Christian Müller (Ed.): Speaker Classification I Fundamentals, Features, and Methods.
Springer, Berlin Heidelberg 2007, LNAI 4343, pp. 138-151.

Speaker Characteristics and Emotion Classification

Anton Batliner[1] and Richard Huber[2]

[1] Lehrstuhl für Mustererkennung, Universität Erlangen–Nürnberg, Martensstr. 3,
91058 Erlangen, Germany
batliner@informatik.uni-erlangen.de
http://www5.informatik.uni-erlangen.de
[2] Sympalog Voice Solutions GmbH, Karl-Zucker-Str. 10,
91052 Erlangen, Germany
huber@sympalog.de

Abstract. In this paper, we address the — interrelated — problems of speaker characteristics (personalization) and suboptimal performance of emotion classification in state-of-the-art modules from two different points of view: first, we focus on a specific phenomenon (irregular phonation or laryngealization) and argue that its inherent multi-functionality and speaker-dependency makes its use as feature in emotion classification less promising than one might expect. Second, we focus on a specific application of emotion recognition in a voice portal and argue that constraints on time and budget often prevent the implementation of an optimal emotion recognition module.

1 Introduction

The modelling, generation, and recognition of emotion has attracted more and more attention during the last years. Most of the time, researchers have typically dealt with prototypical, 'full-blown' emotions and with elicited, prompted, acted speech [1]. Normally, some of the 'big', full-blown emotions have been modelled and classified such as *anger, joy, despair, sadness*. Recognition rates reported are fairly high; [2] for instance report for seven emotions classification rates of up to 71.0% for speaker-independent and 92.7% for speaker-dependent modelling. Nowadays, the voice business is more and more attracted by the possibilities the recognition of user states offers for commercial systems. One main focus of interest is telephony based dialogue systems with spoken input in the broad area of customer care and customer service applications.

One of the general problems is that real life data differ, however, considerably from acted speech. It is way more difficult to collect the data, cf. Labov's well-known observer's paradox [3]: for recording, the subjects have to be observed but if they are aware of that, they are no longer fully spontaneous. Moreover, ethical issues have to be taken into account. To act does not mean the same as to behave: acting refers to a shared concept of emotion expressions — how you imagine someone should behave if they are angry, sad, etc. But in real life, neither the reference is fully clear — is the subject really angry even if we wanted to make them angry with our experimental design — nor the means of expressing specific emotions. In addition, the full range of pure emotions cannot be observed in real life encounters; instead, most of the data are not marked, i.e., neutral, and the non-neutral states are rather emotion-prone/affective in a broader meaning. Last but not least, speaker characteristics can superimpose emotion expressions or interfere with them. Specific applications need specific emotion modelling: for instance, in call center scenarios, we either look for a chance in the user's emotional state or or for a difference in the emotional state of one certain user in contrast to an average application caller.

For the time being, the speaker-independent automatic recognition of emotional user states in realistic, spontaneous speech seems to be 'fossilized' at approx. 80% class-wise computed recognition rate for a 2-class problem, and at approx. 60% for a 4-class problem, cf. [4]. Of course, higher classification performance can be obtained by fine-tuning, for instance, by pre-selection of prototypical cases, cf. [5]. We don't know of any speaker-dependent classifications for realistic, emotional spontaneous speech yet. The reason for that might simply be that it is difficult to collect enough data for one and the same subject because normally, subjects are 'burned' when they have participated in an experiment. And the reason for the low speaker-independent classification performance might be that individual speakers employ different acoustic features in a different way; moreover, features can be multi-functional, and interlabeller agreement is — for spontaneous speech — not very high.

Note that the figures given above are for carefully designed experiments, manually annotated, realistic (real-life, spontaneous) data, speaker-independent modelling, and rather good acoustic conditions. Depending on signal-to-noise ratio and degree of spontaneity, much higher or lower classification rates can be imagined: in a personalized setting (speaker-dependent modelling) with a close-talk microphone in a quiet office surrounding, if the speaker only has to produce a limited amount of commands, and if it is clear when and that they are getting angry, recognition rates well above 90% for two or even more classes can be imagined. On the other hand, in a public, noisy setting with a room microphone, free speech, and speaker-independent modelling, classification performance could drop almost to chance level. This also can happen if you switch to telephony applications where the communication channel is of restricted bandwidth; here the input quality is sometimes rather poor — just think of mobile phone calls. On the other hand, emotion recognition might not be that prone to noise as other speech processing tasks [6].

