Development an Automatic Speech to Facial Animation Conversion for Improve Deaf Lives

S. Hamidreza Kasaei, S. Mohammadreza Kasaei, S. Alireza Kasaei
Young Researchers Club, Isfahan Branch (Khurasgan), Islamic Azad University,Isfahan, Iran
hamidreza_kasaee@yahoo.com

Abstract

In this paper, we propose design and initial implementation of a robust system which can automatically translates voice into text and text to sign language animations. Sign Language Translation Systems could significantly improve deaf lives especially in communications, exchange of information and employment of machine for translation conversations from one language to another has. Therefore, considering these points, it seems necessary to study the speech recognition. Usually, the voice recognition algorithms address three major challenges. The first is extracting feature form speech and the second is when limited sound gallery are available for recognition, and the final challenge is to improve speaker dependent to speaker independent voice recognition. Extracting feature form speech is an important stage in our method. Different procedures are available for extracting feature form speech. One of the commonest of which used in speech recognition systems is Mel-Frequency Cepstral Coefficients (MFCCs). The algorithm starts with preprocessing and signal conditioning. Next extracting feature form speech using Cepstral coefficients will be done. Then the result of this process sends to segmentation part. Finally recognition part recognizes the words and then converting word recognized to facial animation. The project is still in progress and some new interesting methods are described in the current report.

Keywords: Deaf Human, Sign Language Translation Systems, Humatronics, Automatic Speech Recognition

1. Introduction

Today one in 1000 people become deaf before they have acquired speech and may always have a low reading age for written Persian. Sign is their natural language. Persian Sign Language has its own grammar and linguistic structure that is not based on Persian. So voice recognition systems play a very significant role in field of human electronics and its wide applications in deaf live.

This research study was started with several speeches to text experiments to measure the communication skills of deaf people, and to understand their everyday problems better. The primary aim of our project was to develop a communication aid for deaf persons which can be implemented in a mobile telephone. In our system a partially animated face is displayed in interaction with deaf users. They are very useful in much application.

Our system starts with preprocessing and signal conditioning. Next extracting feature form voice using Cepstral Coefficients will be done. Feature extraction is the process that extracts a small amount of data from the voice signal that can later be used to represent each word. Then the result of this process sends to Feature matching, Feature matching involves the actual procedure to identify the unknown word by comparing extracted features from voice input with the ones from a set of known words. Finally recognition part recognizes the words and then converting word recognized to facial animation.

Some of the related research in the field of Automatic Translate Voice to Sign Language Animation is as in Fig. 1.

Attila Andics, James M. McQueen, [1] proposed Neural mechanisms for voice recognition. M. Benzeghiba, R. De Mori, [2] proposed an Automatic speech recognition and speech variability.

Ramin Halavati, Saeed Bagheri Shouraki, [3] proposed a Recognition of human speech phonemes using a novel fuzzy approach.

This paper is organized as follow. Section 2 describe an overview of our system, finally, in section 3 concludes this paper. The project is still in progress and some new interesting methods are described in the current report.

2. The proposed method

All technologies of voice recognition, speaker identification and verification, each has its own advantages and disadvantages and may requires different treatments and techniques. The choice of which technology to use is application-specific. At the highest level, all voice recognition systems contain two main modules: feature extraction and feature matching. Feature extraction is the process that extracts a small amount of data from the voice signal that can later be used to represent each word. Feature matching involves the actual procedure to identify the unknown word by comparing extracted features from his/her voice input with the ones from a set of known words.

A wide range of possibilities exist for parametrically representing the speech signal for the voice recognition task, such as Linear Prediction Coding (LPC), RASTA-PLP and Mel-Frequency Cepstrum Coefficients (MFCC).

