EmoVoice — A framework for online recognition of emotions from voice

Thurid Vogt, Elisabeth André, Nikolaus Bee

Multimedia Concepts and their Applications, University of Augsburg, Germany
{vogt,andre,bee}@informatik.uni-augsburg.de

Abstract. We present EmoVoice, a framework for emotional speech corpus and classifier creation and for offline as well as real-time online speech emotion recognition. The framework is intended to be used by non-experts and therefore comes with an interface to create an own personal or application specific emotion recogniser. Furthermore, we describe some applications and prototypes that already use our framework to track online emotional user states from voice information.

1 Introduction

Research on the automatic recognition of emotions in speech has emerged in the last decade and has since then shifted from purely acted to more natural emotions. So far, most of this research has been concerned with the offline analysis of available or specifically created speech corpora. However, most applications that could make use of affective information would require an online analysis of the user's emotional state. Therefore, the consideration of real-time constraints is also important for emotion recognition. Of course, recognition rates must be expected to be lower than for offline analysis, and tasks should be limited to very few emotional states. In this paper, we present EmoVoice, our framework to building an emotion classifier and to recognising emotions online.

Among possible application scenarios for online speech emotion recognition are call center conversations, by e. g. providing call center employees with information regarding the emotions their voice might portray, or by automatically switching from computer to human operators if a caller exhibits high arousal in his voice, an indication for a problem [Burkhardt et al., 2005b]. Further examples of application scenarios include computer-enhanced learning [Ai et al., 2006] or emotionally aware in-car systems [Schuller et al., 2007a].

As already mentioned, so far only few approaches to online emotion recognition exist. One example is the Jerk-O-Meter that monitors attention (activity and stress) in a phone conversation, based on speech feature analysis, and gives the user feedback allowing her to change her manners if deemed appropriate [Madan, 2005]. Jones and colleagues have explored online emotion detection in games [Jones and Deeming, 2007] and for giving feedback in human-robot interaction[Jones and Sutherland, 2007]. Our own applications will be presented later in section 4.

The rest of this paper is organised as follows: first, in section 2, we give an overview of the modules of EmoVoice, and explain the individual components of the recognition process in detail. Section 3 explains our data acquisition method, while section 4 describes applications and prototypes that already make use of our framework. Finally, we draw some conclusions and indicate our future steps.

2 Components of EmoVoice

The major steps in speech emotion recognition are *audio segmentation*, which means finding appropriate acoustic segments as emotion classification units, *feature extraction* to find those characteristics of the acoustic signal that best describe emotions and to represent each segmented acoustic unit as a (series of) feature vector(s), and lastly the actual *classification* of the feature vectors into emotional states.

EmoVoice, our framework for emotion recognition from speech (see Fig. 1), consists of two modules, one for the offline creation and analysis of an emotional speech corpus, and one for the online tracking of affect in voice while someone is talking. The first module is a set of tools for audio segmentation, feature extraction, feature selection and classification of an emotional speech corpus, and a graphical user interface to easily record speech files and create a classifier. This classifier can then be used for the second module, the online emotion recognition. Here, classification results are obtained continuously during talking, there is no "push-to-talk".

Primarily, the online emotion recognition just outputs the recognised emotions name, so just provides the functionality of emotion recognition. However, it can easily be plugged into other applications visualising or making use of the affective information which is the topic of section 4.

Now we will describe how audio segmentation, feature extraction and classification are addressed in EmoVoice, both in offline and online recognition.

2.1 Audio segmentation

The first step in online emotion recognition is to segment the incoming speech signal into meaningful units that can serve as classification units. Commonly, these are linguistically motivated units such as words or utterances. Though the decision on which kind of unit to take is evidently important, it has not received much attention in past research on emotion recognition. Most approaches so far have dealt with the offline classification of utterances of acted emotions where the choice of unit is obviously just this utterance, a well-defined linguistic unit with no change of emotion within in this case.

