Automatic multi-lingual arousal detection from voice applied to real product testing applications

Florian Eyben, Matthias Unfried, Gerhard Hagerer, Björn Schuller

Angaben zur Veröffentlichung / Publication details:

Eyben, Florian, Matthias Unfried, Gerhard Hagerer, and Björn Schuller. 2017. "Automatic multi-lingual arousal detection from voice applied to real product testing applications." In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 5-9 March 2017, New Orleans, LA, USA*, 5155–59. Piscataway, NJ: IEEE. https://doi.org/10.1109/ICASSP.2017.7953139.

AUTOMATIC MULTI-LINGUAL AROUSAL DETECTION FROM VOICE APPLIED TO REAL PRODUCT TESTING APPLICATIONS

Florian Eyben[1]*, Matthias Unfried*[2]*, Gerhard Hagerer*[1,4]*, Björn Schuller*[1,3]

[1]audEERING GmbH, Gilching, Germany; [2]Gfk-Nürnberg e.V., Nürnberg, Germany
[3]Department of Computing, Imperial College London, London, UK
[4]Chair of Complex & Intelligent Systems, University of Passau, Germany

ABSTRACT

A method is presented which applies Long Short-Term Memory Recurrent Neural Networks on real market-research voice recordings in order to automatically predict emotional arousal from speech. While most previous work has dealt with evaluations of algorithms within the same speech corpus, the novelty of this paper lies in an extensive evaluation across corpora and languages. The approach is evaluated on seven large data sets collected in real tests of TV commercials and new product concepts across four languages. We observe excellent performance within and between the different corpora when compared against the gold standard of arousal ratings by human annotators. Even in the cross-language validation the models show good performance which almost reaches human rater agreement.

Index Terms— Emotion recognition, Marketing research, Arousal, Speech, openSMILE

1. INTRODUCTION

According to [1], emotions emerge from a series of stimulus evaluation checks beginning with the appraisal of the relevance of a certain event based on the individual's goal. Thus, in the first place, an emotion can be considered as relevance detector (see [2]). Stimuli or events that are appraised as relevant for an individual (i.e., that relate to goals, needs, and values) elicit an emotional response by driving changes in action tendencies, expressions, peripheral nervous system, and conscious feeling (see [3]) and according to [4] the arousal level of emotional stimuli is closely linked to their influence on attention. Heightened emotional arousal also increases the likelihood that information is socially shared (see [5]).

Emotional arousal elicited by, e.g., products or commercials, is thus an important indicator to consider in marketing. Unsurprisingly, emotions in general play a central role in marketing and their measurement is of particular importance for market research and monitoring marketing campaigns. Especially, scalable and automated solutions are of great interest. Although there is some literature on the application of software-based inference of emotional states (especially valence) from facial expressions (see [6]) there are hardly any studies on voice-based emotion recognition in real market-research settings. Moreover, arousal can hardly be inferred from facial expressions (see [7]), highlighting the need to infer it from voice.

In this paper a state-of-the-art method is evaluated which predicts emotional arousal from voice recordings of participants of real market-research studies. To build classification models, a huge data base was built containing recordings from tests of TV commercials and new product concepts (Section 2). This data set was manually annotated and was used to train a context-aware deep recurrent neural network (Sections 5 and 6) with state-of-the-art acoustic features from openSMILE [8] as inputs (Section 4). Section 7 reports on extensive evaluations before we conclude in Section 8.

2. CORPORA

The data sets that have been used for the training of the arousal detector have been collected in real market-research studies. In total, seven studies have been conducted to collect voice recordings for different types of market-research applications (test of TV commercials and new product concepts) and languages (German, English, Spanish and Chinese).[1]

2.1. Test of TV commercials

The first study was a central location test (CLT) of TV commercials (TVC), conducted in Germany (hereafter corpus DE1). Respondents were exposed to three TV commercials intended to elicit different types of emotional responses: i) a funny automotive TVC for positive responses with a high level of emotional arousal, ii) a disgusting TVC for a tooth paste for negative responses with a high level of emotional arousal, and iii) a neutral TVC for a dishwasher tab for responses with low level of emotional arousal.

Respondents were asked to answer four neutral questions which could be used for individual calibration and five diagnostic questions on each TVC. During the neutral questions, respondents had to read a short paragraph from a fairy tale, had to introduce themselves, and had to explain how they gather information about fast moving consumer goods as well as durable investment goods. For diagnostic purposes, the respondents were asked to summarize the commercial in their own words, explain which parts they liked and which they disliked, how they felt during the exposure and why (or why not) they like the brand.

