Australian Journal of Basic and Applied Sciences, 9(15) Special 2015, Pages: 107-111

ISSN:1991-8178

Australian Journal of Basic and Applied Sciences

Journal home page: www.ajbasweb.com

Robust Speaker Recognition by Mfcc And Vq with Harmonic Noise Model Based Speech Enhancement

Radhai S, Senthamizh Selvi R, Suresh G.R

Department of Electronics and Communication Engineering, Easwari Engineering College, Chennai-600089

ARTICLE INFO

Article history:
Received 10 March 2015
Received in revised from 20 March
Accepted 25 March 2015
Available online 10 April 2015

Keywords:
HNM, H∞ filter, Mel frequency Cepstral Coefficient(MFCC),Vector Quantization(VQ) Speaker recognition

ABSTRACT

Speaker recognition is a process of identifying a person from a spoken phrase. The process make use of speaker's voice to verify the individual and control access to the services such as biometric security system, voice dialing, telephone banking, telephone shopping and security control for confidential information areas and remote access to the computers. But the recognition system is affected if there is a background noise. HNM based speech enhancement module is used to remove the background noise and make the speaker recognition system robust. In speaker recognition module Mel Frequency Cepstrum Coefficient (MFCC) Method is used for feature extraction and VQ-LBG Algorithm is used to match the feature. With the implementation of HNM based speech enhancement the speaker recognition module is able to recognise the speaker even at 0 dB noise level with an accuracy of 90%.

To Cite This Article: Radhai S, Senthamizh Selvi R, Suresh G.R, Robust Speaker Recognition by Mfcc And Vq with Harmonic Noise Model Based Speech Enhancement. ***Aust. J. Basic & Appl. Sci.,*** *9(15): 107-111, 2015*

INTRODUCTION

Speech enhancement system improves the quality and intelligibility of speech, degraded in the presence of background noise. While Speaker Recognition is a process of automatically recognizing who is speaking on the basis of the individual information included in speech waves. The speech degraded by noise results in reduction of speech discrimination. The objective to enhance the quality and intelligibility of speech, involves in manipulation of the degraded speech signal to alleviate noise effects. Which can make the speaker recognition system resistance to noise. The main advantage of model based speech enhancement is, it reduces the musical noise or musical tones.

In Harmonic Noise Model based approach acoustic parameters like pitch, spectral gain, and spectral envelope extracted from degraded signal (Chen, R, Chan, C. F.2012). Harmonic Noise Model (HNM) with H_∞ filter reconstructs the target speech based on the extracted parameters and it removes the background noise. Speaker Recognition mainly involves two modules namely feature extraction and feature matching.

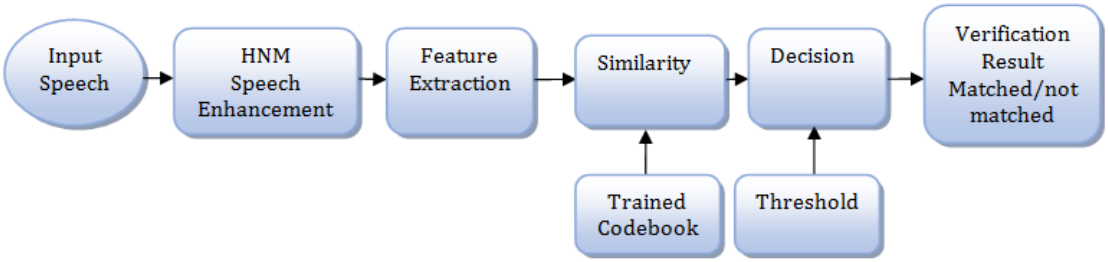

Fig. 1: Speaker Recognition Robust to Background Noise.

Feature extraction is the process that extracts a small amount of data from the speaker's voice signal that can later be used to represent that speaker. Feature matching involves the actual procedure to identify the unknown speaker by comparing the extracted features from his/her voice input. Feature extraction is done by Mel Frequency Cepstrum Coefficients, which are vector quantized using LBG

Corresponding Author: Radhai S, Department of Electronics and Communication, Easwari Engineering College, Bharathi Salai, Ramapuram, Chennai-600089.
E-mail: eradhai@gmail.com

algorithm resulting in the speaker specific codebook (Martinez, J.; Perez, H.; Escamilla, E.; Suzuki, M.M., 2012). In feature matching we find the VQ distortion between the input utterance of an unknown speaker and the codebooks stored in our database. Based on this VQ distortion with the threshold that is set we decide whether to match or not with the unknown speaker's identity.

