Hindawi Publishing Corporation
Advances in Human-Computer Interaction
Volume 2010, Article ID 782802, 15 pages
doi:10.1155/2010/782802

Research Article

Segmenting into Adequate Units for Automatic Recognition of Emotion-Related Episodes: A Speech-Based Approach

Anton Batliner,[1] Dino Seppi,[2] Stefan Steidl,[1] and Björn Schuller[3]

[1] *Pattern Recognition Laboratory, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), D-91058 Erlangen, Germany*
[2] *ESAT, Katholieke Universiteit Leuven, B-3001 Leuven, Belgium*
[3] *Institute for Human-Machine Communication, Technische Universität München (TUM), D-80333 Munich, Germany*

Correspondence should be addressed to Anton Batliner, batliner@informatik.uni-erlangen.de

Received 1 April 2009; Accepted 12 December 2009

Academic Editor: Elisabeth Andre

We deal with the topic of segmenting emotion-related (emotional/affective) episodes into adequate units for analysis and automatic processing/classification—a topic that has not been addressed adequately so far. We concentrate on speech and illustrate promising approaches by using a database with children's emotional speech. We argue in favour of the word as basic unit and map sequences of words on both syntactic and "emotionally consistent" chunks and report classification performances for an exhaustive modelling of our data by mapping word-based paralinguistic emotion labels onto three classes representing valence (positive, neutral, negative), and onto a fourth rest (garbage) class.

1. Introduction

It is not only difficult to define "emotion," it is difficult as well to find out where an emotional episode—whatever it is—begins and where it ends. It is difficult both for theoretical reasons—in order to know where it begins and ends, we have to know what it is—and for methodological/practical reasons as well, which we will detail below. By and large, studies on emotion have bypassed this topic by dealing with episodes delimited by external criteria.

1.1. The Phenomena: Emotions or …. There is definitely no agreement on an extensional or intensional definition of "emotion"—or of any other term that could be used instead such as affect, attitude, and mood, to replace it or to denote similar phenomena that have to be told apart from the core semantics of this term. The core phenomena consist of the big *n* emotions such as despair, anger, joy—*n* being some figures between 4 and 8 or more; this concept is mainly rooted in psychology, and has been challenged, elaborated, and extended amongst others in cognitive psychology, for instance in the OCC model [1]. Perhaps "the" alternative concept is a wider definition, encompassing all the fringe phenomena that are "present in most of life but absent when people are emotionless" [2]; this is the concept of *pervasive emotion*, which often implicitly forms the basis of engineering approaches in a sometimes vague use of the term, addressing states such as interest, stress, and boredom.

These fiercely disputed terminological debates are, however, not relevant for our topic. Segmentation for dealing with such para-linguistic phenomena is pivotal—no matter which definition we use to describe them. They are related to but are not by definition coextensive with linguistic units such as sentences, utterances, dialogue acts, and salience. We will elaborate on this topic in the next subsection. To prevent fruitless debates, we use the rather vague term "emotion-related episodes" in the title to denote both emotions in a strict sense and related phenomena in a broader sense, which are found in the database our experiments are based on. In the text, we will often use "emotion" as the generic term, for better readability. This resembles the use of generic "he" instead of "he/she;" note, however, that in our context, it is not a matter of political correctness that might make a more cumbersome phrasing mandatory, it is only a matter of competing theoretical approaches, which are not the topic of the present paper.

Implicitly, it is normally taken for granted that these states to be modelled are produced and not only perceived.

This difference can be illustrated by the following example: a father can get really angry with his son, and this can be heard in his tone of voice and seen in the outcome of physiological measurements. He can, however, only pretend being angry—because as father, he has to, even if he perhaps likes his son's "misbehaviour." In such a case, we can hear the change in his tone of voice but most likely we will not be able to measure marked physiological changes. The son might notice such a "fake" emotion—if he is clever and has experienced it often enough—or not. We cannot assume that machines are clever enough to notice. Thus, the emotion-related states we are dealing with have to be taken as "perceived" *surface* phenomena, at face value—at least as long as we do not employ physiological measurements, trying to find out a real ground truth. (Strictly speaking, physiological measurements are most likely closer to—but not necessarily constituting—any ground truth.)

The components of speech are vocal expression and linguistic content. Both components can be employed for signalling denotations and semantics, and for constituting illocutions (such as dialogue acts), and for expressing connotations as paralinguistic messages (such as emotions). The same scenario as above can illustrate this usage: the father can get really angry with his son, but instead of expressing his anger in his tone of voice, he simply can say, in a low and calm voice: "*Now I'm really getting angry.*" It could be argued that this is a describing "meta" statement and not an indication of "real" anger. However, the son will be well advised to react as if the father had expressed "real" anger in his tone of voice as well. Moreover, it cannot be argued that this is not an indication of negative valence—note that in this paper, we map our raw labels onto main classes representing positive, neutral, or negative valence. Again, the son can take this at face value and stop his misbehaviour, or he can misconceive his father's anger as pretense because it is not expressed in the father's tone of voice. Again, machines should not try to be too clever; the only possibility they have is to take the linguistic content of the user's utterances at face value.

Thus, both vocal and linguistic expression of emotions should be taken by machines along the lines of Grice's cooperative principle, at face value, and not assuming any indirect use [3]; this excludes, for example, irony, metaphor, and meiosis. (It is sometimes claimed that irony can be recognised by a system; this will never work under real-life conditions, at least not in the foreseeable future.) In this vein, we will employ both acoustic and linguistic features for the automatic classification of emotion-related user states.

1.2. The Need for Segmentation. In this paper, we want to address different possibilities to segment emotional episodes. We will concentrate on speech but, at the same time, we want to argue that in many applications to be imagined, speech will possibly be the best modality to base segmentation upon; of course, this only holds for speech to be found together with other modalities. As has been noted in [4], for all modalities the segmentation into emotion units is one of the most important issues if we aim at real applications but has been "largely unexplored so far." The same holds for another, equally important aspect: even if most authors agree that it is high time to go over from acted data to realistic, spontaneous data, normally, only a subset of the full database is used consisting of somehow clear, that is, more or less prototypical cases. This is not only a clever move to push classification performance, it simply has grown out from the problem of class assignment in emotion processing: there is no simple and unequivocal ground truth. We have elaborated on the use of prototypes and their impact on classification performance in [5, 6].

In the transition from read speech to spontaneous speech in Automatic Speech Recognition (ASR), normally all data have been employed apart from, for instance, nonlinguistic vocalisations, which are treated as garbage; but they are still treated and not removed from the signal before processing. Note that a rough estimate for the difference between read and spontaneous data was that, at least at the beginning, one could expect an error rate for spontaneous data twice the size than the one for read data [7]. Of course, we cannot simply transfer this empirically obtained estimate onto Automatic Emotion Recognition (AER). Yet we definitely will have to deal with a plainly lower classification performance. However, this constitutes the last step before AER, including full ASR and automatic segmentation, really can be used "in the wild," that is, in real applications.

As mentioned in the last subsection, we use "emotion" in a broad sense, following the definition of "pervasive emotions" in the Network of Excellence HUMAINE [2], where emotion is defined as the absence of non-emotion. Thus, it is a foreground-background phenomenon: emotional has to be different from "not emotional", that is, from *neutral* (emotionally *idle*). However, in the four different modalities that have been mostly investigated in emotion studies—speech, vision (i.e., facial gestures), body posture and movements, and physiology—telling apart emotionally idle from emotionally active, that is, not-idle, poses different questions because the modalities themselves behave in a different way: speech versus non-speech is easy to tell apart—at least for a human being. Note that of course, such statements have to be taken with a grain of salt: depending on the signal-to-noise ratio, it can be difficult; telling apart a non-linguistic vocalisation from a linguistic one can be difficult, sometimes; and for a machine, it is more difficult than for a human listener. Yet to start speaking and to finish speaking is a voluntary act, which normally can be detected and delimited by the listener. However, you cannot start or finish your face—you always have it, even while asleep. And you might be able to control your physiological signals, but only up to a certain extent—you cannot stop your heart beats. (Interestingly, perception in these different modalities is different as well: you cannot stop hearing even while sleeping—you only can use some ear protection decreasing the noise—but you can stop looking by simply closing your eyes, while awake or while sleeping.)

If there was an unequivocal ground truth, at least for the reference data used in automatic processing, we could define begin and end of such episodes easily. However, there is none, irrespective of the modalities. Thus, we have to use and rely

on human annotations or on some external criteria; a well-known example for the latter is taking hanging-up the phone abruptly in a telephone human-machine communication as an indication of anger so we know that there has been some anger before—but we do not know yet whether and where it could be noticed. Arousal might be traced back in physiological signals by defining a threshold criterion based on, for example, Feeltrace annotations [8], but this is way more difficult for valence. Moreover, in many applications, physiological signals cannot be recorded.

We mentioned in the beginning that almost all studies have bypassed somehow the decision where to start or to end an emotion. For decades, the bulk of evidence came from acted data in the lab; for such data, beginning and end are given trivially: either some short episodes had to be produced, for example, using a semantically void utterance as carrier, or longer periods have been integrated. Even in a—more or less realistic—verbal human-human or human-machine communication, a dialogue act/move or an "utterance" can be delimited easily by the act of turn taking, when the one partner finishes speaking and the other partner takes over. As long as such a strategy works sufficiently well, there is no pressure to go over to other criteria. The longer such a unit is, however, the higher is the probability that it does not only entail one emotional episode but two or more, and that it is "smeared," that is, not unequivocal. We can compare this situation with dialogue acts: often, one dialogue move (turn) constitutes just one dialogue act. However, it can consist of sequences of dialogue acts as well; for instance, in appointment scheduling dialogues [9] often sequences of rejection (of the dialogue partner's suggestion), statement (of problems/facts), and suggestion (of alternative dates) can be found.

