Naturalistic affective expression classification by a multi-stage approach based on Hidden Markov Models

Hongying Meng and Nadia Bianchi-Berthouze

UCL Interaction Centre, University College London, London, UK
h.meng@ucl.ac.uk,n.berthouze@ucl.ac.uk

Abstract. In naturalistic behaviour, the affective states of a person change at a rate much slower than the typical rate at which video or audio is recorded (e.g. 25fps for video). Hence, there is a high probability that consecutive recorded instants of expressions represent a same affective content. In this paper, a multi-stage automatic affective expression recognition system is proposed which uses Hidden Markov Models (HMMs) to take into account this temporal relationship and finalize the classification process. The hidden states of the HMMs are associated with the levels of affective dimensions to convert the classification problem into a best path finding problem in HMM. The system was tested on the audio data of the Audio/Visual Emotion Challenge (AVEC) datasets showing performance significantly above that of a one-stage classification system that does not take into account the temporal relationship, as well as above the baseline set provided by this Challenge. Due to the generality of the approach, this system could be applied to other types of affective modalities.

Keywords: Emotion recognition, affective computing, multi-stage recognition, affective dimensions, spontaneous emotions, Hidden Markov Models.

1 Introduction

In the affective computing field [12], various studies have been carried out to create systems that can recognize the affective states of their user by analyzing their vocal [1], facial [11] [17], and body expressions [4], and even their physiological changes [6]. Most of the work has been carried out on acted or stereotypical expressions. More recently, there has been an increasing need to move towards naturalistic expressions in order to create systems that can interact with people in their everyday life. Naturalistic expressions, differently from acted ones, change slowly as a person interacts with the environment. The AVEC challenge [13] provides a unique dataset of naturalistic audio and facial expressions to help address this issue. These data have been recorded at a high sampling rate making it possible to capture and analyze the slow transition between affective expressions. The strong relationship between consecutive units (e.g., frames in a video, utterance in a vocal expression) is an important source of information on the basis of which to decide what expression the unit belongs to.

In this paper, we propose to use Hidden Markov Models (HMM) to model this spontaneous process and create a system that is able to recognize the affective content of the expression. Whilst the proposed approach is general, in this paper we test it on the audio dataset in which the units of expression are the way verbal words are expressed. The AVEC dataset uses binary affective dimension levels to label each expression unit, however, our approach can be extended to deal with a larger set of discrete states.

2 Related Work

Our work is not the first work to propose to exploit the temporal relationship existing between recorded observations. Several methods have been proposed for building automatic affective expressions recognition systems from audio and video, with interesting results.

Long Short-Term Memory (LSTM) Recurrent Neural Networks have been successfully used for modelling the relationship between observations [15] [2] [9] [16]. Wöllmer et al. [15] first proposed a method based on LSTM recurrent neural networks for continuous emotion recognition that included modelling of long-range dependencies between observations. This method outperformed techniques such as Support Vector Regression (SVR). Eyben et al. [2] used it for audiovisual classification of vocal outbursts in human conversation and the results showed significant improvements over a static approach based on Support Vector Machines (SVM). Nicolaou et al. [9] also used LSTM networks to outperform SVR due to their ability to learn past and future contexts. Wöllmer et al. [16] used Bidirectional Long Short-Term Memory (BLSTM) networks to exploit long-range contextual information for modelling the evolution of emotions within a conversation.

Eyben et al. [3] proposed a string-based prediction model and multi-model fusion of verbal and nonverbal behavioral events for the automatic prediction of human affect in a continuous dimensional space. Recently, Nicolaou et al. [8] described a dimensional and continuous prediction method for emotions from naturalistic facial expression that augments the traditional output-associative relevance vector machine regression framework by learning non-linear input and output dependencies inherent to the affective data.

