COMPUTATIONAL STUDIES OF EMOTION IN SPEECH AND FEATURE EXTRACTION TECHNIQUES

Mr Nikhil N. Patel[1] , Dr. Himanshu N. Patel[2],

Manager, CPP Software, Bhavnagar, Gujarat, India[1]

Assistant Professor, *Dept. of Computer Science, Dr. Babasaheb Ambedkar Open University, Ahmedabad, Gujarat, India[2].*

nnpatel1979@gmail.com[1], himanshu.patel@baou.edu.in[2]

Abstract—**in this paper we discuss various parameter related to speech emotion. We discuss few systems which approach the goal of recognizing emotion automatically from a speech input. We also describe Emotion Attributions and Features of Deafened People's Speech, Distinctions between Emotional and Neutral Passages Found by this system, Measures Concerning Continuous Acoustic Level.**

Keywords—ASSESS, Simulation, Emotion Recognition, Discriminant Analysis, Speech Parameter

I. INTRODUCTION

The main energy source in speech is vibration of the vocal cords. At any given time, the rate at which vocal cords vibrate determines the fundamental frequency of the acoustic signal, usually abbreviated to F_0. F_0 corresponds to perceived voice pitch. Vocal cord vibration generates a spectrum of harmonics, which is selectively filtered as it passes through the mouth and nose, producing the complex time-varying spectra from which words can be identified. Variations in voice pitch and intensity may also have a linguistic function. The patterns of pitch movement which constitute intonation mark linguistic boundaries and signal functions such as questioning. Linked variation in pitch and intensity mark words as stressed or unstressed. The term prosody refers to the whole class of variations in voice pitch and intensity that have linguistic functions.

Speech presents two broad types of information. It carries linguistic information insofar as it identifies qualitative targets that the speaker has attained in a configuration that conforms to the rules of language. Paralinguistic information is carried by allowed variations in the way that qualitative linguistic targets are realized. These include variations in pitch and intensity having no linguistic function and voice quality, related to spectral properties that aren't relevant to word identity.

The boundary between those streams is a matter of controversy. Linguists assume that there are qualitative targets that are understood intuitively by users, but not yet fully explicated, and that they actually account for a good deal of variation that is classed as paralinguistic. In particular, they tend to look for targets of that kind which underlie the expression of emotion. In contrast, biologists and psychologists tend to assume that the relevant information is defined by continuous variables, which carry paralinguistic information.

Four broad types of speech variable have been related toexpression of emotional states i.e. Pitch, Intensity, Duration and Spectral. Clearly there are relationships among the variables described above. For example, continuous spectral variables relate to voice quality, and the pitch contours described in the experiments must relate to the tune patterns arising from different heads and tones. Table 1 is a summary of relationships between emotion and speech parameters from a review by Murray and Arnott [4].

TABLE 1.
EMOTIONS AND SPEECH PARAMETERS

	Anger	**Happiness**	**Sadness**	**Fear**	**Disgust**
Rate	Slightly faster	Faster or slower	Slightly slower	Much faster	Very much faster
Pitch Average	Very much	Much higher	Slightly lower	Very much	Very much

	higher			higher	lower
Pitch Range	Much wider	Much wider	Slightly narrower	Much wider	Slightly wider
Intensity	Higher	Higher	Lower	Normal	Lower
Voice Quality	Breathy, chest	Breath, blaring tone	Resonant	Irregular voicing	Grumble chest Tone
Pitch Changes	Abrupt on stressed	Smoot, upward inflections	Down-ward inflections	Normal	Wide, down-ward terminal inflects
Articulation	Tense	Normal	Slurring	Precise	Normal

II. Speech Input Emotion Recognizing system

There are relatively few systems which approach the goal of recognizing emotion automatically from a speech input. This section reviews key examples of computational studies of Emotion in Speech.

A. Cowie and Douglas-Cowie ASSESS System

ASSESS [1], [2] is a system which goes part way towards a computational analysis. Automatic analysis routines generate a highly simplified core representation of the speech signal based on a few landmarks peaks and troughs in the profiles of pitch and intensity and boundaries of pauses and fricative bursts. These landmarks can be defined in terms of a few measures. Those measures are then summarized in a standard set of statistics. The result is an automatically generated description of central tendency, spread and centiles for frequency, intensity, and spectral properties.

