An Evaluation of Emotion Units and Feature Types for Real-Time Speech Emotion Recognition

Thurid Vogt · Elisabeth André

Abstract Emotion recognition from speech in real-time is an upcoming research topic and the consideration of real-time constraints concerns all aspects of the recognition system. We present here a comparison of units and feature types for speech emotion recognition. To our knowledge, a comprehensive comparison of many different units on several databases is still missing in the literature and we also discuss units with special emphasis on real-time processing, that is, we do not only consider accuracy but also speed and ease of calculation. For the feature types, we also use only features that can be extracted fully automatically in real-time and look at which types best characterise which emotion classes. Gained insights are used as validation of methodology for our online speech emotion recognition system EmoVoice.

1 Introduction

The consideration of realistic scenarios has come to the fore in recent years in research on speech emotion recognition. For online processing in real-time, which is needed for many realistic applications of emotion recognition, all aspects of the recognition system have to be tuned to a compromise between accuracy and speed. Until now, little attention in research has been paid to emotion units as a design decision for applications, because they arose naturally in acted speech as the single recorded utterances or, in more natural settings, as (manually segmented) dialogue turns. However, when doing online speech emotion recognition consistent and meaningful appropriate units have to be found. For this reason, we present here a systematic comparison of many units and also discuss the units with respect to their online capabilities. Furthermore, we investigate the relevance of various feature types for different tasks by analysing multiple databases of acted as well as realistic emotions which should enable as to achieve very generally valid results. In order to face realistic conditions, we rely only on features that can be calculated automatically and fast exclusively from the acoustic signal. The insights gained on best units and relevant feature types are used for our own online speech emotion recognition system [40] which is part of our open-source emotion recognition toolkit EmoVoice.[1]

In the following, we will first identify possible emotion units and feature types and discuss their usage in the existing literature. Then, we describe the three database that we base our experiments on. Afterwards we present our investigations on emotion units as well as features and conclude with the insights gained from these experiments.

2 Emotion Units

The first step in an automatic speech emotion recognition system is to segment the audio input signal into meaningful units to later derive the actual features from acoustic measurements of those units. The units are usually linguistically

T. Vogt (✉) · E. André
Lab for Human Centered Multimedia, Augsburg University,
Universitätsstraße 6a, 86159 Augsburg, Germany
e-mail: vogt@hcm-lab.de

E. André
e-mail: andre@hcm-lab.de

[1] http://hcm-lab.de/EmoVoice.html.

motivated medium-length time intervals such as utterances (e.g. [16, 19, 30] or words [1, 38]. Though the decision on which kind of unit to take is evidently important, it had initially not received much attention. Most approaches so far have dealt with 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. However, in spontaneous speech this kind of obvious unit does not exist. Neither is the segmentation into utterances straightforward nor can a constant emotion be expected over an utterance. Generally speaking, a good emotion unit has to fulfil certain requirements. In particular, it should be (1) well-defined to be consistently extracted, (2) long enough so that features can reliably be calculated by means of statistical functions (if—as is here—the approach of global statistics features is taken), (3) short enough to guarantee stable acoustic properties with respect to emotions within the segment, (4) consistent with the labelling of the training database. The first point is important because the segmentation in training, test and application should be subject to the same rules, that is, it must have the same characteristics. Thereto the rules for segmentation need to be unequivocally defined. For example, the notion "sentence" does exist for spontaneous speech, however, a segmentation into sentences is often not feasible and if, it is often ambiguous. So a sentence would not be a good unit for spontaneous speech, though it is useful for read speech.

When, as is here and in many other work, the feature extraction approach of calculating global statistics for given time segments is taken, the classification 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 for instance the mean value of a very inhomogeneous segment yields an inadequate description. Thus, a compromise has to be found for these two conflicting requirements. Furthermore, for a given training database, arbitrary units cannot be used, since emotion labels are biased to some extent if they do not exist for exactly this segmentation. Hence a comparison of units on different databases must be drawn with care. For example, if the database is labelled on turn level, not necessarily every word in the turn has this emotion. Some might be neutral if considered individually, especially short words without emphasis. In reverse, careful consideration is needed to derive a turn label from word labels, for example by simple majority voting over word labels. It is often better to reduce the influence of neutral words.

As we aim here at a real-time self-contained emotion recognition system which does not necessarily require further knowledge, for example about words and word boundaries, a further advantage of a good unit would be that it can be automatically computed from the audio signal alone. An alternative to linguistically motivated units can thus be frames with a fixed length, for example 0.5 or 1 seconds. Furthermore, a unit can be considered with its context, that means for example for a word, to consider the preceding and succeeding word(s) as well [1]. In the context of dialogue systems, for which emotion recognition is especially applicable, whole dialogue turns can also serve as emotion classification units. The length and nature of turns, however, strongly depend on the dialogue system. In order to have a unit in spontaneous speech that approximates utterances in acted speech a—manual or automatic—boundary detection can be carried out. A strategy for automatic boundary detection is to segment by pauses, that is sections of low signal energy, that are at least 0.2 to 1 seconds long. For this purpose, voice activity detection algorithms can be used, as breaks in voice activity can mark the boundaries [40]. An example of manually revised boundary detection is the syntactic and prosodic chunking of the FAU Aibo Emotion Corpus [34] which is also described later in this paper. So far, there are some studies comparing various units [4, 23, 27, 28, 33, 38, 39], however, most of them are limited to one database only. Among these, [28] examined fixed time intervals and also combined features based on multiple time scales into a super-vector. These time scales consisted of the full utterance as well as first, second and third, or first, central and last 500 milliseconds. Fixed time intervals turned out to perform worse than utterances, but are very useful in an online application context. The combination of multiple time scales which is also an approach to incorporate timing information exceeded the performance of utterances, showing that timing is evidently relevant for emotion recognition. [4] introduced the "ememe" as smallest meaningful emotional unit and compare words, syntactic chunks and sequences of ememes belonging to the same emotion class on one database, the latter giving best results but is not easily usable in practise. In an earlier study, we compared utterances, words, words in context and fixed time intervals [39] and found longer units to perform better. Generally speaking, it strongly depends on the data and purpose which unit fits best. So far, however, a comprehensive comparison of many types of units on several databases is missing, that is why we investigate this question here with special emphasis on real-time processing.

