Contents

List of Figures

List of Tables

list of tables

Abstract

Spoken communication is very unique virtue with which humans are most effectively interacting with each other.Speech conveys information in many folds, one of that is context, another is how it is delivered, and so on. Research says that major information lies in the way context is said. Under this research we study this aspect of speech as prosody. Prosody is an outstandingly important part to understand verbal communication. This research study focus on prosody in Marathi regional language spoken in India. When human and machines are dealing with each other in spoken medium, the Efforts are towards making our machine understand what is prosody and how to analyze it being in same context, is new research area. The automatic mining of prosodic information is essential for machines to deal speech with human levels of ability ,as prosody convey more meaning compared to context. In this thesis we describe work on the automatic detection and classification of prosodic events in speech segment.features are computed for the prosody detection and classification by the new suggested hybrid approach.Prosody highly depends on the timing information of the signals; therefore an approach to compute features by using Empirical mode decomposition of speech signal is presented.This approach computes the features from the decomposed sub bands based on spectral energy thresholding.It results in higher accuracy in capturing the variations in the speech parameters like jitter, shimmer and HNR.Rest of the features are computed with conventional approach,which includes fundamental frequency,pitch slope,spectral energy and DVB.Last phase of work carries out the transformation of plain discourse into target style of emotion. We present here novel technique, in feature representations and state-of-the-art performance in transforma-

tion of the emotion of an utterance.Conversion of the emotion from one state to other is achieved through Spectral correlative prosodic mapping of the source and target features .Spectral correlative prosodic mapping avoids the mapping of the features in un correlated spectral segments hence features are selectively computed and used for mapping and conversion approach.

Chapter 1

1 introduction

Speech synthesis technology has advanced extremely over the past few years, particularly with regard to segmental naturalness. It leads to enhanced quality of delivery of sound from the speech engines, yet the need of proper prosodic phrasing and distinction assignment becomes all the more obvious. Most commercial text-to-speech systems now a days make use rather simple methods, typically assigning a default sentence accent based on the substance/function word distinction. Prosody is the field of investigating for the parameters like the intonation pattern as well as stress indicators and rhythm content in the human speech. Which is now of a great significance in phonetics, phonology and speech processing . Once it was referred as biddable to study about segmental structure but it is now referred as the ”framework” provider, which represents various levels of phonetic representation .In the recent past views of the phonology of intonation it is observed that the increase of auto segmental or non-linear accounts of phonetic description incorporate metrical formation with phonetic essence (Pierrehumbert, 1981). The accountability of prosody is also changing in speech synthesis. In the processes of synthesis, the accomplishment of concatenative systems with recorded segments of speech are patched together to make narrative utterances - has meant that the key points have shifted from segmental to supra-segmental quality (Klatt., 1987). In speech recognition the rising prominence on discourse systems indicates the more research inclination into the automatic Determination of prosodic composition for the purposes of utterance disam-

biguation (Wightman, 1994)

In the development of speech coding, speech conversion between human and machines is a budding area. Speech conversion has different significance of usage from identification or representation of expression in dialog to forensic discovery to criminal inquiry. The need of future speech synthesis lies on the not just delivery of speech content but to successfully convey meaning of its content through prosody and apply to various critical applications. The medical diagnosis, authentication usage, cinematography , storytelling, Medical field, Effective human Machine interface, Education , Social Environment, Text to Speech Systems Entertainment are few such applications. Discourse alongside feeling uncovers the individual goal. It is subsequently required to know the discourse talked as well as the feeling at which it is spoken as. In The process of expressive synthesis using prosody modeling also has become significant component under research. Wherein prosody models are utilized to distinguish talked discourse feeling, it can likewise be utilized to produce synthetic discourse of emotion from given discourse plain emotion input. Be that as it may, the handling precision of the change from one emotion to other is reliant on the prosodic features and conversion logic. Subsequently it is required to infer the prosody model all the more precisely in order to accomplish legitimate change. Wherein different methodologies were created to accomplish the goal of proper extraction of prosody information from the speech signal, and mapping it in plain discourse for synthetic transformation. It is very evident that if the impacts of noise, environmental disturbances or magnitude varieties are not dealt with correctly it drives wrong change of the discourse. Henceforth it is required to have an exact feature extraction and feature transplantation in emotion synthesis for speech coding.

The issue of emotion transformation in speech segment its conceivable solutions have been drawing in much consideration for fast generation of target expressive styles

Presented work is a simple emotion conversion structure which can be actualized utilizing rules which alter an input utterance deterministically. Finding right intonation, energy content (stress), and duration from composition is the most significant issue for a considerable length of time to come. These features together are called prosodic or supra segmental features and are considered as the rhythm, emphasis of the speech at the perceptual level. The intonation revels how the pitch pattern or fundamental frequency changes amid speech. The prosody of persistent speech relies upon many separate perspectives, for example, the significance of the sentence and the speaker qualities and feelings (Yamagishi, 2003)

Composed text, Unfortunately contains very small information of these features and some of them change dynamically amid discourse. Be that as it may, with some particular control characteristics this information is used to achieve required transformation. Timing at sentence level or grouping of words into phrases correctly is tricky because prosodic choice of words is not always marked in text by punctuation, and phrasal complement is never marked (Santen, 1997). In the event that there is no breath delays in speech or on the off chance that they are in wrong places, the discourse may sound exceptionally unnatural or even the significance of the sentence might be misconstrued. For instance, the information string "John says Peter is a liar" can be talked as two diverse ways giving two unique implications as "John says: Peter is a liar" or "John, says Peter, is a liar". In the first sentence Peter is a liar, and in the second one the liar is John. The work presented here deals in the area of classifying the

speech utterances in to its emotional classes .from the information of the class of the emotion in word level structures along with considering linguistic rules a correct information can be understood from discourse .this processes will further assist our aim of making human machine dialog system more meaningful. Emotion transformation is a process of converting source style of speaking into target style of speaking. Various rule-based endeavors exist in the writing good rules for each expressive style requires manual examination and can just catch an exceptionally restricted arrangement of acoustic prosodic divergences (Schroder, 1999). As of late, information-driven voice transformation techniques have been investigated for demonstrating, modeling and transforming both short-term spectra and prosody (Wu, 2006), (Tsuzuki, 2004).Using explicit control framework in the emotion transformation we have used a new Spectral correlative approach for feature mapping and transformation. Mapped features are transformed on to speech utterance through LPC re synthesis and the TD-PSOLA method. Here the aim in doing this work was to use analysis-by-synthesis to make progress towards a full modeling of Marathi prosody.

1.1 Motivation of the Thesis

Marathi is a official language spoken mainly in the ”Maharashtra” an Indian state, having its population of over 100 million. Marathi is relatively unattended language as far as the speech enabled applications and prosodic study is concerned. However Hindi which is spoken by over 250 million people across globe, Marathi share numerous similarities .These two languages are similar in regard to the script as well as pronuncia-

tion since they are originated from Sanskrit, as being one of the Indo-Aryan languages .Studies on the prosody is done for Hindi (Genzel, 2010) (Harnsberger, 1996)However Hindi is known to be influenced by Persian while Marathi has Dravidian influences in its phonetic.

It is observed that it matters how the things are said than "what" is said .Almost 70 percent meaning is conveyed by prosody. Is it really true to say like this? How much context is important when evaluated against the tone (i.e., prosody) in the view of expressing emotions in speech, and is the processing of prosody differs depending on the emotion in an utterance which is expressed. Investigating that How the processing of emotion change, how over the course of a spoken utterance that contains vocal cues about it and a potential semantic framework for interpreting the speaker's emotion state Emotion recognition bolstered in some way at the intersection of prosody and semantic cues during emotional speech processing. This research study seeks to deal with these questions and makes an attempt to provide insight on how emotion is perfectly recognized or distinguished, and processed further for its conversion from one shed to another with respect to emotions.

The expression is put in plain words as the vocal marker points in different emotional states that mirror in the speech. The diverse expressions and talking patterns are also considered as emotions. In light of this, in the present research work, we have considered distinctive emotions as the expressions and hence we are referring emotions as expressions. The purpose of speech synthesis is to produce speech signal from the transcript. Speech Re-synthesized in different expressions can be used in applications for children like effective story narrating, for drawing attention; it is recommended to deliver speech segments in different styles of expressions

for different contexts. Expressive speech if understood by machine and in reply if appropriate response is also delivered then it proves to be very useful as a part of dialogue system in making the human computer interaction more effective. Further expressive speech analysis can help call centers to spot the emotional state of the customers during conversation and they can be assisted based on their emotional state and maturity. As anxiety and anger emotions reflected by people may get treated differently than to calm and happy customers. Expressive Speech Synthesis finds application in the financial information system to make announcements in special speaking styles to the users (Pitrelli, 2006). To one side from the expressive speech synthesis systems developed based on various speech synthesis approaches, under explicit prosody control, work is done in this research for neutral to target expressive speech conversion. discrete emotions like angry, happy, sad and interrogative, are considered for Re-synthesis. Scaling of sentence energy, F0 ,and duration is done in linear modification

Problems Identified:

1. Prosodic event detection in a speech segment

2. Prosody classification using through Empirical Mode Decomposition feature coding

3. Deploying SVM with hybrid kernel approach for the classification

4. prosody conversion framework using spectral correlative prosodic mapping

1.2 Objectives and Research contributions

1.2.1 Prosodic event detection in a speech segment

The objective of this work is to analyze how acoustic features are related to elements which are responsible for expressivity, and how such prosody pattern differs in the statistical variations of the features in the speech segments to give rise of perception of specific prosody

1.2.2 Novel feature coding for prosody detection in emotional speech and

Feature space is build based on the transformation kernel to a space of lower dimension which allows post processing stage resulting in more useful information. we present a work which does not consider processing of speech signal for feature extraction in frequency domain as in frequency domain it tends to lose the time related clues. Prosody highly depends on the timing information of the signals an approach to compute features by using Empirical mode decomposition of speech signal is presented

1.2.3 Prosody conversion framework using spectral correlative mapping

Prosody in a speech conveys meaningful information along with context. Appropriate prosody modification in synthetic plain Speech plays a vital role in developing an effective speech interface platform. This research study focus on Marathi regional language. It presents a spectral mode decomposition (SMD) approach for Spectral correlative prosodic mapping to achieve emotion transformation

1.2.4 Speech Corpus designed

Recordings for the Marathi speech inventory was carried out in the professional sound recording studio, these are nothing but acoustically treated rooms possessing a moderately non reverberant characteristics. In the outline of LPC analysis and synthesis these reverberant object can show up in the residual part of signal, which in line could foil the concatenation of acoustic units smoothly. Recordings done in a sound studio reduces the influence of the acoustic room conditions to a minimum. A good-quality close-talking condenser microphone was used for recording which shows a low cutoff frequency ,was used so that the inverse-filtered sound signal kept back bordering on the glottal flow derivative The microphone was fitted on a headset so as we ensure the arrangement of the microphone with the voice over artiest constantly and relatively stable all the time .The speech was recorded and digitally stored on .wav format, in mono channel ,at 8 kHz sampling rate and 16 bit accuracy. The speech material used during the expressive speech style transformation experiments was an expressive speech corpus devoted in Marathi ,developed with a twofold purpose: first, to be used for the prosodic event detection in emotional speech and, second, for the prosody conversion of speech unit .For the corpus design, we sought the help of experts in voice over communication artiest .Recording is done in professional studio .Speech styles are defined in 6 different ways portraying basic and most used emotions in daily basis .The texts for each expressive style were read by a professional female speaker in different recording sessions (stimulated speech). The intended style was not performed according to the speaker's criteria for each sentence, but all the utterances of the same style were consecutively recorded in the same session following the previously learned pattern. Thus, the speaker was

able to keep the required expressiveness even with texts whose semantic content was not coherent with the style. Categories are defined as

a. Neutral

b. Question

c. Exclamation

d. Sad

e. Happy

f. Angry

Every emotion is portrayed by 66 sentences. Total database of 390 sentences are recorded. The sound documents are spared in wav design (16 bit mono, tested at 16 kHz).the same information base is utilized to part the sentences in word level sections for the word level examination, its about 1500 words in a corpus. Recording artist is asked to utter the sentence of one emotion category three times so as to catch exact shade of a emotion in utterance .This is done for avoiding the overlapping area in emotions to penetrate in portrayal of specific shade. For example neutral utterance is many times mix its discourse with sad utterance, like this many emotions are having overlapping shades in the expression

Table 1: Table 1 sentences 1a to 1f indicates all six emotions which are recorded for all sentences in a list

Sr.	Emotions	Content
1a	Neutral	???? ?????? ????
1b	Question	???? ?????? ????
1c	Exclamation	???? ?????? ????
1d	Sad	???? ?????? ????
1e	Happy	???? ?????? ????
1f	Angry	???? ?????? ????
2	In All Five Emotions	?? ???? ?????
3	In All Five Emotions	??????? ??????????? ????? ????
4	In All Five Emotions	????? ???????? ??? ????
5	In All Five Emotions	??????? ????? ?? ?????
6	In All Five Emotions	?? ????? ????? ???
7	In All Five Emotions	??? ??? ???? ????
8	In All Five Emotions	?? ???? ?? ???? ????
9	In All Five Emotions	???? ?? ???? ????? ???
10	In All Five Emotions	????????? ??? ??? ????
11	In All Five Emotions	???? ?????? ?????
12	In All Five Emotions	????? ?????? ?????? ???
13	In All Five Emotions	?????? ??? ???? ????
14	In All Five Emotions	??????? ???? ????? ???????? ???

Sr.	Emotions	Content
15	In All Five Emotions	?????? ??
16	In All Five Emotions	??? ????
17	In All Five Emotions	??????? ??? ???? ????
18	In All Five Emotions	??? ????? ??????
19	In All Five Emotions	??? ??????? ????? ?????
20	In All Five Emotions	??????? ??? ????????? ??????
21	In All Five Emotions	?? ???? ????? ???????
22	In All Five Emotions	??? ???? ??? ???
23	In All Five Emotions	?? ???
24	In All Five Emotions	??? ?? ??? ???
25	In All Five Emotions	???? ??????? ??????
26	In All Five Emotions	???? ?? ?? ????

1.3 Outline of the thesis

Organization of the report is as follows section 1 gives Abstract ,introduction and motivation of the work with research contribution.
Section 2 gives detail literature review of the existing work in prosody transformation and classification
section 3 describes about the text to speech synthesis processes by unit selection approach and need for the Expressive speech synthesis
section 4 describes Prosodic event detection in a speech segment for understanding the patterns of features responsible for the prosodic realization.
Section 5 describes about the emotion classification in the hierarchically upper level segment of speech that is sentence. with Novel feature coding for prosody in emotional speech .
section 6 Prosody Modification by spectral correlative prosodic mapping
section 7 concludes the research work with suggestions in future scope

Chapter 2

2 Literature Survey

2.1 Introduction

The effort of speech synthesis is to generate speech signal from the text, referred as TTS, i.e. Text to Speech Synthesis. The Schematic block diagram of a speech synthesis system is shown in Fig. 4. The text provided by user is converted into abstract linguistic depiction by the text processing module .This linguistic adaptation is acquired by performing prosodic annotations on the syntactic, semantic and lexically analyzed text (Klatt., 1987). This phonetic portrayal drives the synthesis routines to get the speech waveform of the input text. In the present work, such a framework is referred as plain Speech Synthesis (PSS) system. In prosodic i.e. expressive speech synthesis, apart from context which is a text, the intended expression indicators additionally works about as a accompanying part for text processing stage. The input context as text is converted into abstract linguistic representation as in PSS. The additional expressive information is implied beside text information and speech is synthesized utilizing the phonetic and expressive information. The discourse is integrated at first with no expressivity, that is, nonpartisan talk discourse and afterward the desired prosodic phrase is transplanted with some conversion logic (J.Tao, 2006) .The expressive speech synthesis methods are broadly classified in the following three ways (Govind, 2013).

Implicit control

Playback approach

Explicit control

2.2 Expressive speech synthesis by implicit approach

In The implicit control model is nothing but expressive speech synthesis system pedals the expressivity by interpolation among two statistical models trained on the diverse expressive databases. The expressive speech is delivered through the system by making interpolation and adaptation of HMM models which are of implicit control paradigm. HMM speech synthesis proposes the adaptation techniques to acclimatize the average style model to a specific style. A style synthesis system based on HMM using a style control vector for each style is projected (Miyanaga, (2004)) During the synthesis the style control vector associated with the target style transforms the mean vectors of the neutral HMM models. if an average model is available, the adaptation method offer flexibility to construct the statistical models even if a data of few minutes is available. The speech synthesized by the way of speaker adaptation techniques are seen to perform more vigorous as compare to speaker dependent work, these adaptation techniques can be used for synthesizing various Styles too. (Yamagishi, 2003). Speech synthesis frame works in HMM offer flexibility to integrate diverse conversation styles or emotions with HMM interpolation or multiple regression of emotion vectors (Barra-Chicote, 2010).though knowing all reward for HMM based speech synthesis systems the noteworthy shortcoming is the over-smoothing of the spectral parameters and excitation parameters (Barra-Chicote, 2010).
Abridged naturalness in the synthesized emotions is caused due to oversmoothing. though, the perceptual tests done by (Barra-Chicote, 2010)

(Jasmine Kaur, 2015), shows that the emotional speech synthesized by HMM based speech synthesis framework and unit selection base speech synthesis system offers more or less comparable expression recognition rates. The present work focuses on the development of expressive speech synthesis systems based on explicit control of prosodic features. Here the issue will be framing of prosodic rules by the analysis of each expression in the individual emotional database and incorporating them into the neutral speech

2.3 Expressive speech synthesis by playback approach

In this method, the expressive speech is synthesized separately using the individual expressive speech corpus . expressive speech synthesis is done just by playing back what is existing in the database of the target expression or by means of the models which are taught using the target expression database. The Hidden Markov Model and unit selection based systems trained on the individual expressive corpus which works on play back approach (Yamagishi, 2003) (Fernandez, 2007) (Iida, 2000) (Pitrelli, 2006).

For better naturalness in the synthesized speech (Iida, 2000), the emotional speech synthesis systems by unit selection was developed to render highly natural emotional speech. System is established by making a large corpus for every emotion. Synthesis of the intended target emotion is then done via , the individual emotional database. Which is used to select the units from it to synthesize the speech in the target emotion. A good sense of speech is produced by speech synthesizer using phrase unit selection from a very big database (Campbell, 2006).

A blended corpus by combination of various emotion databases of happy,

sad, angry as well as neutral speech for incorporating the expressive speech a s per target style (Hofer, 2005). For accomplishing this, cost function is calculated for intended target to give more penalties to pick the units when any other emotions is selected than the intended target emotion. The work done by (Fernandez, 2007) also go subsequent to similar process by integration of the units of various emotions to generate the synthesized speech in intended statistical parametric way to build speech synthesizer are using(HMM) approach is there present in addition to unit selection method .Expressive speech synthesis systems for various speaking styles like plain reading, sorrow, happy and rough is implemented through synthesized speech in the target styles using the separately trained HMM models (Yamagishi, 2003) Some of the synthesized expressive speech samples for happy and angry emotions are available for listen at the following link: http://www.iitg.ernet.in/stud/dgovind/emotionsynthesis.htm.

2.4 Expressive speech synthesis by explicit control

The Expressive Speech Synthesis (ESS) by using explicit control is performed by way of remodeling the neutral speech by way of a signal processing method in accordance to the prosodic rules framed for the target expression. According to (Schroder, 2009) by means of explicit control the expressive speech is produced by adjusting the plain speech in light of the prosodic principles got from the expressive speech database of the different expressions. The examples of explicit control expressive speech synthesis systems are one which developed on formant synthesis (Cahn, 1989) (Sendilmeier, 2000) and second is diaphone concatenation (Vroomen, 1993) (Arnott, 1995)

As the speech synthesizers using formants offers flexibility to manage the different system and source parameters, prior developments of emotional speech synthesis systems were on top of the formant speech synthesis systems (Oliveira, 2006b) (Schröder, 2001). The first endeavor to synthesize expressive speech with a formant synthesizer which is The affect editor, developed by (Cahn, 1989) (Schröder, 2001). formant synthesizer's controlling parameters are manually tuned for every emotion to synthesize the expressive speech. alteration of the control parameters for every emotion is carried out as per the different acoustic profiles discussed.
HAMLET, expressive speech synthesis system created by (Arnott, 1995), is a rule based system developed on marketable formant speech synthesis. The pitch and duration rules and voice quality rules are set in the formant synthesizer known as DECtalk In HAMLET and quality of the synthesized expressions are enhanced heuristically through manual modification. The development of these rules for emotions is as given in (Murray I.R, 1993). The goal of the perceptual experiments performed through Felix Burckhardt turned to , find out the perceptually relevant acoustic features for each emotion by systematically varying those acoustic parameters during the synthesis of the neutral utterances and discover the most appropriate values of each of the acoustic features for the emotional speech synthesis (Sendilmeier, 2000) in line with those perceptual experiments, the pitch parameters like mean pitch and pitch range, supra segmental parameters like speech rate, and voice quality parameters like phonation and vowel precisions, are located to be substantial for successfully synthesizing emotions in the formant synthesizers. The research carried out through (Vroomen, 1993)on seven emotions (neutral, joy, boredom, anger, sadness, fear, indignation) prove that only intonation and duration are

adequate to articulate emotions in the synthesized speech by means of a diaphone synthesizer. The expressive speech synthesis is done by changing the duration and pitch period by Pitch Synchronous Overlap Add (PSOLA) of the plain speech. The significant changes due to variation in fundamental frequency f0 and time duration parameter in expressive speech synthesis is also shown (Montero, 1999) in Spanish by means of diaphone concatenation. This research also showed that the role of prosody and fineness of the rendered voice depends on the expressions to be synthesized. Besides the emotional speech synthesis systems developed based on various speech synthesis approaches, some works made for converting neutral to target expressive speech by means of the explicit control approach.

(J.Tao, 2006) Achieved expressive speech conversion expressive speech conversion by using prosody (pitch and period) change of the neutral expressive speech. This paper compared linear, Gaussian mixture version (GMM) and classification and regression tree (CART) techniques for converting neutral speech to target expressive speech for mandarin language. Apart from discrete emotions like indignant, satisfied, sad and fear, the robust, medium and susceptible versions of each is also taken into consideration for synthesis. Direct scaling of sentence F0 and syllable period is carried out in linear modification model and different acoustic features of F0 contour taken into consideration for change are F0topline, F0baseline, F0avg and intensity. Prosody modification by GMM approach, intonation pattern of each syllable for each expression is considered for making pitch target models. Afterwards this pitch target parameters generated by GMM of the neutral syllable is mapped to that of the information of target expression, trees are built via linguistic information from the con text.

Listening test suggests that the GMM synthesized speech (for small data set) and CART (large data set) sounds more significant compared to linear prosody modification.

Emo Voice system developed (Cabral, 2006a) to include diverse emotions into the expressive speech for German language. Emo Voice system uses the neutral speech further converted to prosodic speech with controlling of prosody parameters (pitch, duration and intensity) and excitation source parameters (jitter, shimmer, and glottal wave parameters) by method of Pitch Synchronous Time Scaling (PSTS) (Cabral, 2006b). The rules for the prosody and voice quality modification are developed by the acoustic outline presented in (Drioli, 2003), (Zovato, 2004) . Prosodic regulations to turn out emotions in the story telling style are developed by (Theune, 2006). Story telling expressions are synthesized by varying the pitch and intensity of various part of the story like suspense, climax etc.

Explicit control method used for ESS is accomplished by changing the plain or null prosody neutral speech by a digital signal processing approach as indicated by the prosodic rules confined for the intended expression. The issues in developing ESS with the method of explicit control approach are accompanying below: ∗ Synthesizing a superior value neutral speech

bullet Study and assessment of expressive features

bullet Integration of expressive features

The synthesized speech either coming from a unit selection concatenative system or HMM based statistical parametric system is of fine intelligibility and convincingly natural. Along these lines any of the two frameworks can be utilized as the neutral speech synthesizer for the present work. The study and evaluation of expression specific parameters of different

emotions are performed on an expressive database. Typical parameters for each expression are examined concerning the neutral expression. At this phase, the concern is of the exact estimation of features to analyze its variation across different expressions. To achieve this selection of digital signal processing tools that precisely calculate the expressive parameters are important for analyzing the expressive parameters. Finally, the result of this study is able to give a set of rules defining parameters related to expressivity. They can further used modify the parameters of neutral speech to synthesize the expressive speech. The last stage in the ESS by explicit control is the consolidation of the rules for every emotion on the parameters of plain neutral speech to obtain the speech in the intended target expression. This task is accomplished by a digital signal processing technique. The issue in consolidating these expressivity contribution rules is to present slightest perceptual distortion without influencing naturalness in the synthesis of expressions in speech discourse. Segment 6 audits different strategies for fusing expressive parameters.

In the process of speech coding, speech transformation is an emerging area. Speech transformations have different criticalness of utilization going from introduction of expression to scientific discovery to crime investigation. The need of future speech synthesis lies on the not simply recognition but rather to approve its substance and apply to various critical applications. The fields of non invasive speech related detection in medical diagnosis, authentication applications, cinematography are few such applications. Wherein speech is an effective way of recognizing user, the spoken format rely its significance. Speech with its emotion reveals the personal objective of the speech. It is hence required not only to know the speech spoken but also the emotion at which it is spoken as. The process

of emotion synthesis was performed using prosody modeling. Wherein prosody model are used to detect spoken speech emotion However, the processing accuracy of the transformation from one emotion to other is dependent on the prosody features. Hence it is required to derive the prosody model more accurately so as to achieve proper transformation. Wherein various approaches were developed to achieve the objective of proper extraction of prosody feature information from the speech signal, and mapping for synthetic transformation, the effect of processing noise, environmental distortion or magnitude variations lead to wrong transformation of the speech signal. Hence it is required to have a accurate feature extraction and transformation for emotion synthesis in speech coding. To observe the performance of the proposing approach, the proposing approach is to be evaluated over variant conditions such as noises, magnitude variation, gender variation, age variation for different emotion transformation. Towards such transformation and feature extraction, various developments were made. Following table highlights the pros and cons of various prosody modeling technique for Text-to-Speech Synthesis Systems (Maheswari, 2012)Table 2 from (Schroder, 2009) summarizes the approaches to parameterization of expressive speech as found in current research on speech synthesis The approach of feature extraction and its usage to classification and transformation is summarized in the following section

Table 2: Current approaches to the parameterization of expressivity in state-of-the-art speech synthesis

	Unit selection -Selection of units	Unit selection - Signal modification	HMM-based synthesis
Explicit acoustic models	hand-crafted prosody rules as targets (Campbell, Beijing, China)	PSOLA with explicit rules (Zovato, 2004) explicit control of glottal spectrum (d'Alessandro, 2003) brute-force estimation of glottal source parameters (Vincent, 2005) (potentially usable for modification)	
Playback	recording separate expressive voice databases (Iida, n.d.) manually labelled symbolic targets (Fernandez, 2007)		train models on style-specific speech data (Yamagishi, 2003) adapt models to style-specific speech data (Yamagishi, 2007)
Implicit acoustic models	HMM models as targets (Campbell, Beijing, China) automatically trained symbolic targets (Fernandez, 2007),	voice conversion between emotional recordings of same speaker (Matsui, 2003) interpolation between recorded or converted voices (Turk, 2005)	interpolate between style specific models (Miyanaga, (2004))

2.5 State of the art in Emotion recognition and Feature extraction

Wherein the process of prosody detection and processing for speech coding is design around the extraction of domain speech features and converting to a specified emotion, the process of recognition, classification, and feature extraction are most important towards its processing accurateness. In the processes of recognition of emotion or its transformation different developments were noted as follows.

2.5.1 Emotion Recognition

For the progress of recognition of prosody for speech transformation in (Jasmine Kaur, 2015) expressive speech synthesis, a review is presented. The review sketches the bottom models for speech conversion for expressive speech transformation (EST) for human computer interaction (HCI). In (Bjorn Granstrom, 2005) an animated format of the prosody feature processing for speech transformation is presented. For a audio-visual representation for expressive speech communication. A realistic speech behavior for speech coding in this application was demonstrated.

In (Pablo Daniel Aguero, 2006.) a multi lingual speech transformation based on prosodic feature to develop a speech-to-speech translation. The prosodic features are derived using unsupervised clustering algorithm and mapped for the target speech transformation based on the targeted speech quality. In (Abhishek Jaywant, 2012.) (Dharaskar, 2010), a cross-modal priming task was employed to perform the categorization of emotions from a speech signal. In this approach, after listening to angry, sad, disgusted, or neutral vocal primes, subjects rendered a facial affect decision about an emotionally congruent or incongruent face target. In (P.tzinger, 2008), a study was performed to investigate the effects of emotion class, sentence type on the speech amplitude and speaker. This approach also studied the recording technique to maintain dynamic information for a complete full blown speech. The results of this approach revealed that the speaker and emotion class are highly significant and later reveals the half of variance.

In (Laba kr. Thakuria, 2014), a hybrid approach was proposed by combining the template parametric manipulation with prosody parametric

manipulation for the purpose of quality improving and also to generate prosody for bodo speech. This approach also aimed to increase the annotation variability of synthesized speech output. This approach also considered prosodic features for emotion detection from speech signal. The extracted prosody is the combination of intensity, pitch and duration. This approach considered anger, happiness, fear and sadness for synthesis evaluation. Through this evaluation this approach found the accurate prosody of bodo speech to confirm perception tests.

In (Klara Vicsi, 2012.), a short introduction was given about speech emotion recognition. This also gives the information about prosody features, speech emotion and speech style and also more other information was also involved. The main uncertainties of speech recognition were also outlined in this.

In (Edmondo Trentin, 2014) a novel network called as probabilistic echo state network (pi-SEN) was proposed to find the density over a variable having length sequences and also multivariate domain vectors. This pi-ESN was formed by combining the parametric density based radial basis function and reservoir fan ESN. At classification stage, this approach used maximum likelihood training process. The feature parameters used for prosody representation are effective to such recognition process.

2.5.2 Feature Description

A technique for changing the pitch and duration of a speech signal based on time-scaling the linear prediction (LP) residual In a time domain analysis for speech transformation based on given input signal (Cabral, 2006b)is presented. The pitch alteration approach was given the procedure of time

based coding for speech signal utilizing linear prediction logic. The linear prediction residual is utilized as a pitch variation parameter for speech prosody change. The procedure of linear prediction additionally helps fit as a altering the shape content of the speech signal.

In (Martin Borchert and Antje Diisterhoft, 2005)new quality features in particular formants, spectral energy distribution in different frequency bands, harmonics-to-noise ratio (in different frequency bands) and irregularities (jitter, shimmer). The procedure of Dimensional approach for emotion classify is completed. It is demonstrated that these quality features are more suitable in portrayal in contrast with various valance levels in dimensional approach in (Rigoulot, 2014), an approach was proposed to concentrate the impact of emotional speech prosody on the participants and fixate features that congruent of an emotional speech of prosody. In this approach, absolutely, 21 members are watched for face expression such as misery, fear, joy while tuning in to en candidly articulated expression talked in an incongruent or consistent prosody. The all members attempted to judge the significance of the voice or face of emotion whether it is same or not. In this review, it was affirmed that the eye movements will play an important role to match this. In (Samantaray, 2015) (Rahul. B. Lanjewar, 2013),

A new approach is suggested with a amalgamation of prosody features (i.e. energy, pitch and Zero crossing rate), derived features (i.e. Linear Predictive Coding Coefficients (LPCC), Mel-Frequency Cepstral Coefficient MFCC)), quality features (i.e. .Spectral features, Formant Frequencies), and dynamic feature (Mel-Energy spectrum dynamic Coefficients (MEDC)) for an proficient automatic recognition of speaker's status of emotion. Multi Support Vector Machine(MSVM) was utilized at

classifier stage to classify the expressions as happy, neutral, fear, sad and anger for a given speech signal. In (Lupu, 2011.), an approach was proposed to find seven types of emotions of a speech signal. They are, fear, anger, sadness, boredom, neutral, disgust and happiness. Various DWT decompositions are used for feature extraction. SVM was used for classification.In (Sudhkar, 2015) a emotion detecting framework was focused to design that performs various actions like feature extraction, speech to text conversion, feature selection and feature classification those are required for emotion detection. This approach uses prosody feature for emotion detection. The classification involves the training of various models and testing of a query sample. This approach extracts the features such that, they will convey measurable level of emotional modulation.