However, in spontaneous speech this kind of obvious unit does not exist. Neither is the segmentation into utterances straight-forward nor can a constant emotion be expected over an utterance. On the other hand, a good unit should be long enough so that features can reliably be calculated by means of statistical functions. Words are often too short for this. Therefore, a suitable unit for spontaneous speech can e. g. be found at the phrase level.

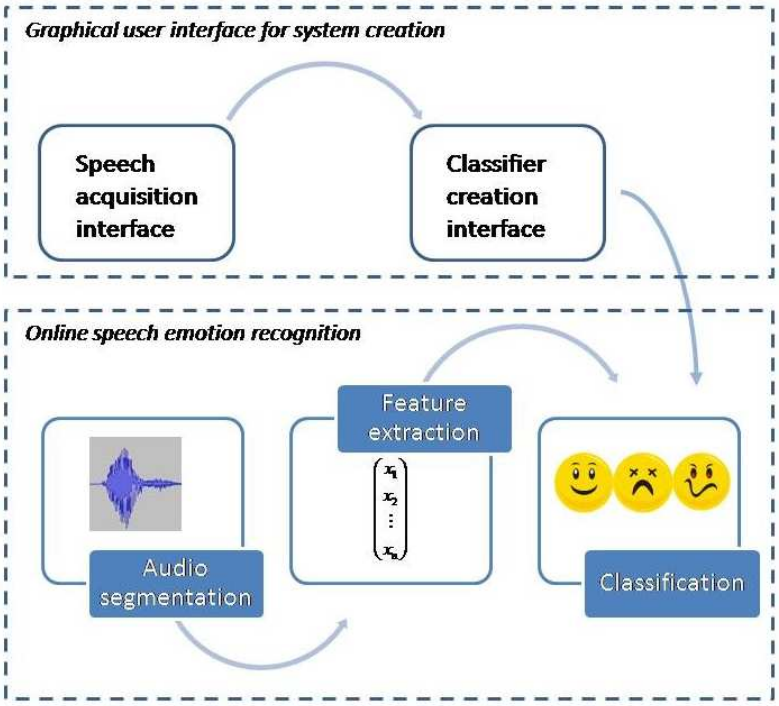

Fig. 1. Overview of the two modules of EmoVoice: the graphical user interface allows for convenient speech corpus and classifier creation. The online recognition module can be plugged into other applications and continuously tracks the emotional state of the user as expressed in his/her voice.

The task is further constrained when dealing with online recognition, as for a word/utterance/phrase segmentation, an automatic speech recogniser (ASR) is required. Though the often faulty ASR output does not necessarily degrade the performance of the emotion recogniser [Schuller et al., 2007b], it is at least time-consuming. For this reason, we use *voice activity detection* to segment by pauses into signal chunks of voice activity without pauses longer than 200 ms within. This method is very fast and comes close to a segmentation into phrases though it does not make use of any linguistic knowledge. Still, no change of emotion can be expected to occur within such a chunk. The algorithm used comes from the Esmeralda environment for speech recognition [Fink, 1999].

For spontaneous speech, voice activity detection yields a very favorable segmentation. When reading speakers usually do not make long enough pauses, even between text sections that differ in terms of content, and emotions. However, in EmoVoice, there is also the option to set a maximum interval for the output

of a classification result if no pause has occurred before. For this interval, 2–3 seconds turned out to be a suitable duration.

2.2 Feature extraction

The goal of the feature extraction is to find those properties of the acoustic signal that best characterise emotions. Common features for speech emotion recognition are based on short-term acoustic observations like pitch or signal energy. Since the specific values of these measures are usually not too expressive *per se*, but rather their change over time, the modeling of the temporal behavior is crucial to the success of the task. Basically, there are two approaches to do this, which depend on the type of classifier that is used. Learning algorithms like HMMs model temporal changes by considering sequences of feature vectors, looking especially at the transitions between the vectors. Thus, a classification unit consists of a series of feature vectors obtaining one label by the classifier. Standard classifiers, however, assign one label to each feature vector. As a result, time needs to be encoded in the features themselves, usually by (optional) transformations of the basic values and applying (statistical) functions like mean calculation, that map a series of values onto a single value. The latter approach is the one followed here.

