Combining speech recognition and acoustic word emotion models for robust text-independent emotion recognition

Björn Schuller, Bogdan Vlasenko, Dejan Arsic, Gerhard Rigoll, Andreas Wendemuth

Angaben zur Veröffentlichung / Publication details:

Schuller, Björn, Bogdan Vlasenko, Dejan Arsic, Gerhard Rigoll, and Andreas Wendemuth. 2008. "Combining speech recognition and acoustic word emotion models for robust text-independent emotion recognition." In *2008 IEEE International Conference on Multimedia and Expo, 23 June - 26 April 2008, Hannover, Germany*, edited by Jörn Ostermann, Touradj Ebrahimi, and Oscar Au, 1333–36. Piscataway, NJ: IEEE. https://doi.org/10.1109/icme.2008.4607689.

COMBINING SPEECH RECOGNITION AND ACOUSTIC WORD EMOTION MODELS FOR ROBUST TEXT-INDEPENDENT EMOTION RECOGNITION

Björn Schuller[1], Bogdan Vlasenko[2], Dejan Arsic[1], Gerhard Rigoll[1], Andreas Wendemuth[2]

[1]Institute for Human-Machine Communication, Technische Universität München, Germany
[2]Cognitive Systems, IESK, Otto-von-Guericke University, Magdeburg, Germany
[1]*schuller@tum.de*, [2]*bogdan.vlasenko@ovgu.de*

ABSTRACT

Recognition of emotion in speech usually uses acoustic models that ignore the spoken content. Likewise one general model per emotion is trained independent of the phonetic structure. Given sufficient data, this approach seemingly works well enough. Yet, this paper tries to answer the question whether acoustic emotion recognition strongly depends on phonetic content, and if models tailored for the spoken unit can lead to higher accuracies. We therefore investigate phoneme-, and word-models by use of a large prosodic, spectral, and voice quality feature space and Support Vector Machines (SVM). Experiments also take the necessity of ASR into account to select appropriate unit-models. Test-runs on the well-known EMO-DB database facing speaker-independence demonstrate superiority of word emotion models over today's common general models provided sufficient occurrences in the training corpus.

Index Terms — Emotion Recognition, Affective Speech, Acoustic Modeling, Word Models

1. INTRODUCTION

Practically every approach to the recognition of emotion in speech ignores the spoken content when it comes to acoustic modeling [1,2,3,4,5]. A general model is trained for each emotion, and applied on test-utterances. While this is a common practice, it seems surprising how well this works, especially considering that many features highly depend on phonetic structure, such as spectral and cepstral features which have become very popular recently [1]. This derives from the high reduction of information: e.g. rather than using the original time-series, higher order statistics, such as means, extremes, deviations, etc., are used. This is also manifested by works that demonstrated lower performance of dynamic modeling, e.g. by HMM, of low-level-descriptors [3]. Apparently, in current approaches phonetic content is over-modeled leading to low generalization capability. In this work we applied phoneme and word models of emotions for phoneme, word and sentence level of analysis.

Yet, the question is if spoken content variance influences emotion recognition performance negatively, and if models trained specifically on the unit at hand, can help. We aim at shedding light on this question by training phoneme-, and word emotion models for the recognition of emotion in the following. Unit-specific models demand knowledge of the phonetic content, opposing "blind" sub-turn entities, as introduced in [2,4,5,6,7]. Likewise, recognition of the spoken content becomes a necessity, in order to choose the correct model each time. Facing real world cases [8], we do not report on transcribed content, as e.g. in [1], but do incorporate an HMM-based state-of-the-art approach to ASR. We report results considering superiority of specific models over general models, and combine emotion and speech recognition in a real system.

2. DATABASE

To demonstrate the effectiveness of unit-specific models, we decided for the popular studio recorded Berlin Emotional Speech Database (EMO-DB) [9], which covers the 'big six' emotion set (MPEG-4) besides boredom instead of surprise, and added neutrality. This database contains acted samples. However, to our best knowledge this is the only public emotional speech database that provides accurate manual syllable boundaries and transcription for model training. Further, the spoken content is pre-defined, thus providing a high number of repeated words in diverse emotions allowing for training of word emotion models. In a later application with spontaneous data, larger corpora may fulfill the same requirement. Finally, these results allow for comparison with the results presented in [7] with respect to sub-turn units.

10 (5f) professional actors speak 10 German emotionally undefined sentences. 494 phrases are marked as min. 60% natural and min. 80% assignable by 20 subjects. 84.3% accuracy is reported for a human perception test. For a comparison with spontaneous data refer to [8].