2.5.3 Vocal tract features and source Excitation features

Picking appropriate components for building up any of the speech frameworks is a important decision. The components are to be picked to accurately describe about intended information. Diverse speech parameters represent different information about the speech segment expression, speech discourse, speaker etc) in exceptionally overlap way. In this way speech area of research work around three significant speech feature areas that is: source excitation, vocal tract transfer system, and prosody features Speech features resulting from glottal source signal are known as source features. Excitation (glottal) source signal parameters are acquire from speech sample, after suppressing vocal tract (VT) characteristics. (Rao, 2012) Review and analysis of different speech parameters about excitation source are enlisted in table 3.below

Table 3: Review and analysis of different speech parameters about excitation source

	Features	Purpose and approach /
1	LP residual energy	Vowel and speaker recognition[Wakita et al (1976)]
2	LP residual	Detection of instants of significant excitation[Rao et al. (2007b)]
3	Higher order relations among LP residual Samples	Categorizing audio documents[Bajpai and Yegnanarayana(2004)]
4	LP residual Environment	Speech enhancement in multi speaker[Yegnanarayana et al. (2009)]
5	LP residual	Characterizing loudness [Bapineedu et al. (2009)]
6	Glottal excitation	Analyzing the relation between [Cummings and Clements (1995)] emotional state of the speaker and glottal activity
7	Glottal excitation	analyze emotion related disorders [Cummings and Clement (1995)]
8	Excitation source	discriminate emotions in continuous speech [Hua et al. (2005)]

Table 4: Review and analysis of different speech parameters about vocal tract transfer function

	Features	Purpose and approach / Ref.
1	MFCC features	Discrimination of speech and music. Higher order MFCCs contain more music specific information and lower number of MFCCs contain more speech specific information. Mubarak et al. (2005)
2	MFCCs, LPCCs RASTA,PLP coefficients	Log frequency power coefficients Classification of 4 emotions in Mandarin language. Anger, happy, neutral and sad emotions are considered in this study.Pao et al. (2005, 2007)
3	Combination of MFCCs and MFCC low features	Emotion classification using Swedish and English emotional speech databases.Neiberg et al. (2006)
4	Spectral features Fourier and Chirp transformations	Emotion classification EnglishLDCandEmo-DB Cnsonant,stressed an databases. Bitouk et al. (2010)

2.5.4 Classification and Coding

In (Anand C1, 2015) (Prashant Aher, 2014), human speech emotion recognition system was developed based on the acoustic features like pitch, energy etc, and spectral feature MFCC. Then SVM and CART has been used as classifier. The complete implementation of proposed approach was

done under two phases: training and testing. This approach was tested over different voice files of different emotions like Anger, Disgust, Fear, Happy, Neutral, Sad and Surprise.

(Odetunji A. odejobi, 2008) Presented the approach of speech transformation using hidden markov model, for phonotic unit detection corresponding to the training dataset. The model uses the dynamic property of HMM for speech recognition. A MFCC based feature extraction is used for the speech feature detection and HMM modeling is used toward decision deriving for emotion detection

(Ryo Aihara, 2012) Presents an approach for text to-speech transformation using prosody model in context to the tone languages. A fuzzy based modeling for speech coding is proposed. A classification and regression model tree (CART) based on testing and modeling of speech duration is proposed. A tree modeling for speech detection to observe the syllable in word and its position in sentence was made. The lengths of the word, peak of the targeted syllable, the phonetic structure of the syllable were also derived for speech coding.

In (Chung-Hsien Wu, 2010) a Gaussian mixture model (GMM) based on emotional voice transformation using the conversion of prosody feature was proposed. The GMM based modeling is used for the conversion of non-linguistic information, while keeping the linguistic data intact. The objective of such conversion was to keep the prosody processing with greater level of speech quality.

An approach of neural network based speech synthesis system was presented in (Rao, 2011). The approach of neural network is carried out for the prosodic feature of the syllables for their position, context and phonological feature. The prosody model developed were evaluated for the text

to speech transformation, used for speech recognition, speech synthesis and speaker recognition. The NN model is developed for the capturing of dialects in Hindi language. Towards database optimization in (Schroder, 2010) (Bethel, 2015) (K.C. Rajeswari, 2014) (P. Gangamohan, 2012) an approach to minimize the dataset collection and processing effort were developed focusing on maintaining acceptable quality and naturalness for speech transformation. A text-to-speech (TTS) framework MARY was developed with voice quality using GMM-based prediction and vocal tract processing.

In (Hirschberg., 2006) (Jianhua Tao, 2006), an approach made an attempt to synthesize emotional speech by using 'strong','medium',and 'weak' classifications. In this approach, various linear models were processed for test like Gaussian Mixture Model (GMM), Linear Modification Model (LMM) and a classification and Regression Tree (CART). This paper also analyzes objective and subjective analyses. (Xiaoyong Lu, 2014) Focused on happiness. This paper computes the emotional speech for happiness for a given speech signal. In this approach, totally 11 types of emotional utterances were developed; each and every utterance was labeled with a PAD value. This paper also proposed a five-scale tone model to model the contour of the pitch. A generalized regression neural network (GRNN) situated prosody conversion model is constructed to recognize the transformation of pitch contour,duration and pause duration of emotional utterance, where the PAD values of emotion and context parameter are adopted to predict the prosodic aspects. Emotional utterance is then re-synthesized with the STRAIGHT algorithm by way of modifying pitch contour, length and pause period.

In (Juan Pablo Arias, 2014), a novel speech synthesis approach was pro-

posed using the subspace constraint. This approach employs Principal Component Analysis (PCA) to reduce the prosodic components dimensions. Then the obtained samples of PCA are trained. The features included in this approach are mainly F0, speech length, power and finally the correlative length. This approach also assumed that the combination of accent type and syllables will determine the correlative dynamics of prosody. This approach successfully worked for synthesized speeches especially, 'disgust', 'boredom', 'joy', 'anger', 'surprise', 'sorrow' and 'depression'. In (Shulan Xia, 2014), an approach was proposed with the aim of speech enhancement and also with the aim of hearing impaired patients. This approach didn't require more training data. This approach also studies the conversion of prosodic samples. This approach used Eigen voice-Gaussian Mixture Model (EV-GMM) to transform the spectral parameters and F0

(Pinheiro, 2015) investigated the effect of musical train on the observation of vocally expressed emotion. This approach takes the base of Event-related Potential correlates for emotional prosody processing. 14 control subjects and 14 emotional musicians are listened to 228 sentences with intelligible semantic content, neutral semantic content, differing in prosody and unintelligible content. This investigation observes that the P50 amplitude was reduced. A difference between SCC and percent stipulations used to be discovered in P50 and N100 amplitude in non-musicians only, and in P200 amplitude in musicians simplest. (Pribil, 2009)represents of micro intonation and spectral characteristics in female and male acted (Pinheiro, 2015)emotional speech. Micro intonation element of speech melody is analyzed related to its spectral and statistical characteristics. Consistent with psychological study of emotional speech, exclusive feel-

ings are accompanied by different spectral noise. We manipulate its amount by using spectral flatness in step with which the excessive frequency noise is mixed in voiced frames throughout cepstral speech synthesis.
In (Gurunath Reddy M, 2015), an emotional speech synthesis framework was proposed to analyze the effect of emotions in the story telling speech. This approach used XML and SABLE languages to synthesize the emotions of text. The SABLE language was used for the purpose of speech quality improvement from the contaminative speech synthesizer. This approach used a set of tags of prosody to synthesize the emotion of the speech from a given text. The prosody correlates of pitch range, pith base and intensity was found by Modified Zero frequency filtered (ZFF) signal. Then the required prosody parameters are stored in a template format. At synthesis stage, prosody tags were replaced by hand annotated text story. The naturalness and quality of the synthesized emotional speech was analyzed through subjective tests.
(Ververidis, 2006) Proposed a new emotional recognition approach with the aim of three main goals. The first one is to update the up to date available emotion speech data collections. Next, goal is to extract the features efficiently such that they are able to find emotion more accurately and also to measure their effect. This approach considered the features as vocal tract cross-section areas, Teager energy operator-ased features, pitch, the mel-frequency cepstral coefficients and the speech rate. The third and last goal is to analyze the techniques for the classification of emotional states. These techniques include, HMM model, ANN mode, LDA model, K-NN model and finally SVM model. In (Klara Vicsi, 2012.), a speech recognition and understanding approach was proposed to model the super segmental characteristics of speech and also for acoustic processing

by including the advanced semantic and syntactic level processing. The proposed approach was completely based on HMM modeling. The HMM is used in this paper to model the speech prosody and to make the initial semantic level processing of input speech. The energy and fundamental frequency were used as acoustic features. An approach was also applied for semantic features extraction. The method was designed to work for fixed-stress languages, and it yields a segmentation of the input speech for syntactically linked word agencies, or even single phrases akin to a syntactic unit (these word companies are regularly referred to as phonological phrases in psycholinguistics, which can include a number of words). These so-referred to as phrase-stress items are marked by means of prosody, and have an associated fundamental frequency and/or energy contour which enables their discovery. This semantic level processing of speech was investigated for the Hungarian and for the German languages.

2.5.5 Medical Processing

In (Dominik R.Bach, 2013) an application of speech prosody is represented ,where a medical diagnosis of unpaired discrimination for different speech emotion were presented. The effect of speech emotion on the detection of medical disorder is outlined. The presented task suggests the usage of speech synthesis on fear detection. In (Swann Pichon, 2013) (Thomas Ethofer, 2008) a generation of expressions on speech coding based on human neural analysis is presented. This approach presents a unique approach of neural based emotional prosody analysis in human speech coding. The bilateral operation of human brain on the prosody modeling and its operations are illustrated

prosody modeling and its operations are illustrated. In (Filomena Castagna, 2013.) (Jia Huanga, 2008) to evaluate the effect of schizophrenia in patients for prosody future extraction, a feature extraction process for emotion feature extraction process is carried out. 94 patient and 51 healthy subjects were taken for the evaluation of the suggested approach. The emotion perception disorder due to the disease patient was analyzed
In (Pell, 2006),an investigation was done towards the hemispheric contributions for emotional speech processing by comparing dults with a focal lesion of left and right hemisphere and also with brain damage. Participants listened to semantically nomalous utterances in three conditions (discrimination, identification, and score) which assessed their attention of 5 prosodic feelings beneath the effect of one of a kind mission- and response-choice demand. Findings printed that correct- and left hemispheric lesions had been associated with impaired comprehension of prosody, even though possibly for distinct motives: correct-hemisphere compromise produced a more pervasive insensitivity to emotive points of prosodic stimuli, whereas left-hemisphere harm yielded better difficulties decoding prosodic representations as a code embedded with language content.
(Yingying Gao, 2016) Used event-related brain potentials(ERPs) to evaluate the time path of emotion processing from on-linguistic vocalizations versus speech prosody, to scan whether vocalizations are dealt with preferentially via the neuron cognitive method. Members passively listened to vocalizations or pseudo-utterances conveying anger, sadness, or happiness because the EEG used to be recorded. Simultaneous effects of vocal expression style and emotion were analyzed for 3 ERP add-ons (N100, P200, late confident factor). Emotional state of a speaker is accompanied with the aid of physiological alterations affecting breathing, phonation,

and articulation. These changes are manifested on the whole in prosodic patterns of F0, energy, and duration, but also in segmental characteristics of speech spectrum. For that reason, a new emotional speech synthesis process proposed in (Pibilová, 2009) is supplemented with spectrum amendment. It includes nonlinear frequency scale transformation of speech spectral envelope, filtering for emphasizing low or excessive frequency range, and controlling of spectral noise via spectral flatness measure in keeping with skills of psychological and phonetic study. The Aprosodia Battery was once developed to distinguish unique patterns of affective-prosodic deficits in sufferers with left versus correct mind injury by way of using affective utterances with incrementally reduced verbal-articulatory demands. A huge, quantitative error evaluation utilizing previous outcome from the Aprosodia (Franco Orsucci, 2013)in patients with left and right brain harm, age-identical controls (historical adults), and a gaggle of younger adults. This inductive evaluation was performed to deal with three foremost disorders in the literature: (1) sex and (2) maturational getting older results in comprehending affective prosody and (3) differential hemispheric lateralization of emotions. This approach found no overall sex effects for comprehension of affective prosody. There have been, nonetheless, scattered results involving a specific have an impact on, suggesting that these variations have been concerning cognitive appraisal alternatively than most important perception. In (Wang, 2015), an emotion recognition approach was proposed especially for children with HFA (n=26, 6-11 years) based the prosody features of them. In this approach, Vineland adaptive behavior scale and communication checklist of children was used to assess social and pragmatic abilities

2.5.6 Language Coding

In (S. S. Agrawa1, 2010) towards a speech transformation based on the effect of emotion transformation in Hindi speech processing a Hindi speech database is developed. Pitch feature were derived for different emotion and used as a transformation metric for speech transformation. Three tests transformation for sad, joy and anger were developed for the transformation from neutral speech signal.

In (Ravi, 2014) a speech transformation process for kannada language was developed.An approach of linear modification model (LMM) was used. This method is used for the conversion of emotion data to its target emotion. A kannada dataset is been created and the effect of change in emotion is evaluated. The effect of pitch on the speech transformation is carried out over the pitch points extracted. The process was carried out over sadnes and fear cases. In [65,66] an approach to speech synthesis for Czech and Slovak emotion speech based on spectral coding for prosodic feature is developed. A Gaussian mixture model (GMM) is used for processing speech classification using GMM training rocess. The selection process of feature and its impact on emotion classification were analyzed. The functionality of a 2 level architecture comprising of gender detection and classification is also developed for speech transformation. The length of the speech signal was varied for different dimension and the effects of speech transformation were observed. The classification accuracy for the developed system is evaluated for detecting the effectiveness of the speech transformation approach.

In the process of speech translation a speech synthesis model for text-to-speech for Marathi language was presented in (Manjare, 2014) . The pitch features were focused based on its magnitude, count and contour used for

transformation. The pitch factor is derived for speech with punctuation marks and processed to derive prosodic features ,the approach of neutral Marathi speech to emotional speech is presented based on pitch modification and word detection is presented. In (Shashidhar G. Koolagudi, 2011), Hindi speech corpus signals have been used for simulated emotion detection. The data base was collected from the Gyanawani FM radio station, Varanasi, India from professional artists. The complete speech corpus for eight emotions, are anger, disgust, fear, happy, neutral, sad, sarcastic and surprise. Emotion classification is performed on the proposed corpus using prosodic and spectral features. Energy, pitch and duration are used to represent prosody information. Mel frequency cepstral coefficients (MFCCs) are used to represent spectral information.

(Schroder, 2003) Proposed an approach to perform the German text to speech synthesis. This approach mainly used XML for internal data representation of system . This approach also provides an interface for users to modify intermediate processing steps without any need of understanding of system technically. This approach also provides a detailed description through examples.

A practical approachto generate the speech of Punjabi language (Kaur, 2014). In Punjabi language, there are so many discontinuities. This work proposed to increase the utterances of speech by overcoming the problem of discontinuity in order to increase the naturalness. This approach was accomplished to increase the quality, natural resound of speech and simulated over various Punjabi audio files.

In (RashmiVerma, 2015) a new prosody rule set was designed to convert the neutral speech to storytelling speech in Hindi language. This approach considered the features as pitch, duration, intensity and tempo to perform

the conversion. For each and every prosodic parameters mentioned above, specific prosodic rules were built to denote the story teller emotions such as anger, sad, fear, neutral and surprise. For this purpose, some professional story tellers were gone to consult. The complete rules are derived both from male and female speakers. With the developments observed in past works, the speech synthesis for transformation are basically been carried out, using prosody parameter extraction and mapping. Wherein most of the works are focused towards development of transformation approaches for feature extraction, emotion recognition, and its transformation, less concentration is given towards the data content. During the process of speech coding it is observed that system noises such as hiss sound, or magnitude difference due to difference in capturing elements or variation in speech data varying with sampling rates, period etc. Though these content impact the speech transformation process less focus is made in this area. With this objective, a system design for robust speech transformation logic is proposed, as outlined in next section.

2.6 System outline

To develop a robust transformation system following prosodic features for emotion transformation, system architecture is outlined. Figure 1 illustrates the block diagram for the proposed speech transformation approach

The process of speech transformation is carried out in two phases, of training and testing. In the process of training, a set of speech utterance for multiple subjects with different emotion will be captured. These samples are recorded with maximum standard environments to obtain highest

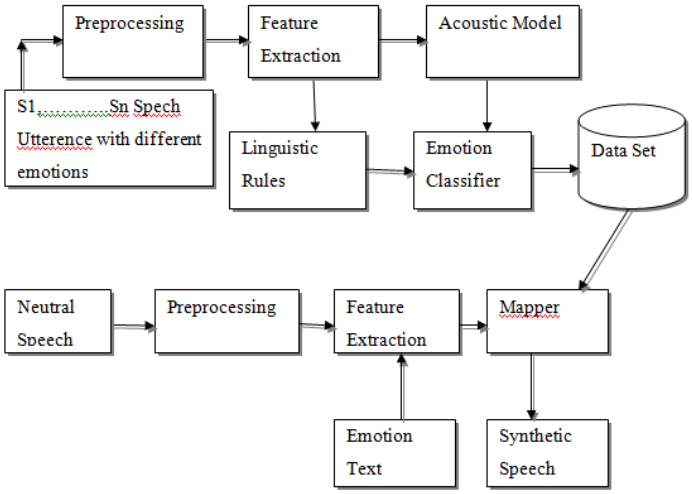

Figure 1: System Architecture for proposed emotion speech transformation

degree of speech clarity in recorded speech. These samples will then be passed for preprocessing, wherein the samples will be linearized to a uniform sampling rate, period and filtration. The windowing technique will be applied and energy spectral coding will be developed towards filtration process. To the processed speech signal, prosody features will be extracted, namely the jitter, shimmer, Harmonic to noise ratio (HNR),degree of voice breaks (DVB), degree of voiceless (DUV), pitch mean, pitch contour, pitch peaks, and pitch variance. These extractions of feature will be performed over the entire training sample and a acoustic model with voice quality feature is developed. The linguistic rules will be extracted from the speech signal, for pause, breaks, and pronounce for each user. These linguistic rules are then passed to a emotion classifier where the emotion classification is performed based on SVM approach. The classified emotions with their corresponding features are then recorded to formulate feature dataset.

During the test process, a neutral sample is passed, which is processed with the same process as carried out during training. The pre-processed sample is then processed for feature extraction and passed to a mapper logic where in based on the text input from the user prosodic feature will be derived from the dataset. The mapper logic transforms the neutral feature to the demanded emotion feature by magnitude and spectral alignment to obtain the transformed synthetic speech.

2.7 Conclusions

This work represents literature review outline in the area of expressive speech synthesis and prosody modeling for emotional speech transformation. Building an artifact which clearly produces a speech message and moreover producing it like human being is the main aim of TTS,called as intelligibility and naturalness. TTS is a superior blend of engineering and science. Researchers builds the models which will give insight on speech and language abilities of humans. There are many procedures for producing the speech signal. There are two ways in synthesis technology first is bottom-up and second is concatenative. In the first bottom-up approach, a speech is generated from scratch, by means using the understanding of how the human speech production organism functions. a basic signal is artificially created and we modify it later, In the very same way that the larynx generate a basic voice signal which is then modified by the mouth in real human speech. In the second approach, concatenative type , recording of some real speech is done, cutting it into small pieces, and then patching it together to form new segment. Concatenative speech synthesis techniques are referred as not ”real” sometimes, in a way that the signal is not generated from scratch,it’s a just patch work. This may or

may not be relevant, but it turns out that, at present concatenative techniques far out perform other techniques, and for this reason concatenative techniques currently dominate

The use of emotions or simply prosody is made by humans to convey the actual meaning or the intelligence of the context. It is evident while listening to person who is speaking in a language not known to us ,In that case expressivity factor quite useful in getting the meanings. The speech coding for emotion recognition and its representation is also reviewed and presented. This chapter has tried to outline the progress in emotional speech synthesis ever since its beginnings, with formant synthesis, by early concatenative systems Relying on explicit models of expressive prosodic speech and 'playback' models, to recent challenges in this domain, which all look to deal with the same issue from diverse angles: how to merge a high level of naturalness with a right level of control over the expression The dominating features for speech signal representation, which have higher impact on the speech transformation, are reviewed and their significance is outlined. Most of the transformation or classifying process are performed using HMM and SVM. In few contexts Advances in Neural networks also suggests for the usage of deep neural network and fuzzy logics for comparable results compared to previous classification techniques used . However as observed, two main streams dealing with prosody transformations are observed .One is using Natural Language Processing NLP. Which gives emphasis on linguistic structures of the Context and processes the prosody transplantation. However these techniques are not able to deliver the good quality output as comprehensive linguistic rule coding is a challenging task. The other side gives emphasis on speech content for transformation process by explicit control. Where sig-

nal processing techniques are utilized over the plain synthesized speech. Conversion algorithm for the prosody transplantation works on two levels , one is on linguistic structure and other is on feature structure. linguistic structure forms verbal communication. Verbal language is seen as a mixture of two systems; one that makes words from phonemes, and another that makes sentences from words. These algorithms considers phonemes or words or sentence as a basic unit for changing the prosody from source emotion to target emotion. It is observed that phoneme level prosody conversion gives good result, compared to word and sentence levels. Second fold of prosody conversion is changing the features of source emotion to realist the target emotion. In this section it is observed that modifying prosody and spectral envelope is considered. When only prosody features are mapped from source to target emotion, results were not that promising and natural which reflects intelligence of the context. When spectral parameter tuning is done along with prosodic parameter variations, then the quality of the transformation is very very good. Conversion algorithm for the prosody transplantation is proposed in this work which will change the one type of emotion of the speech signal to other. We have not considered the linguistic structure and rule base approach in our work though it also affects the transformation quality toward making it general purpose and all comprehensive, in consideration to this system architecture is proposed for robust emotion speech transformation.

The speech coding for emotion recognition and its representation is studied. The dominating extracting features for speech signals, which have higher impact on the speech transformation, are derived. Most of the transformation or classifying process are performed using HMM. Wherein advance intelligence logic such as neural network and fuzzy logics were

also used in few context. However as observed, less emphasis is given on speech content for transformation process. Though these factors also affects the transformation quality, in consideration to this system architecture is proposed for robust emotion speech transformation.

Chapter 3

3 Speech Synthesis

Speech synthesis is the synthetic construction of speech. Speech synthesis initiated from the speaking machine in 1791 (Kempelen, 1791) which was a mechanical system that imitate the physical production of speech. Many enhancements of the original speech synthesizer has taken place by the researchers in the nineteenth century, concluding in the appearance in 1846 of The Euphonia, or Speaking Automaton (faber, Aug 1846). The initial phase of current speech synthesis flourished in the first half of the twentieth century with the advance developments in artificially modeled electrical articulatory and formant speech synthesis systems (Dudley, 1939). Speech synthesis needed human manual interference for managing the output sound production and the deviation within the acoustic speech parameters. Contemporary speech synthesis can be explained as a system which matches to the automatic dealing of context and the advancement of Text-To-Speech (TTS) synthesis systems. The concept of a Text-To Speech synthesis system is to process given text as a context and produce the acoustic speech parameters of a utterance. Text-to-Speech start off with diaphone synthesis, and current Text-to-Speech (TTS) synthesis systems are build with articulatory, formant, unit selection and parametric synthesis systems. TTS synthesis burst out in previous few decades due to emerging corpus-based systems that offer intelligent and human like natural speech. Contemporary speech synthesis systems swathe a extensive variety of use in telecommunications, multimedia, and inventive fields like the creation of simulated avatars, speech interactive systems to

the capability to rebuild the sound of deceased individuals, and the power to craft artificial languages.

3.1 Unit selection Text to speech synthesis

Speech synthesis is the way toward changing over info content into equal discourse frame. Expressive Speech synthesis manages the Speech synthesis alongside including expressivity and talking styles into it. The way of Speech waveform reflects diverse passionate states. Operations on Speech waveform by utilizing signal processing strategies drives it to the expressive Speech. Synthetic voice created from such a framework with included expressivity sounds so regular that, it is hard to recognize from human discourseText To-Speech synthesis (TTS) synthesis frameworks let us change over text into speech discourse waveform and later emotions can be included into it.

The front end text to speech synthesis system serves as the Neutral Speech Synthesis. The parameters of the neutral speech discourse synthesized by the TTS framework are changed by target expression to produce the speech in the target expression. Each TTS has a front end text preparing block, which changes over the text to be synthesized to an abstract linguistic specifications. These theoretical semantic determinations could be a grouping of phonemes or any sub-word unit and furthermore it could be commented on with the prosodic information (Klatt 1987;Clark et al. 2007; King 2011). Text processing stage for the most part incorporates the text standardization and lexical analysis modules. The text processing module is to give an interesting logical portrayal about the sound units over the whole articulation. This abstract linguistic representation drives the waveform module to synthesize the as per the content given. For the

waveform generation from the abstract linguistic representation, there are four methodologies to be specific,

- Articulatory speech synthesis
- Formant speech synthesis
- Concatenative speech synthesis
- Statistical parametric speech synthesis

3.2 Issues in expressive speech synthesis by explicit control

The ESS by explicit control is accomplished by changing the neutral speech by a signal handling approach as indicated by the prosodic rules made for the target expression. The different issues in the ESS by explicit control approach are as below: • Synthesizing a good quality neutral speech • Analysis and estimation of expressive parameters • Incorporation of expressive parameters The a variety of issue and approaches for the development of neutral speech synthesizers are presented here. Based on this review, the speech is synthesized from a unit selection concatenative system of good and reasonably natural. Therefore this system can be used as the neutral speech synthesizer for the present work.2

3.3 Expressive Speech Synthesis Applications

There are many ESS applications. Some of these are story telling application for children, helping hand to dumb persons, dynamic announcements on public places instead of limited domain, good Human Machine Interface (HMI), video games etc. These applications are summarized in

Figure 2: Applications of Expressive Speech Synthesis

Figure 1.image source https://www-expression.irisa.fr/research/expressive-speech-synthesis/

Stephen Hawkins uses TTS for English. In future, ESS can be used to read e-mails, messages, movie dubbing etc.

3.4 Speech production

Speech is the vocalization form of human communication. vocal cords vibrate when the flow of air being expelled from lungs changes during phonation,. vocal cords to open and close due to The air-pressure from lungs. Quasi-periodic vibration of glottis restrict the air passing through the remainder of the vocal tract to form a pulse train. The vocal folds give shape to the pulse train by amplitude modulation. source-filter model is shown for explaining the process of human speech production in Figure 3

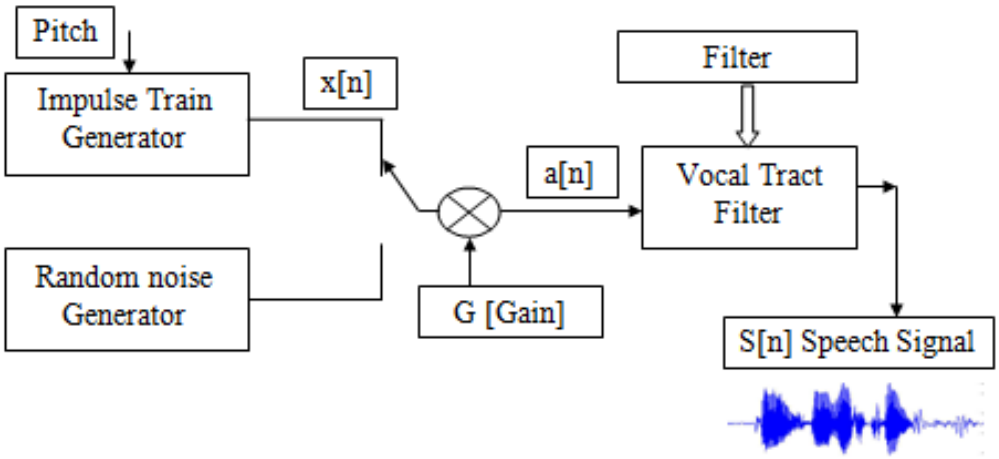

Figure 3: Figure 3:Source Filter model of human speech

In the field of speech synthesis as well as speech analysis The source-filter model is used .The development of the model is due, in large part, to the early work of Gunnar Fant, although others, notably Ken Stevens, have also contributed substantially to the models underlying acoustic analysis of speech and speech synthesis.In implementation of the source-filter model of speech production, the sound source, or excitation signal, is often modeled as a periodic impulse train, for voiced speech, or white noise for unvoiced speech. The vocal tract filter is, in the simplest case, approximated by an all-pole filter, where the coefficients are obtained by performing linear prediction to minimize the mean-squared error in the speech signal to be reproduced. Convolution of the excitation signal with the filter response then produces the synthesized speech

3.5 System Architecture of TTS: Festival Speech Synthesis System

After a study of Marathi language and exploring through different speech synthesizers, it is found that, we cannot use foreign language synthesizer for Marathi. We need to build it on our own. This is done using Festvox also known as Festival Speech Synthesis System

The Festival Speech Synthesis System is a general multi-lingual speech synthesis system originally developed by Alan W. Black at Centre for Speech Technology Research (CSTR) at the University of Edinburgh. Substantial contributions have also been provided by Carnegie Mellon University and other sites. It is distributed under a free software license similar to the BSD License. (Wikipedia, 26 April 2017)

It offers a full text to speech system with various APIs, as well as an environment for development and research of speech synthesis techniques. It is written in C++with a Scheme-like command interpreter for general customization and extension. (Black, n.d.)

Festival is designed to support multiple languages, and comes with support for English (British and American pronunciation), Welsh, and Spanish. Voice packages exist for several other languages, such as Castilian Spanish, Czech, Finnish, Hindi, Italian, Marathi, Polish, Russian and Telugu. (Wikipedia, 26 April 2017) The system architecture is shown in figure4

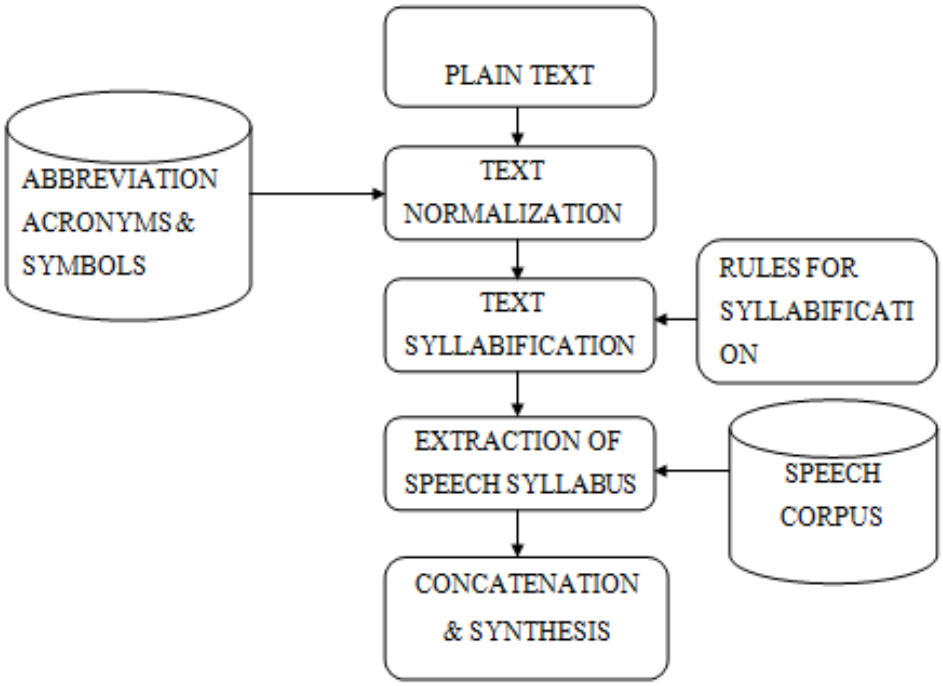

Figure 4: Figure 4 Marathi TTS architecture

Festival is considered as a speech synthesis system for at least three levels of user. First, those who wish to have good quality speech from random text by less effort. Second, people who are working on development of language based system and expects to have context as synthetic output . In such requirements some tailored adjustments are required, like diverse voices, explicit word choice, dialog types etc The third stage is in mounting and testing new synthesis methods (Black, December 2, 2014).

3.6 Set-up for TTS

This is one of the major parts in TTS. Several techniques are available for speech synthesis: articulatory synthesis, formant synthesis, and Concatenative synthesis. The techniques described in Festival are Concatenative synthesis. Concatenative synthesis techniques not only give the most natural sounding speech synthesis. Two techniques are available in Concatenative synthesis: diaphone, unit selection. We used unit selection technique for waveform synthesis (that means to build our voice database). This technique also uses unit selection methodology. It will take 2/3 days to build a new voice if festival is configured properly in Linux PC.

3.6.1 Recording of speech corpus

The synthesis quality is inherently bound to the speech database from which the units are selected. Design and development of a speech corpus balanced in terms of coverage and prosodic diversity of all the units is one of the major issues in such methods (Concatenative synthesis. Recording is one of the big issues in these techniques. So professional speaker should record our voice in a clean environment. This set of selected sentences is recorded in a quiet room with a noise cancelling microphone using the recording facilities of a typical multimedia computer system.

A text prompt of 1000 sentences in Marathi is derived which contains 10,000 words and syllables. The same text is recorded by a native Marathi female speaker, having several years of research experience in speech processing in a professional studio. Such .wav files with 16-bit PCM coding with mono-channel and 16 KHz sampling rate is stored as database. This data is needed for data analysis and modelling. This speech database contains all possibilities of Marathi syllables. This is advantageous when user

will give any input combination of syllables so that, syllabification will take place properly.
which are collected from news papers, considering the criterion of frequently coming sentences and used in it.For generating good quality of speech from TTS system, it necessary to have good speech corpus. We have built our speech corpus on our own comprising 1000 sentences

3.6.2 Text normalization

Text normalization is necessary to remove the symbols. As the system doesn't understand the meaning of symbols, it must be removed if any comes via input. But the character recognizing symbols must be preserved. There are fixed rules for syllabification in Marathi, which is explained in the preceding chapter. Barakhadi rules are used for syllabification. These rules are used for text syllabification. Further, text syllabification will segment the normalized text to syllable unit according to Marathi language rules. Phone set Definition module defines the complete set of phones used in Marathi speech. It also includes feature definitions (ex. vowel/consonant, lip rounding) of these phones. Lexical Analysis module is used to arrive at the phones that make up the pronunciation of a particular word. Since Marathi is agglutinate in nature, we do not require a dictionary for lexical analysis. Instead, this module defines letter-to-sound rules (LTS) which are used to arrive at the speech phones based on the spelling of the word. Syllable extraction and Concatenation module will search for the sequence of syllables obtained from the syllabification module and combine syllable unit sound files to produce a synthesized

speech. Prosodic phrasing in speech synthesis makes the whole speech more understandable. Phrasing is done based on punctuation.
Text normalization is the process of transforming text into a single canonical form that it might not have had before. Text normalization is important in TTS synthesis systems. It removes all punctuation marks (if any) present in the text input (e.g , #,$,%, & etc.). It is necessary because, the system is unknown about the meaning of these symbols. The system can't discern if we put date (15/6/14 or 15 June, 14) or time (11:30am or 1340Hrs). The system can't even understand numbers. For the units such as centimetre, kilograms, they are needed to be written in full-form as the system does not know cm, kg etc.

