Automatic Recognition of Emotions from Speech: A Review of the Literature and Recommendations for Practical Realisation

Thurid Vogt, Elisabeth André, and Johannes Wagner

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

Abstract. In this article we give guidelines on how to address the major technical challenges of automatic emotion recognition from speech in human-computer interfaces, which include audio segmentation to find appropriate units for emotions, extraction of emotion relevant features, classification of emotions, and training databases with emotional speech. Research so far has mostly dealt with offline evaluation of vocal emotions, and online processing has hardly been addressed. Online processing is, however, a necessary prerequisite for the realization of human-computer interfaces that analyze and respond to the user's emotions while he or she is interacting with an application. By means of a sample application, we demonstrate how the challenges arising from online processing may be solved. The overall objective of the paper is to help readers to assess the feasibility of human-computer interfaces that are sensitive to the user's emotional voice and to provide them with guidelines of how to technically realize such interfaces.

1 Introduction

Automatic emotion recognition from speech has in the last decade shifted from a side issue to a major topic in human computer interaction and speech processing. The aim is to enable a very natural interaction with the computer by speaking instead of using traditional input devices and not only have the machine understand the verbal content, but also more subtle cues such as affect that any human listener would easily react to. This can be used in spoken dialogue systems, e.g. in call center applications. However, so far real-time emotion recognition has scarcely been attempted and if so, only in prototypical applications, as there are still many problems that are not yet solved appropriately.

In this article, we focus on technical challenges that arise when equipping human-computer interfaces with the ability to recognize the user's vocal emotions. Therefore, we will start with a short introduction into the acoustic properties of voice that are relevant for emotions as identified by psychological studies and move on to a discussion of databases with emotional speech. To give the reader an idea of how to employ information on the user's emotional state in human-computer interfaces, we then present a number of promising application

C. Peter and R. Beale (Eds.): Affect and Emotion in HCI, LNCS 4868, pp. 75–91, 2008.
© Springer-Verlag Berlin Heidelberg 2008

fields. After that, we address the three main parts of automatic emotion recognition, namely finding appropriate audio units, feature extraction, and classification, that pose the hardest problems. Last, we exemplify the major difficulties of real-time emotion recognition by means of a sample application. The overall objective of the paper is to help readers to assess the feasibility of human-computer interfaces that are sensitive to the user's emotional voice and to provide them with guidelines of how to technically realize such interfaces.

2 Acoustic Measures of Emotion in Speech

Information on emotion is encoded in all aspects of language, in what we say and in how we say it, and the 'how' is even more important than the 'what'. This article focuses on the phonetic and acoustic properties of affective spoken language. The vocal parameters that have been best researched by psychological studies in relation to emotion and which are also intuitively the most important ones are prosody (pitch, intensity, speaking rate) and voice quality. Murray and Arnott [1] wrote an often cited review of literature on emotions in speech and refer to a number of studies which seemingly have identified unambiguous acoustic correlates of emotions as displayed in Table 1. These show prosody and voice quality to be most important to distinguish between emotions according to human perception. In particular pitch and intensity seem to be correlated to activation, so that high pitch and intensity values imply high, low pitch and intensity values low activation.

Table 1. Some variations of acoustic variables observed in relation to emotions, summarised from [1]

Emotion	Pitch	Intensity	Speaking rate	Voice quality
Anger	high mean, wide range	increased	increased	breathy; blaring timbre
Joy	increased mean and range	increased	increased	sometimes breathy; moderately blaring timbre
Sadness	normal or lower than normal mean, narrow range	decreased	slow	resonant timbre

The automatic recognition of emotion seems straight-forward when looking at Table 1. However, this is unfortunately not the case. First of all, psychological studies often get their insights from data of test persons *acting* to be in an emotional state. There, this clear mapping from acoustic variables might even be possible in a number of cases, though even when acting, intra- and inter-speaker variations are higher, as the expressivity of emotions is also dependent

on the personality or the mood. In everyday human computer interaction, however, the occurring emotions are very spontaneous. There, these variations are considerably higher as these are not any more prototypical emotions but may be shaded, mixed, or weak and hardly distinguishable. This makes the task much harder, so that further acoustic features need to be investigated. Of course, personalised emotion recognition, that is from only one speaker, is more reliable. Further evidence of the differences of acted and spontaneous emotions has been supplied by Wilting *et al.* who showed in human listening tests that the perception of acted emotions is different than that from natural emotions [2]. In an experiment based on the Velten mood induction method [3] they had one group of test persons utter a set of positive or negative emotional sentences, and another group that was told to utter the same sentences, but act positively for the negative sentences, and negatively for the positive sentences. After the experiment, the first group stated that they actually felt positive or negative, while the other group felt neutral. Furthermore, in a perception experiment, listeners judged acted emotions to be stronger than natural emotions which suggests that actors tend to exaggerate. So, assumptions that hold for acted emotions do not necessarily transfer to natural emotions which are obviously of greater interest to human-computer interaction.

