MediaTeam Speech Corpus: a first large Finnish emotional speech database

Tapio Seppänen, Juhani Toivanen and Eero Väyrynen
MediaTeam, University of Oulu
Finland
E-mail: tapio.seppanen@ee.oulu.fi, juhani.toivanen@ee.oulu.fi, eero.vayrynen@ee.oulu.fi

ABSTRACT

In this paper, a large Finnish emotional speech database is introduced. The database contains simulated emotional speech, reflecting some of the affective states commonly known as basic emotions. The database has been analyzed instrumentally in terms of some 40 acoustic/prosodic parameters, and is currently being annotated in terms of a number of linguistically relevant contrasts. The annotation involves several hierarchical tiers, relating, for example, to the distribution of sentence accents and different types of pauses in the data. Furthermore, listening tests at several levels of abstraction are being carried out to verify that the speech samples contain the intended emotions. Once the analysis is complete, the results can be utilized in content-based information retrieval from audio databases, for example.

1. INTRODUCTION

It is generally known that social communication, especially the expression of emotion, essentially involves exchanges of non-verbal signs. The importance of emotions in general is nowadays acknowledged across scientific disciplines, as a reaction to “Descartes’ error” [1]. The scientific study of the vocal expression of emotion is now reaching a level of maturity where the focus is on important applications, for example, those involving human-computer interaction [2]. In the area of content-based information retrieval, however, the potential of prosodic information in signaling various affective or global speaker-states is still an under-explored subject. The existing search robots cannot utilize acoustic information to locate speech produced with outward manifestations of a specific affective state. According to some researchers, the next step in information retrieval will be search robots capable of analyzing the semantic content of the (audio) data at a substantially deeper level, including an automatic classification of emotion [3].

The study of the vocal correlates of emotion, and in this respect the development of more intelligent search robots, crucially depend on suitable speech material. In some ways, authentic speech data would be ideal but a number of practical problems usually prevent access to a large-scale corpus of such material. Firstly, speech material reflecting authentic emotions is likely involve a real-life situation which is problematic from the viewpoint of acoustic processing of the data: the “uncontrolled” speech situation may cause the speech signal to be too weak or distorted for acoustic analysis. It is also clear that serious problems may arise if strong authentic emotions are induced in a laboratory environment, for example. Emotions have been induced with psychoactive drugs or by means of cognitively demanding tasks. In some studies, even more drastic methods have been used: experimenters have angered the test subjects by behaving in an arrogant and offensive manner [4]. The problem is that one cannot be certain that experimental situations of this kind will consistently produce the same emotional state in every test subject.

An attractive alternative is to utilize existing emotional speech material produced in the context of radio and television programs, chat shows, documentaries, etc. However, the problem is that the researcher may have no way of ascertaining the intended emotion, its temporal/discoursal domain in the flow of conversation and how strongly the emotion was expressed. Furthermore, if such authentic material is accessed, copyright problems are unavoidable. For example, the data of two current large-scale emotional speech database projects, the Reading-Leeds Emotion in Speech Project [5] and the Belfast Project [6], are not publicly available because of copyright restrictions.

A common strategy to collect emotional speech material is to have a number of speakers (for example, professional actors) simulate emotions with unvarying lexical content. Typically, a relatively short speech unit, for example, a sentence, is used as carrier for different emotions. If listener-judges can perceive the intended emotions afterwards, the material can be considered usable. However, the listeners, while being able to perceive the intended emotion, may be quite aware of the differences between simulated and authentic vocal expressions of emotion. The problem may be especially acute if the acted speech, involving a lengthy monologue, for example, is “read”: such speech may produce a powerful portrayal of the emotion, which is, however, really a caricature [2]. An alternative is to collect “semi-natural” speech material: reading passages with appropriate emotional content instead of the same (semantically neutral) carrier phrase may be a scenario which is more valid from the viewpoint of emotional authenticity/ecology.

While we agree that strictly controlled experimental speech situations are not conducive to the expression of authentic emotions, we must bear in mind that, as for content-based retrieval, restricting the research to simulated emotions in the first stage may be a useful strategy. This is because a significant part of the speech material in digital audio databases is, in fact, acted rather than spontaneous (radio plays, film clips, etc.). Actors, who typically produce such material, can be expected to be aware of the display rules of emotional expression: the (vocal) expression of emotion is to some extent culturally determined, and it can be assumed that professional speakers can vocally produce the language- or culture-specific prototype of each emotion. Thus, although the simulation of emotions is likely to produce unusually intense and prototypical expressions, investigating such data may, as a first line of research, be a profitable taking-off point from the viewpoint of content-based information retrieval.