3.6.3. Building utterance structures

Utterances are needed to be built for unit selection. To make access well defined, we need to build the utterances. It requires labeling for the utterances in the database and the text. These labels are for: intonation events, F0 targets, phrases, words, syllables, segments etc. Here, a waveform synthesizing using unit selection technique is done. From the term 'unit selection' we mean the selection of whole unit right from the phrase/word to the diphone. We must be alert in building unit selection based synthesis system because the segment boundaries are going to under test. So, the quality of segment labels must be good. Generally this is done using automatic speech recognition (ASR) systems. Auto-alignment of the utterances and the corresponding texts is completed with the help of ASR systems. This alignment must be accurate as this can be used to test the accuracy of labels. The basic label types are listed below.

Segment: It labels in the nearer/correct boundaries, in the phoneset.
Syllable: With the stress marking, whose boundaries are closely aligned with the segment boundaries.
Word: These things can be looked in lexicon, explained later.
IntEvent: These are intonation labels; aligned to syllables.
Phrase: A name marking to the end of each prosodic phrase.
Target: the mean F0 value in the midpoint of each segment in the utterance. Unit selection uses 'optimal coupling' wherein, most appropriate joint is found at run time when two units are selected for concatenation.

3.6.4 Extraction of pitch marks, MFCCs and Clustering

For the cluster method defined here it is best to construct more than simply segments, durations and an F0 target. A whole syllabic structure plus word boundaries, intonation events and phrasing allow a much richer set of features to be used for clusters. In order to cluster all similar units in the corpus, we build acoustic representation of them. Here we use Mel-Cepstrum. We do generate it at pitch marks so that we can have parametric spectral representation of each pitch period. The Mel-frequency cepstral coefficients (MFCC) are generated at fixed frames.

3.6.5 Testing and tuning

Finally, each cluster is tested and tuned to proper prompt. Here we confirm proper labelling of the text to the speech waveform. Also, lexicon covers method for finding the pronunciation of a word. The lexicon structure that is basically available in Festival takes both a word and a part of speech (and arbitrary token) to find the given pronunciation. This is either by a lexicon (a large list of words and their pronunciations) or by

some method of letter to sound rules.

3.7 Execution of TTS

When we give an input text to the TTS system, it generates the human speech. FESTVOX platform is utilized for the unit selection synthesis in Marathi language. More detail information can be seen on website http://festvox.org/festival/

Chapter 4

4 Prosody event detection

4.1 Prosodic event detection in a speech segment

Tonal and rhythmic aspects of speech are called as prosody. Since it extends over more than one phoneme segments, prosody is said to be supra-segmental. Prosody conveys emphasis, intend, attitude and emotion. These are important cues for interpretation of speech How emotions are expressed in the voice can be analyzed at three different levels:

1. The physiological stage (e.g., describing nerve impulses or muscle innervations styles of the predominant systems concerned within the speech production system) 2.The phonatory -articulatory stage (e.g., describing the location or motion of the predominant structures example vocal folds) 3. The acoustic level (describing traits of the speech wave shape emanating from the mouth)

The current methods for evaluation at the physiological and phonatory levels are mostly intrusive and needs specialized equipment at very high degree of proficiency. Contradictorily, acoustic prompts of vocal emotion expression may be acquired from subject in a way which is economical, from vocalizations i.e. voice recordings, and allow some conclusion for speech creation and physiological determinants. Therefore, acoustic measurement of voice prompts, involves no particular apparatus, perhaps needed is basic training in voice physiology and acoustics. The acoustic feature measurement method assures the use for interdisciplinary re-

search on speech. Voice prompts are usually characterized into:
(a) fundamental frequency (F0, the perceived pitch),
(b) vocal perturbation (short-term inconsistency in sound production),
(c) voice quality (a correlate of the apparent 'timbre'),
(d) intensity (a correlate of the apparent loudness), and
(e) temporal characteristic of speech (e.g., speaking rate), as well as various combinations of these characteristics (e.g., prosodic features).

4.2 Prosodic corpus

There is barely any point of view which goes against involving a professional voice over artiest speaker for recording task. However, in a study of Marathi language we observed that expressivity patterns of prosody were not as much of preferred by listeners when recorded by non professional sound artists .The influence in favor of a Professional speaker was more, Rather conversely experience shows that professional voice over artiest are likely to be much more stable in their speech productions, ensuing in elevated consistency in the segmental and spectral domains. These sound artists are competent in sustaining a constant amplitude levels across speech utterances and within entire recording process. And their voices not tend to enervate quickly as 'normal' speaker's voices. We have recorded a prosodic database at local language in Marathi for 400 utterances. Recording is done in sound studio under balanced acoustic condition. Professional female voice over artist is involved in recording processes so to ensure good prosodic event incorporation in a speech. Recording is done in mono channel 16 KHz sampling frequency including six emotion colors angry, sad, happy, interrogative, exclamatory and neutral

discourse. Refer table 1 for the detail.

4.3 Acoustic Features

Prosody shows a noticeable effect on supra segmental Acoustic features in Fundamental Frequency(F0), energy, and timing in the locality of the event. Accent and boundary events are marked by overstated movements of the F0 contour. Accented syllables show an increase in the local energy profile. Pre boundary syllable lengthening is a subtle timing variation found in the vicinity of boundary events (Wightman, 1992). We have extracted acoustic features from these prompts, which are variants of supra segmental features .An experimentation work for selecting optimum and important features among all features used for prosody classification is done for giving a rank to acoustic features by its importance . While preparing feature set for classification problem an important issue is selection of appropriate features to make best use of classification algorithms on our feature dataset. The selection of feature is perilously significant as it can stand for the dissimilarity between effectively and meaningfully modeling the problem. This ranking and selecting predominant features was implemented using the WEKA machine learning toolkit

4.3.1 Feature Selection

Feature dataset used for machine learning contains many feature attributes, some of those may be less significant in depicting a particular class or some may be more relevant in making predictions of same classification problem. It's a practical problem to decide about which features to use

and which to discard The process of picking up features from feature set to model a problem is called feature selection. Feature Selection is a practice by which automatic search for the best subset of attributes dataset is carried out. The concept of 'best' is relative, which is based on the classification issue and accuracy expected .In a most useful way a problem is thought of selecting feature is a explore-space search. The explore space is discrete and contains all possible combinations of attributes one can prefer from the feature dataset. The purpose is to steer through the explore space and locate the finest or at least good adequate mixture of features that improves performance of classifiers over selecting all attributes. Three key benefits of performing feature selection are :

1.Reduces Over fitting : fewer redundant features means less opportunity to make decisions based on noise like interfering parameters.

2.Improves accuracy : a lesser amount of misleading features means modeling accuracy improves.

3.Reduces Training Time : Less features means that algorithms get trained quicker.

Feature selection is separated into two parts:

Attribute Evaluator

Search Method.

Each section has many techniques to choose from. The attribute evaluation is the procedure in which every feature in feature set is evaluated in the context of the output class. While the search method attempts a try for diverse combinations of feature in the dataset in order to turn up on a small inventory of selected features. Feature selection techniques in Weka are as follows.

1.Correlation Based Feature Selection Correlation is also known as Pear-

son's correlation coefficient in statistics. In this method we calculate the correlation of the outut with each attribute and select some attributes that have correlation (close to -1 or 1) and discard the attributes with a low correlation (value close to zero).Weka supports correlation based feature selection with the Correlation Attribute Eval technique that requires use of a Ranker search method. (Jason, 2016)

2.Learner Based Feature Selection Different subsets of attributes are created and selected for classification. The subset that results in the best performance is taken as the selected subset. The algorithm used to evaluate the subsets should be generally quick to train and powerful, like a decision tree method. (Jason, 2016)

3. Information Gain Based Feature Selection Another popular feature selection technique is to calculate the information gain.it involves calculating the information gain (also called entropy) for each attribute for the output variable. Entry values vary from 0 (no information) to 1 (maximum information). Those attributes that contribute more information will have a higher information gain value and can be selected, whereas those that do not add much information will have a lower score and can be removed. (Quinlan JR, 1986) Weka supports feature selection via information gain using the Info Gain Attribute Evaluator. Like the correlation technique above, the Ranker Search Method must be used. We have used this method for the work done here.

The information gain of an attribute feature gives us how much relevant information required for doing the classification of the target the attribute gives you. In other words, it computes the difference in information between the cases where we know the value of the attribute and where we don't know the value of the attribute. A information measure is through

Shannon entropy.

It means the information gain is based on two facts ,1st is how much information was offered before knowing the attribute value, and how much was available after. For example, if sample set contains only one class, then the class is without having seen any attribute values and the information gain will always be 0. If, on the other hand, you have no information to start with (because the classes you want to predict are represented in equal quantities in your data), and an attribute splits the data perfectly into the classes, its information gain will be 1. As per experiments conducted to this ranking criterion, F0 related variants and energy related variants play a key role in discrimination between presence or no presence of prosodic event (S. Ananthakrishnan., 2008).

1. Features resulting from F0 incorporate

A Within-speech utterance F0(f0_Average)

B Difference measured from maximum and average in- utterance F0 (f0_maxavg_diff)

C. difference measured in minimum and average within- utterance F0 (f0_avgmin_diff)

2. Number of times the significant change in the pitch slope

3. Features resulting from timing cues consist of

A. duration for every speech utterance (n_dur)

4. Features resulting from energy include

A. within- speech utterance the range of energy (e_range)

B. difference measured between maximum and average in-speech utterance energy (e_maxavg_diff)

C. difference calculated between minimum and average energy n the - utterance (e_avgmin_diff).

4.4 Feature calculations.

4.4.1 Essential prosodic features.

In this section, methods used for evaluating basic feature contours are put in plain words. These basic features (i.e. pitch and energy) will be the initial work done for further aiming towards more complex feature set, which include helpful information for our use. The software package used in this work is ”Praat”.

Looking for the achievement of feasible assessment, and evade the problems caused by the non-stationary type of speech signal, it’s allowed to assume universally that the properties of the signal change relatively slow with time. It leads us to examine of a short-time window of signal to calculate significant features which are supposed to be unchanging within the length of the window. the majority techniques give in parameters averaged over the line of the time of window function. Thus, if dynamic features are expected to be represented, the speech signal should be divided into successive frames with appropriate window function which can be referred as analysis frames, so that the features are computed often enough to follow appropriate changes. therefore, in order to find F0 and energy contours, shorter segments of speech, called frames, are considered.

For each window,the calculation of the F0 and energy values are done. There will be one single value per frame is calculated and extending its calculation for a longer analysis window is employed. in the interior of the analysis window, all the speech signal values are considered and analysis windows are always overlapped to avoid discontinuities in the analysis. Frame durations of 20 ms to 30 ms are commonly used in speech processing techniques, whereas window lengths for F0 and energy calculations

are typically set between duration of 25 ms and 40 ms. The analysis performed in the present work considers frame durations of 25 ms with 30 percent overlap.
Since identifying of voiced/unvoiced region in speech signal is the foundation of the F0 computation, it's the first algorithm in being described within this section. The decision is frame-based, and only over voiced frames, F0 will be estimated.

4.4.2 Voiced/unvoiced decision

Speech signal contains voiced and unvoiced parts. Voiced speech indicates functioning of the vocal folds in response to airflow from the lungs. It is periodic and it is observed separately as the characteristics of the vocal tract. It is periodical signal and measure of the periodicity leads to the determination of the important parameter called fundamental frequency of vibration or also called "pitch". In unvoiced segment of speech the source of speech or sound created as a result of the restriction which is called as a white noise source. It has specifically no dominating periodic components and shows a flat spectrum signifying that all frequencies are represented uniformly (but for some sounds the noise spectrum exhibits a slope of about 6dB/octave). If representation of the time waveform of a noise source is observed then we can claim that, around the zero axis we can see a random pattern of movement. In this perspective when signal is lacking in its periodicity, pitch evaluation gives no logic. Therefore, for F0 assessment it is important to identify which frames are well thought-out as voiced and which one as unvoiced. Voiced frames discriminate them from unvoiced signal s by high amplitude values, a comparatively

low zero-crossing rate and high energy values. Zero-crossing rate is the number for which signal undergoes zero-crossings in one unit time .There are several procedures exists to make a decision between voiced/unvoiced frames .In this work, following method of zero crossing detector for the voiced/unvoiced decision is followed (Raquel Tato, n.d.) Algorithm is as follows (Raquel Tato, n.d.)

1.Zero-crossing rate in Hz:

$$Zcross = n\_cross \cdot \frac{f_s}{N}$$

2.Normalised energy of the signal:

$$EneNorm = \frac{{}^{1}/_{N} \cdot \sum s_n^2}{Range \cdot MaxRange}$$

3.Normalised absolute maximum:

$$MaxNorm = \frac{\max\{|s_n|\}}{Range}$$

Where

fs Sampling frequency in Hz (here 8000)

N Frame length in samples (here 160)

sn n-sample value of the signal

n_cross Amount of zero-crossings during a frame

Range Difference between maximum and minimum value in the signal

MaxRange Maximum feasible range, dependent on the quantification (here 16 Bit means MaxRange=65536)

Normalization in (2) and (3) comes from the fact that the speaker may verbalize at different energy levels at different time. The decision rule is achieved through the comparison of thresholds theoretically based ncross,

EneNorm, MaxNorm with a vector whose components result from equations (1) to (3): If

n_cross < n_cross and

EneNorm > EneNorm and

MaxNorm > MaxNorm

then

sound is Voiced

else

sound is Unvoiced

Where

n_cross ,EneNorm ,MaxNorm, are Vector obtained from (1) - (3)

4.4.3 Calculation of Fundamental frequency with its variants.

For the analysis and synthesis of speech signal in phonetics ”Praat ” is a open ware software package. Paul Boersma and David Weenink from the University of Amsterdam designed it and it is even continuously updated. It has a variety of conventional and non- conventional procedures, along with spectrographic analysis, articulatory synthesis, and artificial neural networks.Features resulting from F0 incorporate calculation of F0 and its variants .algorithm is implemented via Praat as follows.

Before we illustrate how to measure pitch in Praat, let’s discuss what the pitch is and what it used for.Pitch is a term used to refer to variations in fundamental frequency (F0), which serves as an important acoustic cue for tone, lexical stress, and intonation. For example, in Chinese, which is a tone language, each syllable or morpheme may have its own pitch Extracting information about pitch(reference:Praat manual)

1. Display the pitch track: Show pitch

2. At this point, a blue line will be placed on the spectrogram representing the pitch. At this time, you can place the cursor at the point and read the blue number on the right side of the window.

3. Or you can position the cursor in a stable middle part of the blue track and click " Pitch" and then select "Get pitch". A local pitch value will be displayed in a separate window.

4. Select the portion of the sound for which you would like the Maximum, Minimum or Average Pitch

5. Select the proper command for your task from the top menu: Pitch Get Pitch/Get Maximum Pitch/Get Minimum Pitch

6. Improving the pitch contour by adjusting the pitch settings

Sometimes the blue pitch contour jumps up and down, doubling and halving the actual F0, and in many cases, especially where the speaker is creaky, the pitch track will drop out altogether, which is because Praat's default pitch range is not appropriate for the file are analyzing. Therefore, in order to make the pitch track more visible and better reflect the speaker's voice , you may need to adjust some of the pitch settings

The fundamental frequency of the voice (pitch) usually ranges from approximately 30-300 Hz, but this varies according to different speakers: typically male pitch ranges from 50-180Hz and females from 80-250Hz, so we usually set the pitch range to a reasonable range of 50 - 400 for general usage.

If you have a general sense on what the speaker's actual range is (e.g. getting from the previous measuring), you can set the minimum to just under the speaker's lowest F0 and the maximum to just over their highest pitch excursion.

If the pitch contour is too low in the spectrogram, you can increase the maximum value of the pitch range (e.g. increase from 400 to 500); if the pitch contour is too high, you can decrease the maximum value of the pitch range (e.g. increase from 400 to 300).
This part is adapted from Stonham lecture notes (p.13) that is available at http://stonham.dyndns.org/phonetics/handouts/prosod_hndt.pdf

4.4.4 Energy calculation with its variants

Correlating the perception of the loudness along with the acoustic measurement is as intricate as the blending of calculable F0 and the tone pitch perception. The sensation of loudness is both dependent on the frequency of the sound and on the duration, and the other way round, pitch perception depends on the loudness.

fundamental computation trial used for calculation of energy as the acoustic associate of insightful loudness are base on associations involving physical acoustic pressure magnitudes *ps*, measured in Pascal (1Pa= 1N/m2), and the acoustic strength *Is*, whose unit is W/m2. It is defined that *Is* is proportional to *ps2*. With help of the acoustic power reference value, $I0=1pW/m2$, and the acoustic pressure reference value, $p0=20\ \mu Pa$, which shows the human auditory threshold at frequencies.

Amplitude is directly related to the acoustic energy or intensity of a sound. Both amplitude and intensity are related to sound's power. All three of these characteristics have their own related standardized measurements and will be discussed below. Amplitude is measured in the amount of force applied over an area. The most common unit of measurement of force applied to an area for acoustic study is the Newtons

per square meter (N/m2)..A few more relationships between amplitude, intensity and power: intensity equals the square of the amplitude, so if the amplitude of a sound is doubled, its intensity is quadrupled. Power is also proportional to amplitude squared, therefore power and intensity are proportional to each other. .

Energy calculation using PRAAT

1. Similar to what is said about the use of spectrographic techniques outside the 'Edit' window,one can do intensity analyses separately. Again, the 'Edit' options are by far the easiest way to go, but the procedures explained below can be used as an alternative if so desired.

2. Select the original speech object of the vowel .

3. From the main menu, choose the option 'To Intensity.

4. Keep the default values for the selection window (unless your minimal pitch is expected and measured below 100 Hz) and click 'OK'

5. Select again the speech object, and choose 'Edit' from the main menu

6. Determine a selection of your signal, that is, decide on a onset and offset point for an interval for which you want to calculate a mean (and standard deviation) level of intensity. Then, close the window.

7. Select the Intensity object (Intensity _ {name })

8. Choose from the menu the option 'Query-'.

9. Choose the option 'Get mean..' and define the interval according to your previous selection

10. The mean value will appear in a separate ('Info') window. You can save this information by writing it down or copy and paste it into a document

11. Do the same for 'Get standard deviation.

12. FYI, if you select the speech object and choose 'Query' from the

Table 5: speech features for prosodic event detection

	Features derived from F0 comprise	
1	within-utterance F0	f0_Average
2	difference between maximum and average f0 within- utterance	(f0_maxavg_diff)
3	difference between minimum and average f0 within- utterance	(f0_avgmin_diff)
4	Pitch(f0) Slope	df/dt
5	Number of changes in the pitch slope	N
6	Features derived from timing prompts include	
7	Duration for each speech utterance	(n_dur)
8	Features resulting from energy comprise	
9	energy range Within- speech utterance	(e_range)
10	Difference between maximum and average energy in speech utterance	(e_maxavg_diff)
11	Difference between minimum and average energy in speech utterance	(e_avgmin_diff)

main menu, you get a lot more options for calculating levels of energy.

Return value

The intensity in air, expressed in dB relative to the auditory threshold.

Algorithm

The intensity of a sound in air is defined as

$10log_{10}\{1/(T\ {P_0}^2)dt\ x^2(t)$

where x(t) is the sound pressure in units of Pa (Pascal), T is the duration of sound, and $P0 = 2.10^5 Pa$

is the auditory threshold pressure. Following table indicate all the features significantly considered by this research work for the prosody event detection.

4.5 Prosody Analysis Of Emotional Speech

In this section, we cover analysis of emotional speech segments and neutral speech segments for evaluation of various prosody patterns. For this purpose, 400 sentences are considered from single female speaker. Emotional and neutral speech segments have been compared on the basis of selected feature set comprising fundamental frequency ,energy, sentence durations and formants variation w.r.t. fundamental frequency figures 5 and 6 shows variation of the prosodic features in plain discourse and prosodic discourse .In the prosodic discourse we can see the distinct changes and variations of features which indicates the prosodic event. Task accomplished is capturing those variations in transform domain and representing it in the feature space so that it will serve to accomplish the task of detecting the prosodic utterance in this phase of work. Figure 5 and 6 identifies the emotionally stressed word pattern where as figure 7 and 8 indicates whole sentence as a context to represent the prosodic structure. Here we can completely see the prosodic features correlating with spectral features which Indicates that identifying, capturing, and coding the variation in prosodic feature only can't be useful in prosody event detection but spectral correlates are also important variations needs to be accounted and coded so that together it will help in detection of prosody and in further work transformation of the source prosody in to target prosody

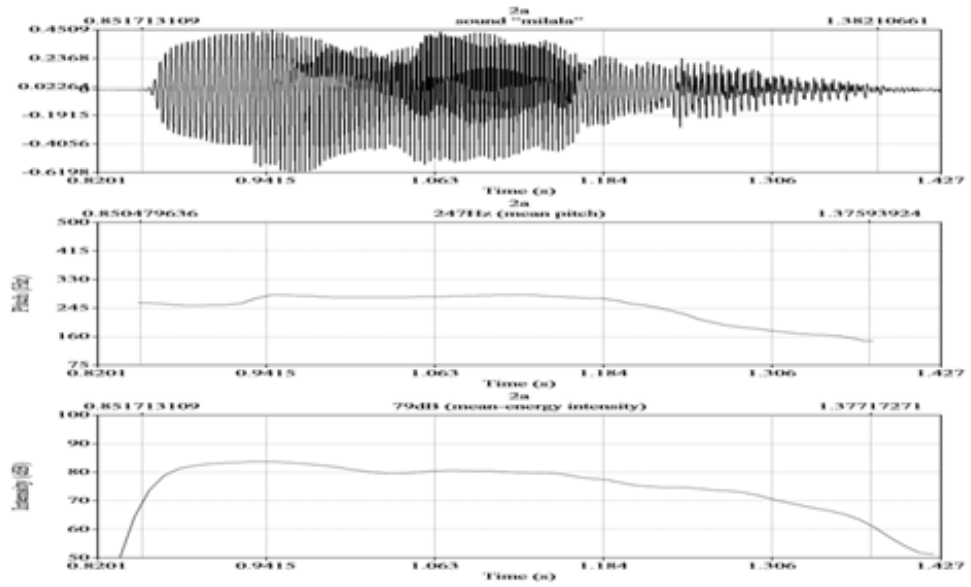

Figure 5: a) (top) speech utterance ”MILALA” Neutral , b)(middle) pitch plot c)(lower) energy

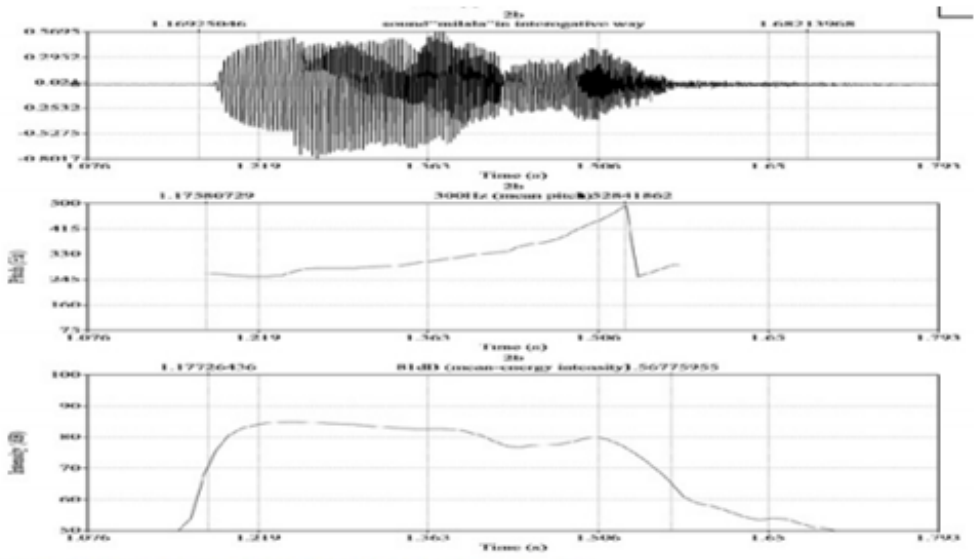

Figure 6: a) (top) speech utterance (prosodic)”MILALA”, b) (middle) pitch plot c) (lower)energy plot

Table 7 below shows the variation in the fundamental frequencies and correspondingly variation in the formants .This variation pattern is analyzed for prosodic and plain discourse.It prominently shows the correlation between spectral parameters and prosodic parameters It leads to make us believe and conclude that when a prosodic parameter makes any variation a backdrop of spectral properties also shows significant change.

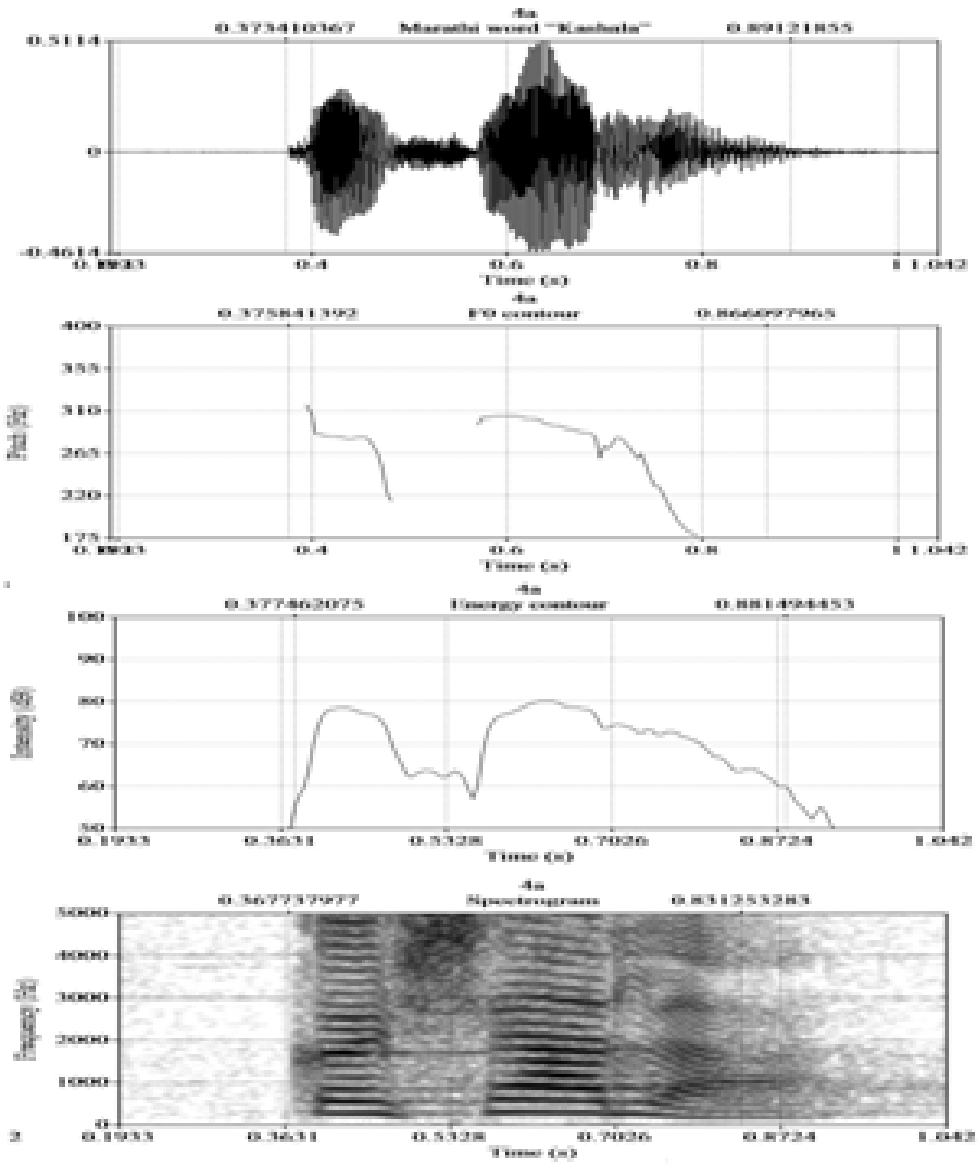

Figure 7: a) (top) speech utterance (prosodic)"MILALA", b) (middle) pitch plot c) (lower)energy plot

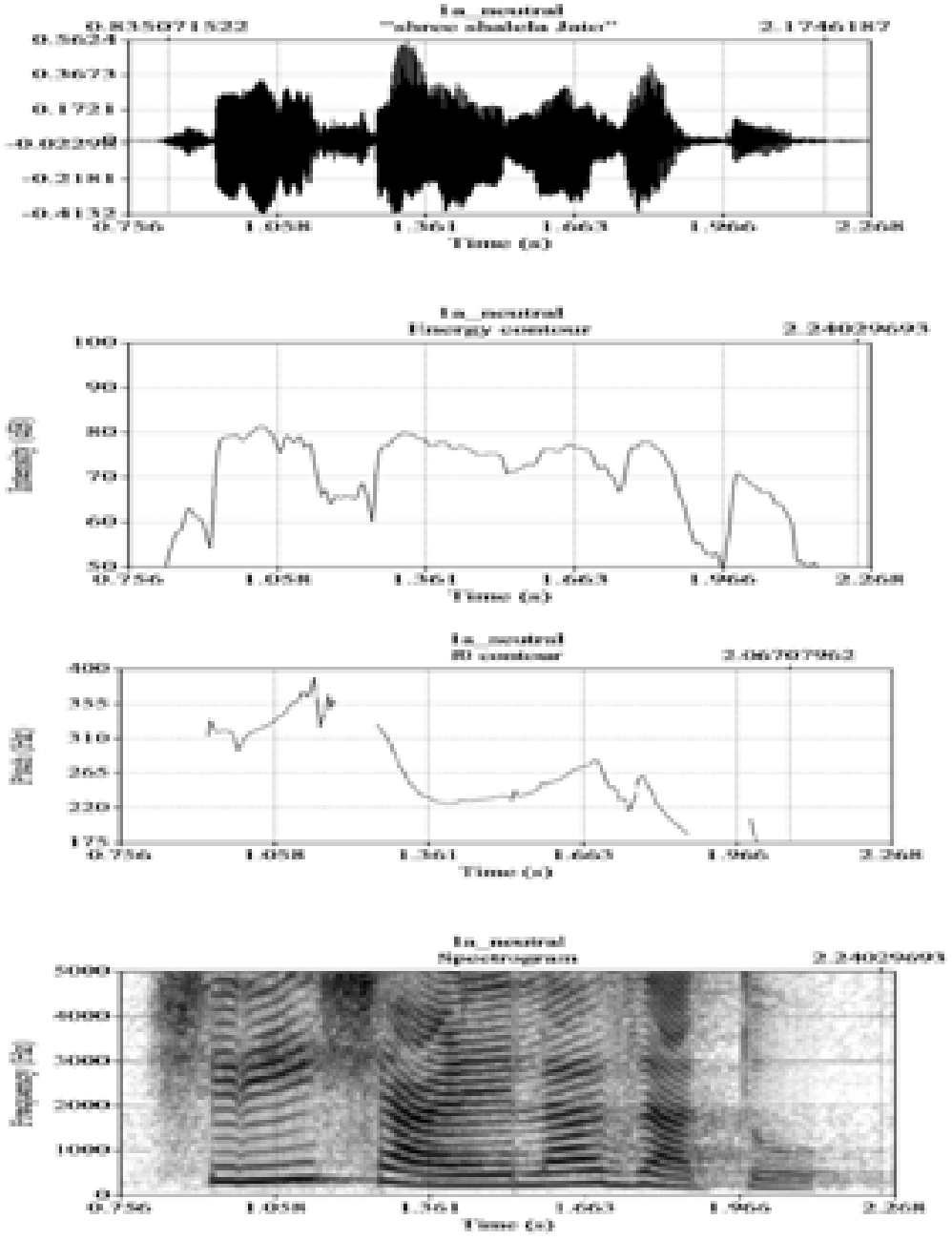

Figure 8: a) (top) speech utterance (prosodic)"KASHALA", b) pitch plot c) energy plot d) (bottom) Spectrogram

Table 6: fundamental Frequency and formant variation

F0 Fundamental Frequency in Hz	Formants	Non prosodic Statement formants in Hz	Prosodic Statement, formants in Hz	Range of, formant variation in Hz
175	F1	871.925158	443.571412	300-400
	F2	1587.972422	1774.279191	200-300
	F3	2745.054532	2984.530219	150-200
	F4	4215.969993	4265.615935	100-150
200	F1	814.803002	512.523414	300-450
	F2	1681.673354	1800.242373	200-300
	F3	2864.712783	3015.838824	150-200
	F4	4309.402785	447.17647	100-150
250	F1	539.603266	952.86551	300-400
	F2	1351.323813	1547.313145	200-300
	F3	2213.129772	2809.744999	300-500
	F4	3368.137737	3479.721831	100-200
275	F1	414.380078	435.881378	50-100
	F2	1928.145161	1755.6136	100-200
	F3	2404.9477	2939.846431	300-500
	F4	3625.889873	3753.714089	100-150
300	F1	544.167013	822.735732	300-400
	F2	1625.407853	1682.038317	100-200
	F3	2580.324714	2810.462495	200-300
	F4	4015.238505	4242.794527	100-150
350	F1	838.1434	940.064487	100-150
	F2	2031.495545	1674.970731	300-400
	F3	2833.955942	2813.645423	50-100
	F4	4581.192956	4147.812335	200-400

