IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 1 Issue 9, November 2014.
www.ijiset.com
ISSN 2348 – 7968

A Deep and Wide Analysis for Speech-Emotion Recognition Using Multilayer Perceptron

T.Jayasankar[1], J.Jayalakshmi[2] K.Rajasekaran[3]

[1] Electronics and Communication Engineering Department, Anna University BIT Campus, Trichy
[2] Electronics and Communication Engineering Department, PABCET Trichy
[3] Electronics and Communication Engineering Department, M.A.R College of Engineering, Trichy

Abstract

This paper reviews a line of research carried out over the last decade in speech-Emotion recognition assisted by some testing & training methods .The particular focus is on the use of deep & wide analysis of speech Emotion recognition such as joy, sad, Anger by using Hidden Markov Model, Multilayer Perceptron. This recognition is based on robust and reliable performance of speech processing, recognition, detection, can also be significant factors. Speech Emotion recognition systems based on the several classifiers. Compare to other classifiers HMM, MLPs are relatively common in speech recognition as it is easy for implementation with accuracy & it has well defined training algorithm.

Keywords: *Machine learning, Multilayer Perceptron, Hidden Markov Model, Speech recognition.*

1. Introduction

Speech is the most natural form of human communication. It is an immensely information-rich signal exploiting frequency-modulated, amplitude-modulated and time-modulated carriers to convey information about words, speaker identity, accent, expression, style of speech, emotion and the state of health of the speaker. The speech signal is produced from the vocal tract system by varying its dimension with the help of articulators and exciting with a time varying source of excitation. The physical structure and dimension of the vocal tract, as well as of the excitation source, are unique for each speaker

2. Speech Recognition

AUTOMATIC speech recognition (ASR) has a long history, minimally dating back to the 1952 Bell Labs paper describing a technique for digit recognition [1].Similarly, machine learning has a long history, with significant development in the branch commonly called neural networks also going back to at least the 1950s. Speech recognition methods converged by 1990 into statistical approaches based on the hidden Markov model (HMM), while artificial neural network (ANN) approaches in common use tended to converge to the multilayer perceptron (MLP) incorporating back-propagation learning. More recently, there has been considerable interest in neural network approaches to phone recognition that incorporate many of the stylistic characteristics of MLPs (multiple layers of units incorporating nonlinearities), but that are not restricted to back propagation for the learning technique [2].

It should be noted in passing that the earliest ANN training methods did not use error back propagation; for instance, the Discriminant Analysis Iterative Design (DAID) technique developed at Cornell in the late 1950s incorporated multiple layers that were separately trained, using Gaussian kernels at the hidden layer and a gradient training technique at the output layer [3]. It has been important for machine intelligence researchers to conduct experiments with modest-sized tasks in order to permit extensive explorations; on the other hand, conclusions from such experiments must be drawn with care, since they often do not scale to larger problems. For this reason, among others, many researchers have gravitated towards large-scale problems such as large-vocabulary speech recognition which often incorporate hundreds of millions of input patterns, and can require the training of tens of millions of learned parameters [4].

Given the maturity of the speech recognition field, competitive performance often requires the use of complicated systems, for which any novel component plays a minor role. Modern speech recognition systems, for instance, incorporate large language models that use prior information to strongly weight hypothesized utterances towards task-specific expectations of what might be said. Thus, it can be difficult to see the advantage of a new method. However, if improving speech recognition is our goal, there is prior speech analysis is important.

This paper will describe various analysing methods developed over the last decade that incorporate multilayers of analysis to provide speech recognition. In each cases the emphasis will be to describe analysis that have exploited structures incorporating both a large number of layers (deep) and multiple streams using MLPs with large hidden layers (wide) . In some cases the underlying model is at least initially generative (as with the maximum likelihood training used in conventional ASR systems prior to

ISSN 2348 – 7968

discriminative training), but in other cases the methods are discriminative from the start.
The focus here will be on what are now classical methods, as well as newer approaches making use of discriminatively trained features. In most cases these systems are inherently heterogeneous, incorporating a sequence of analysing layers that perform differing functions. The class of systems incorporating deep belief networks, which are fundamentally generative nature but also homogeneous in their form and training, will be emphasized in other papers (in this special issue). Speech signal analysing has many efficient and intelligent applications like speech recognition, speaker transformations and text-to-speech systems, Emotion recognition which is continuously gaining a serious attention in automatic speech recognition.

2.1 Emotion Recognition

Emotions would enable a communication virtual agent to both maintain a user's positive emotional state and allow it to judge and refine its current dialogue strategy. One way to determine the emotional state of the user in this context is through an interpretation of the speech signal analysis. In order to use virtual agents in the role of educator, it is necessary to provide them with the capability to sense the emotional state of the human user, so they can offer the appropriate and natural response.