3 Databases

Databases with emotional speech are not only essential for psychological studies, but also for automatic emotion recognition, as standard methods are statistical and need to learn by examples. Generally, research deals with databases of acted, induced or completely spontaneous emotions. Of course, the complexity of the task increases with the naturalness. So at the beginning of the research on automatic vocal emotion recognition, which started seriously in the mid-90s, work began with acted speech [4] and shifts now towards more realistic data [5,6]. Prominent examples for acted databases are the Berlin database of emotional speech [7] and the Danish Emotional Speech corpus (DES) [8] which hold recordings of 10 resp. 4 test persons that were asked to speak sentences of emotionally neutral content in 7 resp. 5 basic emotions. Induced data is for instance the SmartKom corpus [9] and the German Aibo emotion corpus [10] where people where recorded in a lab setting fulfilling a certain task that was intended to elicit e. g. anger or irritation in the subjects without them knowing that their emotional state was of interest. The call center communication dealt with by Devillers and colleagues [5] is fully realistic as it is obtained from live recordings.

The labeled emotions in the databases — and consequently also the emotions that are going to be recognised — can be a classic set of basic emotions like joy, anger, sadness, disgust. Alternatively, emotion states can be placed within a dimensional model of two or three affective dimensions (see Fig. 1). The dimensions are usually valence (from positive to negative) and arousal (from high to low), sometimes a third dimension like stance (from open to close) is added. A dimensional model allows for a continuous description which is very suitable

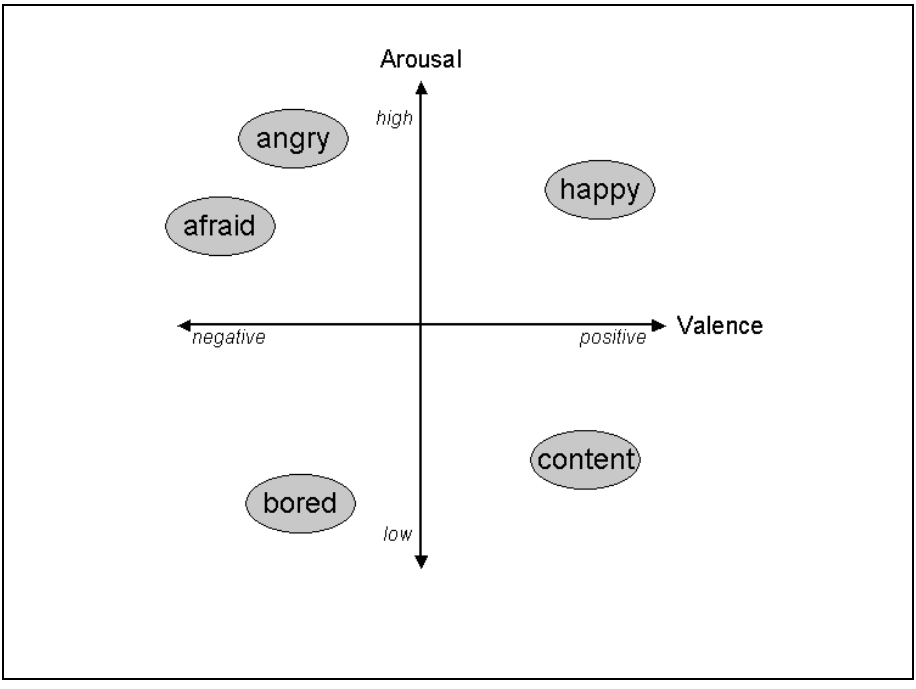

Fig. 1. A two-dimensional emotion space with a valence and an arousal axis. Basic Emotions are marked as areas within the space.

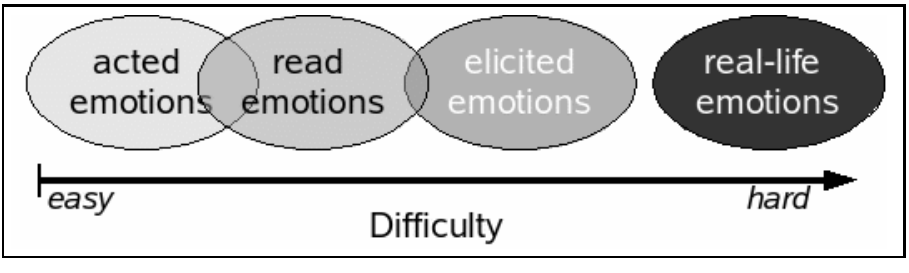

Fig. 2. Types of databases used for emotion recognition and their difficulty

for spontaneous emotions. However, for automatic recognition, this is usually mapped onto the four quadrants positive/high, positive/low, negative/high and negative/low [11,12], since it increases the complexity of the recognition task. Recently, Grimm et al. [13] used a regression technique to classify into a continuous three-dimensional emotion space. There are also only few databases available that are labeled using a dimensional model (e. g. the SAL – Sensitive Artificial Listener database [14]).

Often, the set of emotions is application driven, and can then contain for instance boredom, frustration, or even motherese (baby-talk) [15]. Other less

obvious settings are the distinction between problem and no problem in a dialogue [16], or the detection of engagement [12]. The term "emotion" is thus interpreted widely and would rather comprise all affect related user states that occur in human-computer interaction. However, the more realistic the data is, the smaller is the number of classes that can feasibly be processed. Figure 2 illustrates how the difficulty of emotion recognition increases with the type of data used.