2.2. Concept Tests

Additionally, six studies in four languages (German, English, Chinese, and Spanish) have been conducted in which concepts of new products have been tested. The five main test stimuli depicted concepts for completely new electric tooth brushes.[2]

The German concept test has been conducted online in Germany (corpus DE2), the English one online in the US (corpus US1). Chinese data were collected online in Canada (corpus CN1) and in a

[1]The data have been collected by the authors and are available for academic purposes on request.

[2]In the English test eight additional concepts (cleaning items, sweets) were used.

CLT in the US (corpus CN2); the Spanish tests were conducted online in Mexico (corpus ES1) and in a CLT in the US (corpus ES2).

Similar to the TVC test, the participants were asked to answer a couple of neutral questions in which they had to count from one to ten, read a couple of short sentences and introduce themselves. Subsequently, they were exposed to the test concepts in randomized order and had to explain for all concepts what they think of it.

In total we collected data from 262 German, 520 English, 365 Spanish and 184 Chinese native speakers. Table 1 summarizes all corpora available for model training and evaluation.

Corpus	No. of files	Total hours	Duration (in seconds) min	avg	max
US1	2139	15:11	8.3	25.6	44.9
DE1	1186	07:01	0.7	21.3	169.2
DE2	553	01:32	1.3	10.0	58.2
CN1	165	00:31	1.2	11.3	55.9
CN2	573	02:35	1.6	16.2	65.3
ES1	602	02:11	1.4	13.0	100.7
ES2	538	03:35	1.3	24.0	107.5

Table 1. Overview of available market-research corpora.

3. ANNOTATION

In order to utilize the corpora for model training, the (perceived) emotional content of the recordings was assessed by human annotators. To this end, different student samples were recruited to assess the recordings with respect to three dimensions of the emotional reaction by the respondents.

The main pool of annotators consisted of 24 German psychology students. All corpora have been annotated by students drawn from this pool. [3] In addition, a pool of 24 Chinese (Mandarin) and 21 Spanish native speaking students of social and human sciences repeated the annotations for the respective language in order to check for language effects (which is not discussed in this paper). All recordings have been cut into snippets with a maximum duration of 10 seconds and each of these chunks was rated by at least four raters;[4] a core set in each language was annotated by all raters. The annotators had to assess different emotional dimensions: i) arousal (bipolar 21-point scale from "very low" to "very high"), valence (bipolar 21-point scale from "negative" to "positive"), and iii) interest (unipolar 6-point scale from "low" to "high").

Because a high level of emotional arousal (which could be associated with both positive or negative valence) is considered as an indicator for personal relevance and the likelihood that information is socially shared, from a market research perspective it is considered to be the more important and distinct dimension. Thus, model training was optimised only on this dimension and hence, only the results for emotional arousal are presented in this contribution. For all corpora, fairly high inter-rater correlations were obtained.

Table 2 depicts the range of emotional responses perceived by German raters and their inter-rater correlation.

Except for the English corpus ratings from either the same annotators or from native speakers have been available. Thus, either of both (or a combination) could be used for model training. While using the ratings of native speakers has the advantage that annotators are able to take into account cultural differences in the vocal expression of emotions, it has the disadvantage that the ratings could be biased by the content and that ratings are not comparable between corpora. On the contrary, using ratings from the one annotator group for all corpora could probably make the ratings more coherent but may result in ratings which reflect only the perception of a certain speaker group.

[3]Since all German annotators were English speaking we did not use an additional English native speaking annotator sample.

[4]Except for DE1 where raters assessed whole recordings.

Corpus	Ratings (avg. per snippet) min	mean	max	Inter-rater correlation
US1	-6.8	0.6	6.9	0.69
DE1	-5.8	-0.1	5.0	0.54
DE2	-6.8	0.6	6.6	0.71
ES1	-7.4	0.5	6.6	0.73
ES2	-7.7	0.2	5.8	0.61
CN1	-5.6	0.5	8.2	0.75
CN2	-4.3	1.1	7.0	0.41

Table 2. Range and average of arousal ratings of German annotators; mean inter-rater correlation (Pearson corr. coeff.) per corpus.

However, it should be noted that perception of emotional arousal (and even of valence and interest) is rather similar across rater pools, as seen from the Pearson correlation coefficients between ratings from German and Chinese/ Spanish annotators (0.78 for Spanish and 0.66 for Chinese). Additionally, the German raters had the highest mean inter-rater agreement (pearson correlation coefficient 0.67 vs 0.65 for Spanish and 0.7 vs. 0.62 for Chinese recordings). Thus, only the ratings of the German annotators for all corpora were used.