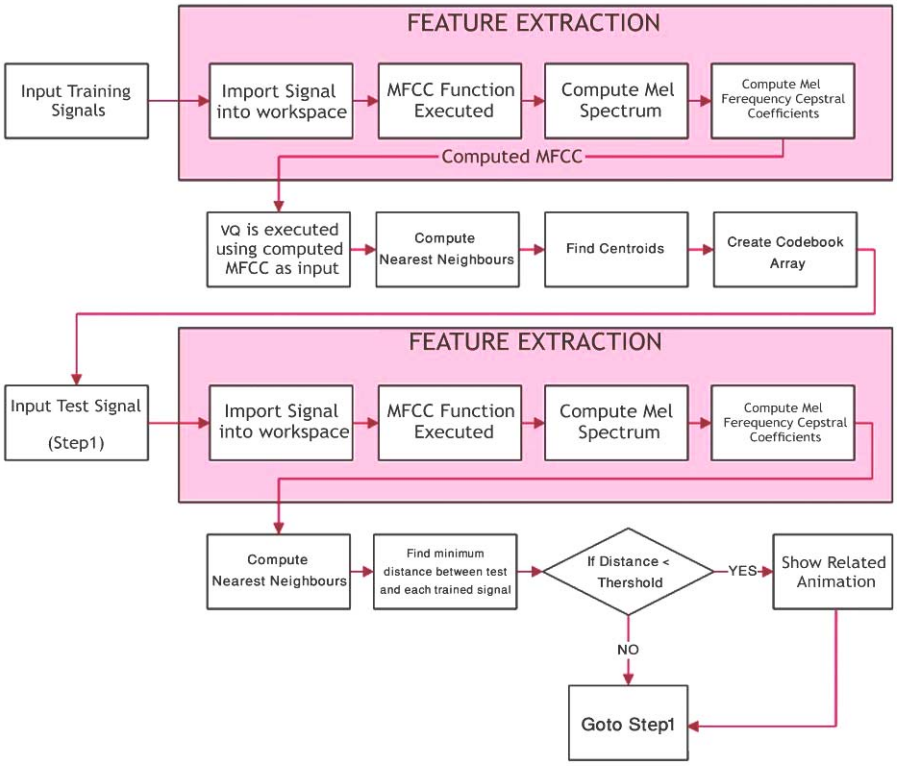

Figure 1. The Structure of the Automatic Translate Voice to Sign Language Animation System

LPC analyzes the speech signal by estimating the formants, removing their effects from the speech signal, and estimating the intensity and frequency of the remaining buzz. The process of removing the formants is called inverse filtering, and the remaining signal is called the residue [4].
Another popular speech feature representation is known as RASTA-PLP, an acronym for Relative Spectral Transform - Perceptual Linear Prediction. PLP was originally proposed by Hynek Hermansky as a way of warping spectra to minimize the differences between speakers while preserving the important speech information. RASTA is a separate technique that applies a band-pass filter to the energy in each frequency subband in order to smooth over short-term noise variations and to remove any constant offset resulting from static spectral coloration in the speech channel e.g. from a telephone line [5].

MFCC's are based on the known variation of the human ear's critical bandwidths with frequency; filters spaced linearly at low frequencies and logarithmically at high frequencies have been used to capture the phonetically important characteristics of speech. This is expressed in the Mel-frequency scale, which is linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz [3]. MFCC is the best known and most popular so we decided to use MFCC in our project. The process of computing MFCCs is described in more detail in the First section of Proposed Method.

In the second, Feature matching Algorithm has been discussed. The goal of pattern recognition is to classify objects of interest into one of a number of categories or classes. The objects of interest are generically called patterns and in our case are sequences of acoustic vectors that are extracted from an input speech using the techniques described in the previous paragraph (Feature Extraction). The classes here refer to individual words. Since the classification procedure in our case is applied on extracted features, it can be also referred to as feature matching. The feature matching techniques used in voice recognition include Dynamic Time Warping (DTW) [6], Hidden Markov Modelling (HMM), Support Vector Machine (SVM) [7], and Vector Quantization (VQ) [8]. There is also another technique which is called Artificial Neural Network (ANN) [9]. In this project, Vector Quantization approach will be used, due to ease of implementation and high accuracy. VQ is a process of mapping vectors from a large vector space to a finite number of regions in that space. Each region is called a cluster and can be represented by its center called a codeword. The Automatic voice Recognition System will compare the voice with the codewords of the trained data. The best matching result will be the desired voice.

