In: Proc. Int. Conf. on Acoustics, Speech, and Signal Processing ICASSP 2009,
Taipei, Taiwan, pp. 4585–4588. IEEE, 2009. 19.-24.04.2009

EMOTION RECOGNITION FROM SPEECH: PUTTING ASR IN THE LOOP

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

[1]Institute for Human-Machine Communication, Technische Universität München, Germany
schuller@IEEE.org
[2]Lehrstuhl für Mustererkennung, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany
batliner@informatik.uni-erlangen.de
[3]Polderland Language & Speech Technology, Nijmegen, The Netherlands, dino@polderland.nl

ABSTRACT

This paper investigates the automatic recognition of emotion from spoken words by vector space modeling vs. string kernels which have not been investigated in this respect, yet. Apart from the spoken content directly, we integrate Part-of-Speech and higher semantic tagging in our analyses. As opposed to most works in the field, we evaluate the performance with an ASR engine in the loop. Extensive experiments are run on the FAU Aibo Emotion Corpus of 4k spontaneous emotional child-robot interactions and show surprisingly low performance degradation with real ASR over transcription-based emotion recognition. In the result, bag of words dominate over all other modeling forms based on the spoken content, directly.

Index Terms— Speech analysis, Speech recognition, Feature extraction

1. INTRODUCTION

The recognition of emotion from speech is mostly realized by means of acoustic feature analysis. However, also the spoken content is well known to carry information on the speaker's emotion [1]. This is usually reflected in the usage of certain words or grammatical alterations - which means in turn, in the usage of specific semantic and pragmatic entities. A number of approaches exist for this analysis: e.g. key-word and phrase spotting [2], rule-based modeling [3], Semantic Trees [4], Latent Semantic Analysis [5], World-knowledge-Modeling [6], among many other more handcrafted solutions. However, two methods prevail, presumably because they are shallow representations of linguistic knowledge and have already been frequently employed in automatic speech processing: *(class-based) N-Grams*, e.g. [7] and *vector space modeling* [8, 1]; these methods will be dealt with in the following sections. Moreover, we introduce string kernels as novel solution in the field.

Considering emotion analysis from spoken text, only few results for emotion recognition rely on ASR output [8] rather than on manual transcription of data. Likewise we had shown the gain obtained by supplementing acoustic features with linguistic ones in [1] based on transcription. However, more knowledge is needed on how actual ASR downgrades performance with respect to individual linguistic feature types.

To actually recognize emotion from text, first a vocabulary has to be established. This vocabulary usually needs to be be reduced somehow, by discarding stop-words. Data-driven approaches as Salience or Information Gain-based reduction are popular. The easiest, yet often effective way, is also stopping by the general minimum frequency of occurrence within a training corpus, as done herein. Further stemming - i.e. clustering of morphological variants as flexions (e.g. by declination or conjugation) reduces the number of entries in the vocabulary while at the same time providing more training instances per class. Thereby also words that were not seen in the training can be mapped upon lexemes, as by simple N-Gram Stemming, or by (Iterated) Lovins, as herein, that bases on suffix lists and rules for their application. A very compact approach to stemming is the use of so called Part-of-Speech (POS) classes, such as nouns, verbs, adjectives, particles [1]. Also *sememes*, i.e. semantic units represented by lexemes, can be clustered into higher semantic concepts such as generally positive or negative terms [1]. In addition, non-linguistic vocalizations like sighs and yawns, laughs, cries, and coughs can easily be integrated into the vocabulary [1]. Next, adequate numerical modeling is required to carry out the actual recognition process. This will be explained in the following two sections for the methods considered, namely vector space modeling in sec. 2 and string kernels in sec. 3. Subsequently we present the data used for experiments in sec. 4 and the Automatic Speech Recognition (ASR) process in sec. 5 prior to results in sec. 6 and the conclusion in sec. 7.

2. VECTOR SPACE MODELING

Bag-of-Words (BOW), a form of vector space modeling, is a well-known numerical representation form of text in automatic document categorization. It has been successfully ported to recognize sentiments or emotion in [8]. Each word in the vocabulary adds a dimension to a linguistic vector representing the term frequency within the actual utterance. Note that usually large feature spaces occur, which requires some kind of feature space reduction. This can be obtained by removing stop words and stemming (for details on works on this set cf. [9]). To smooth Pareto or Zipf distributed feature vectors, frequencies are log-transformed. Furthermore, the term frequency is normalized by utterance length and w.r.t. the overall term frequency within the training corpus. Generally most vector elements resemble zero, as feature vectors are constructed for short utterances rather than for longer texts, as in document retrieval, and only few words of the vocabulary are seen. Support Vector Machines (SVM) show high performance for this task. The possibility of early fusion with acoustic features helped make this technique very popular in speech-based emotion recognition [1].

