On the impact of children's emotional speech on acoustic and language models

Stefan Steidl, Anton Batliner, Dino Seppi, Björn Schuller

Angaben zur Veröffentlichung / Publication details:

Steidl, Stefan, Anton Batliner, Dino Seppi, and Björn Schuller. 2010. "On the impact of children's emotional speech on acoustic and language models." *EURASIP Journal on Audio, Speech, and Music Processing*, 783954. https://doi.org/10.1155/2010/783954.

Hindawi Publishing Corporation
EURASIP Journal on Audio, Speech, and Music Processing
Volume 2010, Article ID 783954, 14 pages
doi:10.1155/2010/783954

Research Article

On the Impact of Children's Emotional Speech on Acoustic and Language Models

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

[1] *Lehrstuhl für Mustererkennung, Friedrich-Alexander-Universität Erlangen-Nürnberg, Martensstraße 3, 91058 Erlangen, Germany*
[2] *ESAT, Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, 3001 Heverlee (Leuven), Belgium*
[3] *Institute for Human-Machine Communication, Technische Universität München, Arcisstraße 21, 80333 München, Germany*

Correspondence should be addressed to Stefan Steidl, stefan.steidl@informatik.uni-erlangen.de

Received 2 June 2009; Revised 9 October 2009; Accepted 23 November 2009

Academic Editor: Georg Stemmer

The automatic recognition of children's speech is well known to be a challenge, and so is the influence of affect that is believed to downgrade performance of a speech recogniser. In this contribution, we investigate the combination of both phenomena. Extensive test runs are carried out for 1 k vocabulary continuous speech recognition on spontaneous *motherese*, *emphatic*, and *angry* children's speech as opposed to *neutral* speech. The experiments address the question how specific emotions influence word accuracy. In a first scenario, "emotional" speech recognisers are compared to a speech recogniser trained on *neutral* speech only. For this comparison, equal amounts of training data are used for each emotion-related state. In a second scenario, a "neutral" speech recogniser trained on large amounts of *neutral* speech is adapted by adding only some emotionally coloured data in the training process. The results show that *emphatic* and *angry* speech is recognised best—even better than *neutral* speech—and that the performance can be improved further by adaptation of the acoustic and linguistic models. In order to show the variability of emotional speech, we visualise the distribution of the four emotion-related states in the MFCC space by applying a Sammon transformation.

1. Introduction

Offering a broad variety of applications, such as literacy and reading tutors [1, 2], speech interfaces for children are an attractive subject of research [3]. However, automatic speech recognition (ASR) is known to be a challenge for the recognition of children's speech [4–8]: characteristics of both acoustics and linguistics differ from those of adults [9], for example, by higher pitch and formant positions or not yet perfectly developed coarticulation. At the same time, these strongly vary for children of different ages due to anatomical and physiological development [10] and learning effects. In [11], voice transformations are applied successfully to increase the performance for children's speech if an adult speech recogniser is used.

Apart from children's speech, also affective speech can be challenging for ASR [12, 13], as acoustic parameters differ considerably under the influence of affect. In [14], acoustic parameters (MFCC and MFB features) are investigated for the 4-class problem *anger*, *sadness*, *happy*, and *neutral* (emotion portrayals) and the 2-class problem *negative* versus *nonnegative* (data of a real call-centre application). It is shown that acoustic models for broad phonetic categories that are trained on neutral speech produce emotional speech with significantly different likelihood scores, which can be used to discriminate emotions. In [15, 16], the influence on ASR of speech under stress as an emotion-related phenomenon is investigated. The two ASR problems *children's speech* and *affective speech* will typically occur in combination when building systems for children-computer interaction by speech: children tend towards natural and spontaneous—and therefore also affective—speech behaviour in interaction with technical systems [17–19]. In [20], we addressed the influence of ASR errors on the performance of an emotion recognition module based on linguistic features. In this paper, it is the other way round: we address the influence of emotion on the recognition of children's speech. As opposed to previous work [21], we study the effect of each

of four emotion-related states individually to answer the main question: how does a particular affect affect speech recognition?

In this paper, we avoid delving into the theoretical debates on the definition of *affect* and *emotion*, and we use both terms interchangeably. Furthermore, as the speakers' states that can be observed in our data are more emotion-related than pure emotions, we opted for the more generic term *emotion-related states*.

The paper is structured as follows. In Section 2, we introduce the FAU Aibo Emotion Corpus, which is a corpus of spontaneous, emotionally coloured children's speech, and briefly describe the scenario to elicit emotional speech. In Section 2.1, we describe the recording settings and the amount of speech data, followed by Section 2.2 where the annotation of the speech data with emotion categories on the word level is described. In this paper, automatic speech recognition is carried out on semantically meaningful "chunk" units that are defined in Section 2.3. Emotion labels for whole chunks are defined in Section 2.4; these labels are based on the manual annotation on the word level. In Section 3, we define subsets of the corpus of equal size for the 4-class problem ***M**otherese*, ***N**eutral*, ***E**mphatic*, and ***A**nger*. Furthermore, we define two ASR scenarios. In the first scenario, which is described in Section 3.1, a speech recogniser trained on neutral speech is compared to speech recognisers that are exclusively trained on the same amount of emotional speech data. In Section 3.2, the second scenario is described, where a speech recogniser trained on large amounts of neutral speech is adapted to emotional speech by adding small amounts of emotional speech data to the training data. For both scenarios, experimental ASR results are presented for ***E**mphatic*, ***A**ngry*, and ***M**otherese* speech compared to ***N**eutral* speech; significant differences in terms of the word accuracy can be observed. The significance tests are described in Section 3.3. In Section 4, the higher variability of emotional speech is illustrated by visualisation of the acoustic feature space. Finally, the major findings of the study are summarised in Section 5.

2. Emotionally Coloured Children's Speech

The experiments described in this paper are based on the FAU Aibo Emotion Corpus, a corpus of German spontaneous speech with recordings of children at the age of 10 to 13 years communicating with a pet robot; it is described in detail in [22].

The general framework for this database of children's speech is child-robot communication and the elicitation of emotion-related speaker states. The robot is Sony's (dog-like) robot Aibo. The basic idea has been to combine children's speech and naturally occurring emotional speech within a Wizard-of-Oz task. The speech is "natural" because children do not disguise their emotions to the same extent as adults do. However, it is not as "natural" as it might be in a nonsupervised setting. Furthermore, the speech is spontaneous, because the children were not told to use specific instructions but to talk to Aibo like they would talk to a friend. In this experimental design, the child is led to believe that Aibo is responding to his or her commands, but the robot is actually being remote-controlled by a human operator, using the "Aibo Navigator" software over a wireless LAN. The existing Aibo speech recognition module is turned off. The wizard causes Aibo to perform a fixed, predetermined sequence of actions, which takes no account of what the child is actually saying. 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 discontinue the experiment. The children believed that Aibo was reacting to their orders—albeit often not immediately. In fact, it was the other way round: Aibo was always strictly following the same screen-plot, and the children had to align their orders to its actions.

2.1. Speech Recordings. The data was collected from 51 children (21 male, 30 female) aged 10 to 13 years from two different schools ("Mont" and "Ohm"); the recordings took place in the respective class-rooms. Speech was transmitted via a wireless head set (Shure UT 14/20 TP UHF series with microphone WH20TQG) and recorded with a DAT-recorder (sampling rate 48 kHz, quantisation 16 bit, down-sampled to 16 kHz). Each recording session took around 30 minutes; in total there are 27.5 hours of data. The recordings contain large amounts of silence, which are due to the reaction time of Aibo. After removing longer pauses, the total amount of speech is equal to 9.2 hours.

2.2. Emotion Labelling on the Word Level. Five labellers (advanced students of linguistics, German native speakers, 4 female, 1 male, 20–26 years old) listened to the recordings in sequential order and annotated independently from each other each word as neutral (default) or as belonging to one of ten other emotion categories. In order to provide context information, the labellers could listen to the whole turn before labelling the single words. The set of emotion categories was defined prior to the labelling process by inspecting the data and the emotional states that can be observed. We resort to majority voting (henceforth MV): if three or more labellers agree, the label is attributed to the word; 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), and *rest*, that is, nonneutral, but not belonging to the other categories (3), *neutral* (39 169). 4707 words had no MV; all in all, the corpus consists of 48 401 words.

