Balancing Spoken Content Adaptation and Unit Length in the Recognition of Emotion and Interest

Bogdan Vlasenko[1]*, Björn Schuller*[2]*, Kinfe Tadesse Mengistu*[1]*, Gerhard Rigoll*[2]*, Andreas Wendemuth*[1]

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

Abstract

Recognition and detection of non-lexical or paralinguistic cues from speech usually uses one general model per event (emotional state, level of interest). Commonly this model is trained independent of the phonetic structure. Given sufficient data, this approach seemingly works well enough. Yet, this paper addresses the question on which phonetic level there is the onset of emotions and level of interest. We therefore compare phoneme-, word- and sentence-level analysis for emotional sentence classification by use of a large prosodic, spectral, and voice quality feature space for SVM and MFCC for HMM/GMM. Experiments also take the necessity of ASR into account to select appropriate unit-models. In experiments on the well-known public EMO-DB database, and the SUSAS and AVIC spontaneous interest corpora, we found that the emotion recognition by sentence level analysis shows the best results. We discuss the implications of these types of analysis on the design of robust emotion and interest recognition of usable human-machine interfaces (HMI).

Index Terms: emotion and interest recognition, affective speech, phoneme and word models

1. Introduction

Detecting non-lexical or paralinguistic cues from speech is one of the major challenges in the development of usable human-machine interfaces (HMI). Notable among these cues are the universal categorical emotional states (e.g. anger, boredom, disgust, fear, joy, neutral, sadness, etc.) and/or level of interest (neutrality, interest, curiosity), prevalent in day-to-day scenarios. Knowing such emotional states and/or levels of interest can help adjust system responses so that the user of such a system can be more engaged and have a more effective interaction with the system.

Practically every approach to the recognition of emotion in speech ignores the spoken content when it comes to acoustic emotion modeling [1, 3, 4, 6, 7]. A general model is trained for each emotion, and applied on test-utterances. As the standard unit for recognition of emotion within speech, a whole turn is commonly used. From an application point of view, this seems appropriate in most cases: a change of emotion during a phrase seems to occur seldom enough for many applications. However, from a recognition point of view, it has often been reported that sub-timing levels seem to be advantageous [3, 5]. Promising results are reached by broad phonetic category analysis of affective speech [12], albeit in speaker dependent and context dependent evaluation. Apart from a few attempts to classify emotions within speech dynamically, current approaches usually employ static feature vectors derived on a sentence, word, or chunk level [9]. In [10] a combination of static and dynamic modeling has shown good results. This derives mostly from the fact, that by (usually statistical) functional application to the Low-Level- Descriptors (LLD) as e.g. pitch, energy, or spectral coefficients an important information reduction takes place, which avoids phonetic (respectively spoken-content) over-modeling. While this is 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 [7]. Apparently, in current approaches phonetic content is over-modeled leading to low generalization capability.

Yet, the question is on which phonetic level there is the onset of emotions and level of interest. What is the optimal phonetic unit of analysis for robust non-lexical or paralinguistic events detection? We aim at shedding light on this question by training phoneme, word, sentence level models for the recognition of emotions and level of interest within speech. Unit-specific models demand knowledge of the phonetic content, opposing “blind” sub-turn entities, as introduced in [5, 6]. Likewise, recognition of the spoken content becomes a necessity, in order to choose the correct model each time. Facing real world cases [9], 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 compare results of different level of analysis for emotion and level of interest recognition tasks.

The paper is structured as follows: in sect. 2 we introduce the databases, in sect. 3, 4, 5 we discuss the diverse models and present results. Sect. 6 discusses findings, and summarizes this paper.

2. Acted and Spontaneous Emotions and Level of Interest Data

To compare the effectiveness of unit-specific models, we decided for the popular studio recorded Berlin Emotional Speech Database (EMO-DB)[2], which covers *anger, boredom, disgust, fear, joy, neutral, sadness* speaker emotions. The spoken content is pre-defined, thus providing a high number of repeated words in diverse emotions allowing for training of word emotion models.

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.