Alternatively, N-grams are applied for emotion recognition to model sequences of words. Following Zipf's principle of least effort stating that irrelevant function words occur very frequently opposing terms of interest that occur sparsely, the number of considered words is reduced to small N in order to prevent over-modeling. Due to

In: Proc. Int. Conf. on Acoustics, Speech, and Signal Processing ICASSP 2009,
Taipei, Taiwan, pp. 4585–4588. IEEE, 2009. 19.-24.04.2009

the typical data sparseness in emotion recognition, mostly uni-grams ($N = 1$) have been applied so far [10], besides bi-grams ($N = 2$) and tri-grams ($N = 3$) [11].

However, only N-Grams with $N > 1$ overcome the missing modeling of word order lost by BOW. As they do not allow for easy integration in the acoustic vector, we suggest to combine these methods by Bag-of-N-Grams (BONG). Thereby the frequency of N-Grams (with $N > 1$) is counted rather than that of words ($N = 1$). Note that by "backing off" we can combine BOW and BONG with diverse N.

3. STRING KERNELS

Alternatively, the string kernel approach makes use of a mapping from text information to a high dimensional feature space without explicit calculation of features. String kernels have proved to be a promising approach in similar tasks like text classification, cross-language document matching, authorship attribution, and text clustering. Based on the theory of SVM, the idea of kernel mapping is extended for strings as input parameters. Thus, a special kernel for text information is used, called the string subsequence kernel (SSK). The idea behind string kernels is to observe small substrings in a given string. For a predefined substring length, all possible substrings form a feature space in which a string (the spoken utterance) can be represented. The numeric value of each feature depends on the substring frequency in the string and on the degree of contiguity. For example the substring "int" exists in the word "international" as well as in "experiment", but with different degree of contiguity. This degree of contiguity is weighted by a decay factor $\lambda \in [0, 1]$ which penalizes non-contiguous substrings. Taking non-continuous substrings into account is a specific characteristic of the string kernel method, not supplied by other approaches.

The transformation of a string s into the feature space is done by a mapping $\Phi(s)$ which can be calculated numerically as described in [12]. In analogy to SVM theory, this mapping does not have to be done explicitly. An implicit calculation is done by using a kernel function:

$$K^{\Phi}(s, t) = \langle \Phi(s), \Phi(t) \rangle \ . \quad (1)$$

This kernel function is part of the decision function for SVM. The inner product calculated by the kernel can be seen as a numeric measure of similarity between two strings s and t. The calculation of this string subsequence kernel can further be simplified due to recursive computation [12], making the procedure practicable.

In order to speed up the computation, which can be quite time consuming for huge databases, we use the lambda pruning technique. This approach is a trade-off between runtime and approximation accuracy and is done by introducing another parameter for the string kernel, called maximum subsequence length Θ. The parameter determines the maximum length to which non-continuous substrings are observed. For a substring length of $l_{substring}$, the Θ is set to $\Theta = 3 \cdot l_{substring}$ which yields to a good trade-off between speed and accuracy. The decay factor λ is always set to 0.5.

4. FAU AIBO EMOTION CORPUS

The database used is the German FAU Aibo Emotion Corpus, a corpus with spontaneous and emotional speech recordings of children communicating with a pet robot; it is described in more detail in [1]. 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 class-rooms. 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, quantization 16 bit, down-sampled to 16 kHz). While each recording session took around 30 minutes, the total amount of speech equals 9.2 hours of speech after removing the pauses. time of the AIBO.

#turns	MONT	OHM	{MONT, OHM}	
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 1. Distribution of turns among emotions and schools.