2.1. Feature Extraction

Preprocessing mostly is necessary to facilitate further high performance recognition. A wide range of possibilities exist for parametrically representing the speech signal for the voice recognition task that we chose MFCC Algorithm in our project.

Mel Frequency Cepstral Coefficients

Mel Frequency Cepstral Coefficients (MFCC) are coefficients that represent audio, based on perception. It is derived from the Fourier Transform (FFT) or the Discrete Cosine Transform (DCT) of the audio clip. The basic difference between the FFT/DCT and the MFCC is that in the MFCC, the frequency bands are positioned logarithmically (on the Mel scale) which approximates the human auditory system's response more closely than the linearly spaced frequency bands of FFT or DCT. This allows for better processing of data. The main purpose of the MFCC processor is to *mimic the behavior of the human ears. Overall the MFCC process has 5 steps that show in figure 2.*

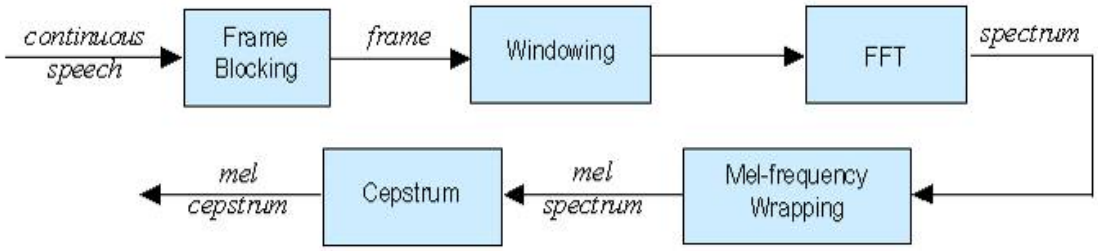

Figure 2. MFCC Block Diagram

Step 1 – Frame Blocking

Step 2 – Windowing

Step 3 – Fast Fourier Transform (FFT)

Step 4 – Mel-frequency Wrapping

Step 5 – Cepstrum Coefficient

At the Frame Blocking Step a continuous speech signal is divided into frames of N samples. Adjacent frames are being separated by M (M<N). The values used are M = 256, and N = 156 the first frame will consist of N samples (i.e. 256 samples).

The next frame will begin M samples (i.e. 156 samples) after the first frame, and it will overlap the first frame by N-M samples (256 – 156 = 100 samples). Then the third frame will start at 2M samples after the first frame and it will overlap first frame by N-2M. The fourth frame will start at 3M samples after the first, and it will overlap it by N-3M. The process will continue until all input signal is accounted for. The result of this step plotted using MATLAB plot command and displayed in Figure 3.