That kind of feature extraction represents a natural first stage for emotion recognition, but in fact ASSESS has not generally been used in that way. Instead the measures described above have been used to test for differences between speech styles, many of them at least indirectly related to emotion. The results indicate the kinds of discrimination that this type of representation could support.

A precursor to ASSESS was applied to speech produced by deafened adults [5]. One of the problems they face is that hearers attribute peculiarities in their speech to emotion-related speaker characteristics. These evaluative reactions were probed in a questionnaire study, and the programs were used to elicit the relevant speech variables.

Table 2 summarizes correlations between emotion attributions and speech features. They suggest that ASSESS-type measures are related to judged emotionality, but also underline the point that the attribution of emotion is dogged by systematic ambiguity.

TABLE 2.
EMOTION ATTRIBUTIONS AND FEATURES OF DEAFENED PEOPLE'S SPEECH.

Response	**Speech Factors**
Judged stability	Relatively slow change in the lower spectrum
Judged poise	Narrow variation in F0 accompanied by wide variation in intensity
Judged warmth	Predominance of relatively simple tunes, change occurring in the mid-spectrum rather than at extremes; low level of consonant errors
Competence	Pattern of changes in the intensity contour

B. Simulation

In a later study, reading passages were used to suggest four archetypal emotions: fear, anger, sadness, and happiness [6]. All were compared to an emotionally neutral passage, and all passages were of comparable lengths.

Speakers were 40 volunteers from the Belfast area, 20 male and 20 female, between the ages of 18 and 69. There was a broad distribution of social status, and accents represented a range of local types. Subjects familiarized themselves with the passages first and then read them aloud using the emotional expression they felt was appropriate.Recordings were analysed using ASSESS. Table 3 summarizes the measures that distinguish the emotionally marked passages from the neutral passage.

TABLE 3.

DISTINCTIONS BETWEEN EMOTIONAL AND NEUTRAL PASSAGES FOUND BY ASSESS.

	Spectrum	**Pitch Movement**		**Intensity**		**Pausing**	
	Mid-point and slope	**Range**	**Timing**	**Marking**	**Duration**	**Total**	**Variability**

Afraid					+	+		
Angry		+	+		+	+		
Happy			+	+	+		+	+
Sad				+		+	+	

Figure 4 presents traces from an individual speakerarbitrarily chosen from the group studied by McGilloway[7] and shows how they relate to the kinds of features suggested by ASSESS analysis.

Figure 4(a)-(e) summarizes the output of initial processing on each of five signals—one neutral and four expressing specified emotions anger, fear, happiness and sadness. Time, in milliseconds, is on the horizontal axis. The heavy lines in each figure show signal intensity (referred to the left-hand scale, in decibels), the light lines represent pitch (referred to the left-hand scale, in hertz). Time scale (on the horizontal axis, in milliseconds) is adjusted to let the whole trace appear on the figure. The patterns are summaries in that inflections and silences have already been identified from the raw input, and the overall contours are represented by a series of straight lines (or gaps) between the resulting points. Several spectrum-like representations have also been computed, but found to contribute relatively little.