In order to obtain, from multiple ratings, a single gold standard for automatic classification, the Evaluator Weighted Estimator (EWE) is used (see [9] and [10]). The EWE is a weighted average of the individual ratings, where the weights are based on the average inter-rater reliability of each rater. All evaluations in this paper are based on the EWE.

4. ACOUSTIC FEATURES

The automatic classification of vocal arousal is based on two standard acoustic feature sets in the field of Computational Paralinguistics, extracted with the openSMILE toolkit version 2.3[5][8]: the ComParE 2016 feature set and the Extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS). The eGeMAPS Set was designed by experts based on theoretical and practical significance of parameters (see [11] for details). It contains a minimal set of 88 acoustic parameters relevant to vocal emotion expressions. The ComParE 2016 feature set is based on the ComParE 2013 baseline feature set [12], which is a brute-force feature combination leading to 6 373 features in total. The ComParE 2016 version contains updates regarding pitch jitter extraction and optimised computation of parameter ratios where the denominator value is very small (thus, in former versions, leading to single large ratio values).

For this study, a reduced set of functionals in the ComParE set is used to reduce complexity. The reduced set of functionals consists of: Linear Regression Slope (coefficient 1), Linear Regression Quadratic Approximation Error (linreqerrQ), Quadratic Regression Steepness (coefficient 1), Quadratic Regression Quadratic Approximation Error (qregerrQ), Arithmetic Mean, Standard Deviation, 6-th percentile, 94-th percentile, range between the two aforementioned

[5]`http://opensmile.audeering.com/`

percentiles. The selection was based on experience in previous work and algorithmic and numeric robustness of the functionals. The feature set contains 1170 features in total.

5. AROUSAL REGRESSION WITH LSTM-RNN

Long Short-Term Memory Recurrent Neural Networks (LSTM-RNN) [13] have been used very successfully in many applications ranging from audio filtering, voice activity detection, over handwriting recognition and speech recognition [14], to speech emotion recognition. The first studies on continuous, dimensional (arousal, valence) emotion recognition from speech on real-life databases with LSTM-RNN are reported in [15], [16], and [17].

LSTM-RNN have memory units and feedback loops with which they are able to model long-range temporal dependencies in time series in an excellent way. The network receives an input at every time-step and computes an output vector based on this input, the previous output, the previous states of the internal layers, and the states of the internal memory.

As the strength of LSTM is the modelling of time series, the feature extraction process had to be adapted. Typically, a single feature vector is computed per speech segment (utterance or sentence) (cf. [18]). However, this discards all temporal information in that segment and prevents LSTM networks from showing their strength of time series modelling.

Thus, a sliding window feature extraction approach was chosen. For each utterance a window (2 seconds long, approximating context of 2-3 words) is shifted forward over each recording at a rate of 1 second. This yields one feature vector every second. For evaluation, it is required to transform the sub-window predictions back to the original units of annotation (recording segments). This was done by averaging the predictions over the recording segments.

Acoustic features are normalised to have zero mean and unit variance on the training set. The means and variances computed for normalisation of the training set are applied to normalise the test data. Normalisation is required in order to ensure that the gradients while training the LSTM-RNN are numerically in an optimal range for tanh functions (symmetric around 0 with variance 1).

5.1. Multi-target modelling

LSTM-RNN are capable of multi-target regression, as they map an N-dimensional input vector to an M-dimensional output vector. Thus, multi-target modelling for arousal was investigated, where arousal, valence and interest are jointly learned, in order to evaluate whether a joint modelling can benefit from mutual information in all three dimensions. Multi-target modelling is compared to single-target modelling of arousal.

5.2. Network topologies

The choice of network topologies is a crucial factor. Based on previous experience (e.g., [17]), two deep network topoplogies were chosen: *Net01* with two hidden layers with 30 hidden units (LSTM cells) in the first layer and 20 in the second layer; *Net02* with two hidden layers with 50 hidden units (LSTM cells) in the first layer and 40 in the second layer. The layers are fully connected.

In addition to the standard (forward) LSTM-RNN, bidirectional LSTM-RNN networks (BLSTM) were investigated. These networks process the sequence of inputs both forwards and backwards and are thus able to make use of contextual information in both directions (cf. [14]). They are not suited for on-line, incremental processing as they require the complete sequence to be available. Regular LSTM, on the contrary, can process every frame individually and incrementally as it arrives.

