International Journal of Electronics Signals and Systems

Volume 1 | Issue 3 Article 7

January 2012

An Interactive and Efficient Voice Processing For Home Automation System

Chiraag Goel
Department of Electronics and Instrumentation SRM University, Chennai, chiraag.goel@gmail.com

An Interactive and Efficient Voice Processing For Home Automation System

Chiraag Goel[#], Kaushik Bhuiya[#], Vinayak Nayyar[#], Abhinav Kumar*

[#]Department of Electronics and Instrumentation
SRM University, Chennai
Email: chiraag.goel@gmail.com

Abstract: ***Home networking has evolved from linked personal computers to a more complex system that encompasses advanced security and automation applications. Once just reserved for high-end luxury homes, home networks are now a regular feature in residences. These networks allow users to consolidate heating, air conditioning, lighting, appliances, entertainment, intercom, telecommunication, surveillance and security systems into an easy-to-operate unified network. Interactive applications operated by voice recognition, for example integrated door security systems and the ability to control home appliances, are key features of home automation networks. This interactive capability depends on high-quality voice processing technology, including acoustic echo cancellation, low signal distortion and noise reduction techniques. A home automation system must also be scalable to allow future evolution, flexible to support field upgrades, interactive, easy-to-use, cost-efficient and reliable. This article introduces some of the voice quality performance issues and design challenges unique to home automation systems. It will discuss home automation network applications that rely on voice processing, and examine some of the critical features and functionality that can help ease design complexity and cost to deliver enhanced performance.***

Keywords— Feature extraction, Mel frequency cepstral coefficients (MFCC), Speaker recognition

1. INTRODUCTION

The home automation market is moving beyond high-end luxury homes to target the mainstream consumer. Even in its infancy, researchers estimate the market is worth over $1 billion. In Asia, Europe and North America the home automation market is growing at an average of 10% per year. In Europe alone, demand for home automation systems is expected to double by 2009 to create a $0.5 billion market. While the market grows, home automation systems themselves are evolving to incorporate technologies such as Bluetooth, Wi-Fi, X10, ZigBee and TCP/IP. As the market and technology matures, high-quality voice processing performance becomes increasingly important for home automation and security applications. Voice is an enabling technology that unifies the home network and is used to control appliances, telecommunication, security and entertainment equipment. Endusers are also more comfortable communicating with a human voice, rather than interacting with a machine.The human speech contains numerous discriminative features that can be used to identify speakers. Speech contains significant energy from zero frequency up to around 5 kHz. The objective of automatic speaker recognition is to extract, characterize and recognize the information about speaker identity. The property of speech signal changes markedly as a function of time. To study the spectral properties of speech signal the concept of time varying Fourier representation is used. However, the temporal properties of speech signal such, as energy, zero crossing, correlation etic are assumed constant over a short period. That is its characteristics are short-time stationary.

In this work, the Mel frequency Cepstrum Coefficient (MFCC) feature has been used for designing a text dependent speaker identification system. The extracted speech features (MFCC's) of a speaker are quantized to a number of centroids using vector quantization algorithm. These centroids constitute the codebook of that speaker. MFCC's are calculated in training phase and again in testing phase. Speakers uttered same words once in a training session and once in a testing session later. The Euclidean distance between the MFCC's of each speaker in training phase to the centroids of individual speaker in testing phase is measured and the speaker is identified according to the minimum Euclidean distance. The code is developed in the MATLAB environment and performs the identification. Poor acoustic echo cancellation, ambient noise and signal distortion make it increasingly difficult for a home automation system to perform reliably. If impended by poor voice performance, voice recognition cannot easily detect commands to turn on/off the appliances and voice authentication has difficulty verifying the user to allow access to the residence.