The state *emphatic* has to be commented on especially: based on our experience with other emotion databases [23], any marked deviation from a neutral speaking style can (but need not) be taken as a possible indication of some (starting) trouble in communication. If a user gets the impression that the machine does not understand him, he tries different strategies—repetitions, reformulations, other wordings, or simply the use of a pronounced, marked speaking style.

Thus, such a style does not necessarily indicate any deviation from a neutral user state but it means a higher probability that the (neutral) user state will possibly be changing soon. Of course, it can be something else as well: a user idiosyncrasy, or a special style such as "computer talk" that some people use while speaking to a computer, or speaking to a nonnative, to a child, or to an elderly person who is hard of hearing. Thus, it can only be found out by analysis of the data whether *emphatic* has to be conceived of as more positive or more negative (cf. the remarks on *surprise* in [24], which can be either negative or positive, depending on the context). In the FAU Aibo Emotion Corpus, *emphatic* can be found between *neutral* and *angry* on the valence scale in a two-dimensional arrangement of the emotional states obtained by *Nonmetric Dimensional Scaling* (NMDS) [17]. There is also another practical argument for the annotation of *emphatic*: if the labellers are allowed to label *emphatic,* it might be less likely that they confuse it with other user states. Note that all the states, especially *emphatic*, had only been annotated if they differed from the (initial) baseline of the speaker.

Some of the labels are very sparse. Therefore, we mapped *touchy* and *reprimanding*, together with *angry*, onto ***A**nger* as these states represent different but closely related kinds of negative attitude. This mapping is corroborated by NMDS analysis presented in [17]. In this paper, we focus on the four-class problem ***M**otherese*, ***N**eutral*, ***E**mphatic*, and ***A**nger* ranging from positive to negative valence. This order is kept constant in all figures and tables of this paper.

Interlabeller agreement is dealt within [22, 25]. On a balanced subset of the FAU Aibo Emotion Corpus, containing only words of the cover classes ***M**otherese*, ***N**eutral*, ***E**mphatic*, and ***A**nger*, weighted kappa values for multirater kappa are reported to be 0.56. Confusion matrices, where the decision of one labeller is compared to the majority vote of all five labellers, allow to judge the similarity of the different emotion categories. Figure 1 shows a graphical representation of the similarity of the four cover classes ***M**otherese*, ***N**eutral*, ***E**mphatic*, and ***A**nger* [17, 22]. The arrangement of these classes in the two-dimensional space is obtained by NMDS. The more likely the classes are to be confused by the human labellers, the closer they are in this arrangement. The quality of the NMDS result is given in Figure 1; it is assessed using Kruskal's stress function *S* and the squared correlation RSQ [26]. The figure is translated such that ***N**eutral* is located in the centre. The negative class ***A**nger* and its prestage ***E**mphatic* are located on the left side, whereas the positive state ***M**otherese* is on the right side. In Section 4 it is shown that the Sammon transformation of the acoustic features (average MFCC features per speaker and emotion) leads to a similar arrangement of the four cover classes; only the position of ***A**nger* is slightly different (closer to ***M**otherese* than to ***E**mphatic*).

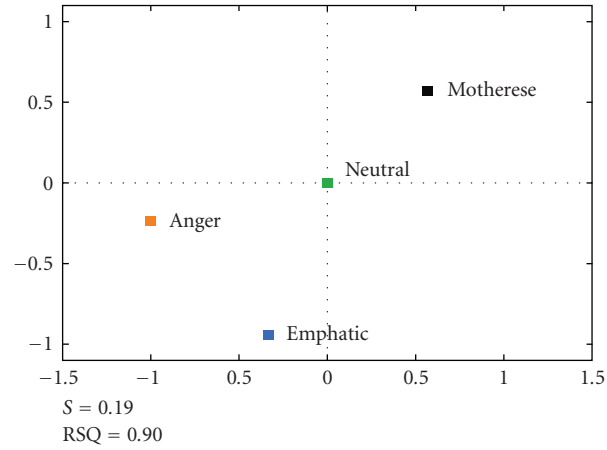

Figure 1: NMDS arrangement of the four cover classes in the 2-dimensional space based on the confusion matrix of the 5 human labellers.

2.3. Definition of Chunks. Finding the best unit of analysis has not posed a problem in studies involving acted speech with different emotions, using segmentally identical utterances, cf. for example, [27, 28]. 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, too much longer utterances. In [23], 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 [29]. Our Aibo 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 Aibo is only acting. Thus, it is not a "tidy" stimulus-response sequence that can be followed by tracking the very same channel. Since we are using only the audio channel of the children, we do not know what Aibo was doing at the corresponding time, or shortly before or after the child's utterance. (This information could be obtained from the video stream that has been recorded for control purposes. However, this information has not been used for chunking.) 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.

A reasonable strategy could be to segment the data in a preprocessing step into such units to be presented to the annotators for labelling emotions. However, this would require an a priori knowledge on how to define the optimal unit—which we do not have yet. In order not to decide beforehand on the units to be processed, we decided in favour of a word-based labelling: each word had to be annotated with one emotion label.

To better process the recordings of the children, the audio files have been split automatically into "turns" at pauses that are at least 1 second long. On average, these turns consist of 3.55 words. Based on the emotion labelling on the word level, emotion labels for turns can be defined without relabelling the whole corpus. A heuristic mapping algorithm is applied which is described in [22]. These turns can certainly be used for automatic speech recognition. Experimental results on

the impact of emotion-related states on the ASR performance using these automatically segmented turns are reported in [30]. Yet, a high inhomogeneity of the emotion-related states within one turn can be observed. The emotional homogeneity is defined as the proportion of raw labels, that is, the decisions of the five human labellers on the word level, that match the emotion label for the whole turn. Whereas the homogeneity is higher for short units and especially for words, larger units of analysis allow to model the context of the words within an utterance. Chunks—an intermediate unit between the word level and the turn level—are a good compromise between the length of the unit of analysis and the homogeneity of the emotion-related state within the unit and are an appropriate unit for ASR as well. For more details on the distribution of the inhomogeneity within turns and chunks, please see [22, Figure 5.18, page 106]. The emotional homogeneity can be taken as a measure of the prototypicality of the emotion. In [31] and [22, Table 7.20, page 172] it is shown how the automatic emotion recognition performance depends on the prototypicality of the chunks.

In our data, 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 Aibo's actions—or pauses segmenting into different dialogue acts—the child waits until he/she reacts to Aibo's actions. Thus, there is much overlap between two different channels: speech produced by the child and vision based on Aibo's actions, which is not used for our annotation. Hence, we 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 milliseconds. By that we try, for example, to tell apart vocatives ("Aibo") that simply function as "relators", from vocatives with specific illocutive functions meaning, for example, "Aibo" in the meaning of "*Hi, I'm talking to you*" or "Aibo!" in the meaning of "*Now I'm getting angry*" (illocution "command": "*Listen to me!*").

Note that in earlier studies, we found out that there is a rather strong correlation higher than 0.90 between prosodic boundaries, syntactic boundaries, and dialogue act boundaries (cf. [29]). Using only prosodic features to automatically classify syntactic or dialogue act boundaries results in a some 5% points lower classification performance compared to a classification based on syntactic or dialogue act information (e.g., information obtained from language models) [29]. 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 automatic speech recognition 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.

The syntactic and pause labels are explained in Table 1. Chunk boundaries are triggered by higher syntactic boundaries after main clauses (`s3`) and after free phrases (`p3`) and by boundaries between vocatives *Aibo Aibo* (`v2v1`) because, here, the second *Aibo* is most likely not simply a relator but is conveying specific illocutions (cf. above). Single instances of vocatives (`v1, v2`) are treated the same way as dislocations (`d2`). If the pauses at those lower syntactic boundaries that are given in Table 1, that is, `s2`, `d2`, `v1`, and `v2`, are at least 500 milliseconds long, we insert a chunk boundary as well. The syntactic boundaries `s3` and `s2` delimit "well-formed" clauses containing a verb; `p3` characterises not-well-formed units, functioning like clauses but without a verb. The boundary `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 pauses at words within all these units were defined as a nontriggering hesitation pauses. Each end-of-turn was redefined as triggering a clause/phrase boundary as well. Note that our turn-triggering threshold of 1 second works well because in the whole database, only 17 end-of-turn (`eot`) triggers were found that obviously denote within clause word boundaries. The boundary `s1` had to be introduced because the German word "*komm*" can function both as a sentence initial particle (corresponding to English "*Well, ...*") and an imperative (corresponding to English "*Come here! ...*"); only the imperative constitutes a clause. For more details on the chunking procedure and the evaluation of different chunking alternatives please see [32].