For bidirectional networks, the same two topologies were chosen in order to keep the number of variables in the models approximately the same. However, it is to note that, these bidirectional networks have independent forward layers and backward layers of half of the size each of the layers in the forward LSTM (e.g., a forward and a backward layer for each hidden layer of size 15 and 10 for *Net01*). This leads to a slightly smaller total number of parameters (due to the fully connected layers) for the BLSTM than for the corresponding LSTM. The bidirectional topologies are referred to as *Net01b* and *Net02b*, respectively.

6. EXPERIMENTS

Evaluations are conducted in leave-one-corpus (or language) out mode to asses the performance on unknown languages (cross-language models), and in leave-one-speaker-group-out (LOSGO) cross-validation mode over all languages, in order to asses the possible performance for matched multi-language models. Also, as a baseline reference, evaluations on single corpora/languages are performed in LOSGO cross-validation. For LOSGO, cross-validation is used in 5 folds to obtain average results over all corpora/languages. Thereby, the speakers of the test corpora are randomnly split into 5 groups. Training partitions contain all data of the test corpora from speakers of 4 out of 5 groups and all data from the remaining training corpora. The data from the fifth speaker group of the test corpora are used as test set in this fold. The procedure is repeated 5 times so that each fold (speaker group) has been the test set once. Results are then averaged over all test folds.

As evaluation metric, the Pearson Correlation Coefficient (ρ) is computed between the network's output (predictions) and the EWE gold standard obtained from the human raters.

7. RESULTS AND DISCUSSION

[ρ]	LSTM		B-LSTM	
Corpus:	eGeMAPS	ComParE	eGeMAPS	ComParE
DE1	0.52	0.58	0.54	**0.61**
DE2	0.53	0.56	0.55	**0.59**
CN1	0.65	0.69	0.67	**0.70**
CN2	0.61	0.60	**0.65**	0.61
ES1	0.62	0.55	**0.65**	0.57
ES2	0.47	0.44	0.47	**0.50**
DE1 + DE2	0.51	**0.60**	0.52	0.59
CN1 + CN2	0.63	0.59	**0.66**	0.62
ES1 + ES2	**0.53**	0.47	0.50	0.51

Table 3. Automatic arousal prediction. Single target intra-corpus results for all languages. Net02 (LSTM) vs. Net02b (BLSTM). Pearson Correlation Coefficients (ρ). Best results per corpus in bold face.

In preliminary evaluations regarding the best network topology, it was found that are only minor differences (many of them not significant on a level of $p = 0.05$ in a two-tailed test) between *Net01* and *Net02*, with a slight preference of *Net02*. In order to not overload the results discussion, we decided to thus produce all results with topology *Net02* and *Net02b*.

First, intra-corpus results (LOSGO within a single corpus, and union of both corpora of the same language) are shown to set benchmarks and broadly asses the quality of each corpus in terms of recordings and annotations. All intra-corpus cross-validation results are given in Table 3. As can be seen, across many languages/corpora, bidirectional networks (BLSTM) are better or at least as good as unidirectional LSTM networks. We could observe the same trend in cross-corpus (leave-one-corpus-out) evaluation. To not overwhelm our reader, we therefore only show and discuss results for BLSTM network *Net02b* in the following.

[ρ]	eGeMAPS		ComParE 2016	
Test corpus:	1-targ	3-targ	1-targ	3-targ
US1	0.616	0.592	**0.633**	0.625
DE1+DE2	0.563	0.561	**0.603**	**0.603**
CN1+CN2	0.549	**0.664**	0.642	0.646
ES1+ES2	**0.563**	0.542	0.559	0.551

Table 4. Automatic arousal prediction. Leave-One-Speaker-Group-Out (LOSGO) modelling with all corpora joint for training and given corpora for testing (column 1). Net02b (B-LSTM). Pearson Correlation Coefficients (ρ); single vs. multi-target (1 vs. 3-targ) modelling.

Comparing the results in Table 4 with the inter-rater correlations in Table 2, we see that BLSTM-RNN are capable of building cross-language regression models for speech arousal of several languages with a performance nearly comparable to human rater agreement.[6] For real applications in marketing research (and also other domains) it is important to generalise to other recording set-ups and to other languages. In order to asses the capability of BLSTM to build a multi-language emotion model training was performed on all corpora together, and evaluation was done on selected languages in LOSGO mode. This means that data from the same corpus (excluding the test data/speakers) was used in training. Results obtained in this way rather give an upper bound performance estimation for a matched scenario (training data available that was recorded under same conditions (room, language, protocol, etc.) as test data). All such LOSGO results are shown in Table 4.

