Context-Sensitive Learning for Enhanced Audiovisual Emotion Classification

Angeliki Metallinou, *Student Member*, *IEEE*, Martin Wöllmer, *Member*, *IEEE*, Athanasios Katsamanis, *Member*, *IEEE*, Florian Eyben, *Student Member*, *IEEE*, Björn Schuller, *Member*, *IEEE*, and Shrikanth Narayanan, *Fellow*, *IEEE*

Abstract—Human emotional expression tends to evolve in a structured manner in the sense that certain emotional evolution patterns, i.e., anger to anger, are more probable than others, e.g., anger to happiness. Furthermore, the perception of an emotional display can be affected by recent emotional displays. Therefore, the emotional content of past and future observations could offer relevant temporal context when classifying the emotional content of an observation. In this work, we focus on audio-visual recognition of the emotional content of improvised emotional interactions at the utterance level. We examine context-sensitive schemes for emotion recognition within a multimodal, hierarchical approach: bidirectional Long Short-Term Memory (BLSTM) neural networks, hierarchical Hidden Markov Model classifiers (HMMs), and hybrid HMM/BLSTM classifiers are considered for modeling emotion evolution within an utterance and between utterances over the course of a dialog. Overall, our experimental results indicate that incorporating long-term temporal context is beneficial for emotion recognition systems that encounter a variety of emotional manifestations. Context-sensitive approaches outperform those without context for classification tasks such as discrimination between valence levels or between clusters in the valence-activation space. The analysis of emotional transitions in our database sheds light into the flow of affective expressions, revealing potentially useful patterns.

Index Terms—Audio-visual emotion recognition, temporal context, Hidden Markov models, bidirectional long short term memory, recurrent neural networks, emotional grammars.

1 Introduction

Human emotional expression is a complex process where a variety of multimodal cues interact to create an emotional display. Furthermore, emotions are usually slowly varying during a conversation and the perception of an emotional display is affected, among others, by recently perceived past emotional displays, which place the expressed emotion into context. Taking into account such contextual information may prove to be advantageous for real-life automatic emotion recognition systems that can process a great variety of complex, vague, or ambiguous emotional displays. The focus of this paper is to investigate learning frameworks for automatic, multimodal emotion recognition that allow the use of information of the structure of past and future evolution of an emotional interaction. The study also considers the flow of emotional expression by examining emotional transitions in a variety of improvised affective interactions.

- *A. Metallinou, A. Katsamanis, and S. Narayanan are with the Department of Electrical Engineering, University of Southern California, 3740 McClintock Ave. EEB 400, Los Angeles, CA 90089-2560. E-mail: metallin@usc.edu, {nkatsam, shri}@sipi.usc.edu.*
- *M. Wöllmer, F. Eyben, and B. Schuller are with the Institute for Human-Machine Communication, Technische Universität München, Theresienstr. 90, 80333 München, Germany. E-mail: {woellmer, eyben, schuller}@tum.de.*

Manuscript received 16 Apr. 2011; revised 1 Dec. 2011; accepted 8 Dec. 2011; published online 27 Dec. 2011.
Recommended for acceptance by R. Calvo.
For information on obtaining reprints of this article, please send e-mail to: taffc@computer.org, and reference IEEECS Log Number TAFFC-2011-04-0033.
Digital Object Identifier no. 10.1109/T-AFFC.2011.40.

Psychology research suggests that human perception of emotion is relative and emotional understanding is influenced by context. Context in human communication can broadly refer to linguistic structural information, discourse information, cultural background and gender of the participants, knowledge of the general setting in which an emotional interaction is taking place, etc. For instance, psychology literature indicates that facial information viewed in isolation might not be sufficient to disambiguate the expressed emotion and humans tend to use context, such as past visual information [1], general situational understanding [2], past verbal information [3], and cultural background [4] to make an emotional decision. Also, emotions are expressed through the interplay of multiple complementary, supplementary, and even conflicting modalities (facial gestures, prosodic information, lexical content), and therefore such multimodal cues provide context for each other [5]. For example, discordance in the emotions expressed by the facial and vocal modalities degrades a subject's ability to correctly identify the emotion expressed by face or voice separately [6]. Furthermore, emotions are usually slowly varying states, typically lasting from under a minute to a few minutes [7]. Therefore, an emotion may span several consecutive utterances of a conversation and emotional transitions are usually smooth. For example, it seems reasonable to assume that an angry utterance is more likely to be succeeded by one displaying anger rather than happiness.

