Emotions and Strategies for Preparation of Emotional Speech Database

Meghana Nagori
Lecturer, Deparment of CSE, Govt. College Of Engg., Aurangabad, M.S., India, +91-9028582137
kshirsagarmeghana@gmail.com

Sarita T. Sawale
M.E. (CSE), Govt. College Of Engg., Aurangabad, M.S., India, +91-9881229017
stsawale@rediffmail.com

V. P. Kshirsagar
Lecturer, Deparment of CSE, Govt. College Of Engg., Aurangabad, M.S., India, +91-9890603949
vkshirsagar@gmail.com

Abstract
The exploration of how we as human beings react to the world and interact with it and each other remains one of the greatest challenges. The ability to recognize emotional states of a person perhaps the most important for successful inter personal social interaction. Automatic emotional speech recognition system can be characterized by the used features, the investigated emotional categories, the methods to collect speech utterances, the languages and the type of the classifier used in the experiment.

Since a well defined database is the necessary precondition for improving the performance Automatic emotional speech recognition systems. This paper explores the theories that explain the social and cognitive roles of emotions and mental states and their expression in human behaviors and communication. The paper describes the planning and accomplishment of a native language emotional speech database of acted emotional speech by number of speakers, recording strategies, conversion etc as well as the alternative approach is briefly addressed. Such database would also contribute to research in intonation and emotion.

Categories and Subject Descriptors

I.5 [PATTERN RECONGINITION]

I.5.4 [APPLICATIONS]

J.4 [SOCIAL AND BEHAVIORAL SCICENE]

H.1.2 [USER / MACHINE SYSTEM]

General Terms

Theory, Human Factors, Languages

Keywords

Keywords are your own designated keywords.

Emotions, emotional speech database, speech emotion recognition system.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

© Copyright 2010 Research Publications, Chikhli, India

Published by Research Publications, Chikhli, India

1 INTRODUCTION
An *emotion* is a mental and physiological state associated with a wide variety of feelings, thoughts, and behavior. Emotions are subjective experiences, often associated with mood, temperament, personality, and disposition. The English word 'emotion' is derived from the French word *émouvoir*. This is based on the Latin *emovere*, where *e-* (variant of *ex-*) means 'out' and *movere* means 'move' [1].

No definitive taxonomy of emotions exists, though numerous taxonomies have been proposed. Some categorizations include:

- Cognitive' versus 'non-cognitive' emotions.
- Instinctual emotions (from the amygdala), versus cognitive emotions (from the prefrontal cortex).
- Basic versus complex: where base emotions lead to more complex ones.
- Categorization based on duration: Some emotions occur over a period of seconds (e.g. surprise) where others can last years (e.g. love).

Ways of expression and recognition of emotions by humans and animals have intrigued researchers for a long time. Charles Darwin published the first monograph devoted to the topic in the nineteenth century [2]. After this milestone work, psychologists have gradually accumulated knowledge in the field offering various descriptions of emotive and affective states (e.g., [3]).

2 THEORIES OF EMOTION
Theories about emotions stretch back at least as far as the Ancient Greek Stoics, as well as Plato and Aristotle. The sophisticated theories are the works of philosophers such as René Descartes, Baruch Spinoza and David Hume. Later theories of emotions tend to be informed by advances in empirical research. Often theories are not mutually exclusive and many researchers incorporate multiple perspectives in their work [1].

2.1 Somatic theories
Somatic theories of emotion claim that bodily responses rather than judgements are essential to emotions. The first modern version of such theories comes from William James in the 1880s. The theory lost favour in the 20th Century, but has regained popularity more recently due largely to theorists such as John Cacioppo, Joseph E. LeDoux and Robert Zajonc who are able to appeal to neurological evidence.

William James, in the article 'What is an Emotion?' [4], argued that emotional experience is largely due to the experience of bodily changes. The Danish psychologist Carl Lange also proposed a similar theory at around the same time, so this position is known as the James-Lange theory.

2.2 Neurobiological theories

Based on discoveries made through neural mapping of the limbic system, the neurobiological explanation of human emotion is that emotion is a pleasant or unpleasant mental state organized in the limbic system of the mammalian brain. If distinguished from reactive responses of reptiles, emotions would then be mammalian elaborations of general vertebrate arousal patterns, in which neurochemicals (e.g., dopamine, noradrenaline, and serotonin) step-up or step-down the brain's activity level, as visible in body movements, gestures, and postures.

