Evdokimova Vera
postmaster@phonetics.pu.ru

Automatic recognition of emotions and physical state of the speaker

Introduction

The speech technology implementation to the new interactive systems attracts the attention not only to the linguistic content of speech but also to its emotional part, which is the important information transmission channel of human communication [16].

The analysis of the "human – human" and "human – computer" dialogue characteristics shows two possible ways to deal with the emotions in speech and consequently two lines of research.

The first is the identification of the emotional state of the speaker. Emotion identification becomes the important part of software in service call centers, life support systems, training systems, interactive games and so on. Emotion identification gives an opportunity to improve the speech signal processing quality, to understand more clearly the user's demands and finally to improve the decision making of the system [23, 25].

Secondly, human emotion investigation opens up the possibilities of the modeling and extracting the emotions into interactive systems to develop the user friendly interface. There are different systems which already use these technologies. It provides for the naturalness of the human – computer communication [17].

The main stages of the phonetic emotion investigation are big speech databases gathering, their classification and analysis, extracting and formalization of the main features of expressing the emotions, these feature's description and modeling, elaborating the algorithms of emotional speech analysis and imitation. Different investigations are devoted to different problems from this list.

Analysis of the emotion and speaker state classifications

The investigation of the phonetic characteristics of the emotional speech can be carried out only after making the proper classification of the emotional states. There is no ideal list for this task to present day.

Emotion classifications of the researchers differ according to the goal of the research and the field, where the emotional speech recognition system can be applied. Also the scientist's opinion about the relevance of dividing different emotions is important.

Some people think that some of the emotions are primary and others are secondary. The secondary emotions can be described as combinations of the primary ones. This idea traces its root back to Descartes [10]. There is no list of basic emotions. However, it is possible to define the list of emotions which are usually chosen as basic ones: anger, disgust, fear, happiness, sadness, surprise.

In other works there is an attempt to use more broad categories of emotions and classify the emotional state of the speaker as negative or positive [10, 13]. Sometimes this division is called neutral speech vs. emotional speech [6, 9]. The majority of works rely on the negative-positive division. It is obvious because this division is a basic characteristic of emotions. The number of emotions in classifications varies from 2 to 14 (not containing the neutral or unemotional speech). For example:

1. Sadness, anger, happiness [5]
2. Anger, fear, surprise, disgust, happiness, sadness [17, 20]
3. Anxiety, disappointment, disgust, excitement, happiness, resignation, satisfaction, sadness [19]

The group of the American scientists even distinguishes the “cold” and “hot” anger [8].

Also the complexity of distinguishing of the emotions lies in the fact that there are almost no pure emotions expressed in speech [2, 24]. Usually there are several emotions and they are mixed up. Also emotions can depend on the extra linguistic factors (such as occurrence, situation, physiological state of the speaker etc.).

The task of the French scientists was to find out how the emotions can be mixed up [24]. They make an assumption that one of the emotions can be always seen as the main one and others as the additional to it. However sometimes it is impossible to distinguish the only one main emotion and there are cases when discordant emotions exist together (for example: compassion and irritation).

Besides that, lots of work is being done in the field of extra linguistic phenomena recognition in human speech. For example, laughter recognition [16]. There are works devoted to recognizing the lie in speech or the level of confidence of the speaker in what he is saying [22, 23].

Collecting the emotional speech databases for the speaker's emotional state recognition

The choice of the speech training material is one of the most important factors of the emotion recognition system quality. The set of the basic speech signal parameters is extracted from this database. The most complex part in gathering such speech corpus is the need to get the spontaneous speech with real spontaneous emotions, not prepared ones (read text, prepared dialogue etc.).

The speech databases can be divided into three kinds according to the way of recording [8]:

1. The first method is the records of the professional actors. All the utterances are said by one speaker expressing different emotions. The actors are usually given time to imagine themselves in different situations. This way of getting the material is widely used [5, 6, 15, 17, 19 etc.] Sometimes it is a database for several languages. The serials and film's dubbing can be considered as one of the ways of recording such speech material [18].

2. The second method is called the "Wizard-Of-Oz" (WOZ). The system which talks to people and puts them into certain emotional state is used. Then the system records all the responses [2, 3, 12, 21]. As an example the DARPA Communicator project could be mentioned where people were asked to make responses to the system to organize a trip somewhere [3]. Although the speakers didn't act any emotions their level of satisfaction or discontent was lower than in real call center's dialogues.

The other way to use WOZ systems is to play with the speaker and record their emotional responses. In one investigation it was a computer version of the game "Who wants to be a millionaire" [21].

In general the WOZ method of recording the emotions doesn't usually give the completely spontaneous emotional speech, however it gives an opportunity to avoid the prepared speech of the actors.