3. TOKEN PASSING SCHEME

To apply grammatical constraints within the token passing scheme [12], these grammar rules are compiled into a set of linked syntax networks of the form illustrated by Fig(1).
The nodes of each syntax network are of three types: links, terminals and non-terminals. Link nodes are used for storing tokens and are the points where recognition decisions are

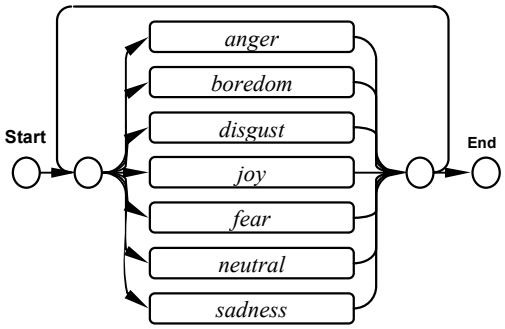

Figure 1. *One-pass Viterbi beam search with token passing.*

recorded. Terminal nodes correspond with emotion acoustic models and non-terminal nodes refer to separate sub-syntax networks representing the right-hand side of the corresponding grammar rule. In our case we did not use non-terminal nodes.

The three types of nodes are combined in such a way that every arc connects a terminal to a link node, or vice versa. Each syntax network has exactly one entry, one exit, and zero or more internal link nodes. Every terminal node has exactly one arc leading into it, whereas each link node may have any number. Link nodes can thus be viewed as filters, which remove all but the best (i.e. lowest cost) tokens passing through them. The main idea is that tokens propagate through the networks. When a token node enters a terminal node, it is transferred to the entry node of the corresponding emotional state model.

4. PHONEME EMOTION MODELS

As a starting point for our experiments we choose phonemes, as these should provide the most flexible basis for unit-specific models: if emotion recognition is feasible on phoneme basis, these units could be most easily re-used for any further content, and high numbers of training instances could be obtained.

We use a simple conceptual model of dynamic emotional state recognition on phoneme level analysis: the full list of 41 phonemes as transcribed for EMO-DB [9] is modeled for each of the 7 emotions contained, independently. As a result 7 x 41 = 287 phoneme emotion models are trained.

Two concepts are tested: first, a time-synchronous one-pass Viterbi-beam search strategy and token passing [12] with direct context free grammar (CFG) are used for decoding. Second, a bi-gram language model (LM) is applied for emotional continuous speech modeling. To apply CFG constraints we used the token passing scheme described in sect. 3.

In the case of emotion recognition via words, as in more recent works [1,8], emotional state models included lists of all emotional phonetic transcriptions for all words contained in the evaluation corpus, that is EMO-DB. We also consider the case of emotion recognition via sentences, which is more commonly used in earlier works [2,3,4,5,7].

For the CFG evaluation no prior LM information is used, as opposed to the second conception, where bi-gram LM are employed for emotion recognition via words. Speech input is processed using a 25 msec Hamming window, with a frame rate of 5 msec. As in typical ASR tasks, we use a 39 dimensional feature vector per frame consisting of 12 MFCC and log frame energy plus speed and acceleration regression coefficients. Cepstral Mean Subtraction and variance normalization are applied to better deal with channel characteristics.

Emotional phonemes are modeled by training three emitting state HMM models in speaker-independent (SI) manner using Baum-Welch re-estimation and up to 32 mixtures, with the named beam search and token passing and direct CFG or bi-gram LM of continuous emotional speech [10].

Test-runs on EMO-DB for phoneme models are carried out in Leave-One-Speaker-Out (LOSO) manner to address speaker independence, as required by most applications. 99.8% sentence speech recognition accuracy can be reported for SI acoustic models. This high accuracy justified the automatic LM application. Yet, speech recognition accuracy on a frame level for phonemes is considerably lower at 79.6%.

In contrast to ASR, partly wrong phoneme models are selected for emotion recognition, and errors of emotional sentence recognition are highly correlated to erroneous recognition of the emotional state of the sentence.

Level	Context	Acc. [%]
sentence	-	66.2
word	bi-gram LM	51.0
word	CFG	32.1
phoneme	bi-gram LM	38.8

Table 1. *Accuracies of emotion recognition on sentence-, word-, and phoneme-level applying phoneme emotion models, dynamic features, HMM, LOSO, on database EMO-DB.*

In tab. 1 results are shown for emotion recognition on a sentence-, word-, and phoneme level in diverse constellations. Overall, the sentence level protrudes, as many errors can be "repaired" on phoneme level. CFG for word level analysis shows many insertions, hence low accuracy. Bi-gram LM can balance the insertions by LM factor, hence higher accuracy. This is also the reason why phoneme-level accuracy is only reported with bi-gram LM: CFG leads here to too high insertion rates.