In this paper, we start with discussing automatic recognition of emotion and user states on a conceptual level. We address some basic challenges and possible reasons why the approaches until now have not been fully successful. Then we report on experiences made in a project where emotion recognition was integrated and applied in a real business environment; constraints in time and budget made it impossible, however, to implement an optimal emotion recognition module.

2 Setting the Scene

2.1 Concepts: Emotion and Speaker Characteristics

In this paper, we use the term ‘emotion’ in a broad sense, encompassing emotional (affective) user states such as bored, interested, stressed, despaired, perplexed as well. Other terms used for such additional states is ‘interpersonal stances’ [7] or ‘social emotions’ [8]. As for speaker characteristics, we want to focus on acoustic features because this field has been more investigated than linguistic features. We do not know of any study dealing with spontaneous, real-life speech, emotions, and in-depth description of speaker-specific traits. Thus we decided to demonstrate the possible impact of speaker-specific characteristics on emotion classification with a sort of ‘gedanken experiment’: how a specific phenomenon (irregular phonation, ‘laryngealization’, cf. below) can affect emotion recognition.

2.2 Two Different Worlds: Generation and Analysis

Synthesis of emotion uses controlled data based on acted speech, and models normally one speaker and/or the same segmental structure, focussing on forced choices in listening experiments for evaluation. Realistic emotion recognition deals with uncontrolled, i.e. spontaneous data based on many speakers, uncontrolled segmental structure and wording; as computation of features, esp. for large databases, is done automatically, extraction errors have to be accepted whose extent can only be estimated roughly.

2.3 Personalization and Data Acquisition: A Problem

Although it had been desirable to develop speaker-independent automatic dictation systems, they have been more or less speaker-dependent (speaker-adaptive) for the last decades. Only the latest versions claim to be really speaker-independent, i.e. a training phase should no longer be necessary. It might be astonishing that for such a complicated problem as emotion recognition, almost all of the studies on emotion recognition in spontaneous speech used speaker-independent modelling. We believe that two factors have been responsible for that: first, the whole speech processing community is oriented towards speaker-independence. Second — and maybe most important in our context — it is difficult to collect enough emotional data from one and the same person, cf. above. We are thus faced with a dilemma: personalization seems to be the only way out towards higher classification performance, but it is way more difficult to obtain than in the case of automatic dictation systems where only subjects are needed with enough patience to read longer stretches of text.

2.4 A Tentative Relevance Hierarchy for Speaker-independent Emotion Recognition in Spontaneous Speech

In this subsection, we want to set up a tentative hierarchy of relevance in speaker-independent emotion recognition in spontaneous speech — as a sort of null hypothesis to be tested in further experiments. This hierarchy is based on own experimental results and on some other studies. Several caveats have to be made: most of the studies on relevant features used acted data; these are not taken into account. The next point is trivial but important: statements on relevant features can only be made on those features that were computed for the respective databases. In some studies, only few features or only features of a certain type are computed; as for other types, no statement can be made. On the other hand, if too many features are computed — nowadays, a set of basic features is often multiplied via different normalization and transformation procedures — it is often not easy to tell apart important from spurious information. And last but not least, it depends of course on the type of data — and by that, on the emotion classes annotated — the features are computed for. Hopefully, results will converge in the future.

Most relevant so far seem to be duration and Mel Frequency Cepstral Coefficient (MFCC) features, then energy and pitch variation (jitter, mean square error of regression). 'Genuine' pitch features such as F0 maximum and minimum — and by that, range — are not that important. MFCC features are 'implicit' spectral features which, however, encode linguistic information as well: they are standard features in word recognition. It is thus difficult to disentangle spectral information itself from word information. Linguistic information depends heavily on the type of data: for uniform speech such as commands, it should not be relevant. On the other hand, it is easy to imagine a full encoding with word information *(this makes me happy/sad/angry/...).* 'Explicit spectral' information on formant band-width or voice quality and/or phonation type can sometimes tell apart specific user states but are, on the whole, less relevant than one should suppose on the basis of acted speech or perception experiments with synthetic speech.