By recognizing and expressing emotions, the interactive educational agent could act in an enthusiastic manner towards the teaching material and also show empathy towards student development. If the interactive agent has an appealing and believable personality, this would have the effect of making the interaction more enjoyable for the students. It is felt that students will relate to virtual agents if they can connect with their personality and their emotional representation.

There has seen a growth in interest in emotional speech recognition using intelligent approaches to aid agent-person interaction. Emotional speech recognition that made use of a emergent unsupervised approach. Emotion recognition system based on a state vector machine within a virtual agent named Greta. The following emotions are

Happiness: The mean pitch of the utterance is high, has a high variance, the rhythm is rather fast, few syllables are accented, the last word is accented, and the contours of all syllables are rising.
Anger: The mean pitch is high, has a high variance, the rhythm is fast, with little variance of phoneme durations, a lot of syllables are accented, the last word is not accented, and the pitch contours of all syllables are falling.
Sadness: The mean pitch is low, has a low variance, the rhythm is slow, with high variance of phoneme durations, very few syllables are accented, the last word is not accented, the contours of all syllables are falling.
The rest of the paper is divided as follows. In Section 3 give the prior work and in Section 4 describes about proposed method, after that report and results in Section 5 elaborate the proposed method and finally conclude this paper with conclusion and future works in Section 6.

3. Prior Work

In existing system the feature extraction (analysis) involves the analysis of speech signal in the pre-processing system. It is mainly based on temporal analysis; it uses the speech waveform itself used for analysis. Rarely based on spectral analysis .The following methods are used for analysing the speech signal for speech recognition. (i) Discrete Cosine Transform (DCT), (ii) Critical band windowing,(iii) IDFT squared Magnitude, (iv) Minimum mean square Error(MMSE), (v)Discrete Fourier Transform (DFT), (vi)Mel-Frequency cepstral coefficients(MFCCs),(vii) Perceptual-Linear-Prediction (PLP).DCT is used for Long segments of speech signals are decomposed into sub bands, windowing performs filter operations of noisy signal, DFT & IDFT square magnitude provides sub band envelope of speech signal, MMSE technique is applied to estimate the clean envelope from the noisy speech, MFCCs representation is approximately holomorphic filter that have smooth transfer function with high quality and sensitivity, PLP is Eigen vector based feature extraction. It defines energy compaction of speech signal. The above speech parameters are used for creating only basic data base for speech recognition parameters are used for creating only basic data base for speech recognition.

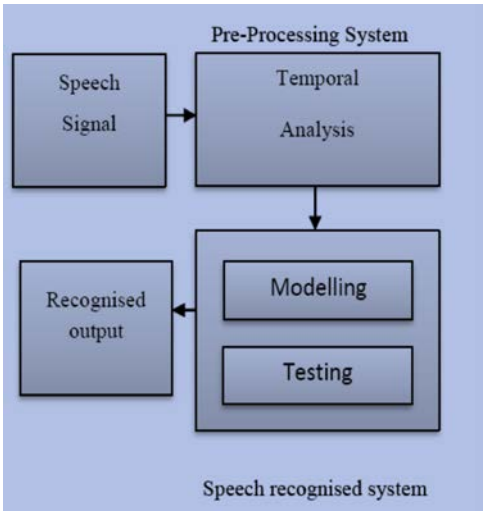


Fig 1.Existing system

ISSN 2348 – 7968

parameters are used for creating only basic data base for speech recognition.

It required more training and testing algorithm for optimizing the performance and complex computation, latency occurs due to more temporal based analysis. Transfer function with high quality and sensitivity, PLP is Eigen vector based feature extraction. It defines energy compaction of speech signal. The above speech parameters are used for creating only basic data base for speech recognition. It required more training and testing algorithm (GMM) Gaussian Markov model is used for optimizing the performance and complex computation, latency occurs due to more temporal based analysis. The speech recognition system became more complex when more and more training data became available. Large amount of information in the waveforms are discarded that is considered to be irrelevant for discrimination. Computation complex due to large temporal analysis. Speech recognizer does not perform well in the presence of low level noise. It requires more separate training and testing algorithm (or models) for speech analysis. This prior reviews the neural networks. It can be very powerful speech signal classifiers. A small set of words could be recognized with some very simplified models. The pre-processing quality is giving the biggest impact on the neural networks performance. In some cases where the spectrogram combined with entropy based endpoint detection is used to observed poor classification performance results, making this combination as a poor strategy for the pre-processing stage.

The system is comprised of three main parts, a pre-processing part, a feature extracting part and a recognition part. In the feature extracting, the parameters such as Fourier Transform uses fixed sized windows, Wavelet Transform to extract the emotion recognition. In recognition part improved HMM as the emotion recognizer. This method is based on the Chinese corpus of emotional speech synthesis database. This method is effective and high speed when it requires more training and testing process otherwise speed very slow.