3. The third kind of gathering the emotional databases is the hardest one. It is the recording of the real human emotions. One of the most widely used ways to get such material is to record the telephone dialogues in call center services [11, 13, 24, 25], such as information services or the emergency services. Also the business negotiations are recorded (ICSI Meeting Recorder Corpus)[16], and the speech of the teachers and the students during the exam is recorded [6, 23].

In the majority of works the preliminary speech material is evaluated through the auditory experiments and evaluated by the listeners. However it is important to mention that the listeners couldn't always give the only one answer about the emotional state of the speakers, even when it was their own speech.

Therefore the high-quality speech emotional database is one of the most important parts of the automatic recognition of the emotional state of the speaker. It is very hard to gather such material. The actor's speech can contain false emotions. It is also very hard to analyze the material, because of the mixture of different emotions in real spontaneous speech. Perhaps, it makes sense to collect the speech using all three kinds of recordings.

The choice of the parameters for speech processing

Despite of the variety of aims and methods of investigation all the works in automatic emotional speech recognition have similar approaches. One of the most important parts of this work is to create the system of the formal parameters for the different emotions in speech. The number of parameters varies greatly. Several groups of these parameters can be marked out. Prosodic features (mainly Pitch and Energy) are classical features, used in a majority of applications and research systems. For accurate emotion detection in natural real-life speech dialogs, lexical, prosodic, disfluency and contextual cues should be considered and not only the prosodic information.

One of them is the group containing the parameters of the fundamental frequency and its variation. They are usually estimated using the short-term or long-time spectrum. Among them are:

F0 contour value, micro-variations, initial value, initial slope, coefficients of F0 contours stylization by first and second degrees splines, cumulated approximation errors of F0 contours stylization, mean absolute slope, minimum, maximum, mean and standard deviation statistics of fundamental frequency etc. Jitter and shimmer are related to the micro-variations of the pitch and power curves. So, they can be estimated as the slope change rate for these curves.

Other group deals with the energy characteristics of the signal:

Energy contour: value, micro-variations, initial slope, spectral flatness measure, spectral barycentre, HF/LF energy ratio, spectral variation, spectral envelope variation etc.

Also the group of temporal parameters is usually extracted. Among them are speaking rate (inverse of the average length of the speech voiced parts), number and length of silences (unvoiced portions between 200-800 ms) etc.

As an example the set of acoustic parameters in the work of Serbian scientist is given bellow [36]:

"An analysis of acoustic features was performed on a small part of the speech database GEES. 15 features are extracted from the selected speech material that characterizes intonation, intensity and durational characteristics of speech, as well as the quality of speech.

The first group of features, that characterize pitch *F0*, present: *F0-mean*, *F0-st.dev*, *d(F0)-mean* and *d(F0)-st.dev*. (Remark: *d(F0)* presents derivative of F0 across the analysis window). The other group consists of features that show intensity contour of each sentence as well as the contour of derivative of intensity across the analysis window: *INT-mean*, *INT-st.dev*, *d(INT)-mean* and *d(INT)-st.dev*. The third group consists of features that characterize the quality of speech: *HNR-mean*, *HNR-st.dev*, *d(HNR)-mean*, *d(HNR)-st.dev*, *SHI(apq5)-mean* and *JIT(loc)-mean* (*HNR* is a sign for *Harmonic-to-Noise Ratio, SHI* is a sign for *shimmer*, *JIT* is a sign for *jitter).* Finally, in the time domain, total duration of utterance (*Duration*) was measured. All features were statistically averaged on the utterance level and so, they show general presentation of a sentence as an emotional utterance, and were named *statistical features"*.

Sometimes the Perceptual Linear Prediction features and MFCC are used to train the GMMs. PLP coding is similar to Linear Predictive Coding (LPC) analysis based on the short-term spectrum of speech with the advantage that PLP is more consistent with human hearing. PLP modifies the short-term spectrum of the speech by several psychophysically based transformations.

Sometimes the set of the acoustic parameters is accomplished by the analysis of the physiological state: heart rate, skin conductivity, blood pressure, temperature and breathe speed. It an raise the level of emotion recognition, but is useless for long-distance measures (the phone calls).

The decision rule

The main algorithm of the decision rule is formed only on the basis of the analysis of the parameter's variations extracted from the collected speech spec database in the certain work. On one hand it is important to make the detailed description of the parameter variations. On the other hand the material description shouldn't be too long [3]. The ideal variant of choosing the main parameters for the decision rule is to leave only the uncorrelated ones [15], i.e. to find the "prosodic cues" for every emotion [12]. The selection could be made using the Sequential Forward Floating Selection (SFFS) [14, 20].

In general, the description of emotions using the system of parameters uses a multidimensional approach and the methods of defining such vectors are used (Support Vector Machines).