Table 7: Quantitive analysis s of prosodic features in sentence level prosody analysis

Feature	Neutral utterance	Prosodic utterance Happy	Prosodic utterance Sad	Prosodic utterance Interrogative
f0_Average	160Hz	317 Hz	230 Hz	300Hz
f0_maxavg_diff	21%	57%	35%	50%
f0_avgmin_diff	30%	30%	55%	15%
N	1 to 4	5 to 6 times	2 to 3 times	6 to 7 times
n_dur	100%	85%	95%	80%
e_range	30db	35 db	30db	25db
e_maxavg_diff	15%	18%	15%	16%
e_avgmin_diff	16%	20%	19%	25%

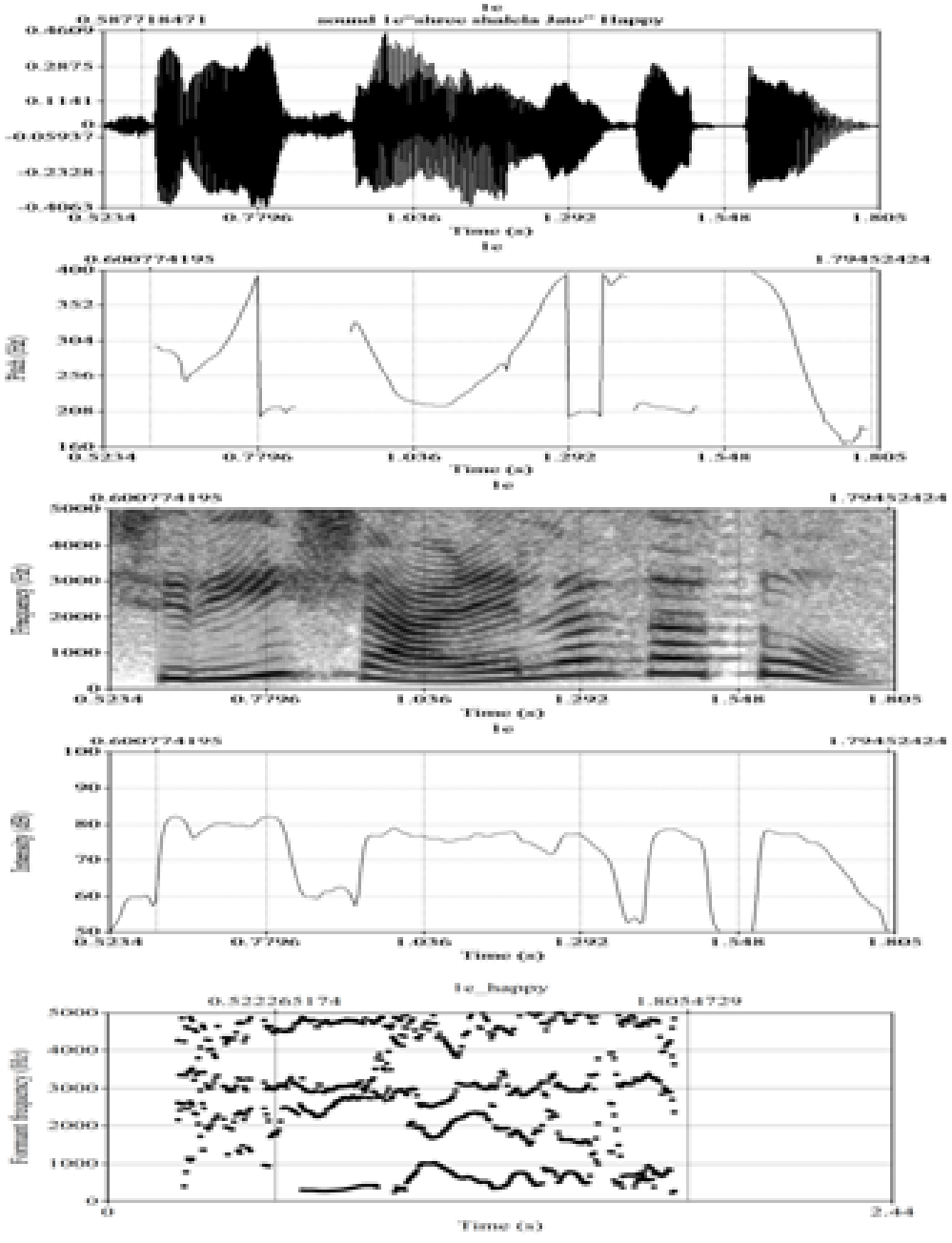

Figure 9: a) (top) speech utterance (prosodic)"shree shalela jato", b) pitch plot c) energy plot d) (bottom) Spectrogram

4.6 Conclusion

Experimental set up revels that expressive-speech holds considerable prosodic variation with respect to neutral speech. Neutral emotions have lower pitch values while interrogative emotions have higher pitch values. Intensity of speech is naturally high for interrogation and anger and happy shades of emotions compared to Neutral and Sad emotions. This holds good in our computations using Mat lab as well using Praat. As deviations from mean in chosen f0 feature is less for Neutral emotion, Standard Deviation and other statistical variations can be used to identify emotion. It means Statistical variations in f0 show lowest value in Neutral emotion. Energy also can be used to classify interrogation, happy and sad part of speech. Speech with neutral emotion is at averagely lower energy. Normally, Duration of interrogative and angry emotion is low compared to Neutral, sad and happy emotions. It is verified using the results. Total number of changes in the pitch slope present in the signal can also be used for classification of emotions. In making the discourse more expressive, humans are making significant sharp changes in the fundamental frequency. It has been validated by WEKA Tool kit feature selection Algorithm based on Information Gain Based Feature Selection. From this analysis, we can get apparent insight of emotion particular information present in the prosodic features. It is also obvious that neutral speech can be altered into emotional speech by reducing prosodic variation between neutral and emotional-speech. Thus, in this work expressive speech is generated by integrating prosodic parameters into neutral speech.

4.7 Contribution of this stage work in the further investigation

For generating better quality of emotional speech it has been observed that emotional-speech contains significant prosodic variations with respect to neutral speech. Here we notably wish to indicate that the energy value remains quite uniform for the entire utterance in the any segment such as Neutral or Prosodic.Of courese the uniform values of spectral energies are different for different emotions . It leads to a pointer towards processing prosody in speech by giving significance to high spectral energy correlation in successive frames. Feature variations are seen to be drastically changing in successive frames per emotions.

Chapter 5

5 Prosody classification with novel feature extraction method.

5.1 Introduction and need

In order to improve synthesis of emotional speech, it is necessary to be able to compare different outputs and to evaluate their quality. So far, the quality is generally assessed through human perception tests. In order to be able to detect even small differences in the quality of two systems, the number of samples as well as the number of human judges has to be sufficiently high. Finding qualified human participants however is difficult and the number of samples that can be presented to one listener should be limited, to avoid fatigue. Hence, human perception tests are time consuming and expensive. These disadvantages are avoided, if automatic emotion identification could be used as an objective measure to evaluate the quality of emotional speech synthesis. The original postulation is that an emotion synthesis scheme is of good quality, if the intended emotion can be predicted correctly by an emotion recognition system which is trained on human voices. Of course, such measures of emotional quality are meant to complement, not replace, existing valuation metrics such as Mel-Cepstral Distortion (MCD),Mean Opinion Scores (MOS) etc.System block diagram is as shown in figure 10..Many merits of the emotion recognition system include application such as effective human computer interaction can be possible if computers are able to distinguish the emotions. Secondly if such systems are implemented in auto answering systems then responses

of the system can be more meaningful if context is understood with the emotions.

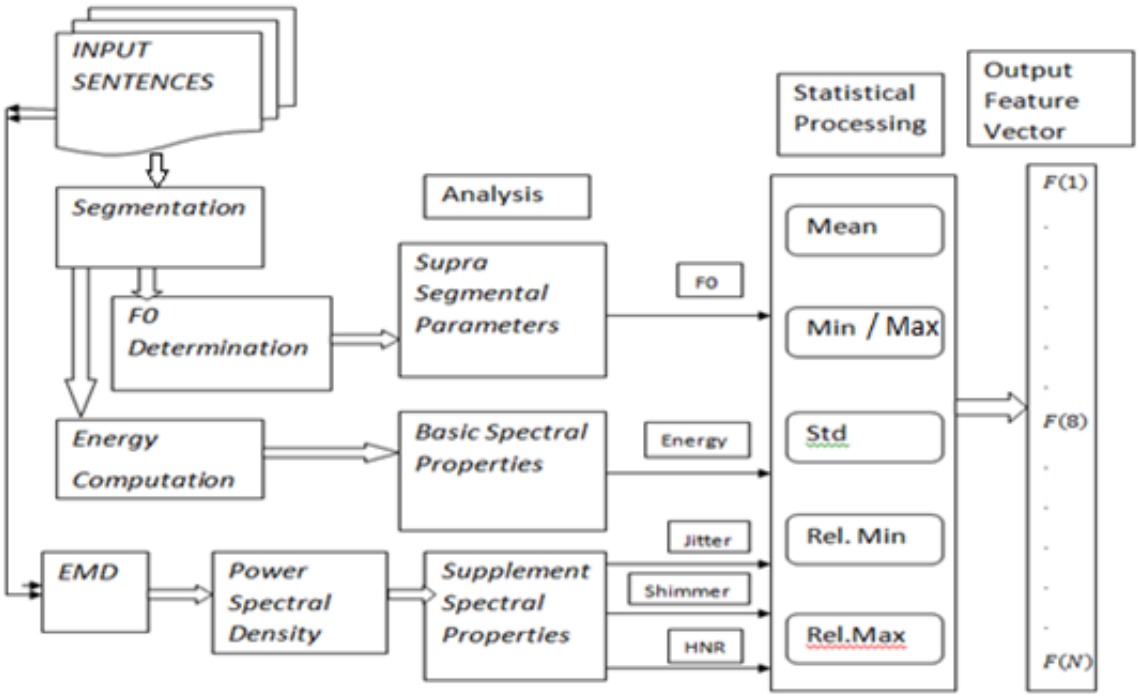

Figure 10: Block Diagram for creation of feature Vector

5.2 Acoustic features evaluated

Speech features are having following categories. For evaluation of acoustic features we use MATLAB .We focus on to extract acoustic features, and use 39 acoustic features in total.

Each sentence recorded is measured in between one and three seconds in length. We separate each utterance into 25 ms frames with 40 % overlap with hamming window. Since human speech contains audible and

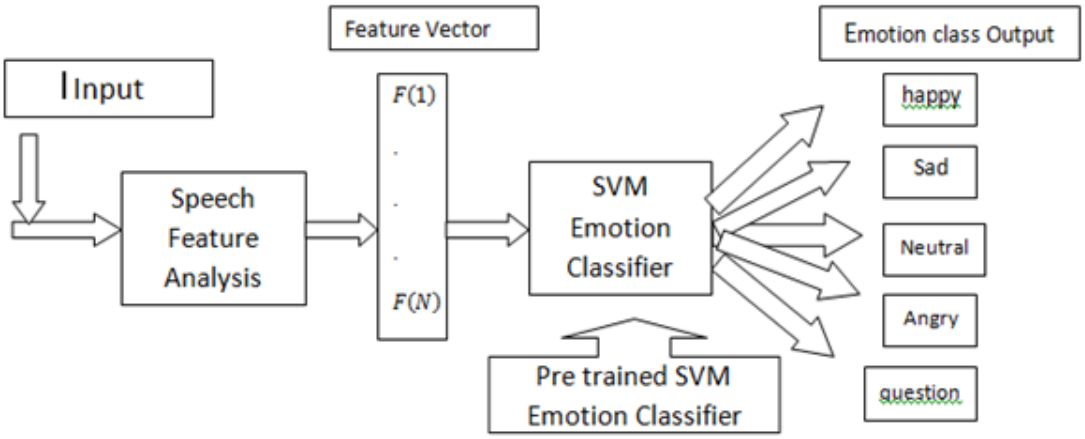

Figure 11: Emotion classification using multi class SVM with hybrid kernel

inaudible slice, we only analyze the acoustic features for the audible segments, and we ignore the inaudible ones. A segment is selected to be an audible segment by the voice activity detection algorithm which separates out sounding and non sounding parts of speech based on zero crossing and energy of signal for reducing the over all computational complexity in the method. We prefer the features: pitch, energy, duration and formants (the first four formants are analyzed for frequency and corresponding bandwidth). seeing as the change in acoustic parameters are also reflects emotional status, we take account of the pitch variation and energy variation as added features.

Choosing suitable features for developing the specific speech based systems is a critical decision. The features are to be preferred to stand for giving intended information. Various speech features represent special speech information (regarding speaker, regarding speech, regarding emotion and so on) in vastly overlapped style. as a result, in speech technology,

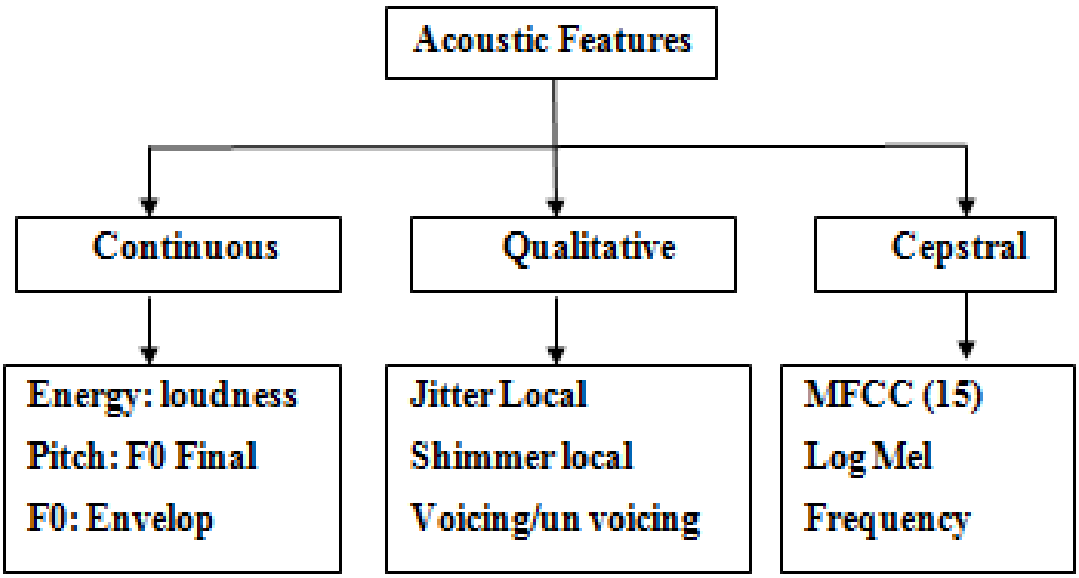

Figure 12: Categories of low-level acoustic features

very frequently features are selected on experimental base, and intermittently using the mathematical approach such as PCA (Principal component analysis). Current trends in research in the domain of speech emotion recognition, emphasized making use of blend of different features to attain the enhancement in the recognition performance. resource, method, and prosodic features discussed in the previous subsections signify mostly mutually exclusive information in the speech signal. as a result, these features are complementary in nature to each other. clever blend of complementary features is likely to improve the proposed performance of the system. Quite a lot of studies on grouping of features, proved to achieve enhanced emotion classification, In contrast to the systems implemented using individual features. Some of the noteworthy works using the blend of different features for speech emotion recognition are discuss below. The

following subsections present the literature on important speech features extracted in conventional way and proposed novel approach of Spectral Resolution coding.

5.3 The features computed and methods

5.3.1 Fundamental frequency F0

Fundamental frequency,also known as ′ pitch′ is lowest frequency component perceived by the ear. It depends on the number of vibrations per second produced by the vocal cords. We use autocorrelation in the frequency domain to extract pitch values.

Some interesting facts about Fundamental frequency
Many modern researchers who approach hearing research on the basis of Ohm's Law or Helmholtz's Resonance Theory may be neglecting the difference between "physical correlate" and "stimulus". In the case of pitch perception, we do not know either. Before we discuss pitch perception, let us get these two points clarified: A physical correlate is any parameter that manifests an invariant (or unchanging) relationship with a target sensation. That parameter helps towards quantitative estimation of the sensation even though it may not be the source of the sensation. For example, the fluctuations of mercury in a thermometer is a correlate of temperature. By graduating the fluctuations, we arrive at a quantification of heat. But it is good to note that the mercury does not cause heat nor is it accessible to the organism. The stimulus that causes heat is something totally different from the physical correlate that is used to measure heat. In the case of hearing, it was thought that frequency of vibration was the physical correlate until it was proved experimentally in the 1930s that pitch may be changed by changing other acoustic parameters even though the frequency of vibration is held constant. Thus, frequency is NOT a correlate of pitch and no one can produce a scientific frequency scale for pitch since the same pitches may be produced using different frequencies of vi-

bration. In the 60s, S. S. Stevens was so disturbed by this matter that he and his colleagues revolted (as it were) and created the MEL scale only to find themselves in the same problem that they thought they had solved. Stevens concluded: "Take for example a field like hearing in which we think we know a lot ... Actually, we know little about the stimulus to pitch." It is not surprising that we know little because, as I have pointed out in this forum, we are looking for pitch (and other auditory phenomena) in frequency of vibration according to the laws of Helmholtz, Ohm and Pythagoras [Akpan Jimmy Essien].

5.3.2. Energy:

energy stand for the loudness of the speech. We compute the energy for each speech segment by making the summation of squared values of the sample's. amplitudes Energy differences defined as " the difference of energy values between two neighboring segments ". More variations may point out towards active emotion, such as happy and anger. Formants (the first four formants frequency, thus four features): formants are determined by the shape of the vocal tract, and are influenced by different emotions.

5.3.3. Formants

(the first four formants are calculated for frequency and bandwidth thus eight features): formants are categorized by the shape of the vocal tract, and are showing prominent variations for different emotions. For instance, elevated arousal consequences in higher mean values for the first formant frequency , while positive valence results in higher mean values for the second formant frequency .

for formant calculation,we make use of linear predictive coding method [18].We find these features for each 20 ms segment of the speech sample, and then we calculate the mean, the maximum, minimum, range, and the standard deviation for each feature, resulting in 80 attributes that are sent to the classifier.

5.3.4. MFCC

It is Male Frequency Cepstral Coefficients (12 MFCC feature values extracted, thus it contributes to 12 features): MFCC coefficient of speech presents a vital component in audio processing because of its ease in calculation, superior capability to extract the feature from speech, it proves to be an proficient method and also has the benefit like anti-noise etc. The actual frequency is measure in the hertz but the pitch frequency is quantify on a range which is well-known as the Mel frequency scale .this scale has the frequency variations less than 0.1 KHz and logarithm variations greater than 0.1 KHz. The frequency on Mel scale is calculated using the formula:

$$MEL(F) = 2595log_{10}\left(\frac{f}{700}\right) \tag{1}$$

Human ear response for the frequency components in a speech signal is not linear so that there is a requirement for a non linear scale which will perfectly simulate the behavior of the cochlea.There are different scales for receiving right perception about frequency content. The scale used in this experiment is the Melody scale abbreviated as Mel Scale.This scale represents a linear frequency component of a pitch to a corresponding pitch component calculated on the non linear Mel scale. The relation of the represented frequency in Hz to frequency in Mel scale is explained in

Eq.2 and vice versa in Eq.3.

$$F(mel) = 2595 log_{10}\left(1 + \frac{f}{700}\right) \tag{2}$$

$$F(Hz) = 700\left(10^{\frac{Fmel}{2595}} - 1\right) \tag{3}$$

Fundamental frequency extraction is done via cepstrum method .FFT is implemented for energy calculation purpose. Table 9 indicates the classification organization for feature in spectral and prosodic class.The prosodic features are determined mainly from the pitch contour. Variations in the pitch and energy contours may be indicators of the underlying stress in the human speech and the emotional state of the speaker. The jitter indicates pitch perturbations extracted from the pitch contour while the shimmer is a measure of the frame by frame variability of the amplitude values extracted from the energy contour. The whole process of estimation of supra-segmental features comprises the following steps:

Determination of F0 mean value from the F0 contour. Calculation of the differential micro intonation signal about F0 are calculated as statistical parameter in terms of Mean ,standard deviation, minimum and maximum value of F0 from the corresponding F0 contours.Calculation of the absolute jitter JAbs values as the average absolute difference between consecutive pitch periods L measured in samples. Calculation of the shimmer from the peak amplitude values An detected within the n th speech frame of the input signal and normalized by the total number NV of voiced frames

5.4 Novel feature coding for prosody in emotional speech synthesis

Further acoustic features like Jitter, Shimmer, HNR(Harmonic to Noise Raio) and DVB (Degree of Voice Breaks) , DUV Degree of Voiceless are derived through the proposed method of SR(Spectral Resolution) coding.In the process of empirical coding, the linearized signal Ls is decomposed into intermediate frequency components using the methodology of Empirical mode decomposition (EMD) [22]. EMD is a popular and effective tool in the area of speech, image and signal processing . Various applications were developed using the approach of EMD for the enhancement of input sample, such as speech de noising, jitter elimination, or noise reduction. EMD is processed in a nonlinear manner for analysis of non-stationary signals. EMD decompose the time domain signal into a set of adaptive basis functions called intrinsic mode function (IMF). The IMF is formulated as the oscillatory components of the signal with no DC element. In the decomposition process, tow extreme are picked and high frequency components are selected between these two points. The left out are defined as the low frequency components. This process is repeated over the residual part repetitively to derive n-IMFs reflecting different frequency elements. A signal x(n) is represented by EMD as,

$$x[n] = \sum_{j=1}^{N} IMF_J[n] + r[n] \tag{6}$$

Where, $r[n]$ is a residual component. The IMFs are varied from high frequency to low frequency content with increase in IMF order. The prosodic features are extracted thanks to the Praat toolkit [23]. In this work, we mean to extract significant prosodic parameters: such as Mean pitch, Jitter and Shimmer, Harmonics to Noise Ratio (HNR), Degree of Voice Breaks (DVB) and Degree of Voiceless (DUV). Mean pitch is parameter repre-

sented by averaging the fundamental frequency (F0) of the utterance with autocorrelation method since F0 is the physical representation of the pitch.

The Jitter gives a measure of pitch variation by frame to frame. It is the average absolute difference between consecutive periods, expressed by:

$$jitter = \frac{\sum_{i=1}^{N-1} | T_i - T_{i+1}|}{N-1} \quad (7)$$

where Ti is the period and N represents the number of periods. The random period variability *frequency perturbation* or *vocal jitter* Vocal jitter increases in voice disorder & is responsible for hoarse, harsh or rough voice quality Jitter is a measurement of vocal stability Normal voices are usually less than 1 % frequency variability

Types of Jitter Measurement:

Mean Absolute Jitter : denote absolute variation between sequential vocal periods calculated during a sustained phonation (measured in seconds or milliseconds)

Mean Percent Jitter & Jitter Ratio: taking the mean absolute jitter & dividing it by the mean vocal period used during the phonation, the proportion is then multiplied by 100 to get a %age. if the proportion is multiplied by 100 it is called jitter ratio and is dimensionless Mean Jitter Factor: mean absolute difference between sequential vocal frequencies divided by the mean frequency of phonation, this proportion is then multiplied by 100

Calculating Jitter:

We consider Period (t) in ms, Jitter in ms

Mean % jitter= Mean absolute. Jitter/mean period x 100

‘

Vocal Shimmer: Amplitude Perturbation

The Shimmer is expressed as the variation of amplitude of a speech seg-

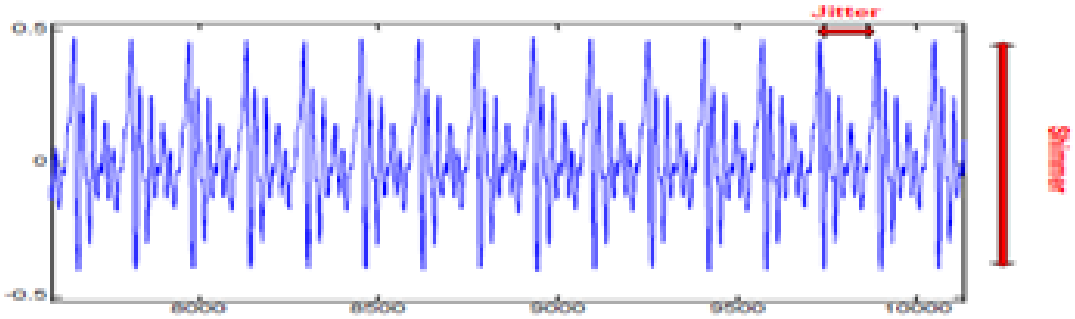

Figure 13: Representations of Jitter and Shimmer perturbation measures in Speech signal

ment frame by frame It is computed as average absolute difference between the amplitude of a signal frame by frames, scaled by factor 20. As shown in equation (8) Frequency of a speakers voice will differ from one cycle to the next

Types of Shimmer

Mean Shimmer in Decibels: mean absolute dB (SPL) difference between sequential vocal amplitudes measured during sustained phonation

Mean Shimmer in Percent: mean absolute cycle-to-cycle difference in vocal amplitude divided by the mean amplitude then multiplied by 100 to yield a % age

$$Shimmer\ (dB) = \frac{1}{N-1} \sum_{i=1}^{N-1} |\ 20\ log\ (A_{i+1} - A_i)| \tag{8}$$

where Ai and Ai+1 are the extracted peak-to-peak magnitude of speech segment and N is the number of extracted F0 periods. Fig represents Jitter and Shimmer perturbation in Speech signal

Harmonics to Noise Ratio (HNR)

This feature measures the harmonicity which signify the quantity of acous-

tic periodicity. HNR can be calculated by dividing the signal energy of periodic parts by the average noise energy, and it is expressed in dB:

$$HNR = 10log\left(\frac{E_p}{E_n}\right) \quad (9)$$ Where Ep and En are the energy component of periodic part and the noise energy.

Degree of Voice Breaks (DVB) is also considered for this work. In this feature we compute the summation of voiceless duration in the speech utterance of signal divided by the total duration of utterence. Voice break is consequence of the stopping of the air flow. DVB estimates only the silence part between the voiced segments. The Degree of Voiceless (DUV) depicts the relative proportion of non-harmonic areas (where fundamental frequency cannot be estimated).

These features are effective enough under ideal speech transformation. As in most cases these speech signals are captured at a closed surrounding. However, system noises such as jitter, harmonic frequency contents were observed in these samples. Wherein a direct feature extraction on these samples will capture feature with noise variation resulting in lower accuracy. To propose the objective of distortion free coding, a spectral domain speech coding is proposed. The suggested approach carries out the prosody generation over the selected frequency components based on spectral power density. As the silence units and noise are considerably of lower magnitude this approach effectively deviates the effect of distortion and results in higher efficiency. In the process of spectral coding, a speech signal x(n) is decomposed into intermediate frequency components using the Spectral mode decomposition. In the decomposition process the time domain signal is divided into a set of adaptive basis functions which is formulated as the oscillatory components of the signal with no DC element. In the decomposition process, two extreme are picked and high frequency

components are selected between these two points. The left out are defined as the low frequency components. This process is repeated over the residual part repetitively to derive n-spectral resolutions reflecting different frequency elements. The decomposition process is defined by,

For a given speech signal x(n) the spectral resolutions (SR) is given as,

$$x[n] = \sum_{j=1}^{N} SR_J[n] + r[n] \tag{10}$$

Where,$r[n]$ denotes as the lower frequency components. The proposed approach of spectral resolution coding is defined as SR algorithm;

5.5 SR Algorithm:

Input: Speech signal, $x[n]$

Output: Feature vector sfi

Step 1: Perform resolution decomposition computation for the speech signal (x(n)).

For SR computation;

Step a: compute the Local maxima (Xmax) and local minima (Xmin) for x(n)

Step b: compute minimum and maximum envelop signal, emin and emax.

Step c: Derive the mean envelop m[n].

Step d: Compute the detail signal d[n].

Step e: Validate for Zero mean stopping criterion.

Step f: Buffer data as SR from the decomposed resolutions.

Step 2: For obtained SR compute spectral density of each resolution using PSD.

Step 3: Select two SR having highest energy density (I1, I2).

Step 4: Compute threshold limit for I1, I2 as 0.6 of max (Ii)

Step 5: Derive the feature vectors (sfi) from these two SR for synthesis.

The operational flow chart for the proposed SR approach is summarized.

The obtained SR {I1 - I4 } are the decomposed detail SR reveling different frequency content at each level. At each decomposition level the residual resolution, r[n], is decomposed in each successive resolution to obtain finer frequency information. Each obtained resolution, reveal a finer frequency content and based on the density of these frequency contents, then

a decision of feature selection is made. This approach of feature selection, results in selection of feature detail, at lower frequency resolutions, which were discarded in the conventional CSS approach. To derive the spectral density of these obtained SR, power spectral densities (PSD) to the obtained SR are computed. PSD is defined as a density operator which defines the variation of power over different content frequencies, in a given signal x(t).

The Power spectral density (PSD) for a given signal x(t) is defined as,

$$PSD, P = lim_{T\to\infty} \frac{1}{2T} \int_{T}^{T} x(t)^2 \, dt \tag{11}$$

Taking each resolution "Ii" as reference, a PSD for each resolution, "PIi" is computed. The PSD features for the 4 obtained SR are then defined by,

$$\text{PSD(Ii), for i = 1 to 4} \tag{12}$$

The resolution PSD's are derived as,

$$PB_i = lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} B_{hi}(t)^2 \, dt \tag{13}$$

From these estimated energy values, SR are calculated and selection is done on a defined criterion, as mentioned,

For obtained PIi, maximum PI is computed, defined by,

$$\text{MPIi=max(PIi)} \tag{14}$$

For i =1 to 4

if (PIi $\geq$ (MPIi / 2))

select _Ii = Ii,

end

For these selected SR, 'Sel _Ii' prosody features are computed. Which,then

processed for transformation process.