More recent research has shown that some of these limbic structures are not as directly related to emotion as others are, while some non-limbic structures have been found to be of greater emotional relevance.

2.3 Cognitive theories

There are some theories on emotions arguing that cognitive activity in the form of judgements, evaluations, or thoughts is necessary in order for an emotion to occur. This, argued by Richard Lazarus, is necessary to capture the fact that emotions are about something or have intentionality. Such cognitive activity may be conscious or unconscious and may or may not take the form of conceptual processing. An influential theory here is that of Lazarus. A prominent philosophical exponent is Robert C. Solomon [5]. The theory proposed by Nico Frijda where appraisal leads to action tendencies is another example. It has also been suggested that emotions (affect heuristics, feelings and gut-feeling reactions) are often used as shortcuts to process information and influence behaviour [6].

3 SPEECH EMOTION RECOGNITION SYSTEM

Like the typical pattern recognition system, speech emotion recognition system contains four main modules: emotional speech input, feature extraction, classification, and recognized emotion output. The structure of system is depicted in Figure 1 [7].

Figure 1: Structure of speech emotion recognition system.

Emotional speech recognition aims at automatically identifying the emotional or physical state of a human being from his or her voice. The emotional and physical states of a speaker are known as emotional aspects of speech and are included in the so called paralinguistic aspects. Although the emotional state does not alter the linguistic content, it is an important factor in human communication, because it provides feedback information in many applications.

3.1 APPLICATION OF SPEECH EMOTION RECOGNITION SYSTEMS

The most important application is in intelligent human-machine interaction. In today's human-machine interaction systems, machines can recognize "what is said" and "who said it" using speech recognition and speaker identification techniques. If equipped with emotion recognition techniques, machines can also know "how it is said" to react more appropriately, and make the interaction more natural. Other applications of automatic emotion recognition include psychiatric diagnosis, intelligent toys, and lie detection [8].

Making a machine to recognize emotions from speech is not a new idea. The first investigations were conducted around the mid-eighties using statistical properties of certain acoustic features [9, 10]. Ten years later, the evolution of computer architectures made the implementation of more complicated emotion recognition algorithms feasible.

Market requirements for automatic services motivate further research. In environments like aircraft cockpits, speech recognition systems were trained by employing stressed speech instead of neutral [11]. The acoustic features were estimated more precisely by iterative algorithms. Advanced classifiers exploiting timing information were proposed [12, 13, and 14]. Nowadays, research is focused on finding powerful combinations of classifiers that advance the classification efficiency in real life applications. The wide use of telecommunication services and multimedia devices paves also the way for new applications. For example, in the projects "Prosody for dialogue systems" and "SmartKom", ticket reservation systems are developed that employ automatic speech recognition being able to recognize the annoyance or frustration of a user and change their response accordingly [15, 16]. Similar scenarios are also presented for call center applications [17, 18] and also presented a two-stream emotion recognition technique for emotion recognition that can be used in call-center monitoring [19]. Emotional speech recognition can be employed by therapists as a diagnostic tool in medicine [20].

In psychology, emotional speech recognition methods can cope with the bulk of enormous speech data in real-time extracting the speech characteristics that convey emotion and attitude in a systematic manner [21].

Now a days Music emotion plays an important role in music retrieval, mood detection and other music-related applications. Many issues for music emotion recognition have been addressed by different disciplines such as physiology, psychology, cognitive science and musicology [22]. One of the first studies of emotion detection in music is presented by Feng *et al,* their work based on Computational Media Aesthetics (CMA), analyzes two dimensions of tempo and articulation which are mapped into four categories of moods: happiness, anger, sadness and fear [23].

Also Mobile emotion measurement (MEM) through physiological signals is a promising tool for both experiments and application. MEM would be of great benefit for mobile HCI [24].

On the other hand, AI and speech technology researchers have made contributions in the following areas: emotional speech

synthesis, recognition of emotions for improving speech recognition systems and for solving applied problems, as well as using agents for decoding and expressing emotions.