HMM is another method typically used to model processes characterized by temporal relationships. Nwe et al. [10] used a four-state fully connected HMM to recognize six archetypical emotions from speech, obtaining recognition performance comparable to subjective observers' ratings. A study by Lee et al. [5] showed that HMMs produce more interesting results when the modelling is not performed at the level of the emotional expression but at the level of the units composing them (phonemes in their case). In this paper, we propose to exploit HMMs to classify units of emotional expressions according to levels of affective dimensions. Differently from previous work, we propose to use the HMMs in a

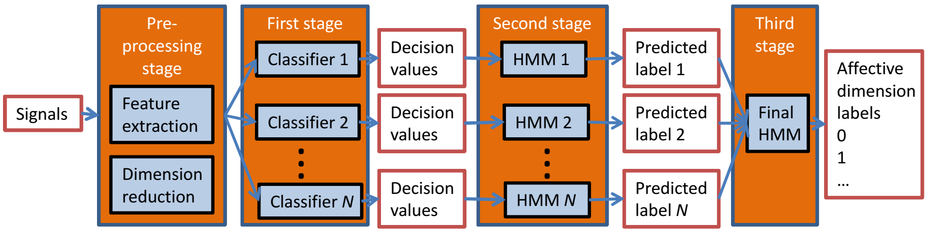

Fig. 1. Overview of the multi-stage automatic affective dimension level classification system. After the pre-processing stage, there are three classification stages, the last two of which are based on HMMs.

second stage of the classification process after a pre-decision has been made by exploiting other local methods. We also combine multiple classifiers into another HMM in the third stage to boost the overall performance. Indeed, it has been shown that multi-classifiers systems can outperform traditional approaches while simultaneously reducing computational requirements (see [1] for a review).

3 A Multi-stage Affective Dimension Level Classification System

3.1 System Overview

We propose the multi-stage automatic affective dimension level classification system shown in Figure 1. The system perform an initial pre-processing stage and three classification stages here described.

In the pre-processing stage, feature extraction and dimension reduction are implemented on each unit (e.g., uttered words in the experiments reported in this paper). Here, the dimension reduction is done by PCA.

In the first classification stage, the system aims to provide a classification of each unit according to the level of affective dimension the unit expresses. Each unit is treated independently from the other units. A set of different classifiers is used to improve the classification of each single unit. The output of each classifier is a set of decision values indicating the likelihood that the classified unit expresses a particular affective dimension level (e.g., the probability to express high arousal). For simplicity, we call this set of values the decision values.

Each classifier of the first stage is paired with an HMM in the second stage. The output of each first-stage classifier is hence used as input to its HMM. Each HMM reclassifies each unit by taking into account the temporal relationship with the other units also classified at the first stage.

Finally, for the third stage, we propose to use a single HMM to combine the predicted labels from all the HMMs of the second stage and reclassify each unit in the sequence according to the affective dimension levels it expresses.

3.2 First-stage Classification

This is a standard pattern recognition system in which every unit of the data is treated as sample. The temporal relationship between these units is not taken into account. For the classification itself, any classifier can be used here. The output of the classifier can be real values like the posterior probabilities in Naive Bayes classification, or the decision values in Support Vector Machines (SVM). Here, we propose to use the K-Nearest neighbour algorithm because it is simple and, as explained in section 4.2, can conveniently produce a very limited set of observed states for each sequence of units to be processed by an HMM in the second stage.

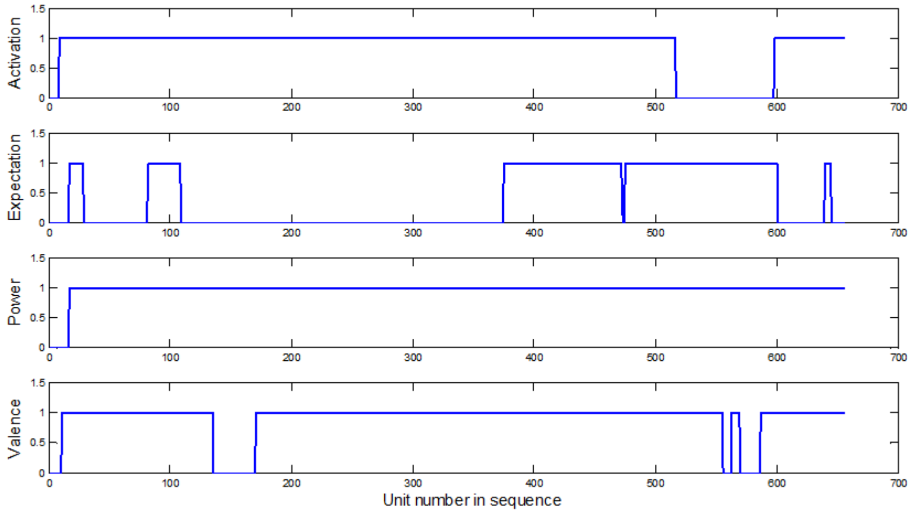

Fig. 2. Four affective dimension level labels (*activation,expectation,power and valence*) in an audio word sequence sample. Most of the affective dimension levels are changing very slowly. Only *expectation* tends to have some faster changes in the whole sequence.