The remainder of this paper is organized as follows. Section II, provides the design of HNM analysis-synthesis framework with H_∞ tracking of spectral envelope , the training phase of the speaker recognition in which computation of MFCC for the speakers listed in the database and create training codebook based on VQ –LBG algorithm where LBQ stands for Linde, Buzo, and Gray (LBG)who proposed the Vector Quantizaion (VQ) design algorithm and the testing phase which recognise the speaker based on the threshold.The system block diagram is illustrated in Fig 1. Section III, illustrates the matching results of the unknown speaker with the trained codebook .Finally, Section IV concludes the paper.

Methodology used:

HNM based speech enhancement is grouped into three stages as HNM analysis, H_∞ tracking and HNM synthesis.And speech feature exraction by computing MFCC which includes frame blocking, windowing ,Fast Fourier transform ,Mel Frequency wrapping and finally convert the log Mel Spectrum back to time. And Feature matching by Vector Quantization. Comparison is done by calculation the Euclidean distance between the codebooks of each speaker with unknown speaker which makes us to identify the corresponding speaker of the speech.

HNM -Based Speech Enhancement:

The degraded speech signal is pre-cleaned, so that it de-emphasizes the portion of spectrum which is corrupted by noise. The speech signal is assumed to be composed of voiced and unvoiced part. The pre- cleaned signal is divided into voiced and unvoiced part (Ephraim,Y. and Malah, D.1984). In the case of voiced segments, the maximum voiced frequency can be determined. The remaining parameters are estimated pitch-synchronously, so the analysis time instants are defined (Cohen, 2002). The magnitude and phase of the harmonics are then calculated. Fig 2 illustrates the above process

Pitch estimation:

Pitch estimation is the first process in HNM analysis. Pitch is estimated by normalized autocorrelation function (Chan, C. F and Yu, E. 1996). The pitch period T_0 is obtained by minimizing error function with respect to searching variable. The pitch period T_0 is obtained by minimizing error function with respect to searching variable T_0 as

$$\mathbf{T_0} = \arg\min \{a(t)\}. \quad (1)$$

Where a (t) is the error function.

Harmonic Magnitude estimation:

The harmonic part of excited speech signal is modeled as

$$E^T S = \sum_{k=1}^{v(m)} M_k(m)\cos(2\Pi(kT_0(m) + \Delta_k(m)) + \theta_k(m)). \quad (2)$$

Where v (m) represents the number of harmonics and T_0 (m) is the pitch, **E** and **S** are the harmonic amplitude vector and the harmonically related sinusoids vector respectively. In clean conditions, harmonic

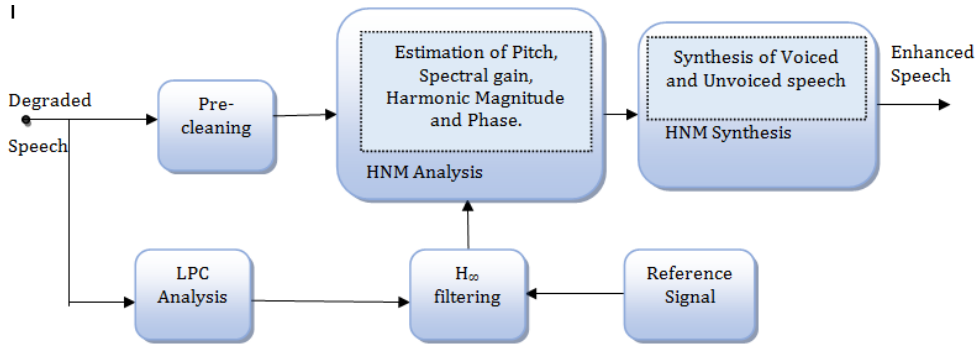

Fig. 2: Block Diagram of HNM Speech Enhancement.

Magnitudes are estimated through the optimum pitch period T_0.

$$e_h(m) = E^T S + d(m). \quad (3)$$

The harmonic magnitude and phase of the speech are modeled as LPC spectral envelope and then it is converted to a relation function (Chen, R, Chan, C. F.2012). In doing so, spectral modification can be achieved by correlating the LPC coefficients. This configuration incorporates the H_∞/ kalman /wiener tracked spectral envelope (Soon ,I. Y. Koh ,S. N. and Yeo C. K.1999).