Note that all this is tentative and based only on some few real-life, spontaneous databases. Anyway, if it proves to be true then two points are more puzzling than the other ones in the above given hierarchy, namely that F0 is not that relevant, and that voice quality and/or phonation type is not that relevant, either. We can imagine two different reasons why: first, dimensionality, second, multi-functionality. Duration and energy are *one-dimensional*: duration on the x- (time-) axis — longer or shorter — and energy on the y- (loudness/decibel) axis — higher or lower. Even if, under certain circumstances, short duration and low energy can encode prominence, at the very most, it is the other way round. (Note that we are speaking here of 'prominence' in a broad meaning, not only of prominence denoting stress/accentuation.) Therefore, we will call these two parameters one-dimensional. F0, however, behaves differently: it is not only high vs. low pitch, it is the whole configuration, i.e. specific tunes, which are prominent. And it might be the same problem for emotion encoding as for accent encoding:

in the tone sequence terminology, accents can be marked by L*H or H*L, i.e. by two ‘opposite’ configurations, whereas almost never, accents are marked by short duration or low energy. Therefore, we will call F0 features *bi-dimensional*. In the next chapter, we will give an example for the multi-functionality of voice quality and phonation type features.

Normally, for emotion classification, acoustic features are extracted automatically by, for instance, doing forced alignment on the spoken word chain. Thus, segmentation is not perfect, and automatic extraction is error prone. Under real-life conditions, if the spoken word chain is not known, there might be more and/or different types of segmentation errors. We do not know much yet about the extent of such extraction errors; as for F0, the ‘technical’ errors amount at least to some few percent points, even under optimal conditions. Often, error rates are higher. (In the emotional database described and processed in [4], octave errors amount to some 6 % of all voiced parts in the words.) In addition, it is not clear yet whether extraction should be close to the signal or close to perception, esp. in the case of irregular phonation, cf. below. The impact of erroneous extraction on emotion recognition is even less clear. It might be the case that MFCCs are that good even at emotion recognition because they are a coarse but robust measure, whereas ‘explicit’ spectral and voice quality measurements are more error prone.

3 An Example: Laryngealizations

The normal speech register ‘modal voice’ comprises an F0 range from about 60 to 250 Hz for male speakers and an F0 range from about 120 to 550 Hz for female speakers. Below this register there is a special phonation type whose mechanisms of production are not totally understood yet and whose linguistic functions are not much investigated until now. There is a variety of different terms for this phenomenon which are used more or less synonymously: irregular phonation, creak, vocal fry, creaky voice, pulse register, laryngealization, etc. We use laryngealization (henceforth LA) as a cover term for all these phenomena that show up as irregular voiced stretches of speech. Normally LAs do not disturb pitch perception but are perceived as suprasegmental irritations modulated onto the pitch curve. Although LAs can be found not only in pathological speech but also in normal conversational speech, most of the time they were not objects of investigation but considered to be an irritating phenomenon that has to be discarded. In [9], five different types of LAs have been established: glottalization, damping, diplophonia, sub-harmonic, and aperiodicity. Voice quality and phonation types such as LAs are known to be utilized in the generation of emotions. We have to keep in mind, however, that the bulk of evidence so far has been obtained from acted speech or from perception experiments with synthesized speech.

Table 1 displays different functions of LAs which can be linguistic or paralinguistic. They can be caused either by higher effort or by relaxation; in the first case, they go together with *accentuation* (prominence) which is, of course, a *local* phenomenon. A typical place for relaxation is *the end of an utterance*;

Table 1. Different Functions of Laryngealizations

phenomenon	**time domain**
linguistic functions: phonotactics, grammar, ...	
accentuation	local
vowels	local
word boundaries	local
native language	local
the end of an utterance, i.e., turn-taking	local
paralinguistic functions: speaker characteristics	
speaker idiosyncrasies	local - global
speaker pathology	global
too many drinks / cigarettes	temporary
competence / power	global / temporary
social class membership	local / global / temporary
emotional states such as boredom, sadness, etc.	short-term or temporary

by that, *turn-taking* can be signalled to the dialogue partner; this is again a *local* phenomenon: [10] report that different types of LAs are used in (British and American) English conversations for holding the floor (filled pauses with glottal closure, no evidence of creaky phonation) and for yielding the floor (filled pauses with lax creaky phonation, no glottal closure). *Word boundaries* in the hiatus, i.e. word final vowel followed by word initial vowel, can be marked by LAs. Boundary marking which is, of course, *local*, with such irregular phonation is dealt with in [11] and [12]. It is well known that back *vowels* such as *[a]* tend to be more laryngealized than front vowels such as *[i]* (*local* phenomenon). A language-specific use of LAs can be either due to phonotactics, as in German, where every vowel in word-initial position is 'glottalized', or phonemes can be creaky, cf. [13]; this is a *local* phenomenon, denoting the *native language*. Normally, specific segments which are laryngealized characterize languages, cf. for vowels [14]; the Danish glottal catch (stød) [15] can be found in vowels and consonants.