3 Feature Extraction

Common features for speech emotion recognition are based on short-term acoustic observations like pitch or energy.

Since the specific values of these measures are usually not too expressive *per se*, but rather their change over time, the modelling of the temporal behaviour 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 (e.g. [29, 41, 43]). Thus, a classification unit consists of a series of feature vectors and obtains 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 them-self, 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 (e.g. [3, 8, 19]). The latter approach is the one followed here. Of course, there is a huge number of possibilities how to transform value sequences to single values. Our approach to feature extraction is for the most part a generative one, which means that we systematically apply statistical functions to derive features from basic observations. By this, we calculate a multitude of possibly relevant features starting from the basic feature types of which the most relevant ones for a specific database or application scenario can be selected. Other examples for such an approach are [9, 25]. In the following, we discuss the feature types that were used for the experiments described later. These are pitch, energy, duration, spectral, cepstral, voiced segments (derived from pitch) and voice quality features. Our main goal is to find and use fast and fully automatic algorithms for feature extraction and classification even if that may degrade recognition accuracy to some extent. For example, we do not use word information as this would require a speech recognition step to precede the emotion recognition. Other approaches comparing feature types include [5, 21]. However, as with units, these were not conducted on different types of databases.

Pitch: Pitch is often considered to be most important for emotion perception. It is included in almost all emotional speech feature sets e.g. [1, 8, 10, 17]. However, though it does definitely have some importance for emotions, its influence is probably not as huge as typically assumed. Generally, a rise in pitch is an indicator for higher arousal, but also the course of the pitch contour reveals information on affect.

Energy: The energy curve depends on many factors, such as phonemes, speaking style, utterance type (e.g. declarative, interrogative, exclamatory), but also on the affective state of the speaker [1, 8, 10, 17]. Again, like pitch, high energy roughly correlates with high arousal, but also variations of the energy curve give hints on the speaker's emotion.

Duration: Timing certainly plays a huge role in the expression of emotion [1, 17, 30]. This concerns the duration of speech units like utterance length or average word length in an utterance, but also the speaking rate. Speaking rate can be measured e.g. by the word or syllable rate, if word information is available. Furthermore, the distribution of voiced and unvoiced segments from pitch calculation approximates speaking rate. Speaking rate and duration itself but also various temporal patterns of words or utterances can help to distinguish between emotional user states. For example, [20] derive rhythm features from vowel duration. Similarly, the duration and distribution of pauses is significant as it makes a difference whether few long pauses or many short pauses occur.

Spectrum: The spectrum of frequencies occurring in a speech signal is also influenced by the affective state of the speaker. Examples for features derived from the spectrum are the spectral slope [14], spectral entropy [18], spectral centre of gravity [32], the ratio of the spectral flatness to the spectral centre of gravity [16], or log frequency power coefficients [20, 24]. Most of these features encode the distribution and weighting of frequencies in the spectrum.

Cepstrum: The most prominent cepstral feature type are MFCCs [7], but also Linear Prediction Cepstral Coefficients (LPCCs) can be found [24]. MFCCs are evaluated in feature sets with other feature types, but they are also relatively often used as sole feature type. Compared to most other types, they have a greater potential to be self-sufficient because they are a parametric representation incorporating many different aspects of the speech signal. MFCC calculation for the feature set used here was again obtained from ESMERALDA [11].

Voiced segments: The length and distribution of voiced and unvoiced segments, as calculated by the pitch algorithm, in a speech signal is related to voice characteristics. These have not been explored as emotional speech features so far.

Voice quality: Relations of voice quality to emotions are diverse, for instance, breathiness may result from excitement, harshness from anger, or a frightened speaker might whisper. Voice quality can be measured in several ways, jitter, shimmer and harmonics-to-noise ratio (HNR) being the most frequent ones in automatic classification approaches. Jitter and shimmer measure the variability in distance and amplitude of the glottal pulses. HNR can be computed on the whole signal or on small parts of it. It relates the energy of the harmonic parts of a signal to the energy of the noise parts and thus measures the pureness of the voice. Though it is intuitively a good indicator for emotions, voice quality features are not part of standard feature sets for emotion recognition, but have been used seldom so far [22, 36].

Feature selection: As already mentioned, in most work with global statistics features a selection process is applied to reduce the dimensionality of the feature space. Feature selection is beneficial for efficiency reasons on the one hand, as a smaller set with comparable results is preferred over a larger set, because training as well as classification times are shorter. On the other hand, in practise a feature selection can actually increase performance, because with the learning algorithms used here, an addition of bad, redundant or correlated features may even deteriorate accuracy. On small training data sets, the "curse of dimensionality", or the effect of overfitting may occur. Furthermore, feature sets should be optimised for the respective application scenario, since it is assumed that good feature sets differ depending on the data type. The most popular search method is sequential (floating) forward selection [17, 30]. [25] alternatively use a genetic search, [37] select those features that significantly modify a linear regression model. [12] showed mutual correlation of features as selection criterion to be inferior in accuracy, but much faster than sequential selection methods.