2. MOTIVATION

Most systems use long–term or short–term parameters that should encode vocal tract features, but contributions of the glottis to these features are largely ignored. Even ifMFCC(Mel Frequency CepstrumCoefficients) are theoretically known to deconvolve the source and the vocal tract; in practice, cepstrum coefficients are affected by high pitched voices (women and infants). One can illustrate the role of pitch when dependence of the source and the vocal tract are maintained. Figure 1 exhibits four spectrograms and

pitch histograms; each column corresponds to a different male speaker, obtained from the YOHO database. All speakers pronounced the same digit utterance 'twenty six'. The pitch range is divided into 56 equal bins of 10 Hz width. The spectrograms show a significant similarity of formant distributions between speakers. The spatial distribution of formants depends on the interspeaker variability as described in [9]. However, the pitch histograms are different and vary from one speaker to another for the same context. If one compares the histograms by taking into account their frequency amplitude and width, it is observed that speaker 2 fromthe second column and speaker 4 from the fourth column do have a similar pitch distribution. On the other hand speaker

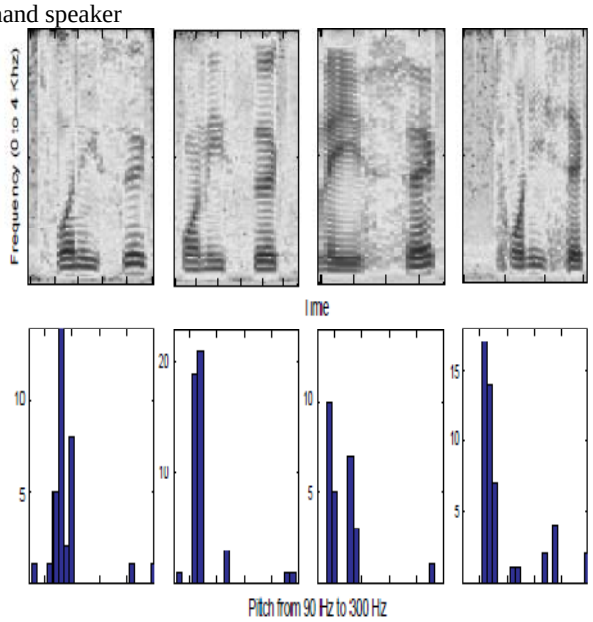


Figure:For each of four male speakers: Spectrograms and pitch histograms for the same English digit utterance '26'.

1 and 3 are characterized by dissimilar pitch histograms. Consequently, if one takes into considerationthe pitch information, the interspeaker variability can be restricted to speakers with similar pitch distributions, and the other speakers will be considered as belonging to other clusters. Speakers with similar pitch will be recognized based on their spectral characteristics. In summary, short–termpitch and vocal tract features can be jointly exploited in order to establish probability models of feature vectors assuming the a priori knowledge of the pitch distribution.

3. SPEAKER RECOGNITION

Anatomical structure of the vocal tract is unique for every person and hence the voice information available in the speech signal can be used to identify the speaker. Anatomical structure of the vocal tract is unique for every person and hence the voice information available in the speech signal can be used to identify the speaker. speaker features of the input speech from test subject will be extracted and matched against the speaker model. A likelihood ratio will evaluate the similarity between the model and the measured observations. The general approach is based on a threshold set for the acoustic likelihood ratio to decide the test speaker is accepted or rejected. Conventional speaker verification systems use hidden Markov models (HMM) or Gaussian mixture model (GMM) to perform the likelihood ratio test [1-6]. These systems make use of a generative model for all speaker models. This will result in over-fitting and maybe cannot maximize the discrimination.

