International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970)
Volume-3 Number-3 Issue-11 September-2013

Emotion Recognition with Speech for Call Centres using LPC and Spectral Analysis

Rashmirekha Ram[1], Hemanta Kumar Palo[2], Mihir Narayan Mohanty[3]

Abstract

Emotion recognition from human speech is a challenge for the researchers. It is mostly considered under ideal acoustic conditions. The performance of such system is degraded while there is existence of environmental mismatches between training and testing phases. For robust speech recognition it requires for reduction of redundancy, variability, and capturing ability of speech signals in noisy environments. Cepstral coefficients are popularly used features and derived from linear predictive coding (LPC). In that case speech signal is assumed to be the output of the all-pole linear filter simulating the vocal tract of a human being. Such recognition systems with LPC-derived cepstrum work well in clean environments, where the performance is carried with these features. The result obtained using this method has a great achievement. In this paper, the overall emotion recognition process has two goals. The first goal is to provide an update record of the available emotional speech data. The number of emotional states, the number of speakers, and the kind of speech are briefly addressed. The second goal is to present the most frequent features used for emotional speech recognition and to assess how the emotion affects them. Two features have been considered for emotion recognition as linear predictive coding (LPC) and spectral analysis.

Keywords

Emotions, emotional speech data collections, emotional speech classification, LPC

1. Introduction

Great progress in speech recognition has yielded many practical applications in recent years, such as user-friendly speech interfaces in control consoles of cars, credit card number recognition and the verbal selection of menus over the telephone [1-5]. Speech is considered one of the most reliable and further more comfortable modalities to automatically estimate a person's emotion [4],

especially as no wiring is needed, and a person may control the amount of emotion shown. Though research efforts and considerable advancement in ASR has been approached by the researchers since two to three decades, still it remains a challenging problem for robustness and accuracy. Ability to understand human emotions is desirable for computer in application such as lie detector, developing learning environments, consumer relations, entertainment etc. Experiments have been conducted for designing intelligent human–machine interaction by simulating emotion intelligence of human brain. Machine can recognize "what is said" and "who said it" using speech recognition and speaker identification techniques [4]. If equipped with emotion recognition techniques, machines can also know "how it is said". In the field of human computer interaction (HCI), apart from facial expression and gestures, speech is a power medium to communicate with emotional intelligence. This paper explores the method to recognize human emotions (sad, surprise, bored, angry, happy) in speech signal. A number of approaches aiming at automatic recognition of emotion out of speech utterances have been presented over the last decade [3-5], [9-10].

2. Human Emotion

Over past two decades study and research on human emotion has been focussed through many fields including medical science, neuroscience, psychology, sociology, and computer science. Numerous theories attempt to explain the origin, neurobiology, experience, and function of emotions have only fostered more intense research on this topic. The present scenario is being conducted about the concept of emotion for development of methods to analyse emotion. Words are not enough to correctly understand the mood and intention of a speaker and thus the introduction of human social skills to

Rashmirekha Ram Siksha 'O' Anusandhan University Bhubaneswar, Odisha, India.
Hemanta Kumar Palo Siksha 'O' Anusandhan University Bhubaneswar, Odisha.
Mihir Narayan Mohanty Siksha 'O' Anusandhan University Bhubaneswar, Odisha

human-machine communication is of paramount importance. This can be achieved by the researching and creating methods of speech modelling and analysis that embrace the signal, linguistic and emotional aspects of communication [7]. Affective computing aims at providing more effective and natural human computer interfaces. The goal of most of the efforts in affective computing is recognizing emotions, such as anger or boredom. However, stress is paid hardly any attention, even though everybody experiences stress at work or in everyday situations. These situations may even be dangerous, for example while driving the car during rush hour. An understanding car management system could be of great importance in these situations, by calming down the driver by playing different music or warning him about his current emotional status. A neutral voice tone has an even, relaxed quality without marked stress on individual syllables. The anger communicates displeasure, irritation, annoyance or frustration. A subject reflects happiness when the voice is high pitched, or has a sing-song tone, that is not whining. Speech is faster or louder than usual, but not angry. An anxious state is expressed when the speaker communicates anxiety, nervousness, fear or embarrassment. An elevated voice volume, accompanied by rapid speech is common. A dysphoric state is evident when the subject communicates sadness and depression with a low voice tone and slow pace of speech. Emotion plays a significant role in cognitive psychology, behavioural sciences and humanoid robot design. The continuing improvements in speech recognition technology have led to many new and fascinating applications in human-computer interaction, context aware computing and computer mediated communication.

