Does Affect Affect Automatic Recognition of Children's Speech?

Björn Schuller
Institute for Human-Machine Communication
Technische Universität München
Munich, Germany
schuller@tum.de

Anton Batliner,
Stefan Steidl
Lehrstuhl für Mustererkennung
Friedrich-Alexander-Universität
Erlangen, Germany
batliner@informatik.uni-erlangen.de

Dino Seppi
Fondazione Bruno Kessler
irst
Trento, Italy
seppi@fbk.eu

ABSTRACT

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 these phenomena: extensive test-runs are carried out for 1k vocabulary continuous speech recognition on spontaneous angry, motherese and emphatic children's speech as opposed to neutral speech. The experiments mainly address the questions how specific emotions influence word accuracy, and whether neutral speech material is sufficient for training as opposed to matched conditions acoustic model adaptation. In the result emphatic and angry speech are best recognised, while neutral speech proves a good choice for training. For the discussion of this effect we further visualise emotion distribution in the MFCC space by Sammon transformation.

1. INTRODUCTION

Offering a broad variety of applications, such as literacy and reading tutors [9], speech interfaces for children are an attractive object of research [11]. However, automatic speech recognition (ASR) is known to be a challenge for the recognition of children's speech [5]: characteristics of both acoustics and linguistics differ from those of adults [7], e.g. by higher pitch and formant positions or not yet perfectly developed co-articulation. At the same time, these strongly vary for children of different ages due to anatomical and physiological development [10] and learning effects.

Apart from children's speech, also affective speech can be challenging for ASR [13], as acoustic parameters differ considerably under the influence of affect. These two problems 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 [3, 1]. We therefore investigate the influence of emotion on the recognition of children's speech. As opposed to previous work [6], we study the effect of each of four emotion-related states individually to answer the two main questions: how does a particular affect affect recognition, and is training acoustic models with neutral speech adequate, as mostly done? The paper is structured as follows: in section 2 we introduce the database used and discuss the mapping from words onto turns with respect to emotion, in section 3, 4 and 5 we present experimental results, an explanation by visualisation of the acoustic space, and conclusions.

2. AFFECTIVE CHILDREN'S SPEECH

The database used is the German FAU Aibo Emotion Corpus, a corpus with recordings of children communicating with a pet robot; it is described in more detail in [4].

The general framework for this spontaneous database is child–robot–communication, and the elicitation of emotion–related speaker states. The robot is Sony's (dog-like) AIBO robot. The basic idea has been to combine a corpus of children's speech with 'natural' emotional speech within a Wizard-of-Oz task. The speech is intended to be 'natural' because children do not disguise their emotions to the same extent as adults do. However, it is of course not fully 'natural' as it might be in a non-supervised setting. Furthermore the speech is spontaneous, because the children were not told to use specific instructions but to talk to the AIBO like they would talk to a friend. In this experimental design, the child is led to believe that the AIBO is responding to his or her 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.

2.1 Speech Recording

The data was collected from 51 children (age 10 - 13 years, 21 male, 30 female) from two different schools ('Mont' and 'Ohm'); the recordings took place in the respective classrooms. Speech was transmitted with a wireless head set (Shure UT 14/20 TP UHF series with microphone WH20 TQG) and recorded with a DAT-recorder (sampling rate 48 kHz, quantisation 16 bit, down-sampled to 16 kHz). While each recording session took around 30 minutes, the total

From: Proc. 1st Workshop on Child, Computer and Interaction, Chania, Greece, 2008

amount of speech equals 9.2 hours of speech after removing the pauses. This derives from a huge amount of silence due to reaction time of the AIBO.

2.2 Emotional Word Labelling

Five labellers (advanced students of linguistics) listened to the recordings and annotated independently from each other each word as neutral (default) or as belonging to one of ten other classes. 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*, i.e., irritated (225), *angry* (84), *motherese* (1260), *bored* (11), *reprimanding* (310), *rest*, i.e. non-neutral, but not belonging to the other categories (3), *neutral* (39169). 4707 words had no MV; all in all, there were 48401 words.