Since an optimal feature set for speech emotion recognition is not yet established, we calculate, starting from basic acoustic observations, a multitude of statistical values for each measure. This is similar to our earlier work on feature extraction in [Vogt and André, 2005,Vogt and André, 2006]. Of course, because of online processing, we use only fully automatically in real-time extractable features which is opposed to most other approaches to speech emotion recognition that rely to some extent on manually annotated information. Our basic observations are logarithmised pitch, signal energy, Mel-frequency cepstral coefficients (MFCCs; 12 coefficients), the short-term frequency spectrum, and the harmonics-to-noise ratio (HNR). The resulting series of values are transformed to different views, and for each of the resulting series mean, maximum, minimum, range, variance, median, first quartile, third quartile and interquartile range are derived (based on [Oudeyer, 2003]). These values constitute the actual features used. The transformations into different views comprise the following:

- *logarithmised pitch:* the series of the local maxima, local minima, the difference, slope, distance between local extrema, the first and second derivation, and of course the basic series;
- *energy:* the basic series and the series of the local maxima, local minima, the difference, slope, distance between local extrema, first and second derivation as well as the series of their local maxima and local minima;
- *MFCCs:* the basic, local maxima, local minima for basic, first and second derivation for each of 12 coefficients alone;
- *frequency spectrum:* the series of the center of gravity, the distance between the 10 and 90 % frequency quantile, the slope between the strongest and the weakest frequency, the linear regression;

– *HNR:* only the basic series.

Additionally, four duration related features are used: segment length in seconds, pause as the proportion of unvoiced frames in a segment obtained from pitch calculation and as the number of voiceless frames in a segment obtained from voice activity detection and the zero-crossings rate. Duration, or speaking rate, is also encoded in the distance between local energy extrema. For pitch, also the positions of the global maximum and minimum in the segment, and the number of local maxima and minima as well as the number of falling resp. rising pitch frames are added as features. For energy, additional features include the position of the global maximum and the number of local maxima. Furthermore, we use jitter, shimmer and the number of glottal pulses of the analysed speech features as voice quality features in addition to HNR.

Our pitch and voice quality calculation are based on the Praat phonetics software [Boersma and Weenink, 2007], energy and MFCC calculation come from the Esmeralda speech recognition environment [Fink, 1999]

Overall, we thus have a feature vector containing 1302 features. Of course, this is a large number of features for fast classification, and it is very likely that some of the features in the set are redundant. We optionally employ a correlation-based feature subset selection [Hall, 1998] from the data-mining software Weka [Witten and Frank, 2005] to reduce our feature set to only uncorrelated features with respect to a specific training audio corpus. This usually means a reduction to 50–200 features, which is a tractable number of features for the classification algorithms.

2.3 Classification

Currently, two classification algorithms are integrated in EmoVoice: a naïve Bayes (NB) classifier and a support vector machine (SVM) classifier (from the LibSVM library [Chang and Lin, 2001]). The NB classifier is very fast, even for high-dimensional feature vectors, and therefore especially suitable for real-time processing. However, it yields slightly lower classification rates than the SVM classifier which is a very common algorithm used in offline emotion recognition. In combination with feature selection and thereby a reduction of the number of features to less than 100, SVM is also feasible in real-time.