Comparing the LOSGO results to the intra-corpus results, shows no major drop in ρ. We can thus conclude that BLSTM, with the given topology *Net02* are well capable of building multi-lingual emotion recognition models from large amounts of data with similar performance on the four languages than intra-corpus models. For German and Spanish test conditions the performance of the multi-language models is even higher than the intra-corpus models. This shows, I) that adding data (even from other languages) is beneficial, and II) arousal affects the voice in similarily across these languages.

The most realistic evaluation setting is disjunctive cross-corpus testing. The results are given in Table 5. In cross-corpus testing, no data of the same language and same corpus (and thus recording set-up, room, microphone, etc.) are present in the training set. I.e., the training set contains all available corpora, excluding those listed as test corpora in column 1 of Table 5.

Due to the size of the US1 corpus and the most realistic setting (cross-language), the results in line 1 of Table 4 are the most significant findings of this study. It can be seen that the ComParE acoustic feature set by far outperforms the much smaller eGeMAPS feature set. This shows that for language independent modelling in realistic acoustic settings, other acoustic parameters than in the eGeMAPS set (and thus studied theoretically in voice sciences so far) are of importance. Over all results no significant trend for multi- vs. single-target learning can be found. This requires more detailed network parameter optimisations and comparisons in future studies.

[6] It should be noted that the recordings were obtained from identical studies across languages (except for DE1).

[ρ]	eGeMAPS		ComParE 2016	
Test corpus:	1-targ	3-targ	1-targ	3-targ
US1	0.426	0.422	**0.541**	0.519
DE1+DE2	0.447	0.452	**0.539**	0.501
CN1+CN2	**0.598**	0.588	0.577	0.569
ES1+ES2	**0.529**	0.507	0.524	0.510

Table 5. Cross-corpus (leave one language out) evaluation for arousal. Net02b (B-LSTM). Pearson Correlation Coefficient metric (ρ); single vs. multi-target (1 vs. 3-targ); clean (cl) vs. multi-condition (mc) training set.

8. CONCLUSION

Overall, it is also a remarkable result that an automatic classifier trained on German, Spanish, and Chinese vocal arousal data, achieves – with its predictions – a correlation of 0.541 (to the EWE gold standard) on completely unknown test data recorded at a completely different site (in the case of US1 even distributed across many remote recording sites) and in a different language. Compared to multi-language models, the cross-corpus result shows only an about 0.09 lower Pearson correlation. It must be noted, that this decrease could also be due to the reduced amount of training data (US1 is the largest corpus). The model evaluated for US1 cross-corpus evaluation is based mainly on the second largest corpora, DE1 and DE2. Intra-corpus evaluation of these two corpora results in $\rho = 0.59$ (with BLSTM and ComParE features).

For practical application, the model opens the possibility to measure emotional arousal in many quantitative domains (e.g., customer satisfaction, product and packaging testing, or user experience) and even in qualitative settings its application is conceivable.

9. RELATION TO PRIOR WORK

As discussed in Section 5, the work presented here bases on the principles in [15], [16], and [17]. For the first time, the concept of dimensional emotion recognition (here: arousal) with LSTM-RNN is applied to a large set of seven corpora collected in real marketing research studies across four different languages. Moreover, excellent cross-corpus and even cross-language performance of the models has been shown. According to earlier studies in [19] and [20], cross-corpus evaluation is a challenge, yet utterly important for real applications of automatic speech emotion recognition technologies in commercial and industry environments.

10. ACKNOWLEDGEMENTS

Research leading to these results has received funding from the European Community's Seventh Framework Programme under grant agreement No. 338164 (ERC Starting Grant iHEARu) and the European Union's Horizon 2020 research and innovation programme under grant agreement No. 680883 (ERC PoC VocEmoApI).

11. REFERENCES

[1] Klaus Scherer, "Emotion as a process: Function, origin, and regulation," *Social Science Information*, vol. 21, no. 4-5, pp. 555–570, 1982.

[2] Klaus Scherer, "What are emotions? and how can they be measured?," *Social Science Information*, vol. 44, no. 4, pp. 695–729, 2005.

[3] David Sander, Didier Grandjean, and Klaus Scherer, "A systems approach to appraisal mechanisms in emotion. neural networks," *Neural networks*, vol. 18, no. 4, pp. 317–352, 2005.