Several different subunits have been investigated as for their impact on improving classification performance such as frame-based processing, or taking some other fixed interval (percentage of whole utterance, n ms, or voiced/unvoiced decisions, just to mention the most important ones, cf. [10]). But all this has rather been independent from higher processing; yet in a full system such as SmartKom [11] or Semaine [12], time constraints make it mandatory not to wait with ASR and other processing modules until the speaker has finished his/her full turn. For a (close to) real-time processing, it might not matter much whether frames, or syllables, or words, or short chunks are processed; when we assume 1.5 real time, for a short chunk lasting 2 seconds, a user has to wait 3 seconds before a system answer has been generated. This can be tolerated. However, for a turn lasting 10 seconds, the user had to wait 15 seconds—which simply is far too long. Taking any unit below chunk level of course results in even shorter processing time.

We can suppose that emotional changes within a word are difficult to produce and therefore very rare. As a further advantage of word-based processing of emotional speech, we see the better dovetailing of emotion and speech processing. We do not have to align the "emotional time axis" in an additional step with Word Hypotheses Graphs (WHG); for instance, each word in a WHG can be annotated with either its individual emotion label or with the label that has been attributed to the higher unit this specific word belongs to. These are practical considerations; yet it might be plausible conceiving the *word* as the "smallest meaningful emotional unit" as well. We thus can speak of an "*ememe*" in analogy to the phoneme, the morpheme, and especially to the sememe and claim that normally, the word and the ememe are coextensive. A sememe consists of either a morpheme or a word indicating both semantic denotation and/or connotation either encoding a holistic meaning or being constituted by a feature bundle. We introduce the ememe as constituting "pure" connotation, indicated both by acoustic and linguistic means. Such a concept definitely makes sense from a practical point of view; thus, we do not have to care too much whether sometimes it might make sense to go over to subword units. Of course, this only holds for speech; there is no equivalent—at least no one that can be defined and segmented easily—in the other modalities. Note that our "ememe" is the *smallest* emotional unit. The same way as it makes more sense to process meaningful sequences of n words (sememes) constituting something like a syntactically meaningful chunk or a dialogue act, the same way it pays off to combine ememes into higher units. The charm of such an approach is that it is relatively easy to find a word, and where it begins and where it ends. In other modalities, it is way more difficult to delimit units.

In this paper, we want to pursue different emotion units based on speech. We will start with the word, and later combine words into syntactically/semantically meaningful chunks or into consistent "ememe sequences," that is, sequences of words belonging to the same emotion class. By that, we model two different approaches: in the one approach, emotion is sort of modelled as being part of linguistics, in the other one, emotion is an independent layer, in parallel to linguistics. The latter one might be more adequate for theoretical reasons—emotion is not (fully or only) part of linguistics. On the other hand, in a communication conveyed partly or mostly via speech, emotion might really be structured along the speech layer; for instance, the emotional load of content words is normally higher than the one of function words, and the "emotional message" might really be coextensive with dialogue acts. To give an example: laughter is a non-linguistic indication of emotions/user states and often co-occurs with joy. It can be stand alone or modulated onto speech (speech laughter). Laughter and speech laughter are mainly found at the end of syntactic units. This does not necessarily mean that laughter and speech/language are processed and generated in the same module, but it demonstrates a close relationship. Moreover, in an end-to-end system, we always need to align emotion and linguistic processing somehow. However, in this paper, we can only deal with performance measures such as classification rates as criteria. Note that we will not deal with data, use cases, or applications without speech. If we take into account more than one modality we always have to align the unit of one modality with the unit(s) found in the other modality/modalities. Of course, this can be done with some criteria for overlapping on the time axis. At least for the time being, it seems to us that speech, if available, is advantageous over the other modalities to start with.

This is, of course, an assumption that has to be validated or falsified. We can imagine that researchers working in other modalities but speech prefer having their own units of analysis and late fusion of channels [13]. From a theoretical point of view, this might be easier to accomplish; from a practical point of view, it will be a matter of performance, of ease of handling, and—perhaps most important—of the weight a modality has in specific scenarios. Thus, a fall-back solution for our approach is, of course, to use it within a uni-modal speech scenario. Time synchronous or adjacent emotional messages conveyed by different modalities can be congruent or incongruent; speech can even distort the emotional message conveyed via facial gesture or bio-signals because of lip and jaw movements. Moreover, we have to tell apart different types of systems: on the one hand, there are end-to-end systems that take into account emotions as a way of "colouring" the intended message, triggering decisions on part of the dialogue manager in, for instance, call-centre interactions. Here we find a high functional load on speech. On the other hand, there are pure "emotion systems" with a low functional load on speech, for instance in video games—here, "non-verbal" grunts and affect bursts might be more relevant, together with facial gestures.

1.3. Overview. In Section 2, we present the database and the annotations performed, as well as the mapping onto main classes used in this paper. Section 3 presents the units of analysis we want to deal with: we start with the word as basic unit, and then discuss two different types of units, one based on syntactic criteria—to be dovetailed with higher processing modules such as dialogue act processing, the other one simply based on "emotional consistency;" adjacent words belonging to the same class are aggregated within the same unit. In Section 4, we describe the acoustic and linguistic features used in this study, as well as the classifier chosen for this task. Classification results are presented in Section 5 and discussed in Section 6. The paper closes with concluding remarks in Section 7.

2. Database and Annotation

The general frame for our FAU Aibo Emotion Corpus is human-robot communication, children's speech, and the elicitation and subsequent recognition of emotional user states. The robot is Sony's (dog-like) robot Aibo. The basic idea is to combine a so far rather neglected type of data (children's speech) with "natural" emotional speech within a Wizard-of-Oz task. The children were not told to use specific instructions but to talk to the Aibo like they would talk to a friend. They were led to believe that the Aibo is responding to their commands, but the robot is actually being controlled by a human operator, using the "Aibo Navigator" software over a wireless LAN (the existing Aibo speech recognition module is not used). The wizard causes the Aibo to perform a fixed, predetermined sequence of actions, which takes no account of what the child says. For the sequence of Aibo's actions, we tried to find a good compromise between obedient and disobedient behaviour: we wanted to provoke the children in order to elicit emotional behaviour but of course we did not want to run the risk that they break off the experiment. The children believed that the Aibo was reacting to their orders—albeit often not immediately. In fact, it was the other way round: the Aibo always strictly followed the same screen-plot, and the children had to align their orders to its actions.

The data was collected from 51 children (age 10–13, 21 male, 30 female). The children were from two different schools, Mont and Ohm. The recordings took place in a classroom at each school. The child, the wizard, and two supervisors were present. The disjoint school recordings will be used to obtain a natural partitioning into train (Ohm) and test (Mont) in the ongoing. Speech was transmitted with a wireless head set (UT 14/20 TP SHURE UHF-series with microphone WH20TQG) and recorded with a DAT-recorder. The sampling rate of the signals is 48 kHz; quantisation is 16 bit. The data is downsampled to 16 kHz. Each recording session took some 30 minutes. The speech data were segmented automatically into speech files (turns), triggering a turn boundary at pauses ≥ 1 second. Note that here, the term "turn" does not imply any linguistic meaning; however, it turned out that only in very few cases, this criterion wrongly decided in favour of a turn boundary instead of (implicitly) modelling a hesitation pause. Because of the experimental setup, these recordings contain a huge amount of silence (reaction time of the Aibo), which caused a noticeable reduction of recorded speech after raw segmentation; finally we obtained about 8.9 hours of speech.

Five labellers (advanced students of linguistics with German as native language, four females, one male) listened to the speech files in sequential order and annotated independently from each other each word as neutral (default) or as belonging to one of ten other classes, which were obtained by inspection of the data. This procedure was iterative and supervised by an expert. The sequential order of the labelling process does not distort the linguistic and paralinguistic message. Needless to say, we do not claim that these classes represent children's emotions (emotion-related user states) in general, only that they are adequate for the modelling of these children's behaviour in this specific scenario. We resort to majority voting (henceforth MV): if three or more labellers agree, the label is attributed to the word; if four or five labellers agree, we assume some sort of prototypes. The following raw labels were used; in parentheses, the number of cases with MV is given: *joyful* (101), *surprised* (0), *emphatic* (2528), *helpless* (3), *touchy*, that is, irritated (225), *angry* (84), *motherese* (1260), *bored* (11), *reprimanding* (310), *rest*, that is, non-neutral, but not belonging to the other categories (3), *neutral* (39169); 4707 words had no MV; all in all, there are 48401 words. *Joyful* and *angry* belong to the "big" emotions, the other ones rather to "emotion-related/emotion-prone" user states but have been listed in more extensive catalogues of emotion/emotion-related terms, for example, "reproach" (i.e., *reprimanding*), *bored*, or *surprised* in [1]. The state *emphatic* has been introduced because it can be seen as a possible indication of some (starting) trouble in communication and by that, as a sort of "pre-emotional," negative state [5, 14, 15]; note that all the states, especially *emphatic*, have only been annotated when they differed from the (initial) baseline of the speaker.

