A Database of German Emotional Speech

F. Burkhardt[1], A. Paeschke[2], M. Rolfes[3], W. Sendlmeier[2], B. Weiss[4]

[1]T-Systems, [2]TU Berlin, Department of Communication Science, [3]LKA Berlin, [4]HU Berlin
Astrid.Paeschke@TU-Berlin.de

Abstract

The article describes a database of emotional speech. Ten actors (5 female and 5 male) simulated the emotions, producing 10 German utterances (5 short and 5 longer sentences) which could be used in everyday communication and are interpretable in all applied emotions.

The recordings were taken in an anechoic chamber with high-quality recording equipment. In addition to the sound electro-glottograms were recorded. The speech material comprises about 800 sentences (seven emotions * ten actors * ten sentences + some second versions).

The complete database was evaluated in a perception test regarding the recognisability of emotions and their naturalness. Utterances recognised better than 80% and judged as natural by more than 60% of the listeners were phonetically labelled in a narrow transcription with special markers for voice-quality, phonatory and articulatory settings and articulatory features.
The database can be accessed by the public via the internet (http://www.expressive-speech.net/emodb/).

1. Introduction

Although having been studied since the 1950's, the investigation of emotional cues in speech is gaining growing attention. This is mainly due to the new developments with respect to human-machine interfaces that see applications of automatic recognition and simulation of emotional expression within reach.

The article describes the planning and accomplishment of a German database of acted emotional speech, containing ten sentences performed in 6 target emotions by ten actors. This database has been the basis for analyses of prosodic features [1], articulatory features [2] and the verification by means of resynthesis [3].

The first part argues why the material was recorded in acted instead of "real life" situations. The second part explains which target emotions were chosen. The next chapter elaborates on the choice of the text material, followed by three sections about the recording, evaluation and labelling of the data. The last part explains a web front-end that can be used to search the database and view prosodic features online.

2. Acted or "Real" Emotions?

We are quite aware that there are many arguments in disfavour of acted emotional expression. As [4] point out, so-called full-blown emotions very rarely appear in the real world. Furthermore there are physical emotional cues that cannot be consciously mimicked. However, as clear emotional expression is not only rare in everyday situations but also the recording of people experiencing full-blown emotions is ethically problematic, it is almost impossible to use natural data if basic emotions are the subject of investigation.

Even more grave, for a clean experimental setup everything except the issue under study should be kept constant. For our needs the following points had to be fulfilled:

- A reasonable number of speakers should perform all emotions to offer generalization over the target group.
- All speakers should utter the same verbal content in order to allow the comparability across emotions and speakers.
- Recordings should be of high audio quality, minimizing background noise. Otherwise spectral measurements would not have been possible.
- Having inverse filtering in mind, an anechoic chamber and laryngographic readings are mandatory.

One way-out of this dilemma for controlled experiments in literature lies in embedding the target sentences in short plays that should arouse the target emotion. We disregarded this approach in favour of simple verbal instructions, as it would have made the whole setup far more complex. Also we would have been faced with the problem that everybody reacts differently with respect to emotional situations. We relied on the performers' ability of self-induction by remembering a situation when the desired emotion had been felt strongly, which is known as the Stanislavski method.

3. Choice of Emotions

In the literature emotions are usually described either by use of "emotion-dimensions" such as "activation", "pleasure" and "dominance" or as discrete concepts such as "anger" or "fear". Distinct terms are the logical choice if one investigates acted emotions, as they are easily understood by the performer as well as by the listener.

In order to be able to compare the results with older studies of our research group [5, 6, 7], the same emotional labels were used, notably (German terms in brackets) ***neutral*** (neutral), ***anger*** (Ärger), ***fear*** (Angst), ***joy*** (Freude), ***sadness*** (Trauer), ***disgust*** (Ekel) and ***boredom*** (Langeweile).

4. Choosing the actors

Bearing in mind that actors learn to express emotions in quite an exaggerated way, we were not convinced that trained actors would be the best choice to perform natural emotional expressions. Therefore it was decided to leave that matter open and search for performers by means of a newspaper advertisement. About 40 people answered and were invited to a preselection session. They had to perform one utterance in each of the target emotions that was recorded in an office directly with a microphone to hard disk. From these 40

sessions, three expert listeners selected 10 people, equally representing the sexes, by judging the naturalness and recognisability of the performance. Interestingly, all but one of the chosen ones had indeed passed an acting schooling.

5. Text material

Setting up the database with speech data in which actors simulate emotions has the advantage that it is possible to control the individual sentences to be spoken. It is important, though, that all these sentences should be interpretable in the emotions under review and that they contain no emotional bias. 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. [8]).
- Normal sentences which could be used in everyday life.