Voiced and Unvoiced Speech Synthesis:

The voiced and unvoiced signals are reconstructed. The deterministic component is synthesized using a sum of sinusoids running at the harmonics of estimated pitch frequency

$$s_l(t) = \sum_{m=1}^{M(\tau_0)} H_{l,m}(t)\cos[\theta_{l,m}(t)]. \quad (4)$$

Where $H_{l,\ m}$ (t) and are the estimated m^{th} harmonic magnitude and phase, respectively. Harmonic phases extracted from degraded input signal are employed in speech enhancement. Harmonic magnitude from each frame index is sampled, (R. F. Chan, C. F. and So .H. C. 2012). The stochastic component is obtained by a direct application of random noise generator. Random Gaussian noise is generated and fed to the synthesis filter to produce unvoiced portion of speech. The spectrum is weighted by the V/UV mixing function.

H_∞ Filtering Algorithm:

The LPC analysis provide the spectral envelope information to the H_∞ filter. It is presumed that within an LPC analysis block, there exists certain inter-frame (between successive clean LSFs (line spectral frequency), i.e., X_k and X_{k+1}) and intra-frame (between noisy and clean LSFs, i.e. Y_k and X_k) linear relationships (Girin, L 2007). The clean LSFs X_k of the Kth frame are modeled as the internal states and the noisy LSFs Y_k are modeled as the observations, and they can be represented in state-space form as

$X_{k+1}=A\ X_k+BW_k$ (System Dynamic Model) (5)

$Y_k = C\ X_k+V_k$ (Measurement Model) (6)

Where W_k is the speech excitation and V_k is the background noise. A is an identity matrix, $B^T=C= [0\ 0\ 0……..1]_{1Xn.}$ No assumption is made on the nature of unknown quantities W_k and $V_{k,}$ (Lim, J. and Oppenheim, A.1978) and estimation of linear combination of X_k, is

$Z_k=LX_k$ (7)

Where $L \in R^{1Xn}$ the design criterion of the filter to provide a uniformly small estimation error $C_k= Z_k-\hat{Z}_{k.}$ and $\hat{Z}_k=L\hat{X}_k$ (Xuemin Shen and Li Deng 1999).

Where

$\hat{X}_k=A\ \hat{X}_{k-1}+H_k\ (Y_K-CA\hat{X}_{k-1})$ (8)

H_k is the gain of $H_{\infty\ filter}$.

$H_k=AP_K\ (1-\gamma^{-2}\bar{I}P_k + C^TV^{-1}CP_K)^{-1}C^TV^{-1}$ (9)

$P_{k+1}=AP_k\ (1-\gamma^{-2}\bar{I}P_k\ +\ C^TV^{-1}CP_K)^{-1}A^{TV-1}+BWB^T$ (10)

$P_0 =0,\ \gamma>0.$ (11)

Feature Extraction:

A block diagram of the structure of an MFCC processor is given in Fig 3. (Martinez, J. Perez, H. 2012)

Pre-processing: The first step involves the conversion of analog speech signal into digital speech signal. Continuous Speech is sampled at a discrete time to form a sample data signal. The process of obtaining a discrete time representation of a continuous time signal through periodic sampling

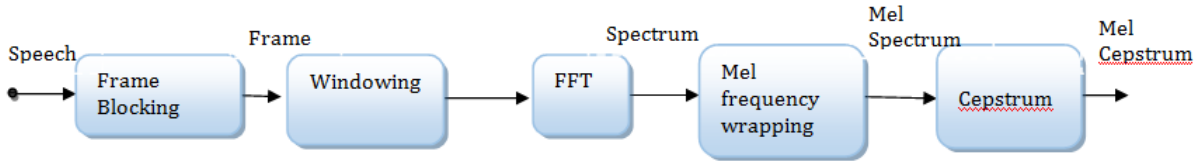

Fig. 3: Block Diagram of Feature Extraction.

Frame Blocking:

Framing is the process of segmenting the speech samples obtained from the analog to digital conversion into small frames with time length in the range of 20ms to 40 ms.

Windowing:

Windowing is performed on each individual frame so as to minimize the signal discontinuities at the beginning and end of each frame. Hamming window is used.

Mel-frequency Wrapping:

For each actual frequency f in Hz, a pitch is measured on a scale called the 'mel' scale. The mel-frequency scale has linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. As a reference point, the pitch of a 1 kHz tone, 40 dB above the perceptual hearing threshold, is defined as 1000 mels.