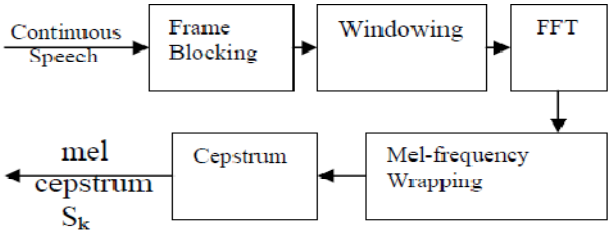
4. SPEECH FEATURE EXTRACTION

The purpose of this module is to convert the speech waveform to some type of parametric representation (at a considerably lower information rate). The speech signal is a slowly time varying signal (it is called *quasi-stationary*). When examined over a sufficiently short period of time (between 5 and 100 ms), its characteristics are fairly stationary. However, over long periods of time (on the order of 0.2s or more) the signal characteristics change to
reflect the different speech sounds being spoken. Therefore, *short-time spectral analysis* is the most common way to characterize the speech signal. A wide range of possibilities exist for parametrically representing the speech signal for the speaker recognition task, such as Linear Prediction Coding (LPC), Mel-Frequency Cepstrum Coefficients (MFCC), and others. MFCC is perhaps the best known and most popular, and this feature has been
used in this paper. MFCC's are based on the known variation of the human ear's critical bandwidths with
frequency. The MFCC technique makes use of two types of filter, namely, linearly spaced filters and logarithmically spaced filters. To capture the phonetically important characteristics of speech, signal is expressed in the Mel frequency scale. This scale has a linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. Normal speech waveform may vary from time to time
depending on the physical condition of speakers' vocal cord. Rather than the speech waveforms themselves, MFFCs are less susceptible to the said variations [1,4].

4.1The MFCC processor

A block diagram of the structure of an MFCC processor is given in Figure 1. The speech input is recorded at a sampling rate of 22050Hz. This sampling frequency is chosen to minimize the effects of *aliasing* in the A/D conversion process.

Figure-shows the block diagram of an MFCC processor .



4.2 Mel-frequency wrapping

The speech signal consists of tones with different frequencies. For each tone with an actual frequency, f, measured in Hz, a subjective pitch is measured on the 'Mel' scale. The *mel-frequency* scale is a linear frequency spacing below 1000Hz and a logarithmic spacing above 1000Hz. As a reference point, the pitch of a 1kHz tone, 40dB above the perceptual hearing threshold, is defined as 1000 mels. Therefore we can use the following formula to compute the mels for a given frequency *f*
in Hz[5]:

mel(f)= 2595*log10(1+f/700)

(1)

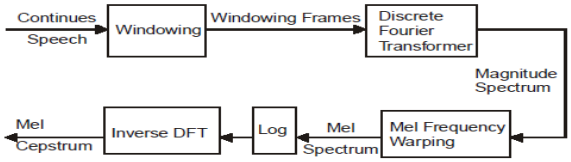
One approach to simulating the subjective spectrum is to use a filter bank, one filter for each desired melfrequency component. The filter bank has a triangular bandpass frequency response, and the spacing as well as the bandwidth is determined by a constant mel-frequency interval.

4.3 CEPSTRUM

In this final step, we convert the log mel spectrum back to time. The result is called the Mel Frequency Cepstrum Coefficients (MFCC).The cepstral representation of the speech spectrum provides a good representation of the local spectral properties of the signal for the given frame analysis. Because the mel spectrum coefficients (and so their logarithm) are real numbers, we can convert them to the time domain using the discrete cosine transform (DCT). In this final step log mel spectrum is converted back to time. The result is called the Mel Frequency Cepstrum Coefficients (MFCC).The discrete cosine transform is done for transforming the mel coefficients back to time domain.

$$C_n = \sum_{k=1}^{k} (\log S_k) \cos\left\{n\left(k-\frac{1}{2}\right)*\frac{\pi}{k}\right\},$$
$$n = 1, 2, \ldots k$$

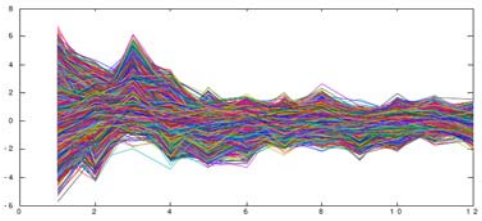
Whereas S_k, $K = 1, 2, \ldots K$ are the outputs of last step. Complete process for the calculation of MFCC is shown in



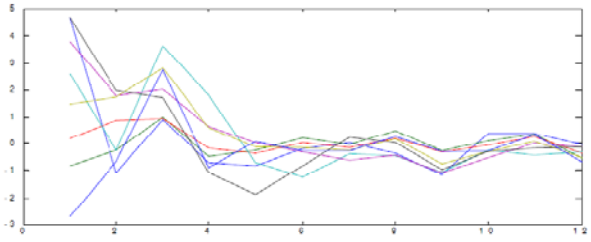
Complete pipeline for MFCC.