It has been created a list of emotions as described in Parrot (2001), where emotions were categorised into a short tree structure.

Table 1:Types of Human Emotions

Primary emotion	Secondary emotion	Tertiary emotions
Love	Affection	Adoration, affection, love, fondness, liking, attraction, caring, tenderness, compassion, sentimentality
	Lust	Arousal, desire, lust, passion, infatuation
	Longing	Longing
Joy	Cheerfulness	Amusement, bliss, cheerfulness, gaiety, glee, jolliness, joviality, joy, delight, enjoyment, gladness, happiness, jubilation, elation, satisfaction, ecstasy, euphoria
	Zest	Enthusiasm, zeal, zest, excitement, thrill, exhilaration
	Contentment	Contentment, pleasure
	Pride	Pride, triumph
	Optimism	Eagerness, hope, optimism
	Enthrallment	Enthrallment, rapture
	Relief	Relief
Surprise	Surprise	Amazement, surprise, astonishment
Anger	Irritation	Aggravation, irritation, agitation, annoyance, grouchiness, grumpiness
	Exasperation	Exasperation, frustration
	Rage	Anger, rage, outrage, fury, wrath, hostility, ferocity, bitterness, hate, loathing, scorn, spite, vengefulness, dislike, resentment
	Disgust	Disgust, revulsion, contempt
	Envy	Envy, jealousy

	Torment	Torment
Sadness	Suffering	Agony, suffering, hurt, anguish
	Sadness	Depression, despair, hopelessness, gloom, glumness, sadness, unhappiness, grief, sorrow, woe, misery, melancholy
	Disappointment	Dismay, disappointment, displeasure
	Shame	Guilt, shame, regret, remorse
	Neglect	Alienation, isolation, neglect, loneliness, rejection, homesickness, defeat, dejection, insecurity, embarrassment, humiliation, insult
	Sympathy	Pity, sympathy
Fear	Horror	Alarm, shock, fear, fright, horror, terror, panic, hysteria, mortification
	Nervousness	Anxiety, nervousness, tenseness, uneasiness, apprehension, worry, distress, dread

3. Methodology

Speech signals are non-stationary in nature. Hence it is a difficult task to recognise the emotion from speech. Again, instead of the whole signal, the major contents can be obtained from its suitable features. Features are the parametric representation of speech signal at a lower information rate. Hence, extraction of efficient features for processing is the important task. As a result, the speech waveforms are commonly split into small frames (typically 5 ms to 40 ms) in which the signal characteristics are considered quasi-stationary to allow for short-term spectral analysis and feature extraction. This parametric representation of speech may be used to generate input feature to the recognition models. The most commonly used speech features include the Linear Predictive Coding (LPC) Coefficients [5-8].

First, decoding of emotions in speech is complex process that is influenced by cultural, social, and intellectual characteristics of subjects. It can not be perfect for decoding such manifest emotions as various types. But, emotion of anger can be easily recognized. It has a great impact on business. But anger has numerous variants such as hot anger, cold anger, etc. It can generate variability into acoustical features as well as influence the accuracy of recognition. Due to such type of variation, the research of emotional voice is still at an early stage for accuracy on real applications. In this work a bit of effort is given to solve such problem to an extent.

Linear Predictive Analysis

Saving bits in speech coders, for example, relies on a perceptual tolerance to acoustic deviations from the original speech. Speech information is extracted from several speech coding and recognition systems.

Emotion recognition system consists of 4 modules (speech input, spectral analysis, error calculation using LPC, recognizing emotion output). Solution to emotion recognition depends on type and purpose of emotion.

This study used a speech database which contains speech sample from 25 different females. The voices of a single speaker with emotions like anger, sad, surprise, bored, happy has been analyzed.

