I.J. Computer Network and Information Security, **2014, 7, 61-77**
Published Online June 2014 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijcnis.2014.07.08

Speaker Emotion Recognition Based on Speech Features and Classification Techniques

J. Sirisha Devi
Assistant Professor, Dept of IT, GRIET, Hyderabad, 501401, Andhra Pradesh, India
E-mail: siri.cse21@gmail.com

Dr. Srinivas Yarramalle
Professor, Dept of IT, GITAM University, Visakhapatnam, 530045, Andhra Pradesh, India
E-mail: sriteja.y@gmail.com

Siva Prasad Nandyala
Research Scholar, Dept of ECE, NIT Warangal, Warangal, 506004, Andhra Pradesh, India
E-mail: nsprasad@nitw.ac.in

Abstract—Speech Processing has been developed as one of the vital provision region of Digital Signal Processing. Speaker recognition is the methodology of immediately distinguishing who is talking dependent upon special aspects held in discourse waves. This strategy makes it conceivable to utilize the speaker's voice to check their character and control access to administrations, for example voice dialing, data administrations, voice send, and security control for secret information.

A review on speaker recognition and emotion recognition is performed based on past ten years of research work. So far iari is done on text independent and dependent speaker recognition. There are many prosodic features of speech signal that depict the emotion of a speaker. A detailed study on these issues is presented in this paper.

Index Terms—Emotion recognition, feature extraction, speaker recognition.

I. Introduction

Speaker recognition is the figuring undertaking of approving a client's asserted personality utilizing qualities removed from their voices. There is a contrast between speaker recognition (distinguishing who is talking) and discourse recognition (distinguishing what is, no doubt said). These two terms are as often as possible confounded, as is voice recognition. Voice recognition is mix of the two where it uses studied parts of a speakers' voice to confirm what is, no doubt said - such a framework can't distinguish discourse from arbitrary speakers quite correctly, however it can arrive at high correctness for unique voices with which it has been prepared. Furthermore, there is a contrast between the demonstration of confirmation (regularly alluded to as speaker check or speaker verification) and ID. At long last, there is a contrast between speaker recognition and speaker diarization (distinguishing when the same speaker is talking).

Speaker recognition has a history going back around four decades and utilizes the acoustic characteristics of discourse that have been discovered to contrast between people. These acoustic examples reflect both life systems (e.g., estimate and state of the throat and mouth) and studied behavioral examples (e.g., voice pitch, talking style).

Speaker recognition alludes to distinguish the individual from their discourse. Exactness of recognition expands if the framework is content free [1]. The discourse indicator holds the message being spoken, the gushing state of the speaker and the data of the speaker. Thusly we can utilize the discourse motion for both discourse and speaker recognition [2]. Speaker recognition extricates the underlying phonetic message in an utterance. It is the procedure of immediately distinguishing who is talking on the support of characteristics present in the discourse indicator. Impact of passionate state of human discourse in speaker recognition is exceptionally high. The expression "feeling" can allude to a greatly perplexing state connected with a wide mixture of mental, physiological and physical events [7] [8].

The spectral feature Mel frequency cepstral coefficients best suited for speaker recognition and the prosodic feature pitch, which is strongly dependent on the emotional state of the speaker [14] [62].

In biometric recognition systems, there are 2 main factors that have created voice a compelling biometric. The telephone system provides an omnipresent, acquainted network of sensors for getting and delivering the speech signal. Speech could be a natural signal to supply that's not thought of threatening by users to produce.

Text dependent speaker recognition is technique where a fixed word, or test –constraint speech is used for recognition. The system has prior knowledge of the text to be spoken. Used for applications with strong control over user. Knowledge of spoken text can improve system performance. Text independent speaker recognition

system do not rely on specific text being spoken. This technique is additional versatile in comparison to previous technique. Used for applications with less or no management over user. Speech recognition will offer data of spoken text. For text-independent speaker recognition, the foremost flourishing chance operate has been GMMs. In text-dependent applications, extra temporal data may be incorporated by victimization Hidden Markov Models (HMMs) for the chance functions.

In section 2, framework for speaker and emotion recognition is given and feature extraction is done, which will be given as input to the classifiers. In section 3, literature review of speaker and emotion recognition [59] is done. In section 4, classification techniques along with their advantages and disadvantages are studied.

II. Theoretical background for speaker and emotion recognition

The speech signal conveys many levels of data to the listener. At the prime grade, speech expresses a note via words. But at other grades talk expresses data about the dialect being voiced and the strong feeling, gender and, usually, the identity of the speaker[10] [30]. While talk acknowledgement aspires at identifying the phrase voiced in talk, the goal of automatic speaker recognition [68] schemes is to extract distinguish and recognize the information in the talk signal conveying speaker persona. Speaker acknowledgement is helpful in the numerous areas [11] as shown in table I.

Speaker specific features will be present in the voiced part of the speech signals. Pre-processing is a process of separating the voiced and unvoiced speech signals [9].

The speech signal is a slowly timed varying signal (it is called quasi-stationary). When examined over a sufficiently short timeframe (between five & 100 msec), its characteristics are stationary. However, over long periods of time (on the order of 1/5 seconds or more) the signal characteristic change to reflect the different speech sounds being spoken.

A. *Noise removal*

While recording the speech signal, it consists of various unwanted piece of signals which are considered to be noise/unvoiced part of speech. This noise part may be due to low quality of source, loss of speech segments, channel fading, and presence of noise in the channel or echo or reverberation.

A low-pass filter allows signal frequencies below the low cut-off frequency to pass and stops frequencies above the cut-off frequency. If speaker specific information is available in the higher frequencies, a high pass filter may be used to remove all the frequencies less than 4 KHz as unvoiced data.

B. *Framing*

Spectral evaluation are reliable in the case of a stationary signal (i.e., a proof whose statistical characteristics area unit invariant with relation to time). Imply that the region is brief enough for the behaviour of (periodicity or noise-like appearance) the signal to be close to constant. In sense, the speech region should be short enough so it will reasonably be assumed to be stationary. Stationary therein region: i.e., the signal characteristics (whether periodicity or noise-like appearance) area unit uniform therein region. Frame period ranges area unit between ten ~ twenty five ms within the case of speech processing.

Each speech signal is split into frame segments of size 30ms (milli second) where the sampling rate is 22.05 KHz. Each segment is extracted at every 50ms interval. This implies that the overlap between segments is 10ms [6].

C. *Windowing*

The aim of characteristic extraction [61] is to supply spectral characteristics that can help us to build phone or sub-phone classifiers. We thus don't desire to extract spectral features from an entire utterance or dialogue, because the spectrum changes very rapidly. Mechanically, we state that talk is a non-stationary signal, significance that its statistical properties are not constant over time. rather than, we want to extract spectral features from a small window of talk that distinguishes a particular sub-phone and for which we can make the (rough) assumption that the signal is stationary (i.e. its statistical properties are unchanging inside this region). We'll do this by utilizing a window which is non-zero interior some district and zero elsewhere, running this window over the talk signal, and extracting the waveform interior this window.

We can distinguish such a windowing procedure by parameters: how wide are the window (in milliseconds), what the offset between successive windows is, & what the shape of the window [18] is. They call the talk extracted from each window a frame, & they call the number of milliseconds in the border the border dimensions & the number of milli-seconds between the left perimeters of successive windows the frame move. The extraction of the pointer takes place by multiplying the worth of the pointer at time n, s[n], with the worth of the window at time n, w[n]:

$$y[n] = w[n]s[n] \quad (1)$$

Fig. 2-1 suggests that these window forms are rectangular, since the extracted windowed pointer examines just like the initial pointer. Really the simplest window is the rectangular window. The rectangular windows can origin troubles, although, be origin it suddenly slashes of the signal at its boundaries. These discontinuities create troubles when we do Fourier analysis. For this cause, a more widespread window used in characteristic extraction is the Hamming window, which shrinks the standards of the pointer in the direction of none at the window boundaries, bypassing discontinuities.

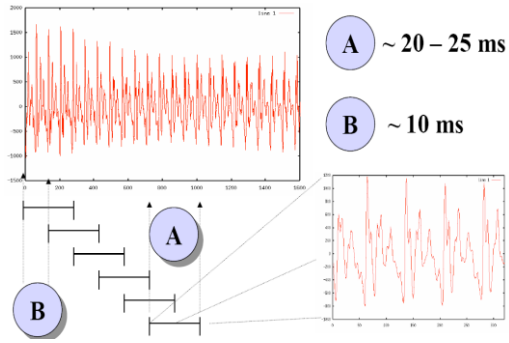

Fig 2.1 Windowing process, showing the frame shift and frame size, assuming a frame shift of 10ms, a frame size of 25 ms, and a rectangular window

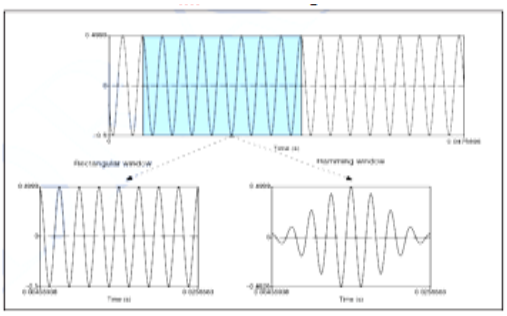

Fig 2.2 Windowing a portion of a pure sine wave with the rectangular and Hamming Windows

Fig. 2.2 shows both of these windows; the equations are as follows (assuming a window that is L frames long):

Rectangular window

$$w[n] = 1; 0 \le n \le L-1 \quad 0\ otherwise \tag{2}$$

hamming window

$$w[n] = 0.54 - 0.46\cos\left(\frac{2\pi n}{L}\right); 0 \le n \le L-1 \quad 0; otherwise \tag{3}$$

D. *Fast Fourier Transform (FFT)*

The next step is to extract spectral information for our windowed signal; we need to know how much energy the signal contains at different frequency bands. The tool for extracting spectral information for discrete frequency bands for a discrete-time (sampled) signal is the Discrete Fourier Transform or DFT.