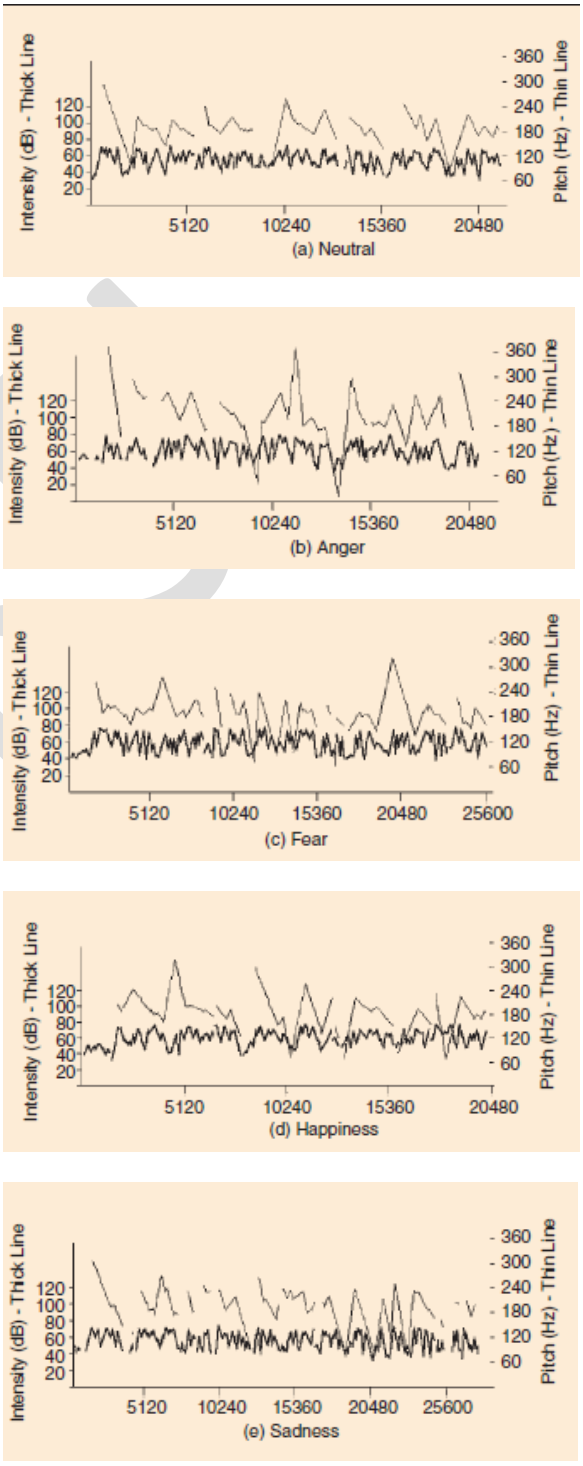
It is not self-evident from inspection that the contours differ systematically, but analysis indicates that they do. Each caption on the left-hand side refers to an output feature whose value in one or more emotional traces is significantly different from its value in the neutral passage. The features are selected from a much larger number that meet that basic criterion. Selection is geared to a) avoiding redundancy, b) representing the main logically distinct areas where differences occur, and c) achieving some formal consistency (e.g., using centile measures to describe central tendency and spread). The graph shows the fit of the speaker's data to a template based on the overall analysis.

The main points are that the kind of analysis embodied in ASSESS generates a range of features that are relevant to discriminating emotion from neutral speech and that different emotions appear to show different profiles. It remains to be seen how reliably individual signals can be assigned to particular emotion categories, but there are grounds for modest optimism.

C. Discriminant Analysis

A new application of the system illustrates the next natural step towards automatic discrimination [8]. Discriminant analysis was used to construct functions which partition speech samples into types associated with different communicative functions, e.g., opening or closing the interaction, conducting business. It is an interesting question how closely that kind of functional difference relates to emotion.

Fig. 4. (a)-(e) Output of initial processing on each of five speech passages



D. Banse and Scherer' s System

Scherer's group has a long record of research on vocal signs of emotion. A key recent paper [3] extracts a systematic battery of measurements from test utterances. The measures fall into four main blocks, reflecting the consensus of Scherer's group has a long record of research on vocal signs of emotion. A key recent paper [3] extracts a systematic battery of measurements from test utterances. The measures fall into four main blocks, reflecting the consensus of Table4). The emotions considered were hot anger, coldanger, panic fear, anxiety, desperation, sadness, elation, happiness, interest, boredom, shame, pride, disgust, and contempt. Discriminant analysis was then used to construct functions which partition speech samples into types associated with different types of expression. Classification by discriminant functions was generally of the order of 50% correct—which was broadly comparable with the performance of human judges.

It is natural to take the techniques described by Banseand Scherer as a baseline for emotion detection fromspeech. They show that automatic detection is a real possibility.

TABLE-4.

BANSE AND SCHERER'S MEASURES CONCERNING CONTINUOUS ACOUSTIC LEVEL.

Fundamental frequency	**Mean F0**	**Standard Deviation of F0**	**25th and 75th Percentiles of F0**
Energy	Mean of log-transformed microphone voltage		
Speech rate	Duration of articulation periods		Duration of voiced periods
Spectral measures	Long-term average spectra of voiced and unvoiced parts of utterances		

III. EMOTION FEATURE EXTRACTION METHODS

This section discussed dominant variables are extracted from the raw input for emotion detection and technique to extract them.

Voice Level: this was considered in previous section.

Voice Pitch: Voice pitch is certainly a key parameter inthe detection of emotion. It is usually equated with F0.Extracting F0 from recordings is a difficult problem, particularlyif recording quality is not ideal. It involves severalsubproblems, such as detecting the presence ofvoicing, the glottal closure instant [9], the harmonicstructure in a brief episode [10], short-term pitch instabilities(jitter and vibrato) [11], and fitting continuouspitch contours to instantaneous data points.