4 Databases

Three databases were used for evaluation purposes. The first one, the Berlin database of emotional speech, is a database of acted read emotions, the other two were recorded in Wizard-of-Oz settings. The SmartKom database has only few emotions and contains speech of adults, while the FAU Aibo Emotion Corpus is more emotional and contains speech of children. All three databases are German, however, all methods evaluated here work language-independently. While classification on the Berlin database is a relatively easy task, it is not very realistic as it contains acted speech obtained under ideal acoustic conditions, a scenario one would scarcely find in an application. Furthermore, it is limited in size. The Aibo and the SmartKom database are much harder tasks, because emotions are not as prototypical and clear, but they are larger and close to realistic conditions. Thus, by evaluating these three databases, that are described in detail in the following, a wide variety of emotions is covered and results can be expected to be largely general.

Berlin Database of Emotional Speech: The Berlin Database of Emotional Speech was recorded at the Technical University of Berlin, Germany [6], and is analysed very often in speech emotion recognition studies. It contains acted emotional German speech of ten carefully chosen speakers (5 male, 5 female) that were asked to pretend six different emotions (anger, joy, sadness, fear, disgust and boredom) as well as a neutral state in ten utterances each of emotionally neutral content. They are characterised by a very high audio quality. After the recordings a listening test was performed with 20 human subjects who should recognise the emotion of every utterance and rate it for its naturalness. Those utterances from the collected material that were misclassified by more than 20% of the test persons or perceived as unnatural by more than 40% were discarded, ending up with 493 utterances (female: 286/male: 207). As can be seen in Table 1 the distribution of emotions on the utterances is, except for anger, relatively equal. The available word boundary information was labelled manually. The database is evaluated here by 5-fold speaker-independent cross-validation leaving always two speakers out (one male, one female).

Table 1 The distribution of emotions in the utterances of the Berlin Database of Emotional Speech

	Joy	Anger	Fear	Disgust	Boredom	Sadness	Neutral	$\sum$
#	64	127	55	38	79	52	78	493

Table 2 The distribution of emotions in those turns of the FAU Aibo Emotion corpus containing at least one AMEN word

	Angry	Motherese	Emphatic	Neutral	$\sum$
#	867	487	1334	1307	3995

FAU Aibo Emotion Corpus: The FAU Aibo Emotion Corpus was recorded at FAU Erlangen, Germany, in a Wizard-of-Oz (WoZ) setting within the EU project PF-Star [34]. It contains about 9.2 hours of speech from 51 children (female: 31, male: 20) in the age of 10 to 13 years interacting with Sony's robot dog Aibo. Each word was annotated by five independent raters with the labels joyful, surprised, emphatic, helpless, touchy/irritated, angry, motherese, bored, reprimanding, rest, and neutral. Word segmentation was obtained from an automatic speech recogniser and then corrected manually. Experiments described here are based on a subcorpus that consists of the turns containing at least one AMEN word as defined for the CEICES initiative [3]. AMEN words are only those words labelled with the four most frequent classes **A**ngry (formed by merging touchy, reprimanding and angry), **M**otherese, **E**mphatic and **N**eutral by at least three of the five labellers. The classes emphatic and neutral were down-sampled to obtain a more balanced distribution of emotions. Thus, the subcorpus based on the turns of the AMEN words finally contains 3995 turns. Table 2 shows the distribution of emotions. In addition to word boundary and turn information, there also exists a semi-automatically obtained segmentation into chunks for the whole Aibo corpus considering pauses and syntactic boundaries. As evaluation strategy for this corpus three-fold cross-validation was chosen, and splits were created with gender, school and emotion distribution (given in order of priority) conserved as much as possible as in [31].

SmartKom Corpus: The SmartKom Corpus was acquired in a Wizard-of-Oz setting at the University of Munich, Germany [26], within the SmartKom project, the goal of which it was to develop a multimodal dialogue system for three different scenarios: a *public* information interface, a *home*-based and a *mobile* communication assistant. 222 subjects were recorded in 447 sessions. Each session was ca. 4.5 minutes in length, however, the speech part is much less. Be-

Table 3 The distribution of emotions in the headset-recorded sessions within the mobile scenario of the SmartKom database

	Positive	Neutral	Negative	$\sum$
#	70	1579	167	1816

cause of better quality, only recordings made with a headset microphone are used here for analysis which restricts the data to 126 sessions of the mobile scenario with 66 speakers (37 female and 34 male). The age ranged from 10 to 65 years, while most speakers were between 12 and 27 years old. Videos were also recorded and supported the later annotation of emotions. Apart from naturally occurring emotions occasionally emotions were elicited by disfunctions of the system such as leading a subject through a movie reservation process and revealing at the last step that this function was not available now. While the occurring emotions can be considered quite realistic, the biggest part of the speech is emotionally neutral. Consequently, this corpus represents a great challenge for automatic emotion recognition based on speech information only. In SmartKom, the emotional user states joy/gratification, surprise, pondering/reflecting, helplessness, anger/irritation, and neutral were labelled [35] based on the video recordings. Therefore, in contrast to the other two databases labels in SmartKom are not based on linguistic units such as words or utterances. Following a scheme from [2], we merged joy and surprise into one class "positive", and pondering, helplessness and anger into one class "negative", thus ending up with three classes positive, neutral, negative. The distribution of emotions, which is obviously very unequal, can be seen in Table 3. Information on word boundaries was available from an automatic speech recogniser. Again, this corpus was evaluated with three-fold speaker-independent and gender-balanced cross-validation.

