COMPARING ONE AND TWO-STAGE ACOUSTIC MODELING IN THE RECOGNITION OF EMOTION IN SPEECH

Björn Schuller[1], Bogdan Vlasenko[2], Ricardo Minguez[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@IEEE.org

ABSTRACT

In the search for a standard unit for use in recognition of emotion in speech, a whole turn, that is the full section of speech by one person in a conversation, is common. Within applications such turns often seem favorable. Yet, high effectiveness of sub-turn entities is known. In this respect a two-stage approach is investigated to provide higher temporal resolution by chunking of speech-turns according to acoustic properties, and multi-instance learning for turn-mapping after individual chunk analysis. For chunking fast pre-segmentation into emotionally quasi-stationary segments by one-pass Viterbi beam search with token passing basing on MFCC is used. Chunk analysis is realized by brute-force large feature space construction with subsequent subset selection, SVM classification, and speaker normalization. Extensive tests reveal differences compared to one-stage processing. Alternatively, syllables are used for chunking.

Index Terms— Emotion Recognition, Affective Computing, Segmentation

1. INTRODUCTION

As standard unit for recognition of emotion within speech a whole turn can be named [1-4]. 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 [4,5,6]. Still, apart from a few attempts to classify emotions within speech dynamically [1,2], current approaches usually employ static feature vectors derived on a turn, word, or chunk level [8]. In [2] such static modeling has also been shown superior to dynamic modeling. 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. Yet, it is also considered received knowledge that thereby important temporal information is lost due to a high degree of abstraction. This led to first successful attempts to integrate information on diverse time levels [3-6].

In this paper we therefore investigate a two-stage approach to acoustic modeling for the recognition of emotion in speech: a first stage segments speech-turns into chunks which are individually analyzed in a second stage. Subsequently, information from a chunk level is mapped on the turn-level. Herein, this step will be realized by multi-instance learning. Additionally, information from the turn-level can be considered.

Alternatively, we will show results for a syntactically motivated chunking, that has not been considered in recognition of emotion within speech, yet: syllables. Likewise, we provide results on a syllable level, automatically computed chunk level, and the mapping on the turn-level.

The paper is structured as follows: in sect. 2 we introduce the automatic chunking of speech turns into acoustically quasi-stationary units from an emotion point of view. In sect. 3 we introduce the feature set for chunk analysis. Sect. 4 discusses the combined processing within two stages. Subsequently, we introduce optimization strategies in sect. 5. In sect. 6 and sect. 7 we introduce extensive experimental results and discussion.

2. AUTOMATIC CHUNKING

This article describes a simple conceptual model of dynamic emotional state recognition. Time-synchronous one-pass Viterbi-beam search and the token passing algorithm with direct context free grammar are used for decoding [7]. To apply context free grammar as constraints within the token passing scheme, 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 to store tokens and are the points where recognition decisions are recorded. Terminal nodes correspond to emotion acoustic models and non-terminal nodes refer to separate sub-syntax networks representing the RHS of the corresponding grammar rule. In our case we did not use non-terminal nodes.

The three types of node are combined in such a way that every arc connects either a terminal or a non-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 and non-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.

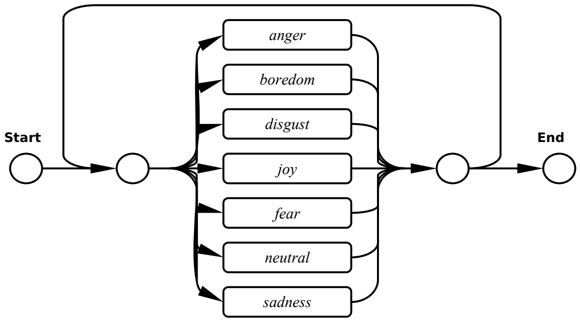

Figure 1. *Stage 1: automatic chunking by acoustic properties and one-pass Viterbi beam search with token passing.*

The main idea is that tokens propagate through the networks just as in the finite state case: when a token node enters a terminal node, it is transferred to the entry node of the corresponding emotional state model.

Speech input is thereby processed using a 25 msec Hamming window, with a frame rate of 10 msec. As in typical speaker or speech recognition tasks we employ a 39 dimensional feature vector per each frame consisting of 12 MFCC and log frame energy plus speed and acceleration coefficients. CMS and variance normalization are applied to better cope with channel characteristics.

