Emotion in Speech through Assamese Language Using GMM Classifier

Dr. Laba Thakuria[1]
Deputy Controller of Examination
AdtU

Mrs Mampi Devi[2]
PhD Scholar, Deptt. of Eng & Technology
AdtU

Abstract : The Emotion recognition is the process of analysis the acoustic difference that occurs on uttering the text similar to the information being conveyed by the speaker under different emotional situations. We experiment on different evaluation approach to speech recognition in multilingual, in which the phone sets, area entirely distinct. The model has parameters not tied to specific states but that are shared across languages. The text-to-speech synthesis systems require an accurate prosody labels to generate natural-sounding speech. This paper presents a method based on Gaussian mixture model (GMM) classifier and Mel-frequency Cepstral Coefficients (MFCC) as features for emotion recognition from Assamese speeches. The experiments are done for the cases of (i) text-independent but speaker-dependent and (ii) text- independent and speaker-independent.

Keywords:Emotion, GMM, Assamese language

1. Introduction

The emotion in speech may be considered as similar kind of stress on all sound events across the speech. An emotional speech describes an Emotional prosody that is characterized as an individual's tone of voice in speech that is conveyed through changes in pitch, loudness, speech rate and pauses which is different from linguistic and semantic information. The prosodic rules of a language evolve with the culture of a community over ages. The emotion expressed and inferred in a speech, depends upon the speaker's community culture and language, gender, age, education, social status, health, physical engagements etc. However emotion recognition from the speech signal is very challenging task for machine, because this requires that the machine should have the sufficient intelligence to recognize human voices and emotion through it.

Emotion recognition from the speaker's speech is very difficult because of the following reasons: In differentiating between various emotions which particular speech features are more useful is not clear. Each emotion includes different portions of the spoken utterance and hence it is very difficult to differentiate among these portions of the utterance. When a speaker is in a 'quiet room' with no task obligations and without any illness, the speech produced by him is 'neutral'. When a speaker feels his environment as different from 'normal', he perceives an emotional arousal in him, and this in turn causes a change of his physiological parameters. The emotional arousal sets the speaker in an emotional state. In such state a speaker normally produces a kind of stressed speech, which is called emotional speech. A particular degree of emotional arousal causes a particular amount of activation level, balance or evaluation level, orientation, etc. Full-blown emotion is generally short lived and intense, where emotional strength has crossed a certain limit, e.g. archetypal emotions such as angry, disgust, fear, happy, sad and surprise.

My present work aims at investigation of the emotion recognition from Assamese speech which will help in proper translation of Assamese speech to the same in any other language.

2. Assamese Language

Assamese, also known as Asamiya is an Eastern Indo-Aryan language spoken mainly in the Indian State of Assam by the Assamese people in general and it serves as a lingua franca in the region. The mixed Aryan culture and the mongoloid culture gave birth to a new culture. So, every community from this region always exhibits their Indigenous culture with diversity. It is the link language for the people living in Assam and adjoining states of Arunachal Pradesh, Meghalaya, and Nagaland etc. This language has come from Sanskrit as its offshoot through different stages of development of Assamese language is as given below.

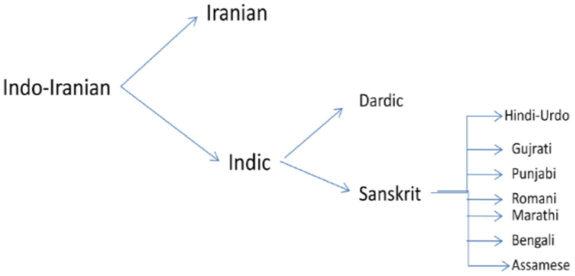
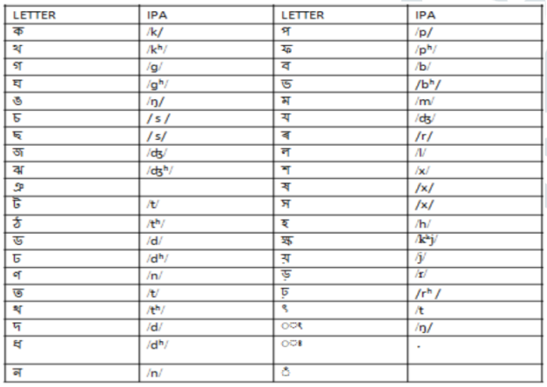