Five labelers (advanced students of linguistics) listened to the recordings and annotated independently from each other each word with respect to emotion. We resort to majority voting (henceforth MV): if three or more labelers agree, the label is attributed to the word. 4707 words had no MV; all in all, there were 48401 words. However, the distribution of classes is very unequal. Therefore, we down-sampled to a more balanced 4-class problem, which we refer to as **MNEA**: it consists of 1224 words for ***M**otherese* (**M**), 1645 for ***N**eutral* (**N**), 1645 words for ***E**mphatic* (**E**), and 1557 words for ***A**ngry* (**A**) . Note that **E** is a pre-stage of **A**, and **E** by definition is a sort of hyper-articulated speech. Throughout, **A** is rather a "pedagogical" **A**—the children are *not* really angry and fully aroused. Weighted kappa for multi-raters is 0.59 for these four classes. 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, more details are described in [1]. 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 2 depicts the distribution of words mapped onto turns. As can be seen from the number of ***N**eutral* words per turn, a typical turn labeled as emotional consists of a considerable percentage of ***N**eutral* words. It seems obvious that this is in particular true for ***E**mphatic* speech, as usually only few words in a turn will be emphazised. This table also depicts the number of words per turn and emotion. ***N**eutral* turns are the longest in terms of the number of words, followed by ***M**otherese* and ***E**mphatic*. ***A**ngry* turns tend to be rather short. Finally, table 2 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 ***N**eutral* 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.

5. SPEECH RECOGNITION

For our experiments, we use an ASR engine based on continuous hidden Markov models (HMM)[13]: 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

In: Proc. Int. Conf. on Acoustics, Speech, and Signal Processing ICASSP 2009,
Taipei, Taiwan, pp. 4585–4588. IEEE, 2009. 19.-24.04.2009

	M	N	E	A	MNEA
#words (w)	2367	6893	5511	2840	17611
#turns (t)	495	1280	1347	868	3990
#w/t	4.8	5.4	4.1	3.3	4.4
N w/t [%]	44.2	94.4	56.7	29.7	65.4
#v(MONT)	99	250	139	107	316
#v(OHM)	190	430	238	173	596
#v(MONT,OHM)	220	514	276	206	698

Table 2. Distribution of emotions and size of the vocabulary (v) across emotions and schools for the Aibo turn set

Train	Test	WA [%]
MONT	OHM	63.5
OHM	MONT	71.0
All		**67.7**

Table 3. Baseline word accuracy (WA) for speaker independent cross-validation training on all of MONT (6653) and testing on all of OHM (6989) turns and vice versa, full FAU Aibo database

tri-phones. Three states and five Gaussian mixtures per state proved to be the optimal parameterization of the phoneme models. Note that we train exclusively on the FAU Aibo Emotion Corpus (and the Aibo turn set, respectively). We use Baum-Welch re-estimation for training and Viterbi decoding. As language model (LM) for ASR 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 order to recognize emotion from spoken content, we first deal with the baseline performance for the generally demanding task of recognizing spontaneous and affective speech: to obtain an upper performance benchmark using only the dataset at hand, we train on all available turns from the one school and test on all turns from the other school, and vice-versa. That way, we ensure maximum learning material while preserving utmost realistic conditions: speakers and acoustic conditions are fully independent. Note that these are considerably more turns than used in the ongoing emotion recognition, as we can employ also turns that could not be assigned an emotional label by a minimum labeler agreement of 3 out of 5. As a general mean of performance evaluation, we use word accuracies (WA), as in [13]. Table 3 shows the according WA and mean (μ) WA for this task. The difference in performance between the two schools is clearly seen. The mean WA of 67.7% demonstrates the difficulty of this task: affective children's speech having word fragments and non-verbal vocalizations such as laughter. However, it can be raised up to a mean WA of 76.9% by speaker adaptation (unpublished experiments). Yet, here we are interested in absolute speaker independence. Note that also a difference in WA per emotion exists: in [13] the following ranking was observed: best recognized is ***E**mphatic* and ***A**ngry* speech, followed by ***N**eutral*, and least ***M**otherese* speech. This seems to derive from the fact that ***E**mphatic* and ***A**ngry* speech are well articulated.