5.6 Performance Analysis

To evaluate the performance of the developed approach and to validate the effect of SR-Coding, different speech utterance were recorded and processed for feature extraction. The impact of spectral noises on these speech signals are evaluated. The speech signal x(n) used for the processing of feature extraction is shown in figure

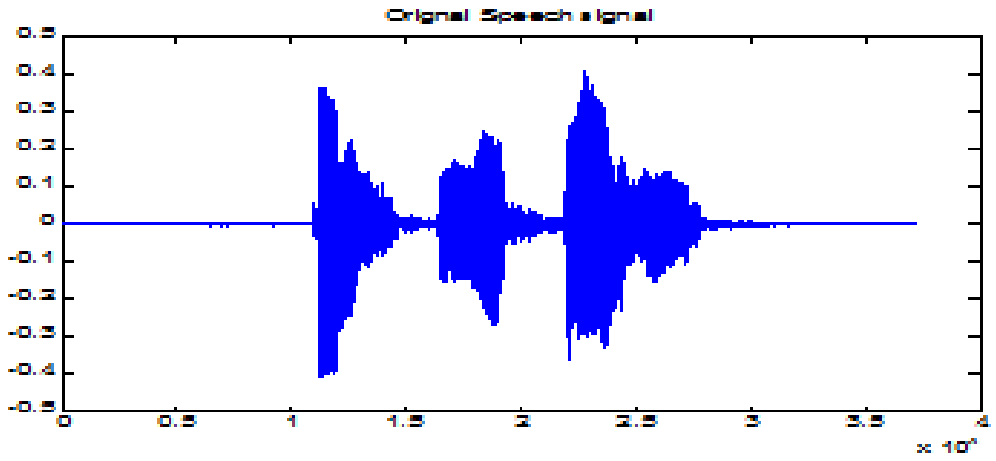

Figure 14: Original speech signal x(n)

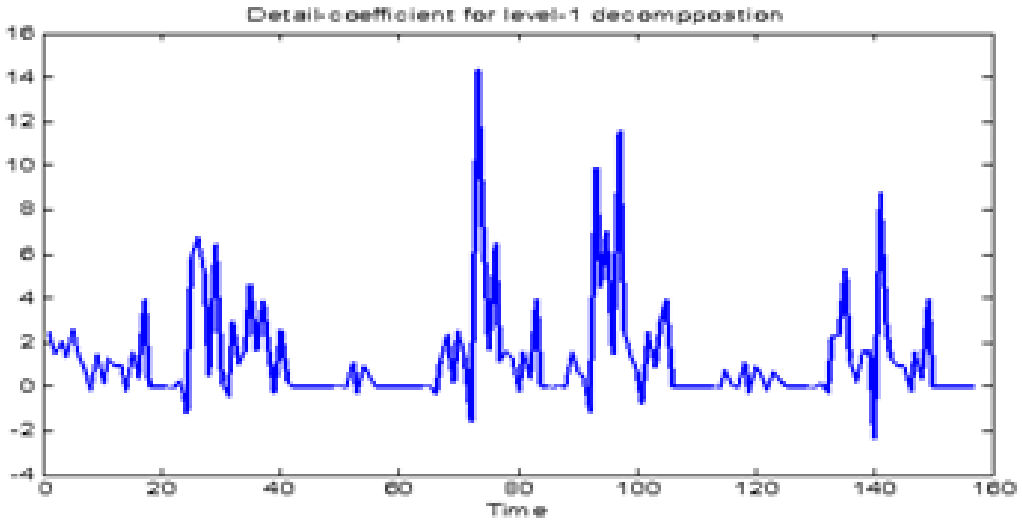

Figure 15: Detail Spectral resolution at level-1 decomposition

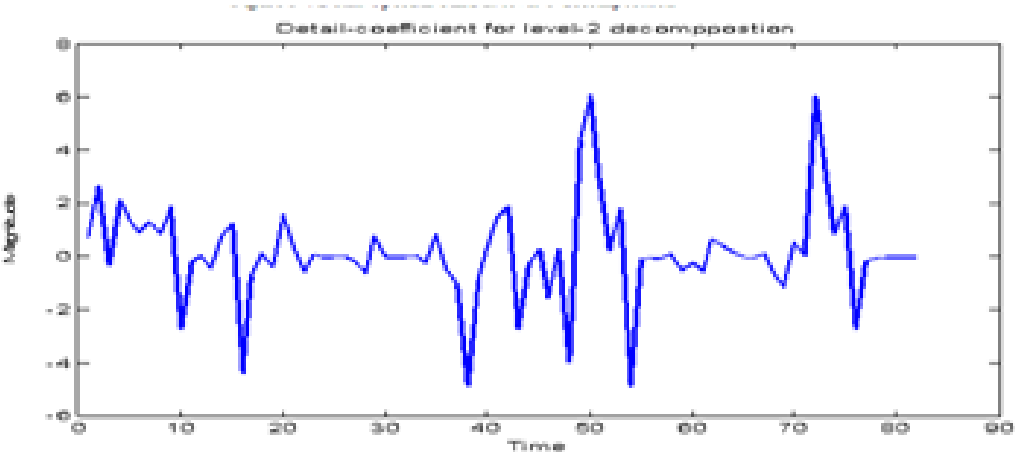

Figure 16: Detail Resolution at level-2 decomposition

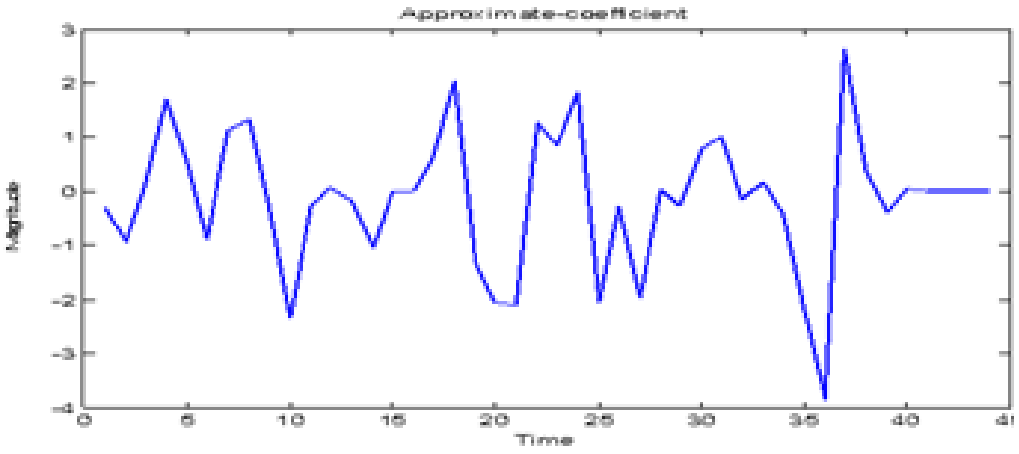

Figure 18: Residual low frequency coefficient for given signal

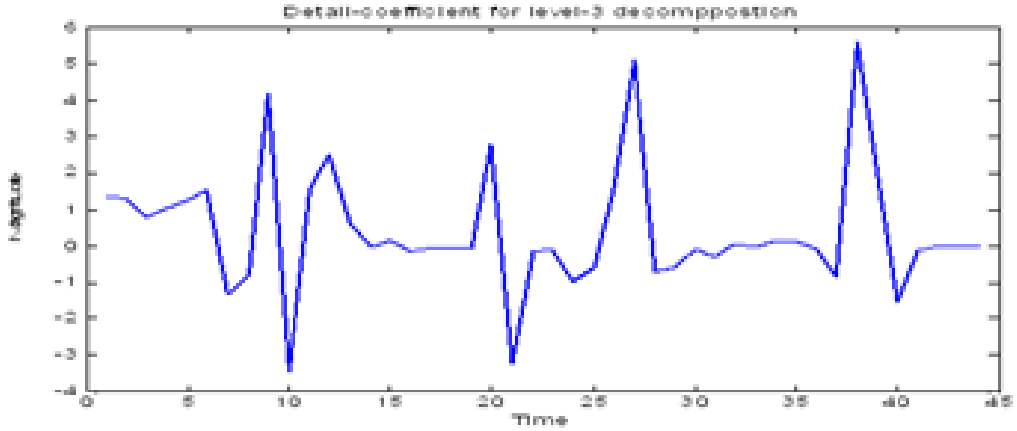

Figure 17: Detail Resolution at level-3 decomposition

To the obtained resolution decomposition is carried out using successive filtration process. The 3 detail resolutions and the approximate resolution obtained are shown in figure 4-6 respectively. It is observed that resolution 1 and resolution 3 exhibits higher coefficients variation than the other two resolutions, hence more curvature information are presented in these two resolutions. To select the required resolutions for feature extraction, a resolution density using power spectral density is used. The energy density for each resolution is as shown in figure 19.

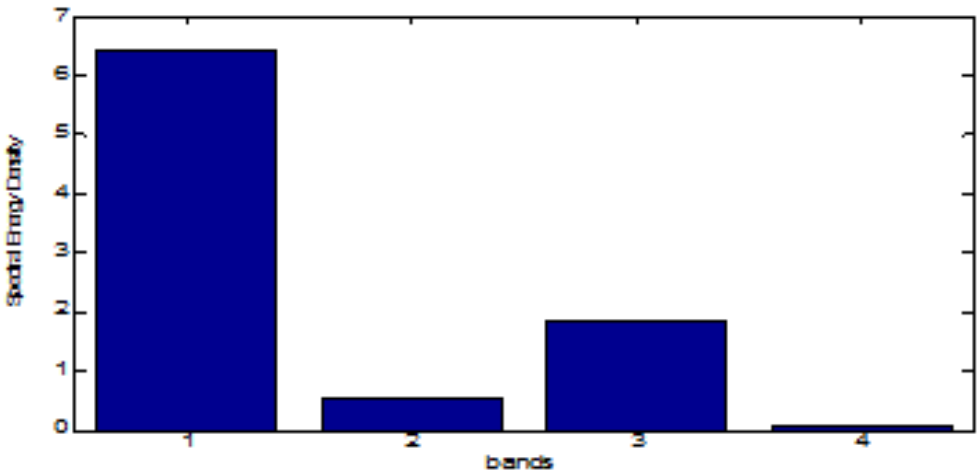

Figure 19: Figure 18 spectral Energy Density for 4-Decomposed Resolutions

The spectral energy density for each resolution is computed using, a power spectral density approach. Each resolution coefficients are averaged by the squared summation of its coefficients and energy is computed. From the resolution energy obtained, it is observed that, resolution 1 and 3 has comparatively higher energy density than the other two resolutions.. Based on the energy derived, two highest energy density resolutions are selected, which is 1 and 3 in this case.

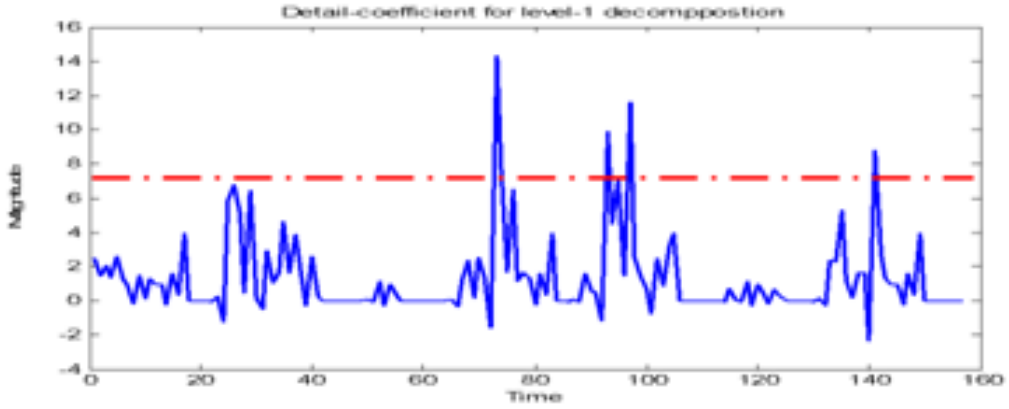

Figure 20: Extraction of Features from selected resolution-1

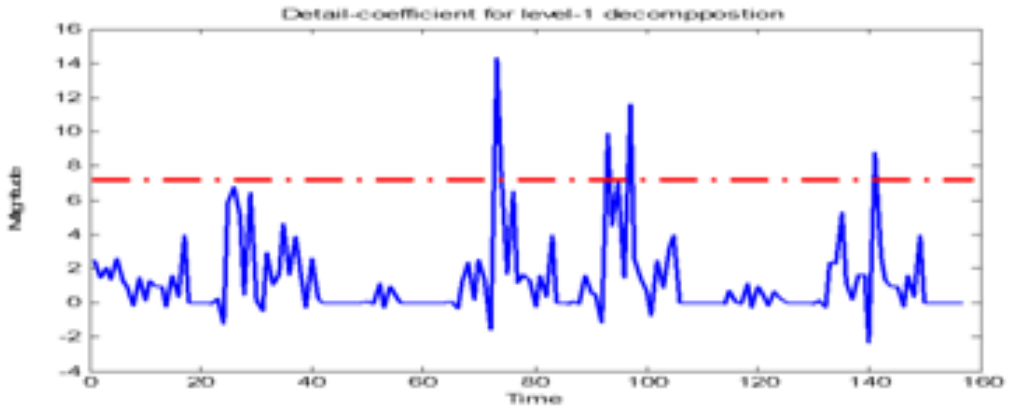

Figure 21: Extraction of Features from selected resolution -3

Figure 19 and 20 shows the two selected resolutions for feature extraction. The prosody features for these selected bands are computed and an average is taken for creating feature information.

following table entries shows Jitter values calculated for only 24 units in database as a sample result out of total 400 sentences.

Table 8: Observed Jitter detected for the given test samples

Test sample	% Jitter (Transform method)	% Jitter (SR method)
1a.wav	20	27
1b.wav	24	33
1c.wav	27	35
1d.wav	24	33
1e.wav	20	31
1f.wav	15	26
2a.wav	19	29
2b.wav	23	34
2c.wav	17	28
2d.wav	23	20
2e.wav	19	30
2f.wav	22	34
3a.wav	19	23
3b.wav	23	32
3c.wav	22	25
3d.wav	23	23
3e.wav	25	34
3f.wav	19	30
4a.wav	20	19
4b.wav	17	29
4c.wav	18	18
4d.wav	20	20
4e.wav	27	36
4f.wav	23	30

Table 9: Observed shimmer detected for the given test samples

Test sample	% Shimmer (Transform method)	% Shimmer (SR method)
1a.wav	12	17
1b.wav	7	9
1c.wav	9	12
1d.wav	11	16
1e.wav	15	21
1f.wav	12	21
2a.wav	6	10
2b.wav	22	26
2c.wav	11	19
2d.wav	9	12
2e.wav	10	11
2f.wav	9	11
3a.wav	12	12
3b.wav	13	15
3c.wav	10	15
3d.wav	11	16
3e.wav	14	15
3f.wav	15	17
4a.wav	12	13
4b.wav	17	15
4c.wav	15	13
4d.wav	9	14
4e.wav	10	15
4f.wav	11	15

Table 10: Observed HNR detected for the given test samples

Test sample	% HNR (Transform method)	% HNR (SR method)
1a.wav	5	9
1b.wav	8	9
1c.wav	2	5
1d.wav	5	8
1e.wav	1	3
1f.wav	8	9
2a.wav	2	6
2b.wav	6	9
2c.wav	8	10
2d.wav	7	9
2e.wav	8	9
2f.wav	5	7
3a.wav	7	6
3b.wav	6	7
3c.wav	3	5
3d.wav	7	6
3e.wav	8	5
3f.wav	6	8
4a.wav	10	9
4b.wav	9	8
4c.wav	4	7
4d.wav	8	9
4e.wav	8	9
4f.wav	8	9

Table 11: Computation Time (sec) measured for the given test samples

Test sample	Computation Time (sec)	
Transform method	SR based coding	Transform method
1a.wav	3	6
1b.wav	2	5
1c.wav	1.5	4
1d.wav	2.5	3.4
1e.wav	2.8	4.9
1f.wav	2.3	4.3
2a.wav	1.9	2.8
2b.wav	2.5	4.3
2c.wav	2.9	4.7
2d.wav	2.3	3.2
2e.wav	2.5	3.1
2f.wav	3.2	4

(table entries shows HNR values calculated for only 24 units in database as a sample result out of total 400 sentences)

The features are extracted using the acoustic-phonetic approach. This Approach considers that there are finite, distinctive phonetic segments in spoken language. These phonetic units are broadly distinguished by a set of properties that are obvious in the speech signal, or its spectrum. Is the speech analysis phase, we here go for spectral representation of the time-varying speech signal. Spectral analysis methods involve filter bank approach for getting spectral decomposition of the speech over time. Next is the stage where we find the features. Spectral measurements are converted in a set of features which delineate comprehensive acoustic properties of the phonetic units It forms a set of detectors which operate in parallel and uses appropriate processing and logic to conform the decision as whether a feature is suppose to be calculated ,or not calculated and if calculated what is the value of it.

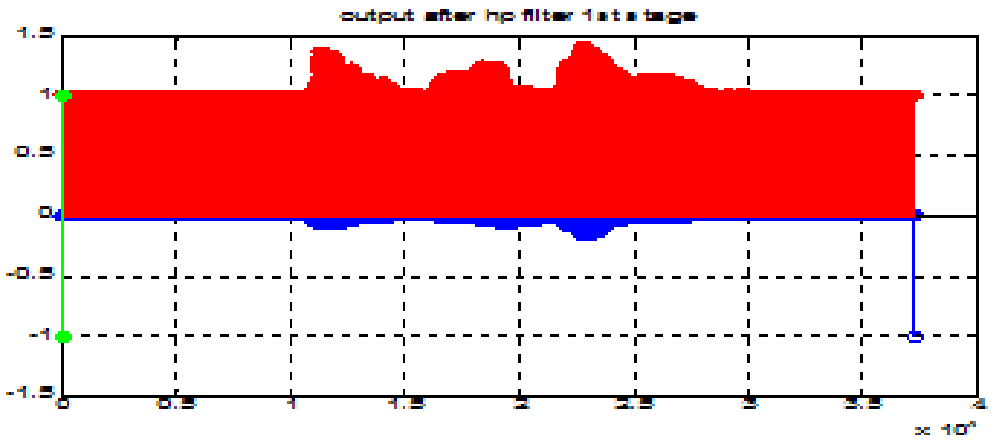

Figure 22: Plot after Decimation 1st stage hp Sub band

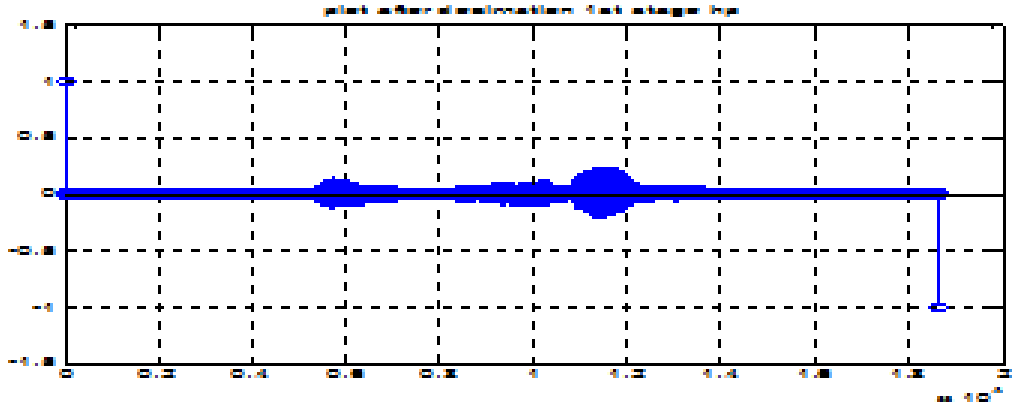

Figure 23: Output after hp 1st stage sub band

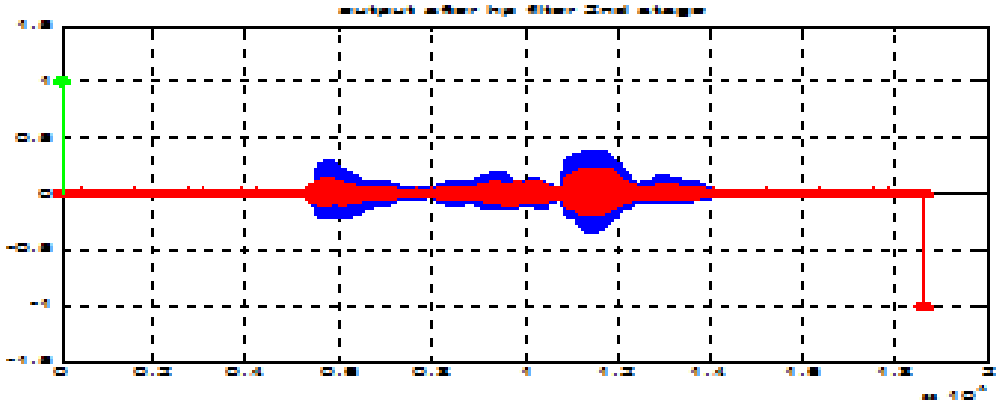

Figure 24: Output after hp filter 2nd stage sub band

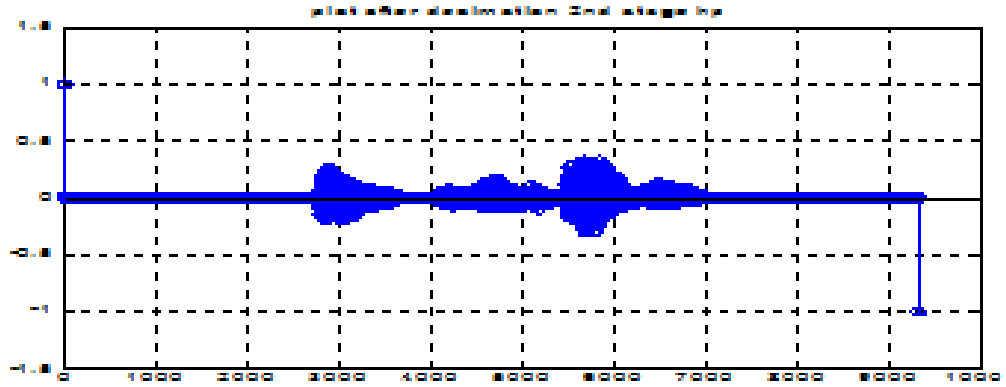

Figure 25: Plot after decimation 2nd stage hp sub band

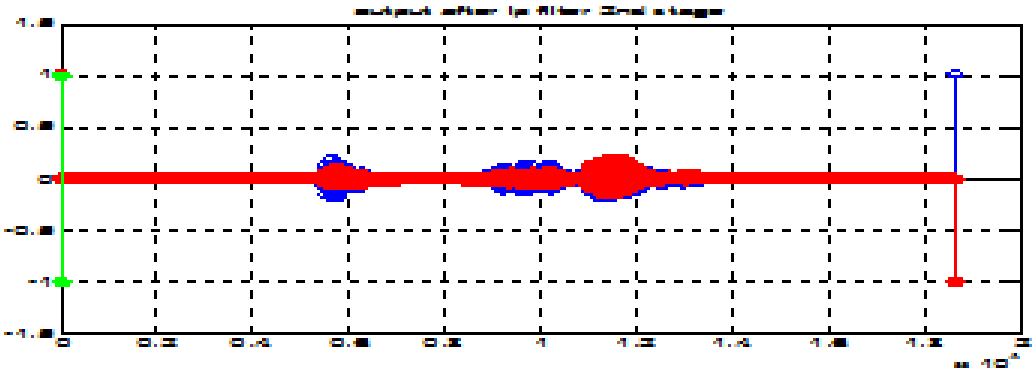

Figure 26: Plot after decimation 2nd stage lp sub band

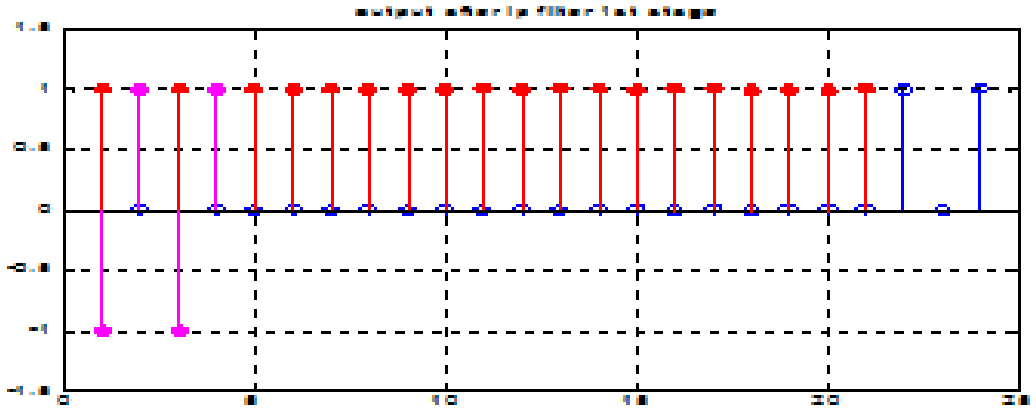

Figure 27: Output after lp filter 1st stage sub band

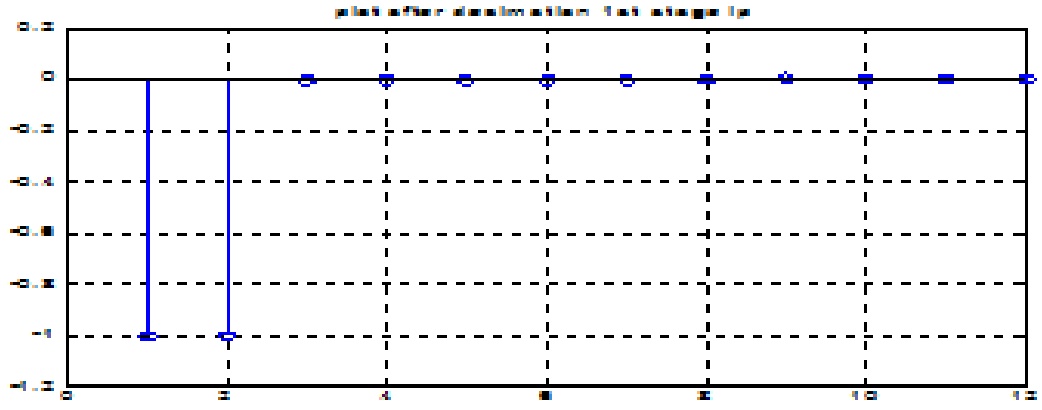

Figure 28: Plot after decimation 3rd stage hp sub band

5.7 Feature Selection

Feature Selection is a practice by which automatic search for the best subset of attributes dataset is carried. The perception of 'best' is relative ,which is based on the classification issue and accuracy expected .In a most useful way a problem is thought of selecting feature is a explore-space search .The explore space is discrete and contains all possible combinations of attributes one can prefer from the feature dataset. The purpose is to steer through the explore space and locate the finest or at least good adequate mixture of features that improves performance of classifiers over selecting all attributes.

Three key benefits of performing feature selection:
Reduces Over fitting: fewer redundant features means less opportunity to make decisions based on noise like interfering parameters. Improves Accuracy: a lesser amount of misleading features means modeling accuracy improves. Reduces Training Time: Less features means that algorithms get trained quicker. Feature selection is separated into two parts:

Attribute Evaluator

Search Method.

Each section has many techniques to choose from. The attribute evaluation is the procedure in which every feature in feature set (also called a column or feature) is evaluated in the context of the output class. While the search method attempts a try for diverse combinations of feature in the dataset in order to turn up on a small inventory of selected features.
Another popular feature selection technique is to calculate the information gain.calculating the information gain for each attribute for the output variable. Entry values vary from 0 (no information) to 1 (maximum information). Those features which contribute extra information will have

a upper information gain value and can be chosen, where as those which do not append much information will have a lower score and can be uninvolved in the processes.

5.8 Hybrid Approach:

Feature selection method is a guide, each method gives different hints about what features might be important. Each feature set/subset can be used as input to train a new model to be compared to a baseline or ensemble together to compete with the baseline.Worse performance after feature selection still has a meaning. For final listing of feature set we represent a hybrid approach as shown in figure 28.

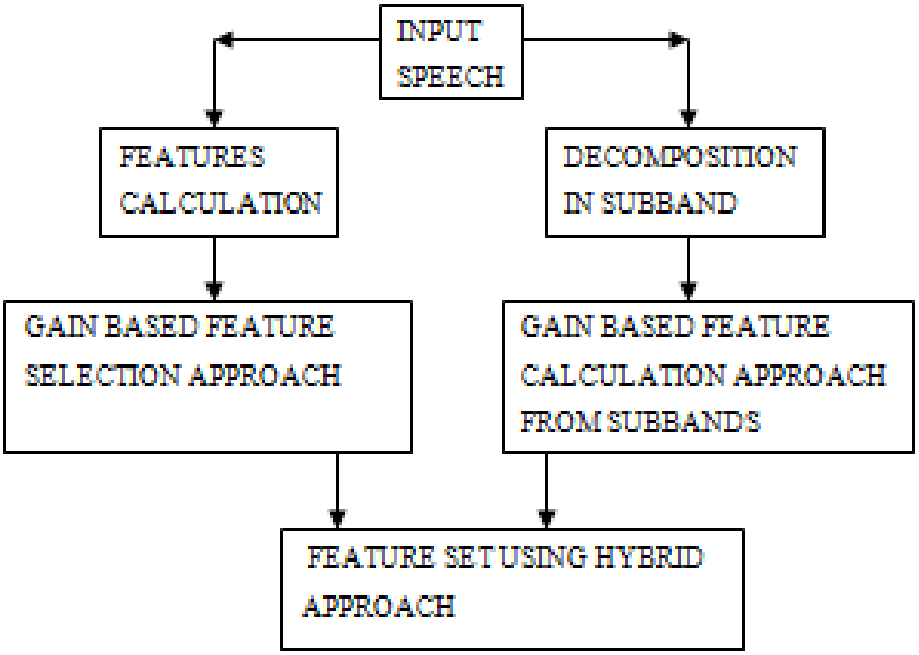

Figure 29: Feature set preparation by hybrid approach

Table 12: Structure of the basic, spectral and supplementary feature set used for evaluation

Feature	Value	Type	Feature	Value	Type
F0	Std	P	Energy	std	B
F0 max	Max	P	Energy max	Max	B
F0 Dev	Std	P	Energy Dev	Std	B
F0 min	min	P	Energy min	min	B
f0_range	Rel min-Rel max	P	Energy _range	Rel min-Rel max	B
f0_maxavg_diff	Std	P	Energy_maxavg_diff	Std	B
f0_avgmin_diff	Std	P	Energy_avgmin_diff	Std	B
Slope	dF0/dt	P	Jitter	abs	P
N	d2F0/dt2	P	Jitter	Median	P
HNR	Mean	S	Shimmer	abs	P
HNR	Std	S	Shimmer	Median	P
MFCC	Std	B	Formants	Mean	B
DVB	Std	S	DUV	std	S

5.9 Classification

In the year 1959, Arthur Samuel explained machine learning as a"Field that gives computers the ability to learn without being explicitly programmed " [Simon, 2013]. Later a more formal definition was provided.A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience [Mitchell, 1997]. The key concept of machine learning is learning from data, since data is what we have [Marsland, 2009]. Now, more and more machine learning algorithms have been developed and they are widely used to help in many applications to investigate multifaceted learning problems; for example, human voice recognition, Auto pricing system, computer gaming and automation in almost everything. All these learning algorithms are classified into basic 5 groups: supervised learning, unsupervised learning, semi-supervised learning, reinforcement learning, and evolutionary learning.
In broader way, pattern recognizers used for speech emotion classification can be categorized into two broad types as showen in fig 29. namely
1. Linear classifiers and
2. Non-linear classifiers.

5.9.1 SVM Classifier:

Support vector machines are supervised learning algorithms in its state-of-the-art .It is used for classification and regression analysis. A support vector machine maps the original dataset to a higher dimensional space. support vector machine builds a hyperplane or set of hyperplanes with adequately high proportions in the new space. The support vector machine

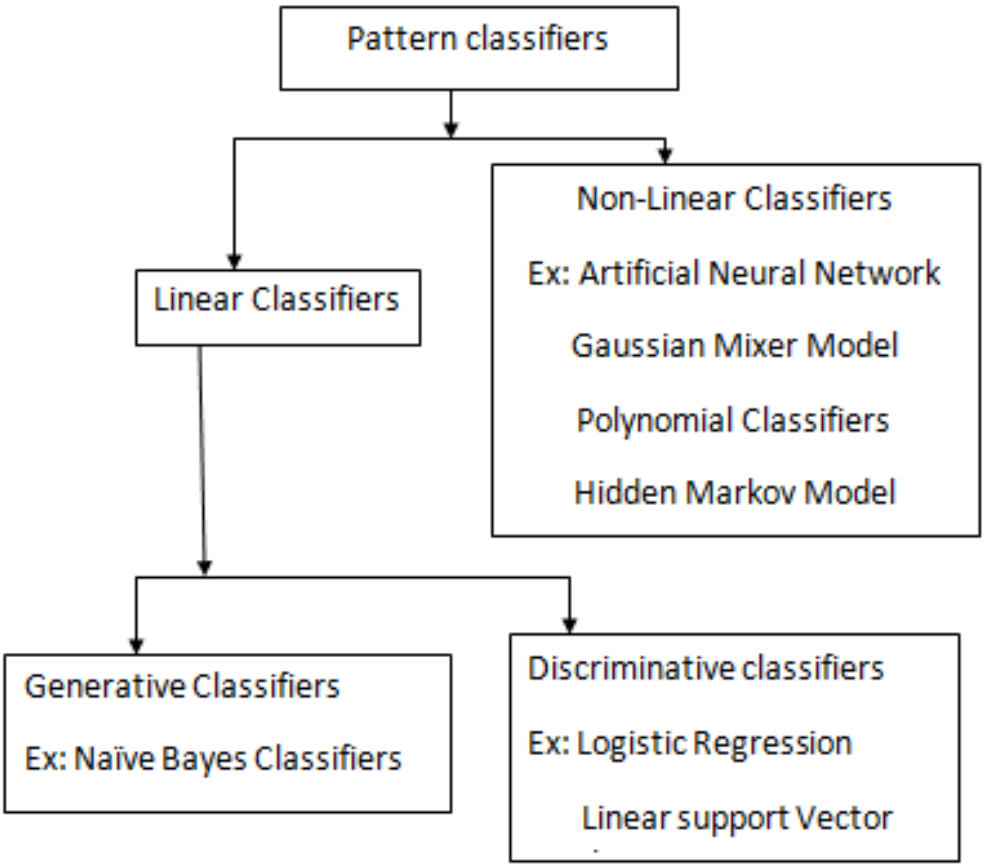

Figure 30: speech emotion classification approach

was first introduced in [Guyon, Boser and Vapnik, 1992]"The key features of SVMs are the use of kernels, the absence of local minima, the sparseness of the solution and the capacity control obtained by optimizing the

margin."Shawe-Taylor and Cristianini (2004).Here, classification of emotion for every speech expression in the database is carried out. Each sound segment recorded ranges averagely from two or three seconds in length. We split every speech sample into 25 ms hamming window segments with 10 ms overlap in successive shifts. As the alter in acoustic features di-

rectly indicate to changed emotional states. The SVM approach is a high dimensional vector supervised learning method that is based on emotion assumptions .It estimate that the occurrence (or absence) of a particular feature of a class is not related to the presence (or absence) of all other features. Its training is very fast and efficient still on very big databases and its correctness is fairly better in comparison to the other available techniques. To start with audio samples are processed and selection of appropriate audio parameters as a feature for extraction is decided. The basic emotions those have to be identified are: angry, happy, neutral, interrogative and sad. We run 3-fold cross-validation tests at the whole dataset of four hundred sentences for all of the built-in kernels (kernel parameters in the parentheses): linear, polynomial (polyorder=3),MLP (default scale [1 -1]), and RBF (sigma=200).Here, we have used 3 fold cross validation in which result obtained are shown. While carring out the training phase, this has been implicit that there must be sufficient amount of learning examples for training model otherwise it leads inaccurate output. As an illustration while the classifier model is learning for all emotions in a database. As per the accuracy level and recall value suitable kernel term for all classifier models have been represented .In our experiments, the SVM toolbox in MATLAB with built-in kernel functions is used. In Our set of experiments, we train classifiers for the interrogative ,Happy ,Angry ,Sad, exclematory,and neutral emotions. We run 3fold cross-validation tests on the entire dataset

Here we have 400 sentences to be classified .we Split randomly the data in K=3 groups with roughly the same size Taking turns using one group as test set and the other k-I as training samples the same size Taking turns

Table 13: 3 fold cross validation

	TRAIN	TEST
RUN1	1,2 (266 samples)	3 (133 samples)
RUN2	1,3 (266 samples)	2 (133 samples)
RUN3	2,3 (266 samples)	1 (133 samples)

using one group as test set and the other k as training samples

Table 14: Hybrid kernel selection based on classifier efficiency

kernels	Classification accuracy	Classification accuracy	Classification accuracy	Classification accuracy	Classification accuracy
	Angry	Happy	Neutral	Sad	interrogative
Linear	78.29	76.45	78	66	79.23
Gaussian	89.88	72.22	83	58.67	80
polynomial	91	81.45	83.65	57.43	85.43
sigmoid	70	62.45	64.65	48.36	75.32
	RBF	RBF	RBF	RBF	RBF

prosodic features are main component in automatic emotion classification.Extracting useful information from prosodic features is yet a difficult task. Therefore, enhanced feature coding procedures need to be developed for emotion classification.There are a many applications which can profit from superior emotion classifiers. In the health care field, emotion classification will help clinicians in real time environment for the evaluation of psychological disorders arising from emotional difficulties in 'Childhood Apprexia Of Speech' disease. Emotion classification also be an probing future towards elaborating context-aware systems for upcoming consumer electronics devices or services. It is observed that no single kernel function provides an optimal solution. Thus, to address the accuracy gap in

existing emotion classification approaches, we propose an emotion classification method based on speech prosodic features and an enhanced Support Vector Machine (SVM) with a hybrid kernel approach and a fusion method with a relative confidence classification threshold. The proposed emotion classification solution extracts the speech signal's pitch, energy and other acoustic features for SVM learning algorithm.