Figure1: Indo-Iranian Family Tree

Assamese language has total eight numbers of oral vowel phonemes, among them three nasalized vowel phonemes. There are total fifteen diphthongs used in Assamese pronunciation. Total number of Assamese consonants phonemes are twenty one.

Table 1: Consonant written scripts used in the language with their IPA symbols.



LETTER	IPA	LETTER	IPA
ক	/k/	প	/p/
খ	/kʰ/	ফ	/pʰ/
গ	/g/	ব	/b/
ঘ	/gʰ/	ভ	/bʰ/
ঙ	/ŋ/	ম	/m/
চ	/ s /	য	/ʤ/
ছ	/ s/	ৰ	/r/
জ	/ʤ/	ল	/l/
ঝ	/ʤʰ/	শ	/x/
ঞ		ষ	/x/
ট	/t/	স	/x/
ঠ	/tʰ/	হ	/h/
ড	/d/	ক্ষ	/kʰj/
ঢ	/dʰ/	য়	/j/
ণ	/n/	ড়	/ɾ/
ত	/t/	ঢ়	/rʰ /
থ	/tʰ/	ৎ	/t
দ	/d/	ং	/ŋ/
ধ	/dʰ/	ঃ	.
ন	/n/	ঁ	

2. Gaussian Mixture Model (Gmm)

A Gaussian Mixture Model (GMM) is a parametric probability density function represented as a weighted sum of Gaussian component densities. GMMs are commonly used as a parametric model of the probability distribution of continuous measurements or features in a biometric system, such as vocal-tract related spectral features in a speaker recognition system. GMM parameters are estimated from training data using the iterative Expectation-Maximization (EM) algorithm.

A Gaussian mixture model is a weighted sum of M component Gaussian densities as given by the equation,

$$\mathrm{P}(\mathrm{x}/\lambda)=wiMi=1\ \ g(x\ \ \mu i,\ \ )\ i \qquad (1)$$

Where x is a D-dimensional continuous-valued data vector (i.e. measurement or features), wi, i = 1, M, are the mixture weights, and $g(x\ \mu i,)i,i= 1,\ldots.,M$ are the component Gaussian densities. Each component density is a D-variate Gaussian function of the form,

$$g\ (,)i =1/2\pi\ D/2\ |\ i\ |1\ 2\ \exp\{-1\ 2\ x-\mu i\ '\ x-\mu i-1\ i\ \} \qquad (2)$$

With mean vector μi and covariance matrix. i the mixture weights satisfy the constraint that $wi\ M\ i=1 = 1$. The complete Gaussian mixture model is parameterized by the mean vectors, covariance matrices and mixture weights from all component densities

3.Prosodic feature extraction:

Statistics related to pitch describes considerable information about the emotional status. For this research pitch is extracted from the speech waveform using RAPT algorithm. Using a frame length of 50ms, the pitch for each frame was calculated and placed in a vector to correspond to that frame. The various statistical features are extracted from the pitch tracked from the samples. We use minimum value, maximum value, range and the moments- mean variance and kurtosis. As a result we get a 7 dimensional feature vector which is appended to the end of the 39 dimensional super vector obtained from the GMM.

Loudness : Loudness is extracted from the samples using MATLAB.