Table 1: Syntactic and pause labels.

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`	Post-vocative
`v1`	Prevocative
`v2v1`	Between "*Aibo*" instances
`0`	Pause 0–249 ms
`1`	Pause 250–499 ms
`2`	Pause 500–749 ms
`3`	Pause 750–1000 ms

If all 13 642 turns of the FAU Aibo Emotion Corpus are split into chunks, the chunk triggering procedure results in a total of 18 216 chunks, which consist of 2.66 words on average.

2.4. Definition of Emotion Labels for Chunks. A heuristic algorithm is used to map the original (raw) labels of the five human labellers on the word level onto one emotion label for the whole chunk. By simple majority voting we would not take into account two main characteristics of our data: firstly, the emotional intensity of our data is rather low due to the fact that we are not dealing with emotion portrayals but with naturally occurring emotions. Secondly, as mentioned, the user state ***E**mphatic* can be seen as some possible prestage of the other user state ***A**nger*.

In the following, the principles of the algorithm are explained. The details can be found in [22]. The algorithm

Table 2: Mapping of the emotion labels on the word level onto emotion labels for chunks: distribution of the emotion categories for the whole FAU Aibo Emotion Corpus.

		Chunk level			
Number of words		**M**	**N**	**E**	**A**[1]
Word level	Motherese	**1165**	94	1	0
	Neutral	298	**37 841**	806	224
	Emphatic	1	674	**1837**	16
	Angry	0	2	1	**81**
	Reprimanding	1	25	8	**201**
	Touchy	0	20	4	**276**
	Joyful	3	91	7	0
	b,h,s,r[2]	1	12	3	1
	No MV[3]	254	1186	1487	1780
	All	1723	39 945	4154	2579

[1] **M**: ***M****otherese*; **N**: ***N****eutral*; **E**: ***E****mphatic*; **A**: ***A****nger*.
[2] **B**ored, **h**elpless, **s**urprised, **r**est.
[3] No majority vote (MV) since less than three labellers agree.

is based on the raw labels of the cover classes ***M****otherese*, ***N****eutral*, ***E****mphatic*, and ***A****nger*. Any labels of the rare other classes are omitted. A chunk is labelled as belonging to ***N****eutral* if at least 60% of the raw labels are ***N****eutral*. If this is not the case, the number of labels ***M****otherese* is compared to the number of labels ***E****mphatic* and ***A****nger*. If ***M****otherese* has the majority and at least 40% of all raw labels within the chunk belong to ***M****otherese*, the chunk is labelled as ***M****otherese*. Otherwise, if there are more ***E****mphatic* and ***A****nger* labels than ***M****otherese* labels, the number of ***E****mphatic* labels is compared to the number of ***A****nger* labels. If there are more ***E****mphatic* labels and if at least 50% of all words within the chunk belong either to ***E****mphatic* or to ***A****nger*, the chunk is labelled as ***E****mphatic*. If it is the other way round, that is, if there are more ***A****nger* labels than ***E****mphatic* labels, the chunk is labelled as ***A****nger*. The different thresholds are defined heuristically by examining the resulting chunk labels.

Table 2 shows which emotion labels on the word level (majority vote of the five human labellers, 11 different user states) are mapped onto which emotion labels on the chunk level (the four cover classes ***M****otherese*, ***N****eutral*, ***E****mphatic*, and ***A****nger*). Note that the chunks of the cover classes ***M****otherese*, ***E****mphatic*, and ***A****nger* contain a considerable proportion of neutral words: 17.3% for ***M****otherese*, 19.4% for ***E****mphatic*, and 8.7% for ***A****nger*. Also the proportion of words where no absolute majority vote exists is very high, especially for ***E****mphatic* and ***A****nger*. Note that the number of words that belong to the cover class ***A****nger* is higher than the sum of the number of words that belong to *angry*, *reprimanding*, or *touchy/irritated*.

3. Emotional Speech Recognition

In this study, we are not interested in maximum word accuracy (WA) but in the impact of affect on the performance of an ASR system. Therefore, we do not evaluate ASR performance on large databases of children's speech but focus only on the FAU Aibo Emotion Corpus, which is rather small but thoroughly annotated with emotion labels. We focus on *two scenarios*.

(1) In the *first scenario*, we compare a standard speech recogniser trained on neutral speech with speech recognisers that are trained exclusively on speech of one emotion/emotion-related state.

(2) In the *second scenario*, we investigate how a standard speech recogniser trained on neutral speech can be improved by adding emotionally coloured speech.

For both scenarios, we use data of one school (Ohm) for training and the data of the other school (Mont) for testing our system. By that, strict speaker independence is guaranteed. To allow a fair comparison of different ASR systems, it is crucial that an equal amount of data is used for training. Therefore, we define the subsets Ohm_N, Ohm_M, Ohm_E, and Ohm_A, which are balanced with respect to the number of words: since the average number of words per chunk varies for the four emotion-relates states, these four subsets contain different numbers of chunks. The statistics are given in Table 3. The "size" of the subsets are given in terms of the number of chunks and the number of words. Additionally, the average number of frames and the average number of words per chunk is given. In general, emotional chunks consist of less words than neutral ones.

In the following, the selection/balancing of the data is described. The classes ***E****mphatic* and ***A****nger* are downsampled by choosing the chunks with the highest emotional homogeneity. The homogeneity is defined as the proportion of raw labels, that is, the decisions of the five human labellers on the word level, that match the emotion label of the whole chunk. There have been selected 772 (of 1289 available) chunks for ***E****mphatic* and 666 (of 721) chunks for ***A****nger*. Chunks of the classes ***E****mphatic* and ***A****nger* that are not included in Ohm_E and Ohm_A, respectively, are discarded for experiments presented in this paper. The samples of the subset Ohm_N (479 chunks) are chosen randomly from the 7383 available neutral chunks. The subset Ohm_base consists of the remaining neutral chunks. All 566 ***M****otherese* chunks fall into the Ohm_M subset. The selection strategies are different for the different emotional states because we aim at almost identical average prototypicality for the three subsets: Ohm_M (0.61), Ohm_E (0.62), and Ohm_A (0.62). Only for neutral speech, the average prototypicality is already clearly higher (0.79) as there are many chunks where all words can be clearly identified as neutral. Figure 2 shows the distribution of the prototypicality of the chunks for the four subsets Ohm_M, Ohm_N, Ohm_E, and Ohm_A.

The evaluation is carried out on the subset Mont. The four classes ***M****otherese*, ***N****eutral*, ***E****mphatic*, and ***A****nger* are highly unbalanced (cf. the subsets Mont_M, Mont_N, Mont_E, and Mont_A in Table 3). Mont_N makes up more than 80% of the test set; consequently, almost all words of Mont are contained in the vocabulary of Mont_N. For the evaluation, the unbalanced distribution is not a problem since we evaluate the ASR performance separately for the four states.

Table 3: Statistics of the various subsets of the FAU Aibo Emotion Corpus: training on the balanced subsets of Ohm, testing on the unbalanced subsets of Mont.

	Ohm_base	Ohm_M	Ohm_N	Ohm_E	Ohm_A
Number of chunks	6904	566	479	772	666
Avg. number of frames	179.5	174.5	184.9	161.9	185.4
Avg. number of words	2.81	2.39	2.82	1.75	2.03
Number of words	19 409	1354	1353	1354	1353

	Mont	Mont_M	Mont_N	Mont_E	Mont_A
Number of chunks	8257	158	6719	848	532
Avg. number of frames	169.8	151.1	170.8	158.4	181.5
Avg. number of words	2.69	2.35	2.86	1.91	2.02
Number of words	22 244	369	19 183	1619	1073

Table 4: Size of the vocabulary for the different training and test subsets of the FAU Aibo Emotion Corpus; training on the balanced subsets of Ohm, testing on the unbalanced subsets of Mont.