4 Applications

Call centre conversations belong to the most popular applications for approaches to the automated recognition of emotions from speech. On the one hand, a system may provide human operators with information regarding the emotions their voice might portray. That is, the system serves as a kind of "Affective Mirror" [17] that helps users to improve their interaction skills. An example includes 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 [18]. On the other hand, mechanisms for detecting emotions may be employed to sort voice messages according to the emotions portrayed by the caller. Among other things, a dialogue system may deploy knowledge on emotional user states to select appropriate conciliation strategies and to decide whether or not to transfer the caller to a human agent. An example includes the emotion-ware voice portal currently under development at T-Systems [19]. Furthermore, information on the caller's emotional state may be used to predict system error rates. Riccardi and Hakkani-Tür [20] investigate how the user's emotional state affects the accuracy of the AT&T "How May I Help You?" spoken dialogue system and conclude that the detection of the caller's emotional state may be beneficial for the adaptation of the system's dialogue strategies. In the case of anger, the performance of the dialogue system tends to go down, for example. This knowledge may again be used to select appropriate repair strategies.

Recently, methods for the recognition of emotions from speech have also been explored within the context of computer-enhanced learning. The motivation behind these approaches is the expectation that the learning process may be improved if a tutoring system adapts its pedagogical strategies to a student's emotional state. For instance, Ai and colleagues [21] consider features extracted from the dialogue between the tutor and the student, such as the prosody of speech, as well as features relating to user and system performance for the emotion recognition process in the ITSpoke tutoring system.

Starting in the last years, research has been conducted to explore the feasibility and potential of emotionally aware in-car systems. This work is motivated by empirical studies that provide evidence of the dependencies between a driver's performance and his or her emotional state. Emotion recognition from speech in cars has so far been investigated e.g. in the Emotive Driver project [22], and in the FERMUS project, a cooperation with the automobile

industry (DaimlerChrysler and BMW) [23] where, however, experiments have been restricted to data collection and evaluation of driving simulators scenarios.

Finally, emotion detection has a high potential in games [24] and for giving feedback in human-robot interaction [25,26].

Summing up, it may be said that there is a large application potential for emotion-aware speech recognition systems. Nevertheless, the wide-spread exploitation of such systems is still impeded by great technical issues that need to be solved. As we will see later, especially online emotion recognition from speech which is in most cases necessary in real-time human-computer interaction poses great challenges. One possibility to mitigate such problems is the reduction to few emotional states. For instance, Burkhardt and colleagues distinguish between low and high anger as well as neutrality while Riccardi and Hakkani-Tür consider positive and negative user states only. The ITSpoke tutoring system uses the additional user and system performance features to enhance robustness.

5 Automatic Speech Emotion Recognition

A speech emotion recognition system consists of three principal parts, as shown in figure 3: signal processing, feature calculation and classification. Signal processing involves digitalisation and potentially acoustic preprocessing like filtering, as well as segmenting the input signal into meaningful units. Feature calculation is concerned with identifying relevant features of the acoustic signal with respect to emotions. Classification, lastly, maps feature vectors onto emotion classes through learning by examples. In the following we will discuss audio segmentation, feature extraction and classification in more detail pointing out differences between acted and spontaneous speech which is obviously of higher relevance for human-computer interfaces.

5.1 Audio Segmentation

The goal of the audio segmentation is to segment a speech signal into units that are representative for emotions. These are usually linguistically motivated middle-length time intervals such as words or utterances. Though the decision on

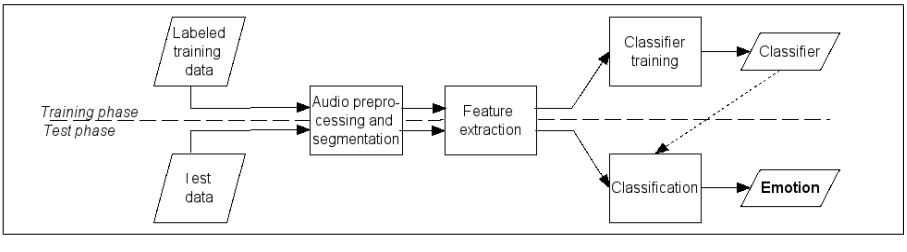

Fig. 3. Overview of a speech emotion recognition system

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 utterances of acted emotions where the choice of unit is obviously just one 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. Generally speaking, a good emotion unit has to fulfill certain requirements. In particular, it should be (1) long enough to reliably calculate features by means of statistical functions and (2) short enough to guarantee stable acoustic properties with respect to emotions within the segment: For features calculated from global statistics over an extraction unit, these units need to have a minimum length. The more values statistical measures are based on, the more expressive they are. On the other hand all alterations of the emotional state should possibly be captured, so the unit should be short enough that no change of emotion is likely to happen within. In addition, it should be so short that the acoustic properties of the segment with respect to emotions are stable, so that expressive features can be derived. This is particularly important for features based on statistical measures, since e. g. the mean value of a very inhomogeneous segment yields an inadequate description. So a compromise has to be found for these two contradicting requirements.