Phrase, Word, Phoneme and Feature Boundaries: DetectingBoundaries are a major, but difficult, issue in speechprocessing. That is why recognition of connected speechlags far behind recognition of discrete words. The issuearises at different levels.

Phrase/Pause Boundaries: The highest level boundarythat is likely to be relevant is between a vocal phrase and apause. Quite sophisticated techniques are available to locatepauses [12]. In [2] a method based on combiningseveral types of evidence is used, and it is reasonably successful.However, the process depends on empiricallychosen parameters, and it would be much better to havethem set by a learning algorithm—or better still, by a contextsensitive process. As noted above, pause length andvariability do seem to be emotionally diagnostic.

Word Boundaries: Speech rate is emotionally diagnostic,and the obvious way to describe it is in words per minute,which depends on recovering word boundaries. That turnsout to be an extremely difficult task, and probably the bestsolution is to look for other measures of speech rate, whichlend themselves better to automatic extraction. Findingsyllable nuclei is a promising option [13], [14].

Phoneme Boundaries: The report by Izzo indicates thatgood use can be made of information about phonemes ifthey can be identified. That directs attention to a large literatureon phoneme recognition [15]-[17].

Feature Boundaries: Some features, such as fricativebursts, are easier to detect than phonemes as such andthey appear to be emotionally diagnostic.

Voice Quality: A wide range of phonetic variables contributeto the subjective impression of voice quality [18].The simplest approach to characterizing it is based onspectral properties [19]. The report by Izzo reflects thattradition. A second uses inverse filtering aimed at recoveringthe glottal waveform (another task where neural nettechniques can be used to set key parameters [20]). Voicequality measures, which have been directly related toemotion, include open-to-closed ratio of the vocal cords,jitter, harmonics-to-noise ratio, and spectral energy distribution[21].

Temporal Structure: This heading refers to measures atthe pitch contour level and related structures in the intensitydomain. ASSESS contains several relevant types ofmeasure. The pitch contour is divided into

simple movements:rises, falls, and level stretches (see Table-5). Describingpitch movement in those terms appears to havesome advantages in the description of emotion overfirst-order descriptions (mean, standard deviation, etc).The intensity contour is treated in a similar way, andagain, descriptions based on intensitymovements seemto improve emotion-related discriminations.

TABLE-5

DURATION FEATURES AND EMOTIONS (MS)..

	Rises	**Falls**	**Tunes**	**Plateau**
	Median	Median	Median	IQR
Fear	82.35	84.8	1265	10.8
Anger	81.66	80.5	1252	10.2
Happiness	78.03	77.4	1404	8.2
Neutral	78.50	77.2	1452	8.4
Sadness	77.28	81.4	1179	11.0

Linguistically Determined Properties: There is a fundamentalreason for considering linguistic content in connectionwith the detection of emotion. On a surface level,it is easy to confound features which signal emotion andemotion-related states with features which are determinedby linguistic rules. The best known example involvesquestions, which give rise to distinctive pitchcontours that could easily be taken as evidence of emotionalityif linguistic context is ignored. Some work hasbeen done on the rules for drawing the distinction [22].

Other linguistic contexts which give rise to distinctivepitch contours are turn taking [23], topic introduction[24], and listing.It is worth noting that these are contexts that are likelyto be quite common in foreseeable interactions withspeech competent computers: systematically misinterpretingthem as evidence of emotionality would be anontrivial problem. The only obvious way to avoid confoundingin these contexts is to incorporate interveningvariables which specify the default linguistic structure andallow the observed speech pattern to be compared with it.

IV. CONCLUSION

We have describe approach techniques of recognizing emotion automatically from a speech input It is natural to take the techniques described above as a baseline for emotion detection fromspeech. They show that automatic detection is a real possibility.The substantial improvementsmight be made by automatic extraction of phonetic variables. The last section also discusses important extraction variable and method for extraction of these variables form row speech input.

REFERENCES

[1] R. Cowie and E. Douglas-Cowie, "Speakers and hearers are people: Reflections on speech deterioration as a consequence of acquired deafness," in *Profound Deafness and Speech Communication*, K-E. Spens and G. Plant, Eds. London, UK: Whurr, 1995, pp. 510-527.

[2] 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.

[3] R. Banse and K. Scherer, "Acoustic profiles in vocal emotion expression," *J. Personality Social Psych*, vol. 70, no. 3, pp. 614-636, 1996.