5.10 Conclusion

In Speech processing, feature space is build based on the transformation kernel to a space of lower dimension which allows post processing stage resulting in more useful information. we present a work which does not consider processing of speech signal for feature extraction in frequency domain as in frequency domain it tends to lose the time related clues. Prosody highly depends on the timing information of the signals An approach to compute features by using Empirical mode decomposition of speech signal is presented. This is effective transformation domain in the field of non stationary signals without losing time information. The effect of distortions observed from the capturing units termed as "system noise" are been eliminated with this approach. This elimination in the distortion content is done in transformed domain(EMD), using spectral density thresholding . Speech coding for accurate prosodic feature extraction is developed by this method. The impact of system noise over the extracted feature and its processing efficiency is been evaluated. The spectral resolution coding derives the prosody features more accurately in comparison to conventional approach of direct application.

Articulation of emotions is a multi-modular action. In this way, different modalities like face expressions, bio-signals might be utilized as the strong proof alongside the speech signal for building up the strong emotion recognition frameworks. The effect of emotion expression likewise relies on the phonetic substance of the discourse. Recognizable proof of emotion striking words from enthusiastic discourse, and the elements extricated from these words alongside other traditional components may improve emotion recognition execution. Real time applications, for example, call investigation in the crisis administrations like rescue vehicle

and fire unit, check of emotion to dissect validity of solicitations is vital. In this specific circumstance, under the system of feeling check suitable components and models can be investigated. Most of the today's emotion recognition frameworks encounter high impact of speaker particular data amid emotion arrangement. An effective method might be created to expel speaker particular data from the speech expressions.

5.11 Contribution of this stage work in the further investigation

Fundamental frequency, Energy, Duration happens to be prosodic features, responsible for introducing prosody in the speech segment Here we significantly wish to indicate that as energy value shows high spectral correlation for the entire utterance in the any segment such as Neutral or Prosodic its absolute value will not much convey the conformation about prosody. It leads to a pointer towards identifying a prosody in speech in by statistical variations in the features

Chapter 6

6. Prosody transformation

prosody in a speech conveys meaningful information along with context. Appropriate prosody modification in synthetic plain Speech plays a vital role in developing an effective speech interface platform.for the applications like human computer interface ,multimodal interface and auto responsive systems. This research study focus on Marathi regional language. It presents a spectral mode decomposition (SMD) approach for Spectral correlative prosodic mapping to achieve emotion transformation. The approach calculates the prosodic features for mapping and before replacing it on target segment its value is adaptively recomputed to minimize the spectral errors between source and target emotion segments. The synthesized speech using prosody modified features with minimized spectral errors gives a better quality of the the target expression. This approach give better quality as the features are calculated and not used for transformation as it is ,they are fine tuned to represent the least spectral error between source and target segment This can also be observed by the waveform, spectrogram and objective measures.

6.1 Transformation system outline

A proposed prosody transformation approach for speech segment is outlined. Figure 30 is the block diagram for the proposed emotion transformation approach.The process of speech prosody transformation is carried out in two fold process, of training and testing. In the training phase,

a database of recorded 400 speech segments of a subject with diverse emotions is recorded.

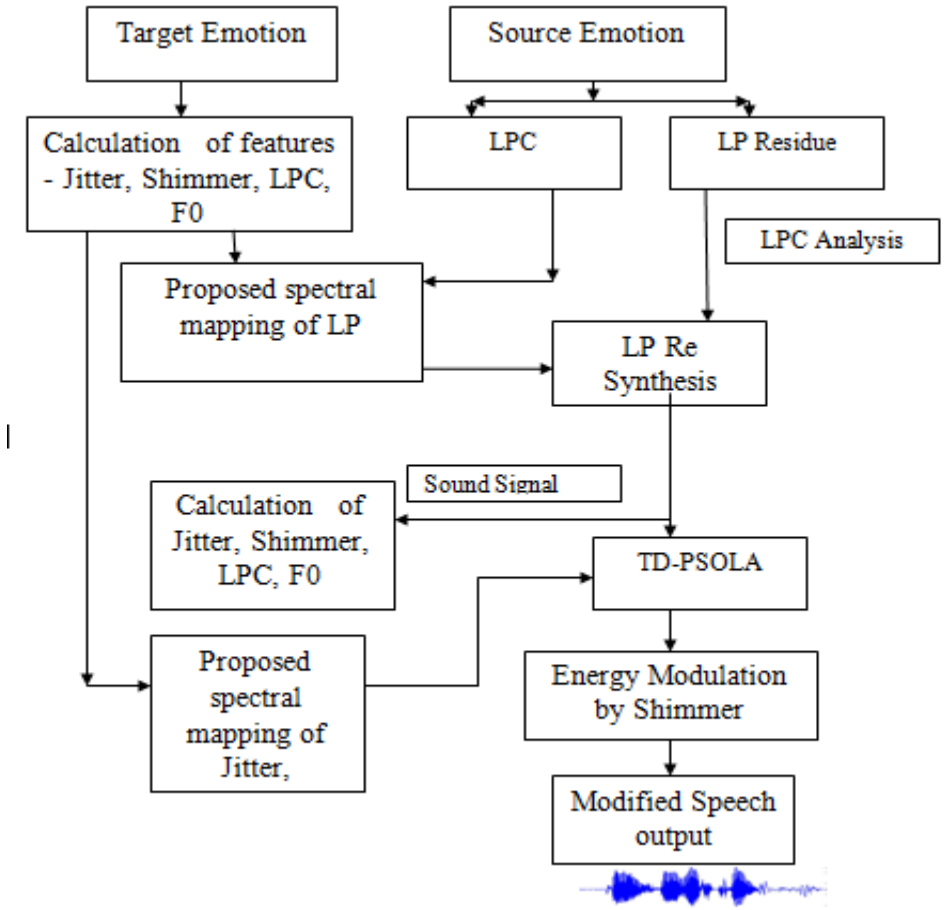

Figure 31: System architecture for proposed emotion transformation

The recording is carried out at acoustically standard environments to obtain highest degree of clarity. These speech samples are recorded with one voice over artiest .recording is done for 6 emotions including neutral.Total database comprises 400 sentences. Sentences are preprocessed, where the samples are processed for silence removal first. Recorded signal is then subjected to energy normalization to a certain level. Threshold of the energy filtration is a function of listeners sensitivity and acoustic

environment at the time of recording. Human hearing range goes from 0 dB to 110 dB.After pre-processing of signal samples, prosodic features are extracted, these features are under two classes one is for the reconstruction of the signal and other category is for supra segmental modification in the transformation. In lieu of speech re synthesis LPC is chosen to be extracted which allow us to modify spectral and prosodic parameters in less complex way. So far in the pitch modification, only the pitch period of the signal was altered by keeping the LP Coefficients (LPCs), unchanged. Therefore the output speech is unnatural due to lack of correspondence between the source and the vocal tract system. To provide naturalness, one should alter the vocal tract shape according to the pitch periodicity. This comes from the fact that if F0 is changed correspondingly formants also changes .That indicates close correlates in the vocal track parameters and the glottal source parameter as F0. For carrying out this, the interrelation between the source and the vocal tract should be known .LP residues are required to be processed for deriving prosodic parameters as source information. Mainly pitch related information is derived for the emotion transformation .some other features responsible for the prosody transformation are Harmonic to noise ratio (HNR), degree of voice breaks (DVB), degree of voiceless (DUV) and jitter, shimmer, these computation of feature will be performed over the entire training sample and an acoustic model with voice quality feature is represented. The linguistic rules will be extracted from the speech signal pronounce. These linguistic clues are then passed to a expression classifier where the expression classification is performed through SVM scheme. The classified emotions with their corresponding features are then recorded to formulate feature dataset. During the test process, a neutral plain sample is

passed to a system, which is processed with the same process as carried out in training. The pre-processed sample is then processed for feature extraction and passed to a mapper logic .Based on the text input clue from the user prosodic feature will be derived from the dataset. The mapper logic transforms the neutral feature to the demanded emotion feature by magnitudal and spectral alignment to obtain the transformed synthetic speech. In the development towards the speech synthesis, in a synthesis approach based on frequency mapping is suggested. The fundamental frequency of neutral speech signal is superimposed with the expression targeted to generate a synthesized expression speech in Hindi language. The approach defines a frequency contour and intensity contour for superimposing the neutral features with the targeted emotion features. The synthesizing tool of PRAT is used as a processing tool for this application. However, the basic observation of the developed approach shows a constraint in transformation accuracy of achieving a MOS of about 4-5 for the developed approach. The superimposition of feature coefficients are effected by harmonic distortion and feature variations, hence a direct mapping results in lower transformation accuracy. To improve the suggested approach, in this paper, a new method of transformation based on recurrent spectral mapping is suggested. The proposed approach is outlined in next section.

6.1.1 lexical analysis :Text based prediction

Speech emotion transformation process takes a speech segment as input which is further transformed in to target prosody. For this task it requires knowledge of target emotion in which transformation of input needs to be done.The features of source are computed for the conversion. Target emo-

tion needs to be conformed so a process of carrying out transformation from the source to target prosody is done. For this task we are doing a lexical analysis.th this phase we classify the text and get prosodic clues from the 6 classes we have already created in database. We are using WEKA Tool kit for accomplishing the task as follows

Framework Tools : We used weka toolkit for performing classification. Weka is an open source toolkit developed by Machine Learning Group at University of Waikato it suggest many algorithms for functioning tasks of classification and clustering as well as it also ropes other functionalities for performing text analytics.

Naive Bayes

Naive Bayes is supported in built classifier in WEKA.It apply Bayes Theorem, "which finds the probability of an event given the probability of another event that has already occurred ". We used Naive Bayes implementation available in Weka toolkit.

In Indian Laguage scenario, past work done for emotion analysis has been done with languages like Bengali and Hindi, rest all the major regional languages are uncharted. The temperament of Indian languages varies a great deal in terms of the script and dialect as well as it surprisingly varies in linguistic characteristics too. So, there is a huge quantity of efforts that needs to be done to understand the behavior and execute the analysis of same consequently. A lexicon that we created as a part of this research work is just the basic structure working on special tags for class identification and a resource in which words form a connected database. We have manually aligned prosodic tags and database corpus for Marathi. In upcoming time , an attempt can be given a try to more specific approach and extra heuristics to develop. figure shows the process flow for lexical

analysis.

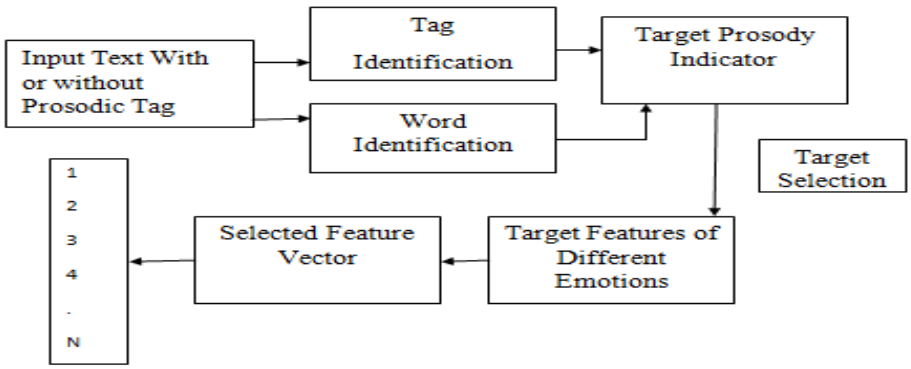

Figure 32: Lexicon classification in text input for target prosody detection

6.2. Prosody conversion framework using spectral correlative prosodic mapping

For prosody conversion using feature mapping approach, the speech signal is coded for feature extraction, these features are then used to do prosody transformation. The features are derived based on our past developed work where, a Empirical mode decomposition (EMD) used to decompose the signal represented in time domain into a set of adaptive basis functions called intrinsic mode function (IMF). The IMF is defined as the oscillatory elements of the signal which contains zero DC elements. Derived n-IMFs which are in tern representing different frequency elements in a signal along with time information. Following figure explains the procedure for finding IMFs.In this effort, we haul out the prosodic features as Harmonics to Noise Ratio (HNR), Jitter and Shimmer. Jitter and Shimmer are second order features representing statistics of variation in pitch and energy . The Jitter is defined as the average absolute difference between two consecutive periods and The Shimmer is defined as the peak to peak

variation of amplitude measure in successive frames. It is the average absolute of the variation between the magnitudes of successive periods. The suggested approach performs the task of prosody generation over the selected frequency components based on spectral power density. It is observed that the noise and silence parts are having noticeably lower spectral magnitude, so this approach successfully deprives the consequence of distortion and results in higher efficiency. In the procedure of spectral coding, a speech signal x (n) is processed through Spectral mode decomposition to obtain intermediate frequency components. The spectral information is defined as the energy contents derived over the speech signal. This approach defines a temporal and spatial localization of transformation model based on Spectral mapping.

In the process of temporal coding, the dataset is categorized into a set of temporal and spectral clusters, where$H_{t,class} \epsilon H_{t,dataset}$. The categorization of dataset leads to faster localization of transformation model in the dataset search. In the first process neutral features are extracted and it is represented as a neutral spectral cluster Hq. During training phase ,a intersection of spectral feature for temporal localization of speech coefficient is used. We define here the spectral intersection $s\left(H_q, H_t\right)$ as follows;

Step1.

Spectral intersection $s\left(H_q, H_t\right)$ $defined\ as$ follows;

$$s\left(H_q, H_t\right) = \sum_{i=1}^{k} \left(\frac{min\left(H_q \ \ H_t\right)}{H_n} \right) \tag{1}$$

Where H_q is energy of query and H_t is energy of cluster in the data set is denoted as Hn

Step 2

Spectrum of the ith sample of speech segment is Hi , N indicates the total number of frames in a segment and M are the total entries in the corpus

$$H_i(k) = [H_i(kN),\ H_i(kN-1), ..., H_i(kN-M+1)] \quad (2)$$

Step 3 The frames are matched in their spectral properties and error is calculated .An error function is defined by,

$$e_{i,H}(k) = H_{i,t}(k) - H_{i,t+1}(k) \quad (3)$$

This error gives the measure of the difference in the spectral energy in two frames, and the successive frames having lower spectral errors $min(e_{i,H}(k))$ are taken for the further processing that is for calculation of the features.

Step 4. To derive the synthetic speech signal, the spectral bins are initially processed to obtain spectral features for target prosody, which are denoted by

$$F(k) = [F_0(k),\ \ F_1(k), ..., F_{M-1}(k)]^T \quad (4)$$

When these feature vectors are transplanted onto the plain sentence result is prosody transplanted speech.

$$\text{Hi new (k)} = \text{Hi(k)F(k)} \quad (5)$$

When superimposing of target features are done on source sentence due to inefficiencies in mapping logic, some coefficients from original sentence appears in the output of the system. This hampers the quality of the transformed speech. $e_{i,H}(k)$ Error function calculates the difference between recorded target and transformed target. We kept permissible error value at 0.1

$$min(e_{i,H}(k)) = H_{i,t}(k) - H_{inew}(k) \quad (6)$$

$$min(e_{i,H}(k)) = H_{i,t}(k) - H_i(k)F'(k) \quad (7)$$

Where Updated spectral features$F'(k)$ for target prosody, are denoted by

$$F'(k) = \left[F'_0(k),\ \ F'_1(k), ..., F'_{M-1}(k)\right]^T \quad (8)$$

Where K represents number of features and M are total recordings in prosodic database.

Individual terms are defined as

$$F'(k) = F(k) + P_s\ (k) \qquad \text{s= \{ 0 ,1,2.... M-1 \}} \tag{9}$$

$$F_0^{'}(k) = F_0(k) + P_0\ (k) \tag{10}$$

$$F_1^{'}(k) = F_1(k) + P_1\ (k) \tag{11}$$

.

.

.

$$F_{M-1}^{'}(k) = F_{M-1}(k) + P_{M-1}\ (k) \tag{12}$$

We define updation coefficient as

$$P_s\ (k) = \mu \sum_{i=0}^{N-1} \frac{H_i^T(k)}{\|H_i(k)\|^2} e_{i,H,init}(k) \tag{13}$$

The deviation in the spectral bin is then integrated over a period of 0 to N defined by,

$$E(H_{i,N}) = \int_0^N \mu \sum_{i=0}^{N-1} \left(2E\left[\frac{H_{i,n}(k)F(k)e_{i,NH}(k)}{\|H_{i,N}(k)\|^2} \right] - \mu E\left[\frac{e_{i,NH}^2(k)}{\|H_{i,N}(k)\|^2} \right] \right) \tag{14}$$

where integrating the estimate, over 'n' observation period accumulates the evaluation for'n' inter frame errors. For each frame with minimum estimate error is then chosen as the selected spectral bin and a intersection bin is then derived from equation (1).

The developed approach is processed into two phase of execution, of training and testing process. Where in training process set of speech sample of different emotions, having angry, sad, happy and exclamation is taken, and these samples are processed for feature selection based on

Spectral features derived. These features are then processed for mapping using a spectral correlation objective, where the correlation factor is defined by,

Correlation factor , $C_i = \sqrt{\sum_{i=1}^{M} H_q - dbH_i}$ (15)

Where, Hq is the neutral feature and, dbHi is the features trained in the data base for different emotions. Subjected to the optimization problem of $min(Cj)$. Coefficient with the spectral correlation with minimum correlation factor is then transformed to the targeted emotion based selected spectral coefficients mapped over the neutral speech signal. To validate the proposed speech transformation a simulation is carried out for different transformations, outlined in next section.

For the evaluation of the suggested approach, a simulation is carried out on a dataset created with 500 recorded speech signals, with 6 emotions, namely, normal, angry, sad, exclamation, question and happy. The performances were observed for the developed approach based on spectral co-relation map coding.

6.3 Simulation Observation

For a subjective evaluation of the proposed approach, a mean opinion score (MOS) test . The MOS observations are carried out as a subjective matter of each volunteer giving a opinion score in 5 scales, as (5: excellent, 4: Good, 3: fair, 2: poor, 1: bad). To evaluate the developed system for this subjective metric, 10 volunteer listeners are played with the transformed speech signals and the individual score is recorded, based on their audibility to the transformed signal in comparison to the original speech signal. The obtained MOS for the developed approach is presented in ta-

ble 18. The observations are presented as a mean observation value for the given score. The simulation observation for the test sample with a neutral speech signal have a voice of ”Aaj pause padto Hai ”, the observations are as illustrated below.

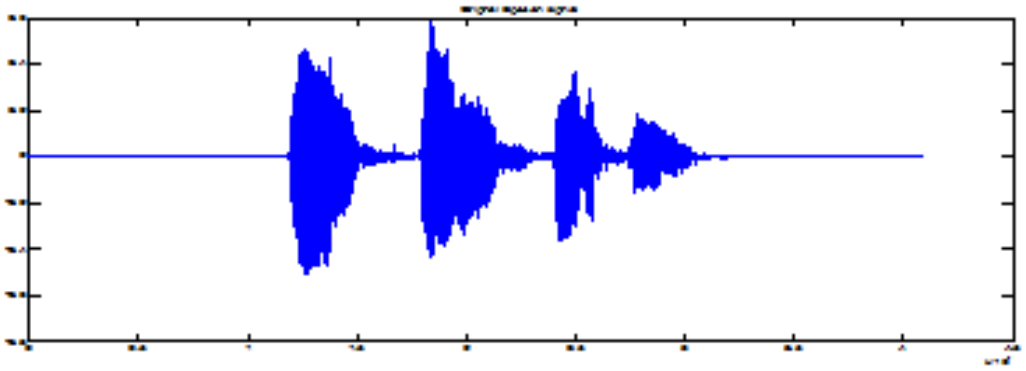

Figure 33: Test sample in neutral format

The test sample read for the testing is illustrated in figure 26. The test sample is read with a coefficient of 45000 coefficient sampled at a sampling rte of 16KHz, 8 bit DPCM coded correlation with minimum correlation factor is then transformed to the targeted emotion based selected spectral coefficients mapped over the neutral speech signal. To validate the proposed speech transformation a simulation is carried out for different transformations, outlined in next section. The test sample read for the testing is illustrated in figure 30. The test sample is read with a coefficient of 45000 coefficient sampled at a sampling rte of 16KHz, 8 bit DPCM coded.

The transformed speech signal for the question formation is shown in figure 31. The correlative mapping for the two test speech signals is shown in figure 34.

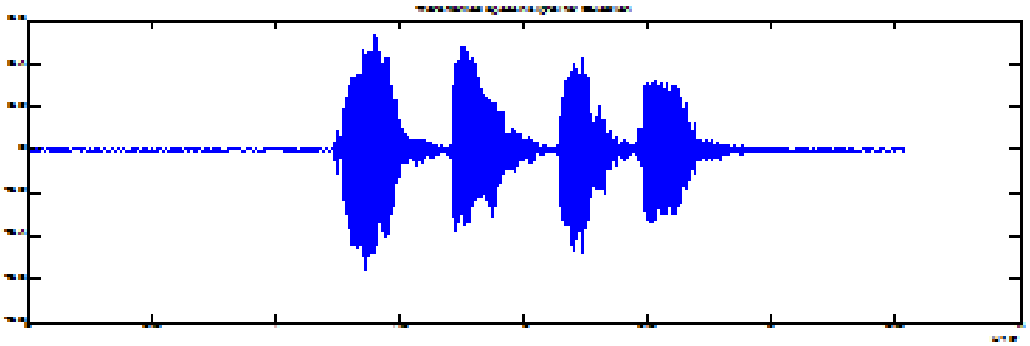

Figure 34: Test sample in Question format

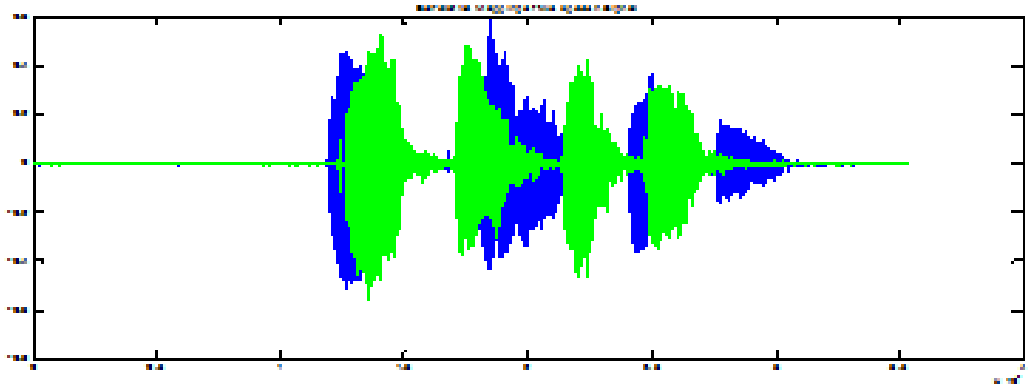

Figure 35: Correlative Mapping of the two samples (neutral and Question)

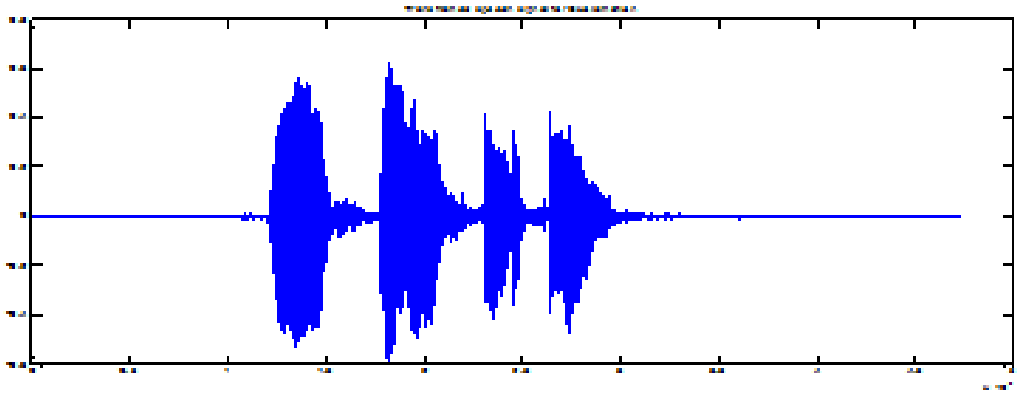

Figure 36: Transformed Speech signal in Exclamation format

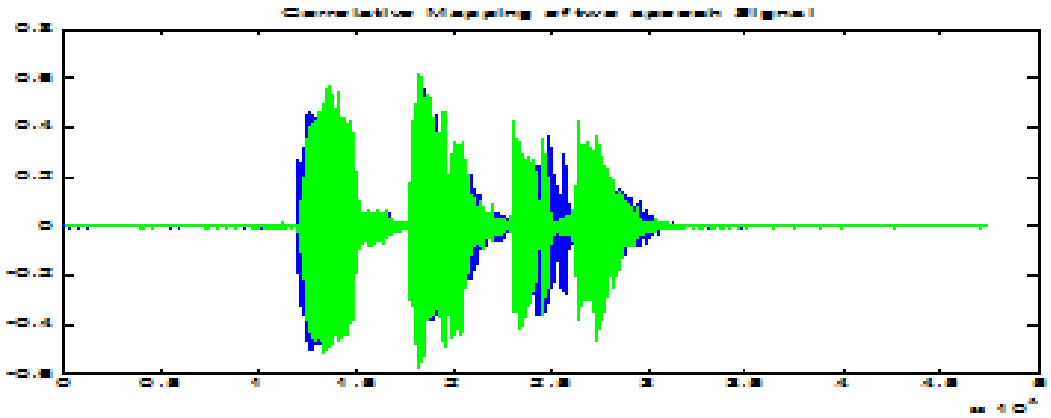

Figure 37: Correlative Mapping of 2samples (neutral and Exclamation)

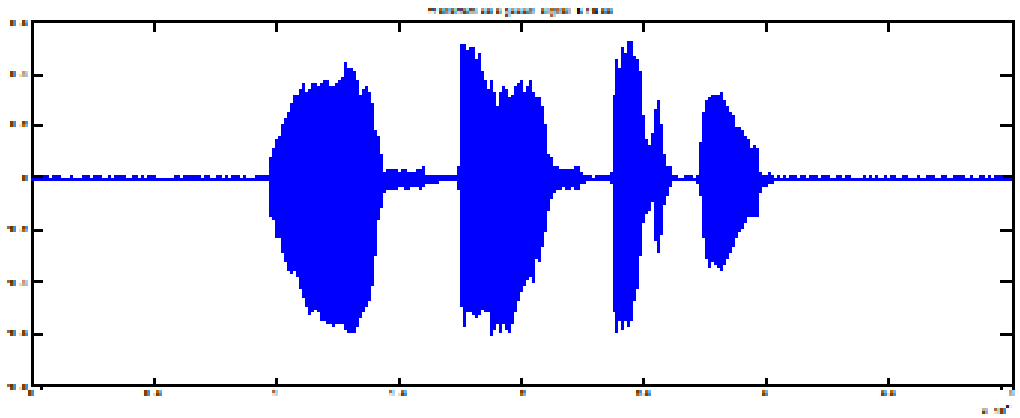

Figure 38: Transformed Speech signal in Sad format

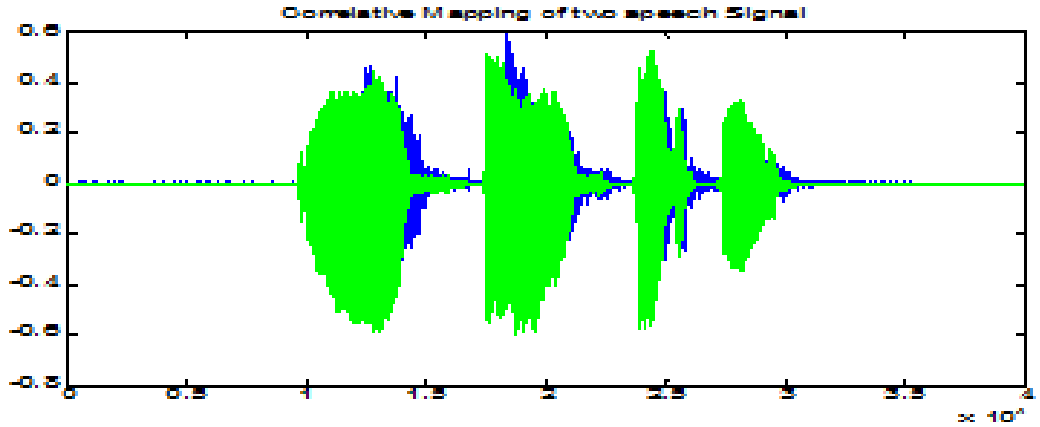

Figure 39: Correlative Mapping of two samples (neutral and sad)

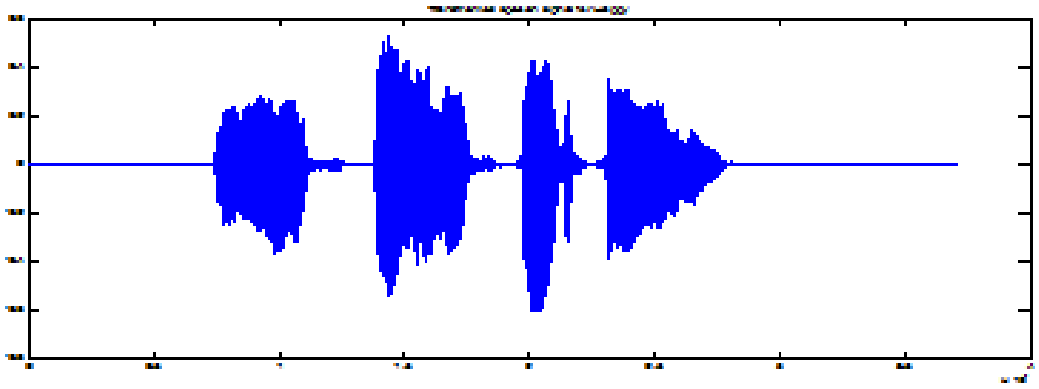

Figure 40: Transformed Speech signal in happy format

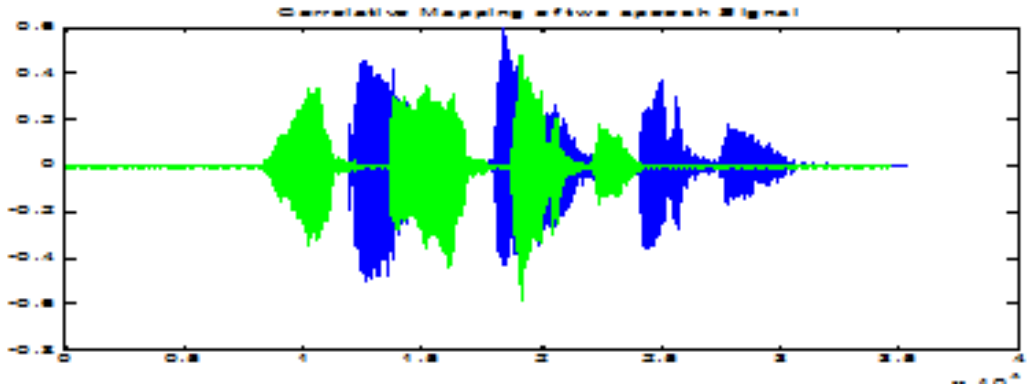

Figure 41: Correlative Mapping of the two samples (neutral and happy)