3 A training procedure for non-experts

Statistical classifiers, as used in EmoVoice, give better results when they are specifically trained on the situation they will be used in. Therefore, in order to facilitate the process of building an own emotion recognition system also for non-experts, we have developed an interface for recording an emotional speech corpus and training a correspondent classifier. The idea of this is also that a normal user could create his/her own speaker dependent recognition system whose accuracy can be expected to be considerably higher than that of a general

recognition system. The method used for emotion elicitation is oriented at the Velten mood induction technique [Velten, 1968] as used in [Wilting et al., 2006] where subjects have to read out loud a set of emotional sentences that should set them into the desired emotional state. We have predefined a set of such sentences for the four quadrants in a two-dimensional emotional space: positive-active, positive-passive, negative-active, negative-passive which we map on the emotions joy, satisfaction, anger, frustration. However, users are encouraged to change sentences according to their own emotional experiences. Though our goal is the recognition of non-acted emotions, of course, this method does not yield truly natural emotions, we could rather call them semi-acted. For offline recognition, research has shifted just in recent years from acted to spontaneous emotions, so that for fully natural low-intensity emotions in online recognition, only low accuracies can be expected. At the current state of the art, rather applications should be considered where expressive speech comes natural, e. g. games or voice training.

We tested this method with 29 students of computer science (8 females, 21 males, aged 20 to 28). The sentence set was as described above and in German, though there were also 10 non-native speakers among them. Students could do the recordings at home, so the audio quality and equipment were not controlled, but all students were told to use a head-set microphone. Offline speaker-dependent accuracies in 10-fold cross-validation for all 4 classes varied — not surprisingly — a lot among speakers and ranged from 24 % to 74 %, with an average of 55 %. This great variation is to a good extent due to the uncontrolled audio recordings which led to very different audio and emotion qualities but this especially makes our setting very realistic with regard to how people cope with the technology on their own.

From all test persons, we selected 10 speakers (5 female, 5 male) that were German native speakers, whose speaker-dependent accuracy was not below 40 % and where audio quality was satisfactory, to train a speaker-independent classifier that could be used as general classifier in many applications responding to emotional states occuring in the recorded set. This resulted in a recognition accuracy of 41 %. All results were obtained with the NB classifier on the full feature set (no selection) and though the figures may not sound high overall, they are well above chance level. Especially in the speaker-independent evaluation, the use of different microphones is responsible to a great extent for low recognition rates. Furthermore, for good results in a realistic setting and online recognition, only 2 or 3 of these classes should be used. For example, we obtained recognition rates between 60 % and 70 % for the speaker independent system when leaving two classes out. Again note that all recognition accuracy figures were obtained offline, though speech data and recording conditions are similar to online conditions. A systematic evaluation of online recognition accuracy has not been done yet, but is empirically 10–20 % lower than the offline accuracy if applied in a scenario similar to the recording conditions. In offline analysis of emotional speech databases, we achieved in our earlier work recognition rates of about 80 % for 7 classes [Vogt and André, 2006] on an actors database [Burkhardt et al., 2005a]

and about 50 % in [Vogt and André, 2005] resp. [Batliner et al., 2006] on two spontaneous emotional speech databases, the SmartKom database (3 emotion classes) [Schiel et al., 2002] and the German Aibo database (4 emotion classes) [Batliner et al., 2004].

It took each speaker about 10–20 minutes to record the 80 sentences, 20 for each emotion. For a good speaker dependent system, however, we recommend at least 40 sentences per emotion.

4 Integration into applications

Of course, knowing the emotion expressed by one's own voice is not very useful *per se*, but only in the context of an application making use of the affective information. The integration of the online recognition tool of EmoVoice into other applications is simple, as the result of the emotion recognition can be continuously transmitted over a socket connection to that application.

EmoVoice has been also successfully integrated in a number of applications or existing architectures. So far, there exist several prototypes and applications that use EmoVoice. Two of them look at whether affective reactions make a robot or virtual agent more believable. These include a scenario of human-robot interaction where a user reads a fairy tale to Barthoc, a humanoid robot, expressing the emotions joy and fear, and the robot mimics these emotions with its facial expressions (see Fig. 2). The other is a virtual agent named Greta