5 Results

We will now describe the experiments we conducted to compare emotion units and feature types for our EmoVoice system. The classifier used in all experiments here is the Naïve Bayes classifier, which we found to be a simple, but powerful algorithm for the task of emotion classification and especially suited for real-time processing. Furthermore, we do not want to lay emphasis on classification at this point. Results are always given as averaged accuracy for each class, as overall accuracy may give false readings of the results because of unbalanced class distribution in two of the databases. Feature selection sets were always obtained on the training set for each cross-validation cycle.

Table 4 Emotion units explored for the three databases

Unit	Berlin	Aibo	SmartKom
Fixed length 0.5 s	√	√	√
Fixed length 1 s	√	√	√
Fixed length 2 s	√	√	√
Automatic pause segmentation by VAD	√	√	√
Word	√	√	√
Word in context (± 1 word)	√	√	√
Syntactic/prosodic chunks	–	√	–
Pause segmentation by ASR	–	–	√
Utterance	√	–	–
Turns	–	√	√

5.1 Comparison of Emotion Units

The first experiment was conducted to compare various types of emotion units in view of their usefulness for different kinds of data, notably fixed length units, words, utterances, segments marked by pauses, and turns. An overview of all units examined in three analysed databases can be found in Table 4 and the units will now be described in detail: Starting with the non-linguistic units, three durations of fixed length units were tested: 0.5, 1 and 2 seconds. These were chosen because units of less than 0.5 seconds were considered as too short for the calculation of statistical measures, while changes of the emotional state may well occur in units longer than 2 seconds. Lastly, as an approximation of a linguistic unit, speech parts segmented by breaks in the voice activity detected automatically and on the acoustic signal only by the voice activity detection (VAD) algorithm integrated into the ESMERALDA framework for automatic speech recognition with HMMs [11], are investigated. All these units are very suitable for an integrated online system.

Linguistically motivated units that we analysed comprise words, words in context, manual and automatic speech recognition (ASR) assisted pause segmentation as well as utterances and dialogue turns. Words are often very short, that is why they are also investigated in the context of one preceding and succeeding word and the potential silent or nonverbal part in between. Furthermore, the difference of adjacent words within an utterance with respect to their emotional tone will scarcely be huge. As higher-level linguistic units, chunks, utterances and turns are examined. On the Aibo database, chunks were obtained by a manually revised detection of syntactic and prosodic boundaries triggered by main clauses, free phrases and between successive occurrences of the word "Aibo", as these repetitions are likely to mark a change in the emotional state [31]. Prosodic boundaries were set when pauses between words exceeded 0.5 seconds. In SmartKom, chunks were defined by pauses longer than 0.5 seconds as detected by an ASR system, which is a fully automatical procedure. However, only a few emotionally neutral turns were affected by this segmentation. Chunking is not available for the Berlin database, instead, utterances were used. Finally, within dialogue turns changes of the emotional state have to be expected, but the classifier has to decide on only one emotion per turn. Turns containing more than one emotion are probably not acoustically homogeneous with respect to emotions so they may produce lower recognition rates. Concerning real-time issues, it is relevant to look at the extraction time of the units. Obviously, fixed length units and, in many cases, turn segmentation, which may be inherent to the resp. application, need least computational effort. VAD depends only on the signal energy and is thus also computationally cheap. Word, word in context and pause segmentation by automatic speech recognition all require considerably more time because the complex algorithms of an automatic speech recognition system have to be deployed. Last, semi-automatic syntactic/prosodic chunking of course cannot be done in real-time by a machine only.

An important point is how labels, which were always available for only one unit per database, are mapped onto other units. Labels in the Berlin database arise from that emotion that the actors were asked to pretend in the superordinate utterance. For the Aibo database a more fine grained label mapping strategy can be applied since labels are available on word level from each of the five annotators. We based our mapping on Steidl's [34] strategy. So, for word and 0.5 seconds, a majority voting of all available labels was used. The label of a word in context is the label of its central word. For VAD based units, chunks and 1.0 seconds, a unit was labelled as neutral, if at least 60% of the votes were neutral; as motherese, if there were at least as many votes for motherese than for emphatic or angry; as angry, if there were equal or more votes for angry than for emphatic; and as emphatic in the remaining cases. For turns and 2.0 seconds, the same strategy was used, but the threshold for neutral was set to 70%. Neutral is treated in a special way because for a whole chunk to be perceived emotional, not all words, especially function words, need to be uttered emotionally. In the case of fixed length and VAD units, labels are weighted by the number of samples occurring in the unit. In the SmartKom database, labels were derived by simple majority voting. If the relative majority of a fixed-length unit pertained to silence, the unit was not used for classification. Words shorter than 0.1 seconds were also discarded. Of course these label mapping strategies are all compromises, for example in the Berlin database, not each word in the utterance may reflect the whole utterance's emotion. Units where the strategy fits more often thus may receive a better evaluation than others even if they are not better per se. However, emotional speech databases come labelled on one or at most two emotion units, so there is just no data available for a better comparison. Furthermore, the mapping strategies are assumedly suitable for most of the segments.