6. EXPERIMENTAL RESULTS

We now describe a number of experiments for the actual recognition of emotion from spoken text, comparing emotion recognition using ASR output with an upper benchmark, i.e. emotion recognition using the spoken word chain. Note that only the first best ASR result is used directly for linguistic analysis in search of affective cues. More elaborate approaches could use acoustic confidences or lattice structures for potentially higher robustness. Further note that in this work, we focus on realistic conditions and refrain from prototyping the dataset as done in almost any other work. We can obtain considerably higher accuracies by picking more prototypical examples as shown in [14], however this does simply not reflect the real-life use-case situation [15]. Table 4 shows results for SVM as classifier with polynomial (upper half) and string kernels (lower half), respectively. In the case of polynomial kernels, we first translate the string into a numeric representation by vector space modeling. In the case of string kernels, the string is used directly. As in section 5 we train and test cross-wisely school vs. school to ensure maximum independence. Reported are means. We consider four cases, each: 1) direct usage of the recognized/transcribed string (Iterated Lovins Stemming and minimum term frequency stopping are applied), 2) Part-of-Speech tagging into six classes based on a lexicon look-up table (in [9] we deal with rather negotiable effects of automatic tagging), 3) as 1) but with 3-Gram tokenization, and 4) higher 3 class semantic tagging by look-up as described above. Furthermore we consider diverse combinations of these by gathering the feature types into super-vectors. As a measurement we use the recognition rate (RR) and the harmonic mean between RR and the unweighted mean of class-wise recall rates F as introduced in [1, 15] to better represent the in-balance of the data-set. In the rightmost double column we show the difference Δ between ground truth and ASR output. Note that in one case this difference is negative for RR yet positive for F due to shifted preferences of majority or minority classes.

As a result we can observe that string kernels fall behind in the majority of cases. Surprisingly, however, the difference between transcription and ASR based is comparably low in most cases, though ASR has a considerable word error rate over 30% (cf. above). The impact of ASR differs significantly between the two different Kernel functions with respect to the different feature types investigated, thus not allowing for easy interpretation which feature type is more robust against ASR confusions. However, a ranking between feature types can be carried out being partly in line with our findings in [16]: first comes using the full string (BOW), then tri-grams (BONG), then first POS prior to higher semantics.

7. CONCLUSION

Altogether, the results show that the string kernel analysis seems not to be an interesting alternative to vector space modeling, especially considering the high computational costs, even if lambda pruning is applied. The only remarkable exception is, maybe, that tri-grams in combination with string kernels perform better than in combination with polynomial kernels. Also, the "incompatibility" with acoustic features for early fusion speaks against this type of analysis. Moreover it is impossible to separate unimportant from meaningful features by applying a feature selection.

Using genuine ASR however led to surprisingly low performance loss, though ASR of emotional speech is a challenge and provides high error rates.

Whether bags-of-(back-off)-N-Grams are advantageous will need more investigation. However, the full spoken word chain clearly outperforms any other representation form, even if it comes to real ASR. The other representation forms (higher semantics and part-of-speech tagging) however helped to slightly improve overall results in combination with the spoken word chain.

In: Proc. Int. Conf. on Acoustics, Speech, and Signal Processing ICASSP 2009,
Taipei, Taiwan, pp. 4585–4588. IEEE, 2009. 19.-24.04.2009

Type				Transcript		ASR		Δ	
B	3	P	H	*RR*	*F*	*RR*	*F*	*RR*	*F*
Polynomial Kernel									
√	-	-	-	48.9	47.2	46.1	44.3	2.8	2.9
-	√	-	-	44.5	41.2	43.3	39.7	1.2	1.4
-	-	√	-	41.1	39.3	34.0	31.7	7.1	7.6
-	-	-	√	35.8	34.9	27.5	26.8	8.3	8.0
√	-	√	-	49.9	48.2	47.1	45.0	2.8	3.2
√	-	-	√	49.3	47.6	47.1	44.6	2.2	3.1
√	-	√	√	49.4	47.9	47.3	45.3	2.1	2.6
-	√	√	√	50.3	48.6	43.8	41.8	6.5	6.9
√	√	√	√	49.9	48.2	47.6	45.6	2.3	2.6
String Kernel									
√	-	-	-	47.7	46.2	43.8	41.2	3.9	5.0
-	√	-	-	47.1	45.1	46.6	43.8	0.5	1.3
-	-	√	-	41.9	39.9	42.9	39.4	-1.0	0.4
-	-	-	√	32.5	31.8	30.0	27.8	2.5	4.0
√	-	√	-	45.8	43.3	45.2	41.6	0.6	1.7
√	-	-	√	49.7	47.1	45.5	43.3	4.2	3.8
√	-	√	√	47.7	44.0	47.6	42.8	0.1	1.2

Table 4. Selected results: Vector Space Modeling with Polynomial Kernel (upper half) and direct string processing with String Kernel (lower half) - ground truth by transcript vs. ASR, Bag-of-Words (B), Part-of-Speech (P), Bag-of-Tri-Grams (3), Higher Semantics (H), and some interesting combinations thereof [%].

