International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

A Review of Comparatively Study of Different Speaker Recognition Techniques

Umer Malik

M.Tech Student, Computer Science & Engineering, Sharda University, Plot No. 32-34, KnowledgePark III, Greater Noida, U.P., India – 201306

Abstract: ***In this paper, we describe a brief overview of the speaker recognition techniques with their processing steps. Speaker recognition has many problems in feature extraction due to the robustness of the speech with noise. Gamma Tone Filter Bank and Wavelet Packet for the speaker recognition have the best performance over the Hidden Markov Model, Mel Frequency Cepstral Coefficient, Dynamic Time Warping and Layered Recurrent Neural Network. The system performance was measured by recognition rate with various signal-to-noise ratios over -10 to 10 dB.***

Keywords: Speaker recognition, neural network, Gamma tone filter, wavelet packet, HMM, MFCC

1. Introduction

Speaker recognition is the process of identifying the speaker with voice of the speaker rather than what they are verbally speaking. Identification of the speaker can be used to verify identity of the speaker as the part of security process. The major problem in speaker recognition systems is the extracting speaker voice with better performance from the noisy speech signal. Error rate can be minimized by keeping the recognition system in a separate box where there are no interfering signals. The best example of speaker and speech recognition is the human ear with our brain memory. Human ear extracts the speech features from the huge noisy signals which make it best speaker and speech recognition system ever. This review paper describes two quantitative models for signal processing in auditory system (i) Gamma Tone Filter Bank (GTFB) and (ii) Wavelet Packet (WP) as frontends for robust speech recognition which are taken from [1]. These auditory feature vectors are used to train neural network. The classifiers are used for the feature vectors by the neural network using Back Propagation (BP) algorithm. The designed system's [1] performance was compared with various types of front-ends and recognition methods such as auditory features with Hidden Markov Model (HMM) & Layered Neural Network (LRNN), auditory features with Mel Frequency Cepstral Coefficient (MFCC) & LRNN and vocal tract model: MFCC & HMM, Dynamic time warping (DTW).

2. Some Conventional Speech Recognition Techniques

2.1 Hidden Markov Model (HMM)

A Markov chain having, that is only partially observable, is said to be HMM. It may also be said that the observations related to the states of a system are not sufficient to determine the states exactly. Markov model is a stochastic model that is used to model randomly changing systems where it is assumed that the future states are dependent only on present states and not on proceeding states. Reasoning and computation are possible only due the assumptions made. Viterbi algorithm and forward algorithm are well known algorithms for HMM. Baum-Welch algorithm is another example of HMM.
Sequence analysis using HMM:

2.1.1 Construct an HMM model.

- Design an HMM generator for the observed sequences.
- Assign hidden states to sequence regions.
- Set up the question to be answered in terms of hidden path way.

2.1.2 Train the HMM

- Supervised or Unsupervised.

2.1.3 Analyze sequences

- Viterbi decoding: Compute most likely hidden path way.
- Forward/Backward: Compute likelihood of sequences.

2.2 Mel Frequency Cepstral Coefficient (MFCC)

The coefficients that collectively make up Mel Frequency Cepstrum (MFC) are said to be MFCC. MFC is the representation of short term power spectrum of sound. It is based on Linear Cosine Transform of a log power spectrum on a non-linear Mel Scale of frequency. MFC is different from Cepstrum as in MFC, the frequency bands are equally spaced on Mel Scale and MFC can study the human auditory system more closely. Noise sensitivity of MFCC is not robust and inefficient in presence of the additional noise.

Thus allows better representation of sound. Here, there are some steps for the MFCC derivation.

- Take Fourier Transform (FT) of sound signal.
- Powers of spectrum obtained in above step are mapped by using Triangular overlapping window.
- Take log of powers at each Mel frequency.
- MFCC's are amplitude of the resulting spectrum.