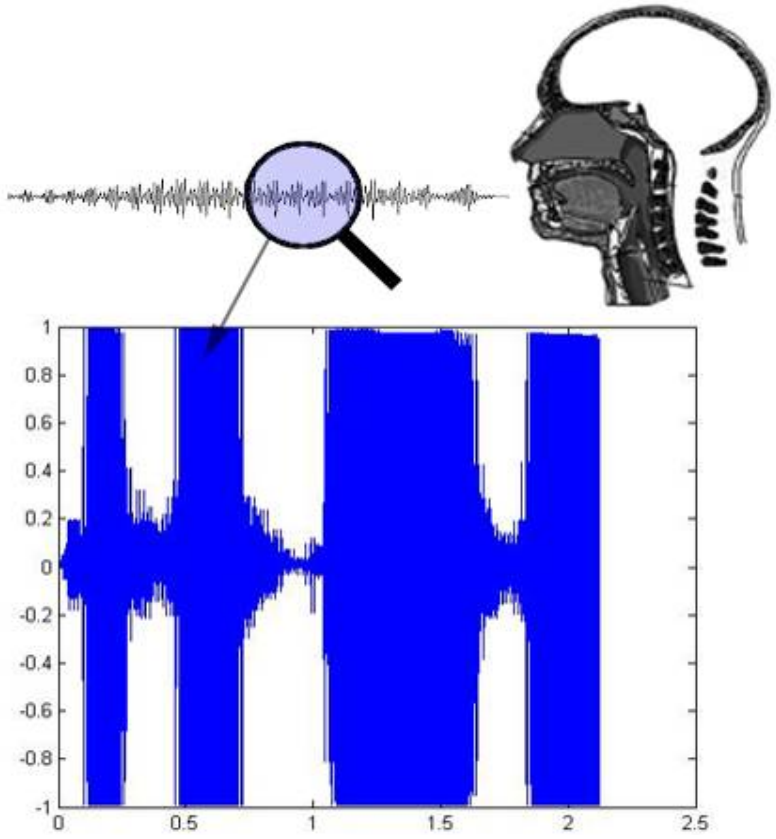

Figure 3 Frame blocking of the speech signal

The next step in the processing is to window each individual frame so as to minimize the signal discontinuities at the beginning and end of each frame. The concept here is to minimize the spectral distortion by using the window to taper the signal to zero at the beginning and end of each frame. If we define the window as $w(n), 0 \le n \le N-1$, where N is the number of samples in each frame, then the result of windowing is the signal

$$y_l(n) = x_l(n)w(n), \quad 0 \le n \le N-1$$

Typically the Hamming window is used, which has the form:

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), \quad 0 \le n \le N-1$$

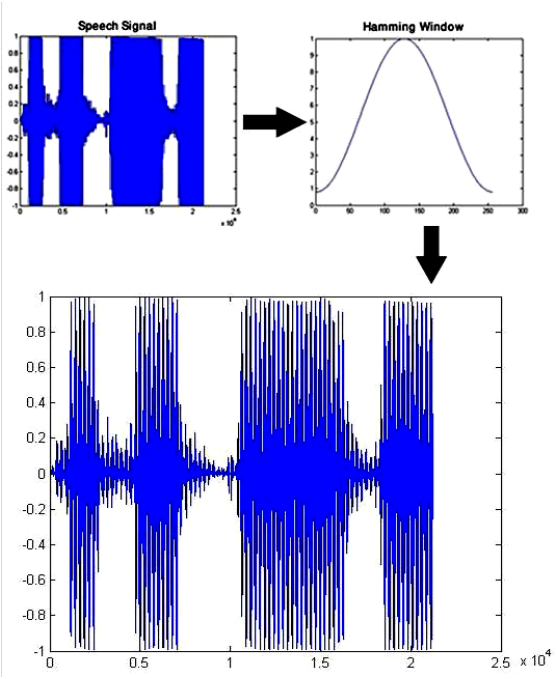

Figure 4. Hamming Window applied to each frame

Use of speech spectrum for modifying work domain on signals from time to frequency is made possible using Fourier coefficients. At such applications the rapid and practical way of estimating the spectrum is use of rapid Fourier changes.

$$X_n = \sum_{k=0}^{N-1} x_k e^{\frac{-2\pi jkn}{N}}$$

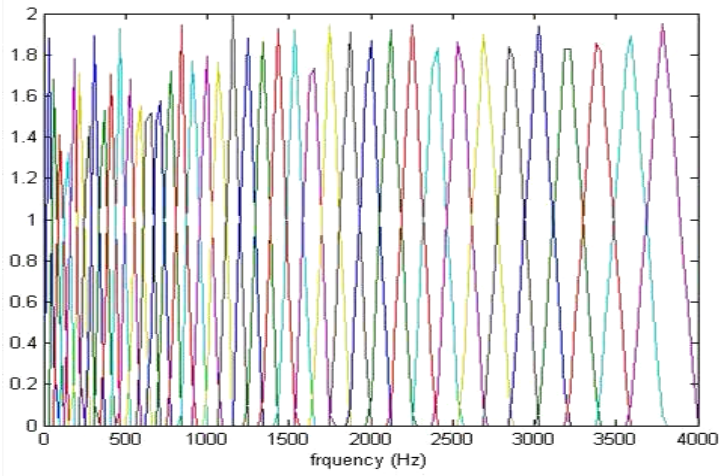

Figure 5. Mel-spaced filter bank

Physiologic changes show that human comprehension of frequency content of sound does not obey a linear space. Therefore for each individual it is computed and measure with a real frequency of sound peak at Mel seal. Using the below equation one can change the frequency at Hertz scale to Mel Scale.