Table 1: Emotion classes and their word-based frequencies.

Class	No. train	No. test	No. total	% total
P(ositive)	1110	299	1409	2.9
I(dle)	20471	18698	39169	80.9
R(est)	685	550	1235	2.5
N(egative)	3891	2697	6588	13.6
Total	26157	22244	48401	100.0

In this paper, we do not preselect a subcorpus out of the whole database but model valence, that is, positive, idle (neutral), and negative, for all the data; the remaining cases are attributed to a rest class. Thus, we map *motherese* and *joyful* onto *P(ositive)*, *neutral* onto *I(dle)*, *emphatic*, *touchy*, *reprimanding*, and *angry* onto *N(egative)*, and *surprised*, *helpless*, *bored*, and *rest* onto *R(est)*. Cases without an MV were mapped onto *R(est)* as well. The confusion matrices and the subsequent one- and two-dimensional plots based on Nonmetric Multidimensional Scaling (NMDS) in [14], both for labelling correspondences and for confusion matrices based on automatic classification, corroborate these mappings.

Table 1 displays the frequencies of these four classes; interestingly, *I(dle)* versus all other classes is Pareto distributed, that is, 80/20, as was the case for the emotion-related user states in the SmartKom [16] and in the AVIC [17] corpus as well.

Our database might seem to be atypical as it deals with children's speech; however, children represent just one of the usual partitions of the world's population into subgroups such as women/men, upper/lower class, or different dialects. Of course, automatic procedures have to adapt to this specific group—children's speech is a challenge for an Automatic Speech Recognition (ASR) system [18, 19], as both acoustic and linguistic characteristics differ from those of adults [20]. However, this necessity to adapt to a specific sub-group is a frequent issue in speech processing. Pitch, formant positions, and not yet fully developed co-articulation vary strongly, especially for younger children due to anatomical and physiological development [21]. Moreover, until the age of five/six, expression and emotion are strongly linked: children express their emotions even if no one else is present; the expression of emotion can be rather intense. Later on, expressions and emotions are decoupled [22] when children start to control their feelings. Thus so far, we found no indication that our children (age 10–13) behave differently from adults in a *principled* way, as far as speech/linguistics in general or emotional states conveyed via speech are concerned.

3. Units: Words, Syntactic Chunks, and Ememe Chunks

In this section, we present the three units we will deal with in the following: the word (ememe) as basic unit, and as higher units syntactic chunks (SCs) consisting of 1 to n words, and consistent sequences of ememes belonging to the same class, as well consisting of 1 to n words. The word (WO) is a straightforward unit, as is the ememe chunk (EC). There are many different syntactic theories yielding different representations of deep structure. However, we resort to a shallow, surface structure, thus, neutralising many of these differences. We are dealing not with syntactically well-formed speech but with spontaneous, natural speech—this will be the rule and not the exception if we aim at applications in real-life scenarios. For such data, no agreed-upon syntactic representation exists; thus, we have to establish one ourselves, based on the phenomena we can observe in our data.

3.1. Words: WO. Based on the orthographic transcription (transliteration) of the speech data, a lexicon has been compiled consisting of 1146 words, of which 333 are word fragments. Beginning and end of each word were presegmented automatically using a forced alignment of the spoken word chain, and eventually manually corrected. Throughout this paper, we will use this manual segmentation for extracting acoustic features. (Different smaller units of analysis for the FAU Aibo Emotion Corpus were pursued in [23].)

3.2. Syntactic Chunks: SC. Finding the appropriate unit of analysis for emotion recognition has not posed a problem in studies involving acted speech with different emotions, using segmentally identical utterances (cf. [24, 25]). In realistic data, a large variety of utterances can be found, from short commands in a well-defined dialogue setting, where the unit of analysis is obvious and identical to a dialogue move, to much longer utterances, and from syntactically well-defined units to all kinds of spontaneous phenomena such as elliptic speech and disfluencies [26]. In [27] it has been shown that in a Wizard-of-Oz-scenario (appointment scheduling dialogues), it is beneficial not to model whole turns but to divide them into smaller, syntactically and semantically meaningful chunks along the lines of [9]. Our scenario differs in one pivotal aspect from most of the other scenarios investigated so far; there is no real dialogue between the two partners; only the child is speaking, and the Aibo is only acting. Thus, it is not a "tidy" stimulus-response sequence that can be followed by tracking the very same channel; we are using only the recordings of the children's speech. Therefore, we do not know what the Aibo is doing at the corresponding time or has been doing shortly before or after the child's utterance. Moreover, the speaking style is rather special; there are not many "well-formed" utterances but a mixture of some long and many short sentences and one- or two-word utterances, which are often commands. The statistics of the observable turn lengths (in terms of the number of words) for the whole database is as follows: 1 word (2538 times), 2 words (2800 times), 3 words (2959 times), 4 words (2134 times), 5 words (1190 times), 6–9 words (1560 times), ≥10 words (461 times). We see that on the one hand, the threshold for segmentation of 1 s is meaningful; on the other hand, there are still many turns having more than 5 words per turn. This means that they tend to be longer than

one intonation unit, one clause, or one elementary dialogue act unit, which are common in this restricted setting "giving commands."

We observe neither "integrating" prosody as in the case of reading, nor "isolating" prosody as in the case of TV reporters. Many pauses of varying length are found, which can be hesitation pauses—the child produces slowly while observing the Aibo's actions—or pauses segmenting into different dialogue acts—the child waits until he/she reacts to the Aibo's actions. Thus, there is much overlap between two different channels: speech produced by the child, and vision, based on the Aibo's actions, which is not used for our annotation. We therefore decided in favour of hybrid syntactic-prosodic criteria: higher syntactic boundaries always trigger chunking; whereas lower syntactic boundaries do so only if the adjacent pause is ≥500 ms. By that, we try, for example, to tell apart vocatives that simply function as "relators" from vocatives with specific illocutive functions meaning, for example, "*Hi, I am talking to you*" or "*Now I am getting angry*" (illocution "command": "*Listen to me!*").

Note that in earlier studies, we found out that there is a rather strong correlation of more than 90% between prosodic boundaries, syntactic boundaries, and dialogue act boundaries [9]. Using only prosodic boundaries as chunk triggers results in (somehow) worse classification performance (in [9], some five percentage points lower). Moreover, from a practical point of view, it would be more cumbersome to time-align the different units—prosodic, that is, acoustic units, and linguistic, that is, syntactic or dialogue units, based on ASR and higher level segmentation—at a later stage in an end-to-end processing system, and to interpret the combination of these two different types of units accordingly. Preliminary experiments with chunks of different granularity, that is, length, showed that using our longer turns actually results in suboptimal classification performance, while the chunking procedure presented below, which was used for the experiments dealt with in this paper, results in better performance. This might partly result from the fact that more training instances are available, but partly as well from the fact that shorter units are more "consistent," [6].

The syntactic and pause labels are explained in Table 2. For this type of data, we could use a simplified version of the full set of syntactic-prosodic boundaries which is described in detail in [9], for both German and English. Chunk boundaries are triggered by higher syntactic boundaries after main clauses (s3) and after free phrases (p3), and boundaries between vocatives *Aibo Aibo* (v2v1) because here, the second *Aibo* is most likely not simply a relator but is conveying specific illocutions, as discussed above. Single instances of vocatives (v1, v2) are treated the same way as dislocations (d2). If the pauses at those lower syntactic boundaries given in Table 2 (s2, d2, v1, and v2) are at least 500 ms long, we insert a chunk boundary as well. s3 and s2 delimit syntactically "well-formed" clauses containing a verb; p3 characterises not-well-formed units, functioning like clauses but without a verb. d2 is annotated between clauses and some dislocated units to the left or to the right, which could have been integrated into the clause as well. Any longer pause at words within all these units was defined as a nontriggering hesitation pause. Each end-of-turn was redefined as triggering a clause/phrase boundary as well. Note that our turn-triggering threshold of 1 s works well because in the whole database, only 17 end-of-turn (<eot>) triggers were found that obviously denote within clause word boundaries as in the case of "*dieses Mal musst du nach* <eot> *links gehen* <eot>" (*this time you have to go to* <eot> *the left* <eot>). The boundary s1 had to be introduced because *komm* can function both as a sentence initial particle (corresponding to English *well, ...*) as well as an imperative (corresponding to English *come! ...*); only the imperative constitutes a clause.

Table 2: Syntactic and pause labels; frequencies given in Table 3.

label	Description
eot	End-of-turn, recoded as s3 (p3)
s3	Main clause/main clause
s2	Main/subord. clause or subord./subord. clause
s1	Sentence-initial particle or imperative "*komm*"
p3	Free phrases/particles
d2	Dislocations to the left/right
v2	Postvocative
v1	Praevocative
v2v1	Between "*Aibo*" instances
0	Pause 0–249 ms
1	Pause 250–499 ms
2	Pause 500–749 ms
3	Pause 750–1000 ms

With our simple and automatic threshold of 1 s between turns, we obtained turns as long as the one below, which we present for illustration. We denote the chunk triggering boundaries described in Tables 2 and 3 with the symbol "|". In Table 3, rows and columns with chunk triggering labels are shaded in grey. Between angle brackets, first the syntactic, and after the colon the pause labels are given; note that pause length has been corrected manually. For the last part, five chunks are defined. The degree of emotional homogeneity (confidence) given at the end of each chunk is simply the number of the labels the whole chunk is attributed to, divided by the number of all labels. These five chunks in this example belong all to *N(egative)*.