4 PREPARATION OF EMOTIONAL SPEECH DATABASE

The voice is especially important where there are no visual signals, such as in communication via radio or telephone, or where there is visual impairment. Several reviews on emotional speech analysis have already appeared. In [25] 64 data collections were reviewed.

There are many problems surrounding database development, some of which may not become obvious until it is too late. The paper describes the planning and accomplishment of a native lenguage database of acted emotional speech, containing different sentences performed in basic target emotions by number of subjects or actors. Such database can be the basis for analyses of prosodic features [26], articulatory features [27] and the verification by means of resynthesis [28].

4.1 Acted emotions

The first part argues why the material is to be recorded in acted instead of “real life” situations. However, as clear emotional expression is not only rare in everyday situations but also the recording of people experiencing absolute emotions is ethically problematic, it is almost impossible to use natural data if basic emotions are the subject of investigation.

4.2 Speakers

For parameter analysis undistorted speech signals without background noise are required. In order to investigate emotional speech as deviation from neutral speech it is necessary to record the same utterance in different emotional situations. In consequence the recording should be done systematically under laboratory conditions. In some psychological experiments, it has been tried to induce specific emotions into test persons, but for ethical reasons it is undesirable to induce negative emotions into test persons. As a consequence the emotional speech is to be spoken by the subjects. This emotions simulated by subjects are a good approximation to true emotional speech.

In recordings of a speaker reporting from a dramatic event is compared with recordings of an actor simulating the reporter's emotional state during the event. Differences between the recordings will be found, but in general the mode of speaking and the fundamental frequency range and variation are alike.

It is however not advisable to use stage actors, because they tend to exaggerate some features to make the emotional content very clear which makes the utterances sound unnatural.

Therefore another approach is to leave this matter open and search for performers by means of a newspaper advertisement or random selection from available subjects. In preselection session the selected subject will perform one utterance in each of the target emotions which will be recorded in an office directly with a microphone to hard disk. From these sessions, expert listeners can select peoples, equally representing the sexes, by judging the naturalness and recognisability of the performance.

4.3 Text material

Emotions can confidentially be recognized in very short utterances like "Yes" or "No". This means short sentences or even single words are appropriate to analyse emotional features in speech. But it can be interesting to analyse passages of "fluent" speech to study pauses and specific emotional sounds like laughter or sighs. It is best to choose some utterances which appear often in everyday communication. Two different kinds of text material would normally meet these requirements:

- Nonsense text material, like for instance haphazard series of figures or letters, or fantasy words (e.g. [29]).
- Normal sentences which could be used in everyday life.

Nonsense material is guaranteed to be emotionally neutral. However, there is the disadvantage that subject will find it difficult to imagine an emotional situation and to produce natural emotional speech spontaneously. This is why nonsense material rather results in stereotyped overacting.

In comparison with poems and nonsense sentences, the use of everyday communication has proved best, because this is the natural form of speech under emotional arousal. Moreover, actors can immediately speak them from memory. There is no need for a longer process of memorising or reading them off a paper, which may lead to a lecturing style [30].

In the construction of the database, priority can be given to the naturalness of speech material and thus everyday sentences could be used as test utterances. A total of five sentences, of length vary from 2 to 5 Seconds can constructed so that they could be interpreted in the target emotions. Moreover, they are utterances which both from their choice of words and their syntactic construction may be used in everyday life. The text is presented to the speakers as a list of separate sentences.

4.4 Recording Strategies

The subject takes then in to the recording room and place in a chair at a table on which the microphone is placed. The prompting text is placed on the table. The actor could be asked to speak in different emotions before the recording in order to set the recording level. It is advisable to read all the utterances with one emotion and then change the emotion and start over again. In this way the actors will not have to change emotions more than target number of emotions.