So far, only few attempts have been undertaken to compare different types of units. In [27], we compared utterances, words, words in context and fixed time intervals and found larger, linguistically motivated units tending to be better. Batliner et al. [16] grounded their features on words with a varying number of context words. In addition to simple word-level recognition, they also mapped word results onto turns and on chunks within the turns using two different strategies. In a qualitative analysis of this they found both advantages and disadvantages of smaller units than turns, but they have not further quantitatively explored it.

Generally, it strongly depends on the data which unit fits best. Most commonly dialogue turns, utterances or phrases as e. g. in [5,28,6,29,30] have been used, but also words [16,31].

5.2 Feature Extraction

The second step for an emotion classification system is the extraction of relevant features. Its aim is to find those properties of the digitised and preprocessed acoustic signal that are characteristic for emotions and to represent them in a n-dimensional feature vector. So far, there is not yet a general agreement on which features are the most important ones and good features seem to be highly data dependent [5,27]. However, a high number of features is often not beneficial because most classifiers are negatively influenced by redundant, correlated or irrelevant features. As a consequence, most approaches compute a high number of features and apply then, in order to reduce the dimensionality of the input data, a feature selection algorithm that chooses the most significant features of the training data for the given task. Alternatively, a feature reduction

algorithm like principal components analysis (PCA) can be used to encode the main information of the feature space more compactly. The start set of features consisted originally mainly of pitch and energy related features, and these continue to be the prominent features. Formants and Mel Frequency Cepstral Coefficients (MFCC) are also frequently found. Durational and pause related features are noted in several papers, as well as different types of voice quality features. Spectral measures and parametric representations other than MFCCs are less common, but include wavelets, Teager energy operator (TEO) based features, log frequency power coefficients (LFPC) and linear prediction cepstral coefficients (LPCC).

The raw pitch, energy, etc. contours can be used as is, and are then called short-term features, or more often, the actual features are derived from these acoustic variables by applying (statistic) functions over the sequence of values within an emotion segment, thus called global statistics features. This could be e g. mean pitch of a word or an utterance; further statistical measures are typically maximum, or minimum, etc. of the segment, but also regression, derivations or other more complex functions. The choice of feature type also determines the type of classifier. For global statistics features, a static classifier like Support Vector Machines (SVM), processing one instance at a time has to be used. Short-term features require a dynamic classifier such as Hidden Markov Models (HMM). One can say, that in the first case, dynamic properties of emotions should be captured by the features, while in the latter case, they are dealt with by the classifier.

Some suprasegmental acoustic phenomena may also be considered as global emotion features. Batliner et al. [32] and Devillers et al. [5] used those, among them hyper-clear speech, pauses inside words, syllable lengthening, off-talk, resp. disfluency cues, inspiration, expiration, mouth noise, laughter, crying, unintelligible voice. Though these have been mainly annotated by hand, automatic extraction would also be possible in some cases.

Furthermore, meta-data can be used to enhance recognition accuracy as e. g. applied by Litman and colleagues [6]: They collected a corpus from a spoken dialogue tutoring system in the physics domain and hence incorporated into their feature set further application dependent knowledge like the respective speaker, the gender and which of five available physics problems was treated.

Unfortunately, it is rarely possible to compare features across published work, since conditions vary a lot and even slight changes in the general set-up can make results incomparable. E. g. most researchers use their own recordings, and different data or particularly data types have a huge impact on the comparability between two approaches. As for now, there don't exist standard databases that could be used for benchmarking. For one database, 50% accuracy may be excellent for a 4-class problem, while for another database, recognition rates of 70% to 80% can be reached. This does not mean that the database in the former case was not well designed, but rather that it is a harder task and that can be due to many factors. A rule of thumb for natural emotions is that recognition rate is not much more than twice chance level, so for a 4-class problem, 50% is good.

Classifiers, target classes, speaker types also differ in the various publications on automatic emotion recognition, so that from a comparison of the literature no general statement can be made on which features are most successful. Of course, comparisons of features within publications are made, e. g. through relevance ranking by the information gain of single features [29,30] or by rank in a sequential selection method [28,33]. Relevance ranking usually has the goal to see the salience of single features, usually per feature type. However, a single feature's relevance does not necessarily imply usefulness in a set of features. Another strategy is to have groups of features (e. g. prosodic, lexical, etc.) and to look at the different performance of the groups or combinations of groups e. g. [6,32,34]. No general conclusion can be drawn from the work on feature evaluation, but pitch features have on various occasions shown not to be that important as previously assumed [27,28,29,30]. As for now, this has however not been confirmed for other emotion classification tasks. The CEICES (Combining Efforts for Improving automatic Classification of Emotional user States) initiative [15] is therefore aimed at finding a more general evaluation of features by providing a database under fixed conditions and having different sites use their own features and classifiers.

5.3 Classification

After the feature calculation, each input unit is represented by a feature vector, and the problem of emotion recognition can now be considered a general data mining problem. So, in principle, any statistical classifier that can deal with high-dimensional data can be used, but static classifiers like support vector machines, neural networks, and decision trees for global statistics features, and HMM for short-term features as a dynamic modeling technique are most commonly found in the literature on emotional speech recognition. All these classifiers need training data to learn parameters.