The input to the DFT is a windowed signal x[n]...x[m], and the output, for each of N discrete frequency bands, is a complex number X[k] representing the magnitude and phase of that frequency component in the original signal. If we plot the magnitude against the frequency, we can visualize the spectrum. For example, fig. 2-3 shows a 25 ms Hamming-windowed portion of a signal and its spectrum as computed by a DFT (with some additional smoothing).

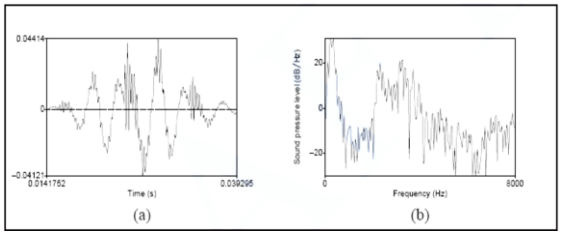

Fig 2.3 A 25ms Hamming-windowed portion of a signal from the vowel [iy] and (b) its spectrum computed by a DFT

We will not introduce the mathematical details of the DFT here, except to note that Fourier analysis in general relies on Euler's formula:

$$e^{ix} = \cos(x) + j\sin(x) \tag{4}$$

As a brief reminder, the DFT is defined as follows:

$$\mathbf{X[k] = \sum_{n=0}^{N-1} x[n] e^{-j2\frac{\Pi}{N}kn}} \tag{5}$$

A routinely used algorithm for computing the DFT is the Fast Fourier Transform or FFT [6]. This implementation of DFT is very efficient, but only works for standards of N which are powers of two. The outcomes can also be examined in 2D plots called "Spectrograms". The following are the Spectrograms for diverse phonemes or rudimentary speech signals, each about 1sec long duration, and fig. 2.4.

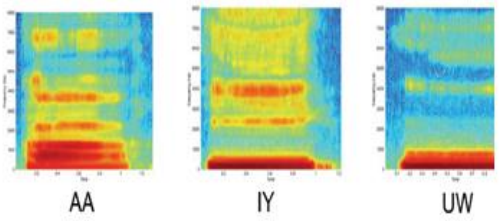

Fig 2.4 Spectrograms for different phonemes

Spectrograms are 2-D time-frequency plots.

- The x-axis is time; *i.e.* the series of analysis windows
- The y-axis is frequency, 0 to 8KHz for a 16KHz sampled signal
- Each point in the 2-D matrix indicates the intensity of the signal in that frequency band, during that time window
- Red is highest intensity, blue is lowest
- The phonemes[9] have a distinct set of dominant frequency bands called formants
- Not all phonemes [11] have distinct formants

The next processing step is the Fast Fourier Transform, which converts each frame of N samples from the time domain into the frequency domain. FFT gets log magnitude spectrum to determine MFCC. We have used 1024 point to get better frequency resolution.

E. The Cepstrum: Inverse Discrete Fourier Transform

While it would be possible to use the mel spectrum by itself as a feature representation for phone detection, the spectrum also has some problems, as we will see. For this reason, the next step in MFCC feature extraction [15] is the computation of the cepstrum. The cepstrum has a number of useful processing advantages and also significantly improves phone recognition performance.

One way to think about the cepstrum is as a useful way of separating the source and filter. The speech waveform is created when a glottal source waveform of a particular fundamental frequency [60] is passed through the vocal tract, which because of its shape has a particular filtering characteristic. But many characteristics of the glottal source (its fundamental frequency, the details of the glottal pulse, etc) are not important for distinguishing different phones. Instead, the most useful information for phone detection is the filter, i.e. the exact position of the vocal tract. If we knew the shape of the vocal tract, we would know which phone was being produced. This suggests that useful features for phone detection would find a way to de-convolve (separate) the source and filter and show us only the vocal tract filter. It turns out that the cepstrum is one way to do this.

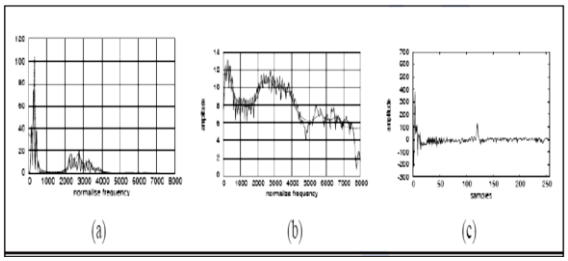

Fig 2.5 magnitude spectrum (a), the log magnitude spectrum (b), and the cepstrum (c). The two spectra have a smoothed spectral enveloped laid on top of them to help visualize the spectrum.

For simplicity, let's ignore the pre-emphasis and mel-warping that are part of the definition of MFCCs, and look just at the basic definition of the cepstrum. The cepstrum can be thought of as the spectrum of the log of the spectrum. This may sound confusing. But let's begin with the easy part: the log of the spectrum. That is, the cepstrum begins with a standard magnitude spectrum, such as the one for a vowel shown in Fig. 2-5(a) from Taylor (2008). We then take the log, i.e. replace each amplitude value in the magnitude spectrum with its log, as shown in Figure 2-5(b).

The next step is to visualize the log spectrum as if it were a waveform. In other words, consider the log spectrum in Fig. 2-5(b). Let's imagine removing the axis labels that tell us that this is a spectrum (frequency on the x-axis) and imagine that we are dealing with just a normal speech signal with time on the x-axis. Now what can we say about the spectrum of this 'pseudo-signal'? Notice that there is a high-frequency repetitive component in this wave: small waves that repeat about 8 times in each 1000 along the x-axis, for a frequency of about 120 Hz. This high-frequency component is caused by the fundamental frequency of the signal, and represents the little peaks in the spectrum at each harmonic of the signal. In addition, there are some lower frequency components in this 'pseudo-signal'; for example the envelope or formant structure has about four large peaks in the window, for a much lower frequency.

Figure 2-5(c) shows the cepstrum: the spectrum that we have been describing of the log spectrum. This cepstrum (the word cepstrum is formed by reversing the first letters of spectrum) is shown with samples along the x-axis. This is because by taking the spectrum of the log spectrum, we have left the frequency domain of the spectrum, and gone back to the time domain. It turns out that the correct unit of a cepstrum is the sample.

Examining this cepstrum, we see that there is indeed a large peak around 120, corresponding to the F0 and representing the glottal pulse. There are other various components at lower values on the x-axis. These represent the vocal tract filter (the position of the tongue and the other articulators). Thus if we are interested in detecting phones, we can make use of just the lower cepstral values. If we are interested in detecting pitch, we can use the higher cepstral values.

For the purposes of MFCC extraction, we generally just take the first 12 cepstral values. These 12 coefficients will represent information solely about the vocal tract filter, cleanly separated from information about the glottal source.

It turns out that cepstral coefficients have the extremely useful property that the variance of the different coefficients tends to be uncorrelated. This is not true for the spectrum, where spectral coefficients at different frequency bands are correlated. The fact that cepstral features are uncorrelated means, that the Gaussian acoustic model (the Gaussian Mixture Model, or GMM) doesn't have to represent the covariance between all the MFCC features, which hugely reduces the number of parameters.

The cepstrum is more formally defined as the inverse DFT of the log magnitude of the DFT of a signal, hence for a windowed frame of speech x[n]:

$$\boldsymbol{c[n]} = \sum_{\boldsymbol{n}=0}^{\boldsymbol{N}-1} \log\left(\left| \sum_{\boldsymbol{n}=0}^{\boldsymbol{N}-1} \boldsymbol{x[n]} \boldsymbol{e}^{-\boldsymbol{j}\frac{2\pi}{\boldsymbol{N}}\boldsymbol{kn}} \right| \right) \boldsymbol{e}^{\boldsymbol{j}\frac{2\pi}{\boldsymbol{N}}\boldsymbol{kn}} \quad (6)$$

F. Deltas and Energy

The extraction of the cepstrum via the Inverse DFT from the previous section results in 12 cepstral coefficients for each frame. We next add a thirteenth feature: the energy from the frame. Energy correlates with phone identity and so is a useful cue for phone detection (vowels and sibilants have more energy than stops, etc). The energy in a frame is the sum over time of the power of the samples in the frame; thus for a signal x in a window from time sample t1 to time sample t2, the energy is:

$$\text{Energy} = \sum_{t=t_1}^{t_2} x^2[t] \qquad (7)$$

Extracted features zero crossing rate, energy entropy, short time energy for angry speech signal are shown in the fig. 2-6, 2.7 and 2.8.