[16] p. 194ff. lists different uses and functions of 'creak' phonation, amongst them the paralinguistic function 'bored resignation' in English RP, 'commiseration and complaint' in Tzeltal, and 'apology or supplication' in an Otomanguean language of Central America. Extra- and paralinguistically, LAs can be a marker of personal identity and/or social class; normally, LAs are a marker of higher class speech. [17] quote evidence that not only for human voices but for mammals in general, 'non-linear phenomena' (i.e. irregular phonation/LA) can denote individuality and status (pitch as an indicator of a large body size and/or social dominance; *"... subharmonic components might be used to mimic a low-sounding voice"*).

Note that all these characteristics which per se are **not** characteristics of single speakers can — maybe apart from the language-specific phonemes —

be used more or less distinctly by different speakers. As for the paralinguistic function of LAs, speakers can simply use them throughout to a higher extent; such *speaker idiosyncrasies* are *local - global.* 'Creaky superstars' like Tom Waits are well-known. The reason might be unknown, or due to one or more of the following factors: *speaker pathology* (*global*), *too many drinks/cigarettes* (*temporary*), *competence/power* (*global / temporary*), or *social class membership* (*local/global/temporary*).

Emotional states such as *despair, boredom, sadness*, etc. are *temporary.* Bad news are communicated with breathy and creaky voice [18], boredom with lax creaky voice, and to a smaller extent, sadness with creaky voice [19]. [20] report for perception experiments with synthesized stimuli that disgust is conveyed with creaky voice. To display boredom or to display upper-class behaviour might coincide; the same can happen if someone who permanently uses LAs as speaker-specific trait, speaks about a sad story. On the other hand, at first sight, speakers who exhibit LAs as an idiosyncratic trait can make a sad impression without actually being sad.

The caveat has to be made that we are speaking of a sort of 'cover phenomenon' covering different sub-phenomena and different temporal traits: some are very short and might rather be perceived as segmental features, i.e. not as supra-segmental, prosodic features that are sort of modulated onto the speech wave. Of course, there are prototypical cases — no LA at all and laryngealized throughout — which easily can be told apart. But we simply do not know yet when people will produce which amount of LA and how an automatic classifier can model it.

It might be safer to find out non-existing/low correlations such as high pitch and fast speech with sadness. Further functions of LAs are reported in [21]. There are only a few studies dealing with the automatic detection of LAs, cf. [22, 23]. We have manually corrected automatically extracted F0 values for one third of the database described in [4, 5] (51 children giving commands to Sony's pet robot Aibo). For some 6% of all voiced frames of all words, we found gross F0 errors denoting LAs; this amounts to some 14.7% words with laryngealized passages. The percentage of laryngealized words per speaker ranges from 0% to 35%; this illustrates a strong speaker dependency. At first sight, the distribution across emotional user states denotes more LAs in emotions with negative valence (*angry*, *touchy* (i.e. irritated), and *reprimanding*) than with neutral or positive valence. This could be a plausible result if we equate indicating negative valence with indicating some kind of superiority. This difference, however, disappears if we compute the distribution separately for words with the initial diphthongs [aU] and [aI] which are prone to be laryngealized more often than other vowels and diphthongs. The reason why is that in our database, some of these words – e.g. the vocative ['?aIbo] – are relatively frequent in the negative valence domain. Note that by that, we did of course not prove that LAs do not signal some emotional states, especially because in our data, emotions such as *sadness* (cf. the database processed in [24]) or *boredom* were not found. We can illustrate, however, the multi-functionality and speaker-dependency of LAs; thus it might be less likely that they are very useful as a generic feature within emotion classification. This might of course be different in a personalized setting.