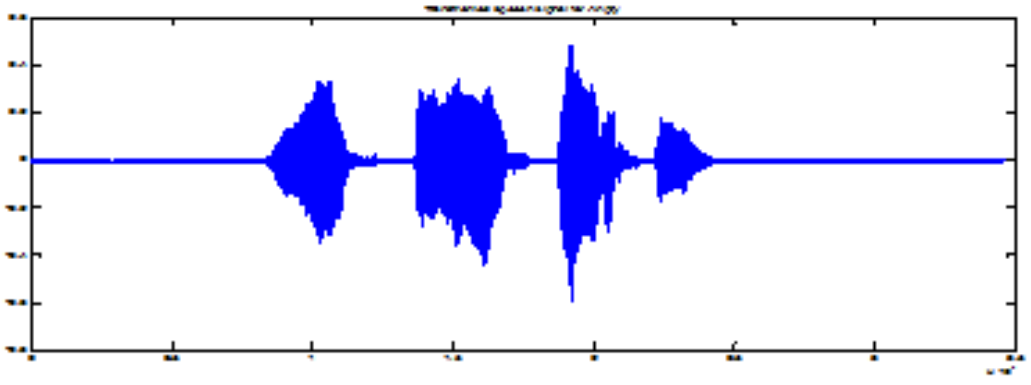

Figure 42: Transformed Speech signal in Angry format

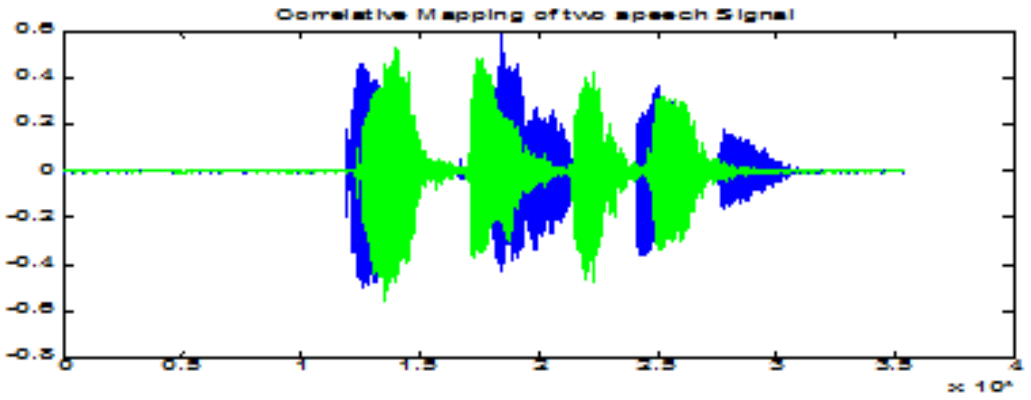

Figure 43: Correlative Mapping of the two samples (neutral and Angry)

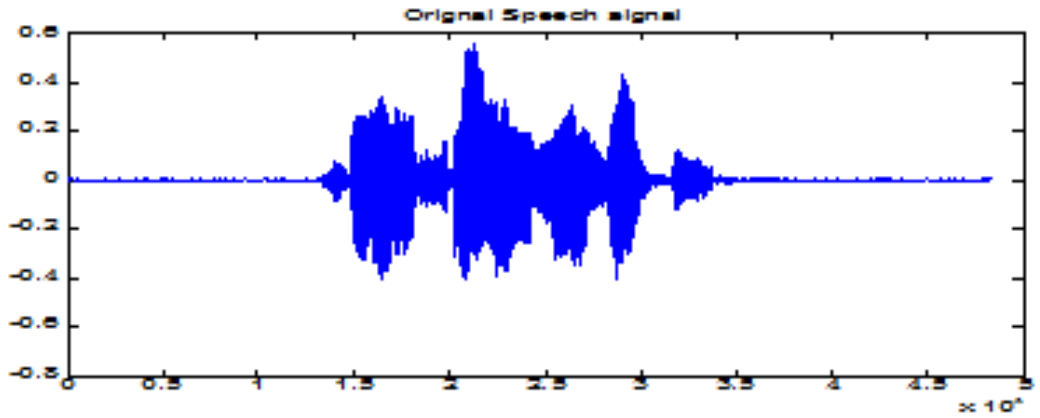

Figure 44: Test sample in neutral format

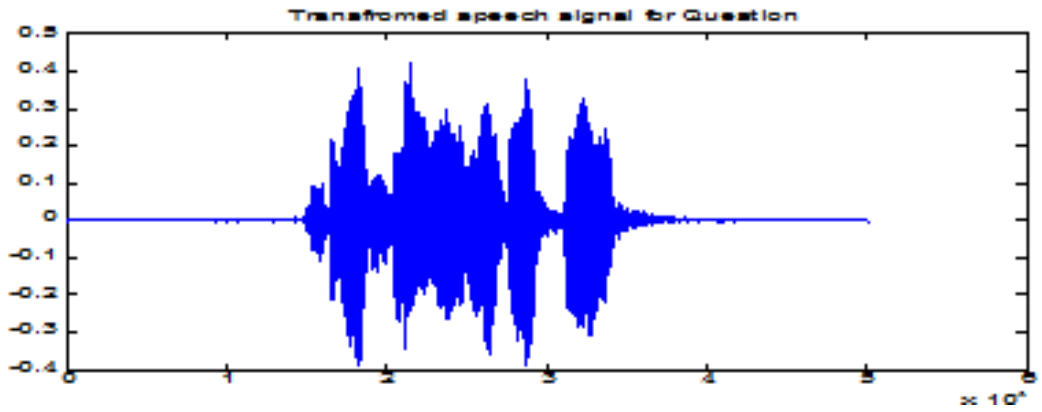

Figure 45: Transformed Speech signal in Question format

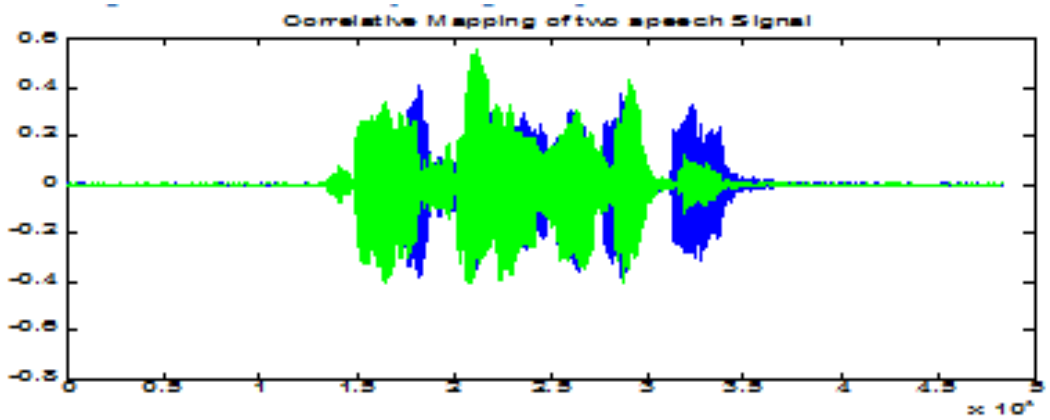

Figure 46: Correlative Mapping of the two samples (neutral and Question)

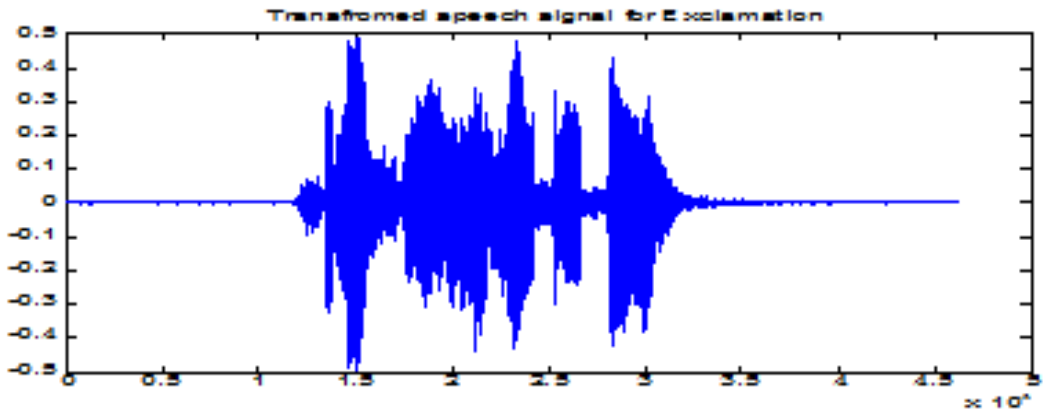

Figure 47: Transformed Speech signal in Exclamation format

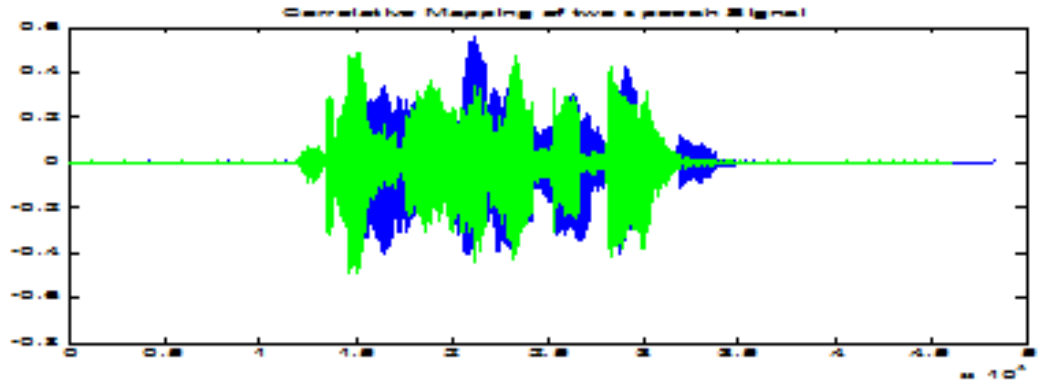

Figure 48: Correlative Mapping of the two samples (neutral and Exclamation)

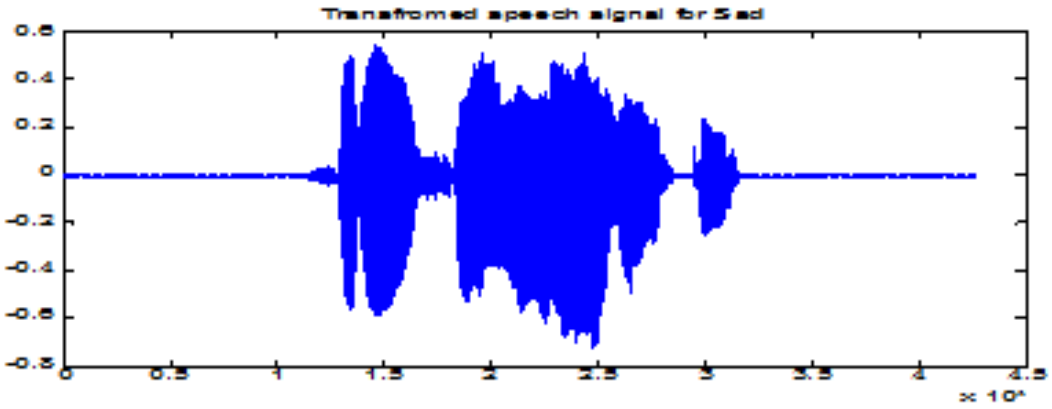

Figure 49: Transformed Speech signal in Sad format

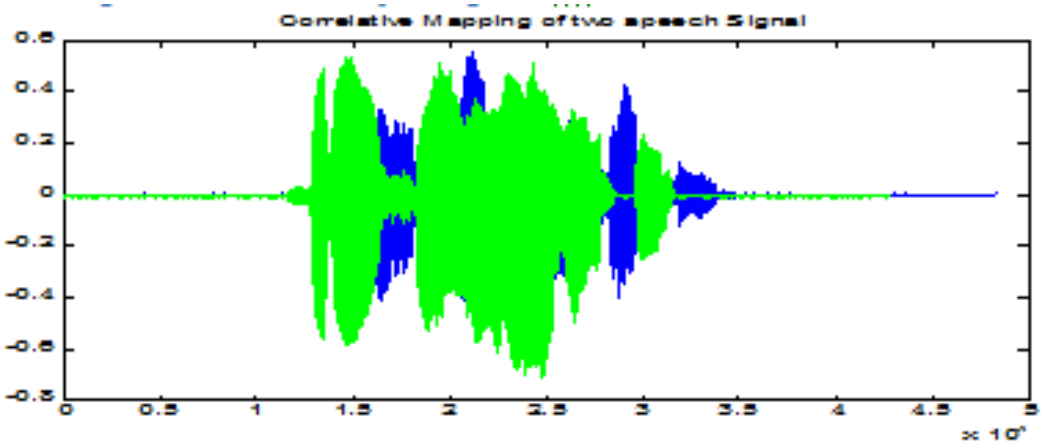

Figure 50: Correlative Mapping of the two samples (neutral and sad)

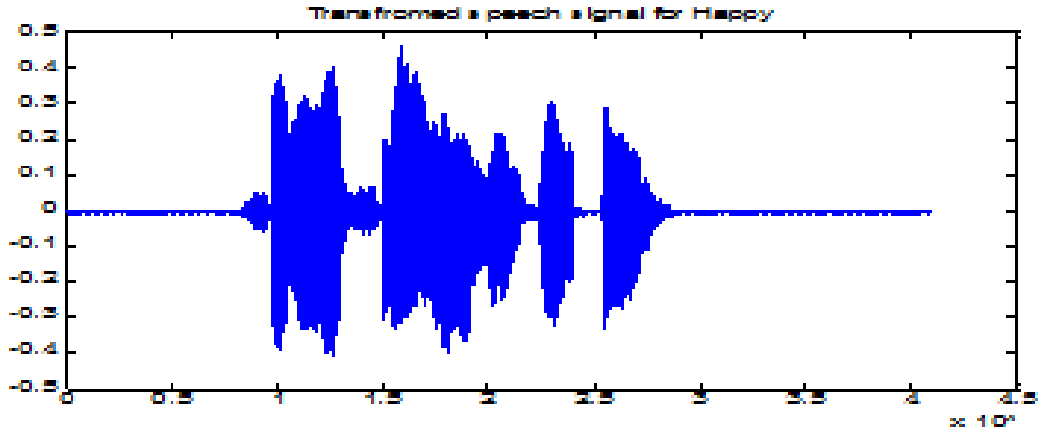

Figure 51: Transformed Speech signal in happy format

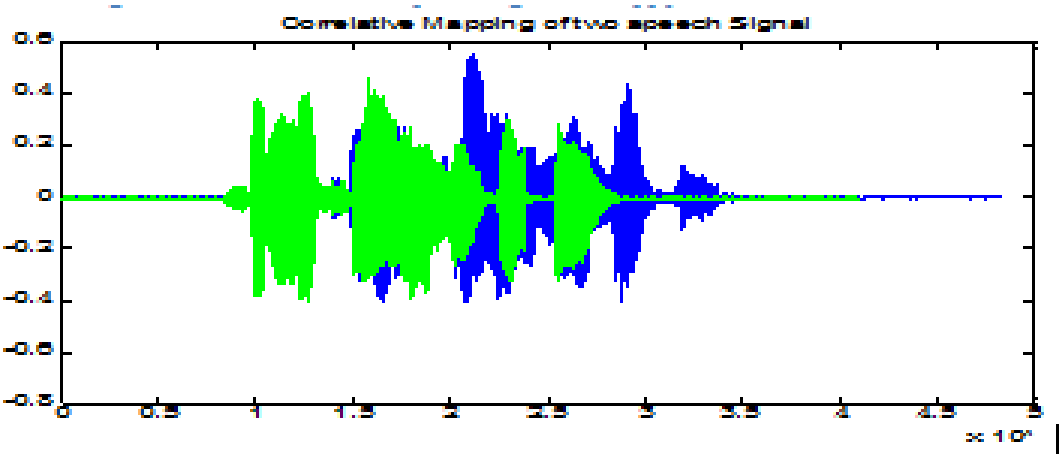

Figure 52: Correlative Mapping of the two samples (neutral and happy)

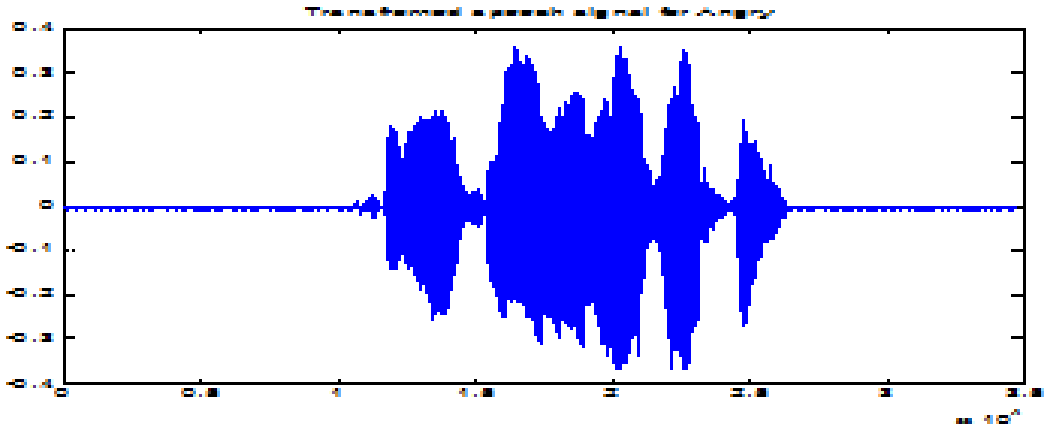

Figure 53: Transformed Speech signal in Angry format

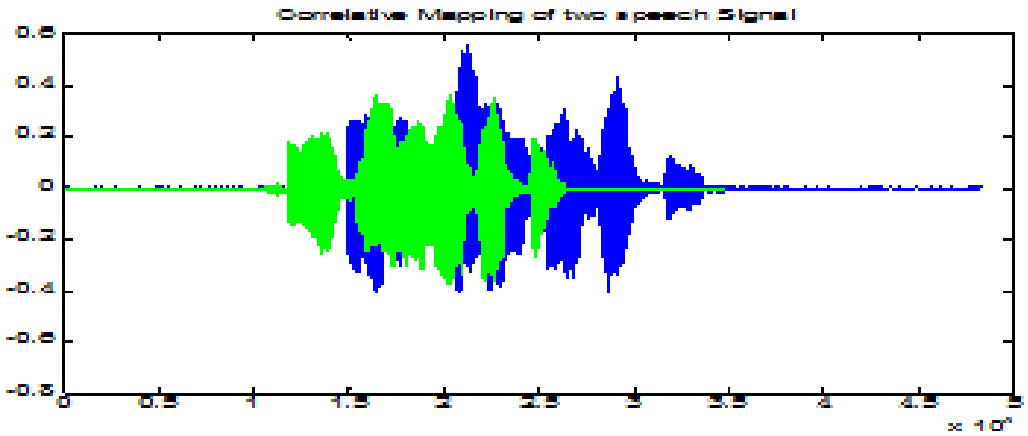

Figure 54: Correlative mapping of the two samples (neutral and Angry)

6.4. Evaluation

In this section we explain the listening test set up and re present the evaluation results. Listening experiment. evaluation of the output emotion categories is achieved by conducting personal listening tests with native listeners. Ten listeners involved in the experiment and each of the listeners was presented with 400 sentences, which consisted of the results
of all modifications as well as the original utterances. The stimuli were offered in arbitrary order in order to eradicate any correlative effects in assessment . use of Headphones with the freedom of the volume adjustment to the current sentence was allowed, however once finished they were not given the chance to come back . The test was structured as a forced-choice experiment, where the raters were necessarily deciding on one of the following five choices: (1) happy, (2) angry, (3) sad, (4) neutral and (5) Interrogative 6)other. The addition of other option was to give a way for including in-between (i.e, fuzzy) emotional categories that can happen as a result of the modifications..following figures shows the results of modified pitch contour and energy contour for prosodical modifications done in the sentence

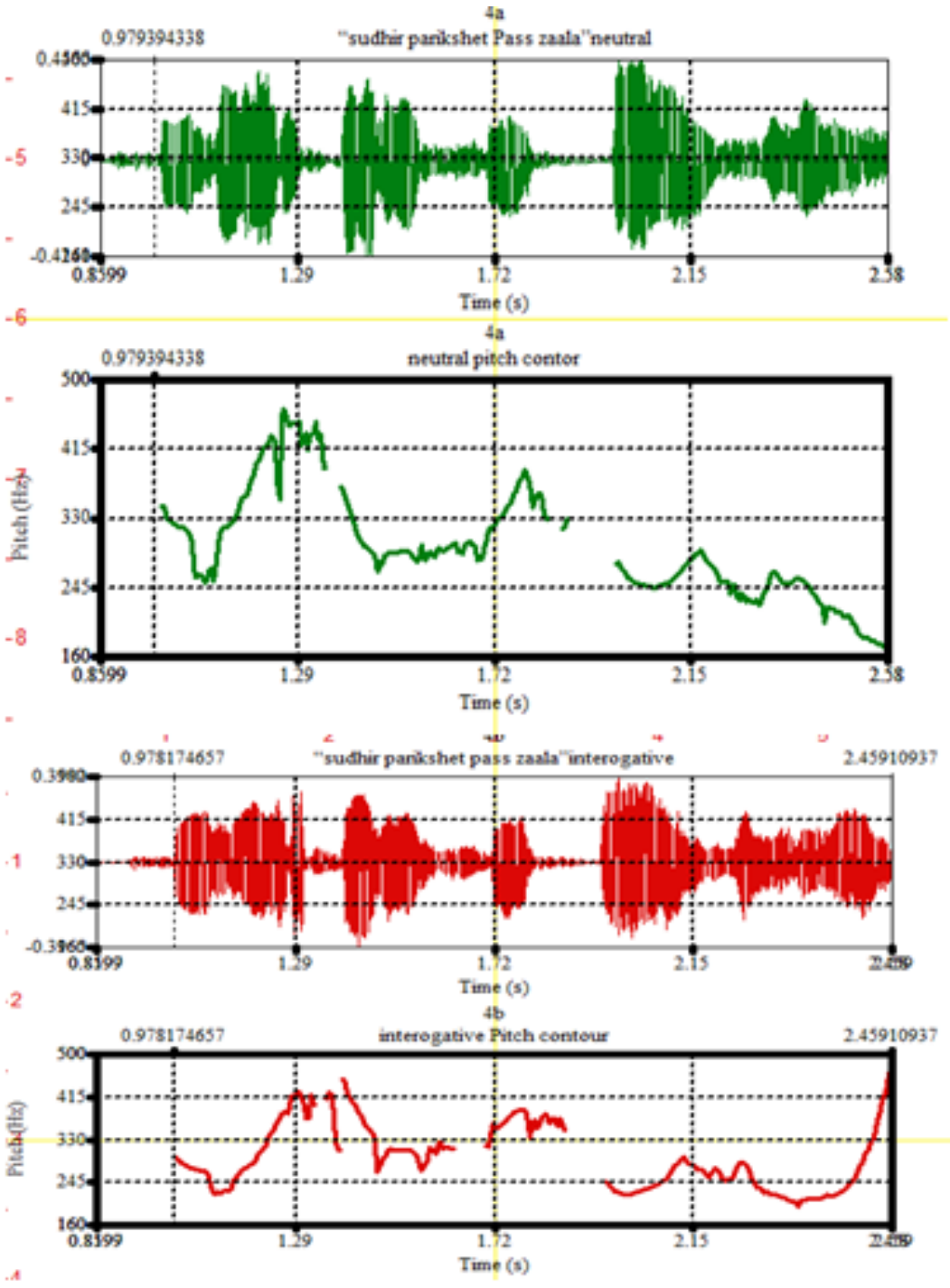

Figure 55: fundamental frequency modification contour of transformed prosodic utterance

Listening test results (in total number) for some selected modifications. h,a,s,e indicate happy, angry, sad and exclamation respectively. The numbers in left parentheses refer to the modification type as explained in section 6.2 .H:A means that source is happy and target is angry

Table 15: Table 18 Result of MOS test

2*Emotion	Neutral To Happy	Neutral to Angry	Neutral to Sad	Neutral to Exclamation	Neutral to Question
	N:H	N:A	N:S	N:E	N:Q
Method					
Frequency mapping [24]	46	39	35	43	45
Proposed Spectral mapping	51	54	37	56	58

Table 16: Listening Test results for Selected modifications (N:H means that source is neutral and target is happy)

Proposed Spectral mapping	Happy	Angry	Sad	Question
N(66) N:H	51	8	0	7
N(66) N:A	2	54	0	10
N(66) N:S	9	5	37	15
N(66) N:Q	0	8	0	58

Table 17: Listening Test results for Selected modifications (H:A means that source is happy and target is angry)

Proposed Spectral mapping	Neutral	Angry	Sad	Question
H(66) H:N	47	2	15	2
H(66) H:A	2	33	5	25
H(66) H:S	25	4	30	6
H(66) H:Q	1	17	2	46

Table 18: Listening Test results for Selected modifications (A:H means that source is angry and target is happy)

Proposed Spectral mapping	Neutral	Happy	Sad	Question
A(66) A:N	47	2	15	2
A(66) A:H	2	27	2	35
A(66) A:S	25	1	39	1
A(66) A:Q	1	17	2	46

parallelly applied prosodic and spectral adaptation present enhanced results than their individual applications. As expected, when all 4 variables (spectrum, LPC , pitch and energy) are modified the results improve. From Table we see that the best results are achieved for following pairs, H:N ,H:Q where as accuracy seen to be low for the pair H:S and N:S.The comparative result of the MOS is shown in figure 55.

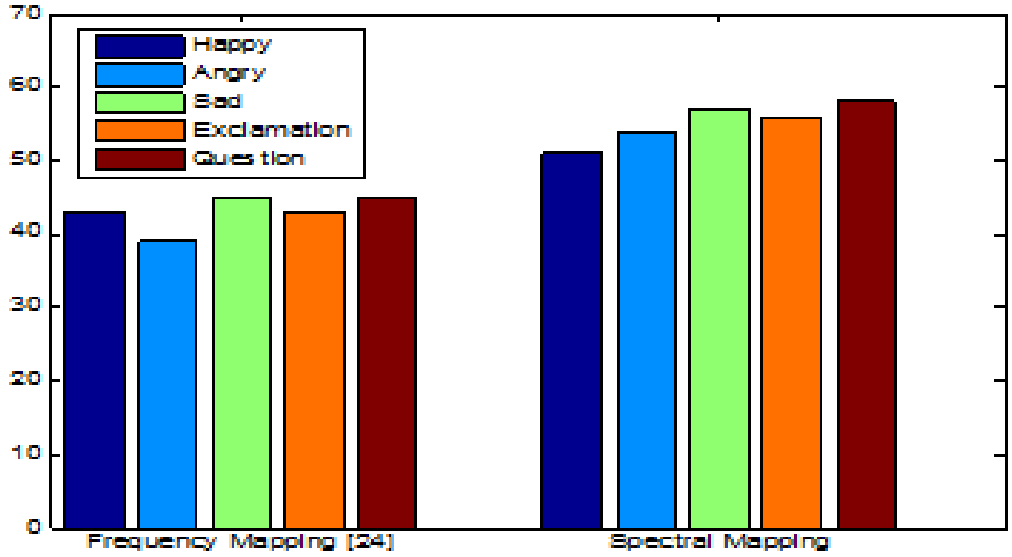

Figure 56: subjective comparison of the test sample under different emotions

6.5.Combination of Prosody and spectrum modification

Our evaluation tests demonstrate a specific pattern in the outcomes that can be related with every parameter independently. For instance, beginning with just spectrally altered sentences, when we changed the pitch the recognition rates enhanced while. A similar perception is not substantial when just pitch alterations were applied to segments which are not spectrally modified.Inclusion of the changes in extra prosodic element, (for example, Jitter, Shimmer), enhanced the outcomes facilitate .The outcomes demonstrate the pattern, as far as creating effective transformation

6.6 Conclusion

This section represents co-relative spectral mapping method for prosody conversion . While generating prosodic speech we are considering prosodic features ,however without considering spectral interrelation it is hard for us to get good prosodic transformation. To solve this problem this phase of work considers the correlation between spectral and prosodic parameters. It involves computation of prosodic features as fundamental frequency, spectral energy, HNR, jitter and shimmer. Before using these features for direct transformation of prosody we update it so that it will compensate the spectral difference between source and target speech signals. An experiment shows that this method has a better ability in tracking prosodic variations in the speech. The converted speech is smoother and reflects desired prosody in a transformed speech. This approach is developed to define a modeling of speech transformation in spectral domain, rather to spatial domain. Results in higher transformation accuracy in comparison to frequency mapping approach. Our proposed method is evaluated for Marathi language. Therefore, it needs to be evaluated for different languages. In future work, for generating better quality of emotional speech, spectral correlations can be also explored at word and syllable level.
These outcomes bolster our theory that consolidating phoneme level and supra-segmental level changes can be a valuable framework to demonstrate the emotional content of synthesized speech Besides, for a prosodic synthesizer to be completely effective extra etymological layers - from phonemic and lexical setting to syntactic and discourse structures-ought to be deliberately represented. A lot of those points of interest are still obscure what's more, are a subject of progressing work.

Chapter 7

7. General Conclusions & Further Directions

7.1 General Conclusions

In this thesis, the concept of prosody in speech segment is handled and it is discussed that how Prosody could be correlated with measurable speech parameters.speech has two fold meaning.a context a sa text and meaning as a prosody.this is very evident that when a speaker changes the way of speaking the same sentence the perceived meaning is completely different.we conclude here that prosody or expressiveness is the intelligence in the speech.Making human machine dialog more intelligent,computers or other devices must learn to deal with this intelligence.So to model the intelligence in speech segment,we investigate the prosody parameters here in this work. The investigation is done over extensive literature in prosody. We found quite a lot of disparity relates in its transformation and representation framework. Realization of prosody can be phonetically done by multiple correlates. These correlates vary in for different languages. The present study involves the investigation of various correlates significant to the prosody of interest.In the section of prosodic event detection we have examined the significant changes in the prosodic features which clearly indicates the onset detection of the prosody.these statistical variations in the feature values will be very useful for machine learning processes where a machine will be taught how to detect prosody or emotion in a said utterance.As the text reflects merely a intonation and stress patterns in the utterance.this leads us to do more investigation on

the prosody event detection from an utterance than a linguistic clues in text.

The efforts put in emotion classification experiments confirm our perception of prosody . The system was required to build that would have been used for the objective evaluation is showing a good connect with the ones using the human perception tests at a much lower time and efforts needs to put by humans for the evaluation. Automatic prosody identification can be used effectively to determine the quality of emotional speech synthesis systems. The process of finding features for classification is nothing but building a feature space. The presented work does not consider processing of speech signal for feature extraction in frequency domain as in frequency domain it tends to lose the time related clues. Prosody highly depends on the timing information of the signals. An approach to compute features by using Empirical mode decomposition of speech signal is presented. This is effective transformation domain in the field of non stationary signals without losing time information. The effect of distortions observed from the capturing units termed as ”system noise” is been successfully not taken into account . This elimination in the distortion content is done in transformed domain(EMD), using spectral density thresholding .this method eliminate the segments of speech considered for the feature calculation which does not show specified spectral energy.thus iluminiting the silence bands ,Noise and signal part with poor spectral energy.Speech coding for accurate prosodic feature extraction is developed by this method and represented by Spectral Resolution (SR) , Algorithm. The spectral resolution coding derives the prosody features more accurately in comparison to conventional approach of direct frequency transform method. Audio samples are processed by SR algorithm to filter out the required frequency bands

for feature calculation . The basic emotions those have to be identified are: angry, happy, neutral, interrogative and sad. The SVM approach used for classification is a high dimensional vector supervised learning method. SVM is deployed with the built-in kernels (kernel parameters in the parentheses): linear, polynomial (polyorder=3),MLP (default scale [1 -1]), and RBF (sigma=200).Here, we have used 3 fold cross validation in which result obtained.

The last phase of work represents co-relative spectral mapping method for prosody conversion. While generating prosodic speech we are considering prosodic features ,however without considering spectral interrelation it is hard for us to get good prosodic transformation. To solve this problem this paper considers the correlation between spectral and prosodic parameters. It involves computation of prosodic features as fundamental frequency, spectral energy, HNR, jitter and shimmer. Before using these features for direct transformation of prosody we update it so that it will compensate the spectral difference between source and target speech signals. An experiment shows that this method has a better ability in tracking prosodic variations in the speech. The converted speech is smoother and reflects desired prosody in a transformed speech. This approach is developed to define a modeling of speech transformation in spectral domain, rather to spatial domain. Results in higher transformation accuracy in comparison to frequency mapping approach. Our proposed method is evaluated for Marathi language. Therefore, it needs to be evaluated for different languages. In future work, for generating better quality of emotional speech, spectral correlations can be also explored at word and syllable level.