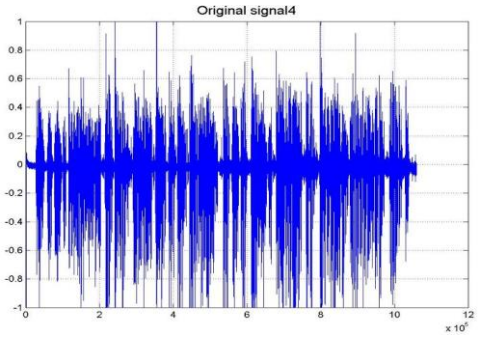

Fig 2.6 Zero Crossing Rate

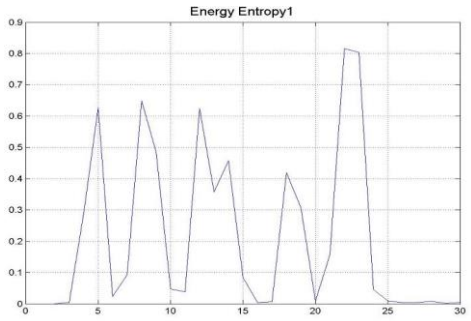

Fig 2.7 Energy Entropy

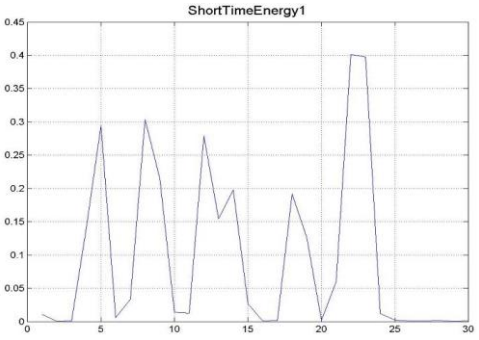

Fig 2.8 Short Time Energy

Another important fact about the speech signal is that it is not constant from frame to frame. This change, such as the slope of a formant at its transitions, or the nature of the change from a stop closure to stop burst, can provide a useful cue for phone identity. For this reason we also add features related to the change in cepstral features over time.

We do this by adding for each of the 13 features (12 cepstral features plus energy) a delta or velocity feature, and a double delta or acceleration feature. Each of the 13 delta features represents the change between frames in the corresponding cepstral energy feature, while each of the 13 double delta features represents the change between frames in the corresponding delta features.

A simple way to compute deltas would be just to compute the difference between frames; thus the delta value d(t) for a particular cepstral value c(t) at time t can be estimated as:

$$d(t) = \frac{c(t+1) - c(t-1)}{2} \qquad (8)$$

Instead of this simple estimate, however, it is more common to make more sophisticated estimates of the slope, using a wider context of frames.

G. *Summary: MFCC*

After adding energy, and then delta and double-delta features to the 12 cepstral features, we end up with 39 MFCC features as shown in fig 2-9:

12 cepstral coefficients
12 delta cepstral coefficients
12 double delta cepstral coefficients
1 energy coefficient
1 delta energy coefficient
1 double delta energy coefficient

Total: 39 MFCC features

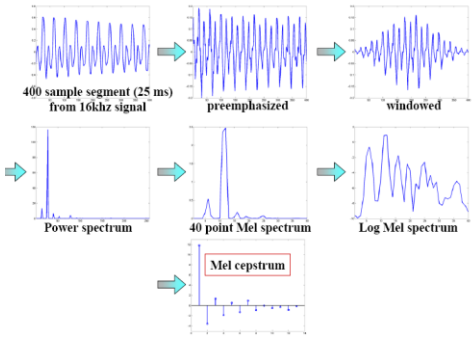

Fig 2.9 Extraction of MFCC for a Frame

H. *Feature Extraction and Matching*

Generally the feature extraction techniques are classified as temporal analysis and spectral analysis technique. In temporal analysis the speech wave shape itself is employed for analysis. In spectral analysis the spectral illustration of speech signal is employed for analysis.

Khalid Saeed and Mohammad Kheir Nammous in 'A Speech-and-Speaker Identification System: Feature Extraction, Description, and Classification of Speech-Signal Image [12] discussed a speech-and-speaker (SAS) identification system based on spoken Arabic digit recognition. The speech signals of the Arabic digits from zero to ten are processed graphically (the signal is treated as an object image for further processing). At the stage of classification, both conventional and neural-network-based methods are used. The techniques used were FFT, Predictive Coding and PCM. The success rate of the speaker-identifying system obtained in the presented experiments for individually uttered words is excellent

and has reached about 98.8% in some cases. The average overall success rate was then 97.45% in recognizing one uttered word and identifying its speaker, and 92.5% in recognizing a three-digit password [12]. Speaker Recognition Using MFCC Front End Analysis and VQ Modeling Technique for Hindi Words using MATLAB" [13] was proposed by Nitisha and Ashu Bansal. This paper introduced text dependent systems that have been trained for a particular user.

The difficulty of speaker acknowledgement pertains to a much broader theme in technical and technology so called pattern acknowledgement. The goal of pattern recognition is to classify things of interest into one of a number of classes or categories. The objects of interest are generically called patterns and in our case are sequences of acoustic vectors that are extracted from an input speech using the methods recounted in the previous part. The classes here mention to individual speakers. Since the classification method in our case is directed on extracted features, it can be also mentioned to as feature matching.

At the most elevated amount, all speaker recognition frameworks hold two principle modules: characteristic extraction and characteristic matching. Feature extraction is the procedure that concentrates a minor measure of information from the voice indicates that can later be utilized to speak to every speaker. Characteristic matching includes the real strategy to recognize the obscure speaker by looking at concentrated characteristics from his/her voice include with the ones from a set of known speakers.

All speaker recognition frameworks need to serve two recognized stages. The first is eluded to the enrolment or preparing stage, while the second one is alluded to as the operational or testing stage. In the preparation stage, each one enrolled speaker needs to give specimens of their discourse with the goal that the framework can manufacture or train a reference show for that speaker. if there should be an occurrence of speaker confirmation frameworks, likewise, a speaker-particular limit is additionally figured from the preparation inspects. In the testing stage, the data discourse is matched with archived reference model(s) and a recognition choice is made.

Besides, if there exists some set of examples that the distinctive classes of which are right now known, then one has an issue in administered example affirmation. These examples contain the educating set and are used to draw from a grouping calculation. The leftover examples are then used to test the characterization calculation; these examples are all things considered alluded to as the test set. Provided that the right classifications of the distinct designs in the check set are besides known, then one can assess the presentation of the calculation.

Various methods for feature extraction are broadly shown in the following table II.

IV LITERATURE REVIEW ON SPEAKER AND EMOTION RECOGNITION

Intensive effort has been made to limit the discussion to area unit as that are most relevant to the work in table III and table IV.

A. CLASSIFICATION TECHNIQUES

In machine studying, the machine gets the information by either examining or by experience. There are two separate sorts of studying strategies: directed and unsupervised. In directed studying, preparing information incorporates both information and fancied effects. The development of a legitimate preparing, approval and test set is essential. These routines are generally quick and exact. In unsupervised studying [67] [68], the machine is not furnished with right comes about throughout the preparation. This study procedure could be utilized to group the data information in classes on the support of their measurable lands just. The naming could be done regardless of the possibility that the marks are accessible for a minor number of items illustrative of the fancied classes [58].

A detailed discussion on the advantages and disadvantages of prominent classifiers based on literature survey is shown in table V.

Limitations on classifiers is depicted in the below table

Classifier	Memory Usage	Training Time	Classification Time
Linear / Logistic Regression	Very low	Fast-Medium	Very fast
Unimodal Gaussian	Very low	Fast-Medium	Fast
Backpropagation	Low	Slow	Very fast
Radial Basis Function	Medium	Medium	Medium
K Nearest Neighbor	High	No training required	Slow
Gaussian Mixture	Medium	Medium-Slow	Medium
Nearest Clustering	Medium	Medium	Fast-medium
Binary / Linear Decision Tree	Low	Fast	Very fast
Projection Pursuit	Low	Medium	Fast
Estimate-Maximize Clustering	Medium	Medium	Medium
MARS	Low	Medium	Fast
GMDH	Low	Fast-Medium	Fast
Parzen's Window	High	Fast	Slow
LVQ	Medium	Slow	Medium

V. CONCLUSION

This paper attempts to provide a complete survey of research on speaker recognition and emotion recognition. Survey says that till date we could not achieve 100% accuracy in recognizing either a speaker or his emotion. When the emotional state of speaker differs in the testing phase the recognition rate decreases significantly. The table shows that the accuracy rate of speaker recognition has been considerably affected when the emotional state

of the speaker was not considered. Pitch is not particularly good for the recognition of neutral tones. Survey of this paper gives us a conclusion the accuracy rate of speaker indirectly depends on the accuracy of emotion recognition.