5. WORD EMOTION MODELS

The larger unit of modeling in the following – words – allows for us to shift to the usual acoustic emotion modeling by large static feature vectors. In order to represent a typical state-of-the-art emotion recognition engine, we use a set of 1,406 acoustic features basing on 37 Low-Level-Descriptors (LLD) as seen in table 2 and their first order delta coefficients [8]. These 37x2 LLDs are next smoothed by low-pass filtering with an Simple Moving Average filter.

In contrast to the formerly introduced dynamic modeling, such systems derive statistics per speaker turn by a projection of each uni-variate time series, respectively LLD, X onto a scalar feature x independent of the length of the turn. This is realized by use of a functional F, as depicted:

$$F : X \rightarrow x \in \mathbf{R}^1 \qquad (1)$$

19 functionals are applied to each contour on the word-level covering extremes, ranges, positions, first four moments and quartiles as also shown in table 2. Note that three functionals are related to position, known as duration in traditional phonetic terminology, as their physical unit is msec.

Low-Level-Descriptors (2x37)	**Functionals (19)**
(Delta) Pitch (Delta) Jitter (Delta) Energy (Delta) Envelope (Delta) Formant 1-5 Amplitude (Delta) Formant 1-5 Bandwidth (Delta) Formant 1-5 Frequency (Delta) MFCC Coefficient 1-16 (Delta) HNR (Delta) Shimmer	Mean, Centroid,Std. Dev. Skewness, Kurtosis Zero-Crossing-Rate Quartile 1,2,3 (Q1,2,3) Q1 – Min, Q2 - Q1, Q3 - Q2,Max - Q3 Max., Min. Value, Rel. Pos.Max., Min. Range Pos. 95% Roll-Off-Point

Table 2. *Overview of Low-Level-Descriptors and functionals for syllable- and word-level analysis.*

For classification we use Support Vector Machines (SVM) with linear Kernel and 1-vs.-1 multi-class discrimination [11]. One could consider the use of 1-state HMM here as well. Yet, SVM have proven the preferred choice in many works to best model static acoustic feature vectors [1,7,8].

The shift to static feature space modeling forces us to use two stage processing in the following, as opposed to the formerly described phoneme emotion models: words have to be recognized by an ASR unit, first.

Next, the corresponding word emotion models have to be selected for emotion recognition. This may lead to a downgrade, if word insertions, deletions or substituions occur, provided the spoken content does influence emotion recognition. Therefore we test emotion recognition in matched word condition (picking only the correct word model) and in mismatched conditions (using all incorrect word models), in contrast to a general model trained on all words. Note that for mismatched condition one vs. one training and testing of each word vs. each other is necessary.

A total of 73 different words are found in EMO-DB. Out of these we select only such that have a minimum frequency of occurrence of 3 within each emotion (likewise having 50 plus instances per word) comprising a total of 41 words with roughly 200 instances per word. 85.0% accuracy is obtained training SI word-models for ASR in a first step in LOSO manner with variable state-number and a maximum frequency of 9 per model. Only 3 mixtures are optimal due to sparse data.

As described, we employ static acoustic features and SVM classification for word emotion models after selection of according words by ASR. Table 3 visualizes the results obtained by two groups of frequency of occurrence in the corpus:

Group 1 (G 1) are high occurrence words that are "worth it" – that is their word emotion model outperforms a general model. These words (10 in total resembling a fourth of the vocabulary) are "*abgeben (give away), am (on), auf (on top of), besucht (visits), gehen (walk), ich (I), sein (to be), sich (oneself), sie (her), sieben (seven)*".

Acc. [%]	**G 1**	**G 2**	**All**
word emotion model			
Matched	**57.2**	46.9	48.9
Mismatched	36.7	37.6	37.4
(training size factor) general emotion model			
1	44.1	42.8	43.1
5	49.7	48.9	49.1
10	46.8	52.9	51.7
all (~100)	**50.7**	56.7	55.5

Table 3. *Accuracies at word-level for word emotion models in matched and mismatched condition compared to general models at diverse relative sizes of training corpora. Static features, SVM, LOSO, on database EMO-DB. Investigated are "worth-it" words (G 1) and "non-worth-it" candidates (G 2), as well as all (All) terms.*

In contrast, group 2 (G 2) is "not worth it" due to low frequency of occurrence in the corpus. Likewise emotion models for these words cannot be trained sufficiently.

Additionally, results for all words are shown (All). Again, we use LOSO evaluation. Note that from now on only rates equivalent to CFG are reported, as no combined decoding is used due to the two-stage processing. In the following, we report results on word level only, which represents the actual variations of emotion within a sentence.

First, matched vs. mismatched conditions are analyzed: spoken content clearly does influence accuracy throughout word-model comparison in any case, as can be seen. In fact, the length of words and phonetic distance are the main influence factors.