[4] I.R. Murray and J.L. Arnott, "Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion," *J. Acoust. Soc. Amer*., vol. 93, no. 2, pp. 1097-1108, 1993.

[5] R. Cowie and E. Douglas-Cowie, *Postlingually Acquired Deafness: Speech Deterioration and the Wider Consequences*. Berlin, Germany: Mouton deGruyter, 1992.

[6] S. McGilloway, R. Cowie, and E. Douglas-Cowie, "Prosodic signs of emotion in speech: Preliminary results from a new technique for automatic statistical analysis," in *Proc. 13th ICPhS*, Stockholm, Sweden, 1995, pp. 1989-1991.

[7] S. McGilloway, "Negative symptoms and speech parameters in schizophrenia," Ph.D. dissertation, Faculty of Medicine, Queen's University, Belfast, UK, 1997.

[8] E. Douglas-Cowie and R. Cowie, "International settings as markers of discourse units in telephone conversations," *Language and Speech (Special Issue, Prosody and Conversation)*, vol. 41, no. 3-4, pp. 351-374, 1998.

[9] EC TMR Project PHYSTA Report, "Hybrid systems for feature to symbolextraction" [Online]. Available: http://www.imge.ece.ntua.gr/physta/face-speech-features, July, 1998.

[10] J. Suzuki, M. Setoh, and T. Shimamura, "Extraction of precise fundamentalfrequency based on harmonic structure of speech," in Proc. 15th Int.Congr. Acoustics, vol. 3, 1995, pp. 161-164.

[11] M.P. Karnell, K.D. Hall, and K.L. Landahl, "Comparison of fundamentalfrequency and perturbation measurements among 3 analysis systems," J.Voice, vol. 9, pp. 383-393, 1995.

[12] B.L. McKinley and G.H. Whipple, "Model based speech pause detection,"IEEE Int. Conf. Acoust., Speech and Signal Processing, 1997, pp. 1179-1182.

[13] H.F. Pfitzinger, S. Burger, and S. Heid, "Syllable detection in read andspontaneous speech," in Proc. ICSLP 96, Philadelphia, PA, pp. 1261-1264.

[14] W. Reichl and G. Ruske, "Syllable segmentation of continuous speechwith artificial neural networks," in Proc. Eurospeech 93, vol. 3. Berlin, Germany,pp. 1771-1774.

[15] Y. Bengio, R. De Mori, G. Flammia, and H. Kompe, "Phonetically motivatedacoustic parameters for continuos speech recognition using neuralnetworks," in Proc. Eurospeech-91, Genova, Italy, pp. 551-554.

[16] A. Esposito, C.E. Ezin, and M. Ceccarelli, "Preprocessing and neural classificationof the English stops [b, d, g, p, t, k]," in Proc. ICSLP 96, vol. 2.Philadelphia, PA, pp. 1249-1252.

[17] A. Esposito and C. Ezin C., "A Rasta-PLP and TDNN based automaticsystem for recognizing stop consonants: Performance studies," VietrisulMare (SA), Italy, IIASS Int. Rep. I9602b, 1996.

[18] J. Laver, The Phonetic Description of Voice Quality. Cambridge, UK: CambridgeUniv. Press, 1980.

[19] B. Hammarberg, B. Fritzell, J. Gauffin, J. Sundberg, and L. Wedin, "Perceptualand acoustic correlates of voice qualities," ActaOtolaryngologica 90,pp. 441-451, 1980.

[20] W. Ding and H. Kasuya, "A novel approach to the estimation of voicesource and vocal tract parameters from speech signals," in Proc. ICSLP 96,Philadelphia, PA, pp. 1257-1260.

[21] G. Klasmeyer and W. Sendlmeier, "Objective voice parameters to characterize the emotional content in speech," in Proc. 13th Int. Congr. PhoneticSciences, vol. 2. Stockholm, Sweden, 1995, pp. 182-185.
[22] G. McRoberts, M. Studdert-Kennedy, and D.P. Shankweiler, "Role offundamental frequency in signalling linguistic stress and affect: Evidence fora dissociation," Perception and Psychophysics 57, pp. 159-174, 1995.
[23] A. Cutler and M. Pearson, "On the analysis of prosodic turn-taking cues,"in Intonation in Discourse, C. Johns-Lewis, Ed. London, UK: Croom Helm,1986, pp. 139-155.
[24] G. Brown, K. Currie, and J. Kenworthy, Questions of Intonation. London,UK: Croom Helm, 1980.