Constructing the algorithm as trained system (for example, Neural Networks) gives an opportunity to train it on the reference material [8, 14].

Some scientists use linear and non-linear discriminant analysis (LDA, QDA), based on the minimization of the different parameters of the sample utterance [9, 15].

Recognition of the physiological state of the speaker

This problem is very close to the emotion recognition task. Mostly the researchers deal with detecting the affect [14], the degree of irritation [3], detecting the lie in speech [22], the degree of confidence of the speaker in their words [23], and the common division of negative vs. positive attitude of the speaker [11].

The physiologic state detection is important in such spheres as public transport, aviation, medicine. The same approaches of the emotional state recognition could be used in this field.

Conclusion

1. The automatic emotion recognition in speech is a well developed part in the Speech Technology field. There lots of works made in this sphere. They take into account the phonetic features of different languages and language families. The wide spectrum of different methods and approaches is used for the processing of the emotional speech.
2. The main aim of all the works is to improve the systems of parameters and characteristics of different emotions and physiological states and to improve the decision rule of the

systems. Some researchers try to widen the number of parameters to make the more detailed description. Others try to leave only the most important features.

3. There are already lots of applications which use the results of those investigations.

Literature

1. Chuang, Ze-Jing / Wu, Chung-Hsien (2002): "Emotion recognition from textual input using an emotional semantic network", In ICSLP-2002, 2033-2036.
2. Tato, Raquel / Santos, Rocнo / Kompe, Ralf / Pardo, J. M. (2002): "Emotional space improves emotion recognition", In ICSLP-2002, 2029-2032.
3. Ang, Jeremy / Dhillon, Rajdip / Krupski, Ashley / Shriberg, Elizabeth / Stolcke, Andreas (2002): "Prosody-based automatic detection of annoyance and frustration in human-computer dialog", In ICSLP-2002, 2037-2040.
4. Makarova, Veronika / Petrushin, Valery A. (2002): "RUSLANA: a database of Russian emotional utterances", In ICSLP-2002, 2041-2044.
5. Seppanen, Tapio / Vayrynen, Eero / Toivanen, Juhani (2003): "Prosody-based classification of emotions in spoken finnish", In EUROSPEECH-2003, 717-720.
6. Rahurkar, Mandar A. / Hansen, John H.L. (2003): "Frequency distribution based weighted sub-band approach for classification of emotional/stressful content in speech", In EUROSPEECH-2003, 721-724.
7. Liscombe, Jackson / Venditti, Jennifer / Hirschberg, Julia (2003): "Classifying subject ratings of emotional speech using acoustic features", In EUROSPEECH-2003, 725-728.
8. Yacoub, Sherif / Simske, Steve / Lin, Xiaofan / Burns, John (2003): "Recognition of emotions in interactive voice response systems", In EUROSPEECH-2003, 729-732.
9. Kwon, Oh-Wook / Chan, Kwokleung / Hao, Jiucang / Lee, Te-Won (2003): "Emotion recognition by speech signals", In EUROSPEECH-2003, 125-128.
10. Hozjan, Vladimir / Kacic, Zdravko (2003): "Improved emotion recognition with large set of statistical features", In EUROSPEECH-2003, 133-136.
11. Lee, Chul Min / Narayanan, Shrikanth (2003): "Emotion recognition using a data-driven fuzzy inference system", In EUROSPEECH-2003, 157-160.
12. Devillers, Laurence / Vasilescu, Ioana (2003): "Prosodic cues for emotion characterization in real-life spoken dialogs", In EUROSPEECH-2003, 189-192.

13. Blouin, C. / Maffiolo, V. (2005): "A study on the automatic detection and characterization of emotion in a voice service context", In INTERSPEECH-2005, 469-472.

14. Fernandez, Raul / Picard, Rosalind W. (2005): "Classical and novel discriminant features for affect recognition from speech", In INTERSPEECH-2005, 473-476.

15. Cichosz, Jaroslaw / Slot, Krzysztof (2005): "Low-dimensional feature space derivation for emotion recognition", In INTERSPEECH-2005, 477-480.

16. Truong, Khiet P. / Leeuwen, David A. van (2005): "Automatic detection of laughter", In INTERSPEECH-2005, 485-488.

17. Luengo, Iker / Navas, Eva / Hernбez, Inmaculada / Sбnchez, Jon (2005): "Automatic emotion recognition using prosodic parameters", In INTERSPEECH-2005, 493-496.

18. Braun, Angelika / Katerbow, Matthias (2005): "Emotions in dubbed speech: an intercultural approach with respect to F0", In INTERSPEECH-2005, 521-524.

19. Audibert, Nicolas / Aubergй, Vйronique / Rilliard, Albert (2005): "The prosodic dimensions of emotion in speech: the relative weights of parameters", In INTERSPEECH-2005, 525-528.