	Ohm_base	Ohm_M	Ohm_N	Ohm_E	Ohm_A
Number of word forms	653	111	180	93	111
Number of fragments	225	32	34	6	14
Total size	878	143	214	99	125

	Mont	Mont_M	Mont_N	Mont_E	Mont_A
Number of word forms	383	69	375	90	72
Number of fragments	158	9	147	9	9
Total size	541	78	522	99	81

For our experiments, we use the ASR engine that has been developed within the speech processing group at the University Erlangen-Nuremberg. A recent overview is given in [33]. The acoustic features are the first 12 standard MFCC features (the first MFCC coefficient is replaced by the sum of the energies of the 22 Mel filterbanks) and their first derivatives. The features are computed every 10 milliseconds over a Hamming window of 16 milliseconds. Our ASR system is based on semicontinuous hidden Markov models (SC-HMM) modelling polyphones, that is, an extension of the well-known triphones to model larger context sizes. A polyphone is modelled by its own HMM if it can be observed at least 50 times in the training set. All HMM states share the same set of Gaussian densities (codebook). By that, a smaller number of densities can be used, which is beneficial if—as in our case—only very little (emotional) training data is available. Yet, full covariance matrices are used in contrast to most systems based on continuous HMMs. We use Baum-Welch reestimation for training and Viterbi decoding. As language model we use back-off bi-grams.

Table 4 displays the size of the vocabulary across emotion-related states and schools. The vocabulary contains word forms as well as word fragments. Apparently, the size of the vocabulary depends on the emotion: the largest vocabulary is observed for ***N**eutral* speech, followed by emotional speech with lower intervariability. Furthermore, a higher vocabulary size is observed for school Ohm, which is a higher education level school. For all experiments, the vocabulary of the ASR systems is kept constant: it contains all word forms (813) of the complete FAU Aibo Emotion Corpus but no word fragments.

For the two scenarios outlined above, three types of experiments are carried out to evaluate the impact of affect on both the acoustic and the linguistic models. In the first experiment, the acoustic models are adapted whereas the linguistic models are kept fixed. In the second experiment, it is the other way round: only the linguistic models are adapted and the acoustic models are kept constant. Finally, both the acoustic and linguistic models are adapted.

3.1. Evaluation of Scenario 1. For the first scenario—comparing a "neutral" speech recogniser with "emotional" speech recognisers—the acoustic and linguistic models of the baseline system are trained on Ohm_N only. Since this subset is rather small, the size of the codebook had to be reduced drastically compared to our standard configuration. Setup experiments showed that a good ASR performance is achieved with 50 Gaussian densities. If evaluated on the different subsets of Mont—which contain only speech of one particular emotion/emotion-related state—the results shown in Table 5 (column "Ohm_N" of the upper table) demonstrate that speech produced in the state ***M**otherese* is recognised clearly worse (43.6% WA) than ***N**eutral* speech

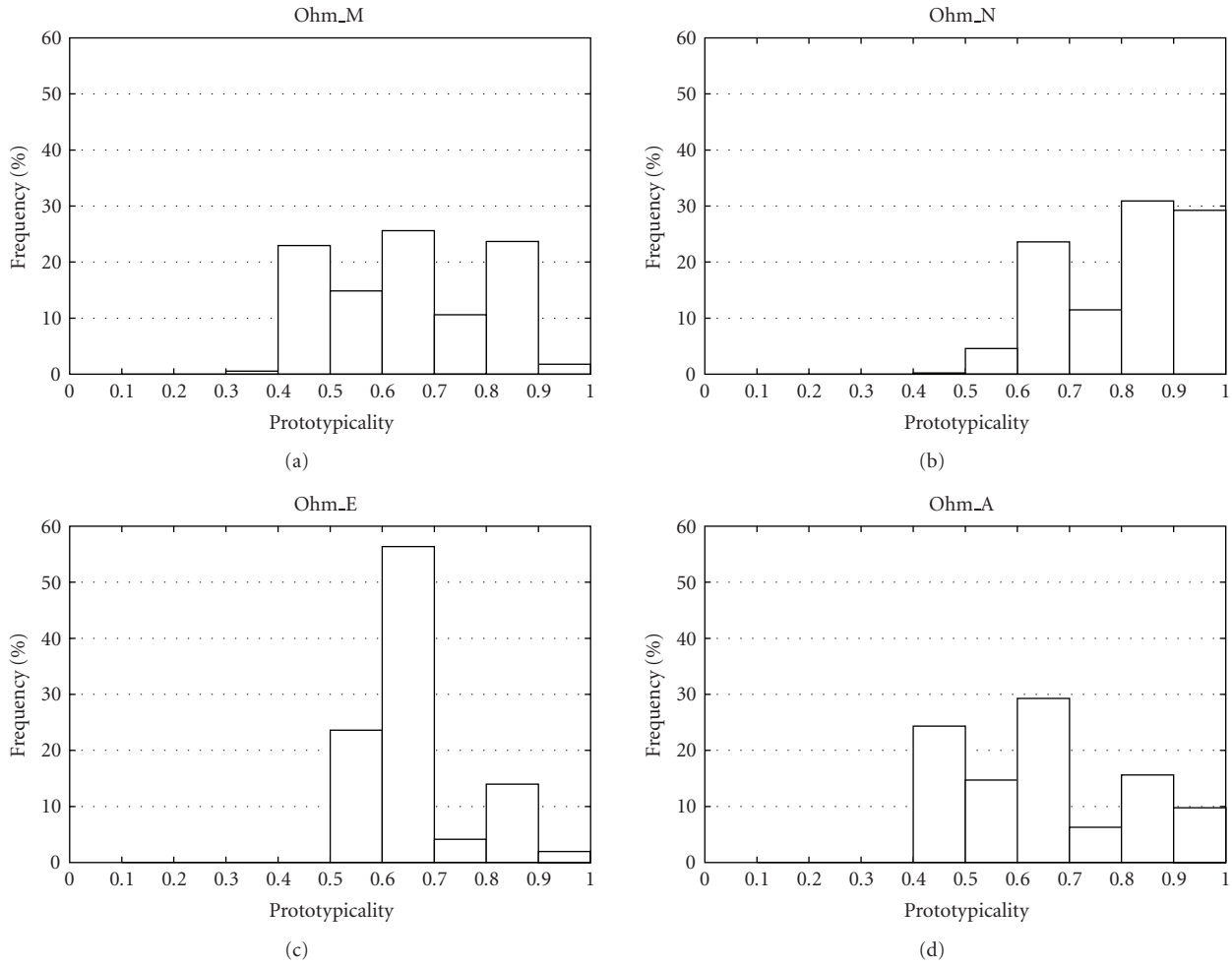

Figure 2: Distribution of the prototypicality of the chunks in the four training sets Ohm_N, Ohm_M, Ohm_E, and Ohm_A. The average level of prototypicality is 0.79 for Ohm_N, 0.61 for Ohm_M, and 0.62 for both Ohm_E and Ohm_A.

(60.3% WA). This is to be expected since the acoustic realisations as well as the linguistic content differ from the neutral training conditions. In contrast, speech produced in the states ***E**mphatic* and ***A**nger* is recognised even slightly better than neutral speech: 61.3% WA and 64.9% WA for ***E**mphatic* and ***A**nger*, respectively. This seems to derive from the fact that ***E**mphatic* and ***A**ngry* speech are articulated more clearly. ***E**mphatic* speech deviates from neutral speech: the child speaks in a pronounced, accentuated, and sometimes even hyperarticulated way. In our scenario, it can be conceived as a possible prestage of anger. Note that the cover class ***A**nger* subsumes three different emotion categories: *angry*, *reprimanding*, and *touchy/irritated*. The emotional intensity is in general rather low and the state is often not comparable to full-blown anger portrayed by actors. Hence, the acoustic realisations seem to differ from ***N**eutral* not as much as the ASR performance would suffer.