The state *emphatic* has to be commented on especially: based on our experience with other emotion databases [2], 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, re-formulations, other wordings, or simply the use of a pronounced, marked speaking style. Such a style does thus 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 – 'computer talk' – that some people use while speaking to a computer, like speaking to a non-native, to a child, or to an elderly person who is hard of hearing. Thus the fact that *emphatic* can be observed can only be interpreted meaningfully if other factors are considered. There is a further – practical – argument for the annotation of *emphatic* in our respect: if the labellers are allowed to label *emphatic* it might be less likely that they confuse it with other user states.

Some of the labels are very sparse. If we only take labels with more than 50 MVs, this 7-class problem is most interesting from a methodological point of view. However, the distribution of classes is very unequal. Therefore, we down-sampled *neutral* and *emphatic* and mapped *touchy* and *reprimanding*, together with *angry*, onto ***Angry*** as representing different but closely related kinds of negative attitude. (The initial letter is given boldfaced; this letter will be used in the following for referring to these cover classes. Note that now, ***Angry*** can consist, for instance, of two *touchy* and one *reprimanding* label; thus the number of ***Angry*** cases is far higher than the sum of *touchy*, *reprimanding*, and *angry* MV cases.). This more balanced 4-class problem, which we refer to as **MNEA**, consists of 1224 words for ***Motherese*** (**M**), 1645 for ***Neutral*** (**N**), 1645 words for ***Emphatic*** (**E**), and 1557 words for ***Angry*** (**A**) [14]. Cases where less than three labellers agreed were omitted as well as those cases where other than these four main classes were labelled. Inter-labeller correspondence is dealt with in [14]; weighted kappa for multi-raters is 0.59 for these four classes.

2.3 Mapping onto the Turn Level

These word-based labels were mapped onto turn-based labels yielding the numbers of instances per emotion and school depicted in Table 1. A turn is thereby simply obtained by automatic cutting at pause lengths greater or equal 1 s.

For the mapping onto turn-based labels, we employed the following strategy: fragments and auxiliaries are used as stop words. In this way, using the turns only that contain our 6071 **MNEA** words, we obtained 17611 words in 3990 turns (6 turns, respectively 7 words were discarded as they contain only stop words). Stop words were 564 fragments and 196 auxiliaries (some were both); this results in 16854 words remaining. Note that of course, we could find some more stop words, but this would be rather data driven and not generic so we refrained from that. For six turns containing only stop words, no turn-based labels were generated. For each turn, we add together the labels given by our 5 labellers (for n words, 5 x n labels). For the turns to be mapped onto neutral, 70% of the labels have to be neutral; *joyful* and the other spurious labels are not taken into account for this computing. If 30% or more are non-neutral, then the turn is **M**, **E**, or **A**. If at least 50% of the non-neutral labels are **M**, the turn is mapped onto **M**. If **A** and **E** are equally distributed, the turn is mapped onto **A**. If the turn is neither **M** or **A**, it is **E**. This simply means that we employ a sort of 'markedness' condition: **M** is more marked than **A**, and **A** is more marked than **E**, and all are more marked than **N**.

More details are described fully in [4]. This subset will be referred to as the "turn set" of the full FAU Aibo Emotion Corpus, in the following denoted as Aibo turn set.

Table 1: Distribution of turns among emotions and schools for the Aibo turn set

#turns	Mont	Ohm	Σ	
M	123	372	495	(12.4 %)
N	670	610	1280	(32.1 %)
E	576	771	1347	(33.8 %)
A	369	499	868	(21.7 %)
{M, N, E, A}	1738	2252	3990	(100.0 %)

Table 2 depicts the distribution of words mapped onto turns by their originally labelled emotion on word-level. As can be seen from the number of ***Neutral*** words per turn, a typical turn labelled as emotional consists of a considerable percentage of ***Neutral*** words (last line in the table). It seems obvious that this is in particular true for ***Emphatic*** speech, as usually only few words in a turn will be emphasised. This table also depicts the number of words per turn and emotion. ***Neutral*** turns are the longest in terms of the number of words, followed by ***Motherese*** and ***Emphatic***. ***Angry*** turns tend to be rather short.