4 Another Example: Pitch

Pitch is multi-functional, maybe up to the same extent as laryngealizations are. People can speak with flat F0 or with marked ups and downs — this is a personality trait. In the high days of intonation models, pitch was held responsible for the marking of word- and sentence accents, of salience, etc. During the last years, however, it has been shown that F0 is of minor importance, in relation to other parameters such as energy and duration, cf. [25–27] and [28]. The same might be true for emotion recognition; again, we do not know yet whether this might be due to pitch simply being less important, or to a combination of extraction errors, speaker specific traits and its bi-dimensionality which is difficult to model, esp. with sparse data.

The manual correction of the database mentioned at the end of section 3 resulted in some 6% gross F0 errors; first experiments on emotion classification with manual corrected F0 values yielded for a four-class problem some 3.5% better classification performance than with automatically extracted – i.e. sometimes erroneous – F0 values. Such a difference which is not very pronounced at first sight might, however, denote a difference between 'somehow relevant' and 'most relevant' feature types.

5 Implications from Applications

As nowadays the automatic recognition of emotion is getting more important for the voice business, several new questions are coming up. Since this single recognition process must become part of a business solution providing voice activated services to customers, we have to deal with integration and performance aspects, we have to figure out how the emotion recognition module can access data, and where the result is needed in which format; i.e. we have to care about interfaces. However, most important is that we have to know about the overall goal of the voice application in general, and we have to get an idea how this goal can be supported using emotion recognition. Last but not least do we have to find a compromise between technical and scientifical implementations on the one hand and budget and time restrictions on the other hand.

Here we will report on experiences we made in a project where emotion recognition was integrated as a component in a voice portal. A detailed description of the system, its capabilities and the results can be found in [29], [30] and [31]. Note that here we do not give details on recognition results, data sets, etc. since we want to concentrate on the fact that in integrated applications, a lot of constraints play a role and influence the emotion recognition, one would not think of in advance.

5.1 Voice Application Setup

Applying emotion recognition for business applications in the first step generates questions miles away from technological and functional aspects concerning the internals of the classifier. These questions concentrate on the business process the emotion recognition should be applied to, the other components working together, and performance and interface issues. For example, if you apply emotion recognition in a telephone based speech dialogue system, you have to care about the performance of the complete system since it has to be avoided that the caller has to wait too long for the system reaction. Thus all the processing has to be performed very fast, so that the system's response is generated within a period of at most 200 milliseconds after the caller stopped speaking.

But let us discuss these problems by means of the concrete project mentioned above. The voice application setup looks like the following: People ring up the voice application which is an information system. In addition to the usual technical components necessary for such an application – speech recognition, dialogue management and speech synthesis – new components for emotion recognition have to be deployed and integrated; constraints based on the specific architecture will be discussed below. In a first step we have to decide what we are really looking for: is it 'general anger' we want to classify in the speech signal or are we simply looking for a situation where the caller's emotional state changes for the worse? Actually, the second situation is the one we are interested in. Then, the speech dialogue system can react in a predefined manner, e.g. try to calm down the caller, transfer her to an agent or, if all agents are currently talking, give her a higher priority in the queue so that she will be transferred earlier than her position in the line would suggest. In this context the assumption usually is that the user behaves neutrally — at least in the first phase of the dialogue. Later on, either if the system makes recognition errors or gives displeasing information (e.g. a high telephone bill, a negative account balance), the caller might loose his good temper and thus change the communication style. Therefore, it is not highly important for the emotion recognition system to detect anger 'per se' in the speech signal but to be able to find those points in the spoken dialogue where the caller's emotional state changes to the worse. Thus the basic setup of the voice application has important implications for the classification task. If the task of the speech dialogue systems is different from the one presented above, it is perhaps very important to classify right from the beginning of the conversation whether the caller is angry or not.

In our example the task clearly is to detect changes in the emotional state of the user so that in case of a change for the worse, the system can apply different strategies to de-escalate. This already points towards a classification algorithm where features are used that characterize the changes in the acoustics between the current utterance and a reference utterance. This reference could either be the first user utterance in the dialogue, the preceeding utterance in the dialogue, or even perhaps an 'average' utterance computed from previous dialogues. The last procedure requires that the system is able to identify the caller, e.g. by means of analysing the calling telephone number, and to have access to a database where these references are stored. We applied such a kind of differential feature extraction algorithm for the project and compared these so called delta features with a classification algorithm using absolute features. The results reported in [29], [30] and [31] show that the delta features clearly outperform the absolute features on a data set from the described setup. The features used are prosodic features based on energy and F0 values, and duration features based on the segmentation of the speech signal into voiced and unvoiced regions. Actually, this should not be taken as proof that differential features based on prosody are generally more appropriate for emotion recognition; however, in the given case and under the constraints described above, they perform better and are the right choice for this task.