To adapt the acoustic models to emotional speech, the acoustic models are trained on Ohm_M, Ohm_E, and Ohm_A, respectively. The linguistic models are trained on Ohm_N and are the same for all three emotion-related states. The results are shown in the upper part of Table 5. The performance for ***E**mphatic* speech increases significantly ($\alpha = .001$) from 61.3% to 74.8% WA if the system is trained on ***E**mphatic* speech instead of neutral speech. Details on the significance test are given in Section 3.3. Training on Ohm_A helps to improve the performance for ***E**mphatic* speech as well albeit the improvement is lower: the performance increases from 61.3% to 67.2%. If the system is trained on Ohm_M, the performance for ***E**mphatic* speech drops to 42.8% WA. Similar results are obtained for speech produced in the state ***A**nger*: both ***A**ngry* and ***E**mphatic* speech help to improve the performance on Mont_A significantly (from 64.9% WA to 75.5% WA and 73.5% WA, resp., $\alpha = .001$), whereas the performance drops to 51.2% WA if the system is trained on Ohm_M. The performance on Mont_M cannot be improved if the system is trained on speech produced in the state ***M**otherese*. The reason might be that the speech in subset Ohm_M is too speaker specific since many instances of ***M**otherese* are produced by only a few speakers. The adapted system is probably more adapted to the acoustic characteristics of these speakers than to the state ***M**otherese* itself. Furthermore, it has to be noted that the test set (Mont_M) is rather small (see Table 3).

Table 5: Scenario 1: adaptation of the acoustic and linguistic models; results are given in terms of *word accuracy* (%). The baseline system (acoustic and linguistic models are trained on Ohm_N) is given in column "Ohm_N" and is identical in all three tables. "∅" denotes the arithmetic (unweighted) mean. The average of the four subsets weighted by the prior probabilities of the four classes is given in row "Mont."

	Acoustic models trained on			
Test set	Ohm_M	Ohm_N	Ohm_E	Ohm_A
Mont_M	**43.1**	**43.6**	34.2	32.8
Mont_N	44.9	**60.3**	54.0	55.8
Mont_E	42.8	**61.3**	**74.8**	67.2
Mont_A	51.2	**64.9**	75.5	**73.5**
∅	45.5	57.5	59.6	57.3
Mont	45.0	60.3	56.2	57.1
	Linguistic bigrams trained on			
Test set	Ohm_M	Ohm_N	Ohm_E	Ohm_A
Mont_M	**49.3**	**43.6**	37.4	38.8
Mont_N	56.0	**60.3**	58.0	59.9
Mont_E	56.3	**61.3**	**67.0**	67.0
Mont_A	60.1	**64.9**	68.0	**68.5**
∅	55.4	57.5	57.6	58.6
Mont	56.1	60.3	58.8	60.5
	Acoustic and linguistic models trained on			
Test set	Ohm_M	Ohm_N	Ohm_E	Ohm_A
Mont_M	**47.4**	**43.6**	32.0	30.6
Mont_N	40.8	**60.3**	52.6	54.7
Mont_E	35.7	**61.3**	**76.5**	70.2
Mont_A	46.0	**64.9**	75.3	**75.3**
∅	42.5	57.5	59.6	57.7
Mont	40.8	60.3	55.1	56.4

Table 6: Scenario 1: *perplexities* of the adapted linguistic models. The baseline system (linguistic models are trained on Ohm_N) is given in column "Ohm_N." "∅" denotes the arithmetic (unweighted) mean. The average of the four subsets weighted by the prior probabilities of the four classes is given in row "Mont."

	Linguistic models trained on			
Test set	Ohm_M	Ohm_N	Ohm_E	Ohm_A
Mont_M	**27.2**	**39.2**	87.4	74.5
Mont_N	38.4	**20.7**	35.8	30.3
Mont_E	31.4	**13.2**	**9.93**	12.4
Mont_A	24.7	**12.6**	12.3	**9.05**
∅	30.4	21.4	36.4	31.6
Mont	36.7	19.7	31.0	26.9

The middle part of Table 5 shows the results of the linguistic adaptation. The linguistic models are adapted by training on Ohm_M, Ohm_E, and Ohm_A, respectively, whereas the acoustic models are always trained on Ohm_N. Again, the performance for ***E**mphatic* and ***A**nger* can be improved by training the linguistic models on Ohm_E and Ohm_A, respectively. Nevertheless, the improvements are smaller than for the acoustic adaptation: the performance increases from 61.3% to 67.0% WA for ***E**mphatic* and from 64.9% to 68.5% WA for ***A**nger*. The improvements are significant at a significance level of 0.001 for ***E**mphatic* and 0.002 for ***A**nger*, respectively. The same improvement for ***E**mphatic* can be obtained if the linguistic models are trained on Ohm_A instead of Ohm_E. Vice versa, linguistic models trained on Ohm_E yield nearly the same improvement for ***A**nger* compared to the models trained on Ohm_A. Obviously, the states ***E**mphatic* and ***A**nger* differ more with respect to their acoustic realisations than with respect to their language models. In contrast, language models trained on Ohm_M are not suited for the word recognition of ***E**mphatic* and ***A**nger* but they are helpful to improve the performance on Mont_M: there, the word accuracy increases from 43.6% to 49.3%. This improvement is significant at a level of 0.05. The performance of an ASR system is always a combination of the influence of the acoustic models and the linguistic models. In order to show the pure impact of the linguistic adaptation on the language models, the results of the linguistic adaptation are reported in Table 6 in terms of the perplexity of the language model. The perplexities are evaluated on the test set Mont and its subsets. After adaptation to the state ***M**otherese*, the perplexity on Mont_M falls from 39.2 to 27.2. If the linguistic models are adapted to ***E**mphatic*, the perplexity on Mont_E decreases from 13.2 to 9.93. If they are adapted to ***A**nger*, the perplexity on Mont_A decreases from 12.6 to 9.05. In terms of the perplexity, the differences between ***E**mphatic* and ***A**nger* are more obvious than in terms of the word accuracy: adaptation to the state ***A**nger* also helps to reduce the perplexity on Mont_E, but the reduction is rather small (from 13.2 to 12.4). Adaptation to the state ***E**mphatic* reduces to perplexity on Mont_A only from 12.6 to 12.3.

In the next experiments, both the acoustic and language models are adapted. The results are reported in the lower part of Table 5. They demonstrate that for ***E**mphatic* and ***A**nger* the improvements of the acoustic adaptation can be further increased by additionally adapting the language models. For ***E**mphatic* the best result that could be obtained is 76.5% WA compared to the baseline of 61.3% WA. For ***A**nger*, the best result is 75.3% WA compared to 64.9% WA in the baseline system. Both improvements are significant at $\alpha = .001$. However, ***E**mphatic* speech has obviously the higher potential for improvements. For ***M**otherese*, the result of the combination of the acoustic and linguistic adaptation is worse than the result of the linguistic adaptation only. This is not surprising since—as mentioned above—the acoustic adaptation alone already resulted in a worse word recognition performance.

The results of all three adaptation methods are summarised in Table 9. They show that the adaptation to one specific emotion yields higher word accuracies for this particular emotion at the expense of higher word error rates for the other emotions. The (unweighted) average word

Table 7: Scenario 2: adaptation of the acoustic and linguistic models; results are given in terms of *word accuracy* (%). The baseline system (acoustic and linguistic models trained on Ohm_base) is given in column "Ohm_base" and is identical in all three tables. "∅" denotes the arithmetic (unweighted) mean. The average of the four subsets weighted by the prior probabilities of the four classes is given in row "Mont."

	Acoustic models trained on				
Test set	Ohm_base (baseline)	Ohm_base + 2x Ohm_M	Ohm_base + 1x Ohm_N	Ohm_base + 3x Ohm_E	Ohm_base + 2x Ohm_A
Mont_M	**65.0**	**64.5**	61.5	59.9	61.3
Mont_N	**77.3**	77.6	**77.5**	77.1	78.0
Mont_E	**81.0**	81.3	80.5	**83.1**	81.2
Mont_A	**79.2**	80.2	78.8	81.4	**83.6**
∅	75.6	75.9	74.6	75.4	76.0
Mont	77.5	77.7	77.5	77.4	78.2
	Linguistic models trained on				
Test set	Ohm_base (baseline)	Ohm_base + 28x Ohm_M	Ohm_base + 1x Ohm_N	Ohm_base + 28x Ohm_E	Ohm_base + 28x Ohm_A
Mont_M	**65.0**	**65.9**	64.5	64.0	64.5
Mont_N	**77.3**	77.0	**77.4**	77.7	77.7
Mont_E	**81.0**	80.1	80.8	**81.6**	81.9
Mont_A	**79.2**	78.9	79.0	79.9	**81.6**
∅	75.6	75.5	75.4	75.8	76.4
Mont	77.5	77.1	77.5	77.8	78.0
	Acoustic models trained on				
	Ohm_base (baseline)	Ohm_base + 0x Ohm_M	Ohm_base + 1x Ohm_N	Ohm_base + 3x Ohm_E	Ohm_base + 2x Ohm_A
	Linguistic models trained on				
Test set	Ohm_base (baseline)	Ohm_base + 28x Ohm_M	Ohm_base + 1x Ohm_N	Ohm_base + 28x Ohm_E	Ohm_base + 28x Ohm_A
Mont_M	**65.0**	**65.9**	61.5	60.4	59.1
Mont_N	**77.3**	77.0	**77.6**	77.4	78.4
Mont_E	**81.0**	80.1	80.4	**84.4**	83.1
Mont_A	**79.2**	78.9	78.7	81.6	**85.1**
∅	75.6	75.5	74.6	76.0	76.4
Mont	77.5	77.1	77.6	77.8	78.7