Static classification has been more prevalent than dynamic classification in the work on emotion recognition. It has proved to be successful for acted data, but for more natural data, recognition accuracy is only useful in a problem with very few emotion classes. Recent approaches try to enhance the recognition accuracy by a multi-layered classification approach, like having several steps of classifying two groups of the target emotion classes and always further splitting the "winning" group in two as in the cascade-bisection process [35] or automatically separating male and female voice before the actual emotion classification [36].

Dynamic classification with HMMs is used less often than static classification, but is thought to be advantageous for better capturing the temporal activity incorporated in speech. So far, HMMs have almost exclusively been applied to acted data, though they might even better be suited for natural emotions. An HMM is a stochastic finite automaton, where each state models some characteristics of the input signal and where the probability to pass to the next state only depends from the previous state (cf. Fig. 4). In order to use HMMs for speech emotion recognition, usually a single HMM is trained for each emotion and an unknown sample is classified according to the model which describes the derived

feature sequence best. Beside the use of appropriate speech features, the architecture of the HMM has main influence on its ability to capture those emotional cues that help to distinguish among different emotions. In [37] we examined the three parameters that are important for the model topology, number of states, connectivity and output probabilities (discrete or continuous and number of mixtures). Although it turned out that finding general tendencies was rather difficult, since on the one hand quite different parameters sometimes gained the same results, whereas on the other hand a slight parameter sometimes caused a very different performance, we could conclude that for the model topology, a medium number of 5 to 10 states per model is most often successful, and for the output probabilities, this was the case for continuous probability densities with a low number of mixtures. With respect to the connectivity of the states, we found high connectivity not necessarily to be more suitable. Results showed also that the network design seems to be relatively independent of the source of speech (acted vs. spontaneous) and the segmentation level (word vs. utterance).

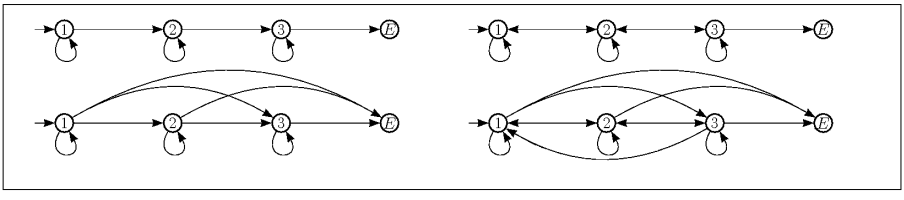

Fig. 4. Some HMM topologies that are suitable for emotion recognition

A direct comparison of static and dynamic classification is difficult since not the same features can be used, so it is difficult to say if just the features have been chosen more favorable or if really the classifier has been superior. Dynamic classification is very promising, but currently, for static classification more feature types can be exploited (e. g. suprasegmental acoustic features like jitter or shimmer to measure voice quality), so that overall, the latter performs better. However, when the feature set is restricted to the same feature types, for instance only MFCCs and energy, HMMs often outperform static modeling techniques [37].

The quality of a classifier can be determined in comparison to human raters in listening tests or to other classification algorithms. The former is more meaningful for practical purposes and shows also the complexity of the particular tasks but it usually involves much effort to conduct such a study. Human rating performance has been reported in different studies to be around 65% [38,39] which is also supported by the findings of [40] in a psychological study, about 70% [41,42], or 80% [43,44]. Interestingly, the automatic classification results presented in these papers reach about the same level or even exceed it. These figures, however, concern acted speech. For spontaneous emotions the gap would supposedly be larger, though figures like this are more difficult to obtain for

these databases. When labeling spontaneous emotions, the "true" emotion is not known but established by agreement between the labelers. A perception test can only tell the relation between this ground truth and further listeners. But, as said before, a good recognition rate for natural emotions is about twice chance level.

As a general tendency it can be observed that sophisticated classifiers do achieve higher recognition rates than simple classifiers but not much. SVM is the most popular and the most often successfully applied algorithm, so it can be considered a kind of standard.

6 Emotion Recognition in Real-Time Human-Computer Interaction: A Practical Example

The previous section focused on a systematic evaluation of static and dynamic classification for robust emotion recognition from speech. Characteristic of these approaches was that emotion recognition was done offline. That is both the training as well as the testing was performed using pre-recorded corpora whereby time was not considered as a critical factor.

In this section, we discuss challenges that arise when analyzing and responding to emotional signals in real-time while the user interacts with an application. We illustrate our ideas by means of a simple demo application where a user converses with an embodied conversational agent. The agent does not analyse the meaning of the user's verbal utterances, but instead just interprets the user's emotive cues from speech and responds to them emotionally, for example, by showing empathy when the user's voice conveys sadness. It is not supposed to take the initiative in a conversation, rather its role is that of a sensitive listening agent that gives facial and verbal feedback according to the emotions the human dialogue partner conveys with her voice. An important consideration is how often and how fast this feedback should be given. The facial feedback directly adapts to each processed emotion unit. If an emotion is directly repeated, the intensity of the facial expression increases. After a random number of repetitions, verbal feedback is given. However, outliers, that is e. g. a single occurrence of boredom in a long sequence of recognised joy, are ignored.