Now for the chunking we train the models in a speaker-independent manner using Baum-Welch re-estimation and 50 mixtures. Afterwards each original turn is chunked by application of the beam-search as described. For the latter processing, only the obtained segment boundaries are used from this stage. The motivation behind this processing is to find an acoustically motivated sub-turn splitting.

After chunking - either using the proposed automatic chunking by acoustic properties or annotation based syllable chunking - each chunk is assigned the emotion of the turn it originates from.

3. CHUNK-LEVEL FEATURE EXTRACTION

In order to represent a typical state-of-the-art emotion recognition engine operating on a turn-level, we use a set of 1,406 acoustic features basing on 37 Low-Level-Descriptors (LLD) as seen in table 1 and their first order delta coefficients [8]. These 37x2 LLDs are next smoothed by Low-pass filtering with an SMA-filter.

As opposed to 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 turn-level covering extremes, ranges, positions, first four moments and quartiles as seen in table 1. Note that three functionals are related to position, known as duration in traditional phonetic terminology, as their physical unit is msec.

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

Table 1. *Stage 2: overview of Low-Level-Descriptors and functionals for chunk-level analysis.*

Low-Level-Descriptors (2x37)	Functionals (19)
Pitch	Mean
Energy	Standard Deviation
Envelope	Zero-Crossing-Rate
Formant 1-5 Amplitude	Quartile 1
Formant 1-5 Bandwidth	Quartile 2
Formant 1-5 Frequency	Quartile 3
MFCC Coefficient 1-16	Quartile 1 - Minimum
Harmonics-to-Noise-Ratio	Quartile 2 - Quartile 1
Shimmer	Quartile 3 - Quartile 2
Jitter	Maximum - Quartile 3
Delta Pitch	Centroid
Delta Energy	Skewness
Delta Envelope	Kurtosis
Delta Formant 1-5 Amplitude	Maximum Value
Delta Formant 1-5 Bandwidth	Relative Maximum Position
Delta Formant 1-5 Frequency	Minimum Value
Delta MFCC Coefficient 1-16	Relative Minimum Position
Delta Harmonics-to-Noise-Ratio	Maximum Minimum Range
Delta Shimmer	Position of 95% Roll-Off-Point
Delta Jitter	

4. TWO-STAGE PROCESSING

In order to map the results of each chunk onto the turn-level, we consider three strategies known from multi instance learning for each chunk:

- an un-weighted majority vote (MV),
- a maximum length vote (ML),
- and maximum classifier prediction score multiplied with the length vote (MSL).

Likewise, we compute the majority label of each turn basing on either the syllable or acoustic chunk level. Note that these levels could be combined to consider information from diverse time levels. Also, information from turn level features can easily be added. In the case of weighted vote, the length in frames is used as multiplicative weighting function. In the MSL case we also use the classifier prediction score for each class as additional weight. Note that in case of un-weighted majority vote turns may occur that cannot be uniquely assigned to a class. This happens, if two or more classes, which are the majority classes, have the same number of chunks. This case will be separately denoted in the ongoing. In the case of time-based weighting this case can almost be ignored, as the majority classes – if there are several – will seldom have the exact same number of frames. This is even truer, if length and prediction score are used for weighting (MSL). As a drawback it has to be mentioned that temporal information is thereby lost. Alternatively, the duration of each chunk can be used as weight. Also, the order of chunks is lost. However, we believe that this information can be neglected under the precondition of constant emotion throughout a turn.

Employing majority voting we can consider three cases: turns that are clearly assignable, and such that have two or more emotions assigned due to a draw. In the second case a further distinction can be considered: turns that have the correct emotion among the majority classes, and such that are simply wrongly assigned.

5. OPTIMIZATION

Next, two optimization strategies are considered: First, speaker normalization (SN) by feature normalization with the whole individual speaker context. Second, feature-space optimization by correlation-based exclusion of highly correlated features (FS) is proposed.

We investigate the benefits of speaker normalization, as we intend to analyze emotion independent of the speaker, herein. SN is thereby realized by a normalization of each feature x by its mean and standard deviation for each speaker individually. Thereby the whole speaker context is used. This has to be seen as an upper benchmark for ideal situations, where a speaker could be observed with a variety of emotions. Yet, it is not necessary to know the actual emotional state of observed utterances at this point.