Nonsense material is guaranteed to be emotionally neutral. However, there is the disadvantage that actors will find it difficult to imagine an emotional situation and to produce natural emotional speech spontaneously. According to [9] 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 [9], 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. In the construction of the database, priority was given to the naturalness of speech material and thus everyday sentences were used as test utterances. A total of ten sentences, five consisting of one phrase and five composed of two phrases were 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.

One of the aims of developing the database was to facilitate analyses of articulatory reduction. Thus, the phonotactic construction of the test sentences had to contain the possibility of different reduction forms. The test sentences were constructed so that they allow all possible deletions and assimilations of segments, according to [10], spread among the 10 sentences. To carry out formant analyses the test sentences had to contain as many vowels as possible.

The following sentences were used:

a01: Der Lappen liegt auf dem Eisschrank. (The cloth is lying on the fridge.)
a02: Das will sie am Mittwoch abgeben. (She will hand it in on Wednesday.)
a04: Heute Abend könnte ich es ihm sagen. (Tonight I could tell him.)
a05: Das schwarze Stück Papier befindet sich da oben neben dem Holzstück. (The black sheet of paper is up there beside the piece of timber.)
a07: In sieben Stunden wird es soweit sein. (In seven hours the time will have come.)
b01: Was sind denn das für Tüten, die da unter dem Tisch stehen? (What are the bags standing there under the table?)
b02: Sie haben es gerade hochgetragen und jetzt gehen sie wieder runter. (They have just carried it upstairs and now they are going down again.)
b03: An den Wochenenden bin ich jetzt immer nach Hause gefahren und habe Agnes besucht. (At the weekends I have always gone home now and seen Agnes.)
b09: Ich will das eben wegbringen und dann mit Karl was trinken gehen. (I just want to take this away and then go for a drink with Karl.)
b10: Die wird auf dem Platz sein, wo wir sie immer hinlegen. (It will be in the place where we always put it.)

6. Recording the data

To achieve a high audio quality the recordings took place in the anechoic chamber of the Technical University Berlin, Technical Acoustics Department using a Sennheiser MKH 40 P 48 microphone and a Tascam DA-P1 portable DAT recorder. In addition to the pure audio data electro-glotto-grams were recorded using the portable laryngograph (Laryngograph Ltd.). Recordings were taken with a sampling frequency of 48 kHz and later downsampled to 16 kHz.

The actors were standing in front of the microphone so they could use body language if desired, only hindered by the cable of the laryngograph and the need to speak in the direction of the microphone with a distance of about 30 cm (see figure 1).

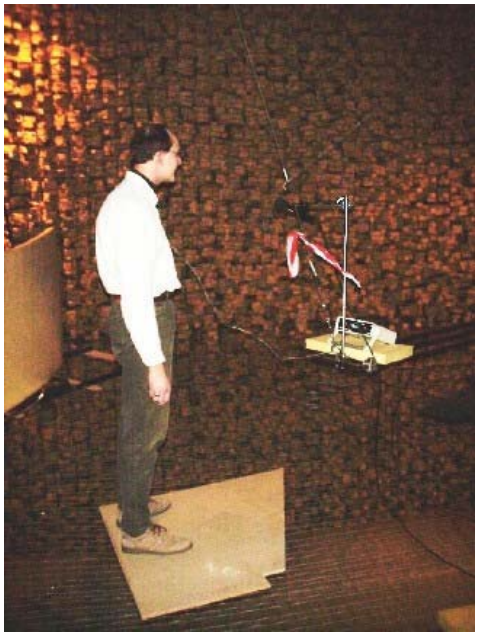

Figure 1: One of the actors during the recordings in the anechoic chamber

There was a single recording session with every actor under supervision of three phoneticians, two of them giving instructions and feedback, one monitoring the recording equipment. Each session lasted about two hours. The text of every utterance was prompted to the actor to avoid a reading intonation style. After that they could choose emotions one after another. They heard a short characterization of this emotion (e.g. happiness after winning a large amount of money in the lottery or sadness caused by losing a very good friend or relative) and got time to put themselves into this

specific emotion. The actors were asked to remember a real situation from their past when they had felt this emotion. By this way we got recordings from actors who re-experienced the emotions and at most favourable terms developed the same physiological effects as in the real situation.

Actors produced each of the sentences as often as they liked. For some combinations we therefore recorded several variants. The actors were instructed not to shout to express anger and to avoid whispering while expressing anxiety. This was necessary in order to get data still analysable regarding voice quality. Attention was paid to a casual pronunciation. A speaking style like on stage should be avoided.

There still remain at least three problems: Because the actors were not only standing but moving in front of the loose-hanging microphone, the distance between mouth and microphone was not constant and that is why the analysis of signal energy would be unreliable. Furthermore the recording level had to be adjusted between very loud speech (mostly anger) and very quiet speech (mostly sadness). Another problem applies to the intonation contour: Actors chose different words for realizing the sentence accent which makes a comparison of fundamental frequency contours more complicated.