2. MEDIATEAM SPEECH CORPUS: DATA AND ACOUSTIC ANALYSIS

To collect speech material, fourteen professional actors (eight men and six women) were recruited to simulate basic emotions in Finnish speech. The age of the speakers varied between 25 and 50. First each speaker was asked to read out a phonetically rich Finnish passage of some 100 words in a "neutral" or "natural" tone of voice. An attempt was made to find a text which would be semantically as neutral as possible (the text dealt with the nutritional value of the Finnish crowberry). Then the speakers were to read out the text simulating the following emotions: "happiness/joy", "sadness", "fear", "anger", "boredom" and "disgust". The actors were encouraged to take their time to prepare for each emotional state and they could retake the reading (as many times as they wanted) if they were not satisfied with the first version. In addition, the speakers acted out two pre-written dialogues containing emotional lines of varying length (the same six emotions, besides the neutral tone, were to be expressed). All the speech material constituting the MediaTeam Speech Corpus was digitally recorded with DAT in an unechoic studio to produce a 48 kHz, 16-bit recording. The data was stored in a PC as wav format files.

The simulated affective states in the database can be considered to represent basic emotions. The question of what is the most appropriate psychological model of emotion is currently intensively debated in psychology. Traditionally, there are at least two major approaches: discrete and dimensional emotion theories. In accordance with the former view, a number of basic or primary emotions is suggested. Currently, the term "big six" is gaining influence, referring to the (universal) existence of six major emotions. However, there is no final agreement on what emotions fall under this heading although fear, anger, happiness, sadness, surprise and disgust are probably the strongest candidates [7]. Theorists supporting the dimensional approach argue that emotions can be conceptualized in terms of a three-dimensional space involving such attributes as pleasant - unpleasant, active - passive and control - lack of control. This is not the place to argue in favor of any particular theoretical model, but we assume that the emotions simulated in the database represent some of the most common affective states at least in the Western cultural context.

All the stored speech data was analyzed acoustically. The analysis was carried out by means of analysis software implemented in the MATLAB language. The software first distinguished between the voiced and the voiceless parts of the signal after which it determined the F0 contour for the voiced parts of the signal. The F0 contours were determined by using cepstral analysis and interpolating waveform matching. By means of the software, some 40 acoustic/prosodic parameters were automatically computed from the speech signal.

The general F0-based parameters were the following: mean F0, median F0, maximum F0, minimum F0, F0 range, 95th value of F0, 5th value of F0, and 5%-95% range. The parameters relating to the "dynamics" of F0 were: average F0 rise during a continuous voiced segment, average F0 fall during a continuous voiced segment, average steepness of F0 rise, average steepness of F0 fall, maximum F0 rise during a continuous voiced segment, maximum F0 fall during a continuous voiced segment, maximum steepness of F0 rise, and maximum steepness of F0 fall. Additional F0-features were: jitter (trend-corrected mean proportional random F0 perturbation), F0 variation, and a parameter describing the F0 variation bandwidth of voiced segments. Intensity-related parameters were: mean RMS intensity, median RMS intensity, maximum RMS intensity, minimum RMS intensity, intensity range, 5th value of RMS intensity, 95th value of RMS intensity, and 5%-95% range. Additional intensity-related features were: shimmer (trend-corrected mean proportional random intensity perturbation), and a parameter describing the width variation of the intensity distribution of voiced segments. Temporal features were: average duration of voiced segments, average duration of voiceless segments shorter than 500 ms, average duration of silent segments shorter than 400 ms, average duration of voiceless segments longer than 500 ms, average duration of silent segments longer than 400 ms, maximum duration of voiced segments, maximum duration of voiceless segments, and maximum duration of silent segments. Parameters indicating different ratios were: silence-to speech ratio and voicing-to-pauses ratio. Finally, spectral features were: proportion of low-frequency energy below 500 Hz, and proportion of low-frequency energy below 1,000 Hz.

It should be pointed out that, above, the term "segment" refers to a part of the signal of varying duration, which may be realized as silence or as voiced or voiceless speech. Thus the term does not describe any linguistic/phonological unit. All the parameters listed above are computed fully automatically from the speech signal, without any manual intervention. The parameters can thus be measured completely independently of the linguistic/phonological structure of the speech data.