The approach is quite similar to the work described in [26] where we equipped an anthropomorphic robot torso with our emotional speech recognition component. The robot mirrors the emotions happiness, fear and neutral as recognised from user speech by facial expressions. A user study revealed that users perceived the emotional mimicry as a sign for the robot reacting more adequately to emotional aspects of a situation and recognising emotion better as compared to a robot reacting without emotion recognition. However, our overall objective is now the creation of rapport between the virtual agent and the human user by providing the agent with emotional sensitivity [45] as opposed to direct mimicry [26]. Furthermore, we analyse continuous dialogue instead of single utterances. Finally, the agent's feedback is not restricted to facial expressions, but may also include verbal comments, such as "That sounds wonderful!".

In Fig. 5, a user is engaged in a dialogue with the Greta agent from Pelachaud and colleagues [46]. The user is reporting on a positive event that happened to her to which Greta responds with a cheerful facial expression.

Fig. 5. Affective conversation with the Greta agent

Real-time emotion recognition is a great challenge for current methodology as apparently demands concerning robustness and accuracy are very high. Additionally, adequate response times are essential for human-computer interaction, as it becomes confusing for the user if he has to wait and there is no direct reaction. Thus, the recognition process needs to be very fast to enhance usability. Note that we are concerned here only with the test phase of the classifier (cf. Fig. 3), as training can be done offline and is not required to be especially fast.

The first issue we faced is the fast segmentation of the continuously incoming audio signal into meaningful, consistent segments. We found a voice activity detection with no in-between pauses longer than 1000 ms to be a good compromise between speed and accuracy. Pauses in the voice activity approximate phrase breaks, though the resulting segments may not be linguistically sound. However, this segmentation requires no further knowledge and is thus very fast. Furthermore, automatic linguistic segmentation by speech recognition, besides being time-consuming, is still very error-prone on spontaneous dialogue, which can easily have negative influence on the emotion recognition, too.

Concerning the features, we decided for global statistics features as more varied feature types can be exploited with this approach. In most related work of offline emotion recognition, some features used to classify emotions rely on manual labeling such as phrase accent annotation or word transcription which is obviously not possible in a fully automatic system. We limited our feature set to only fully automatically computable features without any linguistic knowledge. We computed features based on pitch, energy, MFCCs, the frequency spectrum, duration and pauses which resulted in a vector of 1316 features. Then, in order to reduce dimensionality and to improve and speed up classification, a sequential feature selection was applied ending up with 20 features related to pitch, MFCCs and energy that are most relevant for the given training data. The procedure is equivalent to the one described in [27]. There we showed that good feature sets differ significantly depending on the data type, whether it's acted or spontaneous, so a feature selection is a very beneficial step. Furthermore, we did not find the high degree of automation in the feature extraction to be a disadvantage and we assume the large number of features provided to the selection process to be responsible for this.

For classification, a Naive Bayes classifier was used. Though this is a very simple classifier, it has the advantage of being very fast without performing much worse than more sophisticated classifiers such as support vector machines.

To train the classifier, we used the Berlin emotional speech database [7] which was already introduced in section 3. As mentioned before, it holds very prototypical, acted emotions. The available emotions are fear, anger, joy, boredom, sadness, disgust as well as neutral. However, for our application better results are achieved if the set is restricted, e. g. to joy, boredom, neutral. In general, the more training and test data resemble, the better the results on the test data are. In this case, the training database is quite different from the expected test data, as it is acted and recorded in a very different setting. In our sample application we can expect people to exaggerate their emotions rather than being completely natural. On the other hand, having test speakers in the training data, thus personalising the emotion recognition in some way, would of course improve results. However, data acquisition is laborious and users would often not be willing to do so.

Even though we had to restrict ourselves to few emotion classes and recognition rates were lower than for offline processing, we have evidence that users prefer a listening agent with emotional sensitivity over a listening agent without such a behavior. Due to the more subtle emotional response by the Greta agent, we expect a stronger effect than in the experiment reported in [26] where we compared a robot with and without emotion recognition. A formal experiment testing this hypothesis is currently under preparation.

7 Conclusion

The integration of approaches to the automated evaluation of vocal emotions into human-computer interfaces presents great challenges since emotions have

to be recognized in real-time while the user is interacting with an application. These challenges affect audio segmentation to find appropriate units for emotions, extraction of emotion relevant features, classification of emotions, and training databases with emotional speech. By means of a sample application with a virtual agent giving affective feedback in a dialogue with a human user, we outlined solutions to each of the problems.

Audio segmentation can be performed by voice activity detection which is a fast segmentation method not requiring high-level linguistic knowledge. Feature extraction may only rely on automatically computable properties of the acoustic signal, but this is not a major limitation, if the approach of calculating a multitude of possible features and then selecting the most relevant ones for the given training data is taken. For classification, in principle any statistical classifier can be used with sophisticated classifiers being superior in accuracy, but simple and thus fast classifiers are often sufficient. The speech database used to train the classifier should be adjusted to the particular application as much as possible. Best is a specifically designed database with the same scenario as occurring during testing and possibly even including test speakers. Since this is often not feasible, it is also practical to switch to general databases. Furthermore, the restriction to few (maximally 3) classes is strongly suggested.