To achieve a high audio quality the recordings can take took place in sound proof recording studio using a high quality microphone is used, which did not influence the spectral amplitude or phase characteristics of the speech signal. Such as

- Sennheiser MKH 40 P 48 microphone and a Tascam DA-P1 portable DAT recorder. Recordings were taken with a sampling frequency of 48 kHz and later down sampled to 16 kHz was used for German emotional database.
- A portable Digital Audio Tape recorder Sony TCD-D8 at 48 kHz sampling rate via Sennheiser headphone set. The obtained recordings were converted into monophonic Windows PCM format at 32 kHz sampling

frequency and 16 bits resolution was used for RUSLANA database

- AKG 414 ULS microphone, Amek Angela 36 ch. in - line Mixerdesk ,PANASONIC DAT SV 3500 was for Danish emotional database etc

Also easily available ASM- 580XLR Microphone and USR-60 recorder with a sampling frequency of 32 kHz and later resample to 10 kHz can be used for recording in native language such as English, Marathi, and Hindi etc. Resampling is synonymous with several processes commonly used in manipulating audio, through which a segment of sampled audio termed as sample is manipulated before being stored back to a sampled format.

The text of every utterance was prompted to every subject to avoid reading intonation style. Subjects were instructed to put themselves into specific emotional states and speak the sentence. They are asked to remember a real situation from their past when they had felt this emotion. Few rehearsals are taken before recording every test. The distance between camera and subject can be kept about 30 cm but slight variations in the distance between mouth and microphone, sound intensity and environmental conditions are allowed, so as to make data set more close to the reality.

4.5 Conversion

For feature extraction speech processing toolbox in MATLAB programming language as well as Praat, Sonogram, VoiceBox tools etc can be used. Praat] has a pointy-clicky interface, is high level and does not require deep knowledge in the field of signal processing [31]. From all these MATLAB is most widely used for extracting features from speech as well as image. To cross validate extracted features from MATLAB with advance toolbox Praat; required conversion from MP3 Stereo, Dual Channel files into Mono, Single Channel, and wave format, since MATLAB speech processing functions are supporting for wave file format only. Such converter software's are

- mp3-2-wav
- Super MP3 converter etc.

4.6 Labeling the data

In order to be able to compare the results with older studies of research group [32, 33, 34] the same emotional labels can be used, neutral, anger, fear, joy or happiness, sadness, disgust and surprise.

4.7 Evaluation

Evaluation is very important but time consuming stage in database development. We can employ two procedures for our database evaluation. Both of them require participation of human evaluators who are presented with randomized utterances. The objective of the first procedure is to put the utterance on the activation-evaluation wheel [35], which has been derived from the Plutchik's "emotion wheel" [36]. The activation evaluation wheel is a unit radius circle on activation-evaluation axes. The x-axis is the evaluation or valence axis with positive values on the right side, and the y-axis is the activity axis with high activity on the top. Two radial coordinates specify each point P on the wheel. The emotional state is represented by the angle between the positive y-axis and the vector from the center of the circle to the point P, whereas the strength of emotion corresponds to the distance of P from the center of the circle (Figure 2). The procedure provides a continuous spectrum of emotional states. However, it requires some training before the evaluator starts producing cohesive results. The objective of the second procedure is to estimate how well a subject portrayed the intended emotion. In this case the evaluator knows, which emotion the utterance is supposed to convey, and estimates its quality using the scale from 1 to 5 (where 1 means "very bad - no resemblance" and 5 means "excellent performance").

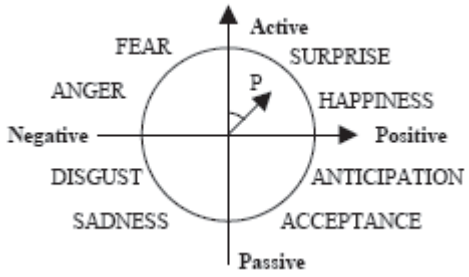

Figure 2: The Activation-evaluation wheel

5 ALTERNATIVE APPROACH

The alternative approach for preparing the database is by cutting the dialogs with length vary from 2 to 5 seconds and simulate the basic emotions from the native language movies by using the softwares such as

- Easy-mp3-cutter
- MP3Cutter
- Mp3splitter etc.

And then following the same procedure like: MP3 to wav conversion, labeling the data, evaluation and so on. Thus we can get the speech emotional database which will serve as input for speech emotional recognition system.

