct. 1996

[11] Juola, P, "Whole-word phonetic distances and the PGPfone alphabet", Fourth International Conference on Spoken Language ICSLP 96, Volume:1, PP: 98 - 101 , Oct. 1996.

[12] Chau, C.W., "Optimization of HMM by a genetic algorithm", IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP-97, Volume:3, PP: 1727 - 1730, Apr. 1997.

[13] Ching-Tang Hsieh, "Distributed fuzzy rules for preprocessing of speech segmentation with genetic algorithm", IEEE International Conference on Fuzzy Systems, Volume:1, PP: 427 - 431, July 1997.

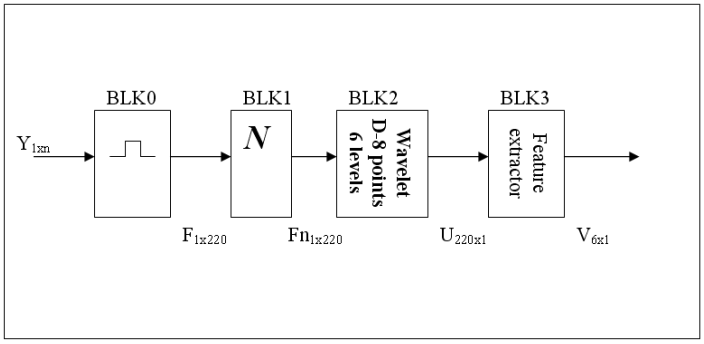

Figure 1 Speech model used in preparing data for PPF design.

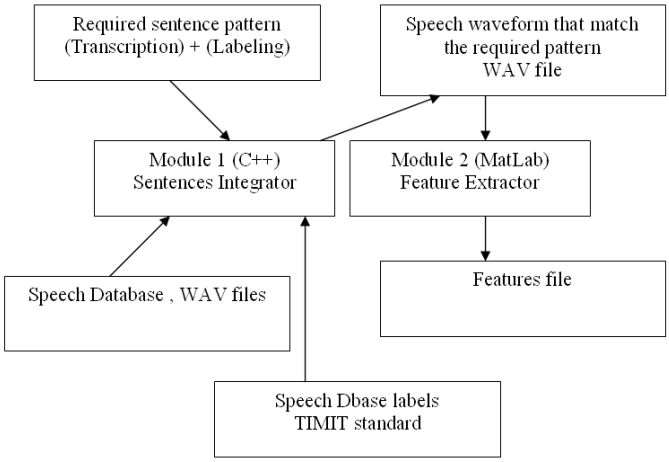

Figure 2 Module integration process

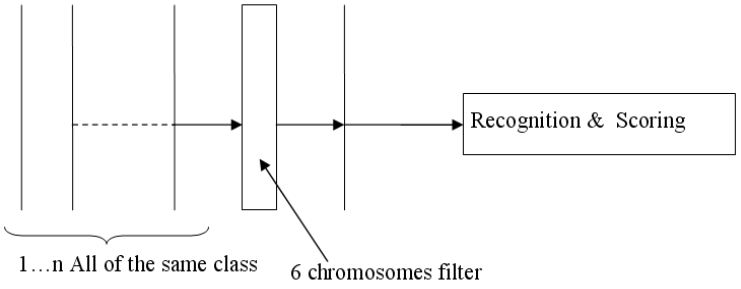

Figure 3 Snapshot of a one complete optimization cycle.