5. Vector Quantization:

A speaker recognition system must able to estimate probability distributions of the computed feature vectors. Storing every single vector that generate from the training mode is impossible, since these distributions are defined over a high-dimensional space. It is often easier to start by quantizing each feature vector to one of a relatively small number of template vectors, with a process called vector quantization. VQ is a process of taking a large set of feature vectors and producing a smaller set of measure vectors that represents the centroids of the distribution.



The technique of VQ consists of extracting a small number of representative feature vectors as an efficient means of characterizing the speaker specific features. By means of VQ, storing every single vector that we generate from the training is impossible.

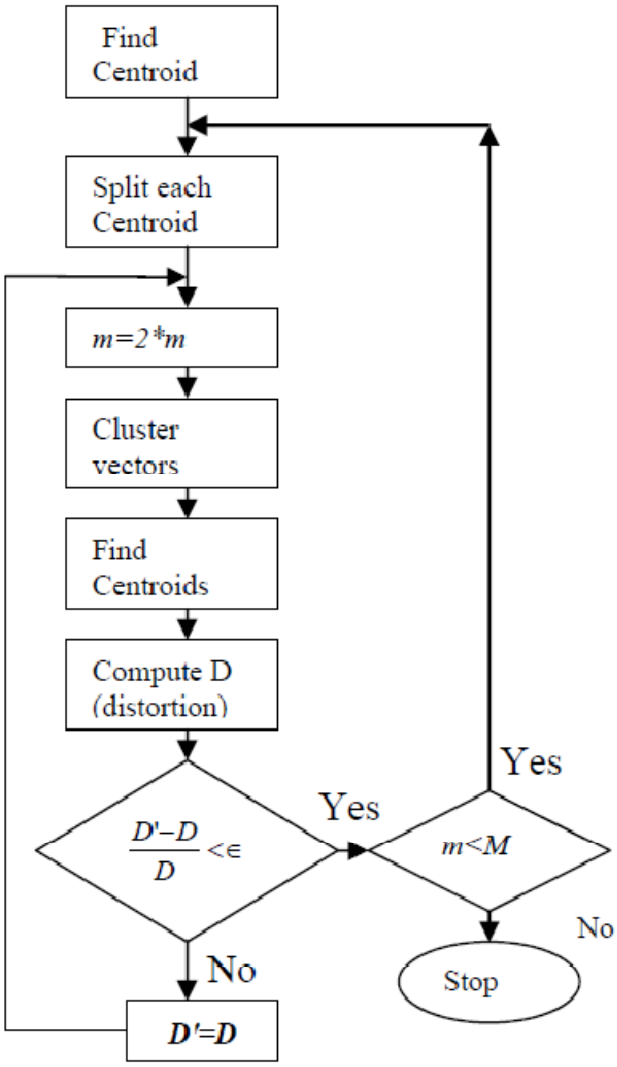


The representative feature vectors resulted after VQ

By using these training data features are clustered to form a codebook for each speaker. In the recognition stage, the data from the tested speaker is compared to the codebook of each speaker and measure the difference. These differences are then use to make the recognition decision.

5.1 LBG design algorithm

The LBG VQ design algorithm is an iterative algorithm which alternatively solves optimality criteria . The algorithm requires an initial codebook. The initial codebook is obtained by the *splitting* method. In this method, an initial codevector is set as the average of the entire training sequence. This codevector is then split into two. The iterative algorithm is run with these two vectors as the initial codebook. The final two codevectors are split into four and the process is repeated until the desired number of codevectors is obtained. The algorithm is summarized in the flowchart of the figure shown below.