TABLE I: Uses of speaker recognition

Security and defense	• Forensic • Looking for suspect in quantity of audio • Waiting online for suspect
Access Control	• Physical facilities • Computer networks & websites
Transaction Authentication	• Telephone banking • Remote purchases
Speech Data Management	• Voice mail browsing • Search in audio archives
Personalization	• Voice-web/device customization • Intelligent answering machine

TABLE II: Feature extraction methods [15]

]**Technique**	**Property**	**Remarks**
Mel-frequency scale analysis	Static feature extraction method, Spectral analysis	Spectral analysis is done with a fixed resolution along a subjective frequency scale i.e. Mel-frequency scale. Remove noisy and redundant features, and improvement in classification error
Spectral Subtraction	Robust Feature extraction method	Restoration of the power spectrum or the magnitude spectrum of a signal observed in additive noise, through subtraction of an estimate of the average noise spectrum from the noisy signal spectrum.
Principal Component Analysis(PCA)	Non linear feature extraction method, Linear map; fast; eigenvector-based	Traditional, eigenvector based method, also known as karhuneu-Loeve expansion; good for Gaussian data.
Independent Component Analysis (ICA)	Non linear feature extraction method, Linear map, iterative non-Gaussian	Blind course separation, used for de-mixing non- Gaussian distributed sources(features)
Power Estimation	Temporal Analysis, provides basis for distinguishing voiced speech segments from unvoiced speech segments	Less computation compared to spectral analysis but limited to simple speech parameters
Fundamental Frequency Estimation	Temporal Analysis, the frequency at which the vocal cords vibrate during a voiced sound.	Processed on logarithmic scale, rather than a linear scale to match the resolution of human auditory system.
Perceptually Based Linear Predictive Analysis (PLP) [63]	Obtaining auditory spectrum, approximating the auditory spectrum by an all pole model	Allows computational and storage saving for speaker recognition
Linear Predictive coding	Static feature extraction method,10 to 16 lower order co-efficient,	Representation of vocal tract configuration by relatively simple computation compared to cepstral analysis.
Cepstral Analysis	Static feature extraction method, Power spectrum	Used to represent spectral envelope
Wavelet [65]	Better time resolution than Fourier Transform	It replaces the fixed bandwidth of Fourier transform with one proportional to frequency which allow better time resolution at high frequencies than Fourier Transform
Linear Discriminate Analysis(LDA)	Non linear feature extraction method, Supervised linear map; fast; eigenvector-based	Better than PCA for classification;
RASTA filtering	For Noisy speech	Preprocessing with both log spectral and the cepstral domain filtering
Cepstral mean subtraction	Robust Feature extraction	Not suitable for use with live recognition
Dynamic feature extractions i)LPC ii)MFCCs [13]	Acceleration and delta coefficients i.e. II and III order derivatives of normal LPC and mfccs coefficients	Synthesized speech is represented in these parameters
Integrated Phoneme subspace method	Transformation based on PCA+LDA+ICA	Higher Accuracy than the existing methods

TABLE III: REVIEW OF SPEAKER RECOGNITION

S. No.	Authors	Title	Features Considered	Classifier Used	Accuracy Rate	Published
1.	Alejandro Bidondo, Shin-ichi Sato, Ezequiel Kinigsberg, Adrián Saavedra, Andrés Sabater, Agustín Arias, Mariano Arouxet, and Ariel Groisman	Speaker recognition analysis using running autocorrelation function parameters [31]	r-ACF (running Autocorrelation Function) microscopic parameters	Euclidean distances vector's distances	No improvement in accuracy rate	POMA - ICA 2013 Montreal Volume 19, pp. 060036 (June 2013)
2.	Taufiq Hasan, Seyed Omid Sadjadi, Gang Liu, Navid Shokouhi, Hynek Boˇril, John H.L. Hansen	CRSS Systems For 2012 NIST Speaker Recognition Evaluation [32]	Mean Hilbert Envelope Coefficients (MHEC), PMVDR Front-End, Rectangular Filter-Bank Cepstral Coefficients (RFCC), MFCC-QCN-RASTALP	L2-Regularized Linear Regression (L2LR), UBS-SVM Anti-Model (UBS-SVM), Score-Averaged PLDA (PLDA-2)	relative improvements in the order of 50 - 60%	ICASSP 2013
3.	Md. Jahangir Alam, Patrick Kenny, Douglas O'Shaughnessy	Low-variance Multitaper Mel-frequency Cepstral Coefficient Features for Speech and Speaker Recognition Systems [55]	mel-frequency cepstral coefficient (MFCC)	low-variance multitaper spectrum estimation methods	compared with the Hamming window technique, the sinusoidal weighted cepstrum estimator, multi-peak, and Thomson multitaper techniques provide a relative improvement of 20.25, 18.73, and 12.83 %, respectively, in equal error rate	Springer-Verlag, j Cognitive Computation December 2012
4.	M. Afzal Hossan ·Mark A. Gregory	Speaker recognition utilizing distributed DCT-II based Mel frequency cepstral coefficients and fuzzy vector quantization [33]	Discrete Cosine Transform (DCT-II) based Mel Frequency Cepstral Coefficients (MFCC)	Fuzzy vector quantization	FVQ have shown better results rather than GMM	Int J Speech Technol (2013), Springer Science+Business Media, LLC 2012
5.	Taufiq Hasan, John H. L. Hansen	Acoustic Factor Analysis for Robust Speaker Verification [34]	PPCA for acoustic factor analysis	i-vector system	16.52%, 14.47% and 14.09% relative improvement in %EER, DCF (old) and DCF (new) respectively	IEEE Transactions On Audio, Speech, And Language Processing, Vol. 21, No. 4, April 2013
6.	David A. van Leeuwen and Rahim Saeidi	Knowing The Non-Target Speakers: The Effect Of The I-Vector Population For PLDA Training In Speaker Recognition [35]	19 MFCC's	PLDA classifier	Improved performance	ICASSP 2013
	Gang Liu, Taufiq Hasan, Hynek Bořil, John H.L. Hansen	An Investigation On Back-End For Speaker Recognition In Multi-Session Enrollment [36]	Mel frequency Cepstral coefficients (MFCC)	Several back-ends on an i-vector system framework.	Relative improvement in EER and minDCF by 56.5% and 49.4%, respectively.	ICASSP 2013
7.	Balaji Vasan Srinivasan, Yuancheng Luo, Daniel Garcia-Romero, Dmitry N. Zotkin, and Ramani Duraiswami	A Symmetric Kernel Partial Least Squares Framework for Speaker Recognition [37]	57 mel-frequency cepstral coefficients (MFCC) features	Kernel partial least squares (KPLS) used for discriminative training in the i-vector space	8.4% performance improvement (relative) in terms of EER	IEEE Transactions On Audio, Speech, And Language Processing, Vol. 21, No. 7, July 2013

8.	Akshay S. Utane, Dr. S. L. Nalbalwar	Emotion Recognition through Speech Using Gaussian Mixture Model and Support Vector Machine [42]	prosodic features like pitch , energy and spectral features such as Mel frequency cepstrum coefficient [3]	Gaussian mixture model and support vector machine	Both the classifiers provide relatively similar accuracy for classification	International Journal of Scientific & Engineering Research, Volume 4, Issue 5, May-2013
9.	Pedro Univaso 1 Miguel Mart ínez Soler 1,2 Jorge A. Gurlekian	Human Assisted Speaker Recognition Using Forced Alignments on HMM [55]	13 MFCC coefficients with delta and acceleration.	HMMs	25.1% equal error rate reduction relative to a GMM baseline system	International Journal of Engineering Research & Technology (IJERT) Vol. 2 Issue 9, September - 2013 IJERTIJERT ISSN: 2278-0181 IJERTV2IS9 0739 www002E
10.	Martinez, J. ; Perez, H. ; Escamilla, E. ; Suzuki, M.M.	Speaker recognition using Mel frequency Cepstral Coefficients (MFCC) and Vector quantization (VQ) techniques [48]	mel-frequency cepstral coefficients (MFCC) features	Vector quantization technique	100% of precision with a database of 10 speakers.	Electrical Communications and Computers (CONIELECOMP), 2012 22nd International Conference
11.	Joder, Cyril ; Schuller	Exploring Nonnegative Matrix Factorization for Audio Classification: Application to Speaker Recognition [49]	mel-frequency cepstral coefficients (MFCC) features	Nonnegative Matrix Factorization (NMF)	significant improvement of accuracy was obtained	Speech Communication; 10. ITG Symposium, 26-28 Sept. 2012
12.	May, T. ; van de Par, S. ; Kohlrausch, A.	Noise-Robust Speaker Recognition Combining Missing Data Techniques and Universal Background Modeling [50]	signal-to-noise ratio (SNR) based mask estimation	universal background model (UBM)	substantial improvements in recognition performance	Audio, Speech, and Language Processing, IEEE Transactions on (Volume: 20 , Issue: 1) Biometrics Compendium , IEEE
13.	Sher, M., Ahmad, N. ; Sher, M.	TESPAR feature based isolated word speaker recognition system [51]	Time Encoded Signal Processing and Recognition (TESPAR) approach	Artificial Neural Network (ANN) classifier	TESPAR features gives better performance with a high recognition rate and low computational complexity as compared with MFCC and LPC based features.	Automation and Computing (ICAC), 2012 18th International Conference
14.	Lei, H.; Meyer, B.T. ; Mirghafori, N.	Spectro-temporal Gabor features for speaker recognition [52]	2D Gabor features (known as spectro-temporal features)	noisy ROSSI database	8% relative EER improvement over MFCC features standalone	Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference