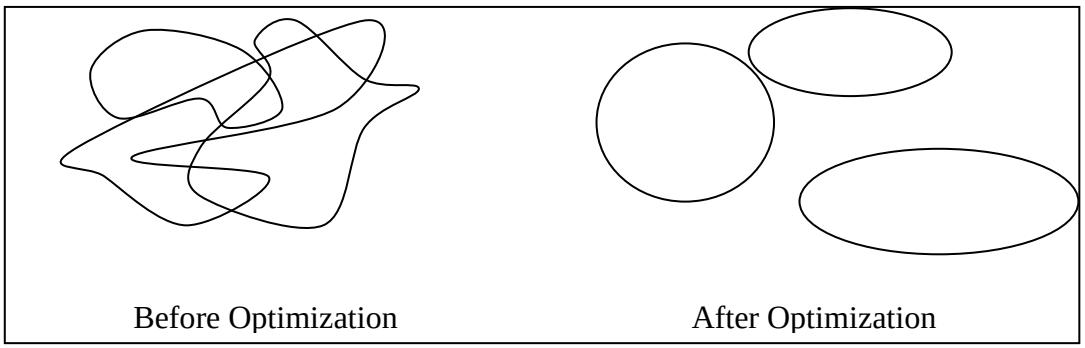

Figure 4 The optimal training results. The classes will be approaching to the mean of the corresponding class is the property of high recognition rate of the class.

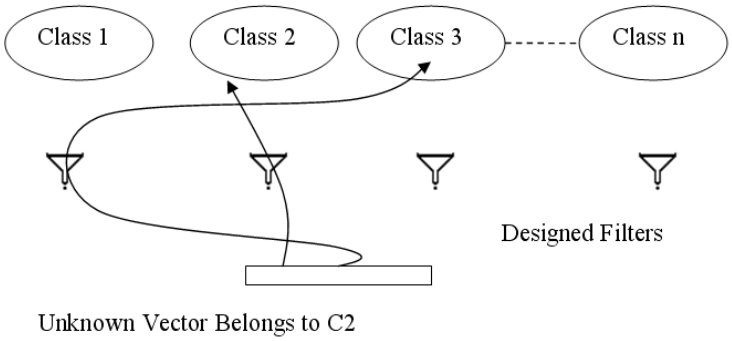

Figure 5 The recognition process of the unknown vector may be failed as the designed filters does not include any constraints regarding the other classes during the training.

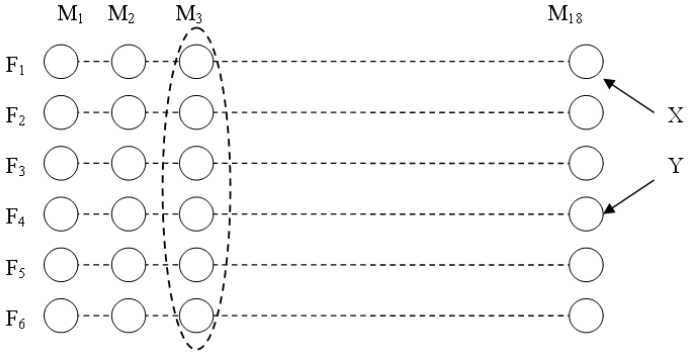

Figure 6 The layout of the Mask array.

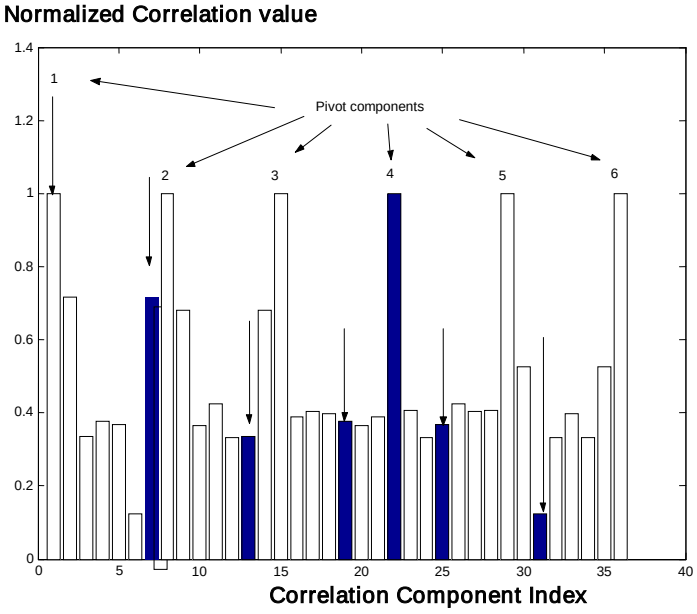

Figure 7 The 6 groups results of Cross correlation between features elements.