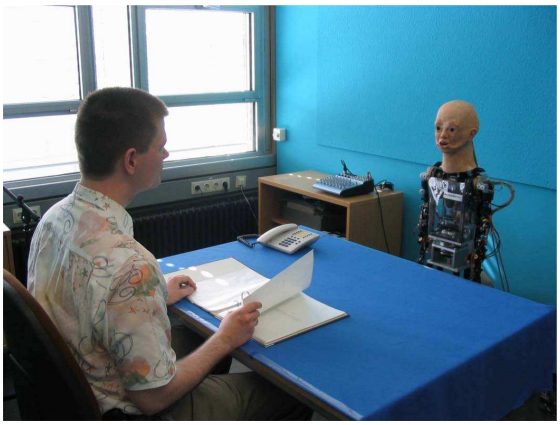

Fig. 2. Emotionally telling a fairy tale to Barthoc, a humanoid robot [Hegel et al., 2006].

[de Rosis et al., 2003] which mirrors a user's emotional state in her face (see Fig. 3) and gives emotionally coloured small-talk feedback, thus showing empathy with the user [Vogt et al., 2007]. Furthermore, through the mirroring behavior, the results of the emotion recognition are made especially clear to the

user. For the first scenario, Hegel et al. [Hegel et al., 2006] show in a user study

Fig. 3. Conversation with a virtual agent showing empathy by mirroring the user's emotional state in her face (upper row from left to right: joy, sadness, anger).

a preference of the emotionally reacting robot over a robot without emotion recognition. For the second scenario, the formal proof of this is still pending but due to the more subtle emotional response by the Greta agent, we expect an even stronger effect.

Other applications are of rather artistic nature having the goal of visualising emotions and allowing users to express themselves emotionally. One of them is an animated kaleidoscope that changes according to a speaker's emotions (see Fig. 4). Within the EU project Callas[1], showcases of interactive art are being developed that respond to the multimodal emotional input of spectators. Some of them are intended to be used primarily by professional actors, and it is assumed and encouraged that users express themselves with strong, possibly exaggerated and acted emotions. For this reason these scenarios are ideally suited for the current state of the art in emotion recognition technology. One of the showcases

[1] http://www.callas-newmedia.eu

Fig. 4. An animated kaleidoscope to visualise online recognised emotional states (examples from left to right: joy, sadness, anger).

integrating EmoVoice is an Augmented Reality application with a tree whose appearance can be changed according to affective input from the user. Among others, the tree can be made grow or shrink and change its color by positive or negative emotions as expressed in the voice [Gilroy et al., 2007] (see Fig. 5). In

Fig. 5. Making E-Tree [Gilroy et al., 2007] grow: influence of neutral, positive active, negative passive emotions from voice (from left to right).

another showcase, a virtual character watches a (horror) movie and reacts to the emotions conveyed in the scenes and by a human spectator [Charles et al., 2007].

Finally, — not in the context of Callas — Rehm et al. [Rehm et al., 2008] have built a virtual Karaoke dancer whose expressiveness can be controlled by expressive singing and gestures.

In the mentioned applications or prototypes, three different languages, German, English and Finnish, were used. This shows that the methodology of EmoVoice is language-independent.

5 Conclusions and future work

We presented EmoVoice, our framework to emotion recognition from speech which allows online tracking of the emotions expressed in a user's voice and comes with an easy-to-use audio acquisition method to quickly build a speaker- or scenario specific recogniser. The framework has been integrated successfully

already in a number of applications. Regarding the recognition process, the biggest contribution of EmoVoice is not the single components used for speech emotion recognition (audio segmentation, feature extraction, classification) but the way — or even the fact that — they are plugged together. Of course, feature extraction and classification should be improved to have higher accuracy and are subject to constant improvements as also in offline emotion recognition, the real breakthrough has not yet been achieved. By combining EmoVoice with other modalities or information sources like visual attention, facial expressions, bio signals or word information, which we have partly already investigate (bio signals [Kim et al., 2005], gender information [Vogt and André, 2006]), we plan to explore further possiblities of recognition rates improvements. Primarily, however, with EmoVoice we have a framework to explore user behavior and acceptance of affective technology. Consequently, one of our future steps will therefore include a thorough user study where we will also assess systematically the accuracy of our system in online usage.