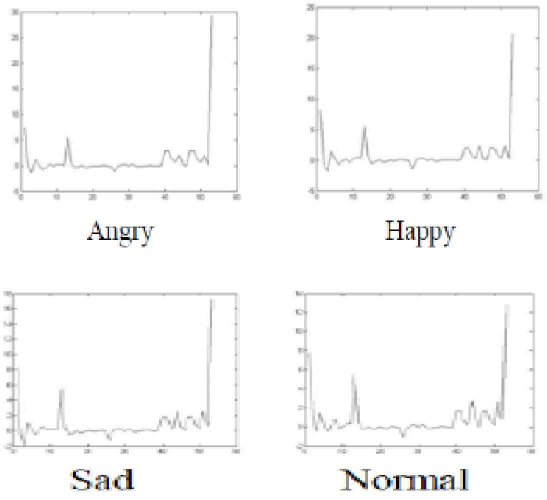


Figure 3: Different Emotion states

The function loudness() returns loudness for each frame length of 50ms and also one single specific loudness value. Now the same minimum value, maximum value, range and the moments- mean, variance and kurtosis statistical features are used to model the loudness vector. Hence we get an 8 dimensional feature vector which is appended to the already obtained 46 dimensional feature vector to obtain the final 54 dimensional feature vector.

There are three kinds of emotional databases with regard to the authenticity of emotion. In general dramatics are asked to speak some given utterances while expressing a certain emotion and the recording is labelled as containing the specified desired emotion. Another kind of databases contains elicited emotions. This kind of emotion is neither real nor simulated. The last types are databases of spontaneous speech containing real emotions. For the speech emotion recognition experiment emotional speech data should be collect from the all types of speaker

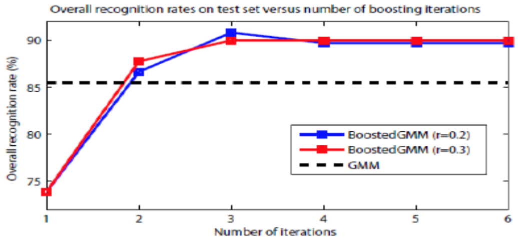


Figure 4: Recognition rate and boosting iterations

(i) Implementation Using Gaussian Mixture Model

The probability density functions of distorted features caused by different emotions are different and hence we can use a set of GMMs to estimate the probability that the observed utterance from a particular emotion.

Maximum Likelihood Estimation: In construction of a Bayesian classifier the class-conditional probability density functions need to be determined. The initial model selection can be done for example by visualizing the training data, but the adjustment of the model parameters requires some measure of goodness, i.e., how well the distribution fits the observed data. Data likelihood is such goodness value .

Assume that there is a set of independent samples X = $\{x1,2,\ldots\ldots xN\}$ drawn from a single distribution described by a probability density function p(x; θ) where θ is the PDF parameter list. The likelihood function

$\mathcal{L}\,(;)= P(xN{:}\theta)\ N\ x{=}1$ (3)

tells the likelihood of the data X given the distribution or, more specifically, given the distribution parameters θ. The goal is to find that maximizes the likelihood:

θ =arg$max\theta$= $\mathcal{L}$ (;) (4)

Usually this function is not maximized directly but the logarithm

$\mathcal{L}$ (;)=ln$\mathcal{L}(X;\theta)$= $lnp(xN{:}\theta)\ N\ n{=}1$ (5)

Called the log-likelihood function which is analytically easier to handle. Because of the monotonicity of the logarithm function the solution to Eq. 8 is the same using $\mathcal{L}\ X;$

(ii) Steps for GMM classification

1] Initialize parameters Expectation step: Compute the posterior probability for i=1, n, k=1 ...K.

$Pi,=\ ak\ (r)\emptyset\ (xi\mu k\ (r),\ k\ (r)\ k{=}1\ (r)\ ak\ (r)\emptyset\ (xi\mu k\ (r),\ k\ (r)$ (6)

2] Maximization step

$ak\ (r{+}1){=}\ i{=}1\ n\ Pi,k\ n$ (7)

$\mu k\ (r{+}1) = i{=}1\ n\ Pi,$ X $i{=}1\ n\ Pi,$ (8)

$\mu k\ (r+1)= k\ (r+1)Pi,k(xi\mu k\ (r+1))(xi\mu k\ (r+1))\mathrm{t}\ i=1$
$n\ Pi,k$ (9)

3] Repeat steps 2) and 3) until convergence.