7.2 Summery of the thesis

Unit selection Text to speech syntesis for marathi language enables us to get better quality of the synthetic speech outputs compare to rule based approach. However so far, no practical general purpose TTS system is developed which deliver the consistent and human like natural synthetic output. In order to improve the performance of speech interactive systems between human and machines we addressed Prosody (emotion) transplantation and detection problems in speech synthesis in this thesis. First major problem addressed is to embed required prosody in plain speech segment. In order to realize this goal we have to address the two further issues in this thesis .One is to detect the prosodic event in speech in order to gain more knowledge about the statistical variation of the speech features in prosodic as well as nonprosodic events. Second is to classify the prosody based on the features which well describe the emotion in speech. We first describe the structure of the TTS in chapter 3 implemented for Marathi language. We used unit selection synthesis approach. This was implemented with FESTVOX Frame work. In chapter 4 we proposed a system through which prosody event detection can be captured in order to get better knowledge about statistical variations of the speech paremeters in the speech utterance. Neutral and emotional 400 utterances are studied for this purpose and results obtained are validated via subjective listening tests. We concluded with the set of features including the one we described as a ”slope” being significant in marathi style of speaking as ”Marathi” being highly inflectional, ergative language. It contains inflectional and derivational morphology. These features observed for describing the prosody onsets in the speech segment. Native

listeners also gave the verdict in the direction which conforms the algorithmic approach for prosody event detection. In chapter 5 Classification of prosody in speech segment was done by referring the results from chapter 4 .as well as we propose novel algorithm. An approach to compute features by using Empirical mode decomposition of speech signal is presented. This is effective transformation domain in the field of non stationary signals without losing time information. The effect of distortions observed from the capturing units termed as "system noise" are been eliminated. This elimination in the distortion content is done in transformed domain (EMD), using spectral density thresholding. The impact of system noise over the extracted feature and its processing efficiency is been evaluated. The spectral resolution coding derives the prosody features more accurately in comparison to conventional approach of direct frequency transform method. In chapter 6 proposed prosody transplantation is carried out over the emotionally plain speech segment .experimental results revealed that the desired prosody is transplanted .we suggest here the novel algorithm of prosodic parameter mapping from one type to other. This section represents co-relative spectral mapping method for prosody conversion. We observed that while generating prosodic speech if spectral interrelation is not considered it is hard to get good prosodic transformation. To address this problem we consider the correlation between spectral and prosodic parameters. Before using these features for direct transformation of prosody we update it so that it will compensate the spectral difference between source and target speech signals. The experimental results lead us to conclude that proposed method has a better ability in tracking prosodic variations in the speech. The converted speech is smoother and reflects desired prosody in a transformed speech.

This approach is developed to define a modeling of speech transformation in spectral domain, rather to spatial domain. Results in higher transformation accuracy in comparison to frequency mapping approach. These outcomes bolster our theory that consolidating phoneme level and suprasegmental level changes can be a valuable framework to demonstrate the emotional content of synthesized speech Besides, for a prosodic synthesizer to be completely effective extra etymological layers - from phonemic and lexical setting to syntactic and discourse structures-ought to be deliberately represented.

7.3 Further Directions

in recent decades vast study on the analysis as well as the modeling of speech prosody has been done by researchers, the modeling of speech prosody continue to be ”on going task ”, owing to the variability and intricacy in it. The current work revels and faces some of concerns that stay behind to be answered in the analysis and statistical modeling of speech prosody.

Subjective assessment of Speech Prosody

To begin with, the proper evaluation procedure needs to be formulated to assess the degree of naturalness or the score of naturalness on pronunciation of synthetic speech prosody. In lieu of this, the evaluation methods be supposed to be in use to weigh up exactness, diversity, and dynamism that all have a say in the perceiving a natural speech prosody. Furthermore, it would be desirable to assess the sufficiency/resemblance inside a prosodic speech segment with a specific subject, especially with reference to the style of speaking of a subject

portrayal of Prosody in Speech

latest studies are pointing towards the hierarchical structure in prosody of speech ,and these hierarchical structures (Weighted Tree Automata) are successfully modeled the prosodic structure and grammar efficiently. In contrast to conventional sequential models, the use of satisfactory statistical technique in context-specific modeling will surely improve the system performance in an explicit version of the hierarchical organization in representation of speech prosody. in addition, the portrayal of speech prosody is always under dispute both from the notional as well as the applicative point of view. Contributors in the enhancement of speech prosody modeling depends on the various factors such as correct definition of prosody dimensions ,correct stylization of prosodic contours and the significance of temporal field used for the depiction of speech prosody variations.

Curve/pattern Modeling

Modeling of accurate patterns about the variations in speech prosody of various temporal areas is at present an accepted developmental strategy in speech synthesis techniques. Being specific, the proposed curve modeling of various speech prosody variations stands on stylization has been evidenced to always model the variations related with particular temporal fields. The precise formulation of the association that is present between syllable contours and long-term curve would ease the problem of long term curve modeling. Then, the line or path modeling will be taken to any random number of temporal domains exclusive of any remarkable alter in intricacy during modeling and synthesis. Typically the projected trajectory modeling is deployed for the comparison of stylization tech-

nique and temporal domains.

Linguistic Context

The luxury of the linguistic portrayal of a text is a major concern in speech prosody. The refinement of the syntactic description and the integration of the higher linguistic description (e.g., semantic and discursive) would provide highly valuable information that could be used to refine the context-dependent modelling of speech prosody, and to improve the variety of the synthesized speech prosody. However, the derivation of a single context-dependent model that accounts simultaneously for the large range of linguistic levels and linguistic information is absolutely unrealistic. A reformulation of context-dependent modelling will be required in the case of very large vocabulary contexts. An appropriate formulation would probably consist of the derivation of several context-dependent models each associated with a specific linguistic dimension, and then to combine context-dependent models adequately during the synthesis of the speech prosody parameters.

Modelling Variability and Alternatives

The explicit modelling of speech prosody alternatives that correspond to the various strategies of a speaker would de facto improve the naturalness and variety of the speech prosody in speech synthesis [Bulyko and Ostendorf, 2001]. The statistical modelling and synthesis needs to be reformulated so as to provide a various alternatives instead of a single prosodic realization. Additionally, the reformulation would probably need to account simultaneously for short and long term variations. The statistical modelling of prosodic variability may be simply achieved with multi-modal

distributions that may be combined during synthesis with more relaxed inference methods such as the General Viterbi Algorithm (GVA).

Unifying the Modelling of Speech Parameters

In most of the current speech synthesis systems, the inference of the speech parameters is achieved iteratively in a top-down process from the symbolic to the acoustic characteristics. For each of the levels, the optimal sequence of parameters is determined with respect to the considered level and the corresponding model. Thus, each of the levels is restricted to the parameters that are inherited from the higher-level, and does not benefit from their variability and the potential alternatives that may correspond to a more natural synthesized speech. A single method that could simultaneously model the symbolic and the acoustic characteristics and the potential alternatives would improve the quality and variety of the synthesized speech [Bulyko and Ostendorf, 2001]. This may additionally be used to vary the speech prosody of a speaker accurately in speech synthesis.

Modelling Speaking Style Speaking style

modelling can be extended to any arbitrary speaking styles associated with emotional states, situations, and sociological and geographical origins. However, a reformulation would 13.2. FURTHER DIRECTIONS 239 be required to manage a large range of para-linguistic and extra-linguistic contexts that are commonly observed in spontaneous speech. In particular, speech disfluencies (e.g., hesitations, reformulations) and para-linguistic non-verbal speech phenomena (e.g., laughter, sighs, inspiration, expiration) require specific processing that are not available in current speech synthesis systems. During the analysis, para-linguistic information has to be automatically labelled. During the training and synthe-

sis, the location and the acoustic characteristics of the para-linguistic phenomena need to be modelled depending on the context. Additionally, the segmentation and the description of speech utterances into different types of narrative and/or discursive sequences (for instance, sports commentary significantly modifies the speech prosody characteristics depending on the more or less degree of implication of the speaker and the intensity of the action being commented on) would qualitatively improve the variety of the synthesized speaking style. Finally, more sophisticated methods have to be employed to adapt finely the characteristics of a speaker to those of a speaking style.

Chapter 8

8. Applications

Synthetic speech generation when done by machine has various applications vary from reading engines that deliver unhindered text input in to simple inquiry scheme that play back prompt-style speech segment as messages. Text to speech synthesis engines can be of huge useful to people with visual disabilities. There is a noteworthy commercial market for speech synthesizers, which run on standard personal computers. For visually impaired people speaking apparatus can support the communication requirements of people with voice handicaps too. There is even a substantial business value for written literature converted on memory chip to render a audio equivalent , Acoustically rendered messages delivered as sound file will be further more useful in numerous situations, where the person is giving attention some other task, or might be drained . There seems, by all accounts, to be an expanding enthusiasm for joining speech synthesizers in estimation and control frameworks, viz in plane cockpits or surgery rooms. Vocal checking is additionally preferred by a few clients of text preparing software to auditory "edit" writings that they have composed. A promising business sector for what's to come is the coordination of synthesis instructive software in education industry, particularly in dialect learning programs, remote dialect direction, and simulation of job training system. On the segmental level, speech clarity and precision can be contended to be on an adequately satisfying state for such applications. Prosody, then again, has frequently been guaranteed to be of deficient quality. Nonetheless, prosody generated by model based

approach is certainty be more steady than human teachers are. In the application situations illustrated till so far, speech synthesis technology must be equipped to deliver any context in any area. Precision have to be as good as possible, and a elevated scale of naturalness will increase the effortlessness of listening to bigger passages as synthetic speech. speech synthesis is included in telecommunication services

proved to be pretty popular since early 1990's (Levinson, Olive, and Tschirgi, 1993). These services comprise: an automated telephone directory system, in which upon receiving the name of customer a telephone number is synthetically delivered

so one can choose whether to pick the call or not ("Who's calling "); a telephone service provider company impart service to facilitate the adaptation with speech or persons having hearing disabilities, putting together text-to-speech and speech-to-text conversion. Increasingly refined integrated messaging service let the user to receive and send messages in many format, e.g., fax, email, or voice mail, and to translate from side to side between any of these format. Speech synthesis systems are also gradually more used to provide access to information retrieval and inquiry services over the (mobile, cellular) telephone. Trendy services in its kind are films and information about films and other neighboring events, hotels and menu information, satellite television program information. Speech output is also expected by consumers in Car navigation systems. it is expected to ne taken a legal initiatives to ban information systems helping drivers in a display-based way while driving a car. The text databases for providing these services are too big and even changing few times. The same cling for more interactive purposes such as hotel, train and airline reservation systems. above applications are changing by the advanced re-

search done in speech so these systems are gradually turning into fully interactive speech dialog systems.

8.1 Interactive dialog systems

Effectively in human-machine dialog systems are rated by its Instructiveness. Such system is defined as computer systems with which human interact on a turn-by-turn basis " (Gibbon, Moore, and Winski, 1997, page 564), i.e., the system in which natural language processing has a key role in achieving meaningful communication process. Speech dialog systems incorporate a large number of components and every component handles a intricate task, such as speech recognition, database management, speech synthesis. A central component, which is called as dialog manager facilitates and provides the dialog between the user and the services offered by the system. A classy speech communication system will build on a speech utterance model, i.e., a model of natural connected spoken speech segments, and on a grammar, i.e., a model of the construction of speech dialogs, often implemented in the form of state machines for sequencing turns in the discourse. The sophisticated systems employ a hybrid-initiative dialog approach, in which mutually the user and the dialog manager can take over the initiative and proceed ahead the communication process, which ideally results in a joint give and take dialog (Chu-Carroll, 1996).

8.2 Multimodal communication

Speech is a natural means of communication between a human and a machine, but it is always fascinating and even more intuitive task for researchers to combine speech with other modalities, both on the input and on the output side of the system. such a multimodal dialog system are under development for realizing the goal of adapting the user profile and inclination which will lead to prefer between available output modalities depending on the user and user requirements about the task, and the specific application. Spoken utterance is a vital modality in all such scenarios. Multimodal human machine interfaces need to combine information and technology from a extensive choice of technical disciplines, such as intelligent systems, speech technology, information representation and multimedia, needs to be implemented for speech, images, gestures and even hand-writing. Ahead of mere appreciation, pattern understanding and semiotics should be applied to, again, speech, images, and mimic; implementation of such systems would require a research and knowledge in spoken dialog and communication. In the area of information processing, identification of the user's objective or intention is an significant aim, and it needs to be converted to presentation language and graphics generation and synchronization. User models and system evaluation depends on outcome from cognitive science. Finally, hardware, software, and application trade will engage a significant role in the realization, predominantly for system design and incorporation

Bibliography

References

[1] A. Iida, N. Campbell, F. Higuchi and M. Yasumura, "A corpus-based speech synthesis system with emotion", Speech Communication., Vol 40, n,1. pp189-212 ,Elsevier,2003

[2] A. Jaywant, M. D. Pell, "Categorical processing of negative emotions from speech prosody", speech communication, 54(1), pp1-10,Elsevier, 2012.

[3] Alessandro.C,Doval.B., 2003. "Voice quality modification for emotional speech synthesis." In Proc. Eurospeech. Geneva, Switzerland, 2003.

[4] Anand C, Devendran B., 2015. "Speech Emotion Recognition using CART algorithm." International Research Journal of Engineering and Technology, 2(03),pp 870-874.

[5] Barra-Chicote, R., Yamagishi, J., King, S., Montero, J. M., Macias-Guarasa, J. (2010). Analysis of statistical parametric and unit selection speech synthesis systems applied to emotional speech.Speech Communication, 52(5), pp 394-404.

[6] Bjorn Granstrom, D.H., 2005. "Audio visual representation of prosody in expressive speech communication", Speech Communication. Speech Communication Elsevier, 46.

[7] Black, A.W., December 2,2014.www.festvox.org.http://www.festvox.org /docs/manual-2.4.0

[8] Burkhardt.F, W.F.Sendilmeier, "Verification of acoustical correlates of emotional speech using formant synthesis" Proc. ISCA workshop on speech emotion , Northon Irland, pp. 151-156,2000.

[9] C. Gobl,Chasaide.A, "The role of voice quality in communicating emotion, mood, and attitude, Speech Communication" Vol 40, n,1. pp189-212 ,Elsevier,2003.

[10] Cabral, L. C. Oliveira., 2006a. "Emo voice: a system to generate emotions in speech." in proc Interspeech.

[11] Cabral.J ,L. C. Oliveira, "Pitch-synchronous time-scaling for prosodic and voice quality transformations," Proc. of the 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, 2005, pp. 1137-1140.

[12] Cahn, J.E., 1989. "Generation of affect in synthesized speech". American voice I/O society.8,pp1-19

[13] Campbell, N., 2006. "Conversational speech synthesis and the need for some laughter". IEEE Transactions on Audio, Speech, and Language, 17(4), p.1117-1179.

[14] Campell, N., Hamza, W., Hog, H., Tao, J. (2006). Editorial special section on expressive speech synthesis. IEEE Transactions on Audio, Speech, and Language Processing, 14, 1097-1098.

[15] Chung Wu, C. H. Wu, H. Lee, M. C. Lin, "Hierarchical Prosody Conversion Using Regression-Based Clustering for Emotional Speech Synthesis," in IEEE Transactions on Audio, Speech, and Language Processing, 2010. 18(6)pp. 1394-1405.

[16] CLEMENTS, G. N. The geometry of phonological features. Phonology Yearbook 2 (1985),225-252.

[17] D. Erro, D. Erro,E. Navas and I. Hernaez,I. Saratxaga, 2010." Emotion conversion based on prosodic unit selection." IEEE Transactions on Audio, Speech, and Language Processing, 18, pp.974-83.

[18] D.Erickson,"Expressive speech: production perception and application to speech synthesis," Acoustical Science and Technology Vol 26, n. 4, pp 317-325, 2005

[19] Deshpande.P.S, Chitode.J.s, 2016. "Novel Feature Coding In Prosody Generation For Emotional Speech Synthesis." National journal IETE.

[20] Dharaskar.M,B.Chandak Dr. Rajiv, 2010. "Emotion Extractor: A methodology to implement prosodyfeatures in Speech Synthesis". In electronic computer technology (ICECT)., 2010.

[21] Dominik R.Bach, R.H.R.J.D., 2013. "Unimpaired discrimination of fearful prosody after amygdala lesion." Neuropsychologia,Elsevier, 51.

[22] Drioli.C,Graziano.T,Piero C, 2003. "Emotions and voice quality: experiments with sinusoidal modeling." In VOQUAL'03, Geneva,2003,pp 127-132.

[23] Dudley, H..R.R..a.W.S., 1939." A synthetic speaker. Journal of the Franklin Institute," 227(6), pp.739-764.

[24] Jainath Yadav,K. Sreenivasa Rao "Generation of emotional speech by prosody imposition on Sentence, Word and Syllable level fragments of neutral speech"' Cognitive computing and information processing (CCIP), international conference, IEEE,Noida. 2015,PP1-5.

[25] faber, Aug 1846. The Euphonia. Vol. 9. london: Illustrated London News.

[26] Fernandez.R,2007." Automatic exploration of corpus specific properties for expressive text-to-speech: a case study." In In Proc. ISCA workshop on speech synthesis., 2007.

[27] ,Orsucci, F., Petrosino, R., Paoloni, G., 2013." Prosody and synchronization in cognitive neuroscience". EPJ Nonlinear Biomedical Physics, Springer.

[28] Genzel, S.a.K., 2010. "The prosodic expression of contrast in Hindi. In 5th International Conference of Speech Prosody." Chicago, USA, 2010.

[29] Govind, D..a.S.R.M.P., 2013. "Expressive speech synthesis: a review". International Journal of Speech TechnologySpringer Science+Business Media New York 2012, pp.237-60.

[30] Group, A., 2010. "Acapela Speech Synthesis System". [Online] Acapela Group.

[31] Gurunath Reddy M, H.D.M.K.S.a.M.K.E., 2015. Telugu Emotional Story Speech Synthesis using SABLE Markup Language. In SPACES Dept of ECE, K L University, 2015.

[32] Harnsberger, J.D.., 1996."Towards an intonational phonology of Hindi". In Proc of Fifth Conference on Laboratory Phonology. Northwestern University, 1996.

[33] Hass, P.J., 2003. An Acoustics Primer,Chapter6.www.indiana.edu/ emusic/acoustics/amplitude.htm.

[34] Hofer, G..Richmond.K...C.R. ”Informed blending of databases for emotional speech synthesis.” Interspeech,2005.

[35] I.Hadhami, B Aicha,” Speech Signal Enhancement Using Empirical Mode Decomposition and Adaptive Method Based on the Signal for Noise Ratio Objective Evaluation,” International Review on Computers and Software (IRECOS),Vol 9 ,n.8,pp.1461-1467,2014.

[36] J. Crumpton, C.L.Bethel,2015.”Validation of Vocal Prosody Modifications to Communicate Emotion in Robot Speech.”In collaboration technologies and systems (CTS) international conference IEEE., 2015.

[37] J.Tao, ”Nonlinear emotional prosody generation and annotation.” In ISCSLP . BERLIN, 2006.

[38] Jasmine Kaur, P.S., 2015. ”Review on Expressive Speech Synthesis.” International Journal of Electrical, Electronics and Computer Systems, 3(10).

[39] Jason, B., 2016. weka machine learning. [Online] Available at: http://machinelearningmastery.com/perform-feature-selection-machine-learning-data-weka/.

[40] Jia Huanga, R.C.K.C.X.L.Z.M.Z.L.Q.-y.G., 2008. ”An exploratory study of the influence of conversation prosody on emotion and intention identification in schizophrenia.” Elsevier Brain Research, 58.pp 58-63

[41] Jianhua Tao, Y.K.a.A.L.P.,”Prosody Conversion From Neutral Speech to Emotional Speech.” IEEE Transactions on audio, Speech, and Language Processing, 2006,14(4).

[42] Juan Pablo Arias, C.B.N.B.Y., 2014. "Shape-based modeling of the fundamental frequency contour for emotion detection in speech." Computer Speech and Language , Elsevier, pp.278-94.

[43] J. R. Quinlan. 1986." Induction of Decision Trees." Mach. Learn. 1, 1 (March 1986), 81-106.

[44] J.Vroomen., R.Collier, S.J.L.Mozziconacci, "Duration and intonation in emotional speech", Proc. Eurospeech, pp. 577-580,1993.

[45] J.M.Montero, J. Gutierrez-Arriola, Colas, J., .Enriquez, E., Pardo,J. M. "Analysis and modeling of emotional speech in Spanish",1999, Proc. ICPhS ,pp. 671-674.

[46] J. F. Pitrelli, R. Bakis, E. M. Eide, R. Fernandez, W. Hamza, and M. A.Picheny, "The ibm expressive text to speech synthesis system for American english," IEEE Trans. Audio, Speech and Language Processing ,Vol. 14,n. 4, pp. 1099-1109, July 2006.

[47] K.C. Rajeswari, G.K.P., 2014. "Developing Intonation Pattern for Tamil Text To Speech Synthesis System." international Journal of Innovative Research in Computer and Communication Engineering, 2(1).

[48] Kawanami, I., n.d. "GMM-based Voice Conversion Applied to Emotional Speech Synthesis." IEEE Trans. Speech and Audio Proc, 7(6), pp.697-708.

[49] Klara Vicsi, 2012. "Thinking about present and future of the complex speech recognition.",3rd IEEE International Conference on Cognitive Info communications, Proceedings.Cog InfoCom.2012.371-376.

[50] Klatt., D.H., 1987. "Review of text-to-speech conversion for English," Journal of the Acoustical Society of America, 82(3), pp.737-793.

[51] Klatt, D., 1980."Software for cascade or formant synthesizer."Journal of acoustical society of America, 67, pp.971-75.

[52] Laba kr. Thakuria, 2014. "Integrating Rule andTemplate- based Approaches to Prosody Generationfor Emotional BODO Speech Synthesis". In IEEE,International Conference on Communication Systemsand Network Technologies., 2014.

[53] Latorre, J.a.A.M., 2008. " Multilevel parametric-base F0 model for speech synthesis." In Interspeech. Brisbane, Australia., 2008.

[54] Lupu, S.E.E., 2011. "Improving Speech Emotion Recognition Using Frequency and Time Domain Acoustic Features". In Proceedings of SPAMEC, Cluj-Napoca, EURASIP. Romania, 2011.

[55] Maheswari, R.K.C.a.U., 2012. "Prosody Modeling Techniques for Text-to-Speech Synthesis Systems A Survey." International Journal of Computer Applications , 39(16), pp.8-11.

[56] Murray, Iain L. Arnott, John.1995. "Implementation and testing of a system for producing emotion-by-rule in synthetic speech." Speech-Communication , 16, pp.359-68.

[57] Manjare, C.A.S.D..S., 2014. "Speech Modification for Prosody Conversion in Expressive Marathi Text-to-Speech Synthesis". In International Conference on Signal Processing and Integrated Networks, IEEE, 2014.

[58] Borchert.M,Diisterhoft.A, 2005. ”Emotions in Speech - Experiments with Prosody and Quality Features in Speech for Use in Categorical and Dimensional Emotion Recognition Environments”. In Proceeding of NLP-KE’05.Proceedings of IEEE international conference. 2005.

[59] Miyanaga, K..M.T..K.T., (2004). ”A style control techniques for hmm-based speech synthesis.” Transactions on Information and Systems, E90-D(9),2007,Pp.1406-1413.

[60] Montero, Gutiérrez.A.,Colás,Macías.G,uarasa, J.,Pardo,J, 1999. ”Development of an Emotional Speech Synthesizer in Spanish”. In Proceedings of Eurospeech. Budapest, Hungary, 1999.

[61] Murray I.R, A.J.I., 1993.” Towards simulation of emotion in synthetic speech:A review of literature on human vocal emotion”. journal of Acoustical Society Of America, pp.1097-108.

[62] Matsui, H..K.H., 2003. ”Investigation of emotionally morphed speech perception and its structure using a high quality speech manipulation system.” In Proc.of Eurospeech. Geneva, Switzerland, 2003.

[63] N. Campbell,”Developments in corpus-based speech synthesis:Approaching natural conversational speech,”IEICE Transactions on information and analysis, Vol E88, n.3,pp497-500, 2005

[64] Odetunji A. odejobi, S.H.S.W.a.A.J.B., 2008. ”A modular holistic approach to prosody modeling for Standard Yoruba speech synthesis”. Computer Speech and Language ELSEVIER, pp.39-68.

[65] Oliveira, J.P.C.a.L, 2006. ”Emovoice: a system to generate emotions in speech”. In INTERSPEECH, 200. Pittsburgh, Pennsylvania, 2006.

[66] Oliveira, J.P.C.a.L.C., 2006b. ”Pitch-synchronous time scaling for prosodic and voice quality transformations.” Interspeech, pp.1137-40.

[67] Oliveira, J.P.C.a.L.”2006. Emovoice: a system to generate emotions in speech.” In interspeech, 2000. Pittsburgh, Pennsylvania, 2006.

[68] P. Gangamohan,Mittal,V,B.Yegnanarayana, 2012. ”A Flexible Analysis Synthesis Tool (FAST) for studying the characteristic features of emotion in speech.” In 9th Annual IEEE Consumer Communications and Networking Conference., 2012.

[69] Pablo Daniel Aguero, J.A.a.A., 2006. translation ”prosody generation for speech”. In ICASSP, international conference IEEE, 2006.

[70] Pell, M.D., 2006.” Cerebral mechanisms for understanding emotional prosody in speech.” Brain and Language Elsevier.

[71] Pierrehumbert, J, (1980). ”The Phonology and Phonetics of English Intonation.” Ph.D thesis. MIT.

[72] Pierrehumbert, J., 1981. ”Synthesizing Intonation.” Journal of the Acoustical Society of America, (70), pp.985-995.

[73] Pribil, J., 2009.”Statistical Analysis of Spectral Properties and Prosodic Parameters of Emotional Speech”, Measurement Science Review. 9(4).

[74] Pibilová, A., 2009. ”Spectrum Modification for Emotional Speech Synthesis”. LNAI 5398, Springer.

[75] P.tzinger, H.R., 2008.” Amplitude and Amplitude Variation of Emotional Speech.” In Inter Speech., 2008.

[76] Pinheiro, A.P., 2015. "The music of language: An ERP investigation of the effects of musical training on emotional prosody processing." Brain and Language, Elsevier.

[77] Pitrelli, J.R., 2006. "Expressive text to speech synthesis system for English". IEEE trans. Audio Speech Language processing, 14(4), pp.1099-108.

[78] Prashant Aher, A.C., 2014. "Comparative Analysis of Speech Features for Speech Emotion Recognition". International Journal of Advanced Research in Computer and Communication Engineering, 2003.

[79] Rahul. B. Lanjewar, D.S.C., 2013. "Speech Emotion Recognition: A Review". International Journal of Innovative Technology and Exploring Engineering, 2(4).

[80] Rao, K.S., 2010. "Voice conversion by mapping the speaker-specific features using pitch synchronous approach". Computer Speech Language, 24, pp.474-94.

[81] Rao, S.G.K.a.S., 2012. "Emotion recognition from speech: a review." International Journal of Speech Technology, 15(2), pp.99-117.

[82] K. Rao, "Role of neural network models for developing speech systems," Sadhana, Vol. 36, Part 5, Indian Academy of Sciences, pp.783-836, 2011

[83] R.Verma, 2015." Conversion of Neutral Speech to Storytelling Style Speech." In Advances in pattern recognition (ICAPR), eighth international conference IEEE., 2015.

[84] Ravi, G.A.a.D.D.J., 2014. "Transformation of Emotions using Pitch as a Parameter for Kannada Speech."In International Conference on Recent Trends in Signal Processing Image Processing and VLSI, Association of Computer Electronics and Electrical Engineering., 2014.

[85] Rigoulot, S., 2014. "Emotion in the voice influences the way we scan emotional faces". speech communication, ,Elsevier , 65.

[86] Rao, K.S., 2011. "Role of neural network models for developing speech systems." Indian Academy of Sciences, Sadhana, 36(5).

[87] Ryo Aihara, R.T.T.T.Y.A., 2012." GMM-Based Emotional Voice Conversion Using Spectrum and Prosody Features". American Journal of Signal Processing, 2(5).

[88] S. Ananthakrishnan., S.S.N., 2008. "Automatic Prosodic Event Detection Using Acoustic, Lexical, and Syntactic Evidence." IEEE Trans Audio Speech Lang Processing., pp.216-28.

[89] S. S. Agrawa1, N.P.a.A.J., 2010. "Transformation of emotion based on acoustic features of intonation patterns for Hindi speech." African Journal of Mathematics and Computer Science Research, 3(10).

[90] Samantaray, A.K., 2015." A novel approach of Speech Emotion Recognition with prosody, quality and derived features using SVM classifier for a class of North-Eastern Languages." In 2nd International Conference on Recent Trends in Information Systems, IEEE, 2015.

[91] Santen 2003. "prosodic modeling in text to speech synthesis." Proc EUROSPEECH,

[92] Schroder, M., 2003. "The German Text-to-Speech Synthesis System MARY: A Tool for Research, Development and Teaching." International Journal of Speech Technology, 6.

[93] Schroder, M., 2009." Expressive speech synthesis: past, present and possible futures." Affective Information Processing, 2, pp.111-26.

[94] Schroder, O.T.a.M., 2010. "Evaluation of Expressive Speech Synthesis With Voice Conversion and Copy Resynthesis Techniques." IEEE Transactions on Audio, Speech, and Language Processing, 18(5).

[95] Sendilmeier, B.a., 2000. "Verification of acoustical correlates of emotional speech using formant synthesis." In ICSA workshop on speech and emotion. Newcastle, 2000.

[96] Shashidhar G. Koolagudi, 2011. IITKGP-SEHSC: Hindi speech corpus formation analysis. In Devices and communications (ICDeCom), international conference, IEEE., 2011.

[97] Shulan Xia, 2014. "A speech emotion enhancement method for hearing aid". Computer Modeling & New Technology, 18(11).

[98] Sudhkar, R.S., 2015. "Analysis of Speech Features for Emotion Detection: A review". In ,International Conference on Computing Communication Control and Automation, IEE., 2015.

[99] Swann Pichon, a.C.A.K., 2013. "Affective and Sensory motor Components of Emotional Prosody Generation". The Journal of Neuroscience, 33.

[100] S. Takeda, G. Ohyama, A. Tochitani and Y. Nishizawa, "Analysis of

prosodic features of ”anger” expressions in Japanese speech,” Journal of Acoust. Soc. Japan. Vol.58,n.9 ,pp. 561-568 ,2002.

[101] Teppereman, J.a.N.S., 2008. ”Tree grammars as models of prosodic structure”. In Interspeech. Brisbane, Australia., 2008.

[102] Theune, M,Meijs.K,Heylen.D,Ordelman.R,2006.”Expressive speech for story telling applications.” IEEE Transactions Audio, Speech, and Language Processing, 14(4), pp.1099-108.

[103] Thomas.E,Kreifelts.B,Wiethoff.S,Jonathan,W et.al., 2008. ”Differential Influences of Emotion, Task, and Novelty on Brain Regions Underlying the Processing of Speech Melody”. Journal of Cognitive Neuroscience, 21(7).pp 1255-68.

[104] Turk, O.Schröder.M.Baris.B..Arshalan.L.,2005.”Voice quality interpolation for emotional text-to-speech synthesis”. In: Proc. Interspeech. Lisbon, Portugal, 2005.

[105] Tsuzuki, R., Zen, H., Tokuda, K., Kitamura, T., Bulut, M., & Narayanan, S.S. (2004). ”Constructing emotional speech synthesizers with limited speech database.” Proceedings of InterSpeech pp 1185-1188.Korea.

[106] Vroomen, J..C.R.a.M.S., 1993.”Duration and intonation in emotional speech,” In Proceedings of Eurospeech. Berlin, Germany, 1993. pp.577-80

[107] Wightman, C.W., 1992.” Segmental durations in the vicinity of prosodic phrase boundaries. ”The Journal of the Acoustical Society of America , pp.1707-17.

[108] Wu, C.H..H., 2006. ”Voice Conversion Using Duration-Embedded Bi-HMMs for Expressive Speech Synthesis.” IEEE Trans. Audio, Speech and Language Proc, 14(4), pp.1109-16.

[109] Wightman,M.Ostendorf., 1994. ”Automatic labeling of prosodic patterns.”. IEEE Trans. on Speech,2(4),469-481

[110] Xiaoyong Lu, H.Y.A.Z., 2014. ”Applying PAD Three Dimensional Emotion Model to Convert Prosody of Emotional Speech.” In Orange Technologies (ICOT), international conference IEEE,2014.

[111] Yadav, J., 2015. ”Generation of emotional speech by prosody imposition on Sentence, Word and Syllable level fragments of neutral speech”. In Cognitive computing and information processing (CCIP), international conference IEEE, 2015.

[112] Yamagishi, J..O.K..M.T..K.T, ”Modeling of various speaking styles and emotions for HMM-Based Speech Synthesis”. Proc. EUROSPEECH, 2003, pp.2461-64.

[113] Y. Hayashi, ”Recognition of vocal expression of mental attitudes in Japanese: Using the interjection ”eh”,” Proc. Int.Congr. Phonetic Sciences, San Francisco, 1999, pp. 2355-2358

[114] Zovato, E..P.A..Q.S..S.S., 2004. ”Towards emotional speech synthesis :A rule based approach.”In Proc. 5th ISCA Speech Synthesis Workshop. Pittsburgh, 2004.

Publications

[1] Deshpade P.S.,Chitode J.S., "Spectral correlative mapping approach for Transformation of Expressivity in Marathi Speech"Paper is accepted for the volume 8 issue 1 February 2018,by International Review on communication,antenna and propagation.(IRECAP),Press Worthy Prize Publication.

[2] Deshpade P.S.,Chitode J.S., "Emotion Classification Technique in Speech Signal for Marathi" IJARSE,6(11),2017,338-353.

[3] Deshpade P.S.,Chitode J.S., "Transformation coding for emotion speech translation: a review"IJEEER,Vol 6,n.1,2016.

[4] Deshpade P.S.,Chitode J.S., "Novel feature coding using prosody generation for emotional speech synthesis," IETE National Journal of innovation and research,Vol.3,n.1,pp.38-44,2015.

[5] R. S. Deo and P. S. Deshpande, "Pitch contour modeling and modification for expressive Marathi speech synthesis," 2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI), New Delhi, 2014, pp. 2455-2458.