4 Hidden Markov Model

4.1 HMM Design

The main reason for considering HMM as modelling approach is that in a naturalistic affective expression labeled as a sequence of affective dimension levels we can observe the Markov property. Figure 2 shows an example of an audio recording whose units have been labeled according to levels over four different affective dimensions. Every expressed word (i.e., each unit) was labeled with a set of levels, one for each affective dimension. The levels considered in the AVEC database are binary: '0','1'. '1' means high level (e.g., high arousal) and '0' denotes low level (e.g., low arousal). The level of an affective dimension of one word

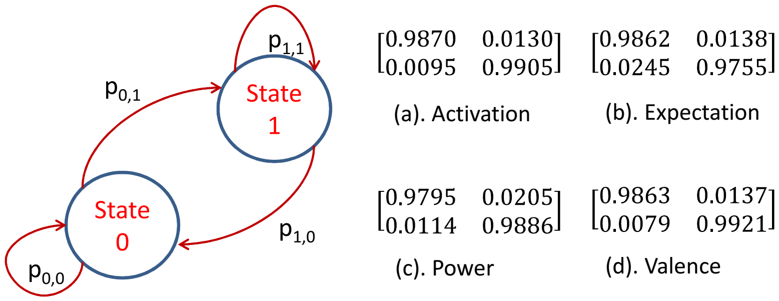

Fig. 3. Two hidden states in the HMMs and their transition matrices for four affective dimensions computed on a subset of the AVEC audio dataset. These two states are associated with the two levels of an affective dimension.

is very likely to be the same as that of the previous word. In the case of a more refined description of number of levels per dimension, we could expect that the level assigned to two consecutive words would be highly identical or very similar.

Based on this typical Markov property presented in these sequences, for each affective dimension, we design HMMs with two hidden states: '0' and '1' as shown in Figure 3. These two states are exactly associated with the two levels of an affective dimension. These hidden states capture the temporal structure of the data. $p_{0,0}$ and $p_{1,1}$ are the probabilities the system remains in the current state and $p_{0,1}$ and $p_{1,0}$ are the transition probabilities between states. For each dimension, a typical transition matrix is represented in Figure 3.

4.2 HMMs in the Second Stage Classification

For the HMMs in the second stage, the observed sequence was obtained based on the decision values from the first-stage classification. Each classifier of the first stage is paired with a HMM in this second stage. These decision values output by the first stage can be continuous values, or discrete values depending on the classifier used. When the decision values are continuous, Gaussian Mixture Models can be used to estimate their probability distribution. When the decision values are discrete, discrete probability matrices can be estimated for each symbol on each state.

In this paper, KNN was used in the first stage. Discrete HMMs can be built based on the decision function from KNN as shown in Figure 4. For example, when K=5, the neighbours of a sample are 5 samples with label '1' or '0'. The probability of the label of this sample will be '0' depending on the number of '0' in its neighbours' labels. The number of '0' can be counted and there are only 6 possibilities: 0,1,2,3,4,5. Here, we simply choose this count number as observed value of the HMMs, as shown in Figure 4.

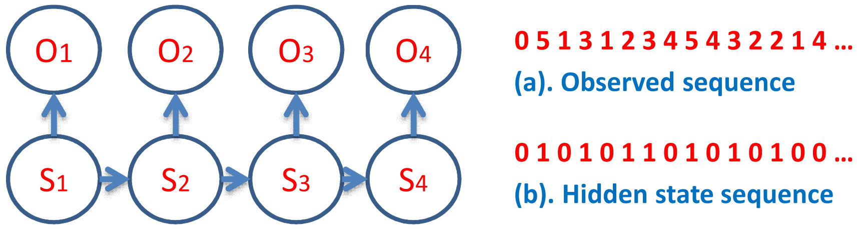

Fig. 4. HMMs used in the second stage. The observed values O_i are obtained from the decision values from the first classification stage. This example shows a possible observed sequence outputted by a KNN when K=5. The hidden states S_i are chosen from states '0' and '1' as the hidden state sequence illustrates.