Linear Predictive Coding (LPC) determines the coefficients of a forward linear predictor by minimizing the prediction error in the least square sense. LP models are used for speech coding, recognition and enhancement. A LP model as in [6], can be represented mathematically,

$$\sum_{k=1}^{p} a_k \, x(m-k) + e(m)$$

where, x (m) is speech signal, a_k are the LP parameters and $e(m)$ is speech excitation. Note that the coefficients a_k gives the correlation of each sample with the previous P samples whereas $e(m)$

models the part of speech that cannot be predicted from the past *P* samples.
The differences in the two waveforms create a need for the use of different input signals for the LPC filter in the synthesis or decoding. One input signal is for particular emotional sounds and the other is for rest. The LPC encoder notifies the decoder regarding various emotions by sending a single bit.

The basic steps of LPC processor include the following:

1) *Preemphasis*: The digitized speech signal, *s*(*n*), is put through a low order digital system, to spectrally flatten the signal and to make it less susceptible to finite precision effects later in the signal processing.
2) *Frame Blocking*: The output of preemphasis step, ~*s* (*n*) , is blocked into frames of *N* samples, with adjacent frames being separated by *M* samples.
3) *Windowing*: After frame blocking, the next step is to window each individual frame so as to minimize the signal discontinuities at the beginning and end of each frame.
4) *Autocorrelation Analysis*: The next step is to auto correlate each frame of windowed signal.
5) *LPC Analysis*: The next processing step is the LPC analysis, which converts each frame of $p + 1$ autocorrelations into LPC parameter set by using Durbin's method.
6) LPC Parameter Conversion to Cepstral Coefficients: LPC cepstral coefficients, is a very important LPC.

The LPC cepstral coefficients are the features that are extracted from speech signal. In this system, voice signal is sampled using sampling frequency of 8 kHz and the signal is sampled within 0.5 seconds, therefore, the sampling process results 4000 data.

Spectral Analysis
Linear prediction is a generally accepted method for obtaining all-pole representation of speech. However, in some cases, spectral zeros are important and a more general modelling procedure is required. The goal of SPECTRAL ESTIMATION is to describe the distribution (over frequency) of the power contained in a signal, based on a finite set of data. Estimation of power spectra is useful in a variety of applications, including the detection of signals buried in wideband noise. The POWER SPECTRAL DENSITY (PSD) of a stationary random process X_n is mathematically related to the autocorrelation sequence by the discrete-time Fourier transform [1]. In terms of normalized frequency, this is given by

$$p_{xx(}\omega_{)} = \sum_{m=-\infty}^{\infty} R_{xx}(m) e^{-j\omega m}$$

Parametric Methods are those in which the PSD is estimated from a signal that is assumed to be output of a linear system driven by white noise [7]. The basic and most popular method as Fourier Transform is used in this case to obtain spectrums of signals and from the spectrum the peak value is calculated for each signal.
The basic steps of spectral analysis include the following:

1. The preprocessing of speech signal has to be done first.
2. After completing speech signal preprocessing, the next step is carried out for the extraction of the numeric feature, which is the normalized power spectral density. Human speech has emotion, intonation, stresses, and therefore the signal takes significant fluctuations in energy, even within the same phoneme, which may lead to incorrect segmentation. Energy normalization is used to reduce the influence of the mentioned factors: the energy of each time segment should be equal to unity, which allows taking into account only the energy distribution in frequency, rather than its absolute value.
3. In the first step the segmentation of the speech signal into portions with 50% overlap is carried out, which improves the temporal localization of spectral change.
4. At the next step the time segment undergoes a discrete Fourier transform to obtain the signal spectrum:
5. In the next step the signal energy is normalized according to the expression.
6. Then the distance between *i*th and *j*th PSDs is defined.

The bar graph represents different magnitudes of different signals which are shown in Figure 2.