Designers of human-computer interfaces should consider carefully how to use information on the user's emotional state in the envisioned application context and what impact erroneous recognition results may have. In this paper, we have sketched some application fields, such as self monitoring systems, where the automated evaluation of vocal emotions seems promising and beneficial.

Acknowledgements

This work was partially supported by the European Community (EC) within the network of excellence Humaine IST-507422, the eCIRCUS project IST-4-027656-STP and the Callas project IST-034800. 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

1. Murray, I., Arnott, J.: Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. Journal of the Acoustical Society of America 93(2), 1097–1108 (1993)
2. Wilting, J., Krahmer, E., Swerts, M.: Real vs. acted emotional speech. In: Proceedings of Interspeech 2006 — ICSLP, Pittsburgh, PA, USA (2006)
3. Velten, E.: A laboratory task for induction of mood states. Behavior Research & Therapy 6, 473–482 (1968)
4. Dellaert, F., Polzin, T., Waibel, A.: Recognizing emotion in speech. In: Proceedings of ICSLP, Philadelphia, USA (1996)

5. Devillers, L., Vidrascu, L., Lamel, L.: Challenges in real-life emotion annotation and machine learning based detection. Neural Networks 18(4), 407–422 (2005)
6. Litman, D.J., Forbes-Riley, K.: Predicting student emotions in computer-human tutoring dialogues. In: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL), Barcelona, Spain (2004)
7. Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W.F., Weiss, B.: A database of German emotional speech. In: Proceedings of Interspeech 2005, Lisbon, Portugal (2005)
8. Engberg, I.S., Hansen, A.V.: Documentation of the Danish Emotional Speech Database (DES). Technical report. Aalborg University, Aalborg, Denmark (1996)
9. Schiel, F., Steininger, S., Türk, U.: The SmartKom multimodal corpus at BAS. In: Proceedings of the 3rd Language Resources & Evaluation Conference (LREC) 2002, Las Palmas, Gran Canaria, Spain, pp. 200–206 (2002)
10. Batliner, A., Hacker, C., Steidl, S., Nöth, E., D'Arcy, S., Russell, M., Wong, M.: "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, Lisbon, pp. 171–174 (2004)
11. Tato, R., Santos, R., Kompe, R., Pardo, J.M.: Emotional space improves emotion recognition. In: Proceedings International Conference on Spoken Language Processing, Denver, pp. 2029–2032 (2002)
12. Yu, C., Aoki, P.M., Woodruff, A.: Detecting user engagement in everyday conversations. In: Proceedings of Interspeech 2004 — ICSLP, Jeju, Korea, pp. 1329–1332 (2004)
13. Grimm, M., Kroschel, K., Harris, H., Nass, C., Schuller, B., Rigoll, G., Moosmayr, T.: On the necessity and feasibility of detecting a driver's emotional state while driving. In: International Conference on Affective Computing and Intelligent Interaction, Lisbon, Portugal, pp. 126–138 (2007)
14. Kollias, S.: ERMIS — Emotionally Rich Man-machine Intelligent System. (2002) retrieved: 09.02.2007, http://www.image.ntua.gr/ermis/
15. Batliner, A., Steidl, S., Schuller, B., Seppi, D., Laskowski, K., Vogt, T., Devillers, L., Vidrascu, L., Amir, N., Kessous, L., Aharonson, V.: Combining efforts for improving automatic classification of emotional user states. In: IS-LTC 2006, Ljubljana, Slovenia (2006)
16. Batliner, A., Fischer, K., Huber, R., Spilker, J., Nöth, E.: How to find trouble in communication. Speech Communication 40, 117–143 (2003)
17. Picard, R.W.: Affective Computing. MIT Press, Cambridge (1998)
18. Madan, A.: Jerk-O-Meter: Speech-Feature Analysis Provides Feedback on Your Phone Interactions (2005), retrieved: 28.06.2007, http://www.media.mit.edu/press/jerk-o-meter/
19. Burkhardt, F., van Ballegooy, M., Englert, R., Huber, R.: An emotion-aware voice portal. In: Electronic Speech Signal Processing Conference, Prague, Czech Republic (2005)
20. Riccardi, G., Hakkani-Tür, D.: Grounding emotions in human-machine conversational systems. In: Proceedings of Intelligent Technologies for Interactive Entertainment, INTETAIN, Madonna di Campiglio, Italy (2005)
21. Ai, H., Litman, D.J., Forbes-Riley, K., Rotaru, M., Tetreault, J., Purandare, A.: Using system and user performance features to improve emotion detection in spoken tutoring dialogs. In: Proceedings of Interspeech 2006 — ICSLP, Pittsburgh, PA, USA (2006)