4.3 HMM in the Third Stage Classification

Another HMM was built for the third stage classification based on the predicted labels ('0' or '1') from the multiple HMMs in the second stage. This is a decision fusion stage where the Markov property of temporal relationships in the sequences is taken into account. As in the second stage, the count number of how many '0' were predicted is used as observed value. For example. a second stage with 5 HMMs is equivalent to a KNN with K=5 in the first stage for HMM modelling.

4.4 HMM Implementation

For the HMM training, the state transition matrix can be directly estimated from the labels in the training set. The state emission matrices can be estimated from the discrete probability distribution of the decision values from previous classifications. For the HMM testing, the classification problem is converted into a best path finding problem for the decision value sequence. The Viterbi algorithm [14] was used to produce the best match label sequence.

Although there are some commercial or free HMM software packages available from Internet, this study did not use them because our model is simple and the matrices can be estimated directly from the data. The HMM can be designed to be more complex when the decision values are vectors. In the following experiments, only these simple HMMs were used. They were trained separately for four affective dimensions as a possible relationship between these dimensions was not taken into consideration.

5 Experimental Results

5.1 Dataset and features

The challenge data is constructed from the SEMAINE database, which consists of a large number of emotionally coloured interactions between a user and an

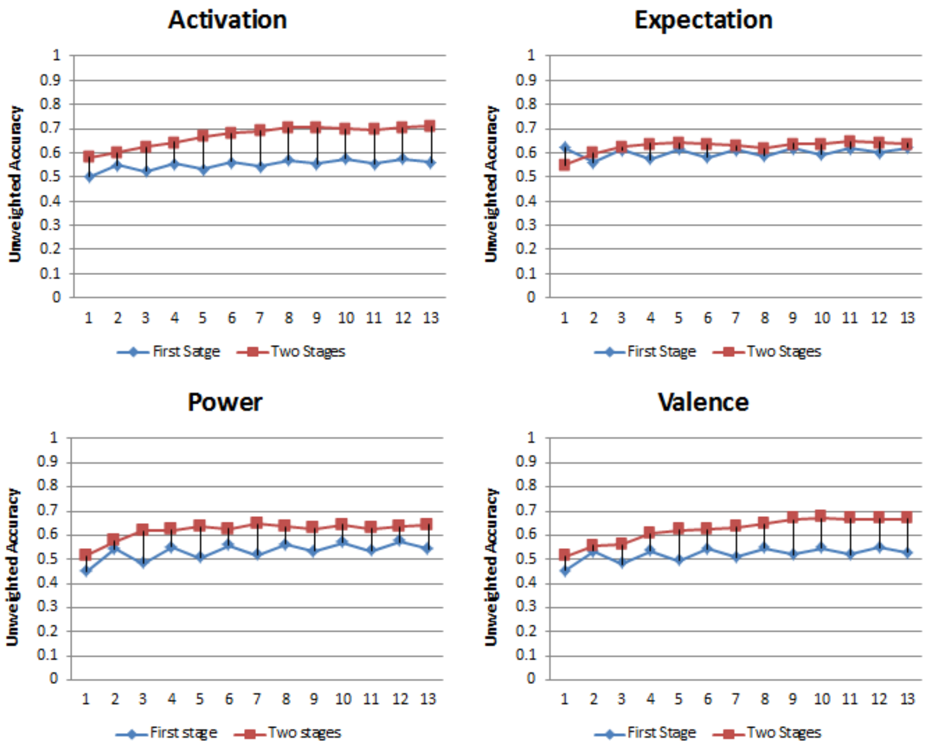

Fig. 5. Unweighted accuracy comparison for four affective dimension levels. 13 classifiers were used to train both a first-stage only classification system and a two-stage classification system. The performance was improved significantly for affective dimensions *activation, power and valence*. There is only marginal improvement in *expectation*.

emotionally stereotyped character. More specifically, the AVEC2011 dataset has been created from the first 140 operator-user interactions, which constitutes the Solid-SAL partition of the SEMAINE database. The Solid-SAL partition consists of users interacting with another person who plays the role of the emotionally stereotyped character. There are 31 sequences in training set and 32 and 11 sequences in development and test dataset. The SEMAINE database is fully described in [7]. Only those features provided by the challenge organizer were used. For the audio dataset, each uttered word is described by a vector of 1941 features and a set of four labels representing the level of *activation, valence, expectation* and *power*. The detailed description of the features can be found in [13]. In order to reduce the dimension of the feature vector, PCA was used and only 100 principle components were selected in the following experiments as they covered most of the variance.