2.3 Dynamic Time Warping (DTW)

Under certain boundary conditions, similarity between any two given time dependent sequences can be found by very well-known algorithm called as DTW. In order to make a comparison, the sequences are warped in Non Linear

Pattern. DTW was basically used to compare the speech patterns in automatic speaker recognition. Lately it was applied to fields like information retrieval from audios and videos and data mining. DTW can analyze any data that can be turned into linear sequence. Speaker recognition is a well-known example of it. DTW has some limitations like it has quadratic time and space complexity that limits its use to small time series.

2.4 Wavelet Packet Filter Bank using Neural Network

Most of the research work going on in the field of speech recognition is to improve the performance of the recognition of the noisy speech. Human mind processes the signals using neural networks which work very fast due to parallel processing. So, neural network is better than other techniques for recognizing speech signals.The speaker independent system is improved by utilizing a different cochlea model which is designed with a high resolution Wavelet Packet Filter Bank (WPFB). A speaker dependent recognition system using Gamma Tone Filter Bank (GTFB) as the front-end and Back Propagation Neural Network (BPNN) [1] as the recognition method has also been developed for better performance.

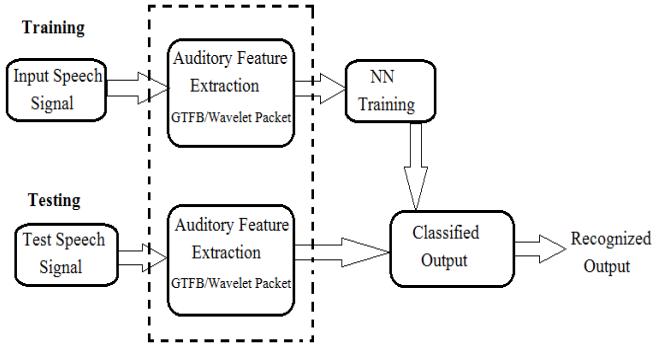

Figure 1: Functional block

Functional block of this technique is shown in Fig. 1. It contains two main stages; (a) Training stage (b) Testing stage. During the training, the input speech signal samples from 2-3 male voice uttering of each five times were collected as source database. The speech signals will be given as input to auditory feature extraction unit. The feature vectors will be extracted in two ways - (a) GTFB front-end (b) WP front-end. By utilizing these auditory feature vectors, BPNN will be trained until reaching the target goal. The second stage of this project is testing, when the test inputs given to system, feature vectors will be extracted and will be given as input to an already trained network. The neural network performs the classification.

2.4.1 Auditory Feature Extraction

Signal processing front end for extracting the feature set is an important stage in any speech recognition system. This technique is based on the human auditory system characteristics named as Model of auditory Periphery. It relies on the GTFB to emulate the cochlea frequency resolution.

2.4.1.1 Model of the Auditory Periphery

The model of auditory perception [3] was designed as a model of the 'effective' signal processing that takes place in the auditory periphery transforming the acoustic signal into its 'internal representation'. Fig. 2(a) and 2(b) represent this model.