7. Evaluating the data

To ensure the emotional quality and naturalness of the utterances a perception test was carried out. 20 subjects took part in this test. They were presented with the utterances in random order in front of a computer monitor. They were allowed to listen to each sample only once before they had to decide in which emotional state the speaker had been and how convincing the performance was .

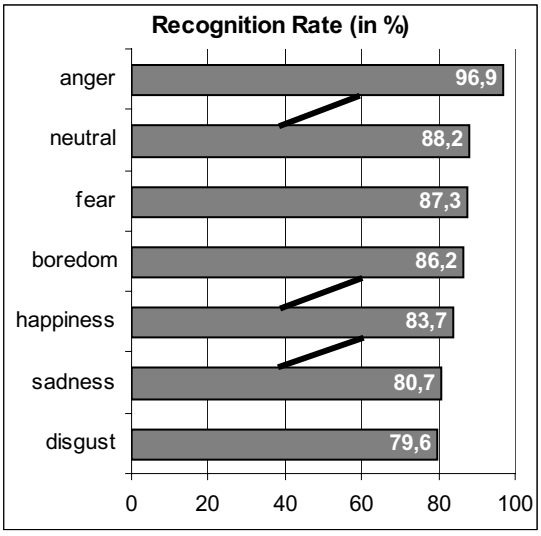

Figure 2: Recognition rates and significant differences between emotions

Mean recognition rates are shown in figure 2. Connecting lines between the bars show significant differences between emotions ($p < 0.05$). Utterances with a recognition rate better than 80% and naturalness better than 60% were chosen for further analysis. On the whole, about 500 utterances out of 800 were left out.

Two more perception tests were carried out: In one test subjects were asked to rate the strength of the displayed emotion for each utterance, in the other subjects had to judge the syllable stress of every utterance. In both tests subjects had the possibility to listen to the utterances as often as they liked before giving their rating. Results were that most emotions are moderate up to strong. Emotional strength was then used as a control variable in statistical analyses. Evaluating the syllable stress was necessary for the analysis of stress because objective measurements of stress are not available. This test was the only one in which only phonetically trained subjects took part because other people felt not to be able to rate the syllable stress.

8. Labeling the data

Utterances were annotated using ESPS/waves+. Two label files in ASCII format were created for every utterance. The first label file contains a narrow phonetic transcription that is based on auditive judgement supported by visual analysis of oscillogram and spectrogram (see figure 3 for an example including the electro-glottogram). For the transcription the SAMPA phonetic alphabet was used. Emotional characteristics of voice and manner of speaking were labelled with additional characterisations, namely annotations of articulatory settings like *harsh voice* or *whispery voice* [11, 12]. While phonemic segments were labelled with SAMPA symbols, settings and diacritics were marked with German abbreviations (e.g. “nas” for nasal).

Segment and pause boundaries were labelled, too. Each single sound was transcribed with one symbol, except for diphthongs and plosives. Diphthongs were treated as one segment. Additional symbols were assigned to burst and aspiration phases of plosives. Diacritics were not associated with regular segment boundaries. The exact time of occurrence and disappearance was labelled with “+ / -” (e.g. “+nas” until “-nas” for a nasalised part). The used IPA diacritics are the following: *voiceless*, *voiced*, *aspirated*, *rounded*, *centralized*, *syllabic*, *breathy voice*, *creaky voice*, *labialised*, *palatalised*, *velarised*, *pharyngealised*, *raised*, *lowered*, *dental*, *apical*, *laminal*, *nasalised*, *nasal release*, *no audible release*, *advanced*, *retracted*. In addition, the suprasegmental *long* is used as well as *labial spreading* and *denasal* (from ExtIpa), *lateral*, *fricative* and the articulatory settings *harsh voice*, *falsetto*, *whispery voice*, *faucalized voice*, *shouted* and *laughing*.

Plosives starting at the beginning of an utterance or right after a pause were especially marked because the exact starting time of these segments is, of course, not determinable.

The second label file contains a segmentation into syllables and markings of four different levels of stress (sentence stress, primary, secondary stress, unstressed). These levels were verified in a perception test of eight trained phoneticians to make the data more reliable. The segmentation of the syllables is based on the boundaries of the narrow phonetic transcription in the first label file. Solely in the case of ambisyllabic segments the syllable boundary is set within a sound.