In future work we will investigate usage of knowledge sources to cope with "out-of-vocabulary" (OOV) occurrences with respect to the emotion analysis. Likewise a word that is recognized by the ASR unit, but was not seen throughout linguistic analysis can be replaced by a (web-based) description. This may help recover the missing word if the description does not inherit too many novel OOVs. Further we aim at usage of the recognized word chain for acoustic modeling with respect to emotion recognition: emotion models trained text-independent of the spoken content can then be replaced by fitting models. Finally, we plan to use confidence scores to improve the results.

8. ACKNOWLEDGEMENTS

This work originated in a co-operation between several sites dealing with classification of emotional user states conveyed via speech. This initiative was taken in the European Network of Excellence HUMAINE under the name CEICES (**C**ombining **E**fforts for **I**mproving automatic **C**lassification of **E**motional user **S**tates).

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.

9. REFERENCES

[1] 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, pp. 240–245.

[2] R. Cowie, E. Douglas-Cowie, B. Apolloni, J. Taylor, A. Romano, and W. Fellenz, "What a neural net needs to know about emotion words," *Journal of Computational Intelligence and Applications*, pp. 109–114, 1999.

[3] D. Litman and K. Forbes, "Recognizing emotions from student speech in tutoring dialogues," in *Proc. ASRU*, Virgin Island, 2003, pp. 25–30.

[4] X. Zhe and A.C. Boucouvalas, "Text-to-emotion engine for real time internet communication," in *Proc. the Int. Symposium on Communication Systems, Networks, and DSPs*, Staffordshire University, 2002, pp. 164–168.

[5] B. Goertzel, K. Silverman, C. Hartley, S. Bugaj, and M. Ross, "The baby webmind project," in *Proc. The Annual Conf. of The Society for the Study of Artificial Intelligence and the Simulation of Behaviour (AISB)*, 2000.

[6] H. Liu, H. Liebermann, and T. Selker, "A model of textual affect sensing using real-world knowledge," in *Proc. 7th International Conference on Intelligent User Interfaces (IUI 2003)*, 2003, pp. 125–132.

[7] T. S. Polzin and A. Waibel, "Emotion-sensitive human-computer interfaces," in *Proc. ISCA Workshop on Speech and Emotion 2000*, 2000, pp. 201–206.

[8] B. Schuller, R. Müller, M. Lang, and G. Rigoll, "Speaker Independent Emotion Recognition by Early Fusion of Acoustic and Linguistic Features within Ensembles," in *Proc. 9th Eurospeech - Interspeech 2005*, Lisbon, 2005, pp. 805–809.

[9] D. Seppi, M. Gerosa, B. Schuller, A. Batliner, and S. Steidl, "Detecting Problems in Spoken Child-Computer Interaction," in *Proc. 1st ISCA Workshop on Child, Computer and Interaction*, Chania, Crete, Greece, 2008.

[10] C. M. Lee, S. S. Narayanan, and R. Pieraccini, "Combining acoustic and language information for emotion recognition," in *Proc. ICSLP*, 2002, pp. 873–376.

[11] J. Ang, R. Dhillon, E. Shriberg, and A. Stolcke, "Prosody-based automatic detection of annoyance and frustration in human-computer dialog," in *Proc. ICSLP*, 2002, pp. 2037–2040.

[12] H. Lodhi, J. Shawe-Taylor, N. Cristianini, and C. Watkins, "Text classification using string kernels," *Journal of Machine Learning Research*, pp. 419–444, 2002.

[13] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Does Affect Affect Automatic Recognition of Children's Speech?," in *Proc. 1st ISCA Workshop on Child, Computer and Interaction*, Chania, Crete, Greece, 2008.

[14] D. Seppi, A. Batliner, B. Schuller, S. Steidl, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, N. Amir, and V. Aharonson, "Patterns, Prototypes, Performance: Classifying Emotional User States," in *Proceedings 9th INTERSPEECH 2008*, Brisbane, Australia, 2008.

[15] B. Schuller, D. Seppi, A. Batliner, A. Meier, and S. Steidl, "Towards more Reality in the Recognition of Emotional Speech," in *Proc. of ICASSP 2007*, Honolulu, 2007, pp. 941–944.

[16] B. Schuller, A. Batliner, D. Seppi, S. Steidl, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, N. Amir, L. Kessous, and V. Aharonson, "The Relevance of Feature Type for the Automatic Classification of Emotional User States: Low Level Descriptors and Functionals," in *Proc. Interspeech*, Antwerp, 2007, pp. 2253–2256.