22. Jones, C., Jonsson, I.: Using Paralinguistic Cues in Speech to Recognise Emotions in Older Car Drivers. In: Peter, C., Beale, R. (eds.) Affect and Emotion in Human-Computer Interaction. LNCS, vol. 4868. Springer, Heidelberg (2008)
23. Schuller, B., Rigoll, G., Grimm, M., Kroschel, K., Moosmayr, T., Ruske, G.: Effects of in-car noise-conditions on the recognition of emotion within speech. In: Proc. of the DAGA 2007, Stuttgart, Germany (2007)
24. Jones, C., Sutherland, J.: Acoustic Emotion Recognition for Affective Computer Gaming. In: Peter, C., Beale, R. (eds.) Affect and Emotion in Human-Computer Interaction. LNCS, vol. 4868. Springer, Heidelberg (2008)
25. Jones, C., Deeming, A.: Affective Human-Robotic Interaction. In: Peter, C., Beale, R. (eds.) Affect and Emotion in Human-Computer Interaction. LNCS, vol. 4868. Springer, Heidelberg (2008)
26. Hegel, F., Spexard, T., Vogt, T., Horstmann, G., Wrede, B.: Playing a different imitation game: Interaction with an empathic android robot. In: Proc. 2006 IEEE-RAS International Conference on Humanoid Robots (Humanoids 2006) (2006)
27. Vogt, T., André, E.: 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 (2005)
28. Fernandez, R., Picard, R.W.: Classical and novel discriminant features for affect recognition from speech. In: Proceedings of Interspeech 2005, Lisbon, Portugal (2005)
29. Oudeyer, P.Y.: The production and recognition of emotions in speech: features and algorithms. International Journal of Human-Computer Studies 59(1–2), 157–183 (2003)
30. Schuller, B., Müller, R., Lang, M., Rigoll, G.: Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles. In: Proceedings of Interspeech 2005, Lisbon, Portugal (2005)
31. Nicholas, G., Rotaru, M., Litman, D.J.: Exploiting word-level features for emotion recognition. In: Proceedings of the IEEE/ACL Workshop on Spoken Language Technology, Aruba (2006)
32. Batliner, A., Zeißler, V., Frank, C., Adelhardt, J., Shi, R.P., Nöth, E.: We are not amused - but how do you know? User states in a multi-modal dialogue system. In: Proceedings of Eurospeech 2003, Geneva, Switzerland, pp. 733–736 (2003)
33. Kwon, O.W., Chan, K., Hao, J., Lee, T.W.: Emotion recognition by speech signals. In: Proceedings of Eurospeech 2003, Geneva, Switzerland, pp. 125–128 (2003)
34. Lee, C.M., Narayanan, S.S.: Toward detecting emotions in spoken dialogs. IEEE Transaction on speech and audio processing 13(2), 293–303 (2005)
35. Zhang, S.,, P.: C.C., Kong, F.: Automatic emotion recognition of speech signal in mandarin. In: Proceedings of Interspeech 2006 — ICSLP, Pittsburgh, PA, USA (2006)
36. Vogt, T., André, E.: Improving automatic emotion recognition from speech via gender differentiation. In: Proc. Language Resources and Evaluation Conference (LREC 2006), Genoa (2006)
37. Wagner, J., Vogt, T., André, E.: A systematic comparison of different hmm designs for emotion recognition from acted and spontaneous speech. In: International Conference on Affective Computing and Intelligent Interaction (ACII), Lisbon, Portugal, pp. 114–125 (2007)
38. Nwe, T.L., Foo, S.W., De Silva, L.C.: Speech emotion recognition using hidden markov models. Speech Communication 41, 603–623 (2003)

39. Petrushin, V.A.: Creating emotion recognition agents for speech signal. In: Dautenhahn, K., Bond, A.H., Canamero, L., Edmonds, B. (eds.) Socially Intelligent Agents. Creating Relationships with Computers and Robots, pp. 77–84. Kluwer Academic Publishers, Dordrecht (2002)
40. Scherer, K.R., Banse, R., Walbott, H.G., Goldbeck, T.: Vocal clues in emotion encoding and decoding. Motivation and Emotion 15, 123–148 (1991)
41. Polzin, T.S., Waibel, A.H.: Detecting emotions in speech. In: Proceedings of Cooperative Multimodal Communications, Tilburg, The Netherlands (1998)
42. Polzin, T.S., Waibel, A.H.: Emotion-sensitive human-computer interfaces. In: Workshop on Speech and Emotion, Newcastle, Northern Ireland, UK, pp. 201–206 (2000)
43. Lee, C.M., Yildirim, S., Bulut, M., Kazemzadeh, A.: Emotion recognition based on phoneme classes. In: Proceedings of Interspeech 2004 — ICSLP, Jeju, Korea (2004)
44. Nogueiras, A., Moreno, A., Bonafonte, A., No, J.M.: Speech emotion recognition using hidden markov models. In: Proceedings of Eurospeech, Aalborg, Denmark (2001)
45. Gratch, J., Okhmatovskaia, A., Lamothe, F., Marsella, S., Morales, M., van der Werf, R.J., Morency, L.P.: Virtual rapport. In: 6th International Conference on Intelligent Virtual Agents, Marina del Rey, USA (2006)
46. de Rosis, F., Pelachaud, C., Poggi, I., Carofiglio, V., de Carolis, B.: 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 (2003)