accuracy over all four emotion-related states (denoted as "∅" in Table 5) remains nearly constant if the neutral acoustic and/or linguistic models are adapted to speech produced in the states ***E**mphatic* or ***A**nger*. If the acoustic models are adapted to ***M**otherese*, the average word accuracy drops clearly. If the a priori probabilities of the four different emotion-related states are taken into account, that is, the word accuracy is evaluated on the whole test set Mont, the best results in terms of the weighted average word accuracy are achieved if the acoustic models are trained on neutral speech due to the high a priori probability of the state ***N**eutral* (cf. Table 3).

3.2. Evaluation of Scenario 2. In the second scenario, the ASR performance for emotionally coloured speech is tried to be improved by adding emotionally coloured data to a baseline speech recogniser that is trained on neutral speech. For this purpose, the acoustic and linguistic models of the baseline system are trained on Ohm_base. Due to the size of Ohm_base, the codebook of the baseline system now contains 500 Gaussian densities—ten times more than the ASR systems trained for Scenario 1. The larger size of Ohm_base compared to Ohm_N yields clearly higher word accuracies on Mont as shown in Table 7 (column "Ohm_base" of the upper table). ***N**eutral* speech is now recognised with 77.3% WA compared to 60.3% in Scenario 1. Speech produced in the state ***E**mphatic* is recognised best (81.0% WA), followed closely by ***A**nger* (79.2%). ***M**otherese* is still recognised clearly worse (65.0% WA) than ***N**eutral* speech. Hence, the ranking—the negative states ***E**mphatic* and ***A**nger* on the top, ***N**eutral* in the middle, and ***M**otherese* on the bottom—is the same in both scenarios.

Table 8: Scenario 2: *perplexities* of the adapted linguistic models. The baseline system (linguistic models are trained on Ohm_base) is given in column "Ohm_base". "∅" denotes the arithmetic (unweighted) mean. The average of the four subsets weighted by the prior probabilities of the four classes is given in row "Mont."

	Linguistic models trained on				
Test set	Ohm_base (baseline)	Ohm_base + 2x Ohm_M	Ohm_base + 1x Ohm_N	Ohm_base + 3x Ohm_E	Ohm_base + 2x Ohm_A
Mont_M	**24.4**	**20.8**	24.3	32.3	29.0
Mont_N	**14.9**	18.5	**14.8**	17.2	16.8
Mont_E	**10.2**	14.1	10.2	**8.28**	9.77
Mont_A	**9.87**	12.7	9.86	9.64	**7.66**
∅	14.8	16.5	14.8	16.9	15.8
Mont	14.2	17.8	14.2	15.9	15.6

Again, the acoustic models and the linguistic models are adapted separately before their combination is evaluated. The upper part of Table 7 shows the results of the adaptation of the acoustic models. Certainly, there are different well-known strategies such as MAP and MLLR to adapt the acoustic models of a speech recogniser to new data. Due to the small amounts of emotionally coloured data, we preferred to adapt the acoustic models of the speech recogniser by adding emotionally coloured data (Ohm_M, Ohm_N, Ohm_E, and Ohm_A) to the training data of the baseline system (Ohm_base). Best results were not obtained by adding the emotionally coloured data once, but several times increasing the weight of the new data. In experiments not reported here, the best factor has been optimised. For ***N**eutral*, the optimal factor is 1. This makes sense since the training data of the baseline system is already ***N**eutral* speech. The optimal factor is 3 for ***E**mphatic* and 2 for ***A**nger*. The ASR performance cannot be increased any further by adding the new data more often. It actually decreases if the factor is too high. By that, the performance on Mont_A can be increased significantly ($\alpha = .001$) by adding Ohm_A twice from 79.2% to 83.6% WA. Adding Ohm_E also helps to improve the performance on Mont_A albeit the improvements are lower. The best improvement on Mont_A adding Ohm_E (81.4% WA) is achieved if Ohm_E is added three times. The performance on Mont_E can be increased from 81.0% to 83.1% WA by adding Ohm_E three times. This improvement is significant at a level of 0.05. Even better results (83.9% WA) are obtained by adding Ohm_A once to Ohm_base (results not shown in Table 7). The slight increase of the performance on Mont_N by adding Ohm_N once is not significant. As for Scenario 1, the adaptation of the acoustic models could not improve the speech recognition results for ***M**otherese*. Instead, the word accuracy slightly drops probably due to speaker adaptation instead of the adaptation to the state ***M**otherese* itself. The least (nonsignificant) decrease is obtained by adding Ohm_M twice.

The results of the linguistic adaptation are shown in the middle part of Table 7. In contrast to the adaptation of the acoustic models, the emotionally coloured data has to be added much more often. Best results for ***M**otherese*, ***E**mphatic*, and ***A**nger* are obtained, if twice as much (in terms of the number of words) emotionally coloured data is added to Ohm_base, that is, a factor of 28. Naturally, the optimal factor for ***N**eutral* is 1 since Ohm_base already consists of ***N**eutral* speech and (almost) no new information about the state ***N**eutral* is added. However, the improvements of the word accuracy are rather small and not significant: on Mont_M from 65.0% to 65.9% by adding Ohm_M, on Mont_N from 77.3% to 77.4% by adding Ohm_N, and on Mont_E from 81.0% to 81.6% by adding Ohm_E. A bigger and significant improvement ($\alpha = .001$) is only achieved on Mont_A (from 79.2% to 81.6% WA) by adding Ohm_A. Again, the pure influence on the language models is given in terms of the perplexity of the language models in Table 8. Since the language models are trained on more data, the perplexities are in general lower compared to the ones of the first scenario. After adaptation to the state ***M**otherese*, the perplexity on Mont_M decreases from 24.4 to 20.8. Adapting to ***E**mphatic* reduces the perplexity on Mont_E from 10.2 to 8.28. If the language models are adapted to ***A**nger*, the perplexity on Mont_A is 7.66 compared to 9.87 in the baseline system. As it could be observed in the first scenario, the differences between ***E**mphatic* and ***A**nger* are more obvious in terms of the perplexity than in terms of the word accuracy. In terms of the perplexity, the best adaptation results are always obtained, if data of the same state is used for adaptation of the language models; that is, although ***A**nger* also helps to reduce the perplexity on Mont_E (from 10.2 to 9.77), the best adaptation results are obtained with ***E**mphatic* speech (8.28). However, the improvements on Mont_E in terms of the word accuracy were not significant and the adaptation to ***A**nger* even resulted in a better word accuracy on Mont_E (81.9%) than the adaptation to ***E**mphatic* (81.6).

In the last experiments shown in the bottom part of Table 7, the combined adaptation of acoustic and linguistic models is carried out. By that, the improvements obtained by the acoustic adaptation can be increased further. After the adaptation, ***N**eutral* is recognised with a word accuracy of 77.6%. The recognition of speech produced in the states ***E**mphatic* and ***A**nger* can profit significantly from the adaptation of both the acoustic and linguistic models compared to the baseline system: the word accuracy for ***E**mphatic* speech is now 84.4% compared to 81.0% of the baseline system and the one for ***A**nger* is now 85.1% compared to the baseline of 79.2% WA. In this second scenario, the models for ***A**ngry* speech could profit more by the adaptation than the ones for

Table 9: Summary of the performance gains by adaptation of the baseline system to the three emotion-related states ***M**otherese*, ***E**mphatic*, and ***A**nger*. The performance is given in terms of word accuracy (WA) (%). The filled bullets indicate at which level the improvements w.r.t. the baseline system are significant.