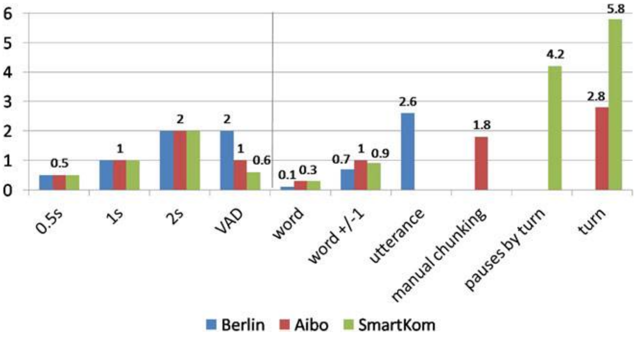

Fig. 1 Average duration of units in seconds for the three databases

Figure 1 shows the average length of each unit. Words are shortest in Berlin, and VAD based units are shorter in SmartKom and Aibo as in Berlin. The reason for that is that Berlin contains read speech: the shorter word length can be explained by a higher speaking rate as speakers do not have to plan what to say and by the limited vocabulary of the 10 sentences which does not contain extraordinarily long words. Furthermore, when reading one makes fewer pauses than when speaking freely and spontaneously, so that there are less breaks in the voice activity. Thus VAD based units are longer in Berlin than in Aibo or SmartKom. The speech in SmartKom is especially characterised by pondering with many pauses so it is no surprise that here, VAD units are shortest. It can further be observed that VAD units are shorter than utterances or chunks (manually or by ASR), so they seem to be rather on a segmental level between these and words. Comparing automatic with linguistic units in respect to length, 1.0 s approximately matches the word in context of ±1 word, and 2.0 s approximates utterances or manual chunking. A word is even shorter than 0.5 s.

Table 5 shows recognition results for the different segmentation units for the three databases. For the Berlin database, the difference in results between the units is most dramatically: the best unit (utterances) is almost twice as good as the worst (words). This may be due to the very short average word length: 0.1 s is obviously too short for global statistics features. A dynamic classification approach might give better results here. In the SmartKom database, words also score worse than longer units, though the difference is not as big. In contrast, results for words and chunks on the Aibo corpus are very similar, but here it is important to notice that the labelling was based on words. With regard to the question whether turns are too long because they may often contain several emotions, results indicate that this effect is not huge as they score quite good in both Aibo and SmartKom. VAD units on average come off best which argues in favour of not overly long units. Even though this unit is rather targeted at non-prompted speech as in Aibo or

Table 5 Comparison of different segmentation levels by averaged class accuracy in %. Best and insignificantly worse results for each database in bold

Unit	Berlin	Aibo	SmartKom
0.5 s	44.4	47.5	43.9
1 s	57.6	48.7	44.0
2 s	64.9	48.8	42.4
VAD	60.0	**51.0**	**48.5**
Word	38.8	51.0	44.3
Word ± 1	51.6	**50.7**	45.1
Chunks	–	**50.3**	–
Pauses by ASR	–	–	45.1
Utterance	**73.4**	–	–
Turn	–	49.3	46.1

SmartKom, VAD units score worse only than utterances and 2.0 s segments in the Berlin database. So they apparently do not give bad results on prompted speech either and are very suitable for online recognition at the same time because they are fast to extract. For this reason, VAD is chosen as unit in the experiments on features in the following section.

Concluding, the best type of unit is different in each database. Apart from the characteristics of the speech data (read, spontaneous), this is probably due to the labelling, which is always based on only one unit and is different in each database.

5.2 Feature Type Evaluation

In this section, we aim to evaluate the significance of the feature types described in Sect. 3, especially with respect to their ability to predict particular classes which is assumed to be different for different data types. For this end, we composed a feature set on all three databases using VAD based units, in which all features can be extracted fast, reliably and fully automatically. Most features were extracted as one of 9 statistical functions (mean, minimum, maximum, range, standard deviation, median, first quartile, third quartile and interquartile range) from basic acoustic observation series or transformations of them in a generative way. For pitch, these series were the raw pitch, the logarithmised pitch, the normalised pitch by subtracting the median from the logarithmised pitch values [15], transformations of these as the series of the local maxima, the local minima the difference, distance and slope between adjacent local extrema as well as first and second derivation. Additionally, the unlogarithmised pitch mean, median, first and third quartile values were normalised by minimum and maximum pitch of the respective segment. Further pitch features were the position of the overall pitch maximum, which approximates the main chunk accent, and the position of the overall pitch minimum. As indicators for pitch contours, the number of minima, maxima, falling and rising values were obtained. All these values were normalised by the number of pitch values in the segment.

From energy, as for pitch, the series of only the local maxima and only the local minima were created, as well as difference, distance and slope between adjacent local extrema, furthermore first and second order derivation together with the series of their local maxima and local minima. Also, the position of the global maximum and the number of local maxima, both normalised by the number of frames in the segment, were calculated. Duration features fell out of the generative approach. They include the chunk length, measured in seconds, the zero-crossing rate to roughly decode speaking rate, and pause as the proportion of non-speech calculated by the voice activity detection algorithm from the signal energy [11] and also approximated by the ratio of unvoiced pitch frames to the total number of pitch frames in the unit. With respect to spectral features, especially information on the slope of the spectrum was regarded as important, so for each short-term spectrum, the distance between the 10th and the 90th percentile, the slope between weakest and strongest frequency, the centre of gravity as well as two linear regression coefficients were calculated by ordinary least-square estimation. Each of these 5 values yielded a new 1-dimensional time series, to which the statistical functions listed earlier were applied. For MFCC related features, 12 coefficients and their average, plus their first and second derivatives were calculated as time series, and also transformed into the series of local maxima and minima. Voiced segments were considered further as counts of the lengths of both the voiced and the unvoiced segments in a unit as time series, and additionally, the number of voiced segments normalised by the number of pitch frames in the chunk was calculated. Voice quality was modelled by jitter and shimmer of the glottal pulses of the whole segment, and the number of glottal pulses normalised by the segment length in seconds. Furthermore, statistics were derived from the Harmonics-to-Noise ratio. A final summary of the number of features per type can be found in Table 6.