Table 3 displays the size of the vocabulary across emotions and schools. Apparently, the size of the vocabulary is dependent on the emotion: in the case of ***Neutral*** speech it is highest, followed by emotional speech with lower inter-variability. Further a higher vocabulary size is observed for the Ohm school, which is a higher education level school.

Table 2: Mapping words onto turns: distribution of emotions; Aibo turn set

		turn level			
	#words	**M**	**N**	**E**	**A**
word level	**motherese**	**1134**	62	50	13
	neutral	1046	**6507**	3126	844
	emphatic	13	213	**1739**	59
	angry	16	42	133	**1430**
	repremanding	2	0	1	0
	joyful	2	10	1	0
	-	154	59	461	494
#words		2367	6893	5511	2840
#turns		495	1280	1347	868
#words/turn		4.8	5.4	4.1	3.3
N words/turn [%]		44.2	94.4	56.7	29.7

Table 3: Size of the vocabulary across emotions and schools for the Aibo turn set

#entries	**M**	**N**	**E**	**A**	**{M,N,E,A}**
Mont	99	250	139	107	316
Ohm	190	430	238	173	596
{Mont,Ohm}	220	514	276	206	698

3. CHILDREN'S SPEECH RECOGNITION

For our experiments, we use an ASR engine based on continuous hidden Markov models (HMM): [15] a 30 ms Hamming window is applied with 50% overlap to extract the MFCC coefficients 0-12 and their first and second order regression coefficients. We use a tied-state acoustic model (AM) with 41 phonemes, and 1979 back-off triphones. Three states and five Gaussian mixtures per state proved to be the optimal parameterisation of the phoneme models. Note that we train exclusively on the FAU Aibo Emotion Corpus (and the Aibo turn set, respectively), as we are not interested in maximum accuracy, but rather in the effect of affect. We use Baum-Welch reestimation for training and Viterbi decoding. As language model (LM) we use back-off bi-grams. Both AM and LM are trained and tested speaker independently on data of one school, exclusively. Note that better results are obtained for testing on MONT, as more instances are available for training, and the vocabulary size is lower. In all experiments the LM is kept fixed: we focus on the impact of affect on the AM.

In the following experiments we want to shed light on the effect of individual affects: the AM is therefore trained exclusively with turns belonging to one emotion. Tests are carried out separately for each emotion. Table 4 shows the word accuracies (WA) for testing independently of speaker and school. Training is considered in particular with ***Neutral*** and ***Emphatic*** speech, as these are more or less balanced with respect to instances (cf. Table 2). Results with training on ***Motherese*** and ***Angry*** speech fell comparably behind due to data sparseness. However, this reflects the true distribution: there simply will be more neutral words available in most application scenarios. The following ranking can be observed: best recognised is ***Emphatic*** and ***Angry*** speech, followed by ***Neutral***, and least ***Motherese*** speech. This seems to derive from the fact that ***Emphatic*** and ***Angry*** speech are well articulated. This is in accordance with findings and explanations in [6], where children's emotional speech as a whole was compared with children's neutral speech.