As a high number of features is used throughout static modeling, feature space optimization seems a must in view of performance and real-time-capability. In order to optimize a set of features rather than combining attributes of single high relevance, we use a correlation-based analysis, herein. Thereby features of high class-correlation and low inter-feature correlation are kept [9]. This does not employ the target-classifier in the loop. Likewise it mostly reduces correlation within the feature space rather than evaluation of the benefit of single attributes. Still, this helps to obtain a very compact representation of the feature space, which usually leads to an improvement of accuracy while reducing feature extraction effort at the same time.

6. DATABASE

To demonstrate effectiveness of each chunking and the fusion on the turn-level, we decided for the popular studio recorded Berlin Emotional Speech Database (EMODB) [10], which covers the 'big six' emotion set (MPEG-4) besides boredom instead of surprise, and added neutrality. This database contains acted samples of an emotionally undefined predefined spoken content. However, to our best knowledge this is the only public emotional speech database that provides accurate syllable boundaries. Also, these results allow for comparison with the results presented in [6]. 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.

7. EXPERIMENTAL RESULTS

Within this section we present a number of results carried out on EMO-DB. Test-runs are done in Leave-One-Speaker-Out (LOSO) manner for speaker-independent tests, and in j-fold Stratified-Cross-Validation (SCV) for speaker-dependent tests. Table 2 first depicts the baseline results for speaker-independent classification on the turn-level employing standard turn-wise derived acoustic features. The table also shows diverse optimization strategies as described in sect. 5. As can be seen, both speaker normalization, and feature selection help to improve overall performance. Note that features are thereby optimized over all speakers, as this is a speaker-independent task. Table 3 shows baseline results for speaker-dependent analysis. Speaker normalization is hereby superfluous and likewise omitted.

Table 2. *Baseline results by turn-level analysis skipping stage 1. Accuracies for EMO-DB, turn-wise feature extraction, considering speaker-normalization (SN), and feature selection (FS) for optimization, speaker-independent (SI) LOSO evaluation with SVM.*

SI Accuracy [%]	SN	FS	EMODB
Turn	-	-	74.9
Turn	√	-	79.6
Turn	√	√	**83.2**

Table 3. *Baseline results by turn-level analysis skipping stage 1. Accuracies for EMO-DB, turn-wise feature extraction, considering full set and feature selection (FS) for optimization, speaker-dependent (SD) evaluation with SVM.*

SD Accuracy [%]	FS	EMODB
Turn	-	80.0
Turn	√	**95.1**

Here again, a significant boost is obtained by feature space optimization. This time however, this step is carried out for each speaker individually resulting in very high mean accuracy. The mean optimal number of features is 57 with 42 as minimum and 86 as maximum. The comparison of table 2 and table 3 clearly show the difference between speaker-dependent and independent analysis.

Next, table 4 provides detailed number of chunks and syllables per emotion obtained by the chunking as described in sect. 2. Note that an almost constant factor of chunks per emotion resembling 3 is obtained. Disgust however shows a slightly different behavior.

Table 4. *Distribution among emotions, database EMO-DB. Considered are turns, automatically extracted chunks and syllables.*

[#]	Turns	Chunks	Syllables
Anger	127	269	1843
Boredom	79	225	1151
Disgust	38	173	516
Fear	55	160	794
Joy	64	179	927
Neutral	78	213	1093
Sadness	53	143	823
Sum	**494**	**1362**	**7147**

Apart from the mean number of chunks and syllables per emotion, table 5 depicts their frequencies in more detail. As can be seen, roughly a third of the turns is not chunked, but kept as turn.

Table 5. *Distribution among emotions, database EMO-DB. Considered are turns, automatically extracted chunks and syllables.*

[#]	Chunks	Syllables
1	167	-
2	86	-
3	95	-
4	65	-
5-9	78	94
10-14	3	135
15-19	-	156
20-29	-	109

The next table 6 shows first classification results for the aimed at sub-turn entities chunks and syllables as introduced in sect. 4. As for the base-line turn-level features, speaker normalization and feature space optimization are considered for optimization. In table 7 the same results as in table 6 are shown for the speaker-dependent case. Finally, we show results for the mapping of chunks or syllables onto turns by the diverse strategies introduced in sect. 4. Thereby only the optimal cases with speaker normalization and feature space optimization are considered, as chunk-level accuracy is crucial for the overall success. First, we investigate the speaker-independent-case in table 8. Thereby the three strategies majority vote (MV), maximum length (ML) and maximum length times prediction score (MLS) are considered.