4. Result

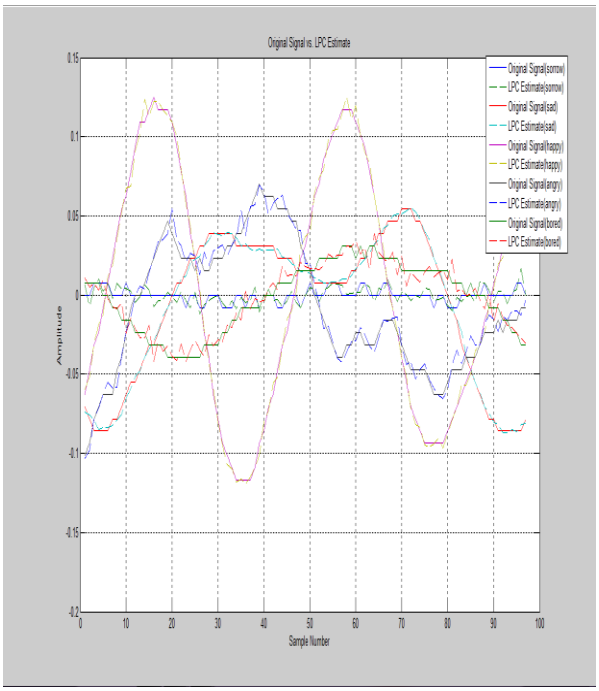

Fig:1: LPC Estimation of Five different Emotional Speech

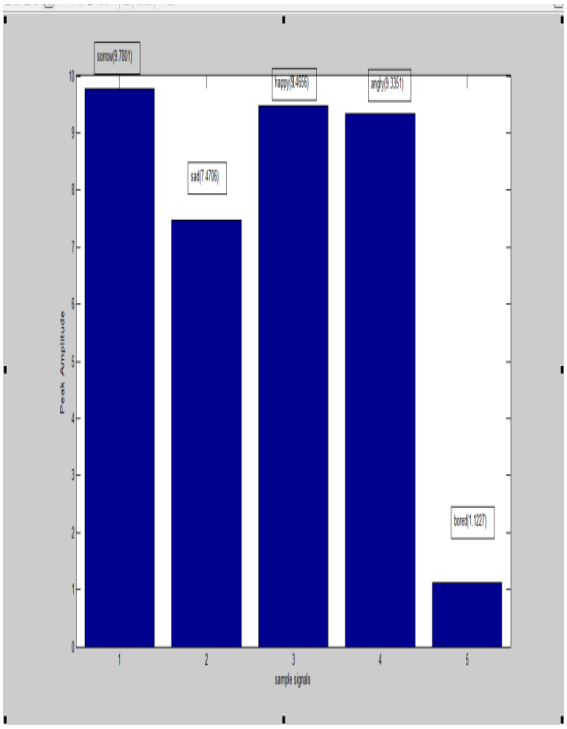

Fig:2. Spectral Analysis of five different Emotional Speech

Table 2: Prediction Error of signals

SPEECH SIGNAL	MINIMUM	MAXIMUM	RANGE
Sad	-0.0152	0.0122	0.0292
Surprise	-0.0389	0.1663	0.1967
Angry	-0.0456	0.0487	0.0842
Bore	-0.0328	0.0196	0.0434
Happy	-0.045	0.0365	0.0799

In this paper the emotional speech samples as angry, bored, happy, surprise, sad of female speakers is collected as the database. In the later stage both the method have been verified and is shown in Fig. 1 and 2 separately. As automatic emotion recognition based on speech matures, new challenges can be faced. We therefore address the major aspects in view of potential applications in the field, to benchmark today's emotion recognition systems and bridge the gap between commercial interest and current performances. These speech samples from the database were made to test and a comparison has been done. Following conclusions were drawn upon in these experiments.

From LPC calculation of a speech signal we found that 'sad' emotion has lowest prediction error than other emotions. 'Bored' emotion is in second position having lower prediction error and 'Surprise' emotion has highest error rate among all the emotional states.

Form spectral analysis we found that 'Surprise' emotion has highest magnitude i.e. it is having more energy than other emotions. 'Happy' emotion comes in second position having more energy. 'Bored' emotion has lower peak value i.e. it is having least energy than other emotions. For analysis verification we have tested through Sony Soundforge 7.0. and Gold wave standard software. The result is even accurate than these.

5. Conclusion

Since schemes for low bit-rate coding rely on signal manipulations that spread over durations of several tens of ms, and since schemes for speech recognition rely on phonemic articulatory information that extend over similar time intervals, it is concluded that the shortcomings are due mainly to a perceptually related rules over durations of 50-100 ms. These observations suggest a need for a study aimed at understanding how auditory nerve activity is integrated over time intervals of that duration. Information associated with quality of speech may be required and considered as an alternative for coding.