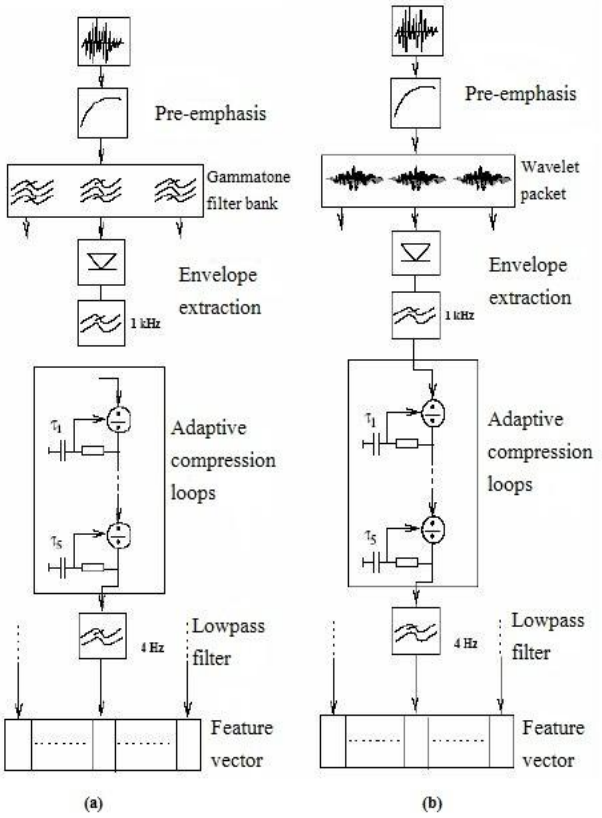

Figure 2: Processing stages of the auditory model (a) Gamma Tone Filter Bank front-end (b) Wavelet Packet front-end

- *Pre-emphasis:* The first processing step is a pre-emphasis of the input signal with a first-order high pass filter. This increases the relative energy of the higher frequency spectrum. The transfer function of such afilter is
  $H(z) = 1 - \alpha z - 1$
  where α is constant and $0.9 \leq \alpha \leq 1$.
- *GTFF:* It models the cochlea by a bank of overlapping band pass filters. The impulse response of each filter follows the Gamma tone function shape. This function was introduced by Aertsen and Johannesma [4]. It has the following classical form:
  $$h(t) = \gamma(n,b)t^{n-1}e^{-bt}\cos(\omega t + \phi)u(t)$$
  Here $\gamma(n,b)$ is a normalization constant depending on the order, n, and the bandwidth related factor, b, w is the radian center frequency, φ is the phase shift and u(t) is a unit step function.
- *Envelope extraction:* After gamma tone filtering, each frequency channel is half wave-rectified and first-order low pass filtered with a cut-off frequency of 1 kHz for envelope extraction, which reflects the limiting phase-locking for auditory nerve fibers above 1 kHz.
- **Adaptive Compression Loop:** Amplitude compression is performed in a subsequent processing step. In contrast to conventional bank-of-filters front ends, the amplitude compression of the auditory model is not static (e.g., instantaneously logarithmic) but adaptive, which is realized by an adaptation circuit consisting of five consecutive non-linear adaptation loops. Each of these

loops consists of a divider and an RC low pass filter with an individual time constant ranging from 5 to 500ms.

- **Feature Extraction***:* The last processing step of the auditory model is a first-order low pass filter with a cut-off frequency of 4 Hz. It attenuates fast envelope fluctuations of the signal in each frequency channel. Very slow envelope fluctuation suppressed by the adaptation loops and attenuation of fast fluctuations by the low pass filter results in a band pass characteristic the amplitude modulation transfer function of the auditory model with a maximum at about 4 Hz.

3. Results and Conclusions

Two speaker independent robust recognition systems using auditory features for isolated words (1-5) were developed using Gamma Tone Filter Bank (GTFB) as front-end and Back Propagation NeuralNetwork (BPNN) as the recognition method, Wavelet Packet (WP) filter bank as the front-end – BPNN as the recognition method.Auditory based methods give better performance for low SNRs compared to feature extraction using source model. High resolution wavelet packet filter bank gives highest recognition rate compared to the other auditory based methods since the impulse response of the filters used in this implementation is the wavelet function (‘db4’) which is having compact support and suitable for non-stationary signals like speech.The system performance was measured by recognition rate with various signal-to-noise ratios over -10 to 10 dB.The comparison of recognition rate of proposed (i) GTFB front-end –BPNN and (ii) Wavelet Packet- BPNN with different front-end and recognition techniques are shown in Fig. 3 and Fig. 4.