Table 6. *Results by chunk-level analysis. Accuracies for EMO-DB, chunk-wise feature extraction, considering speaker-normalization (SN), and feature selection (FS) for optimization, speaker-independent LOSO evaluation with SVM.*

SI Accuracy [%]	SN	FS	EMODB
Chunk	-	-	42.6
Chunk	√	-	46.7
Chunk	√	√	**51.4**
Syllable	-	-	42.1
Syllable	√	-	44.6
Syllable	√	√	**47.6**

Table 7. *Results by chunk-level analysis. Accuracies for EMO-DB, chunk-wise feature extraction, considering feature selection (FS) for optimization, speaker-dependent 10-fold SCV evaluation with SVM.*

SD Accuracy [%]	FS	EMODB
Chunk	-	58.2
Chunk	√	**66.6**
Syllable	-	58.6
Syllable	√	**59.5**

Table 8. *Results by turn-level mapping. Accuracies for EMO-DB, chunk-wise features with speaker-normalization and feature selection, considering correct and correct* cases, by addition of non-unique winning-classes, speaker-independent (SI) LOSO evaluation with SVM.*

SI Acc. [%]	Strategy	Correct	Correct*
Chunk	MV	45.3	64.2
	ML	60.1	64.2
	MLS	70.6	**70.6**
Syllable	MV	42.8	60.1
	ML	56.9	60.1
	MLS	67.8	**67.8**

As can be seen, we discriminate between correct assignment, and cases, where the correct class has been the winner class among one or more other classes. Second, the same results are observed in the speaker-dependent case in table 9. Apparently, the chunk level accuracy is crucial for the mapping on the turn level: in the speaker-independent case roughly every second chunk is correct. Likewise, mapping cannot "repair" too many cases. This differs for the case of speaker dependent analysis, where around 2 out of 3 chunks are correct. Here, the mapping is closer to the accuracy obtained by turn-level features.

Table 9. *Results by turn-level mapping. Accuracies for EMO-DB, chunk-wise features with speaker-normalization and feature selection, considering correct and correct* cases, by addition of non-unique winning-classes, speaker-dependent (SD) LOSO evaluation with SVM.*

SD Acc. [%]	Strategy	Correct	Correct*
Chunk	MV	69.9	79.8
	ML	78.2	79.8
	MLS	88.4	**88.4**
Syllable	MV	66.3	80.0
	ML	79.9	80.0
	MLS	87.5	**87.5**

The main outcomes of these results are that the proposed chunking seems superior to annotation-based syllable chunking. However, turn-level features cannot be reached. This holds even after mapping on the turn-level by the investigated three different strategies.

8. DISCUSSION

In this work we analyzed emotion recognition within speech by sub-turn entities. An automatic chunking was introduced as opposed to annotation-based syllable chunking. The introduced approach was superior to syllables both for speaker-dependent, as well as for independent analysis. This may be due to the fact that it produces roughly 5 times longer segments, though at the same time 5 times fewer instances are obtained for robust training. However, both these sub-turn entities clearly fall behind turn-level analysis.

We secondly investigated mapping of these chunks on the turn-level by multi-instance learning. Yet, as a result for the used data-base no advantage over direct turn-level feature extraction can be reported. However, no turn-level feature information was integrated, which may lead to an advantage as in our former related experiments reported in [6], where chunk- and turn-level features were integrated in one super-vector. However, on databases as the AIBO set [8] changes of emotion are observed within a turn. At this point chunking may reveal its advantage.

Apart from these findings speaker normalization and correlation-based feature space optimization could be proven highly beneficial. Furthermore significant improvement of speaker-dependent over independent analysis was demonstrated.

In future works we aim at modeling of the chunk order, e.g. by means of dynamic modeling as Dynamic Bayesian Nets. Also, diverse training strategies as combined chunk learning or turn-level features learning independent of the unit may reveal facts about potential performance boost and incremental processing. Also analysis on further data-sets as SUSAS, DES, eNTERFACE, or SMARTKOM may show different potential of the investigated 2-stage processing. Finally we aim at investigation of the benefit that can be obtained by training of syllable-dependent models.

9. ACKNOWLEDGMENTS

This work has been conducted within the framework of the NIMITEK project (Sachsen-Anhalt Federal State funding). Bogdan Vlasenko acknowledges support by a graduate grant of the Federal State of Sachsen-Anhalt.

10. REFERENCES

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

[2] 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.

[3] Lee Z; Zhao Y.: "Recognizing emotions in speech using short-term and long-term features", *Proc. ICSLP 98*, pp. 2255-2558, 1998.

[4] 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.

[5] 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.

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

[7] 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.

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

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

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