20. Schuller, Bjцrn / Mьller, Ronald / Lang, Manfred / Rigoll, Gerhard (2005): "Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles", In INTERSPEECH-2005, 805-808.

21. Kim, Jonghwa / Andrй, Elisabeth / Rehm, Matthias / Vogt, Thurid / Wagner, Johannes (2005): "Integrating information from speech and physiological signals to achieve emotional sensitivity", In INTERSPEECH-2005, 809-812.

22. Hirschberg, Julia / Benus, Stefan / Brenier, Jason M. / Enos, Frank / Friedman, Sarah / Gilman, Sarah / Girand, Cynthia / Graciarena, Martin / Kathol, Andreas / Michaelis, Laura / Pellom, Bryan L. / Shriberg, Elizabeth / Stolcke, Andreas (2005): "Distinguishing deceptive from non-deceptive speech", In INTERSPEECH-2005, 1833-1836.

23. Liscombe, Jackson / Hirschberg, Julia / Venditti, Jennifer J. (2005): "Detecting certainness in spoken tutorial dialogues", In INTERSPEECH-2005, 1837-1840.

24. Vidrascu, Laurence / Devillers, Laurence (2005): "Detection of real-life emotions in call centers", In INTERSPEECH-2005, 1841-1844.

25. Liscombe, Jackson / Riccardi, Giuseppe / Hakkani-Тьr, Dilek (2005): "Using context to improve emotion detection in spoken dialog systems", In INTERSPEECH-2005, 1845-1848.

26. Takahashi, Toru / Fujii, Takeshi / Nishi, Masashi / Banno, Hideki / Irino, Toshio / Kawahara, Hideki (2005): "Voice and emotional expression transformation based on statistics of vowel parameters in an emotional speech database", In INTERSPEECH-2005, 1853-1856.

27. Campbell, Nick (2004): "Perception of affect in speech - towards an automatic processing of paralinguistic information in spoken conversation", In INTERSPEECH-2004, 881-884.

28. Chateau, Noel / Maffiolo, Valerie / Blouin, Christophe (2004): "Analysis of emotional speech in voice mail messages: the influence of speakers' gender", In INTERSPEECH-2004, 885-888.

29. Lee, Chul Min / Yildirim, Serdar / Bulut, Murtaza / kazemzadeh, Abe / Busso, Carlos / Deng, Zhigang / Lee, Sungbok / Narayanan, Shrikanth (2004): "Emotion recognition based on phoneme classes", In INTERSPEECH-2004, 889-892.

30. Jiang, Dan-Ning / Cai, Lian-Hong (2004): "Classifying emotion in Chinese speech by decomposing prosodic features", In INTERSPEECH-2004, 1325-1328.

31. Fujisawa, Takashi / Cook, Norman D. (2004): "Identifying emotion in speech prosody using acoustical cues of harmony", In INTERSPEECH-2004, 1333-1336.

32. Tao, Jianhua (2004): "Context based emotion detection from text input", In INTERSPEECH-2004, 1337-1340.

33. Iwai, Atsushi / Yano, Yoshikazu / Okuma, Shigeru (2004): "Complex emotion recognition system for a specific user using SOM based on prosodic features", In INTERSPEECH-2004, 1341-1344.

34. Cho, Hoon-Young / Yao, Kaisheng / Lee, Te-Won (2004): "Emotion verification for emotion detection and unknown emotion rejection", In INTERSPEECH-2004, 1345-1348.

35. Cernak, Milos /Wellekens, Christian (2006):" EMOTIONAL ASPECTS OF INTRINSIC SPEECH VARIABILITIES IN AUTOMATIC SPEECH RECOGNITION", In SPECOM'2006, 405-408.

36. Jovicic, Slobodan/ Rajkovic, Mirjana/ Djordjevic, Miodrag/ Kasic, Zorka (2006):" PERCEPTUAL AND STATISTICAL ANALYSIS OF EMOTIONAL SPEECH IN MAN-COMPUTER COMMUNICATION", In SPECOM'2006, 409-414.

37. Kamath Narsimh (2006): “A PITCH BASED ALGORITHM FOR INDEXING OF HUMOUR IN CONVERSATIONS” , In SPECOM’2006, 415-418.

38. Jang, Kwang-Dong/ Kwon, Oh-Wook (2006): ” SPEECH EMOTION RECOGNITION FOR AFFECTIVE HUMAN-ROBOT INTERACTION”, In SPECOM’2006, 419-422.

39. Petrushin, Valery/ Makarova, Veronika (2006):” PARAMETERS OF FRICATIVES AND AFFRICATES IN RUSSIAN EMOTIONAL SPEECH”, In SPECOM’2006, 423-426.