$$mel(f) = 2595 * \log_{10}(1 + \frac{f}{700})$$

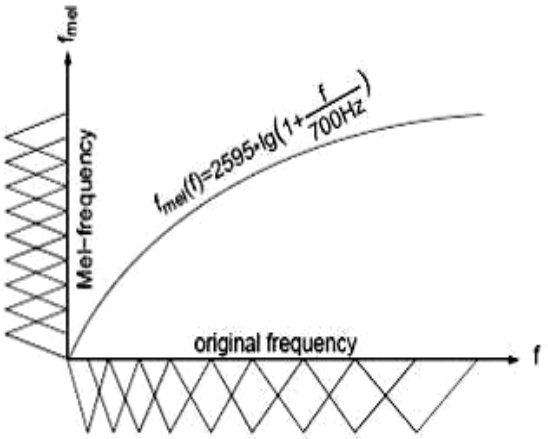

Figure 6. Change the frequency at Hertz scale to Mel Scale.

Following computing the spectrum **power** and applying the above equation on frequency axis, some mediating filters equal to identical overlapping are applied on the scaled spectrum and each filters energy is computed as particularity. This is because conception of a particular frequency by the auditory system is affected by a critical band of frequencies surrounding it. Number of filters is usually between 20 and 30. Logarithmic non linear change operations on obtained particulates for adjusting amplified of particularities and their important though coordinating them with the structure of the auditory system following computing energy of each filter is done as follows: in the following equation F1 is filter in I th, and e (i) is logarithm of energy at ith band.

$$E(i) = \sum_{k=0}^{M/2} (\log | H(k,m) |).F_i(\frac{2k\pi}{M})$$

There are two general stages in commuting Cepstral Coefficients.

- Computation of logarithm of exit amplitude from the signal

This stage is performed through computing spectrum energy of the signed passed through a filter band and via logarithmic conversion.

- Computation of cosine conversion of a filter out put bank.

In computing coefficients usually logarithmic conversion is used for applying on the spectrum. Conventionally this conversion is used following application of filter banks. Through translocation of this logarithmic conversion in the field of temporal array, interesting results have been obtained. Also, other conversions have been studied. The reason for studying other conversions is that in such occasions low energy parts of the signal were considered as noise, that using such energy conversions this effect becomes less significant. M, number of the region, holds the most significant portion of energy contained in the sound signal.

$$Log(x) = \left(\log_m^x\right)^2 * \log^m$$

$$sigm(x) = \frac{1}{1 + 0.0004 * e^{x/m*a+5}} * \log(m)$$

Use of speech spectrum for changing the work domain on the signal from time to frequency is used via Fourier conventions. The rapid and practical way of estimating Spectrum in such applications is employment of rapid Fourier transform. The final stage is extracting particularities is use of discrete cosine to return particularities to the time domain and converse FFT approximation. Major advantage of this method is decrease of number of particularities of number of filter from N_f to N_c in which $N_c \le N_f$. In addition, doing so, includes making independent of the obtained particularity and rendering them non dependant which leads matrix covariance features to become axial. The following equation shows this point.

$$C_i = \sum E(j).\cos(\frac{i\pi}{N_f}(j - \frac{1}{2})) \quad 1 \le i \le N_c$$

Output of this transform is called Cepstrum Coefficients. Final number of features can be equal to the number of filters however the Cepstrum coefficient is more miscorrelated and its low

components indicate more important information compared with higher components which contain less information in speech recognition and contain only, minor information of spectrum and frequency omission of which can be even effective in improving the system precision.