[4] Ulrich Schimmack, "Attentional interference effects of emotional pictures: threat, negativity, or arousal?," *Emotion*, vol. 5, no. 1, pp. 55–66, 2005.

[5] Jonah Berger, "Arousal increases social transmission of information," *Psychological Science*, vol. 22, no. 7, pp. 891–893, 2011.

[6] Jens-Uwe Garbas, Tobias Ruf, Matthias Unfried, and Anja Dieckmann, "Towards robust real-time valence recognition from facial expressions for market research applications," in *Proc. of the 5th International Conference on Affective Computing and Intelligent Interaction (ACII 2013)*, Geneva, Switzerland, Sept. 2013, pp. 570–575, IEEE.

[7] Marc Mehu and Klaus Scherer, "Emotion categories and dimensions in the facial communication of affect: An integrated approach," *Emotion*, vol. 15, no. 6, pp. 798–811, 2015.

[8] Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller, "Recent developments in openSMILE, the munich open-source multimedia feature extractor," in *Proc. of ACM Multimedia 2013*, Barcelona, Spain, 2013, pp. 835–838, ACM.

[9] Michael Grimm and Kristian Kroschel, "Evaluation of Natural Emotions Using Self Assessment Manikins," in *Proc. of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU) 2005*, Cancun, Maxico, Nov. 2005, pp. 381–385, IEEE.

[10] Michael Grimm, Emily Mower, Kristian Kroschel, and Shrikanth Narayanan, "Primitives based estimation and evaluation of emotions in speech," *Speech Communication*, vol. 49, pp. 787–800, 2007.

[11] Florian Eyben, Klaus Scherer, Björn Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Devillers, Julien Epps, Petri Laukka, Shrikanth Narayanan, and Khiet Truong, "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, July 2015.

[12] Björn Schuller, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus Scherer, Fabien Ringeval, Mohamed Chetouani, et al., "The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism," in *Proc. of INTERSPEECH 2013*, Lyon, France, 2013, pp. 148–152, ISCA.

[13] Stefan Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[14] Alex Graves, *Supervised sequence labelling with recurrent neural networks*, Doctoral thesis, Technische Universität München, Munich, Germany, 2008.

[15] Martin Wöllmer, Florian Eyben, Stephan Reiter, Björn Schuller, Cate Cox, Ellen Douglas-Cowie, and Roddy Cowie, "Abandoning Emotion Classes – Towards Continuous Emotion Recognition with Modelling of Long-Range Dependencies," in *Proc. of INTERSPEECH 2008*, Brisbane, Australia, Sept. 2008, pp. 597–600, ISCA.

[16] Florian Eyben, Martin Wöllmer, Alex Graves, Björn Schuller, Ellen Douglas-Cowie, and Roddy Cowie, "On-line emotion recognition in a 3-D activation-valence-time continuum using acoustic and linguistic cues," *Journal on Multimodal User Interfaces (JMUI)*, vol. 3, no. 1-2, pp. 7–19, Mar. 2010.

[17] Florian Eyben, Martin Wöllmer, and Björn Schuller, "A Multi-Task Approach to Continuous Five-Dimensional Affect Sensing in Natural Speech," *ACM Transactions on Interactive Intelligent Systems, Special Issue on Affective Interaction in Natural Environments*, vol. 2, no. 1, Mar. 2012, Article No. 6, 29 pages.

[18] Dimitrios Ververidis and Constantine Kotropoulos, "Emotional speech recognition: Resources, features, and methods," *Speech Communication*, vol. 48, no. 9, pp. 1162–1181, Sept. 2006.

[19] Florian Eyben, Anton Batliner, Björn Schuller, Dino Seppi, and Stefan Steidl, "Cross-Corpus Classification of Realistic Emotions - Some Pilot Experiments," in *Proc. of the 3rd International Workshop on EMOTION: Corpora for Research on Emotion and Affect, satellite of LREC 2010*, Laurence Devillers, Björn Schuller, Roddy Cowie, Ellen Douglas-Cowie, and Anton Batliner, Eds., Valletta, Malta, May 2010, pp. 77–82, European Language Resources Association (ELRA).

[20] Björn Schuller, Bogdan Vlasenko, Florian Eyben, Martin Wöllmer, André Stuhlsatz, Andreas Wendemuth, and Gerhard Rigoll, "Cross-corpus acoustic emotion recognition: Variances and strategies," *IEEE Transactions on Affective Computing (TAC)*, vol. 1, no. 2, pp. 119–131, Aug. 2010.