Flowchart of VQ-LBG algorithm

6. Distance measure

In the speaker recognition phase, an unknown speaker's voice is represented by a sequence of feature vector {x1, x2 ….xi), and then it is compared with the codebooks from the database. In order to identify the unknown speaker, this can be done by measuring the distortion distance of two vector sets based on The Euclidean distance between two points P = (p1, p2…pn) and Q = (q1, q2...qn)

$$\sqrt{(p_1-q_1)^2+(p_2-q_2)^2+\cdots+(p_n-q_n)^2}=\sqrt{\sum_{i=1}^{n}(p_i-q_i)^2}$$

,

minimizing the Euclidean distance.The Euclidean distance is the "ordinary" distance between the two points that one would measure with a ruler, which can be proven by repeated application of the Pythagorean Theorem. The formula used to calculate the Euclidean distance can be defined as following: The speaker with the lowest distortion distance is chosen to be identified as the unknown person. A vector quantizer maps k-dimensional vectors in the vector space R^k into a finite set of vectors $Y = \{yi: i=1,2….N\}$.Each vector y_i is called a code vector or a code word and the set of all the code words is called a codebook. Codebook of each speaker and measure the difference. These differences are then used to make the recognition decision.

7. EXPERIMENTAL RESULTS

The experimental results of the proposed text-dependent speaker verification system are achieved by using 20 male and 20 female speakers selected from the Aurora 2 database [11]. All of the test speech signals are noisy-free and are sampled at 8000 Hz with 16-bit resolution. Each test speech signal consists of 2~8 English digital numbers or English alphabets. Speaker verification performance will be reported using the false acceptance rate (FAR), the false rejection rate (FRR), and the equal error rate (EER). The definitions of FAR and FRR are given as follows:

$$\text{FAR} = \frac{\#\,\text{aceepted imposter claims}}{\#\,\text{imposter accesses}} \times 100\% \qquad (10)$$

$$\text{FRR} = \frac{\#\,\text{rejected genuine claims}}{\#\,\text{genuine accesses}} \times 100\% \qquad (11)$$

Once the receiver operating characteristic (ROC) curve of FAR vs. FRR is obtained, one can determine the EER, which FAR and the FRR at this point is the same for both of them. In this paper, the different settings of MFCC order are studied experimentally for speaker verification. It follows from [8] that the higher-order MFCC does not further reduce the error rate in comparison with the lower-order MFCC. Hence, this paper compared the results obtained on the SVM based speaker verification system with 13 settings of MFCC order, namely $p = 2q$, q = 1~13. An impostor model was trained on all the MFCCs in the impostor data set while the speaker model was built using the corresponding speaker data set. During speaker verification task, a likelihood ratio was computed between the speaker model and the impostor model. The likelihood ratio was defined as:

$$LR = \log P(x \mid \text{speaker model}) - \log P(x \mid \text{impostor model}) \qquad (12)$$

where x is the input test MFCCs vector. Table 1 shows a summary of the experimental results of the proposed text-dependent speaker verification systems. It follows from Table 1 that the better performance could be obtained when MFCC order p = 22. An EER of 0% and average accuracy rate of 95.1% are achieved using the proposed system. The ROC plots of FRR and FAR with MFCC order = 10 and 22 are shown in Figs. 7 and 8, respectively.

Table . Comparison of SVM based text-dependent speaker verification system with different MFCC orders

MFCC order	Average accuracy rate	EER
2	72.1%	12.2%
4	83.9%	5.8%
6	86.7%	2.2%
8	87.7%	2.7%
10	90.7%	2.0%
12	92.5%	0.7%
14	93.1%	1.3%
16	94.0%	0.4%
18	94.4%	0.0%
20	94.7%	0.2%
22	95.1%	0.0%
24	95.0%	0.0%
26	94.8%	0.4%