One long turn, German original word sequence with syntactic and pause labels, and chunk boundaries:

> *und stopp* <v1:1> *Aibo* <v2:0> *stehenbleiben* <s3:0> | *darein musst du laufen* <s3:0> | *da links in die Straße* <p3:2> | *so is gut* <v1:0> *Aibo* <v2:0> *und jetzt laufen* <s3:3> | *fein gemacht* <v1:0> *Aibo* <v2:2> | *und* <A> *weiter* <p3:3> <A> | *und jetzt da nach links in die Straße abbiegen* <d2:0> | *zu dem blauen Napf* <s3:1> | *nein* <v1:0> *Aibo* <v2:0> *nein* <p3:0> | *stopp* <v1:1> *Aibo* <v2:0> *nicht* <p3:2> <G> | *nein* <v1:0> *Aibo* <v2:0> *stopp* <s3:0> | *stehenbleiben* <v1:3> | *Aibo* <v2:1> *stehenbleiben* <s3:eot> |

Table 3: Frequencies of syntactic and pause labels in the full database (48401 words), grey rows and columns indicate chunk triggering boundaries; "% tr." displays the percentage of triggering boundaries within the specific row/column.

Label	Pause length				Sum	% tr.
	0	1	2	3		
eot					13642	100
s3	801	407	340	273	1821	100
p3	885	276	183	135	1479	100
s2	165	10	12	10	197	11
s1	328	48	32	28	436	25
d2	56	5	7	2	70	13
v2	3278	498	376	228	4380	14
v1	3226	217	204	178	3825	10
v2v1	20	49	59	69	197	100
Sum	8759	1510	1213	923		
% tr.	19%	48%	100%	100%		

English translation with chunk boundaries:

> *and stop Aibo stand still | go this way | to the left towards the street | well done Aibo and now go on | well done Aibo | and further on | and now turn into the street to the left | to the blue cup | no Aibo no | stop Aibo no | no Aibo stop | stand still | Aibo stand still |*

Last part, German original, emotion labels per word, syntactic and pause labels, chunk boundaries, and "confidence":

> *nein* NNNNI <v1:0> *Aibo* NNNNI <v2:0> *nein* NNNNN <p3:0> 0.87 | *stopp* NNNNN <v1:1> *Aibo* NNNNI <v2:0> *nicht* NNNNI <p3:2> 0.87 | *nein* NNNIN <v1:0> *Aibo* NNNII <v2:0> *stopp* NNNNN <s3:0> 0.80 | *stehenbleiben* NNNNN <v1:3> 1.0 | *Aibo* NNNNN <v2:1> *stehenbleiben* NNNNN <s3:eot> 1.0 |

If all 13642 turns are split into chunks, the chunk triggering procedure results in a total of 18216 chunks. Note that the chunking rules have been determined in a heuristic, iterative procedure; we corroborated our initial hypotheses, for instance, that pauses between adjacent vocatives are longer on average than pauses after or before single vocatives, with the descriptive statistics given in Table 3. The basic criteria have been formulated in [9]; of course, other thresholds could be imagined if backed by empirical results. The rules for these procedures can be automated fully; in [9] multilayer perceptrons and language models have successfully been employed for an automatic recognition of similar syntactic-prosodic boundaries, yielding a classwise average recognition rate of 90% for two classes (boundary versus no boundary). Our criteria are "external" and objective and are not based on intuitive notions of an "emotional" unit of analysis as in the studies by [28–30]. Moreover, using syntactically motivated units makes processing in an end-to-end system more straightforward and adequate.

In order to obtain emotion labels for the chunks, we first mapped the word level decisions of the five labellers (the raw labels) onto the four main classes *P(ositive)*, *I(dle)*, *N(egative)*, and *R(est)*. A whole chunk is considered to be *P(ositive)* if either the absolute majority (≥50%) of all raw labels is positive or if the proportion of positive raw labels is at least one third and the remaining raw labels are mostly neutral, that is, the positive and the neutral raw labels make up at least 90% of all raw labels. By that, chunks that are mostly neutral but where some words clearly signal the subject's positive state are considered to be *P(ositive)* as well. The heuristic thresholds are adjusted by inspecting the resulting chunk labels. *N(egative)* and *R(est)* chunks are defined along the same lines. If according to these definitions a chunk does not belong to one of these three main classes and the proportion of neutral raw labels is at least 90%, the chunk is considered to be neutral, that is, *I(dle)*. If the proportion of neutral raw labels is lower but at least 50% and raw labels of only one other main class appear, the chunk is assigned *I(dle)* as well. These are the cases where single words signal one nonneutral main class but where the proportion of these words is too low. In all other cases, the raw labels belong to too many main classes and the whole chunk is assigned to the *R(est)* class. The frequencies of the four main classes on the chunk level are given in Table 4.

Our word-based labelling makes it possible to try out different types and sizes of chunks. The other way round would be to attribute the same label to a word that the chunk it belongs to has been annotated with. This has two disadvantages: first, there is only one possibility to map chunk labels onto word labels—each word has to be annotated with the chunk label. Thus, we could, for instance, not contrast SC with EC. Second, the result would be "smeared" because of the contra-factual assumption that all words belonging to a chunk necessarily belong to the same emotion class. This can be, but need not be the case. (Of course, sometimes chunking together words belonging to different classes to the rest class, as we do, results in some "smearing" as well—but at least we do know where and up to what extent. Thus, thresholds can be altered and more or less prototypical cases can be established [5, 6].)

3.3. Eme Chunks: EC. The last unit of analysis investigated in this work consists of eme chunks (ECs). ECs are obtained from the eme sequence by clustering together adjacent emes belonging to the *same* main class. An EC is

therefore an i-tuple of ememes characterised by an identical emotional content. In general, from an utterance of n ememes, we can obtain from 1 to n EC. The practical motivation behind EC is that homogeneous groups of ememes might be easier to classify than single ememes; for the recognition of each emotional class, we are exploiting the largest amount of contextual information, that is, the entire EC. From a theoretical point of view, this approach might be most adequate when we model emotional episodes fully independently from linguistic processing.

In the example below, we draw the ECs that belong to the same utterance described in the previous section. Chunks of ememes (denoted as pairs of spoken word and emotion label, that is, "*word* `emotion_label`") are delimited by markers (symbol "|").

Sequence of ememe chunks:

> *und* `I` | *stopp* `R` | *Aibo* `N` *stehenbleiben* `N` | *darein* `I` *musst* `I` *du* `I` *laufen* `I` *da* `I` | *links* `N` | *in* `I` *die* `I` *Straße* `I` *so* `I` **is* `I` *gut* `I` *Aibo* `I` *und* `I` *jetzt* `I` *laufen* `I` *fein* `I` *gemacht* `I` *Aibo* `I` *und* `I` *weiter* `I` *und* `I` *jetzt* `I` *da* `I` *nach* `I` *links* `I` *in* `I` *die* `I` *Straße* `I` *abbiegen* `I` *zu* `I` *dem* `I` | *blauen* `N` | *Napf* `I` | *nein* `N` *Aibo* `N` *nein* `N` *stopp* `N` *Aibo* `N` *nicht* `N` *nein* `N` *Aibo* `N` *stopp* `N` *stehenbleiben* `N` *Aibo* `N` *stehenbleiben* `N` |

Combining ememes into EC is trivial that way; we use a simple finite state automaton. However, this is only the case because our processing is sort of trivial; we are able to map mixed cases onto one class because we have previously performed MV and adopted threshold criteria (Section 2). Taking into account other dimensions or mixtures of annotations [29] would have required a more sophisticated clustering strategy and would not have been feasible for our data, due to the severe sparse data problem. A mixed case in our data—albeit a rather seldom one—is this sequence of words: "*so* `PNNPI` *weit* `PNNPI` **simma* `PNNPI` *noch* `PNNPI` *nicht* `PNNPI`" (*we ain't that far yet*) which is attributed to *R(est)* both as SC and EC. In fact, this is a good example of a mixture of *motherese* and *reprimanding*, the latter being indicated by the wording, the former by the tone of voice. However, as these are very rare cases, we cannot model them reliably for automatic processing and have to map them onto *R(est)*.

Furthermore, in this paper we will not deal with the problem of automatically obtaining EC given a set of features. Instead, we will assume the emotional labels as given. Thereby, we are avoiding segmentation errors for EC, as we do for SC, in both cases assuming a 100% correct segmentation; this can be considered as an upper bound for classification performance. Note that in preliminary experiments, we found out that Hidden Markov Models trained on EC obtained on the training set lead to a segmentation of the test data that achieves a classification performance comparable to the SC approach. To keep the two approaches fully apart, we do not combine sequences of SC with the same label into one higher "SC/EC-unit."

Table 4: Emotion classes and their SC-based frequencies.

Class	No. train	No. test	No. total	% total
P(ositive)	674	215	889	4.9
I(dle)	5260	5083	10343	56.8
R(est)	667	494	1161	6.4
N(egative)	3358	2465	5823	32.0
Total	9959	8257	18216	100.0

Table 5: Emotion classes and their EC-based frequencies.