These results encourage to have a more detailed look on personalization and on those features dealt with in sections 2.4, 3 and 4; due to restrictions in time and budget, this has, however, not been possible for this specific project.

5.2 System Architecture

In this section we will have a detailed look at the constraints imposed by the chosen system architecture, the applied modules and the existing interfaces, and, especially, the features employed: as sketched above, we have a regular telephony based speech application using a speech recognizer, a dialogue management component, and a synthesis module; these three elements are standard modules also employed in other voice application environments. In our case we have to add the following components: two emotion recognition modules, one working exclusively on the recognized word chain (e.g. looking for swearwords) and the other one using only the speech signal as data source to compute its decision. Additionally, there is a decision module which takes the results of the two emotion recognizers and merges them into one classification result which is handed over to plan the necessary reaction. The speech recognition module is a standard product from the market with a predefined set of different interfaces and functions. Unfortunately, with these interfaces it is not possible to access all necessary (desired) information. It is for instance not possible to get the time alignment for the best word chain from the recognition engine, we do not have access to the features computed from the speech signal, and it is even not possible to get the incoming speech signal incrementally. Thus we have to wait until the end of the user's input before the waveform is accessible by other modules.

Looking at this system architecture, some interesting questions arise:

- Why do we use an 'off-the-shelf' recognizer engine with that many unwanted side effects?
- Why are there two separate emotion recognition modules, one using only acoustic, the other one only linguistic information?
- What would be a more appropriate system architecture and processing?

The application was planned to be installed and to go online either with or without emotion recognition. Basically, the operator of this information hotline wanted to have this specific automatic speech dialogue system. After in-depth discussion, they agreed with emotion recognition as additional component, because of the possible benefits. Nevertheless, they required that the resulting system should also work properly without emotion recognition and that it has to meet their internal administrative requirements. There were already other speech applications running based on the VoiceXML standard (cf. www.w3c.org/voice or www.voicexml.org for detailed information), using specific components and system architectures. Hence, the new system had to be based on this standard architecture, with the modules already in use in this company. The use of our own speech recognition module which would allow to have access to time alignment and spectral (MFCC) features was thus not possible.

If we look at the system architecture, the imposed restrictions, and the demand that the dialogue system has to react immediately after the user stopped speaking, it is obvious that the feature extraction for the emotion recognition does not have that much time. Additionally, at that time point when the emotion recognizer can start working, a recognized word chain is almost already available from the standard recognizer. Therefore it makes sense to separate the linguistic emotion classification from the acoustic processing; this is one of the reason why there are two emotion modules in the resulting system. Actually, the linguistic component is more or less ‘integrated’ in the recognition engine since the used grammars have to model also utterances containing swearwords and other phrases expressing emotion.

As for the acoustic emotion recognition, time constraints made it necessary to look for features and classification procedures which operate rather fast. Budget restriction made it impossible to spend additional time on the implementation of new types of features. Thus, we decided in favour of an already existing feature set based on energy, F0 values, and duration features based on the segmentation of the speech signal in voiced and unvoiced regions; as for details, cf. [31]. From the speech signal we computed one feature vector of defined length, usually a mixture of absolute and delta features. We decided to apply a rather simple Gaussian mixture model (GMM) approach for classification. For the training of the individual components of this GMM, five students manually annotated the utterances; for each utterance, a majority voting was applied. All this resulted not in an optimal but in a very good solution — given the constraints addressed above — for this specific application.

6 Concluding Remarks

In this paper, we dealt with those factors that are, in our opinion, most relevant for the — suboptimal — state of the art in emotion recognition: results obtained using acted speech and/or perception experiments with synthesized speech cannot be transferred onto real-life data; the sparse data problem prevents us from having enough training data both for speaker-independent and esp. for speaker-dependent modelling of spontaneous, real-life data; even if in theory, applications as described above could provide us with optimal classifiers and enough data for training, constraints imposed by time and budget prevent this. Of course, there are many other possible applications for emotion recognition [32, 33], not only the call center scenario dealt with in this paper, which might impose other (types of) constraints on the implementation of an emotion recognition module.