An effort is made in this paper to recognize human emotion by means of spectral analysis in which we can judge human emotion speech due to their different spectral peak. LPC analysis is another method to determine human emotion by giving different prediction error for different emotions.

References

[1] Mihir Narayan Mohanty, Bhagyalaxmi Jena, "Analysis of stressed human speech", Int. J. of Computational Vision and Robotics , Vol.2, No.2 , pp.180 – 187, 2011.

[2] Mihir Narayan Mohanty Aurobinda Routray Prithviraj Kabisatpathy, "A Statistical Approach for Voiced Speech Detection", Special Issue of IJCCT Vol. 2 Issue 2, 3, 4; 2010 for International Conference [ICCT-2010], 3rd-5th December 2010.

[3] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," IEEE Signal Processing Magazine, vol. 18, no. 1, pp. 32–80, January 2001.

[4] C. M. Lee and S. S. Narayanan, "Toward Detecting Emotions in Spoken Dialogs," IEEE Transactions on Speech and Audio Processing, vol. 13, no. 2, pp. 293–303, 2005.

[5] B. S. Atal and M. R. Schroeder, "Predictive coding of speech signals and subjective error criteria", IEEE Trans. Acoust., Speech, Signal Proc., vol. ASSP-27, pp. 247-254, June 1979.

[6] Thomas F. Quatieri, "Discrete-Time Speech Signal Processing", Prentice-Hall, third edition, 1996.

[7] Jacob Benesty, M. Mohan Sondhi, Yiteng Huang (Eds.), " Springer Handbook of Speech Processing", Springer-Verlag Berlin Heidelberg 2008.

[8] L. R. Rabiner and B. H. Juang, "Fundamental of Speech Recognition," 1st ed., Pearson Education, Delhi, 2003.

[9] B. S. Atal and S. L. Hanauer, "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," J. Acoust. Soc. Am, Aug. 1971, Vol. 50, No. 2, pp. 637-655.

[10] J. Makhoul, "Linear prediction: A tutorial review," Proceedings of the IEEE, vol. 63, pp. 561-580, April 1975.

Rashmirekha Ram received her Bachelor's Degree from Seemanta Engineering College, Mayurbhanj under Biju Patnaik University. She received her Master's degree from ITER under Siksha O" Anusandhan University , Bhubaneswar, Odisha. She is presently working as Lecturer in Zenith Institute of Science & Technology, Bhubaneswar, Odisha.

Hemanta Kumar Palo has completed his 'A.M.I.E.' from Institute of Engineers, India in 1997 and his 'Master of Engineering' from Birla Institute of Technology, Mesra, Ranchi in 2011. He completed his 'Diploma in Rail Transport and Management' from Institute of Rail Transport, India in 2003. He is having 20 years of experience in the field of Electronics and Communication Engineering from 1990 to 2010 in Indian Air Force and was an Assistant Professor in Gandhi Academy of Technology and Engineering, Odisha, in the Department of ECE from 2010 to 2012. He is the life associate member of IEI, India and is the memberof IEEE. Currently he is serving as an Assistant Professor in the Institute of Technical Education and Research, Siksha 'O' Anusandhan University, Bhubaneswar, Odisha.

Mihir Narayan Mohanty is presently working as an Associate Professor in the Department of Electronics and Communication Engineering, Institute of Technical Education and Research, Siksha 'O' Anusandhan University, Bhubaneswar, Odisha. He has published over 80 papers in International/National Journals and Conferences along with approximately 20 years of teaching experience. He is the active member of many professional societies like IEEE, IET, IETE, EMC & EMI Engineers India, IE (I), ISCA, ACEEE, IAEng etc. He has received his M.Tech. degree in Communication System Engineering from the Sambalpur University, Sambalpur, Odisha. Now he has done his Ph.D. work in Applied Signal Processing. He is currently working as Associate Professor and was Head in the Department of Electronics and Instrumentation Engineering, Institute of Technical Education and Research, Siksha O' Anusandhan University, Bhubaneswar, Odisha. His area of research interests includes Applied Signal and image Processing, Digital Signal/Image Processing, Biomedical Signal Processing, Microwave Communication Engineering and Bioinformatics. Some students are working under his guidance for research work in above mentioned fields, both from state and out of state.