Class	No. train	No. test	No. total	% total
P(ositive)	518	199	717	3.9
I(dle)	6410	6185	12595	69.0
R(est)	479	391	870	4.8
N(egative)	2276	1789	4065	22.3
Total	9683	8564	18247	100

3.4. Ememe Chunks versus Syntactic Chunks. Tables 4 and 5 reveal that the overall frequencies of the two chunk types are almost the same; however, there are 2.2 k more *I(dle)* EC-chunks than SC-chunks, counterbalanced by more *P(ositive)*, *N(egative)*, and *R(est)* SC-chunks. Figure 1 displays for each of the four classes and for all classes taken together, frequencies in percent for SC and EC with the length 1 to n words. The same information is given in the stacked histograms of Figure 2; in Figure 1, relationships within classes and differences between type of chunk can be seen, whereas Figure 2 concentrates on frequencies across classes within one plot. One-word ECs are more frequent than one-word SCs; especially for the three "marked" classes *P(ositive)*, *N(egative)*, and *R(est)*, there is a decline in frequencies, especially for chunks with 2, 3, or 4 words, which display higher frequencies for SC than for EC. These differences can be traced back to the different MV and thresholds. EC in our case are "pure," that is, after the initial, word-based MV, the labels are fixed, and only adjacent words with identical labels are combined into EC; as we mentioned above, this is not a necessary condition for EC but had to be chosen for our database, to avoid the sparse data problem. In contrast, if it is a chunk with more than one word, individual words belonging to the very same SC can be attributed to different classes but the combined threshold for the whole SC overrides such differences.

4. Features and Classifiers

4.1. Acoustic Features. The main focus has been on prosodic features in the past, in particular pitch, durations, and intensity [31]. Comparably small feature sets (10–100) were first utilised. In only a few studies, low-level feature modelling on a frame level was pursued, usually by Hidden Markov Models (HMMs) or Gaussian Mixture Models (GMMs). The higher success of static feature vectors derived by projection of the Low-Level Descriptors (LLDs) such as pitch or energy by descriptive statistical functional application such as lower order moments (mean, standard deviation) or

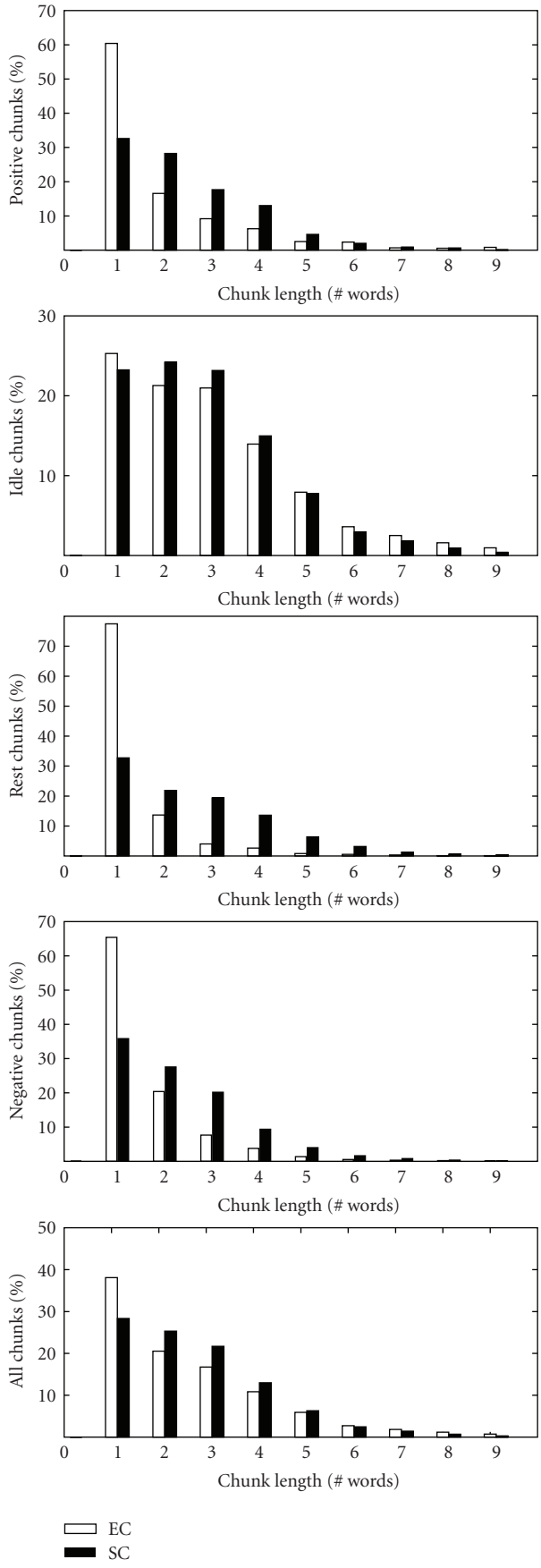

Figure 1: Chunk histogram with frequencies.

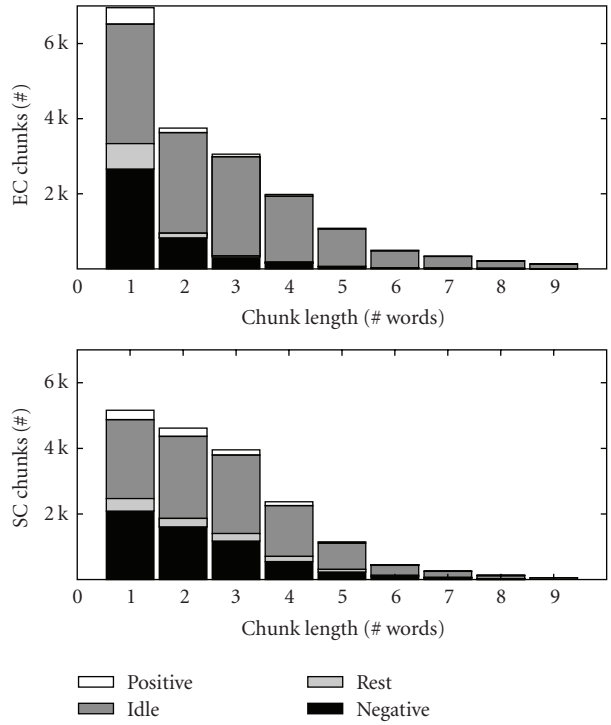

Figure 2: Chunk histogram with frequencies, stacked.

extrema is probably justified by the supra-segmental nature of the phenomena occurring with respect to emotional content in speech. In more recent research, also voice quality features such as Harmonics-to-Noise Ratio (HNR), jitter, or shimmer, and spectral and cepstral features such as formants and Mel-Frequency Cepstral Coefficients (MFCCs) have been successfully added to prosodic features. At the same time, brute-forcing of features (1000 up to 50000), for example, by analytical feature generation, partly also in combination with evolutionary generation, has become popular. It seems as if this (slightly) outperforms hand-crafted features while the individual worth of automatically generated features seems to be lower. Within expert-based hand-crafted features, perceptually more adequate features have been investigated, reaching from simple log-pitch to Teager energy or more complex features such as articulatory features (e.g., (de-)centralisation of vowels).

In this study, a feature set is employed that shall best cover the described gained knowledge. We therefore stick to the findings in [32] by choosing the most common and at the same time promising feature types and functionals covering prosodic, spectral, and voice quality features. Furthermore, we limit to a systematic generation of features. For the highest transparency, we utilise the open source openSMILE feature extraction and choose the basic set used in the only official challenge on emotion recognition from speech to the present day (cf. "Classifier Sub-Challenge" [33]). In detail, the 16 Low-Level Descriptors chosen are: zero-crossing-rate (ZCR) from the time signal, root mean square (RMS) frame energy, pitch frequency (normalised to 500 Hz), HNR by autocorrelation function, and MFCCs 1–12. To each of

Table 6: Acoustic Low-Level Descriptors (LLD) and functionals.

LLD (16 · 3)	Functionals (12)
(Δ, ΔΔ) ZCR	Mean
(Δ, ΔΔ) RMS Energy	Standard deviation
(Δ, ΔΔ) F0	Kurtosis, skewness
(Δ, ΔΔ) HNR	Extremes (max, min): value (2), rel. position (2), range (1)
(Δ, ΔΔ) MFCC 1-12	Linear regression: offset, slope, MSE

these, the delta coefficients are additionally computed. We further add double-delta coefficients for a better modelling of context. Then the 12 functionals mean, standard deviation, kurtosis, skewness, minimum and maximum value, relative position, and range as well as two linear regression coefficients with their Mean Square Error (MSE) are applied on a chunk basis as depicted in Table 6. Thus, the total feature vector per chunk contains $16 \cdot 3 \cdot 12 = 576$ attributes.

4.2. Linguistic Features. Spoken or written text also carries information on the underlying affective state [34–36]. This is usually reflected in the usage of certain words or grammatical alterations—which means in turn, in the usage of specific higher semantic and pragmatic entities.

From the many approaches existing we chose vector space modelling, that is, bag of words [37]. This is a well-known numerical representation form of text in automatic document categorisation introduced in [38]. It has been successfully ported to recognise sentiments in [39] or emotion and interest in [40]. The possibility of early fusion with acoustic features helped make this technique very popular as shown in [37].