A possible marking of one specific type of emotional state can be superimposed or hampered by at least these factors: several linguistic and paralinguistic functions such as given in Table 1, and by some extraction errors. Emotions are temporary phenomena and should be signalled not only locally at some specific (phonotactic) positions (cf. the linguistic functions in Table 1), and not globally as in the case of some paralinguistic functions. It might be possible to disentangle these functions on the time domain — but only with a personalized, speaker-dependent modelling. As is, the normal strategy in emotion recognition to classify speaker-independently short stretches of speech (at least syllables, sometimes words, most of the time phrases or turns/utterances) is possibly severely impaired because it is, at the time of the classification, not clear whether the marker is due to linguistic/paralinguistic factors, or to the signalling of emotions.

For many of the statements given above, there are no hard facts yet to prove or to invalidate them. Single studies will not do, converging results are the only way.

7 Acknowledgments

This work was partly funded by the EU in the framework of the NoE HUMAINE under grant IST-2002-507422, and by the German Federal Ministry of Education and Research (*BMBF*) in the framework of the SmartWeb project (Grant 01IMD01F). The responsibility for the contents of this study lies with the authors.

References

1. Cowie, R., Cornelius, R.: Describing the emotional states that are expressed in speech. Speech Communication **40** (2003) 5–32
2. Schuller, B., Müller, R., Lang, M., Rigoll, G.: Speaker Independent Emotion Recognition by Early Fusion of Acoustic and Linguistic Features within Ensembles. In: Proc. 9th Eurospeech - Interspeech 2005, Lisbon (2005) 805–808
3. Labov, W.: The Study of Language in its Social Context. Studium Generale **3** (1970) 30–87
4. 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: Proceedings of IS-LTC 2006, Ljubliana (2006) 240–245
5. Batliner, A., Steidl, S., Hacker, C., Nöth, E., Niemann, H.: Tales of Tuning – Prototyping for Automatic Classification of Emotional User States. In: Proc. 9th Eurospeech - Interspeech 2005, Lisbon (2005) 489–492
6. Schuller, B., Seppi, D., Batliner, A., Meier, A., Steidl, S.: Towards more Reality in the Recognition of Emotional Speech. In: Proc. of ICASSP 2007, Honolulu (2007) to appear.
7. Scherer, K.: Vocal communication of emotion: A review of research paradigms. Speech Communication **40** (2003) 227–256
8. Poggi, I., Pelachaud, C., de Carolis, B.: To Display or Not To Display? Towards the Architecture of a Reflexive Agent. In: Proceedings of the 2nd Workshop on Attitude, "Personality and Emotions in User-adapted Interaction", User Modeling 2001. (2001) 7 pages, no pagination
9. Batliner, A., Burger, S., Johne, B., Kießling, A.: MÜSLI: A Classification Scheme For Laryngealizations. In House, D., Touati, P., eds.: Proc. of an ESCA Workshop on Prosody. Lund University, Department of Linguistics, Lund (1993) 176–179
10. Local, J., Kelly, J.: Projection and 'silences': notes on phonetic and conversational structure. Human Studies **9** (1986) 185–204
11. Kushan, S., Slifka, J.: Is irregular phonation a reliable cue towards the segmentation of continuous speech in American English? In: Proc. of Speech Prosody 2006, Dresden (2006) 795–798
12. Ní Chasaide, A., Gobl, C.: Voice Quality and f_0 in Prosody: Towards a Holistic Account. In: Proc. of Speech Prosody 2004, Nara, Japan (2004) 4 pages, no pagination.
13. Ladefoged, P., Maddieson, I.: The Sound of the World's Languages. Blackwell, Oxford (1996)
14. Gerfen, C., Baker, K.: The production and perception of laryngealized vowels in Coatzospan Mixtec. Journal of Phonetics (2005) 311–334
15. Fischer-Jørgensen, E.: Phonetic analysis of the stød in standard Danish. Phonetica **46** (1989) 1–59
16. Laver, J.: Principles of Phonetics. Cambridge University Press, Cambridge (1994)
17. Wilden, I., Herzel, H., Peters, G., Tembrock, G.: Subharmonics, biphonation, and deterministic chaos in mammal vocalization. Bioacoustics **9** (1998) 171–196
18. Freese, J., Maynard, D.W.: Prosodic features of bad news and good news in conversation. Language in Society **27** (1998) 195–219
19. Gobl, C., Ní Chasaide, A.: The role of voice quality in communicating emotion, mood and attitude. Speech Communication **40**(1-2) (2003) 189–212