One problem with the experiment is that the number of features is very unequal for different types (see Table 6). A large number can be positive for a feature type, as the chance of catching meaningful properties is higher, or negative, if redundant features interfere with each other. In order to make the feature set sizes a bit more equal and to restrict the sets to meaningful and discriminative features, correlation based feature subset selection (CFS, [13]) from the Weka data mining toolkit [42] was carried out for each type. The goal of CFS is to find a subset where the correlation of each feature with the class is maximised, while the correlation of the features among each other is low. This strategy is especially beneficial for the Naïve Bayes classifier which performs bad when features are highly correlated since it assumes features to be independent for simplification reasons. Each database is analysed with automatic voice activity detection as unit as this was found to be the best compromise between accuracy and fast computation in the previous comparison of emotion units and is thus especially suited for our goal of real-time emotion recognition.

Table 7 shows the results of feature type evaluation obtained for the Berlin database. Anger, neutral, disgust and sadness achieve high accuracies while joy and fear seem difficult to distinguish from other classes. Anger can be detected very accurately by all feature types except for voicing and pitch, by the latter, however, with at least medium accuracy. Neutral is recognised well by pitch, MFCC, duration and spectral features, but achieves low accuracies with the other feature types. Energy features contribute most to the recognition of disgust, whereas for sadness, MFCC features achieve the highest accuracy. Boredom is discriminated moderately well by pitch and MFCC features. Voicing features are the best indicator for joy, pitch and MFCC for sadness, though for both classes, only low accuracies can be achieved. From the perspective of the feature types, not of the classes, it can be observed that pitch detects neutral better than other classes, while energy is a good indicator for anger and disgust, so rather for negative emotions with high activation, but not for other classes. MFCCs have the broadest aptitude and are best for neutral, anger and sadness. Duration features are mainly good at detecting anger, spectral features at detecting neutral and anger. Voicing features seem to have discriminating ability only for joy, and voice quality for anger. With regard to the recognition rates for all classes, only MFCCs classify correctly in more than half of the cases, though all feature types classify above chance level. Duration, voicing and voice quality feature are, however, considerably worse than the other feature types. Obviously, MFCCs have the best generalising power though they do not perform best in each class. A positive result is that each feature type contributes to the recognition accuracy, though for most classes, single types stand out, while most perform bad. In the Aibo database (see Table 8), the classes angry, neutral and motherese can be recognised very well, only emphatic is difficult to detect. Voice quality, pitch and MFCC features are relevant for anger; for emphatic, only MFCCs give useful results. Neutral is best discriminated by duration and voicing features which in turn have considerably lower scores for the other classes. However, due to the unbalanced class distribution in the Aibo database, it is

Table 6 Number of features per types

Feature type	Number of features
Pitch	208
Energy	110
Duration	4
Spectrum	45
Cepstrum	1053
Voiced segments	19
Voice quality	12
$\sum$	1451

Table 7 Feature type comparison for Berlin by individual and mean class accuracy, best or insignificantly worse results for each class in bold

Berlin	Pitch	Energy	MFCC	Duration	Spectral	Voicing	Voice quality	All
Joy	38.6	12.3	42.1	5.3	15.8	**47.4**	10.5	*49.1*
Neutral	**67.2**	34.4	**68.8**	55.7	60.7	6.6	13.1	*80.3*
Anger	51.6	71.1	64.8	72.7	**77.3**	16.4	**78.1**	*70.3*
Fear	**40.7**	11.1	**40.7**	1.9	20.4	29.6	3.7	*53.7*
Disgust	21.8	65.2	43.5	6.5	21.7	13.0	15.2	*56.5*
Sadness	26.2	36.9	**63.1**	40.5	46.4	16.7	34.5	*41.7*
Boredom	**51.5**	32.4	**54.4**	23.5	20.6	17.6	20.6	*64.7*
Average	42.5	37.6	**53.9**	29.4	37.6	21.1	25.1	*59.5*

Table 8 Feature type comparison for Aibo by individual and mean class accuracy, best or insignificantly worse results for each class in bold

Aibo	Pitch	Energy	MFCC	Duration	Spectral	Voicing	Voice quality	All
Angry	74.2	42.5	61.3	22.1	34.0	26.8	**79.7**	*59.0*
Emphatic	27.9	37.0	**49.5**	32.8	33.6	29.3	14.6	*44.5*
Neutral	26.8	29.1	44.9	**77.5**	28.3	73.7	12.2	*42.5*
Motherese	6.5	63.8	45.3	0.0	**74.5**	5.3	18.7	*58.3*
AA	33.9	43.1	**50.3**	33.1	42.6	33.8	31.3	*51.1*

Table 9 Feature type comparison for SmartKom by individual and mean class accuracy, best or insignificantly worse results for each class in bold.