For the FAU Aibo Emotion Corpus, the vocabulary size is 1146 entries. But only a fraction of these words conveys relevant information about the underlying emotional states of a person. In order to reduce the information in a meaningful way, two methods can be applied: stopping and stemming. Stopping uses simple rules or a data-driven evaluation to exclude single words from the vocabulary. A simple yet popular method for reducing the vocabulary is exploited: a minimum training database word frequency, here two, determines the necessary minimum number of occurrences for a word in the database for being part of the vocabulary. Very rare words are therefore discarded. Stemming instead is a method for reducing different morphological forms of a word to its base form. The Iterated Lovins Stemmer [41] is used for the experiments in this paper.

The main idea of the bag of words approach is the representation of words (or lexemes if stemming is applied) as numeric features. For each word (i.e., term) in the vocabulary, a corresponding feature that represents its frequency of occurrence in the unit exists, resulting in a high-dimensional feature vector space. Each unit can therefore be mapped to a vector in this feature space.

This frequency can be transformed in various ways [42, page 311], [38]. The logarithmic term frequency normalised to the inverse document (here database) frequency (TFIDF) in combination with normalisation to the unit length proved to be the best in our experiments.

Within this paper, the linguistic analysis is based on the correct transcription of the spoken content. Therefore it describes the performance under perfect speech recognition conditions. This follows the typical reporting of linguistic analysis results in emotion recognition, as it allows for better comparability of results [37]; the corpus comes with the transcription, while speech recognition results would differ from site to site. Also, some practical relevance exists: consider media retrieval from broadcasts; here the close captions are usually available. However, to close the gap to the real world where spoken content has to be determined by an ASR engine first, we had carried out experiments employing ASR for this corpus in other studies: though recognition of affect related speech is a rather difficult problem which has not been solved yet to complete satisfaction [43], this did not yield marked differences, as reported, for example, in [44, 45] for this corpus. This derives from the fact that not the perfect word chain is needed as, for example, in transcription of speech. Some minor mistakes are caught by stemming and stopping, and not all words are necessarily needed. Insertions and substitutions are only critical if they change the "tone" of the affective content. As additional features in linguistic analysis, we utilise each word's start and end time, as well as the derived duration. This is motivated by the fact that an ASR engine would also provide this information.

4.3. Classifiers. Classifiers typically used for the recognition of emotion from speech comprise a broad variety: depending on the feature type considered for classification either dynamic algorithms as Hidden Markov Models [46] or Multiinstance Learning techniques [10] for processing on a frame-level, and static classifiers for processing on the supra-segmental functional level are found. With respect to static classification the list of classifiers seems endless: Neural Networks (mostly Multilayer Perceptrons), Naïve Bayes, Bayesian Networks, Gaussian Mixture Models, Decision Trees, Random Forests, Linear Discriminant Classifiers, k-Nearest Neighbour distance classifiers, and Support Vector Machines are found most often [4]. Also a selection of ensemble techniques has been applied, as Boosting, Bagging, Multiboosting, and Stacking with and without confidences [47]. Finally, the two general types may also be mixed by fusion of dynamic and static classification [48].

As we consider acoustic and linguistic information, the two information streams need to be integrated. In this respect, all experiments found in the literature use static classification techniques [17, 37, 49]: an early fusion is usually the best choice for preserving all information prior to the final decision. Thus, the acoustic features introduced in Section 4.1 and the linguistic ones introduced in Section 4.2 are combined in one feature vector on the respective unit level (i.e., word or chunk), which demands for static classification.

The classifier of choice to this aim in this paper is a discriminatively learned simple Bayesian Network, namely Discriminative Multinomial Naïve Bayes (DMNB) [50]

instead of Support-Vector Machines (SVMs) and Random Forests as applied in our previous investigations [23, 32, 33]. The reason is twofold: first, DMNB only requires lower memory and only a fraction of the computation time of SVM. (Sequential Minimal Optimisation training of SVM with linear Kernel demanded 200 times higher computation time than DMNB in parameterisation as below using [42] on an 8 GB RAM, 2.4 GHz, 64 Bit industry PC.) At the same time, the mean recall values resulted in a slight absolute improvement over SVM in our experiments on the FAU Aibo Emotion Corpus (−0.9/+1.3 weighted/unweighted average recall on average for acoustic features; +6.9/+2.3 for linguistic features). Second, the parameter learning is carried out by discriminative frequency estimation, whereby the likelihood information and the prediction error are considered. Thus, a combination of generative and discriminative learning is employed. This method is known to work well in highly correlated spaces (as in our case), to converge quickly, and not to suffer from overfitting.

For optimal results we found it best to ignore the frequency information in the data and select a number of ten iterations. Numeric variables are discretised using unsupervised ten-bin discretisation [42]. Multiclass decision is obtained by transformation into binary problems by taking the two largest classes, each.

5. Classification

As mentioned above and carried out within [33], we split the FAU Aibo Emotion Corpus into train and test partitions by schools of recording. Thus, utmost independence of the speaker, room acoustics, general intonation and articulation patterns, and wording of the children is ensured. To better cope with this variety, all features are standardised per partition (speaker group normalisation). Due to the high imbalance among classes (cf. Table 1), balancing of the training instances is further mandatory to achieve reasonable values of unweighted recall and thus avoid overfitting of strong classes (here *I(dle)* and *N(egative)* [17]. The chosen straightforward strategy is random mixed up-sampling of sparse and downsampling of majority classes' instances enforcing unit distribution while preserving the total number of instances. Note that the order of operations has an influence on (un)weighted recall figures [33]; we first standardise and then balance the training. Next we classify with DMNB as described. At this point constant parameterisation is preferred over individual optimisation; thus, no alterations are undertaken with respect to number of iterations, quantisation, and so forth, among the different units and feature types to be classified.

Table 7 displays weighted average recall (WA), that is, the overall recognition rate (RR) or recall (number of correctly classified cases divided by total number of cases), and unweighted average recall (UA) (or "class-wise" computed recognition rate (CL)), that is, the mean along the diagonal of the confusion matrix in percent, for three sets of features: only acoustic features, only linguistic features, and both acoustic and linguistic features (early fusion).

Table 7: Evaluation in percent correct; (un-)weighted average recall (UA/WA). Note that the chunk- and the word-based evaluations of word units coincide; words can be seen a the smallest possible chunk.

Unit	Acoustic		Linguistic		Ac. + ling.	
	WA	UA	WA	UA	WA	UA
Basic unit of evaluation: the chunk						
WO	49.73	46.15	44.27	45.50	53.59	**48.56**
SC	46.43	44.40	43.04	42.30	50.02	**46.33**
EC	57.23	51.63	62.42	51.80	64.89	**55.38**
Basic unit of evaluation: the word						
WO	49.73	46.15	44.27	45.50	53.59	**48.56**
SC	44.56	45.10	40.63	44.37	48.82	**48.35**
EC	65.84	53.10	71.98	53.00	73.66	**56.77**

There are two different types of evaluation: first we evaluate for the whole units WO, SC, and EC; note that the total number of chunks is different for each of these units. Then, we evaluate WO, SC, and EC by checking each word in these units whether it has been classified correctly, that is, attributed to the class the higher unit it belongs to has been annotated with. Obviously, the evaluation of the unit WO is identical under the two methods, as words can be seen as smallest possible chunks. WA tends to be higher because of the bias in class distribution; UA is more relevant for applications which are, most of the time, more interested in the "marked" classes, that is, in our case, not in the frequent *I(dle)* class; thus, we concentrate on the interpretation of UA. All results are above chance level (25% correct). The chunk-based figures might be more relevant if we have applications in mind, the word-based figures are more balanced. We can see that the early fusion of acoustic and linguistic features pays off, always yielding higher WA and UA. As there are more—and especially longer *I(dle)* chunks containing more words, it could be expected that word-based evaluation for SC and EC yields better results; the differences are, however, not marked. As ECs are more consistent—all words belonging to an EC belong to the same class—UA for EC is higher than for SC.

On average, the unit "word" contains less information than the units SC and EC; each unit consists of only one word whereas SC and EC mostly consist of more than one word. The number of SC and EC is in the same range, as can be seen in Tables 4 and 5, although EC should be more consistent than SC. In Table 8, we display classification results for cross-unit evaluation, that is, we use different units for the training and for testing partitions. This could only be done for acoustic features because of the unbalanced distribution of linguistic features in the different units. We see that performance is really worst when we use the unit "word" both as train or test unit, with UA being consistently below 40 % correct. Overall, there is again almost no difference between chunk-based and word-based evaluation of UA. Although the figures are of course lower than for the within-unit evaluation displayed in Table 7, it is reassuring that

Table 8: Cross unit: train ≠ test; evaluation in percent correct; (un-)weighted average recall (UA/WA); acoustics only.

Train	Test	Chunk based		Word based	
		UA	WA	WA	UA
EC	SC	47.77	42.15	51.51	42.10
WO	SC	44.01	38.43	46.83	39.90
SC	EC	45.22	45.40	47.85	46.98
WO	EC	32.53	30.73	31.23	33.08
EC	WO	37.81	38.30	37.81	38.30
SC	WO	27.50	39.35	27.50	39.35

Table 9: Confusion matrix, acoustics + linguistics, word evaluation in percent correct.

Class. as →	*P*	*I*	*R*	*N*	Total
P(ositive)	**47.49**	24.41	16.38	11.70	299
I(dle)	9.60	**51.76**	12.95	25.67	18698
R(est)	21.27	26.54	**21.63**	30.54	550
N(egative)	3.15	13.12	10.30	**73.41**	2697

performance does not break down when we train with EC and test with SC or vice versa.