Second, we selected the Speech Under Simulated and Actual Stress (SUSAS) database [14] as a reference for spontaneous recordings. As additional challenge speech is partly masked by field noise. It consists of five domains, encompassing a wide variety of stresses and emotions. We decided for the 3,663 actual stress speech samples recorded in subject motion fear and stress tasks, as acted samples are already covered by EMO-DB in this work. 7 speakers, 3 of them female, in roller coaster and free fall actual stress situations are contained in this set. Two different stress conditions have been collected: *medium stress, and high stress*. Within the further samples also *neutral samples, fear during freefall* and *screaming* are contained as classes. Likewise a total of five emotions, respectively speaking styles, are covered. SUSAS samples are constrained to a 35 words vocabulary of short aircraft communication commands. All files are sampled in 8 kHz, 16 bit. The recordings are partly overlaid with heavy noise and background over-talk. However, this resembles realistic acoustic recording conditions, as also given in many related scenarios of interest such as automotive speech interfaces or public transport surveillance.

To find an optimal phonetic level of analysis for different level of interest among sentences, we decided for the AVIC (Audiovisual Interest Corpus)[15]. In the scenario setup, an experimenter and a subject are sitting on both sides of a desk. The experimenter plays the role of a product presenter and leads the subject through a commercial presentation. The subjects role is to listen to explanations and topic presentations of the experimenter, ask several questions of her/his interest, and actively interact with the experimenter considering his/her interest to the addressed topics without respect to politeness.

The level of interest (LOI) is annotated for every subspeaker turn. 5 LOI were distinguished in the first place: 1 - *Disinterest* (subject is bored with listening and talking about the topic, very passive, does not follow the discourse), 2 - *Indifference* (subject is passive, does not give much feedback to the experimenters explanations, unmotivated questions if any), 3 - *Neutrality* (subject follows and participates in the discourse, it can not be recognized, if she/he is interested or indifferent in the topic), 4 - *Interest* (subject wants to discuss the topic, closely follows the explanations, asks some questions), 5 - *Curiosity* (strong wish of the subject to talk and learn more about the topic). For automatic processing a fusion of these LOIs to a Master LOI was automatically fulfilled as described in [15]. Additionally, the spoken content and nonverbal interjections have been labeled. As too few items for LOI 1 and 2 have been seen, these were clustered together with LOI 3, and the LOI scale was shifted to LOI 0-2. For our evaluation we use 996 phrases.

3. Phoneme-Level Analysis

As a starting point for our experiments we choose phonemes, as these should provide the most flexible basis for unit-specific models: if emotion and level of interest 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 is modeled for each of the 7 emotions contained, independently. As a result 7 x 41 = 287 phoneme emotion (PE - speaker's emotional state dependent phoneme) models are trained. For SUSAS the full list of 35 phonemes is modeled for each of 5 emotions contained, independently. As a result 5 x 35 = 165 PE models are trained. For level of interest recognition on phoneme level analysis: the full list of 39 phonemes as transcribed for AVIC is modeled for each of the 3 levels of interest contained, independently. As result 3 x 39 = 117 phoneme level of interest (PLOI - speaker's level of interest dependent phoneme) models are trained.

An HMM of three emitting states and 16 mixtures of Gaussians was built for each PE and PLOI models. The HTK toolkit was used to build these models, using standard techniques such as forward-backward and Baum-Welch re-estimation algorithms [13]. After a high-frequency pre-emphasis of the speech signal, MFCC feature vectors were estimated. Speech input is processed using a 25 msec Hamming window, with a frame rate of 100 fps. 13 coefficients were estimated with cepstral mean normalization (CMN). The velocity and acceleration of these coefficients were included forming the "classical" 39 dimensional feature vector.

For a start we are using an Automatic Speech Recognition (ASR) engine adapted for affective speech to recognize a unit (sentence, word). After this we are generating possible emotional or level of interest phonetic transcriptions for the recognized sentence or words by using the corresponding phoneme set (PE or PLOI). In case of EMO-DB we are considering 7 PE transcriptions, 5 PE transcription for SUSAS and 3 PLOI transcriptions for AVIC. Then we employ the Viterbi algorithm [13] to choose the most appropriate PE or PLOI transcription for each recognized sentence. As output the most appropriate emotional state or level of interest are chosen.

Test-runs on EMO-DB, SUSAS and AVIC for phoneme level models are carried out in a Leave-One-Speaker-Out (LOSO) manner to address speaker independence (SI), as required by most applications.