Mel Filterbank:

In this magnitude frequency response is multipled by a set of 20 triangular band pass filters to get the log energy of each triangular band pass filter. The positions of these filters are equally spaced along the Mel frequency, which is related to the common linear frequency f by the following equation

mel (f) =1125*ln (1+f /700) (12)

Cepstrum:

In this final step, the conversion of the log Mel spectrum back to time. The result is called the Mel frequency cepstrum coefficients (MFCC). Discrete Cosine Transform (DCT) is used to convert them back to the time domain. Therefore if we denote those Mel power spectrum coefficients that are the result of the last step are $\tilde{S}_0, k = 0,2,\ldots,K-1$, we can calculate the MFCC's, $\tilde{c}_n$, as

$$\tilde{c}_n = \sum_{k=1}^{K}(\log\ \tilde{S}_k)\cos\left[n\left(k-\frac{1}{2}\right)\frac{\pi}{K}\right] \quad (13)$$

Feature Matching:

LBG-VQ Algorithm: The LBG-VQ design

algorithm is an iterative algorithm which alternatively solves the optimality criteria. The algorithm requires an initial codebook. The algorithm is formally implemented by the following recursive procedure:

1. Design a 1-vector codebook; this is the centroid of the entire set of training vectors
2. Double the size of the codebook by splitting each current codebook **y** according to the rule

$$\mathbf{y}_n^+ = \mathbf{y}_n(1+\varepsilon)$$

$$\mathbf{y}_n^- = \mathbf{y}_n(1-\varepsilon)$$

where n varies from 1 to the current size of the codebook, and ε is a splitting parameter

3. Nearest-Neighbor Search: for each training vector, find the codeword in the current codebook that is closest, and assign that vector to the corresponding cell.
4. Centroid Update: update the codeword in each cell using the centroid of the training vectors assigned to that cell.
5. Iteration 1: repeat steps 3 and 4 until the average distance falls below a preset threshold

Iteration 2: repeat steps 2, 3 and 4 until a codebook size of M is designed. In the recognition phase, an input utterance of an unknown voice is "vector-quantized" using each trained codebook and the total VQ distortion is computed. The speaker corresponding to the VQ codebook with smallest total distortion is identified

RESULTS AND DISCUSSIONS

To analyze and evaluate the performance of the Speaker Recognition technique with speech enhancement module, TIMIT database has been used. The clean speech data is taken from:

Table 1: Recognition Result without Speech Enhancement.

Samples	Recognition Result with-out Speech Enhancement Algorithm								
	Car Noise (0 dB)	Car Noise (5 dB)	Car Noise (10 dB)	Babble Noise (0 dB)	Babble Noise (5dB)	Babble Noise (10 dB)	Airport Noise (0 dB)	Airport Noise (5 dB)	Airport Noise (10 dB)
Speaker 1	Not Matched	Not Matched	Matched	Not Matched	Not Matched	Matched	Not Matched	Not Match	Matched
Speaker 2	Not Matched	Matched	Matched	Not Matched	Not Matched	Matched	Matched	Not Match	Matched
Speaker 3	Not Matched	Not Matched	Matched	Not Matched	Not Matched	Matched	Not Matched	Not Match	Matched
Speaker 4	Not Matched	Not Matched	Matched	Not Matched	Not Matched	Matched	Not Matched	Not Match	Matched
Speaker 5	Not Matched	Not Matched	Matched	Not Matched	Not Matched	Matched	Not Matched	Not Match	Matched
Speaker 6	Not Matched	Not Matched	Matched	Not Matched	Not Matched	Matched	Matched	Not Match	Matched
Speaker 7	Not Matched	Matched	Matched	Not Matched	Not Matched	Matched	Not Matched	Not Match	Matched
Speaker 8	Not Matched	Not Matched	Matched	Not Matched	Not Matched	Matched	Not Matched	Not Match	Matched
Speaker 9	Not Matched	Not Matched	Matched	Not Matched	Not Matched	Matched	Not Matched	Not Match	Matched
Speaker 10	Not Matched	Not Matched	Matched	Not Matched	Matched	Matched	Not Matched	Not Match	Matched

TIMIT acoustic phonetics speech corpus. 10 different speakers are selected and trained. The speech data are originally sampled at 16 kHz and quantized to 16 bits. The data is appropriately filtered and down sampled to 8 kHz to obtain the narrow band speech which are used in our investigation. Performance of the speech enhancement with speaker recognition system is evaluated in this section. Three types of noise namely babble noise, car noise, airport noise, are employed. Clean speech is manually corrupted at SNR level of 0, 5, and 10 dB.

Recognition Result with Speech Enhancement Algorithm:

With the implementation of HNM based Speech enhancement algorithm the Speaker is matched even under very high noise level (0 dB noise level) at an accuracy of 90% with the improvement in the SNR level as shown in Table 2. The Euclidean distance plot which shows the VQ distortion between the input speech with the speakers in the trained codebook which is illustrated in Fig 4.