SmartKom	Pitch	Energy	MFCC	Duration	Spectral	Voicing	Voice quality	All
Positive	4.8	8.1	12.1	0.0	**26.6**	14.5	0.0	*19.4*
Neutral	86.9	68.5	65.3	95.9	49.5	72.3	**96.8**	*67.4*
Negative	19.1	57.0	58.3	14.2	**68.7**	23.2	6.3	*58.6*
AA	36.9	44.5	45.2	36.7	**48.3**	36.6	34.4	*48.5*

hard to tell whether these features are especially suited to recognise neutral voices or whether they just classify in the most frequent class. Motherese is recognised best by spectral and energy features. Again looking at the feature types, pitch, MFCCs and voice quality detect anger well, energy and spectral motherese, duration and voicing neutral. Each feature type excels in only one class. Overall, again MFCCs are best (again with about half of the instances classified correctly), followed by energy and spectral features. The problem of unbalanced class distribution is even more serious in the SmartKom database (see Table 9). Especially the difference in recognition rate for the most and least frequent classes, neutral and positive, is extreme. Results for neutral and the overall recognition rate are therefore not overly meaningful. For positive emotions, spectral features are by far the best, but still below chance level. Spectral features, and to a lesser extent also MFCC and energy features, are most important for negative emotions. The importance of spectral features is noticeable. Again, MFCCs show the best performance over all classes.

Comparing all three databases, MFCC features generally prove to be the most descriptive type of features. However, it is hard to say whether this is due to the quality of the features or just to their high number (even after correlation analysis). Furthermore, each feature type contributes at least at one point so it is not wise to drop any type completely. Neither is it possible to make a general statement valid across database types which feature types are especially suited for certain emotions. For example, voice quality is important for anger in the Berlin and Aibo databases, but not for negative in SmartKom. Apparently, however, the less emotional the data is, the higher is the importance of MFCCs. With an ideal fusion scheme relying always on the best feature type for each class, even a higher result could be obtained by a multi-level classification split into feature types than with all types together. Of course, relevant features and feature types do not only depend on the emotional classes that are considered; for other databases, units or classifiers, different features may be relevant than were found here.

6 Conclusion

In this paper, we investigated strategies for emotion units and feature type evaluation. With respect to emotion units, to our knowledge no such comprehensive and systematic comparison of units for emotion recognition from speech has been conducted so far. We found good units to some extent to be dependent on the particular database which we think is mainly due to what the labelling was based on. In general, however, accuracy increased with the length of the unit. Furthermore, our comparison showed that non-linguistic units may also have very good performance, though so far, rather the opposite has been implicitly assumed, as there are almost no studies using such units. This positive result holds at least under the specific conditions (e. g. databases, features, classifier, evaluation strategies) of the experiments here. In particular, VAD based units have not yet been systematically examined and we found them to compare very well to traditional units in all three databases. We think that they are especially suited for online processing and therefore use them for our online emotion recognition framework EmoVoice.

The feature type evaluation showed that the relevance of types is indeed very database dependent. While for read, acted speech, pitch features were very relevant, the less emotional the data was, the higher was the importance of MFCC and spectral features. Evaluating each class individually, MFCCs also proved to be the most generally successful feature type across databases. But, as MFCCs are very general and actually intended to filter out non-linguistic influences in speech, there is still more potential in the search for good feature types. The characteristics of each class were usually best described by one or two feature types only while all others were considerably less suited and all feature types were most relevant for at least one class. This lets us argue in favour of using many feature types as they capture very different aspects of emotions and we integrated all feature types described here into our EmoVoice framework.

Acknowledgements This work was partially funded by the EC in the Integrated Projects CALLAS (IST-34800) and CEEDS (258749) and the IRIS Network of Excellence (231824).

References

1. Batliner A et al (2003) How to find trouble in communication. Speech Commun 40(1–2):117–143
2. Batliner A et al (2003) We are not amused—but how do you know? User states in a multi-modal dialogue system. In: Proc of INTERSPEECH, Geneva, Switzerland