4. Proposed System

This work is based on deep and wide analysis of speech signal which is recognized by using some (HMM,MLP) training & testing methods. Deep and wide analysis is one of the important feature extraction techniques in speech (Emotion) recognition. The objective of the project is to design an efficient pre-processing system for automatic speech (Emotion) recognition to obtain the perceptually meaningful parameters (Joy,Sad,Angry) that capture spectral dynamics of speech signal or changes of spectrum with time The parameters which are robust to the variations, according to the recognition system.

The goal of speech recognition is for a machine to be able to "hear," understand," and "act upon" spoken information and recognition is to analyze, extract characterize and recognize information about the speaker identity. The speaker recognition system may be viewed as working in a four stages.
(i)Pre-Processing, (ii) Feature Extraction(analysis), (iii) Modeling, (iv) Testing.

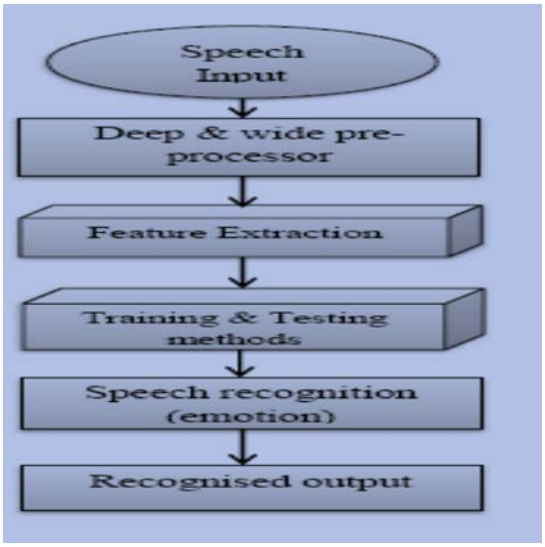


Fig.2. Proposed system

4.1 Pre-processing

This step is the first step to create feature vectors. The objective in the pre-processing is to modify the speech signal, so that it will be more suitable for feature extraction analysis. The pre-processing operation based on noise cancelling, pre-emphasis and voice activation detection. Two commonly used noise reduction algorithm in the field of speech context is spectral subtraction& adaptive noise cancellation.

(i) Temporal Analysis: Temporal analysis is the simple deep and wide analysis .Temporal features are easy to extract and have easy physical interpretation. Temporal features are Spectrum analysis, (i) Linear Predictive Coding analysis, (ii) Cepstrum analysis.

(ii) Spectral Analysis: Spectral analysis gives lot of information about the speech signal. Time domain data is converted to frequency domain by applying Fourier Transform it. This process gives the spectral information. Spectral information is the energy levels at different frequencies in given window. Thus the features like frequency with maximum energy, distance between frequencies of maximum and minimum energies can be extracted. The spectral features are(i)Energy

ISSN 2348 – 7968

normalization, (ii)Zero-crossing rate, (iii) Short term energy,(iv) Sub frame rate.

4.2 Feature Extraction (Analysis)

The goal of feature extraction is to find a set of properties of an utterance that have acoustic correlations to the speech-signal, that is parameters that can somehow be computed or estimated through processing of the signal waveform. Such parameters are termed as features. It includes the process of measuring some important characteristic of the signal such as energy or frequency response, augmenting these measurements with some perceptually meaningful derived measurements and statically conditioning these numbers to form observation vectors.

Modelling: The objective of modeling technique is to generate speech recognition models using specific feature vector. The speech recognition is also divided into two parts that means speaker dependant and speaker independent. In the speaker independent mode of the speech recognition the computer should ignore the speaker specific characteristics of the speech signal and extract the intended message .on the other hand in case of speaker recognition machine should extract speaker characteristics in the acoustic signal. The main aim of speaker identification is comparing a speech. The deep and wide analysis pre-processor consists of two analyses. (i) Temporal Analysis. (ii) Spectral Analysis.

4.3. Linear Predictive Coding Analysis

Linear prediction is a method for signal source modeling dominant in speech signal processing and having wide application in speech processing. This method has become predominant technique for estimating the basic speech parameter such as Pitch, Formants, Spectra, Vocal tract area function. It is used for representing the speech for low bit rate transmission or storage.

The Importance of the method lies both in its ability to provide extremely accurate estimates of the speech parameters and in its relative speed of computation. Starting with a demonstration of the relationship between linear prediction and the general difference equation for linear systems, the unit shows how the linear prediction equations are formulated and solved. The unit then discusses the use of linear prediction for modeling the source of a signal and the signal spectrum.

X (n) = ∑ αk x (n-k) for some values of p=1 to k

4.4 LPC Co-Efficient Estimation

Linear Predictive Coding (LPC), also known as LPC analysis or Auto-Regressive (AR) modelling .This method is widely used because it is fast and simple, yet an effective way of estimating the main parameters of speech signals. Basic idea of linear prediction: current speech sample can be closely approximated as a linear combination of past samples, i.e. The predictor coefficients are determined (computed) by minimizing the sum of squared differences (over the finite interval) between the actual speech samples and the linearly predicted ones.