Table 1: *Accuracies of emotion and level of interest recognition on sentence-, and word-level applying phoneme-level analysis, MFCC, HMM/GMM, LOSO.*

Classification unit [%]	EMO-DB	SUSAS	AVIC
word	51.0	49.5	45.8
sentence	66.2	49.5	54.1

Note that in case of SUSAS only one word is contained per sentence. All databases are annotated on sentence level only. Detailed results of EMO-DB and AVIC evaluations show that some words within a sentence are classified erroneously when the whole sentence is classified correctly. This means that emotional and level of interest trace is distributed irregularly among words inside a sentence. As a result phonemes which belong to the different words within a sentence have diverse emotion and level of interest saturation. Consequently, we are not able to train reliable PE and PLOI models.

4. Word-Level Analysis

The next level of analysis (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 based on 37 Low-Level-Descriptors (LLD) as seen in Table 2 and their first order delta coefficients [9]. These 37x2 LLDs are next smoothed by Low-pass filtering with a SMA-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:

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

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.

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

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

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 vector classification [1, 6, 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 substitutions 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 those 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. For EMO-DB these words (10 out of 41) 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)”*. For AVIC these words (7 out of 50) are *”ah, but, is, it, mh, not, you”*. For SUSAS this word (1 out of 11) is *”fifty”*. In contrast, group 2 (G 2) is “not worth it” due to low frequency of word 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. No combined decoding is used due to the two-stage processing. In the following, we stick to words as unit of analysis, which allow for incremental emotion recognition.

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 Table 3. In fact, detailed analysis shows that the length of words and phonetic distance are the main influence factors.

Table 3: *Accuracies at word-level for word emotion models in matched and mismatched condition. Static features, SVM, LOSO. Investigated are “worth-it” words (G 1) and “non-worth-it” candidates (G 2), as well as all (All) terms.*

Model description	Acc. [%]	G 1	G 2	All
EMO-DB	matched	57.2	46.9	48.9
	mismatched	36.6	37.7	37.4
SUSAS	matched	64.6	60.3	60.7
	mismatched	52.4	54.4	55.2
AVIC	matched	79.7	57.8	60.9
	mismatched	49.2	51.3	50.1

Analyzing results of word level analysis for acted and spontaneous emotion and spontaneous level of interest, we found notable differences between matched and mismatched condition for words from group G1 and G2. As can be seen from Table 3, in matched cases word dependent models for words from the group G1 provide better performance then general emotion models. This confirms our assumption about irregular trace of emotion and level of interest within words in sentence.

Table 4: *Accuracies at word-level for word emotion models for general models at diverse relative sizes of training corpora. Static features, SVM, LOSO.*

Training size factor	1%	2%	5%	10%	100%
EMO-DB	43.1	44.7	49.1	51.7	55.5
SUSAS	50.6	56.1	60.7	61.5	64.7
AVIC	58.0	62.6	65.2	68.6	68.6

As mis-selection of word emotion models would apparently significantly downgrade performance, we next address the question how a general model trained on the whole corpus (the common state-of-art) would perform.

We set this in relation to the amount of training data available for each word specific emotion model by the relative training size factor by random down-sampling preserving class-balance, see Table 4

Every word will occur with an average frequency reaching from 1.0% to 2.0% for EMO-DB, SUSAS, AVIC. It can be seen that for all databases a general model with that training size factor will perform between matched and mismatched models for all words. With more training material available, the general model outperforms the matched case picking "All", and approaches the "G1" matched case. Without "G1" selection it seems preferable to decide for the general emotion model, simply as more data is available. With "G1" matched cases accuracy of emotion recognition with word level models outperform the general model with 100% training size factor. This shows the usefulness of selection of "worth-it" words with high frequency of occurrence.

5. Sentence-Level Analysis

For the sentence as level of analysis we trained emotion and level of interest models on whole sentences. We used two different emotion classification engines as described: firstly based on GMM analysis of MFFC, secondly 1.4k large-feature-space SVM.