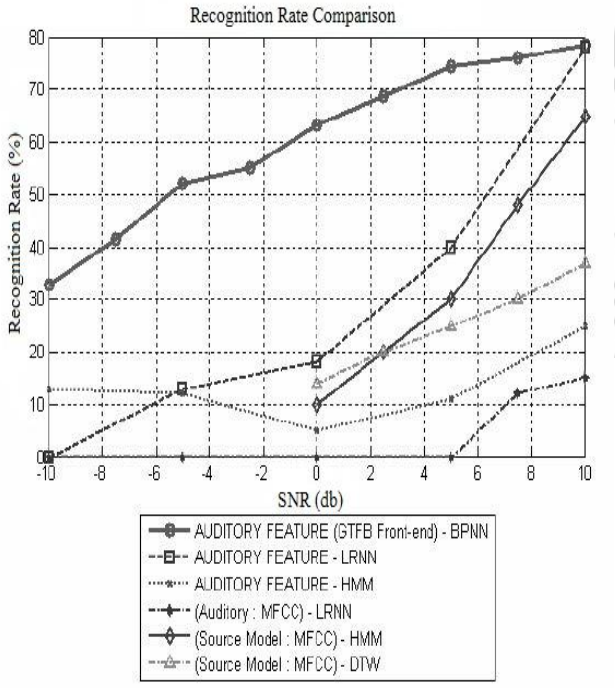

Figure 3: (Speaker Independent) recognition rate comparisons with GTFB as front-end vs different front-ends and recognition techniques

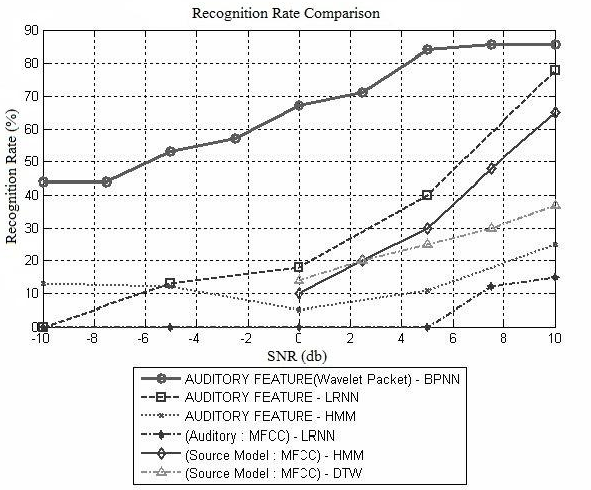

Figure 4: (Speaker Independent) recognition rate comparisons with Wavelet Packets as front-end vs different front-ends and recognition techniques

References

[1] R.Gandhiraj, Dr.P.S.Sathidevi, "Auditory-based Wavelet Packet Filter bank for Speech Recognition using Neural Network," 15th International IEEE Conference on Advanced Computing and Communications, pp. 666-671, 2007

[2] M. Kleinschmidt, J. Tchorz and B. Kollmeier, “Combining speech enhancement and auditory feature extraction for robust speech recognition”, Speech Communication 34 (2001), 75-91.

[3] T.Dau, D. Puschel, and A. Kohlrausch, “A quantitative model of the effective signal processing in the auditory system: I. Model structure.” J. Acoust. Soc. Am. 99 (6), 3615-3622, 1996

[4] T.Dau, D. Puschel, and A. Kohlrausch, “A quantitative model of the effective signal processing in the auditory system: II. Simulations and measurements”, J. Acoust. Soc. Am. 99 (6), 3623-3631, 1996

[5] W.H. Abdulla, “Auditory based feature vectors for speech recognition systems”, Electrical and Electronic Engineering Department , The University of Auckland, 20 Symonds Street, Auckland, New Zealand.

[6] T.H. Hwung, L.M. Lee and H.C. Wung, “Feature adaptation using deviation vector for robust speech recognition in noisy environment”, IEEE International Conference on Acoustics, Speech, and Signal Processing 1997.ICASSP-97., 1997.