The solution of the Yule-Walker equations can be achieved with any standard matrix inversion package. Because of the special form of the matrix here, some efficient solutions are possible, as described below. Also, each solution offers a different insight represent three different algorithms: Covariance method, Autocorrelation method , Lattice method .

The coefficients are calculated by minimising the short-term mean-squared prediction error is

$$E\{e^2[n]\} = E\left\{\left(s[n] + \sum_{k=1}^{p} a_k s[n-k]\right)^2\right\}$$

It type, with subscripts and superscripts in a slightly smaller font size. This is acceptable.

The LPC coefficients can be solved using the set of linear equations known as the Yule-Walker equations:

$$\begin{bmatrix} R_n(0) & R_n(1) & \cdots & R_n(p-1) \\ R_n(1) & R_n(0) & \cdots & R_n(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ R_n(p-1) & R_n(p-2) & \cdots & R_n(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} = \begin{bmatrix} R_n(1) \\ R_n(2) \\ \vdots \\ R_n(p) \end{bmatrix}$$

$$R_n(j) = \sum_{m=0}^{N-1-j} s_n[m]s_n[m+j]; \quad j = 0,\ldots, p.$$

The above matrix is a Toeplitz matrix, and therefore the equation is usually solved by a recursive algorithm known as Durbin's algorithm.

4.5 Cepstrum Analysis

Cepstrum is the result of taking the inverse Fourier transform (IFT) of the logarithm of the estimated spectrum of a signal. The power cepstrum in particular finds applications in the analysis of human speech as shown Fig 3.

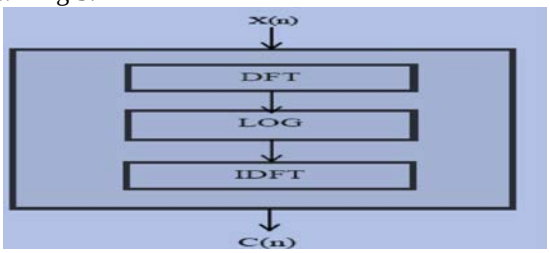


Fig.3. Block diagram of cepstral analysis

ISSN 2348 – 7968

Operations on cepstra are based on quefrency analysis, lightering, or cepstral analysis $c(n) = \sum_{i=0}^{n-1} log\,[\sum_{i=0}^{n-1} x(n)\, e^{-i\omega k}]$ The complex cepstrum is defined as the inverse Fourier transform of the logarithm of the phase unwrapped spectrum.

$$c_r[\tau] = DFT^{-1}\{\log(|S[k]|)\}$$

The real cepstrum is simply obtained by discarding the phase information.

$$c_c[\tau] = DFT^{-1}\{\log(S[k]) + i2\pi m\}$$

Both complex and real cepstrum are real valued, as s[n] is real, and S[k] is complex conjugate symmetric. The main

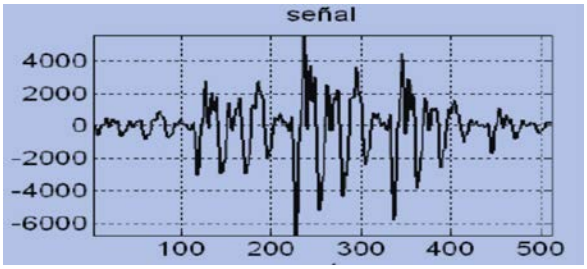
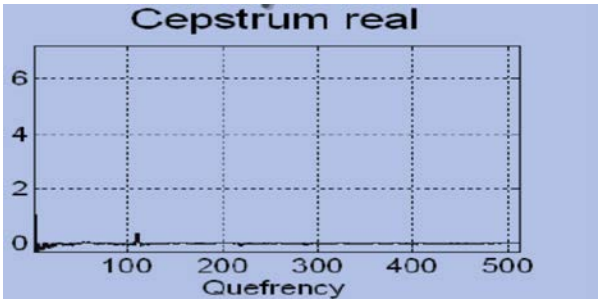


Fig.4. Complex cepstrum

drawback with the real cepstrum is that the phase information is discarded and it is not possible to reconstruct the original spectrum, while it is possible to do this from the complex cepstrum by using an inverse DFT.



F.g.5. Real cepstrum

4.6 Energy Representation

Energy or Intensity is sound energy transmitted per second (power) through a unit area in a sound field. Intensity is proportional to the square of the pressure variation.

Normalized Energy = $\dfrac{\sum_{n=t}^{t+N-1} x_n^2}{N}$

x_n = signal x at time sample number n, N time samples.