20. Drioli, C., Tisato, G., Cosi, P., Tesser, F.: Emotions and Voice Quality: Experiments with Sinusoidal Modeling. In: Proceedings of VOQUAL'03, Geneva (2003) 127–132
21. Ishi, C., Ishiguro, H., Hagita, N.: Using Prosodic and Voice Quality Features for Paralinguistic Information Extraction. In: Proc. of Speech Prosody 2006, Dresden (2006) 883–886
22. Kießling, A., Kompe, R., Niemann, H., Nöth, E., Batliner, A.: Voice Source State as a Source of Information in Speech Recognition: Detection of Laryngealizations. In Rubio Ayuso, A., López Soler, J., eds.: Speech Recognition and Coding. New Advances and Trends. Volume 147 of NATO ASI Series F. Springer, Berlin (1995) 329–332
23. Ishi, C., Ishiguro, H., Hagita, N.: Proposal of Acoustic Measures for Automatic Detection of Vocal Fry. In: Proc. 9th Eurospeech - Interspeech 2005, Lisbon (2005) 481–484
24. Devillers, L., Vidrascu, L.: Real-life Emotion Recognition in Speech. In Müller, C., ed.: Speaker Classification. Volume 4343 of Lecture Notes in Computer Science / Artificial Intelligence. Springer, Heidelberg - Berlin - New York (2007) this issue.
25. Batliner, A., Buckow, J., Huber, R., Warnke, V., Nöth, E., Niemann, H.: Prosodic Feature Evaluation: Brute Force or Well Designed? In: Proc. of the 14th Int. Congress of Phonetic Sciences. Volume 3., San Francisco (1999) 2315–2318
26. Batliner, A., Buckow, J., Huber, R., Warnke, V., Nöth, E., Niemann, H.: Boiling down Prosody for the Classification of Boundaries and Accents in German and English. In: In Proc. 7th Eurospeech, Aalborg (2001) 2781–2784
27. Batliner, A., Möbius, B.: Prosodic Models, Automatic Speech Understanding, and Speech Synthesis: Towards the Common Ground? In Barry, W., Dommelen, W., eds.: The Integration of Phonetic Knowledge in Speech Technology. Springer, Dordrecht (2005) 21–44
28. Kochanski, G., Grabe, E., Coleman, J., Rosner, B.: Loudness predicts Prominence; Fundamental Frequency lends little. Journal of Acoustical Society of America **11** (2005) 1038–1054
29. Burkhardt, F., van Ballegooy, M., Englert, R., Huber, R.: An emotion-aware voice portal. In: Proc. Electronic Speech Signal Processing ESSP. (2005)
30. Burkhardt, F., Stegmann, J., Ballegooy, M.V.: A voiceportal enhanced by semantic processing and affect awareness. [34] 582–586
31. Huber, R., Gallwitz, F., Warnke, V.: Verbesserung eines Voiceportals mit Hilfe akustischer Klassifikation von Emotion. [34] 577–581
32. Batliner, A., Burkhardt, F., van Ballegooy, M., Nöth, E.: A Taxonomy of Applications that Utilize Emotional Awareness. In: Proceedings of IS-LTC 2006, Ljubliana (2006) 246–250
33. Burkhardt, F., Huber, R., Batliner, A.: Application of Speaker Classification in Human Machine Dialog Systems. In Müller, C., ed.: Speaker Classification. Volume 4343 of Lecture Notes in Computer Science / Artificial Intelligence. Springer, Heidelberg - Berlin - New York (2007) this issue.
34. Cremers, A.B., Manthey, R., Martini, P., Steinhage, V., eds.: INFORMATIK 2005 - Informatik LIVE! Band 2, Beiträge der 35. Jahrestagung der Gesellschaft für Informatik e.V. (GI), Bonn, 19. bis 22. September 2005. In Cremers, A.B., Manthey, R., Martini, P., Steinhage, V., eds.: GI Jahrestagung (2). Volume 68 of LNI., GI (2005)