3. ANNOTATION OF THE DATA

The acoustic analysis of the data described above reveals something about the global prosodic features of the speech material but it does not directly describe intonation in any systematic way; it can be argued that the parameters represent the paralinguistic correlates of emotion at the signal level of speech communication. For example, the parameters relating to the "dynamics of F0" do not describe features of "speech melody" at the utterance level. Similarly, the parameters indicating the "steepness" of F0 movements are connected with general average features of F0 in the whole speech situation, instead of being directly connected with "nuclear" F0 contours or tone types. A related point is that sentence accentuation (focus structure) cannot be inferred from global intensity-related or spectral parameters. It can be assumed that the distribution of tone types vis-à-vis sentence modality and focus structure, as well as signaling the informational/syntactic structure, reflect also the emotional content of the speech. It is clear that, in order to describe the data in terms of linguistically relevant variation in F0, a different kind of analysis is necessary. The aim is to find out how the functionally important parameters that cannot be directly measured from the acoustic signal are related to different emotions. The speech data is currently being analyzed in terms of a number of linguistically relevant prosodic contrasts; in this respect, the analysis is carried out at the symbolic level. Below, the major principles of the linguistic analysis are outlined.

The analysis is carried out with the Praat program, utilizing its features of labeling and segmentation of speech [8]. The general idea is to create interlinear transcription, where each word is annotated with phonetic information of many kinds. The analysis will involve several hierarchically organized tiers, with multiple transcription windows time-linked to the sound recording and the text.

For intonation, the ToBI framework is used: ToBI is a multi-tiered general framework for developing generally agreed-upon conventions for transcribing the intonation of spoken language [9]. The ToBI notation is used to transcribe pitch direction and the pitch of accented syllables. In the annotation used for the present investigation, we follow Välimaa-Blum's [10] suggestion in that two complex accents, L+H* and L*+H, two boundary tones, H% and L%, and two initial boundary tones, %H and %L, are recognized. In addition to the prosodic transcription based on the ToBi system, a more traditional "contour" model is used, distinguishing between the "non-emphatic pattern" (basically a falling contour), the "expressive pattern" (basically a rising contour) and the "progredient pattern" (intonation terminating at mid pitch). This inventory of global pitch patterns has been proposed for Finnish [11]. With the contour model analysis, additional attributes such as high/low pitch and wide/narrow tone are possible.

Each word is annotated in terms of accentuation/prominence applying the descriptive system used by Suomi et al. [12]. Each word is (auditorily) analyzed as "unaccented", "moderately accented" or "strongly accented".

Pauses with a minimum duration of 100 ms are annotated according to the following principles. At the first level of description, pauses are sentence-internal or sentence-final. Both types of pauses may be filled or unfilled. Filled pauses are classified as "vocalizations" (e.g. *uh*, *um*) or "vocal noises" (inhalations, exhalations, laughter, sobbing, etc.); the latter category basically represents "respiratory reflexes" and "voice qualifications" as defined by Crystal and Quirk [13]. At the second level of description, the (filled or unfilled) pauses are "syntactically motivated" or "syntactically unmotivated": syntactically motivated pauses occur at syntactic junctures (e.g. between clauses and after long noun phrases). For each pause, the duration and type is described.

Auditory features of voice quality are included in the annotation. For each clause/sentence or pause-defined unit, a categorization of voice quality is chosen. The descriptors include "modal voice", "falsetto", "creak", "whisper", "tense" and "rough"; these labels are basically those suggested by Laver [14] for the description of different phonation types. Features of rhythm are also annotated. While the absolute speed of speech can be measured in terms of phonemes/syllables per second, utilizing automatic speech recognition, for example, more complex qualities are difficult to quantify. Following Roach et al. [5], we use the following labels, in addition to "neutral", to describe the rhythm of speech: "fast", "slow", "accelerating", "decelerating", "clipped", "drawled", "precise", and "slurred".

All in all, the annotation involves multilevel transcriptions. The first layer is used to transcribe the speech material orthographically (although all the speakers read the same text, additions, elisions and mispronunciations of various kinds were common). The second layer is used to code tones in terms of the ToBI model, and the third layer complements this analysis by describing the pitch pattern in terms of contours. The fourth layer represents lexical and sentence stress, distinguishing between three degrees of accentuation. The fifth layer describes the duration and type of each pause in the data. The sixth layer is used to code features of voice quality, and the seventh layer is reserved for features of the rhythm of speech. It must be pointed out that, in the annotation, especially features of voice quality and rhythm will be difficult to quantify or define physically. The perceptual qualities cannot be automatically measured or detected from speech although there are obviously some correlation between voice qualities and the spectral features of speech, for example.