4.7 Zero Crossing Rate

The zero-crossing rate is the rate of sign-changes along a signal, i.e., the rate at which the signal changes from positive to negative or back. This feature has been used heavily in both speech recognition and music information retrieval, being a key feature to classify percussive sounds

ZCR is defined formally as

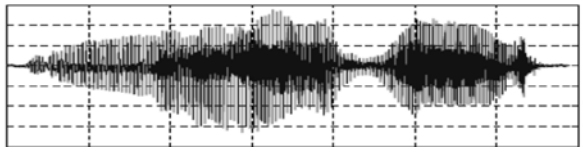
$$zcr = \frac{1}{T-1}\sum_{t=1}^{T-1} \mathbb{I}\{s_t s_{t-1} < 0\}$$

where S is a signal of length T and the indicator function $\mathbb{I}\{A\}$ is 1 if its argument A is true and 0 otherwise.In some cases only the "positive-going" or "negative-going" crossings are counted, rather than all the crossings - since, logically, between a pair of adjacent positive zero-crossings there must be one and only one negative zero-crossing**. Fig.8. short time zero crossing rate**

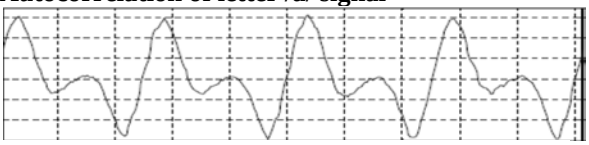
4.8 Short Term Energy

Short Term Processing of speech can be performed either in time domain or in frequency domain. The particular domain of processing depends on the information from the speech that we are interested in. For instance, parameters like short term energy, short term zero crossing rate and short term autocorrelation can be computed from the time domain processing of speech. Alternatively, short term Fourier transform can be computed from the frequency domain processing of speech. The short time energy representation is shown in figure 6.

Short term spectrum of letter /a/



Autocorrelation of letter /a/ signal



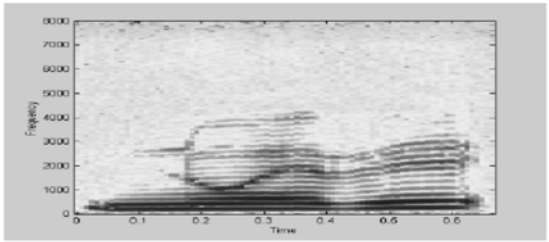


Fig.6 short term energy spectrum**.**

ISSN 2348 – 7968

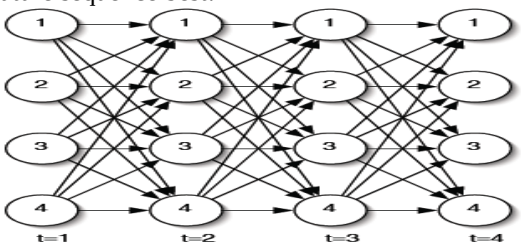
Each of these parameters gives different information about speech that can be used for automatic processing. The energy associated with speech is time varying in nature. Hence the interest for any automatic processing of speech is to know how the energy is varying with time and to be more specific, energy associated with short term region of speech. By the nature of production, the speech signal consist of voiced, unvoiced and silence regions. Further the energy associated with voiced region is large compared to unvoiced region.

4.9 Training & Testing Methods

In this speech emotion recognition system after the calculation of the features, the best features are provided to the classifier. A classifier recognizes the emotion in the speaker's speech utterance. Various types of classifier have been proposed for the task of speech emotion recognition. Gaussian Mixtures Model (GMM), K-nearest neighbours (KNN), Hidden Markov Model (HMM) and Support Vector Machine (SVM), Artificial Neural Network (ANN), etc. are the classifiers used in the speech emotion recognition system. Each classifier has some advantages and limitations over the others. But Hidden Markov Model (HMM, MLP) with multilayer perceptron has high level advantages for emotion recognition compare to others.

4.10 Hidden Markov Model

A simple model that can be used to approximate non-stationary, non-memoryless sources is the hidden Markov model as shown fig 7..In speech recognition system like isolated word recognition and speech emotion recognition, hidden markov model is generally used. The main reason is its physical relation with the speech signals production mechanism. In speech emotion recognition system, HMM has achieved great success for modelling temporal information in the speech spectrum. The HMM is doubly stochastic process consist of first order markov chain whose states are buried from the observer. For speech emotion recognition typically a single HMM is trained for each emotion and an unknown sample is classified according to the model which illustrates the derived feature sequence best.