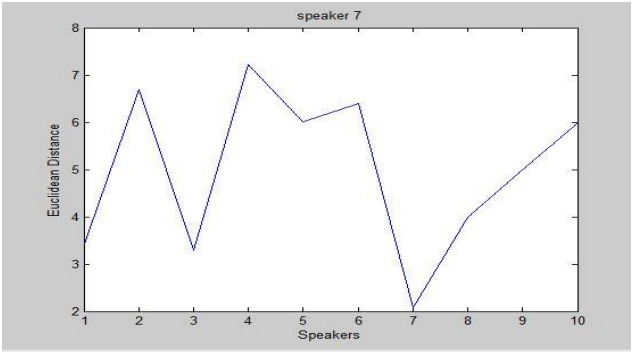

Fig 4: Euclidean distance plot.

Table 2: Recognition Result at 0 dB SNR with Speech Enhancement.

Samples	Recognition Result		
	Car Noise (0 dB)	Babble Noise (0 dB)	AirportNoise (0 dB)
Speaker 1	Matched	Matched	Matched
Speaker 2	Matched	Matched	Matched
Speaker 3	Matched	Matched	Matched
Speaker 4	Matched	Matched	Matched
Speaker 5	Matched	Matched	Matched
Speaker 6	Matched	Matched	Matched
Speaker 7	Matched	Matched	Matched
Speaker 8	Matched	Matched	Matched
Speaker 9	Not Matched	Not Matched	Not Matched
Speaker 10	Matched	Matched	Matched

Conclusion:

We have proposed a robust Speaker recognition System using Mel Frequency Cepstral coefficients and vector Quantization along with HNM based speech enhancement. The HNM speech enhancement module improves the SNR of the degraded speech signal and makes it suitable for speaker recognition. The proposed speaker recognition system can recognize the speaker even at very high noise level (0 dB SNR) with 90% accuracy for background noises like car, babble and airport. This process can be extended for n number of speakers with stationary and non-stationary noises.

REFERENCES

Chan, C.F. and E.W.M. Yu, 1996. 'Improving pitch estimation for efficient multiband excitation coding of speech', Electron. Lett., 32(10): 870–872.

Chen, R., C.F. Chan, 2012. 'Model-Based Speech Enhancement With Improved Spectral Envelope Estimation via Dynamics Tracking', IEEE Trans. Audio, Speech, Lang. Process, 20(4): 1324-1336.

Chen, R.F., C.F. Chan and HC. So, 2012. 'Noise suppression based on an analysis–synthesis approach', in Proc. Eur. Signal Process. Conf. (EUSIPCO) IEEE Transactions On Audio, Speech, And Language Processing, 20(4): 1539-1543.

Cohen, 2002. 'Optimal speech enhancement under signal presence uncer- tainty using log-spectral amplitude estimator, IEEE Signal Process. Lett., 9(4): 113–116.

Ephraim, Y. and D. Malah, 1984. 'Speech enhancement using a minimum mean square error short-time spectral amplitude estimator',IEEE Trans.. Acoust., Speech, Signal Process., 32(6): 1109–1121.

Gardner, W. and B. Rao, 1995. 'Theoretical analysis of the high-rate vector quantization of LPC parameters', IEEE Trans. Speech Audio Process, 3(5): 367–381.

Girin, L., 2007. 'Long-term quantization of speech LSF parameters', in Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP), 4: 845–848.

Laroche, J., Y. Stylianou and E. Moulines, 1993. 'HNM: A simple, efficient harmonic plus noise model for speech', Proc. Workshop Applicat. Signal Process. Audio Acoust. (WASPAA), pp: 169-172.

Lim, J. and A. Oppenheim, 1978. 'All-pole modeling of degraded speech', IEEE Trans. Acoust., Speech, Signal Process., 26(3): 197–210.

Martinez, J., H. Perez, E. Escamilla, M.M. Suzuki, 2012. "Speaker recognition using Mel frequency Cepstral Coefficients (MFCC) and Vector quantization (VQ) techniques," International Conference on Electrical Communications and Computers, pp: 248-251.

Soon, I.Y., S.N. Koh and C.K. Yeo, 1999. 'Improved noise suppression filter using self-adaptive estimator of probability of speech absence', Signal Processing, 75(2): 151–159.

Xuemin Shen and Li. Deng, 1999. 'A Dynamic System Approach to Speech Enhancement Using the Filtering Algorithm', IEEE Trans. Audio, Speech, Lang. Process, 7(4): 391-399.