Considering the impact of training on different emotions, it can be summarised that the best overall choice is to train on ***Neutral*** speech (mean over all test emotions 59.75% WA), next on ***Emphatic*** (mean 59.06% WA), then ***Angry*** (mean 53.82% WA), and finally ***Motherese*** (mean 51.33% WA) speech. This seems not exclusively to be dependent on the amount of speech available for training the AM, as in the case of ***Emphatic*** speech, which is best recognised if trained on itself, though having only roughly 80% of words for training if compared to ***Neutral*** speech (cf. Table 2). Interestingly, in the case of training on OHM and testing on MONT, it is also optimal for the recognition of ***Angry*** speech to train on ***Emphatic*** speech. However, for ***Motherese*** and ***Neutral*** speech ***Neutral*** speech seems to be the best choice for training. And apart from the fact that for ***Neutral*** speech most material is available, this certainly also derives from the fact that – as shown before – even in a turn that is labelled emotional as a whole, a considerable number of ***Neutral*** words are contained. These ***Neutral*** words are already well modelled by training on ***Neutral*** speech. The fact that ***Emphatic*** speech is second best for training lies well in line with this explanation: it is the emotion with the highest frequency of ***Neutral*** words.

4. ACOUSTIC SPACE VISUALISATION

Figure 1 shows the benefit of training on ***Neutral*** speech: we visualise the distribution of emotions in the MFCC space by Sammon transformation to a 2D space [12]. 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. The latter topology is defined by the distances between emotions or speakers [8]: therefore the MFCC features as used for the ASR engine are averaged over the 51 children per emotion and speaker. Next, a distance matrix is calculated by Euclidean distance between the mean vectors of all speakers. Clusters are preserved by the subsequent Sammon transformation.

As can be seen, ***Neutral*** speech is found "in the middle" of the projected MFCC space and most compactly clustered compared to emotional speech. As all other clusters lay around ***Neutral*** speech and as ***Neutral*** speech possesses the largest overlap with any other type of emotional speech or cluster, ***Neutral*** speech forms the optimal subset for training if only one emotion is available. ***Motherese*** speech on the other hand shows the highest acoustic variability in the MFCC space; this in turn explains why it is difficult to be recognised robustly.

Table 4: Word Accuracies (WA) training the AM on Ohm and testing on Mont, Aibo turn set. Baseline using all turns of Ohm for training: 71.00% WA

WA [%]	Train **N**	Train **E**
Test **M**	**50.09**	39.66
Test **N**	**53.16**	51.13
Test **E**	69.65	**71.10**
Test **A**	65.03	**69.01**
mean	**59.75**	**59.06**

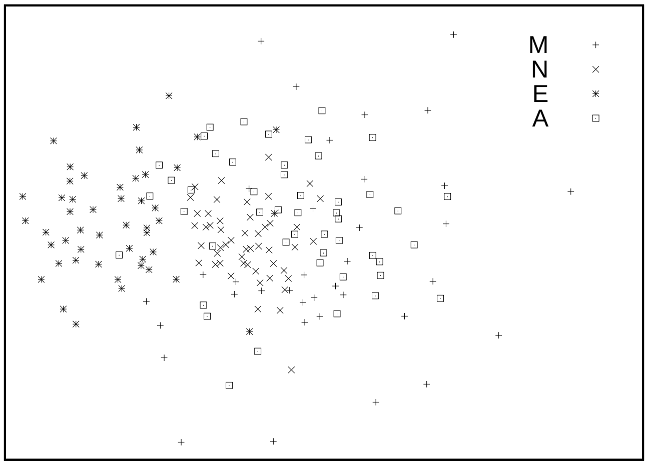

Figure 1: Visualisation of the distribution of emotions in a high-to-low dimensional Sammon-transform of the MFCC space: framewise MFCC 0-12 calculation, averaging over the 51 speakers per emotion and calculation of the matrix of Euclidean distances between static mean vectors of all speakers.

5. CONCLUSION

Our results demonstrate the difficulty of recognising children's speech, especially in the case of spontaneous and affective speech: independent of the speaker, 71.00% WA are obtained with optimal parameters and maximum training material. ***E****mphatic* and ***A****ngry* speech is thereby recognised best, followed by ***N****eutral*; worst recognition is found for ***M****otherese* speech. This finding is independent of the differences in the amount of training material per emotion.