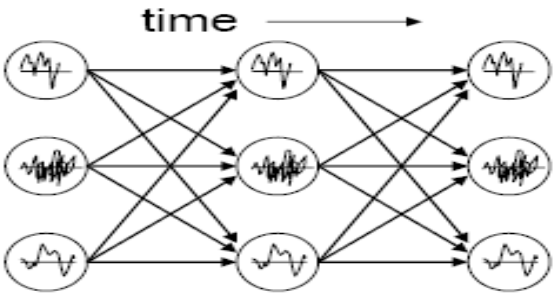


Fig.7 .Diagram of Hidden Markov Model

HMM has the important advantage that the temporal and spectral dynamics of speech features can be stable &second accessibility of the well-established procedure for optimizing the recognition frame work. In this project HMM with MLP performs both male and female emotion speech recognition. Each state represents a process of measurable observations such as pitch, intensity, sample period, range. Inter-process transition is governed by a finite state Markov chain Processes are stochastic and individual observations do not immediately identify the state.

HMM probability of an observed sequence

The task is to compute, given the parameters of the model, the probability of a particular output sequence. This requires summation over all possible state sequences:

The probability of observing a sequence $Y = y(0), y(1), \ldots, y(L-1)$ of length L is given by

$$P(Y) = \sum_X P(Y \mid X) P(X),$$

where the sum runs over all possible hidden-node sequences

$$X = x(0), x(1), \ldots, x(L-1).$$

Applying the principle of dynamic programming, this problem, too, can be handled efficiently using the forward algorithm.

HMM Probability of the latent variables

A number of related tasks ask about the probability of one or more of the latent variables, given the model's parameters and a sequence of observations $y(1), \ldots, y(t)$.

Filtering

The task is to compute, given the model's parameters and a sequence of observations, the distribution over hidden states of the last latent variable at the end of the sequence,

ISSN 2348 – 7968

i.e. to compute $P(x(t) \mid y(1), \ldots, y(t))$. This task is normally used when the sequence of latent variables is thought of as the underlying states that a process moves through at a sequence of points of time, with corresponding observations at each point in time. Then, it is natural to ask about the state of the process at the end. This problem can be handled efficiently using the forward algorithm.

Smoothing

This is similar to filtering but asks about the distribution of a latent variable somewhere in the middle of a sequence, i.e. to compute $P(x(k) \mid y(1), \ldots, y(t))$ for some $k < t$. From the perspective described above, this can be thought of as the probability distribution over hidden states for a point in time k in the past, relative to time t.

Forward-Backward Algorithm

It is an efficient method for computing the smoothed values for all hidden state variables. The task, unlike the previous two, asks about the joint probability of the entire sequence of hidden states that generated a particular sequence of observations. This task is generally applicable when HMM's are applied to different sorts of problems from those for which the tasks of filtering and smoothing are applicable. An example is part-of-speech tagging, where the hidden states represent the underlying parts of speech corresponding to an observed sequence of words. In this case, what is of interest is the entire sequence of parts of speech, rather than simply the part of speech for a single word, as filtering or smoothing would compute. This task requires finding a maximum over all possible state sequences, and can be solved efficiently by the Viterbi algorithm.

4.11. Multilayer Perceptron

MLP is a class of artificial neural network and it consists of a set of process units (simple perceptron's) arranged in layers. In the MLP, the nodes are fully connected between layers without connections between units in the same layer. The input vector (feature vector) feeds into each of the first layer perceptron's, the outputs of this layer feed into each of the second layer perceptron's, and so on. The output of the neuron is the weighted sum of the inputs plus the bias term, and its activation is a function (linear or nonlinear) as In order to apply HMM and MLP, Tests increasing the number of Gaussian components in the mixtures were performed to find the optimal structure.

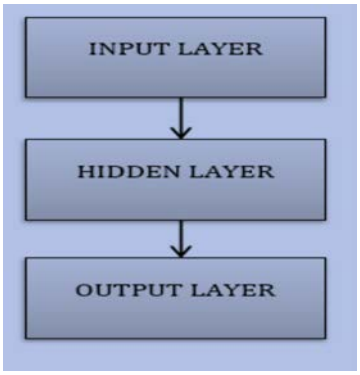


Fig.8.Diagram of Multilayer Perceptron

In order to optimize the MLP performance, different numbers of inputs and neurons in the hidden layer were tested. The estimation of recognition rate can be biased if only one training and one test partition is used. To avoid these estimation biases, a cross-validation with the leave-k-out method was performed. After the design phase, ten data partitions were generated for the test phase. In MLP experiments, 60% of data was randomly selected for training, 20% was used for the generalization test and the remaining 20% was left for validation3. The MLP training was stopped when the network reached the generalization peak with test data. In HMM cases, the 20% used for tests were added to the standard train set. In MLP experiments each partition has 196 utterances for training, 63 utterances for generalization and 63 utterances for validation.

$$\frac{1}{2}\sum_{k=1}^{p}(d_k(t) - O_k(t))^2$$

5. Results

The energy content reviews the intensity or pitch of audio wave and unit is represented by decibels. Audio signals are analysed by 1.LPC co-efficient, 2.LPC spectrum estimation,3.cepstrum, 4. Auto-correlation,5.Zero crossing rate, 6.Short-term average mean value, 7.Absolute short-time energy, 8.Average sub-frame rate.