15.	McLaren, M, van Leeuwen, D	Source-Normalized LDA for Robust Speaker Recognition Using i-Vectors From Multiple Speech Sources [53]	using a cosine kernel, i-vectors are projected into an linear discriminant analysis (LDA) space	source-normalized (SN) LDA algorithm	SN-LDA demonstrated relative improvements of up to 38% in equal error rate (EER) and 44% in minimum DCF over LDA	Audio, Speech, and Language Processing, IEEE Transactions on (Volume: 20 , Issue: 3) Biometrics Compendium , IEEE
16.	Garcia-Romero, D. Xinhui Zhou ; Espy-Wilson, Carol Y.	Multicondition training of Gaussian PLDA models in i-vector space for noise and reverberation robust speaker recognition [54]	speech utterances with adverse noise	Gaussian Probabilistic Linear Discriminant Analysis (PLDA) modeling	No specific improvement	Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference
17.	Tomi Kinnunen, Rahim Saeidi, Filip Sedlák, Kong Aik Lee, Johan Sandberg, Maria Hansson-Sandsten, Haizhou Li	Low-Variance Multitaper MFCC Features: A Case Study in Robust Speaker Verification [38]	mel-frequency cepstral coefficients (MFCCs)	Three Gaussian mixture model based classifiers with universal background model (GMM-UBM), support vector machine (GMM-SVM) and joint factor analysis (GMM-JFA)	20.4% (GMM-SVM), 13.7% (GMM-JFA)	IEEE Transactions On Audio, Speech, And Language Processing, Vol. 20, No. 7, September 2012
18.	P. M. Ghate, Shraddha Chadha, Aparna Sundar, Ankita Kambale	Automatic Speaker Recognition System [47]	Mel Frequency Cepstral Coefficients (MFCC)	Dynamic Time Warping (DTW) algorithm	Improvement is shown	International Conference on Advances in Computing Advances in Intelligent Systems and Computing Volume 174, 2012, p p 1037-1044
19.	Tobias May, Steven van de Par, and Armin Kohlrausch	Noise-Robust Speaker Recognition Combining Missing Data Techniques and Universal Background Modeling [39]	Spectral features	universal background model (UBM)	substantial improvements in recognition performance, especially in the presence of highly non-stationary background noise at low SNRs	IEEE Transactions On Audio, Speech, And Language Processing, Vol. 20, No. 1, January 2012
20.	Wen Wang, Andreas Kathol, Harry Bratt	Automatic Detection of Speaker Attributes Based in Utterance Text [40]	lexical features(n-grams) as well as features inspired by Linguistic Inquiry	linear kernel SVM	SVM performance on native/nonnative detection is 88.25 %, comparable to the 88.24% accuracy from the best single system.	INTERSPEECH, page 2361-2364. ISCA, (2011)

Table IV: REVIEW OF EMOTION RECOGNITION

S. No.	Authors	Title	Features considered	Classifier used	Accuracy rate	Published
1.	Garg, Vipul; Kumar, Harsh ; Sinha, Rohit	Speech based Emotion Recognition based on hierarchical decision tree with SVM, BLG and SVR classifiers [64]	hierarchical decision tree for the GMM means supervector based feature set	SVM, BLG and SVR	83% recognition results for closed set	Communications (NCC), 2013 National Conference

2.	Shashidhar g. Koolagudi · k. Sreenivasa rao	Emotion recognition from speech using source, system, and prosodic features [16]	Prosodic features in isolation and in combination	ANNs, GMMs, and SVMs	Combinations are found to improve the emotion recognition indicating the complementary nature of the features.	Springer science+business media, llc 2012
3.	Jae-bok kim, jeong-sik park, yung-hwan oh	Speaker-characterized emotion recognition using online and iterative speaker adaptation [17]	Log energy, 12-dimensional MFCCs, pitch, and their first and second derivatives as a feature vector	Modified maximum likelihood linear regression (mllr)	Results demonstrated the feasibility of applying ser to personal devices	Springer science+business media, llc 2012
4.	Yong-soo seol, han-woo kim and dong-joo kim	emotion recognition from textual modality using a situational personalized emotion model [18]	Situation model using lexical and syntactic information	Kbann(knowledge -based artificial neural network	Average accuracy 70.47%	International journal of hybrid information technology vol. 5, no. 2, april, 2012
5.	Kartik audhkhasi, shrikanth s. Narayanan	Emotion classification from speech using evaluator reliability-weighted combination of ranked lists [19]	13 mel filter bank (mfb) coefficients	Borda count and schulze's method	Results showed reliability weighted versions of the two ranked list voting methods performing the best.	ICASSP 2011
6.	Prof .sujata pathak , prof .arun kulkarni	Recognising emotions from speech [20]	Linear prediction coefficients (lpc)	Neural network (nn)	Accuracy rate was 46%	IEEE transactions on audio, speech, and language processing , 2011 ieee
7.	Muzaffar khan, tirupati goskula, mohmmed nasiruddin ,ruhin a quazi	Comparison between k-nn and svm method for speech emotion recognition [21]	Formant frequencies fo to log entropy	K-nearest neighbor (k-nn) and support vector Machine (svm) classifier	K-nearest neighbor-91.71% Svm – 76.57%	International journal on computer science and engineering, vol. 3 no. 2 feb 2011
8.	Priyanka abhang, shashibala rao, bharti w. Gawali, pramod rokade	Emotion recognition using speech and eeg signal – a review [22]	Electroencephalogram (eeg) brain signal and speech, the mel-frequency cepstral coefficients (MFCC)	Higher order crossings (hoc), empirical mode decomposition (emd)	A review on the combined efforts of eeg brain signal and speech to recognize the emotions in humans	International journal of computer applications (0975 – 8887) volume 15– no.3, february 2011
9.	N. Murali krishna, p.v. lakshmi, y. Srinivas j.sirisha devi	Emotion recognition using dynamic time warping technique for isolated words [23]	MFCC, delta coefficients (δMFCC) and delta delta coefficients (δδMFCC)	Dynamic time warping (dtw), Svm classifier	Recognition rates are above 78%	IJCSI international journal of computer science issues, vol. 8, issue 5, no 1, september 2011
10.	Krishna mohan kudiri, gyanendra k verma and bakul gohel	Relative amplitude based feature for emotion detection from speech [24]	Relative amplitude	Rbfc approach was used for segmentation of speech signal	71.2% with back-propagation neural networks 72% with svm classifier	IEEE transactions on audio, speech, and language processing , 2010
11.	Tsang-long pao, jun-heng yeh, yao-wei tsai	Recognition and analysis of emotion transition in mandarin speech signal [25]	Group of features(uniform, end detection, and whole sentence segmentation.)	Conventional k-nearest neighbour and weighted discrete –knn	The average accuracy rate attained was 73%	IEEE transactions on audio, speech, and language processing , 2010

12.	Emily mower, maja j mataric, shrikanth narayanan	A framework for automatic human emotion classification using emotion profiles [26]	Mel filterbank cepstral coefficients (MFCCs)	Ep-svm classification	Overall accuracy rate 68.2%	IEEE transactions on audio, speech, and language processing, vol. 19, no. 5, july 2011
13.	Marc lanze ivan c. Dy, ivan vener l. Espinosa, paul patrick v. Go, charles martin m. Mendez, jocelynn w. Cu	Multimodal emotion recognition using a Spontaneous filipino emotion database [43]	Both speech as well as facial features	SVM as classifier	40% using speech, 86% using face and 80% using the combination of speech and face.	IEEE transactions on audio, speech, and language processing , 2010
14.	Ying shi , weihua song	Speech emotion recognition based on data mining technology [27]	16-dimentional MFCC , 16-delta MFCC	BP neural networks for classification	Average recognition rate was 54.2%	2010 sixth international conference on natural computation (icnc 2010)
15.	Sheguo wang, xuxiong ling, fuliang zhang, jianing tong	Speech emotion recognition based on principal component analysis and Back Propagation neural network [44]	Prosodic features, such as Fundamental frequency, formant, intensity, duration time, And features that derive from these. Such as mean, Maximum, minimum, change rate, variance, median	Principal component analysis and back propagation neural network	Improved accuracy rate ranging from 52%-62%.	2010 international conference on measuring technology and mechatronics automation
16.	Sanghamitra mohanty , basanta kumar swain	Emotion recognition using fuzzy k-means from Oriya speech [45]	Incorporating mean pitch, first two formants, jitter, shimmer and energy as feature vectors.	Fuzzy k-means algorithm	Accuracy of 65.16%	2010 for international conference [accta-2010], 3-5 august 2010, special issue of IJCCT
17.	S. Das1, A. Halder, P. Bhowmik, A. Chakraborty, A. Konar, R. Janarthanan	A Support Vector Machine Classifier of Emotion from Voice and Facial Expression Data [57]	First three formants: F1, F2, and F3, and Respective powers at those formants, and pitch	Linear Support Vector Machine classifier	Recognition accuracy of emotion up to a level of 95%	IEEE transactions on audio, speech, and language processing , 2009
18.	Firoz shah.a, raji sukumar.a, babu anto.p	Automatic emotion recognition from speech Using artificial neural networks with gender-Dependent databases [46]	MFCCs, lpcs, fundamental frequency f0, formants, voice energy	Discrete wavelet transform for feature vector	Accuracy of 72.055% of accuracy for male database and 65.5% of accuracy for female database	2009 international conference on advances in computing, control, and telecommunication technologies
19.	Aditya bihar kandali, aurobinda routray, tapan kumar basu	Emotion recognition from assamese speeches using MFCC features and gmm classifier [28]	43 - mel-frequency cepstral coefficients (MFCC)	Gaussian mixture model (gmm)	74.4% with m=10 and fd=15 ms, 76.5% with m=12 and fd=23.22 ms, and 66.9% with m=11 and fd=40 ms	TENCON 2008 - 2008 ieee region 10 conference
20.	Daniel neiberg, kjell elenius, inger karlsson1, and kornel laskowski	Emotion recognition in spontaneous speech [29]	Mel-frequency cepstral Coefficients, MFCCs, and a variant, MFCC-low, that is calculated between 20 and 300 hz in Order to model pitch	Gaussian mixture models	Two corpora were used: telephone services and meetings. Results show that frame level gmms are useful for emotion classification	Lund university, centre for languages & literature, dept. Of linguistics & phonetics working papers 52 (2006), 101–104