Table 5: *Accuracies of emotion recognition on sentence-level applying sentence-level analysis, LOSO, on databases EMO-DB and AVIC. SN and FS represent speaker normalization and feature selection.*

Model description [%]	EMO-DB	SUSAS	AVIC
SVM + LLD	74.9	62.0	69.4
SVM + LLD + SN	79.6	63.3	70.9
SVM + LLD + SN + FS	83.2	64.7	72.3
GMM + MFCC	77.1	47.2	69.9
GMM + MFCC + VTLN	82.9	51.2	72.1

Vocal Tract Length Normalization (VTLN) thereby improves emotion recognition accuracy for GMM based analysis. Over all, the sentence level analysis shows the best accuracy of emotion and level of interest recognition. However, clearly this level of analysis is not able to detect emotional state changes within a sentence. On a sample of SUSAS evaluations one can see that LLD with SVM provides higher accuracy for data recorded in heavily noisy environment.

6. Discussion

This work compared emotion recognition on the sentence level by phoneme-, word-, sentence-level analysis. As shown in sect. 3, 4 and 5, and in accordance with earlier results [10], 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. With more training material available, a general model outperforms the matched case picking all words and approaches the "worth it" words matched case. Appearance of word level labeled corpora can improve current performance of phoneme and word level emotion and level of interest models. We found that emotional and level of interest saturation is distributed irregularly among words inside a sentence. For example in AVIC, accuracy of level of interest recognition for the words *"ah, but, is, it, mh, not, you"* by word dependent models exceeds accuracy of level of interest detection by general models. This is not the case for other databases. Speaker normalization, Vocal Tract Length Normalization and feature space optimization clearly help to improve overall results.

In future work we plan to investigate cross-corpora and cross language non-lexical and paralinguistic events detection. We thereby also aim at investigation of robust fusion of static LLD with dynamic MFCC analysis combined with an ASR engine. Furthermore, analysis of non-lexical and paralinguistic cues can help to find an optimal adaptation method for robust affective speech recognition.

7. Acknowledgements

The work has been conducted in the framework of the NIMITEK project (Sachsen-Anhalt Federal State funding) FKZ XN3621H/1005. This project is associated and supported by the Magdeburg Center for Behavioral Brain Sciences (Neuroscience Excellence Cluster). 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., and Aharonson, V., "Combining Efforts for Improving Automatic Classification of Emotional User States", Proc. 1st Int. Language Technologies Conference IS-LTC, Ljubljana, Slovenia, 240-245, 2006.

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

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

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

[5] Murray, L.R. and 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, 1097-1108, 1993.

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

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

[8] Schuller, B. and Rigoll, G., "Timing Levels in Segment-Based Speech Emotion Recognition, Proc. INTERSPEECH 2006, ICSLP, ISCA, 1818-1821, 2006.

[9] Schuller, B., Seppi, D., Batliner, A., Maier, A. and Steidl, S., "Towards More Reality in the Recognition of Emotional Speech", Proc. ICASSP, Vol. IV, 941-944, 2007.

[10] Schuller, B., Vlasenko, B., Minguez, R., Rigoll1, G. and Wendemuth, A., "Comparing One and Two-Stage Acoustic Modeling in the Recognition of Emotion in Speech" Proc. ASRU 2007, 596-600, 2007.

[11] Witten, I.H. and Frank, E., "Data Mining: Practical machine learning tools with Java implementations", Morgan Kaufmann, 133-145, 2000.

[12] Busso, C., Lee, S., Narayanan S. S., "Using Neutral Speech Models for Emotional Speech Analysis" Proc. INTERSPEECH, ISCA, Antwerp, Belgium, 2225-2228, 2007.

[13] Young, S., Evermann, G., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V. and Woodland P., The HTK-Book 3.4, Cambridge University, Cambridge, England, 2006.

[14] Hansen, J.H.L., Bou-Ghazale, S., Getting Started with SUSAS: A Speech Under Simulated and Actual Stress Database, Proc. EUROSPEECH-97, Rhodes, Greece, Vol. 4, 1743-1746, 1997.

[15] Schuller, B., Müller, R., Hörnler, B., Höthker, A., Konosu, H. and Rigoll., G., "Audiovisual recognition of spontaneous interest within conversations." In Proc. 9th Int. Conf. on Multimodal Interfaces (ICMI), Special Session on Multimodal Analysis of Human Spontaneous Behaviour, Nagoya, Japan, ACM SIGCHI, 30-37, 2007.