Speech-Emotion recognition such as Anger, Joy, and Sad which is recognised by HMM with MLPs .The iterative process is used to compare the classifieds and several emotion states on acoustic features such as iteration points, Sample time and Sample periods.

ISSN 2348 – 7968

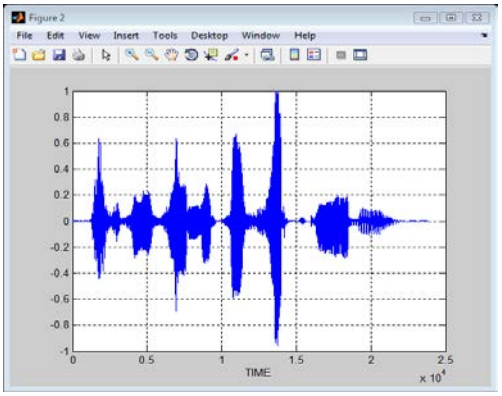


Figure.9. Energy content of audio wave

Figure.10 shows the iterative process of audio wave ,the iterative points are estimating the pitch value, it is used for calculating the amplitude with respect to sample time.

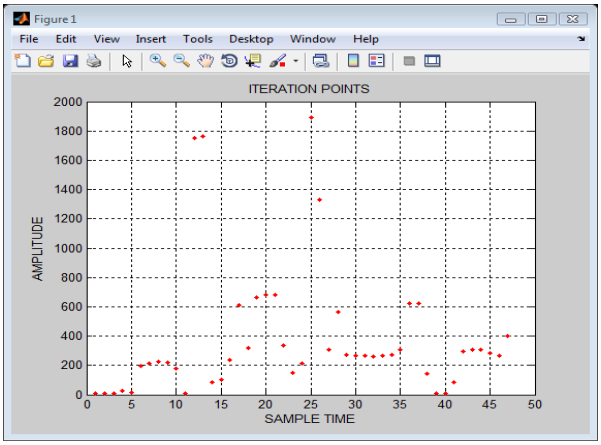


Fig.10. Iteration points of audio wave

Figure.11 shows the effects of several emotion states on selected acoustic features such as Anger, Sad, & Joy.

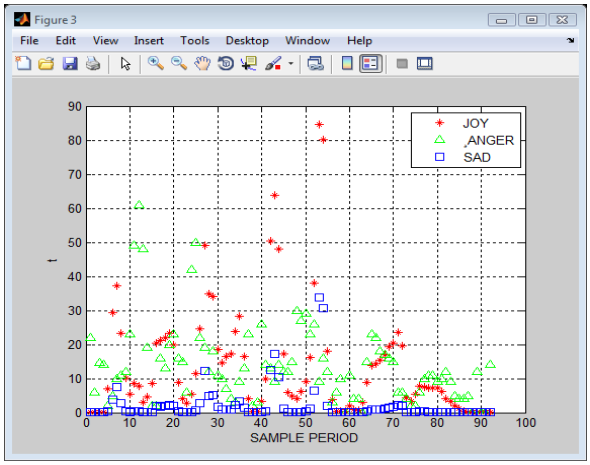


Fig.11.Comparison of Emotional states

Parameter	Speaker 1			Speaker 2			Speaker 3		
	IP	ST	SP	IP	ST	SP	IP	ST	SP
Anger	**0 to 450**	**0 to 70**	**0 to 140**	**0 to 2000**	**0 to 60**	**0 to 95**	**0 to 1200**	**0 to 70**	**0 to 120**
Joy	**0 to 2000**	**0 to 90**	**0 to 140**	**0 to 1200**	**0 to 90**	**0 to 90**	**0 to 1400**	**0 to 90**	**0 to 120**
Sad	**0 to 200**	**0 to 30**	**0 to 140**	**0 to 1800**	**0 to 1800**	**0 to 95**	**0 to 1800**	**0 to 40**	**0 to 2000**

Table 1 Summary of the effects of several emotion states on selected acoustic features

The comparison states are used for calculating the Sample time and Sample periods which is used for recognition.This recognition is based on robust and reliable performance of speech processing. Each of these parameters gives different information about speech that can be used for automatic speech processing.