21.	Oh-wook kwon, kwokleung chan, jiucang hao, te-won lee	Emotion recognition by speech signals [41]	Pitch, log energy, formant, mel-band energies, and mel frequency cepstral coefficients (MFCCs), velocity/ acceleration of pitch and MFCCs to form feature streams	Quadratic discriminant analysis (qda) and support vector machine (svm)	Accuracy of 96.3% for stressed/neutral style classification and 70.1% for 4-class speaking style classification using gaussian svm. Speaker-independent aibo database, they achieved 42.3% accuracy for 5-class emotion recognition	Eurospeech 2003 – geneva

TABLE V: ADVANTAGES AND DISADVANTAGES OF PROMINENT CLASSIFIERS

Classifier	Description	Advantage	Disadvantage
Binary Decision Tree	Decision Tree is a stream design drawing like structure in which inward hub talks to check on a quality, every extension talks to deduction of check and every leaf hub talks to class title (choice taken in the wake of registering all traits). A way from root to leaf speaks to arrangement runs the display. In alternative examination a conclusion tree and the almost identified influence journal is utilized as a visual and scientific choice support device, where the usual qualities (or required utility) of arguing options are computed.	Easy implementation, easy explanation of input and output relationship Can handle high dimensional data Easy to interpret for small sized trees The learning and classification steps of induction are simple and fast Accuracy is comparable to other classification techniques for many simple data sets Convertible to simple and easy to understand classification rules	Decision-tree learners can create over-complex trees that do not generalize the facts and figures well. Decision trees can be unstable because small variations in the facts and figures might outcome in a absolutely different tree being developed. This difficulty is mitigated by using decision trees inside an ensemble. The difficulty of discovering an optimal decision tree is known to be NP-complete under several facets of optimality and even for easy concepts. Consequently, functional decision-tree learning algorithms are founded on heuristic algorithms such as the greedy algorithm where locally optimal decisions are made at each node. Such algorithms will not assurance to return the globally optimal decision tree. There are concepts that are hard to discover because decision trees do not articulate them effortlessly, such as XOR, parity or multiplexer troubles. conclusion tree learners conceive biased trees if some categories dominate. It is thus suggested to balance the dataset prior to fitting with the conclusion tree
Artificial Neural Network	An Artificial Neural mesh (ANN) is a facts and figures organising standard that is inspired by the way biotic anxious structures, for example the cerebrum, process facts and figures. The key constituent of this ideal model is the innovative structure of the facts and figures handling structure. It is made out of countless interconnected changing components (neurones) employed as one to tackle specific issues. ANNs, for demonstration persons, study by illustration. An ANN is designed for a specific provision, for demonstration design acknowledgement or information characterization, through a revising method. revising in living structures includes acclimations to the synaptic associations that exist between the neurones. This is accurate of ANNs besides	They can both about any convoluted conclusion supplied that enough nodes are utilised. Neural systems are rather easy to implement (you do not need a good linear algebra solver as for examples for SVNs). Neural networks often exhibit patterns alike to those exhibited by humans. although this is more of interest in cognitive sciences than for functional examples	Long preparing time The VC measurement of neural systems is indistinct. This is extremely critical when you need to think about how exceptional an answer could be. Neural systems can't be retrained. Provided that you include information later, this is just about difficult to add to an existing system. Taking care of time arrangement information in neural systems is an exceptionally confounded point.

Bayesian Network	Structured, graphical representation of probabilistic connections between some random variables. Explicit representation of dependent independencies. Missing arcs encode dependent independence	It can accommodate a kind of knowledge sources and data types can gladly handle incomplete details and numbers sets. Bayesian systems permit one to discover about causal attachments Bayesian procedures supply an effective method for halting the over fitting of details and figures (there is no need for pre-processing)	Information theoretically infeasible. It turns out that specifying a prior is extremely difficult. Computationally infeasible
Support Vector Machine [66]	Support Vector appliance (SVM) is mainly a classes method that performs classification jobs by assembling hyperplanes during a dimensional house that divides situations of various class labels. SVM carries each regression and classification jobs and may handle multiple relentless and categorical variables. For categorical variables a dummy variable is formed with case standards as either none or one.	Adaptability in picking a closeness capacity Inadequacy of result when managing expansive information sets just underpin vectors are utilized to determine the differentiating hyperplane Capacity to handle substantial characteristic spaces multifaceted nature does not hinge on upon the dimensionality of the characteristic space. Overfitting might be regulated by delicate edge approach Nice math property: a simple convex optimization problem which is guaranteed to converge to a single global solution Good generalization capability	It is sensitive to noise A relatively small number of mislabeled examples can dramatically decrease the performance It only considers two classes
Hidden Markov Model	Hidden Markov forms (HMMs) are a formal establishment for making probabilistic forms of directly arrangement calling" issues. They give a theoretical device stash for building complex illustrates just by drawing an instinctive picture. They are at the heart of a different continue of schemes, incorporating gene finding, profile hunts for, various grouping placement and administrative location recognizable verification. HMMs are the Legos of computational grouping examination	Modularity: HMMs can be combined into bigger HMMs Transparency of the form: presuming architecture with a good conceive. Persons can read the form and make sense of it. The form itself can help boost comprehending. Incorporation of former information: Incorporate former knowledge into the architecture. Initialize the form close to certain thing believed to be correct. Use former information to constrain teaching process	Markov Chains: States are supposed to be unaligned. P(y) must be unaligned of P(x), and vice versa. This generally isn't factual. Can get round it when connections are localized. Not good for RNA bending problems. Standard appliance discovering Problems. Watch out for local maxima: form may not converge to a really optimal parameter set for a given training set. by pass over-fitting: You're only as good as your teaching set. More teaching is not habitually good.
Gaussian Mixture Model	Mixture Models are a sort of width form which embodies diverse part capabilities, usually Gaussian. These segment capabilities are consolidated to furnish a multimodal thickness. They might be utilized to form the hues of an article holding in mind the end goal to present undertakings, for demonstration unchanging color-based following and partition. These undertakings may be made mightier by producing a blend model matching to foundation colors notwithstanding a forefront model, and utilizing Bayes' hypothesis to present pixel order. blend forms are likewise agreeable to ample schemes for on-line places to stay of models to acclimatize to step-by-step fluctuating lighting conditions	Very simple implementation, self-assurance level can be got from the posterior probabilities. Even though Gaussian constituents, the overhead architectures can approximate arbitrary convoluted circulation. Advantages of expectation-maximization: • It is the fastest algorithm for learning mixture forms. • As this algorithm maximizes only the prospect, it will not bias the means in the direction of none, or bias the cluster dimensions to have exact organizations that might or might not request	Sample distributions may not be Gaussian. Disadvantages of expectation-maximization: **Singularities:** when has insufficiently lots of points per mixture, estimating the covariance matrices becomes difficult, & the algorithm is known to diverge & find solutions with boundless likelihood unless regularizes the covariances artificially. This algorithm will always use all the parts it's access to, needing complex held-out knowledge criteria to select how lots of parts to make use of without outside cues.