3. Batliner A et al (2006) Combining efforts for improving automatic classification of emotional user states. In: Proc of IS-LTC, Ljubljana, Slovenia
4. Batliner A et al (2010) Segmenting into adequate units for automatic recognition of emotion-related episodes: a speech-based approach. Adv Hum Comput Interact. doi:10.1155/2010/782802
5. Batliner A et al (2011) Whodunnit—searching for the most important feature types signalling emotion-related user states in speech. Comput Speech Lang 25:4–28
6. Burkhardt F et al (2005) A database of German emotional speech. In: Proc of INTERSPEECH, Lisbon, Portugal
7. Busso C et al (2007) Using neutral speech models for emotional speech analysis. In: Proc of INTERSPEECH, Antwerp Belgium
8. Devillers L et al (2005) Challenges in real-life emotion annotation and machine learning based detection. Neural Netw 18(4):407–422
9. Eyben F et al (2009) openEAR—introducing the Munich open-source emotion and affect recognition toolkit. In: Proc of ACII, Amsterdam, The Netherlands
10. Fernandez R, Picard RW (2005) Classical and novel discriminant features for affect recognition from speech. In: Proc of INTERSPEECH, Lisbon, Portugal, pp 473–476
11. Fink G (1999) Developing HMM-based recognizers with ESMERALDA. In: Proc of TSD, Plzen, Czech Republic
12. Haindl M et al (2006) Feature selection based on mutual correlation. In: Proc of IBEROAM, Congress on pattern recognit, Cancun, Mexico
13. Hall MA (1998) Correlation-based feature subset selection for machine learning. Master's thesis, University of Waikato, Hamilton, New Zealand
14. Huang R, Ma C (2006) Toward a speaker-independent real-time affect detection system. In: Proc of ICPR, Hong Kong, China
15. Kießling A (1996) Extraktion und Klassifikation prosodischer Merkmale in der automatischen Sprachverarbeitung. PhD thesis, University Erlangen-Nuremberg, Germany
16. Kim EH et al (2009) Improved emotion recognition with a novel speaker-independent feature. IEEE/ASME Trans Mechatron 14(3):317–325
17. Lee CM, Narayanan S (2005) Toward detecting emotions in spoken dialogs. IEEE Trans Speech Audio Process 13(2):293–303
18. Lee WS et al (2008) Speech emotion recognition using spectral entropy. In: Proc of ICIRA, Wuhan, China
19. Litman D, Forbes-Riley K (2006) Recognizing student emotions and attitudes on the basis of utterances in spoken tutoring dialogues with both human and computer tutors. Speech Commun 48(5):559–590
20. Luengo I et al (2009) Combining spectral and prosodic information for emotion recognition in the INTERSPEECH 2009 emotion challenge. In: Proc of INTERSPEECH, Brighton, UK
21. Lugger M, Yang B (2007) An incremental analysis of different feature groups in speaker independent emotion recognition. In: Proc of ICPhS, Saarbrücken, Germany, pp 2149–2152
22. Lugger M, Yang B (2009) On the relevance of high-level features for speaker independent emotion recognition of spontaneous speech. In: Proc of INTERSPEECH, Brighton, UK
23. Mansoorizadeh M, Charkari NM (2007) Speech emotion recognition: Comparison of speech segmentation approaches. In: Proc of IKT, Mashad, Iran
24. Nwe TL et al (2003) Speech emotion recognition using hidden Markov models. Speech Commun 41(4):603–623
25. Oudeyer PY (2003) The production and recognition of emotions in speech: features and algorithms. Int J Hum-Comput Stud 59(1–2):157–183
26. Schiel F et al (2002) The SmartKom multimodal corpus at BAS. In: Proc of LREC, Las Palmas, Gran Canaria, Spain
27. Schuller B, Devillers L (2010) Incremental acoustic valence recognition: an inter-corpus perspective on features, matching, and performance in a gating paradigm. In: Proc of Interspeech, Makuhari, Japan
28. Schuller B, Rigoll G (2006) Timing levels in segment-based speech emotion recognition. In: Proc of INTERSPEECH, Pittsburgh, PA, USA
29. Schuller B et al (2003) Hidden Markov Model-based speech emotion recognition. In: Proc of ICASSP, Hong Kong, China
30. Schuller B et al (2005) Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles. In: Proc of INTERSPEECH, Lisbon, Portugal
31. Schuller B et al (2007) The relevance of feature type for the automatic classification of emotional user states: low level descriptors and functionals. In: Proc of INTERSPEECH, Antwerp. Belgium
32. Sethu V et al (2009) Pitch contour parameterisation based on linear stylisation for emotion recognition. In: Proc of INTERSPEECH, Brighton, UK
33. Shami MT, Kamel MS (2005) Segment-based approach to the recognition of emotions in speech. In: Proc of ICME, Amsterdam, The Netherlands
34. Steidl S (2009) Automatic classification of emotion-related user states in spontaneous children's speech. Logos Verlag, Berlin
35. Steininger S et al (2002) Development of user-state conventions for the multimodal corpus in SmartKom. In: Proc of LREC workshop on multimodal res and multimodal syst eval, Las Palmas, Gran Canaria, Spain
36. Sun R et al (2009) Investigating glottal parameters for differentiating emotional categories with similar prosodics. In: Proc of ICASSP, Taipei, Taiwan
37. Tato R et al (2002) Emotional space improves emotion recognition. In: Proc of INTERSPEECH, Denver, CO, USA
38. Vlasenko B et al (2008) Balancing spoken content adaptation and unit length in the recognition of emotion and interest. In: Proc of INTERSPEECH, Brisbane, Australia
39. Vogt T, André E (2005) Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition. In: Proc of ICME, Amsterdam, The Netherlands
40. Vogt T et al (2008) EmoVoice—a framework for online recognition of emotions from voice. In: Proc of PIT, Irsee, Germany, pp 188–199
41. Wagner J et al (2007) A systematic comparison of different HMM designs for emotion recognition from acted and spontaneous speech. In: Proc of ACII, Lisbon, Portugal
42. Witten IH, Frank E (2005) Data mining: practical machine learning tools and techniques. Kaufmann, San Francisco
43. Yun S, Yoo CD (2009) Speech emotion recognition via a max-margin framework incorporating a loss function based on the Watson and Tellegen's emotion model. In: Proc of ICASSP, Taipei, Taiwan

Thurid Vogt is a software developer at BMW AG, Munich. Prior to that she worked as a researcher in several EU projects at the University of Augsburg and had a graduate scholarship of the DFG at the University of Bielefeld. In 2010, she obtained her Ph.D. in computer science from the University of Bielefeld.

Elisabeth André is full professor of Computer Science at Augsburg University and Chair of the Laboratory for Human-Centered Multimedia. Prior to that, she worked as a principal researcher at DFKI GmbH where she has been leading various academic and industrial projects in the area of intelligent user interfaces.