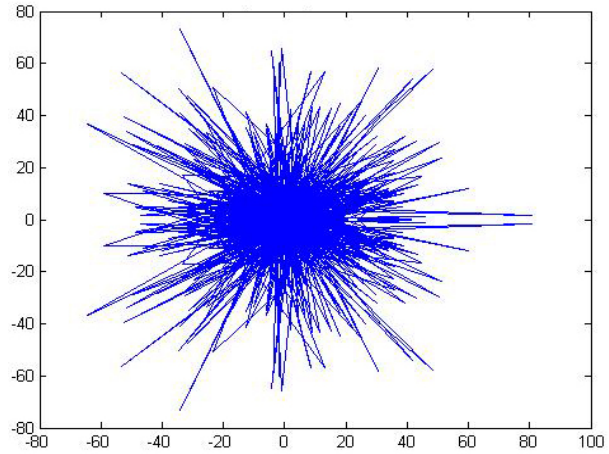

Figure 7. Speech signal after FFT

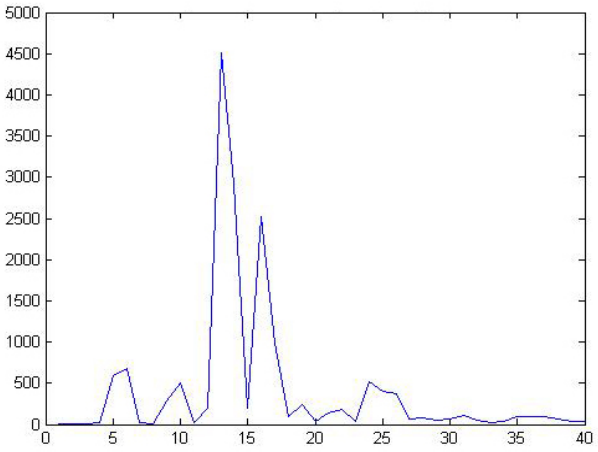

Figure 8. Speech signal after frequency wrapping

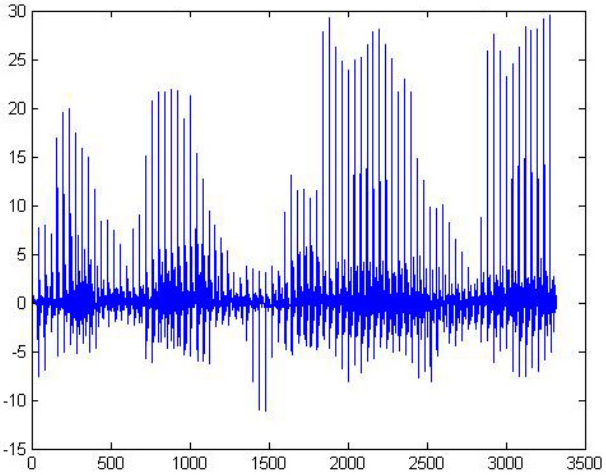

Figure 9. Mel Cepstral Coefficients in time domain

2.2. Feature Matching

In this section, Feature matching Algorithm has been discussed. The goal of feature matching is to classify objects into one of a number of categories or classes. In this project, Vector Quantization approach will be used and the best matching result will be the desired voice.

Vector Quantization Method

Speech recognition systems are inherent of a database, which stores information used to compare the test voice against a set of trained voices. Ideally, storing as much data obtained from feature extraction techniques is advised to ensure a high degree of accuracy, but realistically this cannot be achieved. The number of feature vectors would be so large that storing and accessing this information using current technology would be unfeasible and impractical.[8] VQ is a quantization technique used to compress the information and manipulate the data such in a way to maintain the most prominent characteristics. VQ is a process of mapping vectors from a large vector space to a finite number of regions in that space. Each region is called a cluster and can be represented by its center called a codeword. Figure 10 shows a conceptual diagram to illustrate this recognition process. In the figure, only two word voices with four dimensions of the acoustic space are shown. The result codewords (centroids) are shown by black circles and black triangles for word 1 and 2, respectively. The distance from a feature to the closest codeword is called a VQ-distortion. The speech is authenticated by calculating the feature vector of voice in the input speech using VQ algorithm and measuring the similarity between the calculated feature vector and the feature vector of voice retrieved from the database.