Acknowledgements

Thanks to Stephen Gilroy from Univ. of Teesside, UK, for providing the pictures in Fig. 5.

This work was partially supported by the European Community (EC) within the Callas project IST-034800, the eCIRCUS project IST-4-027656-STP and the network of excellence Humaine IST-507422. The authors are solely responsible for the content of this publication. It does not represent the opinion of the EC, and the EC is not responsible for any use that might be made of data appearing therein.

References

[Ai et al., 2006] Ai, H., Litman, D. J., Forbes-Riley, K., Rotaru, M., Tetreault, J., and Purandare, A. (2006). Using system and user performance features to improve emotion detection in spoken tutoring dialogs. In [Interspeech, 2006].

[Batliner et al., 2004] Batliner, A., Hacker, C., Steidl, S., Nöth, E., D'Arcy, S., Russell, M., and Wong, M. (2004). "You stupid tin box" - children interacting with the AIBO robot: A cross-linguistic emotional speech corpus. In *Proceedings of the 4th International Conference of Language Resources and Evaluation LREC 2004* , pages 171–174, Lisbon.

[Batliner et al., 2006] Batliner, A., Steidl, S., Schuller, B., Seppi, D., Laskowski, K., Vogt, T., Devillers, L., Vidrascu, L., Amir, N., Kessous, L., and Aharonson, V. (2006). Combining efforts for improving automatic classification of emotional user states. In *Proc. IS-LTC 2006*, Ljubljana, Slovenia.

[Boersma and Weenink, 2007] Boersma, P. and Weenink, D. (2007). Praat: doing phonetics by computer (version 4.5.15) [computer program]. http://www.praat.org/. Retrieved 24.02.2007.

[Burkhardt et al., 2005a] Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., and Weiss, B. (2005a). A database of German emotional speech. In [Interspeech, 2005].

[Burkhardt et al., 2005b] Burkhardt, F., van Ballegooy, M., Englert, R., and Huber, R. (2005b). An emotion-aware voice portal. In *Electronic Speech Signal Processing Conference*, Prague, Czech Republic.

[Chang and Lin, 2001] Chang, C.-C. and Lin, C.-J. (2001). *LIBSVM: a library for support vector machines.* Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[Charles et al., 2007] Charles, F., Lemercier, S., Vogt, T., Bee, N., Mancini, M., Urbain, J., Price, M., André, E., Pelachaud, C., and Cavazza, M. (2007). Affective interactive narrative in the callas project. In *Demo paper in Proceedings of the 4th International Conference on Virtual Storytelling*, Saint Malo, France.

[de Rosis et al., 2003] de Rosis, F., Pelachaud, C., Poggi, I., Carofiglio, V., and de Carolis, B. (2003). From Greta's mind to her face: modelling the dynamics of affective states in a conversational embodied agent. *International Journal of Human-Computer Studies*, 59:81–118.

[Fink, 1999] Fink, G. (1999). Developing HMM-based recognizers with ESMERALDA. In Matoušek, V. et al., editors, *Lecture notes in Artificial Intelligence*, volume 1962, pages 229–234. Springer, Berlin, Heidelberg.

[Gilroy et al., 2007] Gilroy, S. W., Cavazza, M., Chaignon, R., Mäkelä, S.-M., Niiranen, M., André, E., Vogt, T., Billinghurst, M., Seichter, H., and Benayoun, M. (2007). An emotionally responsive AR art installation. In *Proceedings of ISMAR Workshop 2: Mixed Reality Entertainment and Art*, Nara, Japan.

[Hall, 1998] Hall, M. A. (1998). Correlation-based feature subset selection for machine learning. Master's thesis, University of Waikato, New Zealand.

[Hegel et al., 2006] Hegel, F., Spexard, T., Vogt, T., Horstmann, G., and Wrede, B. (2006). Playing a different imitation game: Interaction with an empathic android robot. In *Proc. 2006 IEEE-RAS International Conference on Humanoid Robots (Humanoids'06)*.