K- Nearest Neighbor	All occurrences contrast to focuses in a n-dimensional Euclidean space. Alignment is postponed till another occasion reaches. Alignment finished by investigating emphasizes vectors of the characteristic focuses. goal capacity may be discrete or authentic esteemed	No training is required, confidence level can be obtained	Classification correctness is reduced is convoluted decision-region boundary lives, sizable storage needed KNN algorithm is that it is a slovenly learner, i.e. it does not discover any thing from the teaching data and easily values the teaching data itself for classification The algorithm should compute the expanse and sort all the teaching data at each proposition, which can be slow if there are a large number of teaching demonstrations The algorithm does not learn anything from the teaching facts and figures, which can outcome in the algorithm not generalizing well and furthermore not being robust to loud data
Vector Quantization [4] [5]	A vector quantizer maps k-dimensional vectors in the vector space Rk into a limited set of vectors Y = {yi: i = 1, 2, ..., N}. Each vector yi is known as a code vector or a codeword. Furthermore the set of every last one of codewords is known as a codebook. Associated with every codeword, yi, is a closest neighbor area called Voronoi locale.	Donates smallest spectral distortion for a given bit-rate as the correlation that lives between the trials of a vector is maintained With smaller codebooks computational complexity and recollection obligations are reduced to a very great extent	Gives smallest spectral distortion for a granted bit-rate as the association that exists between the trials of a vector is preserved High complexity, recollection obligations and the lifetime of codebook is tough task as vectors of full length are used for quantization without any functional constraint. Due to splitting the linear and non linear dependencies that exist between the trials of a vector is lost and the shape of the quantizer cells is affected. As a result the spectral distortion rises slightly
Naïve Bayes classifier	The Naive Bayes Classifier procedure is reliant upon the purported Bayesian hypothesis and is particularly suited when the dimensionality of the inputs is high. Notwithstanding its effortlessness, Naive Bayes can frequently outflank more refined grouping strategies	Fast to train (single scan). Fast to classify Not sensitive to irrelevant features Handles real and discrete data Handles streaming data well	Assumes independence of features

REFERENCES

[1] Gish, H., Schmidt, M., 1994. Text-independent speaker recognition. IEEE Signal Process. Magazine (October), 18–32.

[2] Furui, S., 1981. Cepstral analysis technique for automatic speaker verification. IEEE Trans. Acoust. Speech Signal sProcess. 29 (2), 254–272.

[3] Gold, B. and Rabiner, L.R, Parallel processing techniques for estimating pitch periods of speech in time-domain.

[4] Y. Linde, A. Buzo, and R. M. Gray.: 'An algorithm for vector quantizer design," IEEE Trans. Commun.', vol. COM-28, no. 1, pp. 84-95, 1980.

[5] A. Gersho, R.M. Gray.: 'Vector Quantization and Signal Compression', Kluwer Academic Publishers, Boston, MA, 1991.

[6] Lawrence Rabiner, Biing-Hwang Juang and B.Yegnanarayana, "Fundamental of Speech Recognition ", Prentice-Hall, Englewood Cliffs, 2009.

[7] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge," in Interspeech (2009), ISCA, Brighton, UK, 2009.

[8] Serajul Haque, Roberto Togneri and, Anthony Zaknich," Zero Crossings with Peak Amplitudes and Perceptual Features for Robust Speech Recognition", http://www.ee.uwa.edu.au/~roberto/research/theses/tr0 6-01.pdf, March 2012.

[9] K. R. Scherer, "How emotion is expressed in speech and singing," in Proceedings of XIIIth International Congress of Phonetic Sciences., pp. 90-96. 1995.

[10] M. Kockmann, L. Ferrer, L. Burget, E. Shriberg, and J. H. Cernocký, "Recent progress in prosodic speaker verification," in Proc. IEEE ICASSP, (Prague), pp. 4556--4559, May 2011.

[11] Kumar, K. S., Reddy, M. S. H., Murty, K. S. R., & Yegnanarayana, B. (2009). Analysis of laugh signals for detecting in continuous speech. In INTERSPEECH-09, Brighton, UK, September 6–10 (pp. 1591–1594).

[12] Koolagudi, S. G., & Rao, K. S. (2009). Exploring speech features for classifying emotions along valence dimension. In S. Chandhury, et al. (Eds.), LNCS. The 3rd international conference on pattern recognition and machine intelligence (PreMI-09), IIT Delhi, December 2009 (pp. 537–542). Heidelberg: Springer.

[13] Bitouk, D., Verma, R., & Nenkova, A. (2010). Class-level spectral features for emotion recognition. Speech Communication, 52(7–8), 613–625.

[14] Iliou, T., & Anagnostopoulos, C. N. (2009). Statistical evaluation of speech features for emotion recognition. In Fourth international conference on digital telecommunications, Colmar, France, July 2009 (pp. 121–126).

[15] Kamaruddin, N., & Wahab, A. (2009). Features extraction for speech emotion. Journal of Computational Methods in Science and Engineering, 9(9), 1–12.

[16] Rao, K. S., & Yegnanarayana, B. (2009). Intonation modeling for Indian languages. Computer Speech and Language, 23, 240–256.

[17] [Rao, K. S., Prasanna, S. R. M., & Yegnanarayana, B. (2007). Determination of instants of significant excitation

in speech using Hilbert envelope and group delay function. IEEE Signal Processing Letters, 14, 762–765.

[18] Prasanna, S. R. M., Reddy, B. V. S., & Krishnamoorthy, P. (2009). Vowel onset point detection using source, spectral peaks, and modulation spectrum energies. IEEE Transactions on Audio, Speech, and Language Processing, 17, 556–565.

[19] Yegnanarayana, B., Swamy, R. K., & Murty, K. S. R. (2009). Determining mixing parameters from multispeaker data using speech specific information. IEEE Transactions on Audio, Speech, and Language Processing, 17(6), 1196–1207.

[20] Lugger, M., & Yang, B. (2007). The relevance of voice quality features in speaker independent emotion recognition. In ICASSP, Honolulu, Hawaii, USA, May 2007 (pp. IV17–IV20). New York: IEEE.

[21] Zhang, S. (2008). Emotion recognition in Chinese natural speech by combining prosody and voice quality features. In Sun, et al. (Eds.), Lecture notes in computer science. Advances in neural networks (pp. 457–464). Berlin: Springer.

[22] W. Wang, A. Kathol, and H. Bratt, "Automatic detection of speaker attributes based on utterance text," in Proc. Interspeech, (Florence, Italy), pp. 2361- 2364, August 2011.

[23] Khalid Saeed and Mohammad Kheir Nammous, "A Speech-and-Speaker Identification System: Feature Extraction, Description, and Classification of Speech-Signal Image", IEEE Transactions on Industrial Electronics Vol. 54, No.2, April 2007, pp. 887-897.

[24] Nitisha and Ashu Bansal, "Speaker Recognition Using MFCC Front End Analysis and VQ Modelling Technique for Hindi Words using MATLAB", Hindu College of Engineering, Haryana, India.

[25] Marius Vasile Ghiurcau , Corneliu Rusu,Jaakko Astola, "A Study Of The Effect Of Emotional State Upon Text-Independent Speaker Identification", ICASSP 2011.

[26] M.A.Anusuya , S.K.Katti ,"Speech Recognition by Machine: A Review", International Journal of Computer Science and Information Security, Vol. 6, No. 3, 2009.

[27] Shashidhar G. Koolagudi · K. Sreenivasa Rao," Emotion recognition from speech using source, system,and prosodic features", Springer Science+Business Media, LLC 2012.

[28] Jae-Bok Kim, Jeong-Sik Park, Yung Hwan Oh," Speaker-Characterized Emotion Recognition using Online and Iterative Speaker Adaptation", Springer Science+Business Media, LLC 2012.

[29] Yong-Soo Seol, Han-Woo Kim and Dong-Joo Kim, "Emotion Recognition from Textual Modality Using a Situational Personalized Emotion Model", International Journal of Hybrid Information Technology Vol. 5, No. 2, April, 2012.

[30] Kartik Audhkhasi, Shrikanth S. Narayanan," Emotion classification from speech using evaluator reliability-weighted combination of ranked lists", ICASSP 2011.

[31] Prof .Sujata Pathak , Prof .Arun Kulkarni ," Recognising emotions from speech", IEEE Transactions On Audio, Speech, And Language Processing , 2011 IEEE.

[32] Muzaffar Khan, Tirupati Goskula, Mohmmed Nasiruddin ,Ruhina Quazi," Comparison between k-nn and svm method for speech emotion recognition", International Journal On Computer Science And Engineering, Vol. 3 No. 2 Feb 2011.

[33] Priyanka Abhang, Shashibala Rao, Bharti W. Gawali, Pramod Rokade, "Emotion Recognition using Speech and EEG Signal – A Review", International Journal Of Computer Applications (0975 – 8887) Volume 15– No.3, February 2011.

[34] N. Murali Krishna, P.V. Lakshmi, Y. Srinivas J.Sirisha Devi, "Emotion Recognition using Dynamic Time Warping Technique for Isolated Words", IJCSI International Journal Of Computer Science Issues, Vol. 8, Issue 5, No 1, September 2011.

[35] Krishna Mohan Kudiri, Gyanendra K Verma and Bakul Gohel," Relative amplitude based feature for emotion detection from speech", IEEE Transactions On Audio, Speech, And Language Processing , 2010.