(a) Scenario 1: "neutral versus emotional ASR engine"

	M	E	A
Baseline system	43.6	61.3	64.9
Adapted systems			
Acoustic models	43.1 ○ ○ ○ ○ ○	74.8 ● ● ● ● ●	73.5 ● ● ● ● ●
Linguistic models	49.3 ● ○ ○ ○ ○	67.0 ● ● ● ● ●	68.5 ● ● ● ● ○
Both	47.4 ○ ○ ○ ○ ○	76.5 ● ● ● ● ●	75.3 ● ● ● ● ●

(b) Scenario 2: "adaptation of neutral ASR engine"

	M	E	A
Baseline system	65.0	81.0	79.2
Adapted systems			
Acoustic models	64.5 ○ ○ ○ ○ ○	83.1 ● ○ ○ ○ ○	83.6 ● ● ● ● ●
Linguistic models	65.9 ○ ○ ○ ○ ○	81.6 ○ ○ ○ ○ ○	81.6 ● ● ● ● ●
Both	65.9 ○ ○ ○ ○ ○	84.4 ● ● ● ● ●	85.1 ● ● ● ● ●

Levels of significance (adjusted according to [34]):
● ○ ○ ○ ○ 0.05 ● ● ○ ○ ○ 0.01 ● ● ● ○ ○ 0.005 ● ● ● ● ○ 0.002 ● ● ● ● ● 0.001.

***E**mphatic* speech. The gap between ***E**mphatic* and ***A**nger* on the one hand and ***N**eutral* on the other hand has widened clearly. The results of the adaptation are summarised in the bottom part of Table 7.

3.3. Significance Tests for ASR. The significance between pairs of ASR recognition scores has been investigated by applying the *matched-pairs t-test*. According to [35], this statistical test gives accurate results when (1) the recognition of pairs of utterances is carried out under almost identical conditions, (2) the errors made by the two ASR engines in different utterances are independent, and (3) the number of utterances is sufficiently large. All these conditions are certainly met.

The test can be briefly described as follows. For each couple of utterances transcribed by two ASR systems, the Levenshtein distance of each of them to the (same) reference is computed. This step basically coincides to the alignment needed for computing the word accuracies. Then, the significance of the difference between these two sequences is examined by applying a one tailed t-test: we want to see if the proposed algorithm is *better* than the baseline one. The test is eventually accomplished for all couples of experiments sharing the same test segments. To cope with the *multiplicity effect*, that is, the chances of getting an increased number of significant results due to multiple tests, we adjusted the α values as described in [34].

Table 9 summarises the performance gains obtained by the adaptation of the baseline system to the speech of the three emotion-related states ***M**otherese*, ***E**mphatic*, and ***A**nger* and shows which of the improvements are significant at a level of at least 0.05.

4. Feature Space Visualisation

To visualise the ASR feature space, the 12-dimensional static MFCC feature vectors are averaged over all words produced by one speaker in a particular emotion-related state. By averaging over all MFCC feature vectors of all words, the average MFCC feature vector contains not only acoustic but also linguistic information to some degree. Based on the Euclidean distances between the average MFCC feature vectors, a Sammon transformation [36] is applied to map the points from the original, 12-dimensional feature space to a low-dimensional space with—in the presented case—only two dimensions. The Sammon mapping performs a topology preserving reduction of data dimension by minimising a stress function between the topology of the low-dimensional Sammon map and the high-dimensional original data. More details can be found in [22, 37].

Thus, each point in Figure 3(a) represents one speaker in a particular emotion. The four emotion-related states form four different clusters in this two-dimensional space. As can be seen, ***N**eutral* speech is found "in the middle" of the projected MFCC space and is most compactly clustered compared to emotional speech. The other three clusters are located around the ***N**eutral* one. For a better visualisation, the clusters are modelled by two-dimensional Gaussian probability density functions. These are illustrated in Figure 3(b) by their mean vector and an ellipse representing their covariance matrix. ***E**mphatic* speech also forms a rather compact cluster with almost no overlap with the other three clusters. In contrast, speech produced in the state ***A**nger* shows a clearly higher acoustic variability resulting in a large overlap between the cluster of ***A**nger* and the ones of ***M**otherese* and ***N**eutral*. The overlap with ***M**otherese* is partly due to the acoustic similarity of *reprimanding*, which is mapped onto ***A**nger*, and *motherese*. The highest variability in the MFCC space can be observed for ***M**otherese* speech; this in turn explains why it is difficult to recognise it robustly.

5. Conclusion and Discussion

Our results demonstrate the difficulty of the automatic recognition of children's speech, especially in the case of spontaneous and affective speech. The evaluation shows clearly that affect *does* affect recognition of children's speech. Thereby, ***E**mphatic* and ***A**ngry* speech is recognised best—even better than **N**eutral speech, although the baseline ASR system is trained on ***N**eutral* speech only. The reasons could be that *emphatic* speech or speech produced in slight forms of *anger* is articulated clearly and that the acoustic realisations are obviously quite similar to those of *neutral* speech. This does not hold for ***M**otherese* speech resulting in high word error rates.

The ASR performance can be increased by adaptation of the acoustic and linguistic models. Best results are obtained for speech produced in the states ***E**mphatic* and ***A**nger*. Training material consisting of ***E**mphatic* speech—*emphatic* being a prestage of *anger*—does not only help to improve the recognition of ***E**mphatic* but also helps to increase the performance on ***A**nger*. Vice versa, speech produced in ***A**nger*

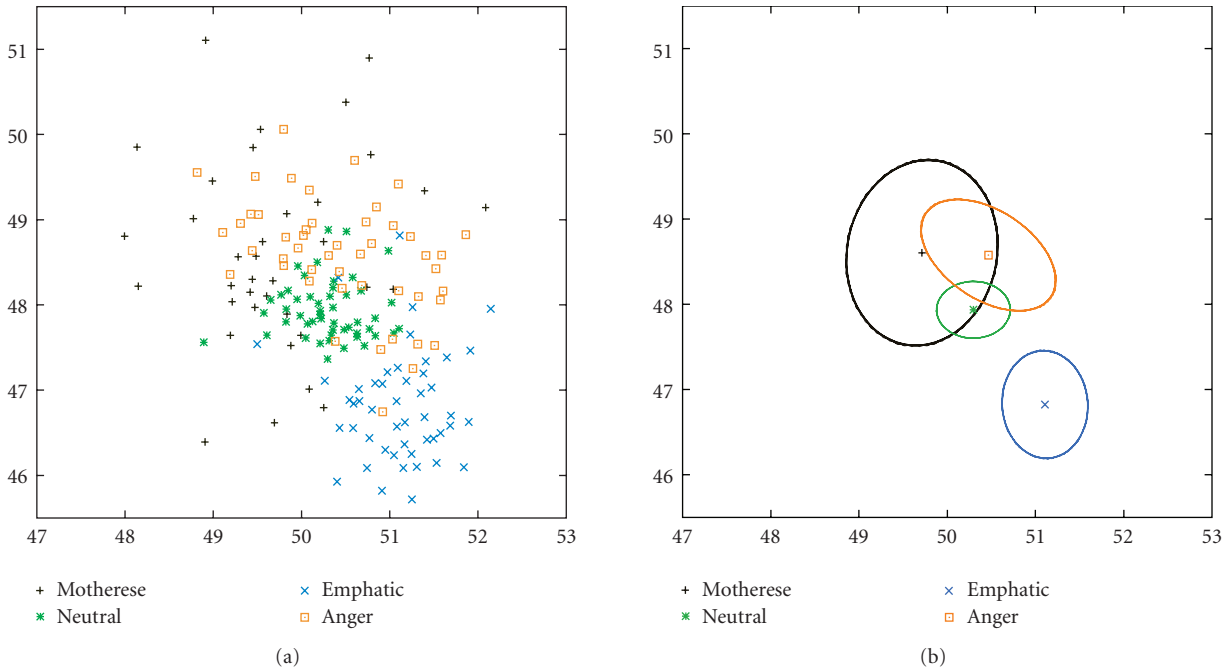

Figure 3: Visualisation of the distribution of emotions in a high-to-low-dimensional Sammon transformation of the MFCC space: each point (a) represents speech of one speaker in one particular emotion-related state. The four emotion-related states form clusters that are modelled by Gaussian densities (b).

also helps to improve the ASR performance on ***Emphatic*** speech. For speech produced in ***Motherese***, the adaptation of the acoustic models was not successful—probably due to the high interspeaker variability and the dominance of one single speaker. However, the results could be improved by adaptation of the linguistic models.