There are three kinds of emotional databases with regard to the authenticity of emotion. Databases with acted speech include portrayals of emotions by professional or amateur actors. In general actors are asked to speak some given utterances while expressing a certain emotion and the recording is labelled as containing the specified desired emotion. Another kind of databases contains elicited emotions. This kind of emotion is neither real, nor simulated. The last types are databases of spontaneous speech which contain real emotions. Therefore during the collection of the emotional speech database for the speech emotion recognition system care has to be taken that there should be naturalness in the speech recording. For the speech emotion recognition experiment emotional speech data should be collect from the all type of speaker.

4. Data Collection

The performance of the speech emotion recognition system is based on the naturalness of the speech from which emotion has to be extracted. A typical set of emotions contains 200 emotional states. Therefore to classify such a great number of emotions is very complicated. Primary emotions are anger, disgust, fear, joy, sadness and surprise. This required emotional acting by the speaker and a short emotional story was narrated to him to sufficiently arouse the same emotion in him. This is termed as simulated emotion. The set of utterances were recorded in two different sessions. In one session the utterances corresponding to sad, disgust and fear emotions were recorded in the same order, which generally have similar acoustic characteristics. In the other session the utterances corresponding to happy, surprise and angry emotions were recorded in the same order, which also generally have similar acoustic characteristics. The neutral utterances were recorded at the beginning of any of the above sessions.

There are three kinds of emotional databases with regard to the authenticity of emotion. In general actors are asked to speak some given utterances while expressing a certain emotion and the recording is labelled as containing the specified desired emotion. Another kind of databases contains elicited emotions. This kind of emotion is neither real, nor simulated. The last types are databases of spontaneous speech which contain real emotions.

5. Experiments

In this study we have designed two experiments: (i) Speaker dependent but text independent and (ii) Speaker independent and text independent. In experiment (i) the simulated emotion and the induced emotion utterances of all the speakers are considered for training and testing respectively. In the training of GMM classifier by expectation-maximization (EM) algorithm, its mean-vectors are randomly initialized. Hence, it is quite likely that in two or more different runs of the training program the GMM parameters (i.e. mean-vectors, variances and coefficients) converges to different values. Hence, the experiment (i) is repeated 10 times and finally average success score is computed. In experiment (ii), the simulated emotion utterances of a set of 24 speakers (12 Male and 12 Female) are considered for training and the induced emotion utterances of remaining 15 speakers are considered for testing. The experiment (ii) is repeated 10 times to consider all the speakers for training and testing at least once. Also, for each set of training and testing speakers the experiment (ii) is repeated 10 times and finally average success score for all the 10 sets of speakers is computed.

i) Research on Emotional Intonation

The research works related with Emotion from speech has been done till date to examine how vocal emotions are encoded, from which we know that emotional meanings in the voice are conveyed by changes in several acoustic parameters of speech, including to fundamental frequency, duration, rhythm, and different aspects of voice quality. The fact that most of the researchers have measured changes in pitch, intensity, and speech rate implies that these parameters are critical features of emotional expressions. In particular case, a speaker's pitch level, pitch range, and speech rate appear to differentiate among discrete emotion categories in both acoustic and perceptual terms. For example, expressions of sadness tend to be produced with a relatively low pitch and slow speaking rate, whereas expressions of anger and happiness tend to be produced with a moderate or high mean pitch and fast speaking rate.

ii) Functional and Emotional Recognition

The most usage of any speech corpora is for training a speech recognizer. When we listen to any speech, we can recognize the speaker's interactive function and emotional state. Since the human speech conveys the speakers' interactive intention and emotional state. A given emotion is considered as a result of the interaction among acoustical, psychological, and physiological features. Emotion is experienced at a time when something unexpected happens, arising suddenly in response to a particular event. For natural human-machine interaction, there is a requirement of machine based functional and emotional intelligence. For satisfactory responses to human interactive functions and emotions, Computer systems need accurate function and emotion recognition and that correct recognition of human interactive function and emotion improves efficiency of human-machine interaction.