Table 9 displays the confusion matrix for the fusion of acoustics and linguistics for the word evaluation (cf. the two last columns in Table 7, row WO), giving an impression of the confusion between classes. The confusion between the classes is as expected; no much confusion between *P(ositive)* and *N(egative)*, most confusion between *I(dle)* (and partly *R(est)*) and the other classes. Table 10 displays recall-rates (i.e., only the figures of the diagonal of the confusion matrices) as correctly classified cases per class, for the remaining four constellations from Table 7. This gives an impression of the performance per class across all constellations. Basically, the picture is always the same: the mixed *R(est)* class is recognised worst and almost evenly smeared across all classes. The highest recognition rates can be observed for the rather acoustically and linguistically marked *N(egative)* instances—both for SC and EC; *P(ositive)* is in between.

6. Discussion

6.1. Classification Performance: The Reality Shock. The scientific community has been used to good or almost perfect classification performance in emotion recognition; it is such figures that are remembered and implicitly defined as standard. We have to realise, however, that such figures have been obtained only within specific constellations: acted data [51], prototypical cases preselected out of the whole database, or a focus on one specific class, modelling all other classes as rest/garbage; for this last constellation, high recall can be obtained if we can live with many false alarms in the rest classes. Normally, the data have not been processed fully automatically but the experiments have been based on the spoken word chain. In the present study, we aim at realistic conditions—apart from the last step to use fully automatic ASR. In [23] we could show that—depending on the recording conditions and the feature set used—ASR errors do not always deteriorate emotion recognition.

Table 10: Classwise recall values (i.e., diagonal values of confusion matrices), acoustics + linguistics; chunk evaluation in percent correct. Note that the chunk- and the word-based evaluations of WO coincide.

Constellation	*P(ositive)*	*I(dle)*	*R(est)*	*N(egative)*
WO, chunk based	47.49	51.76	21.63	73.41
SC, chunk based	51.62	44.26	22.06	67.34
EC, chunk based	44.72	65.40	40.66	70.65
WO, word based	47.49	51.76	21.64	73.41
SC, word based	55.18	46.43	21.45	70.26
EC, word based	48.49	75.93	33.81	68.85

We simply do not know yet which type of realistic databases—amongst them our FAU Aibo Emotion Corpus—could be conceived as being representative, as far as distinctiveness of classes and by that, goodness of performance is concerned. Chances are, however, that we will never achieve such high performance as we did, using only acted and/or prototypical data, and that approaches such as the present one—trying to model all phenomena present is a database—will give way to more focused approaches, aiming at specific classes for specific application tasks.

6.2. Deciding between Types of Units. At least three aspects are relevant for deciding between WO, SC, and EC—or any other type of sequencing emotional episodes: first, performance; second, adequacy in real-life applications; third, perceptual, cognitive adequacy.

Performance has been significantly better for EC than for SC. Note that in this paper, we used the spoken word chain simulating 100% correct word recognition, and a manual segmentation into SC and EC. For a fair comparison between SC and EC, this had to be done automatically. We know that SC can be established automatically with high reliability even for spontaneous speech [9]. As for EC, this might look like a "Münchhausen" approach; finding the boundaries of phenomena we afterwards want to recognise; however, preliminary experiments showed that it can be done using an HMM approach, albeit yielding lower classification performance in the range of SC. Semantically "rich" words, that is, content words such as nouns, adjectives, and verbs, tend to be marked emotionally to a higher extent than function words such as particles. For instance, in our data, more EC (22.5%) than SC (21.3%) consist only of 1 to n content words. A modelling of part-of-speech (POS) sequences yields a classification performance, not much lower than one obtained with acoustic modelling [32]. POS modelling is rather robust because ASR confusions between words within one POS class have no effect. Factors like these make it likely that LM modelling of EC is as promising as LM modelling of SC. Thus, it is an empirical question to be addressed

whether EC will be classified better than SC, if the process is fully automated. The best compromise between automation and performance seems to be WO. Here, we obtain better results than for SC—but still worse results than for EC. Word segmentation is obtained for free if ASR has been applied. However, due to the reasons sketched passim and in the following, the single ememe, that is, the word, might not be the optimal unit, if it comes to processing in both higher linguistic and emotion modules.

Applications are different. If we look at the attempt towards a taxonomy of applications in [52], most important for segmentation might be the difference between online and offline applications. We mentioned in the beginning that for online applications such as SmartKom [11] or Semaine [12], incremental processing will be mandatory because of time constraints. Matters are similar in any interaction between users and Embodied Conversational Agents (ECAs) or robots. In such online applications, there is normally an interaction between system and user. The system does not only monitor somehow the user's emotional states but has to recognise and process linguistic content and semantics and illocutions (dialogue acts) in order to react appropriately. This makes a close dovetailing of linguistics and para-linguistics such as monitoring emotional states most adequate, and this in turn might favour the processing of SC instead of EC or WO. It is different in offline applications; the processing of movie databases in search for emotional episodes needs not be incremental and can be done in several passes. Thus, we can imagine one pass for emotion monitoring within the whole movie, and then a second pass for segmentation, and so on.

To our knowledge, there are not many studies on the relationship between human speech/linguistic processing and human emotion processing. We know, however, that phonetic/psycholinguistic studies on the localisation of nonverbal signals within speech showed that listeners tend to structure the perception of these phenomena along the perception and comprehension of linguistic phenomena (sentence processing) [53]. Unpublished studies on the localisation of laughter in our data showed that this is the case for the production of paralinguistic events as well. Thus, it might be that linguistics and emotions are more intertwined—at least within interactions where emotional and non-emotional episodes alternate. If this is the case, the modelling of SC seems to be most adequate also from the point of view of cognition and comprehension.

7. Concluding Remarks

The unique contribution of the present study is the use of word-based annotations and the subsequent mapping onto different types of higher units, to investigate promising possibilities of segmenting emotional episodes. However, word-based annotation is very time-consuming and thus expensive. Perhaps it should not be established as a new standard but only be used for basic research. The higher units "syntactic chunk" and "emotion/ememe chunk" introduced in this study are, in our opinion, representative for two different types of most promising units. However, a great variety of different thresholds or mapping procedures can be imagined. Most of them will not differ considerably, as far as usability or performance is concerned. Although being a truism, we definitely need more realistic databases for deciding between such alternative approaches.

Acknowledgments

This work originated in the CEICES initiative (Combining Efforts for Improving Automatic Classification of Emotional User States) taken in the European Network of Excellence HUMAINE [37]. The research leading to these results has received funding from the European Community under Grant (FP7/2007–2013) no. 211486 (SEMAINE), Grant no. IST-2002–50742 (HUMAINE), and Grant no. IST-2001-37599 (PF-STAR). The responsibility lies with the authors.

References

[1] A. Ortony, G. L. Clore, and A. Collins, *The Cognitive Structure of Emotions*, Cambridge University Press, Cambridge, NY, USA, 1988.

[2] R. Cowie, N. Sussman, and A. Ben-Ze'ev, "Emotions: concepts and definitions," in *Humaine Handbook on Emotion*, P. Petta, Ed., Springer, Berlin, Germany, 2010, to appear.

[3] H. Grice, "Logic and conversation," in *Syntax and Semantics*, P. Cole and J. Morgan, Eds., vol. 3 of *Speech Acts*, pp. 41–58, Academic Press, New York, NY, USA, 1975.

[4] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: audio, visual, and spontaneous expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, 2009.

[5] A. Batliner, S. Steidl, C. Hacker, E. Nöth, and H. Niemann, "Tales of tuning—prototyping for automatic classification of emotional user states," in *Proceedings of the 9th European Conference on Speech Communication and Technology (Interspeech '05)*, pp. 489–492, Lisbon, Portugal, 2005.

[6] D. Seppi, A. Batliner, B. Schuller, et al., "Patterns, prototypes, performance: classifying emotional user states," in *Proceedings of the 9th International Conference on Spoken Language Processing (Interspeech '08)*, pp. 601–604, Brisbane, Australia, September 2008.

[7] R. P. Lippmann, "Speech recognition by machines and humans," *Speech Communication*, vol. 22, no. 1, pp. 1–15, 1997.

[8] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schröder, "Feeltrace: an instrument for recording perceived emotion in real time," in *Proceedings of the ISCA Workshop on Speech and Emotion*, pp. 19–24, Newcastle, Northern Ireland, 2000.

[9] A. Batliner, R. Kompe, A. Kießling, M. Mast, H. Niemann, and E. Nöth, "M = Syntax + Prosody: a syntactic-prosodic labelling scheme for large spontaneous speech databases," *Speech Communication*, vol. 25, no. 4, pp. 193–222, 1998.

[10] M. Shami and W. Verhelst, "Automatic classification of expressiveness in speech: a multi-corpus study," in *Speaker Classification II*, C. Müller, Ed., vol. 4441 of *Lecture Notes in Computer Science/Lecture Notes in Artificial Intelligence*, pp. 43–56, Springer, Berlin, Germany, 2007.

[11] M. Streit, A. Batliner, and T. Portele, "Emotions analysis and emotion-handling subdialogues," in *SmartKom: Foundations of Multimodal Dialogue Systems*, W. Wahlster, Ed., pp. 317–332, Springer, Berlin, Germany, 2006.