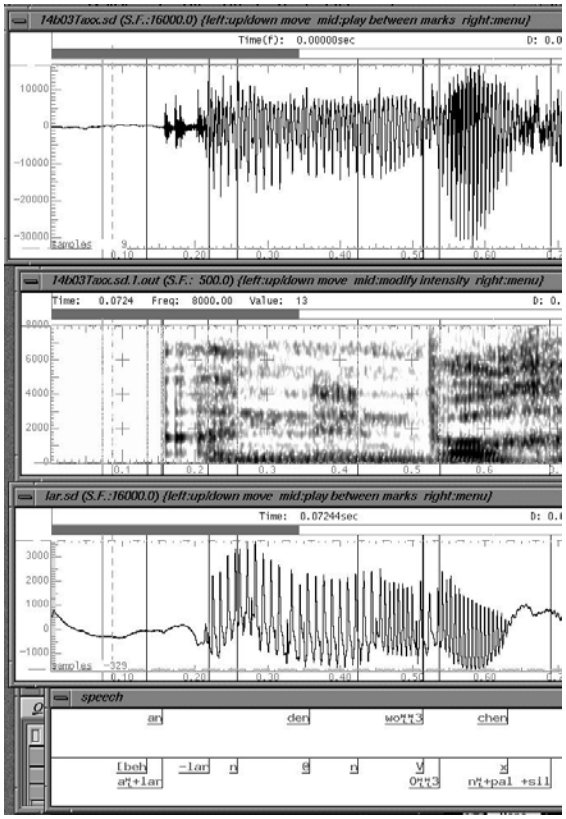

Figure 3: Top to bottom: screenshot of the signals: oscillogram, spectrogram, electro-glottogram and two label files

9. Presenting the data

A web interface was developed to present the database of emotional speech. All the available information of the speech database can be accessed via the internet: http://www.expressive-speech.net/emoDB/.

One can select utterances from the database and listen to them. The selection can be done according to speaker, spoken text and displayed emotion. Along with the syllable labels and duration information intonation contours, global trends, histograms of fundamental frequency, energy and loudness curves as well as rhythm events (calculated by Zwickers loudness based rhythm model [13]) can be displayed. Results of the evaluation tests are also displayed. As a special feature one can listen to and download different re-synthesised (MBROLA) versions of the original utterance.

The web page also provides a number of analysis results, mainly statistics concerning duration and stress as well as fundamental frequency histograms.

The recorded and labelled speech samples can be downloaded as a package from there.

10. Summary

We recorded an emotional database with the "big four" basic emotions plus boredom and disgust. The emotional utterances were performed by German actors. The material was evaluated in an automated listening test and each utterance was judged by 20 listeners with respect to recognisability and naturalness of the displayed emotion. The database has already served as a basis for numerous studies and is freely available from the internet.

11. Acknowledgements

The project was financially supported by the Deutsche Forschungsgemeinschaft DFG (German Research Community), grant nr. SE 462/3-1 to W. Sendlmeier.

12. References

[1] Paeschke, A., „Prosodische Analyse emotionaler Sprechweise“, *Reihe Mündliche Kommunikation*, Band 1, Logos Berlin, 2003

[2] Kienast, M., „Phonetische Veränderungen in emotionaler Sprechweise“, Shaker, Aachen, 2002

[3] Burkhardt, F., „Simulation emotionaler Sprechweise mit Sprachsyntheseverfahren“. *Reihe Berichte aus der Kommunikationstechnik*, Shaker, Aachen, 2001

[4] Douglas-Cowie, E., Cowie, R., Schröder, M., “A New Emotion Database: Considerations, Sources and Scope”, *Proceedings of the ISCA Workshop on Speech and Emotion*, Newcastle, 2000

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

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

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

[8] 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

[9] 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

[10] Kohler, K. J., “Articulatory Reduction in Different Speaking Styles”, *Proceedings ICPhS ‘95*, Stockholm, Vol. 2, p. 12-19, 1995

[11] Laver, J., “The Phonetic Description of Voice Quality”, Cambridge University Press, Cambridge, 1980

[12] Laver, J., “The Gift of Speech”, University Press, Edinburgh, 1991

[13] Zwicker, E., Fastl, H., “Psychoacoustics“, p. 245f., Springer-Verlag, Berlin, 1990

[14] Burkhardt, F., Sendlmeier, W., “Verification of Acoustical Correlates of Emotional Speech Using Formant Synthesis”, *Proceedings of the ISCA Workshop on Speech and Emotion*, Newcastle, 2000

[15] Sendlmeier, W., „Phonetische Variation als Funktion unterschiedlicher Sprechstile“, *Elektronische Sprach-signalverarbeitung*, w.e.b. Dresden, p. 23-25, 2001

[16] Sendlmeier, W., „Stimmliche und phonetische Manifestation emotionaler Sprechweise“, in: H. Geißner (ed.) *Stimmkulturen*, Röhrig Universitätsverlag, St. Ingbert, p. 39-49, 2002