As mis-selection of word emotion models would apparently significantly downgrade performance, we next address the question how a general model trained on any word in the corpus – the common state-of-the-art – would perform. We set this in relation to the amount of training data available for each word specific emotion model, by indicating the relative training size factor by random down-sampling preserving class-balance.

As can be seen, a minimum of 5 times as many data is needed in average to outperform word emotion models of the low occurrence frequency group 2. Further, even using the whole corpus, general models could not outperform matched word specific emotion models of the "worth it" high occurrence frequency group 1. This shows usefulness of selection these words.

6. DISCUSSION

We try to compare emotion recognition on the phoneme and word level. As shown in sect. 3 and 4, and in accordance with earlier results [7], larger units seem to be beneficial for emotion recognition. However, the introduced unit-specific emotion models clearly outperformed common general models provided enough training material per unit. Sadly however, this is more likely given smaller units.

In a fully automatic system that recognizes units prior to their selection, only word-models seem to be reasonable. These should be limited to words having sufficient occurrences in the training material, as e.g. function words. As we considered a closed vocabulary setting, reliable word-spotting will be needed to handle out-of-vocabulary (OOV) occurrences with respect to the word emotion model inventory.

Further, acoustic confidences should be integrated to select general models in case of low confidence, as wrong models tend to downgrade performance. Still, acoustic confusions can typically be expected to show similar phonetic content, which resulted in a comparably small downgrade, herein.

As a general finding it can be stated that word emotion models seem to be worth the extra effort of a previous ASR stage and training of word emotion models.

Future works will deal with use of word emotion models on larger and spontaneous corpora, such as AIBO [8], OOV-emotion model handling, and optimization by integration of acoustic confidences in the selection process.

7. ACKNOWLEDGMENTS

The work has been conducted in the framework of the NIMITEK project (Sachsen-Anhalt) FKZ XN3621H/1005. This project is associated and supported by the Magdeburg Center for Behavioral Brain Sciences. Bogdan Vlasenko acknowledges support by a graduate grant of the Federal State of Sachsen-Anhalt.

8. REFERENCES

[1] Batliner, A.; Steidl, S.; Schuller, B.; Seppi, D.; Laskowski, K.; Vogt, T.; Devillers, L.; Vidrascu, L.; Amir, N.; Kessous, L.; Aharonson, V.: "Combining Efforts for Improving Automatic Classification of Emotional User States," Proc. IS-LTC 2006, Ljubljana, Slovenia, 2006.

[2] Polzin, T.S.; Waibel, A.: "Detecting emotions in speech", Proc. Cooperative Multimodal Communication, 2nd Int. Conf. 98, 1998.

[3] Schuller, B.; Rigoll, G.; Lang, M.: "Hidden Markov Model-Based Speech Emotion Recognition," Proc. ICASSP 2003, IEEE, Vol. II, pp. 1-4, Hong Kong, China, 2003.

[4] Lee Z; Zhao Y.: "Recognizing emotions in speech using short-term and long-term features", Proc. ICSLP, pp. 2255-2558, 1998.

[5] Jiang, D. N.; Cai, L.-H.: "Speech emotion classification with the combination of statistic features and temporal features," Proc. ICME 2004, IEEE, Taipei, Taiwan, pp. 1967-1971, 2004.

[6] Murray L.R.; Arnot, I.L.: "Toward the simulation of emotion in synthetic speech: A review of the literature of humans vocal emotion," JASA, Vol. 93, issue 2, pp.1097-1108, 1993.

[7] Schuller, B.; Rigoll, G.: "Timing Levels in Segment-Based Speech Emotion Recognition," Proc. INTERSPEECH 2006, ICSLP, ISCA, pp. 1818-1821, Pittsburgh, PA, 2006.

[8] Schuller, B.; Seppi, D.; Batliner, A.; Maier, A.; Steidl, S.: "Towards More Reality in the Recognition of Emotional Speech," Proc. ICASSP, Vol. IV, pp. 941-944, Honolulu, Hawaii, 2007.

[9] Burkhardt, F.; Paeschke, A.; Rolfes, M.; Sendlmeier, W.; Weiss, B.: "A Database of German Emotional Speech," Proc. INTERSPEECH, ISCA, Lisbon, Portugal, pp.1517-1520, 2005.

[10] Young, S.; Evermann, G.; Kershaw, D.; Moore, G.; Odell, J.; Ollason, D.; Povey, D.; Valtchev, V.; Woodland, P.: The HTK-Book 3.2, Cambridge University, Cambridge, England, 2002.

[11] Witten, I.H.; Frank, E.: Data Mining: Practical machine learning tools with Java implementations, Morgan Kaufmann, San Francisco, pp. 133, 2000.

[12] Young, S.; Russel, N.; Thornton, J.; "Token Passing: a Simple Conceptual Model for Connected Speech Recognition Systems", Cambridge University Engineering Department. July 1989.