[Interspeech, 2005] Interspeech (2005). *Proceedings of Interspeech 2005*, Lisbon, Portugal.

[Interspeech, 2006] Interspeech (2006). *Proceedings of Interspeech 2006 — ICSLP*, Pittsburgh, PA, USA.

[Jones and Deeming, 2007] Jones, C. and Deeming, A. (2007). Affective human-robotic interaction. In Peter, C. and Beale, R., editors, *Affect and Emotion in Human-Computer Interaction*, volume 4868 of *LNCS*. Springer, Heidelberg, Germany.

[Jones and Sutherland, 2007] Jones, C. and Sutherland, J. (2007). Acoustic emotion recognition for affective computer gaming. In Peter, C. and Beale, R., editors, *Affect and Emotion in Human-Computer Interaction*, volume 4868 of *LNCS*. Springer, Heidelberg, Germany.

[Kim et al., 2005] Kim, J., André, E., Rehm, M., Vogt, T., and Wagner, J. (2005). Integrating information from speech and physiological signals to achieve emotional sensitivity. In [Interspeech, 2005].

[Madan, 2005] Madan, A. (2005). Jerk-O-Meter: Speech-Feature Analysis Provides Feedback on Your Phone Interactions. http://www.media.mit.edu/press/jerk-o-meter/. retrieved: 28.06.2007.

[Oudeyer, 2003] Oudeyer, P.-Y. (2003). The production and recognition of emotions in speech: features and algorithms. *International Journal of Human-Computer Studies*, 59(1–2):157–183.

[Rehm et al., 2008] Rehm, M., Vogt, T., Wissner, M., and Bee, N. (2008). Dancing the night away — controlling a virtual karaoke dancer by multimodal expressive cues. In *Proceedings of AAMAS'08*.

[Schiel et al., 2002] Schiel, F., Steininger, S., and Türk, U. (2002). The SmartKom multimodal corpus at BAS. In *Proceedings of the 3rd Language Resources & Evaluation Conference (LREC) 2002*, pages 200–206, Las Palmas, Gran Canaria, Spain.

[Schuller et al., 2007a] Schuller, B., Rigoll, G., Grimm, M., Kroschel, K., Moosmayr, T., and Ruske, G. (2007a). Effects of in-car noise-conditions on the recognition of emotion within speech. In *Proc. of the DAGA 2007*, Stuttgart, Germany.

[Schuller et al., 2007b] Schuller, B., Seppi, D., Batliner, A., Maier, A., and Steidl, S. (2007b). Towards more reality in the recognition of emotional speech. In IEEE, editor, *Proc. ICASSP 2007*, volume 2, pages 941–944, Honolulu, Hawaii, USA.

[Velten, 1968] Velten, E. (1968). A laboratory task for induction of mood states. *Behavior Research & Therapy*, (6):473–482.

[Vogt and André, 2005] Vogt, T. and André, E. (2005). Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition. In *Proceedings of International Conference on Multimedia & Expo*, Amsterdam, The Netherlands.

[Vogt and André, 2006] Vogt, T. and André, E. (2006). Improving automatic emotion recognition from speech via gender differentiation. In *Proc. Language Resources and Evaluation Conference (LREC 2006)*, Genoa.

[Vogt et al., 2007] Vogt, T., André, E., and Wagner, J. (2007). Automatic recognition of emotions from speech: a review of the literature and recommendations for practical realisation. In Peter, C. and Beale, R., editors, *Affect and Emotion in Human-Computer Interaction*, volume 4868 of *LNCS*. Springer, Heidelberg, Germany.

[Wilting et al., 2006] Wilting, J., Krahmer, E., and Swerts, M. (2006). Real vs. acted emotional speech. In [Interspeech, 2006].

[Witten and Frank, 2005] Witten, I. H. and Frank, E. (2005). *Data Mining: Practical machine learning tools with Java implementations*. Morgan Kaufmann, San Francisco, 2nd edition.