[12] M. Schröder, R. Cowie, D. Heylen, M. Pantic, C. Pelachaud, and B. Schuller, "Towards responsive sensitive artificial listeners," in *Proceedings of the 4th International Workshop on Human-Computer Conversation*, Bellagio, Italy, 2008.

[13] M. Pantic, N. Sebe, J. F. Cohn, and T. Huang, "Affective multimodal human-computer interaction," in *Proceedings of ACM Multimedia*, pp. 669–676, Singapore, 2005.

[14] A. Batliner, S. Steidl, C. Hacker, and E. Nöth, "Private emotions versus social interaction: a data-driven approach towards analysing emotion in speech," *User Modelling and User-Adapted Interaction*, vol. 18, no. 1-2, pp. 175–206, 2008.

[15] S. Steidl, *Automatic classification of emotion-related user states in spontaneous children's speech*, Ph.D. thesis, Logos, Berlin, Germany, 2009.

[16] A. Batliner, V. Zeissler, C. Frank, J. Adelhardt, R. P. Shi, and E. Nöth, "We are not amused—but how do you know? User states in a multi-modal dialogue system," in *Proceedings of the 8th European Conference on Speech Communication and Technology (Interspeech '03)*, pp. 733–736, Geneva, Switzerland, September 2003.

[17] B. Schuller, R. Müller, F. Eyben, et al., "Being bored? Recognising natural interest by extensive audiovisual integration for real-life application," *Image and Vision Computing*, vol. 27, no. 12, pp. 1760–1774, 2009.

[18] M. Blomberg and D. Elenius, "Collection and recognition of children's speech in the PF-Star project," in *Proceedings of the Swedish Phonetics Conference (Fonetik '00)*, pp. 81–84, Umeå, Sweden, 2003.

[19] M. Russell, S. D'Arcy, and L. Qun, "The effects of bandwidth reduction on human and computer recognition of children's speech," *IEEE Signal Processing Letters*, vol. 14, no. 12, pp. 1044–1046, 2007.

[20] D. Giuliani and M. Gerosa, "Investigating recognition of children's speech," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '03)*, vol. 2, pp. 137–140, Hong Kong, 2003.

[21] S. Lee, A. Potamianos, and S. Narayanan, "Acoustics of children's speech: developmental changes of temporal and spectral parameters," *Journal of the Acoustical Society of America*, vol. 105, no. 3, pp. 1455–1468, 1999.

[22] M. Holodynski and W. Friedlmeier, *Development of Emotions and Emotion Regulation*, Springer, New York, NY, USA, 2006.

[23] B. Schuller, D. Seppi, A. Batliner, A. Maier, and S. Steidl, "Towards more reality in the recognition of emotional speech," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '07)*, vol. 4, pp. 941–944, Honolulu, Hawaii, USA, 2007.

[24] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proceedings of the 9th European Conference on Speech Communication and Technology (Interspeech '05)*, pp. 1517–1520, Lisbon, Portugal, 2005.

[25] I. S. Engberg, A. V. Hansen, O. Andersen, and P. Dalsgaard, "Design, recording and verification of a Danish emotional speech database," in *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech '97)*, pp. 1695–1698, Rhodes, Greece, 1997.

[26] J. Schwitalla, *Gesprochenes Deutsch: Eine Einführung*, Erich Schmidt, Berlin, Germany, 1997.

[27] A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Nöth, "How to find trouble in communication," *Speech Communication*, vol. 40, no. 1-2, pp. 117–143, 2003.

[28] Z. Inanoglu and R. Caneel, "Emotive alert: HMM-based emotion detection in voicemail messages," in *Proceedings of the 10th International Conference on Intelligent User Interfaces (IUI '05)*, pp. 251–253, San Diego, Calif, USA, 2005.

[29] L. Devillers, L. Vidrascu, and L. Lamel, "Challenges in real-life emotion annotation and machine learning based detection," *Neural Networks*, vol. 18, no. 4, pp. 407–422, 2005.

[30] F. de Rosis, A. Batliner, N. Novielli, and S. Steidl, "'You are sooo cool, Valentina!' recognizing social attitude in speech-based dialogues with an ECA," in *Affective Computing and Intelligent Interaction*, A. Paiva, R. Prada, and R. W. Picard, Eds., pp. 179–190, Springer, Berlin, Germany, 2007.

[31] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, et al., "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, 2001.

[32] B. Schuller, A. Batliner, D. Seppi, et al., "The relevance of feature type for the automatic classification of emotional user states: low level descriptors and functionals," in *Proceedings of the 8th Annual Conference of the International Speech Communication Association (Interspeech '07)*, vol. 2, pp. 2253–2256, Antwerp, Belgium, August 2007.

[33] B. Schuller, S. Steidl, and A. Batliner, "The Interspeech 2009 emotion challenge," in *Proceedings of the 10th Annual Conference of the International Speech Communication Association (Interspeech '09)*, pp. 312–315, Brighton, UK, 2009.

[34] S. Arunachalam, D. Gould, E. Anderson, D. Byrd, and S. Narayanan, "Politeness and frustration language in child-machine interactions," in *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech '01)*, pp. 2675–2678, Aalborg, Denmark, September 2001.

[35] Z.-J. Chuang and C.-H. Wu, "Emotion recognition using acoustic features and textual content," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME '04)*, vol. 1, pp. 53–56, Taipei, Taiwan, 2004.

[36] K. Dupuis and K. Pichora-Fuller, "Use of lexical and affective prosodic cues to emotion by younger and older adults," in *Proceedings of the 8th Annual Conference of the International Speech Communication Association (Interspeech '07)*, vol. 2, pp. 2237–2240, Antwerp, Belgium, August 2007.

[37] A. Batliner, S. Steidl, B. Schuller, et al., "Combining efforts for improving automatic classification of emotional user states," in *Proceedings of the 1st International Language Technologies Conference (IS-LTC '06)*, pp. 240–245, Ljubljana, Slovenia, 2006.

[38] T. Joachims, "Text categorization with support vector machines: learning with many relevant features," in *Proceedings of the 10th European Conference on Machine Learning (ECML '98)*, C. Nédellec and C. Rouveirol, Eds., pp. 137–142, Chemnitz, Germany, 1998.

[39] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '02)*, pp. 79–86, Philadelphia, Pa, USA, 2002.

[40] B. Schuller, N. Köhler, R. Müller, and G. Rigoll, "Recognition of interest in human conversational speech," in *Proceedings of the 9th International Conference on Spoken Language Processing (ICSLP '06)*, vol. 2, pp. 793–796, Pittsburgh, Pa, USA, 2006.

[41] J. B. Lovins, "Development of a stemming algorithm," *Mechanical Translation and Computational Linguistics*, vol. 11, pp. 22–31, 1968.

[42] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, San Francisco, Calif, USA, 2nd edition, 2005.

[43] T. Athanaselis, S. Bakamidis, I. Dologlou, R. Cowie, E. Douglas-Cowie, and C. Cox, "ASR for emotional speech: clarifying the issues and enhancing performance," *Neural Networks*, vol. 18, no. 4, pp. 437–444, 2005.

[44] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Does affect affect automatic recognition of children's speech?" in *Proceedings of the 1st Workshop on Child, Computer and Interaction*, Chania, Greece, 2008.

[45] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Emotion recognition from speech: putting ASR in the loop," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '09)*, pp. 4585–4588, Taipei, Taiwan, 2009.

[46] T. L. Nwe, S. W. Foo, and L. C. De Silva, "Speech emotion recognition using hidden Markov models," *Speech Communication*, vol. 41, no. 4, pp. 603–623, 2003.

[47] B. Schuller, R. Jiménez Villar, G. Rigoll, and M. Lang, "Meta-classifiers in acoustic and linguistic feature fusion-based affect recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '05)*, vol. 1, pp. 325–328, Philadelphia, Pa, USA, March 2005.

[48] B. Vlasenko, B. Schuller, A. Wendemuth, and G. Rigoll, "Combining frame and turn-level information for robust recognition ofemotions within speech," in *Proceedings of the 8th Annual Conference of the International Speech Communication Association (Interspeech '07)*, vol. 4, pp. 2249–2252, Antwerp, Belgium, August 2007.

[49] C. M. Lee and S. S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 2, pp. 293–303, 2005.

[50] J. Su, H. Zhang, C. X. Ling, and S. Matwin, "Discriminative parameter learning for Bayesian networks," in *Proceedings of the 25th International Conference on Machine Learning (ICML '08)*, pp. 1016–1023, Helsinki, Sweden, 2008.

[51] A. Batliner, K. Fischer, R. Huber, J. Spilker, and R. Nöth, "Desperately seeking emotions: actors, wizards, and human beings," in *Proceedings of the ISCA Workshop on Speech and Emotion*, pp. 195–200, Newcastle, Northern Ireland, 2000.

[52] A. Batliner, F. Burkhardt, M. van Ballegooy, and E. Nöth, "A taxonomy of applications that utilize emotional awareness," in *Proceedings of the 1st International Language Technologies Conference (IS-LTC '06)*, pp. 246–250, Ljubljana, Slovenia, 2006.

[53] M. Garrett, T. Bever, and J. Fodor, "The active use of grammar in speech perception," *Perception and Psychophysics*, vol. 1, pp. 30–32, 1966.