8. CONCLUSION

Once just a feature of luxurious high-end residences, home automation is now bridging to mainstream residences and will become a standard for new and existing homes. Traditionally, home automation systems have included very basic voice processing techniques that provide half-duplex speakerphone performance. As these terminals integrate speakerphone functionality, and home security systems rely on voice verification and recognition technologies, high-performing voice processing solutions become a key element in home automation system design. To maintain a low bill-of-material and achieve high performance, single-microphone systems need a voice processing solution with advanced noise reduction techniques (e.g. psychoacoustic noise reduction) that provide an improved performance over traditional noise reduction and limit distortion The MFCC technique has been applied for speaker identification. VQ is used to minimize the data of the extracted feature. The study reveals that as number of centroids increases, identification rate of the system increases. It has been found that combination of Mel frequency and Hamming window gives the best performance. It also suggests that in order to obtain satisfactory result, the number of centroids has to be increased as the number of speakers increases. The study

shows that the linear scale can also have a reasonable identification rate if a comparatively higher number of centroids is used. However, the recognition rate using a linear scale would be much lower if the number of speakers increases. Mel scale is also less vulnerable to the changes of speaker's vocal cord in course of time. The present study is still ongoing, which may include following further works. HMM may be used to improve the efficiency and precision of the segmentation to deal with crosstalk, laughter and uncharacteristic speech sounds. A more effective normalization algorithm can be adopted on extracted parametric representations of the acoustic signal, which would improve the identification rate further. Finally, a combination of features (MFCC, LPC, LPCC, Formant etc) may be used to implement a robust parametric representation for speaker identification. To achieve high performance in a small plastic enclosure, the designer has to drive a small speaker in a nonlinear range to meet audio requirements. An algorithm that can cancel non-linear echo and handle gain and distortion in the echo path solves the designer's problem. High integration, field upgradability and flexibility are also key criteria for a viable voice processing solution. Integration eases the design by reducing the complexity of interfacing multiple components and reduces the bill-of-material cost. Field upgradability and flexibility allow the designer to continuously enhance the feature set and add functionality without changing hardware.

References

[1] B. S. Atal, "Automatic recognition of speakers from their voices," in *Proc. IEEE*, 1976, vol. 64, pp. 460– 475.

[2] Douglas O'Shaughnessy and Hesham Tolba, "Towards a robust/fast continuous speech recognition systemusing a voiced-unvoiced decision," pp. 413– 416, ICASSP, 1999.

[3] C.R Jankowski Jr., T.F. Quatieri, and D.A. Reynolds, "Measuring fine structure in speech: Application to speaker idendification," in *IEEEICASSP*, 1995, pp. 325–328.

[4] Kemal S¨onmez, Larry Heck, Mitchel Weintraub, and Elisabeth Shriberg, "A lognormal tied mixture model of pitch for prosody–based speaker recognition," in *Proc. of EUROSPEECH*, september 1997, pp. 1391-1394.

[5] Kemal S¨onmez, Elisabeth Shriberg, Larry Heck, and Mitchel Weintraub, "Modeling dynamic prosodic variation for speaker verification," in Proc. of the International Conference on Spoken Language Processing, 1998, pp. 3189-3192.

[6] M.M Homayounpour and I. Rezaian, "Robust Speaker Verification Based on Multi Stage Vector Quantization of MFCC Parameters on Narrow Bandwidth Channels," ICACT 2008, vol 1, pp.336-340, Feb. 2008

[7] C.C. Lin, S.H. Chen, T. K. Truong, and Yukon Chang, "Audio Classification and Categorization Based on Wavelets and Support Vector Machine," IEEE Trans. on Speech and Audio Processing, Vol. 13, No. 5, pp. 644-651, Sept. 2005.

[8] L. Rabiner and B. H. Juang, Fundamentals of Speech Recognition, Prentice Hall, 1993

[9] S. B. Davis and P. Mermelstein, "Comparison of Parametric Repre- sentation for Monosyllabic Word Recognition in Continuously Spoken Sentences", IEEE Trans. On ASSP, vol. ASSP 28, no. 4, pp. 357-365, Aug. 1980.