Whereas the ASR performance on speech of a particular emotion-related state could be improved by the adaptation to this particular state, the performance on speech produced in other states decreased in general. Hence, an emotion classification module could be used to dynamically select an emotion dependent speech recogniser such that matched conditions between the training and the testing of the speech recogniser are preserved.

ASR performance is influenced by many factors. For this study, we have tried to keep as many factors as possible constant. We have defined subsets of equal size for each emotional state. The average prototypicality of the emotional states is comparable for the three subsets Ohm_M, Ohm_E, and Ohm_A; only for Ohm_N the average prototypicality is higher. The experiments show that ***Neutral*** speech is recognised worse than speech produced in the states ***Emphatic*** and ***Anger***. This is certainly not due to the higher prototypicality of the ***Neutral*** chunks. However, the influence of prototypicality on the ASR performance has not been studied yet. For all experiments in this study, speech recognisers have been trained that have the same vocabulary. However, the four different subsets of the test set differ with respect to the size of the vocabulary that is actually used in the different emotional states. As this is spontaneous speech, this factor cannot be controlled. It remains unclear how ASR performance is affected by these different vocabularies. It may be that words of the vocabulary that are acoustically similar can be more often observed in the state ***Neutral*** than in the other two states ***Emphatic*** and ***Anger***. Furthermore, the acoustic realisations of ***Motherese*** in the training set seemed to be too different from those in the test set such that the acoustic models could not be adapted successfully.

Acknowledgments

This work originated in the CEICES initiative (**C**ombining **E**fforts for **I**mproving automatic **C**lassification of **E**motional user **S**tates) taken in the European Network of Excellence HUMAINE. The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under Grant agreement no. 211486 (SEMAINE), the projects PF-STAR under Grant IST-2001-37599, and HUMAINE under Grant IST-2002-50742. The responsibility lies with the authors.

References

[1] A. Hagen, B. Pellom, and R. Cole, "Children's speech recognition with application to interactive books and tutors," in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU '03)*, pp. 186–191, St. Thomas, Virgin Islands, USA, 2003.

[2] A. Hagen, B. Pellom, and R. Cole, "Highly accurate children's speech recognition for interactive reading tutors using subword units," *Speech Communication*, vol. 49, no. 12, pp. 861–873, 2007.

[3] S. Narayanan and A. Potamianos, "Creating conversational interfaces for children," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 2, pp. 65–78, 2002.

[4] J. G. Wilpon and C. N. Jacobsen, "A study of speech recognition for children and the elderly," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '96)*, vol. 1, pp. 349–352, Atlanta, Ga, USA, 1996.

[5] S. Das, D. Nix, and M. Picheny, "Improvements in children's speech recognition performance," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '98)*, vol. 1, pp. 433–436, Seattle, Wash, USA, 1998.

[6] A. Potamianos, S. Narayanan, and S. Lee, "Automatic speech recognition for children," in *Proceedings of the 5th European Conference on Speech Communication and Technology (Eurospeech '97)*, pp. 2371–2374, Rhodes, Greece, September 1997.

[7] M. Blomberg and D. Elenius, "Collection and recognition of children's speech in the PF-Star project," in *Proceedings of the 16th Swedish Phonetics Conference (FONETIK '03)*, pp. 81–84, Umeå, Sweden, 2003.

[8] M. Wöllmer, F. Eyben, J. Keshet, A. Graves, B. Schuller, and G. Rigoll, "Robust discriminative keyword spotting for emotionally colored spontaneous speech using bidirectional LSTM networks," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '09)*, pp. 3949–3952, Taipei, Taiwan, 2009.

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

[10] 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.

[11] J. Gustafson and K. Sjölander, "Voice transformations for improving children's speech recognition in a publicly available dialogue system," in *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP '02)*, pp. 297–300, Denver, Colo, USA, September 2002.

[12] B. Schuller, J. Stadermann, and G. Rigoll, "Affect-robust speech recognition by dynamic emotional adaptation," in *Proceedings of Speech Prosody*, Dresden, Germany, 2006.

[13] 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.

[14] C. Busso, S. Lee, and S. S. Narayanan, "Using neutral speech models for emotional speech analysis," in *Proceedings of the 8th Annual Conference of the International Speech Communication Association (Interspeech '07)*, vol. 4, pp. 2304–2307, Antwerp, Belgium, 2007.

[15] H. L. Hansen, *Analysis and compensation of stressed and noisy speech with application to robust automatic recognition*, Ph.D. thesis, Georgia Institute of Technology, Atlanta, Ga, USA, 1988.

[16] H. J. M. Steeneken and J. H. L. Hansen, "Speech under stress conditions: overview of the effect on speech production and on system performance," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '99)*, pp. 2079–2082, Phoenix, Ariz, USA, 1999.

[17] 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.

[18] S. Yildirim, C. M. Lee, S. Lee, A. Potamianos, and S. Narayanan, "Detecting politeness and frustration state of a child in a conversational computer game," in *Proceedings of the 9th European Conference on Speech Communication and Technology (Eurospeech '05)*, pp. 2209–2212, Lisbon, Portugal, September 2005.

[19] S. Arunachalam, D. Gould, E. Andersen, 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. 2675–2679, Aalborg, Denmark, September 2001.

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

[21] S. M. D'Arcy, L. P. Wong, and M. J. Russell, "Recognition of read and spontaneous children's speech using two new corpora," in *Proceedings of the 8th International Conference on Spoken Language Processing (ICSLP '04)*, Jeju Island, South Korea, October 2004.

[22] S. Steidl, *Automatic Classification of Emotion-Related User States in Spontaneous Speech*, Logos, Berlin, Germany, 2009, http://www5.cs.fau.de/en/our-team/steidl-stefan/dissertation/.

[23] 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.

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

[25] S. Steidl, M. Levit, A. Batliner, E. Nöth, and H. Niemann, "'Of all things the measure is man': Automatic classification of emotions and inter-labeler consistency," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '05)*, vol. 1, pp. 317–320, Philadelphia, Pa, USA, March 2005.

[26] J. B. Kruskal, "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis," *Psychometrika*, vol. 29, no. 1, pp. 1–27, 1964.

[27] 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*, pp. 1517–1520, Lisbon, Portugal, September 2005.

[28] 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 5th European Conference on Speech Communication and Technology (Eurospeech '97)*, pp. 1695–1698, Rhodes, Greece, September 1997.

[29] 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.

[30] 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 (WOCCI '08)*, Chania, Greece, October 2008.

[31] S. Steidl, B. Schuller, A. Batliner, and D. Seppi, "The hinterland of emotions: facing the open-microphone challenge," in

Proceedings of International Conference on Affective Computing and Intelligent Interaction (ACII '09), pp. 690–697, Amsterdam, The Netherlands, September 2009.

[32] A. Batliner, D. Seppi, S. Steidl, and B. Schuller, "Segmenting into adequate units for automatic recognition of emotion-related episodes: a speech-based approach," to appear in *Advances in Human-Computer Interaction*.

[33] G. Stemmer, *Modeling Variability in Speech Recognition*, Logos, Berlin, Germany, 2005.

[34] S. L. Salzberg, "On comparing classifiers: pitfalls to avoid and a recommended approach," *Data Mining and Knowledge Discovery*, vol. 1, no. 3, pp. 317–328, 1997.

[35] L. Gillick and S. J. Cox, "Some statistical issues in the comparison of speech recognition algorithms," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '89)*, vol. 1, pp. 532–535, Glasgow, UK, 1989.

[36] J. Sammon, "A nonlinear mapping for data structure analysis," *IEEE Transactions on Computers*, vol. 18, no. 5, pp. 401–409, 1969.

[37] T. Haderlein, D. Zorn, S. Steidl, E. Nöth, M. Shozakai, and M. Schuster, "Visualization of voice disorders using the sammon transform," in *Proceedings of the 9th International Conference on Text, Speech and Dialogue (TSD '06)*, vol. 4188 of *Lecture Notes in Computer Science*, pp. 589–596, Springer, Brno, Czech Republic, September 2006.