6. Result

i) Recognition Accuracy

The analysis part signifies the recognition accuracy in percentage for each known test speech input to the total trained emotional speech data.

Accuracy= (Correctly detected Emotions inputs/ Total trained emotions inputs) x 100%

The accuracy for each classifier for the six emotions is calculated on the basis of above relation. Percentage success scores by GMM classifier with 43 MFCC features.

Table 1 Recognition accuracy (In percentage)

Emotion	Angry	Happy	Sad	Natural	Fear
GMM	100	67	89	73	50

Table 2: Confusion matrix for GMM classifier (In percentage)

Responded	Angry	Happy	Sad	Natural	Fear
Angry	100				
Happy		67			
Sad			87	11	
Natural		19		73	8
Fear					50

7. Conclusion

Through this paper, another design and development of a database of functional and emotional intonation in Assamese is described. The main purpose of our database is to work as a resource for studying functional and emotional intonation and its recognition and synthesis. It is based on conversations from movies and TV lays of about 100 hours, with utterances segmented, information recorded. Moreover, syllable and prosody annotation is done and pitch is extracted and manually corrected. The database contains large number of utterances and their transcriptions, pitch values etc. by various speakers.

8. References

[1] A. Razak, A. H. M. Isa and R. Komiya, "A Neural Network Approach for Emotion Recognition in Speech", Proc. 2nd Int. Conf. Art. Intell. In Engineering & Technology, Aug 3-5, 2004, Kota Kinabalu, Sabah, Malaysia.

[2] A. Hanson and T. H. Applebaum, "Robust Speaker- Independent Word Recognition Using Static, Dynamic and Acceleration Features: Experiments With Lombard and Noisy Speech", .ICASSP, 1990, pp.857-860.

[3] D.Womack and J. H. L. Hansen, "N-channel hidden Markov models for combined stressed speech classification and recognition," IEEE Trans. Speech Audio Process., vol. 7, no. 6, pp. 668-676, Nov. 1999.

[4] A. Reynolds and R. C. Rose,"Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models", IEEE Trans Speech Audio Process, Vol. 3, No. 1. Jan. 1995, pp. 72 -83.

[5] Ververidis, C. Kotropoulos, I. Pitas, "Automatic Emotional Speech Classification", ICASSP, 2004, pp (I- 593) - (I-596).

[6] Zhou, J. H. L. Hansen, and J. F. Kaiser, "Nonlinear feature based classification of speech under stress," IEEE Trans. Speech Audio Process., vol. 9, no. 3, pp. 201-216, Mar. 2001.

[7] **http://www.ciil.org/Main/languages/index.htm**

[8] http://www.wikipedia.com

[9] J. H. L. Hansen and B. Womack, "Feature analysis and neural network based classification of speech under stress," IEEE

[10] Trans. Speech Audio Process., vol. 4, no. 4, pp. 307- 313, Jul.1996.

[11] L.R. Rabiner and M.R. Sambur, "An Algorithm for Determining the Endpoints for Isolated Utterances",

[12] The Bell System Technical Journal, Vol. 54, No. 2, Feb. 1975, pp.297-315

[13] R. Cowie and E. Douglas-Cowie, "Automatic statistical analysis of the signal and prosodic signs of emotion in speech," in Proc. 4th Int. Conf. Spoken Language Processing. Philadelphia, PA, 1996, pp.1989-1992.

[14] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G.Votsis,

[15] S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction", IEEE Signal Process. Mag., vol. 18, no. 1. pp.32-80, Jan.2001.

[16] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", IEEE Trans. Audio Speech and Signal Process., vol. 28, no. 4, Aug. 1980, pp 357 -365.