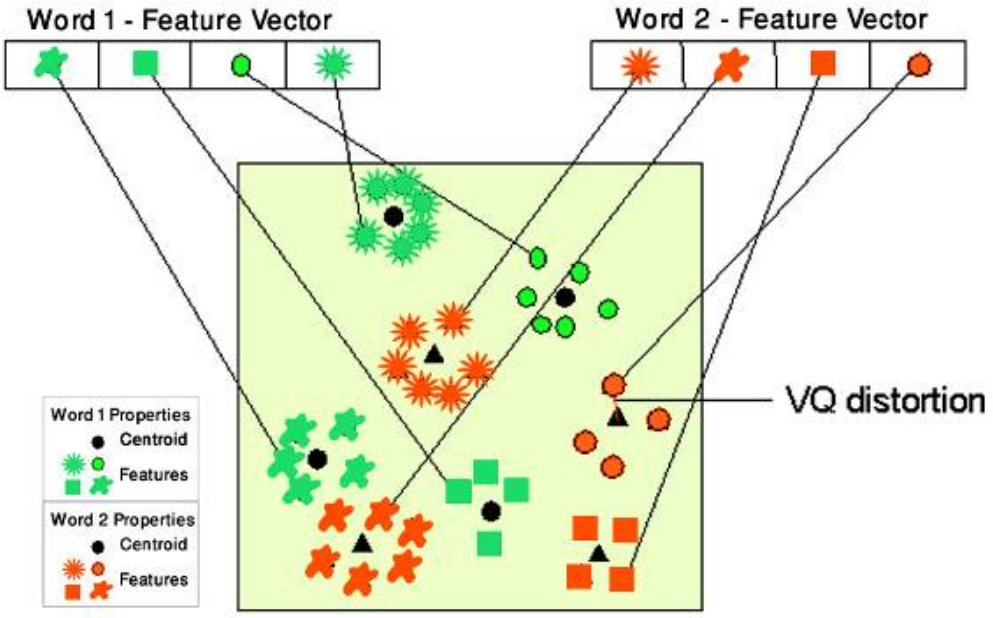

Figure 10. Conceptual diagram illustrating vector quantization codeword formation.

After the conversion process converts the analogue speech signal into a series words, then converting recognized word to facial animation **based on VRML.**

3. Conclusions

To increase the autonomy of deaf and hard of hearing people in their day-to-day professional and social lives, in this paper design and initial implementation of a new approach based on MFCC and Vector Quantization Method is described. This approach includes analyses of speech to animate the talking head. Our future work will include the conception of new test types and performance patterns. We are particularly interested in extending this approach to testing to include implementation of applications under real-time constraints.

Figure 11. A sample of speech to facial animation system

References

[1] Attila Andics, James M. McQueen, "Neural mechanisms for voice recognition" NeuroImage, Volume 52, Issue 4, 1 October 2010, Pages 1528-1540.

[2] M. Benzeghiba, R. De Mori, O. Deroo, etc, "Automatic speech recognition and speech variability: A review" Speech Communication, Volume 49, Issues 10-11, October-November 2007, Pages 763-786.

[3] Ramin Halavati, Saeed Bagheri Shouraki, "Recognition of human speech phonemes using a novel fuzzy approach" Applied Soft Computing, Volume 7, Issue 3, June 2007, Pages 828-839.

[4] http://www.otolith.com/otolith/olt/lpc.html

[5] Hynek Hermansky, Nelson Morgan "RASTA-PLP speech analysis technique", ICASSP'92 Proceedings of the 1992 IEEE international conference on Acoustics, speech and signal processing 1992.

[6] François Petitjean, Alain Ketterlin, etc. "A global averaging method for dynamic time warping, with applications to clustering" Pattern Recognition, Volume 44, Issue 3, March 2011, Pages 678-693.

[7] Jingwei Liu, Zuoying Wang, Xi Xiao "A hybrid SVM/DDBHMM decision fusion modeling for robust continuous digital speech recognition" Pattern Recognition Letters, Volume 28, Issue 8, 1 June 2007, Pages 912-920

[8] Jim Z.C. Lai, Yi-Ching Liaw "A novel encoding algorithm for vector quantization using transformed codebook" Pattern Recognition, Volume 42, Issue 11, November 2009, Pages 3065-3070.

[9] Edmondo Trentin, Marco Gori "A survey of hybrid ANN/HMM models for automatic speech recognition" Neurocomputing, Volume 37, Issues 1-4, April 2001, Pages 91-126.

[11] Balci, K. Xface, "MPEG-4 based open source toolkit for 3D facial animation". In Proc. Advance Visual Interface, 2004.