6 SUMMARY AND CONCLUSIONS

In fact a well defined database is the preliminary necessary prerequisite for improving the performance of speech emotion recognition systems. This paper gives the general guidelines for preparing the emotional database in native language with the basic emotions contain emotional utterances performed by native language subjects. Normal sentences which could be used in everyday life can be used for building such database. The recording in the sound proof studio is suggested to reduce the noise in the audio files. The material could be evaluated in an automated listening test and each utterance can judge by listeners with respect to recognisability and naturalness of the displayed emotion. Such database could serve as a basis for numerous studies.

A discrepancy appears between really experienced emotions and emotions mugged by the speaker under the influence of cultural tradition, language, situation, discoursal roles and other factors. Such database preparation is also facing more general issues: the types of emotions expressed in speech, and the relationship

between genuine and simulated emotion. Hope such database would also contribute to research in intonation and emotion.

7 REFERENCES

[1] http://en.wikipedia.org/wiki/Emotion

[2] Darwin, Ch. The expression of the emotions in man andanimals. Chicago: University of Chicago Press, 1965 (Original work published in 1872).

[3] Plutchik, R., Emotion: A Psychoevolutionary Synthesis.New York: Harper and Row, 1980.

[4] James, William (1884). "What is Emotion". Mind 9: 188–205. Cited by Tao and Tan.

[5] Solomon, R. (1993). The Passions: Emotions and the Meaning of Life. Indianapolis: Hackett Publishing.

[6] Drake, R. A. (1987). Effects of gaze manipulation on aesthetic judgments: Hemisphere priming of affect. Acta Psychologica, 65, 91-99

[7] YI-Lin and G. Wei "Speech emotion recognition based on HMM and SVM" Proceedings of the Fourth International Conference on Machine Learning and Cybernetics, Guangzhou, 18-21 August 2005.

[8] Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., and Taylor, J. G., "Emotion recognition in human-computer interaction", IEEE Signal Processing magazine, Vol. 18, No. 1, pp. 32-80, Jan. 2001.

[9] Van Bezooijen, R., 1984. The Characterisitcs and Recognizability of Vocal Expression of Emotions. Drodrecht,The Netherlands: Foris.

[10] Tolkmitt, F. J., Scherer, K. R., 1986. Effect of experimentally induced stress on vocal parameters. J. Experimental Psychology: Human Perception and Performance 12 (3), 302–313.

[11] Hansen, J. H. L., Cairns, D. A., 1995. ICARUS: Source generator based real-time recognition of speech in noisy stressful and Lombard effect enviroments. Speech Communication 16, 391–422.

[12] Deller, J. R., Hansen, J. H. L., Proakis, J. G., 2000. Discete-Time Processing of Speech Signals. N.Y.: Wiley.

[13] Womack, B. D., Hansen, J. H. L., 1996. Classification of speech under stress using target driven features. Speech Communication 20, 131–150.

[14] Polzin, T. S., Waibel, A. H., 1998. Detecting emotions in speech. In: Proc. Cooperative Multimodal Communication (CMC '98).

[15] Ang, J., Dhillon, R., Krupski, A., Shriberg, E., Stolcke, A., 2002. Prosody-based automatic detection of annoyance and frustration in human-computer dialog. In: Proc. Int. Conf. Spoken Language Processing (ICSLP '02). Vol. 3.pp. 2037–2040.

[16] Schiel, F., Steininger, S., Turk, U., 2002. The Smartkom multimodal corpus at BAS. In: Proc. Language Resources and Evaluation (LREC '02).

[17] Petrushin, V. A., 1999. Emotion in speech recognition and application to call centers. In: Proc. Artificial Neural Networks in Engineering (ANNIE 99). Vol. 1. pp. 7–10.

[18] Lee, C. M., Narayanan, S. S., 2005. Toward detecting emotions in spoken dialogs. IEEE Trans. Speech and Audio Process. 13 (2), 293–303.

[19] Purnima Gupta, Nitendra Rajput "Two-Stream Emotion Recognition For Call Center Monitoring," in INTERSPEECH 2007 August 27-31, Antwerp, Belgium

[20] France, D. J., Shiavi, R. G., Silverman, S., Silverman, M., Wilkes, M., 2000. Acoustical properties of speech as indicators of depression and suicidal risk. IEEE Trans. Biomedical Engineering 7, 829–837.