Summing up and answering our original questions, affect *does* affect recognition of children's speech. However, and surprisingly, it seems to be indeed sufficient to some extent to train on ***N****eutral* speech, even if confronted with ***A****ngry* or ***M****otherese* children's speech. For ***E****mphatic* and ***A****ngry* speech, however, a slight gain could be obtained by training on ***E****mphatic* speech – being a pre-stage of ***A****ngry* speech.

In future work we aim at adaptation of neutrally trained acoustic models for emotional speech. This could be combined with emotion recognition to dynamically adapt to the present specific emotion – resembling matched condition. Further, investigation of effects of affect on the language model seems an interesting topic.

6. ACKNOWLEDGEMENTS

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), and the projects PF-STAR under grant IST-2001-37599, and HUMAINE under grant IST-2002-50742. The responsibility lies with the authors.

7. REFERENCES

[1] S. Arunachalam, D. Gould, E. Andersen, D. Byrd, and S. Narayanan. Politeness and frustration language in child-machine interactions. In *Proc. of Eurospeech 2001*, pages 2675–2679, Aalborg, Denmark, 2001.

[2] A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Nöth. How to find trouble in communication. *Speech Communication*, 40:117–143, 2003.

[3] A. Batliner, S. Steidl, C. Hacker, and E. Nöth. Private emotions vs. social interaction — a data-driven approach towards analysing emotions in speech. *User Modeling and User-Adapted Interaction*, 18:175–206, 2008.

[4] A. Batliner, S. Steidl, B. Schuller, D. Seppi, K. Laskowski, T. Vogt, L. Devillers, L. Vidrascu, N. Amir, L. Kessous, and V. Aharonson. Combining efforts for improving automatic classification of emotional user states. In *Proc. IS-LTC 2006*, Ljubliana, 2006.

[5] M. Blomberg and D. Elenius. Collection and recognition of children's speech in the pf-star project. In *Proc. of Fonetik 2003*, pages 81–84, Umeå, Sweden, 2003.

[6] S. M. D'Arcy, L. P. Wong, and M. J. Russell. Recognition of read and spontaneous children's speech using two new corpora. In *Proc. of ICSLP 2004*, Jeju Island, South Korea, 2004.

[7] D. Giuliani and M. Gerosa. Investigating recognition of children's speech. In *Proc. of ICASSP 2003*, volume 2, pages 137–140, Hong Kong, China, 2003.

[8] T. Haderlein, D. Zorn, S. Steidl, E. Nöth, M. Shozakai, and M. Schuster. Visualization of voice disorders using the sammon transform. In *Proc. of Text, Speech and Dialogue (TSD) 2006*, pages 589–596, Brno, Czech Republic, 2006. Springer, LNAI 4188.

[9] A. Hagen, B. Pellom, and R. Cole. Highly accurate children's speech recognition for interactive reading tutors using subword units. *Speech Communication*, 49(12):861–873, 2007.

[10] S. Lee, A. Potamianos, and S. Narayanan. Acoustic of children's speech: Developmental changes of temporal and spectral parameters. *Journal of the Acoustic Society of America (JASA)*, 105(3):1455–1468, 1999.

[11] S. Narayanan and A. Potamianos. Creating conversational interfaces for children. *IEEE Transactions on Speech and Audio Processing*, 10(2):65–78, 2002.

[12] J. Sammon. A nonlinear mapping for data structure analysis. *IEEE Trans. Computers*, C-18:401–409, 1969.

[13] B. Schuller, J. Stadermann, and G. Rigoll. Affect-robust speech recognition by dynamic emotional adaptation. In *Proc. of Speech Prosody 2006*, Dresden, Germany, 2006.

[14] 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 *Proc. ICASSP 2005*, Philadelphia, U. S. A., 2005.

[15] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland. *The HTK book (v3.4)*. Cambridge University Press, Cambridge, UK, 2006.