5.2 Results

The system was trained on the training dataset and firstly tested on the development dataset. 13 different KNN classifiers were used for the first-stage classification tasks. The output of the KNN classifiers were inputted to the HMM models and their corresponding two-stage results were obtained. All samples in the training dataset were used for training and all samples in the development

Table 1. Unweighted accuracy for first-stage and two-stage classifications on the development dataset. 13 KNN classifiers (K=2,3,...,14) were used in this experiment.

Accuracy	first-stage classification				Two-stage classification			
%	activation	expectation	power	valence	activation	expectation	power	valence
KNN-02	50.00	62.09	44.90	45.36	58.21	54.84	51.49	51.09
KNN-03	54.96	55.99	54.30	53.36	60.13	59.74	57.54	55.74
KNN-04	52.31	61.34	48.45	47.94	62.35	62.37	61.72	56.18
KNN-05	55.82	57.17	55.18	53.49	64.28	63.80	62.06	60.61
KNN-06	53.41	61.52	50.74	49.45	66.65	64.21	63.64	62.14
KNN-07	56.47	57.98	55.90	54.42	68.02	63.56	62.60	62.45
KNN-08	54.65	61.37	51.89	50.88	69.10	62.96	64.57	63.30
KNN-09	57.00	58.67	56.33	54.75	70.47	61.94	63.56	64.83
KNN-10	55.50	61.79	53.39	51.91	70.39	63.75	62.83	66.38
KNN-11	57.48	59.38	56.88	54.71	69.77	63.69	64.38	67.06
KNN-12	55.56	61.90	53.48	52.02	69.63	64.75	62.77	66.59
KNN-13	57.26	60.06	57.12	55.04	70.55	64.37	63.65	66.72
KNN-14	56.09	62.16	54.44	52.78	70.94	63.82	64.28	66.76

dataset were used for testing. In this first testing phase, only the first two stages of the system are considered and compared in order to evaluate the contribution made by the second stage to the first stage for each individual classifier.

The unweighted accuracy results for thirteen KNNs are shown in Figure 5 and detailed values are shown in Table 1. We can observe from the table a clear improvement in recognition rate between two-stage classification and first-stage classification for *activation*, *power* and *valence* dimensions. Instead, there is only marginal improvement for the *expectation* dimension. A 1-tailed paired t-test and a 1-tailed non-parametric Wilcoxon signed-ranked test confirmed the significance of the improvement between first and second stage with p-values < 0.0001 for *activation*, *power* and *valence* dimensions and p-values $= 0.01$ for the *expectation* dimension.

Table 2. Comparison of recognition rates on test dataset between the proposed three-stage method and the baseline method.

Accuracy	Activation		Expectation		Power		Valence	
%	WA	UA	WA	UA	WA	UA	WA	UA
Baseline	55.0	57.0	52.9	54.5	28.0	49.1	44.3	47.2
Multi-stage	**64.3**	**66.2**	**57.0**	**58.6**	**41.3**	**54.4**	**50.5**	**51.4**

The multi-stage method (including the third stage) was finally fully tested on the test dataset of the AVEC audio sub-challenge that contains 11 samples of audio sequences. The overall performances are shown in Table 2 and compared with the baseline performance provided with the AVEC dataset. Our results clearly show our method outperforms the baseline rates for all the affective dimensions.

6 Discussion and Conclusion

In this paper, HMMs were proposed to model the classification of a unit of affective expression by taking into account the naturally slow changes in terms of levels of affective dimensions occurring within an affective expression. A multi-stage automatic affective dimension level classification system was built. The process could be reduced into a two-stage system if the classifier in stage one was not used. In this case, the dimensionality of the feature vectors would have to be significantly reduced in order to be processed by the HMMs. The computing load would then be much higher than with the multi-stage system proposed here. Indeed, the classifier in the first stage reduces the dimensionality of the feature vector to one dimension, thus making the system faster.