Where,
IP-Iteration points, ST-Sample time, SP-Sample period ,> high, < low, = equal, M –male, F- female

EMOTION	PITCH			INTENSITY			TIMING
	Mean	Range	variance	Mean	Range	Variance	Sampling Rate
Anger	>>	>	>>	>> **M,** **>F**	>	>>	**<M,> F**
Joy	>	**>M,** **<F**	>	>	>	<	**>M,** **<F**
Sad	<	<	<	<	<	>	=

ISSN 2348 – 7968

5. Conclusion &Future Enhancements

The aim of work to developed a speech analysis process for speech (Emotion recognition) recognition. The work carried in deep and wide analysis of speech signal. It carried both temporal analysis and spectral analysis effectively and Emotion recognition such as joy, angry, sad by using testing & training methods (HMM, MLP). This recognition is based on robust and reliable performance of speech processing. Speech emotion recognition systems based on the several classifiers is illustrated. The important issues in speech emotion recognition system are the signal processing unit in which appropriate features are extracted from available speech signal and another is a classifier which recognizes emotions from the speech signal. The average accuracy of the most of the classifiers for speaker independent system is less than that for the speaker dependent.

This proposal ultimately concludes that while the deep and wide analysis of speech processing structures can provide improvements for this genre, features and the structures with which they are combined with various speech –Emotion recognition, detection, phonetics tracking can also significant factors.

Automatic emotion recognitions from the human speech are increasing now a day because it results in the better interactions between human and machine. To improve the emotion recognition process, combinations of the given methods can be derived. Also by extracting more effective features of speech, accuracy of the speech emotion recognition system can be enhanced. Despite the research carried out into emotional speech recognition, there is little agreement on how toper forms feature selection, and so this project will offer an indication of an approach to do so.

References

[1] J. P. Pinto, “Multilayer perceptron based hierarchical acoustic modelingfor automatic speech recognition,” Ph.D. dissertation, EPFL, Lausanne,Switzerland, 2010.
[2] S. Zhao, S. Ravuri, and N. Morgan, “Multi-stream to many-stream:Usingspectro-temporal features for ASR,” in Proc. Interspeech,Brighton, UK, 2009, pp. 2951–2954..
[3] L. Chase, “Error-responsive feedback mechanisms for speech recognizers,”Ph.D. dissertation, Robotics Inst., Carnegie Mellon Univ.,Pittsburgh, PA, 1997.
[4]N. Amir, "Classifying emotions in speech: a comparisonof methods", Eurospeech 2001, Poster Proceedings, Scandinavia, 2001, pp. 127-130.
[5] A. de Cheveigné, H. Kawahara, “Comparative evaluationof F0 estimation algorithms”, Eurospeech 2001, Arlborg, Denmark, 2001.
[6] S. S. Viglione, “Applications of pattern recognition technology,” inAdaptive Learning and Pattern Recognition, J. M. Mendel and K. S.Fu, Eds. New York: Academic, 1970, pp. 115–161.
[7] D. Ellis and N. Morgan, “Size matters: An empirical study of neuralnetwork training for large vocabulary continuous speech recognition,”inProc. ICASPP, 1999, pp. 1013–1016.
[8] Y.-H. Chiu, B. Raj, and R. Stern, “Learning based auditory encodingfor robust speech recognition,” in Proc. ICASSP, 2010, pp. 428–4281.
[9]S. Arunachalam, D. Gould, E. Andersen, D. Byrd, and S.Narayanan, "Politeness and Frustration Language in Child-Machine Interactions", Paper Proc. Eurospeech2001, Proceedings, Scandinavia, 2001, pp. 2675.
[10] B. Heuft, T. Portele, and M. Rauth, “Emotions in timedomain synthesis”, in Proc. of ICSLP, Philadelphia, Oct. 1996, pp. 1974–1977.
[11] C. J. Leggetter and P. C. Woodland, “Maximum likelihood linear regressionfor speaker adaptation of continuous density HMMs,” SpeechCommun., vol. 9, pp. 171–186, 1995.
[12] A. Stolcke, B. Chen, H. Franco, V. R. R. Gadde, M. Graciarena, M.-Y.Hwang, K. Kirchhoff, A. Mandal, N. Morgan, X. Lei, T. Ng, M. Ostendorf,K. Sonmez, A. Venkataraman, D. Vergyri, W.Wang, J. Zheng,and Q. Zhu, “Recent innovations in speech to-Text transcription at SRIICSI-UW,” IEEE Trans. Audio, Speech, Lang. Process., vol. 14, no. 5,pp. 1729–1744, Sep. 2006.

First Author

T. Jayasankar. Assistant Professor Department Of Electronics and Communication Engineering, , Anna University BIT Campus Trichy. His research interests in speech processing and Text to Speech synthesis. Email: jayasankar_t@rediffmail.com, t.jayasankar@mail.aubit.edu.in.

Second Author

J.Jayalakshmi, received the B.E. degree in Electronics and Communication Engineering from Anna University, Trichy in 2011 and M.E. degree at Anna University, Chennai in Email. san.jayalakshmi@gmail.com

Third Author

K.Rajasekaran Assistant Professor Department Of Electronics and Communication Engineering, M.A.R College of Engineering and Technology, His research interests in speech recognition and Wireless Networks
Email:mvkraja@yahoo.com.