[36] Tsang-Long Pao, Jun-Heng Yeh, Yao-Wei Tsai," Recognition and analysis of emotion transition in Mandarin speech signal", IEEE Transactions On Audio, Speech, And Language Processing , 2010.

[37] Emily Mower, Maja J Mataric, Shrikanth Narayanan," A framework for automatic human emotion classification using emotion profiles", IEEE Transactions On Audio, Speech, And Language Processing, Vol. 19, No. 5, July 2011.

[38] Ying Shi, Weihua SONG," Speech emotion recognition based on data mining technology", 2010 Sixth International Conference on Natural Computation (ICNC 2010).

[39] Aditya Bihar Kandali, Aurobinda Routray, Tapan Kumar Basu," Emotion recognition from Assamese speeches using MFCC features and GMM classifier", Tencon 2008 - 2008 IEEE Region 10 Conference.

[40] Daniel Neiberg, Kjell Elenius, Inger Karlsson1, and Kornel Laskowski," Emotion Recognition in Spontaneous Speech", Lund University, Centre For Languages & Literature, Dept. Of Linguistics & Phonetics Working Papers 52 (2006), 101–104.

[41] Oh-Wook Kwon, Kwokleung Chan, Jiucang Hao, Te-Won Lee," Emotion Recognition by Speech Signals", Eurospeech 2003 – Geneva.

[42] Alejandro Bidondo, Shin-ichi Sato, Ezequiel Kinigsberg, Adrián Saavedra, Andrés Sabater, Agustín Arias, Mariano Arouxet, and Ariel Groisman, "Speaker recognition analysis using running autocorrelation function parameters", POMA - ICA 2013 Montreal Volume 19, pp. 060036 (June 2013).

[43] Taufiq Hasan, Seyed Omid Sadjadi, Gang Liu, Navid Shokouhi, Hynek Boˇril, John H.L. Hansen," CRSS SYSTEMS FOR 2012 NIST SPEAKER RECOGNITION EVALUATION", ICASSP 2013.

[44] M. Afzal Hossan · Mark A. Gregory," Speaker recognition utilizing distributed DCT-II based Mel frequency cepstral coefficients and fuzzy vector quantization", Int J Speech Technol (2013), Springer Science+Business Media, LLC 2012.

[45] Taufiq Hasan, John H. L. Hansen," Acoustic Factor Analysis for Robust Speaker Verification", IEEE Transactions On Audio, Speech, And Language Processing, Vol. 21, No. 4, April 2013.

[46] David A. van Leeuwen and Rahim Saeidi," Knowing The Non-Target Speakers: The Effect Of The I-Vector Population For Plda Training In Speaker Recognition", ICASSP 2013.

[47] Gang Liu, Taufiq Hasan, Hynek Bořil, John H.L. Hansen," An Investigation On Back-End For Speaker Recognition In Multi-Session Enrollment", ICASSP 2013.

[48] Balaji Vasan Srinivasan, Yuancheng Luo, Daniel Garcia-Romero, Dmitry N. Zotkin, and Ramani Duraiswami," A Symmetric Kernel Partial Least Squares Framework for Speaker Recognition", IEEE Transactions On Audio,

Speech, And Language Processing, Vol. 21, No. 7, July 2013.
[49] Tomi Kinnunen, Rahim Saeidi, Filip Sedlåk, Kong Aik Lee, Johan Sandberg, Maria Hansson-Sandsten, Haizhou Li," Low-Variance Multitaper MFCC Features: A Case Study in Robust Speaker Verification", IEEE Transactions On Audio, Speech, And Language Processing, Vol. 20, No. 7, September 2012.
[50] Tobias May, Steven van de Par, and Armin Kohlrausch," Noise-Robust Speaker Recognition Combining Missing Data Techniques and Universal Background Modeling", IEEE Transactions On Audio, Speech, And Language Processing, Vol. 20, No. 1, January 2012.
[51] Wen Wang, Andreas Kathol, Harry Bratt," Automatic Detection of Speaker Attributes Based in Utterance Text", INTERSPEECH, page 2361-2364. ISCA, (2011).
[52] Akshay S. Utane, Dr. S. L. Nalbalwar, "Emotion Recognition through Speech Using Gaussian Mixture Model and Support Vector Machine", International Journal of Scientific & Engineering Research, Volume 4, Issue 5, May-2013.
[53] Marc Lanze Ivan C. Dy, Ivan Vener L. Espinosa, Paul Patrick V. Go, Charles Martin M. Mendez, Jocelynn W. Cu, "Multimodal Emotion Recognition Using a Spontaneous Filipino Emotion Database", IEEE Transactions On Audio, Speech, And Language Processing, 2010.
[54] Sheguo Wang,Xuxiong Ling,Fuliang Zhang,Jianing Tong," Speech Emotion Recognition Based on Principal Component Analysis and Back Propagation Neural Network", 2010 International Conference on Measuring Technology and Mechatronics Automation.
[55] Sanghamitra Mohanty , Basanta Kumar Swain," Emotion Recognition using Fuzzy K-Means from Oriya Speech", 2010 for International Conference [ACCTA-2010], 3-5 August 2010, Special Issue of IJCCT.
[56] Firoz Shah.A, Raji Sukumar.A, Babu Anto.P," Automatic Emotion Recognition from Speech using Artificial Neural Networks with Gender- Dependent Databases", 2009 International Conference on Advances in Computing, Control, and Telecommunication Technologies.
[57] S. Das1, A. Halder, P. Bhowmik, A. Chakraborty, A. Konar, R. Janarthanan," A Support Vector Machine Classifier of Emotion from Voice and Facial Expression Data ",IEEE Transactions On Audio, Speech, And Language Processing 2009.
[58] M.D. Zeiler, M. Ranzato, R. Monga, M. Mao, K. Yang, Q.V. Le,P. Nguyen, A. Senior, V. Vanhoucke, J. Dean, , G.E. Hinton, " ON RECTIFIED LINEAR UNITS FOR SPEECH PROCESSING", Zeileretal_ICASSP13.
[59] A. B. Ingale, D. S. Chaudhari,"Speech Emotion Recognition", Int'l Journal of Soft Computing and Engineering, vol-2, Issue-1, pp 235-238, Mar. 2012.
[60] C. Busso, S. Lee and S. Narayanan, "Analysis of Emotionally Salient Aspects of Fundamental Frequency for Emotion Detection", IEEE Trans. on Audio, Speech and Language processing, vol. 17, no. 4, pp 582-596,May 2009.
[61] I. Luengo, E. Navas, I. Hernáez , "Feature Analysis and Evaluation for Automatic Emotion Identification in Speech", IEEE Trans. on Multimedia, vol. 12, no. 6,pp 1117-1127, Oct. 2010.
[62] Chung-Hsien Wu and Wei-Bin Liang, "Emotion Recognition of Affective Speech Based on Multiple Classifiers Using Acoustic-Prosodic Information and Semantic Labels", IEEE Trans. on Affective Computing, vol. 2, no. 1, pp 10-21,Jan-Mar 2011.
[63] Manish P. Kesarkar, Prof. Preeti Rao, "FEATURE EXTRACTION FOR SPEECH RECOGNITON", M.Tech. Credit Seminar Report, Electronic Systems Group, EE. Dept, IIT Bombay, Submitted November2003.
[64] Garg, Vipul, Kumar, Harsh; Sinha, Rohit," Speech based Emotion Recognition based on hierarchical decision tree with SVM, BLG and SVR classifiers" Communications (NCC), 2013 National Conference.
[65] Nitin Trivedi, Dr. Vikesh Kumar, Saurab. Singh, Sachin Ahuja, Raman Chadha, "Speech Recognition by Wavelet Analysis", International Journal of Computer Applications (0975 – 8887) Volume 15– No.8, February 2011.
[66] W.M.Campbell, J.P.Campbell, T.P. Gleason, D.A. Reynolds, and T.R.Leek,'High-Level Speaker Verification With Support Vector Machines,' ICASSP, 2004.
[67] R. Schwartz, J. Campbell, W. Shen, D. E. Sturim, W. M. Campbell, F. S. Richardson, R. B. Dunn et al. USSSMITLL 2010 human assisted speaker recognition. Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference, pp. 5904–5907, (2011).

Dr. Srinivas Yarramalle was awarded M. Tech (Computer Science & Technology) from Andhra University -1999. He was awarded Doctorate in Computer Science with Specialization in Image Processing from Acharya Nagarjuna University, Guntur.- 21-1-2008. He received Best Teacher Award from JNTU University, Sept-5-2010 and from B.C. Corporation in 2005. He prepared 4 Monographs for School of Correspondence Courses, Andhra University. He received SASTRA Award from Vignan' s Institute of Information Technology, 10-Jan-2008, for Research publications. Received SASTRA Award from Vignan's Institute of Information Technology, 10-Jan- 2009, for Research publications.

How to cite this paper: J. Sirisha Devi, Srinivas Yarramalle, Siva Prasad Nandyala,"Speaker emotion recognition based on speech features and classification techniques", IJIGSP, vol.6, no.7, pp. 61-77, 2014.DOI: 10.5815/ijigsp.2014.07.08