The data must also being annotated "psychologically": listening tests are carried out to chart the extent to which listener-judges can hear the intended emotions. The psychological annotation also includes several layers. Firstly, each monologue is used as an emotional speech sample on its own, and, perceptually, discrimination (deciding between alternatives) is studied. Initially, the options are limited to those emotional states which were actually expressed; in later listening tests, distracters are used. Secondly, recognition, as opposed to discrimination, is investigated: are listener-judges able to recognize a particular category in its own right? Thirdly, the speech samples are divided into smaller units. As each text is divided into units of two sentences: there will be 490 short emotional speech samples to be evaluated in listening tests (with discrimination or recognition scenarios). As the listener-judges will hear the short samples in a random order, the stability of the vocal expression of emotion for each long monologue can be tested. Listening tests with designs described above are currently under way.

4. CONCLUDING REMARKS

The MediaTeam Speech Corpus is currently the largest Finnish database of emotional speech, containing linguistic units with specific emotional content which range from one-word exclamations to monologues of nearly one minute in duration. In the vocal emotion literature, the focus has so far been on such major languages as English, German and French, and very little is known about the vocal correlates of emotion in continuous spoken Finnish. In this respect, the investigation outlined here clearly covers new ground. The database, once fully annotated at several tiers, will be a valuable tool in basic research, enabling a theoretically coherent multi-parametric description of vocal correlates of emotions in Finnish. On an applied side, the algorithms developed for prosodic parameter estimation, the statistical classification of emotional speech, and the estimation of the degree of emotions can open up new opportunities for future Internet database technologies and methods for content-based information retrieval. To develop “affective computing” for Finnish in the form of algorithms that are able to understand (and respond to) human emotions, language-specific information on the vocal correlates of emotions is necessary. It is our hope that the database can serve these research purposes.

ACKNOWLEDGEMENT

This study was financially supported by Academy of Finland (project 52268: “Prosodic Information in Content-based Information Retrieval”).

REFERENCES

[1] A. Damasio, *Descartes’ Error*, New York: Grosset/Putnam, 1994.

[2] E. Douglas-Cowie, N. Campbell, R. Cowie and P. Roach, “Emotional speech: Towards a new generation of databases,” *Speech Communication*, vol. 40, pp. 33-60, 2003.

[3] F. Yu, E. Chang, Y.-Q. Xu and H.-Y. Shum, “Emotion detection from speech to enrich multimedia content,” *Proceedings of the Second IEEE Pacific Rim Conference on Multimedia*, pp. 550-557, Peking, 2001.

[4] G. Stemmler, M. Heldmann, C.A. Pauls and T. Scherer, “Constraints for emotion specificity in fear and anger: The context counts,” *Psychophysiology*, vol. 38, pp. 275-291, 2001.

[5] P. Roach, R. Stibbard, J. Osborne, S. Arnfield and J. Setter, “Transcription of prosodic and paralinguistic features of emotional speech,” *Journal of the International Phonetic Association*, vol. 28, pp. 83-94, 1998.

[6] E. Douglas-Cowie, R. Cowie and M. Schroeder, “A new emotion database: considerations, sources and scope,” *Proceedings of the ISCA ITRW on Speech and Emotion*, pp. 39-44, Belfast, 2000.

[7] R.R. Cornelius, *The Science of Emotion. Research and Tradition in the Psychology of Emotion*, Upper Saddle River, New Jersey: Prentice-Hall, 1996.

[8] P. Boersma and D. Weenink, “Praat, a System for doing Phonetics by Computer,” *Institute of Phonetic Sciences of the University of Amsterdam, Report 132*, 1996.

[9] M.E. Beckman and G.M. Ayers, *Guidelines for ToBI labeling*, Columbus: Ohio State University, Department of Linguistics, 1994.

[10] R. Välimaa-Blum, “A Pitch Accent Analysis of Intonation in Finnish,” *Ural-Altaische Jahrbucher*, vol. 12, pp. 82-94, 1993.

[11] A. Iivonen, “Intonation in Finnish,” in *Intonation Systems: A Survey of Twenty Languages*, D. Hirst and A. Di Cristo, Eds., pp. 311-327. Cambridge: Cambridge University Press, 1998.

[12] K. Suomi, J. Toivanen and R. Ylitalo, “Durational and tonal correlates of stress in Finnish,” *Journal of Phonetics*, in press.

[13] D. Crystal and R. Quirk, *Systems of Prosodic and Paralinguistic Features in English*, The Hague: Mouton, 1964.

[14] J. Laver, *Principles of Phonetics*, Cambridge: Cambridge University Press, 1994.