[21] Mozziconacci, S. J. L., Hermes, D. J., 2000. Expression of emotion and attitude through temporal speech variations. In: Proc. Int. Conf. Spoken Language Processing (ICSLP '00). Vol. 2. Beijing, pp. 373–378.

[22] Byeong-jun Han, Seungmin Rho Roger B. Dannenberg Eenjun Hwang "SMERS: Music Emotion Recognition Using Support Vector Regression" 10th International Society for Music Information Retrieval Conference (ISMIR 2009).

[23] Y. Feng, Y. Zhuang, Y. Pan : "Music information retrieval by detecting mood via computational media aesthetics," Proc. of IEEE/WIC Intl. Conf., Web Intelligence, pp. 235-241, 2003.

[24] Joris H. Janssen, Egon L. van den Broek "Guidelines for Mobile Emotion Measurement" MobileHCI09, September 15 - 18, 2009, Bonn, Germany ACM 978-1-60558-281-8/09/09

[25] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: resources, features, and methods," Speech Communication, vol. 48, no.9, pp.1163-1181, Sep. 2006.

[26] Paeschke, A., "Prosodische Analyse emotionaler Sprechweise", Reihe Mündliche Kommunikation, Band 1, Logos Berlin, 2003

[27] Kienast, M., "Phonetische Veränderungen in emotionaler Sprechweise", Shaker, Aachen, 2002

[28] Burkhardt, F., "Simulation emotionaler Sprechweise mit Sprachsyntheseverfahren". Reihe Berichte aus der Kommunikationstechnik, Shaker, Aachen, 2001

[29] Banse, R. & Scherer, K. R., "Acoustic Profiles in Vocal Emotion Expression", Journal of Personality and Social Psychology, Vol. 70, No. 3, p. 614-636, 1996

[30] Scherer, K. R., "Speech and Emotional States", in: Darby, J. K. (ed.), The Evaluation of Speech in Psychiatry, New York: Grune & Stratton, p. 189-220, 1981

[31] Boersma, P., Weenink, D.: Praat: doing phonetics by computer (Version 4.3.14) [Computer program]. 2005.

[32] Klasmeyer, G., „Akustische Korrelate des stimmlich-emotionalen Ausdrucks in der Lautsprache", Forum Phoneticum 67, Hector-Verlag, Frankfurt, 1999

[33] Sendlmeier, W.; Klasmeyer, G., "Voice and Emotional States", in: Voice Quality Measurement, p. 339-357, Singular, San Diego, CA, 2000

[34] Sendlmeier, W., „Phonetische Reduktion und Elaboration bei emotionaler Sprechweise“, in: Von Sprechkunst und Normphonetik, p. 169-177. Verlag Werner Dausien, Hanau, Halle, 1997

[35] R. Cowie et al. “Emotion recognition in Human-Computer Interaction”, IEEE Signal Processing Magazine, Jan 2001, vol. 18, No. 1, pp. 32-80.

[36] Plutchik, R., Emotion: A Psychoevolutionary Synthesis.New York: Harper and Row, 1980.

Author Biographies

Meghana Nogori received her Masters Degree in Computer Science & Enginerring from Thapar Institute of Engineering, Chandigarh, Panjab. She is currently working as a lecturer at the Department of Computer Science & Engineering at Government College of Engineering, Aurangabad [MS], India. She is having 12 years teaching experience and life time membership of CSI, ISTE. Her area of interest for research is algorithms, data mining. She has presented 04 papers in International Conference and more than 10 in national.

Sarita T. Sawale received her Bachelor’s Degree in Computer Science & Engineering from the Amravati university and perceiving Masters in same from Government College of Engineering, Aurangabad [MS], India. Her research interest includes operating system, network security, and pattern recognition. She has presented 01 paper in international conference, 02 in national conference.

Vivek Kshirsagar received his Masters Degree in Computer Science & Enginerring from Thapar Institute of Engineering, Chandigarh, Panjab. He is currently working as a lecturer at the Department of Computer Science & Engineering at Government College of Engineering, Aurangabad [MS], India. He is having 15 years teaching experience and life time membership of CSI, ISTE. His area of interest for research is Operating system, Networking. He has presented 04 papers in International Conference and more than 10 in national.