The key idea of the paper is that the hidden states of the HMMs are associated with the levels of affective dimensions. Therefore, the classification problem is converted into a best path finding problem in HMMs. The Viterbi algorithm can be used to produce the best match label sequences. Our system was tested on the audio data of the AVEC challenge datasets and performance was shown to improve significantly in comparison to a one-stage classifier that does not consider the temporal information. In our tests with the development set, performance improved significantly for almost all 13 classifiers used. For the test dataset, our method outperformed the baseline method significantly.

An interesting development of this approach will be to take into consideration possible correlations between affective dimensions. Furthermore, given the generality of the approach, it will be interesting to test it on other modalities such as the video database, replacing words in audio with frames in videos.

7 Acknowledgments

This work was supported by EPSRC grant EP/G043507/1: Pain rehabilitation: E/Motion-based automated coaching.

References

1. Moataz El Ayadi, Mohamed S. Kamel, and Fakhri Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572 – 587, 2011.
2. F. Eyben, S. Petridis, B. Schuller, G. Tzimiropoulos, and S. Zafeiriou. Audiovisual classification of vocal outbursts in human conversation using long-short-term memory networks. In *Proceedings of IEEE Intl Conf. Acoustics, Speech and Signal Processing (ICASSP11)*, Prague, Czech Republic, May 2011.

3. F. Eyben, M. Wollmer, M.F. Valstar, H. Gunes, B. Schuller, and M. Pantic. String-based audiovisual fusion of behavioural events for the assessment of dimensional affect. In *Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition (FG'11)*, Santa Barbara, CA, USA, March 2011.
4. A. Kleinsmith, N. Bianchi-Berthouze, and A. Steed. Automatic recognition of non-acted affective postures. *IEEE Transactions on Systems, Man and Cybernetics, Part B.*, 2011. In press.
5. Chul Min Lee, Serdar Yildirim, Murtaza Bulut, Abe Kazemzadeh, Carlos Busso, Zhigang Deng, Sungbok Lee, and Shrikanth Narayanan. Emotion recognition based on phoneme classes. In *Proc. ICSLP04*, pages 889–892, 2004.
6. Regan L. Mandryk, Kori M. Inkpen, and Thomas W. Calvert. Using psychophysiological techniques to measure user experience with entertainment technologies. *Behaviour & IT*, 25(2):141–158, 2006.
7. G. Mckeown, M.F. Valstar, R. Cowie, and M. Pantic. The semaine corpus of emotionally coloured character interactions. In *Proceedings of IEEE Int'l Conf. Multimedia, Expo (ICME'10), Singapore*, pages 1079–1084, July 2010.
8. M.A. Nicolaou, H. Gunes, and M. Pantic. Output-associative rvm regression for dimensional and continuous emotion prediction. In *Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition (FG'11)*, Santa Barbara, CA, USA, March 2011.
9. Mihalis A. Nicolaou, Hatice Gunes, and Maja Pantic. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE Transactions on Affective Computing*, 2:92–105, 2011.
10. Tin Lay Nwe, Say Wei Foo, and Liyanage C. De Silva. Speech emotion recognition using hidden markov models. *Speech Communication*, 41(4):603 – 623, 2003.
11. Maja Pantic and Leon J.M. Rothkrantz. Automatic analysis of facial expressions: The state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:1424–1445, 2000.
12. Rosalind W. Picard. *Affective Computing*. The MIT Press, 1997.
13. B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic. Avec 2011 the first international audio/visual emotion challenge. In *Proceedings of International HUMAINE Association Conference on Affective Computing and Intelligent Interaction 2011 (ACII 2011)*, Memphis, TN,, USA, Oct. 2011.
14. A. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2):260–269, April 1967.
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 *INTERSPEECH*, pages 597–600, 2008.
16. Martin Wöllmer, Angeliki Metallinou, Florian Eyben, Björn Schuller, and Shrikanth S. Narayanan. Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional lstm modeling. In *INTERSPEECH*, pages 2362–2365, 2010.
17. Zhihong Zeng, Maja Pantic, Glenn I. Roisman, and Thomas S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(1):39–58, 2009.