Context awareness is recognized as an important element in human-computer interfaces and can be broadly defined as an understanding of the location and identity of

the user as well as the type and timing of the human-computer interaction [8]. In the emotion recognition literature, relatively few works make use of contextual information and generally use diverse context definitions. In [9], the authors propose a unimodal framework for short-term context modeling in dyadic interactions, where speech cues from the past utterance of the speaker and his interlocutor are taken into account during emotion recognition. In [10], lexical and dialog act features are used in addition to acoustic (segmental/prosodic) features. In [11], the authors make use of prosodic, lexical, and dialog act features from the past two turns of a speaker for recognizing the speaker's current emotional state. In [12], the author describes a framework for building a tutoring virtual agent, where the tutor's behavior takes into account the student's recent emotional state as well as a variety of contextual variables such as the student's personality and the tutor's goal. In [13] and [14], the authors use the formalization of domain ontologies to describe generic frameworks for defining relations between emotion and context concepts such as environment, physiological cues, cultural information, etc. In our previous work, we have used neural network (NNs) architectures such as Bidirectional Long Short Term Memory (BLSTM) neural networks that take into account an arbitrary amount of past and future audio-visual emotional expressions to recognize the current emotion of a speaker [15].

Following our previous work [16], we define context to be information about the emotional content of audio-visual displays that happen before or after the observation that we examine. We focus on emotion recognition at the utterance level. Here, an utterance is loosely defined as a chunk of speech where the speaker utters a thought or idea. The phrases that we examine have been manually segmented from longer dyadic conversations and usually last a few seconds. In addition to the current utterance's audio-visual cues, we exploit information from an arbitrary number of neighboring utterances that could range from one past or future utterance to all the utterances of the conversation. Apart from this definition of context, which is our primary focus, we could also interpret the use of audio-visual cues as another form of context where the interplay within the multimodal streams provides context for one another and offers a fuller picture of the expressed emotion.

We investigate three alternative multimodal and hierarchical schemes for incorporating contextual information in emotion recognition systems by modeling emotional evolution at two levels: within an emotional utterance and between emotional utterances of a conversation. Specifically, we examine the use of hierarchical Hidden Markov Model (HMM) classifiers [17], of Recurrent Neural Networks (RNNs), and specifically BLSTM neural networks [18], [19] as well as the use of a hybrid BLSTM/HMM approach. The HMM-based classification is inspired by the Automatic Speech Recognition (ASR) literature, where algorithms exploit context at multiple levels within a Markov model structure: from phonetic details including coarticulation in speech production to word transitions reflecting language-based statistics [20], [21]. We hypothesize that similar within and across model transitions can be advantageously used to capture the dynamics in the evolution of emotional states, including within and across emotional categories. Alternatively, RNN architectures are a powerful, discriminative framework that enable modeling the emotional flow of a conversation without making Markov assumptions about emotional transitions. Here, we apply BLSTM neural networks which overcome the vanishing gradient problem of conventional RNNs and are able to learn from an arbitrarily large amount of past and future contextual information [15].

For our experiments we use a large multimodal and multisubject database of dyadic interactions between actors, namely, the IEMOCAP database [22], which contains detailed facial information, obtained from facial Motion Capture (MoCap) as well as speech information. The IEMOCAP database consists of dyadic conversations that are elicited so as to contain emotional manifestations that are nonprototypical and resemble real-life emotional expression. Our goal is to obtain a realistic assessment of an emotion recognition performance when our system is required to make a decision about the emotional content of all possible input utterances, including those containing subtle or ambiguous emotions.

We focus on the recognition of dimensional emotional descriptions, i.e., valence and activation levels, instead of categorical emotional tags, such as "anger" or "happiness." Valence describes how positive versus negative and activation how calm versus excited is the expressed emotion. We derive a dimensional label for all available utterances by averaging the decisions of multiple annotators. In addition to classifying the degree of valence and activation separately, we also investigate their joint modeling by classifying among clusters in the two-dimensional valence-activation space [23]. Our analysis of the relation between clusters and categorical labels shows that each cluster tends to contain certain categorical emotional manifestations, which allows us to draw association between a cluster and a generic categorical emotional content. Modeling of emotional transitions between utterances of an interaction could be formulated equivalently using the concept of probabilistic emotional grammars, which could inform us about the structure of emotional evolution during affective conversations.

Our experimental results show that incorporating temporal context in emotion classification systems generally leads to improvement in average performance for our classification tasks, except for the case of activation. For most of our context-sensitive classifiers we consistently observe an increase in performance compared to classifiers that do not take context into account. Such improvements are statistically significant for the valence and the three cluster classification task for classifiers such as hierarchical HMM and hybrid HMM/BLSTM. These results suggest that context-sensitive approaches could pave the way for better performing emotion recognition systems.

2 Context-Sensitive Frameworks

2.1 Hierarchical Context Sensitive Frameworks

Our problem can be posed as a two-level modeling problem of an emotional conversation. At the higher level, an emotional conversation is modeled as a sequence of emotional utterances, while at the lower level, each such utterance is modeled as a sequence of audiovisual observations. We

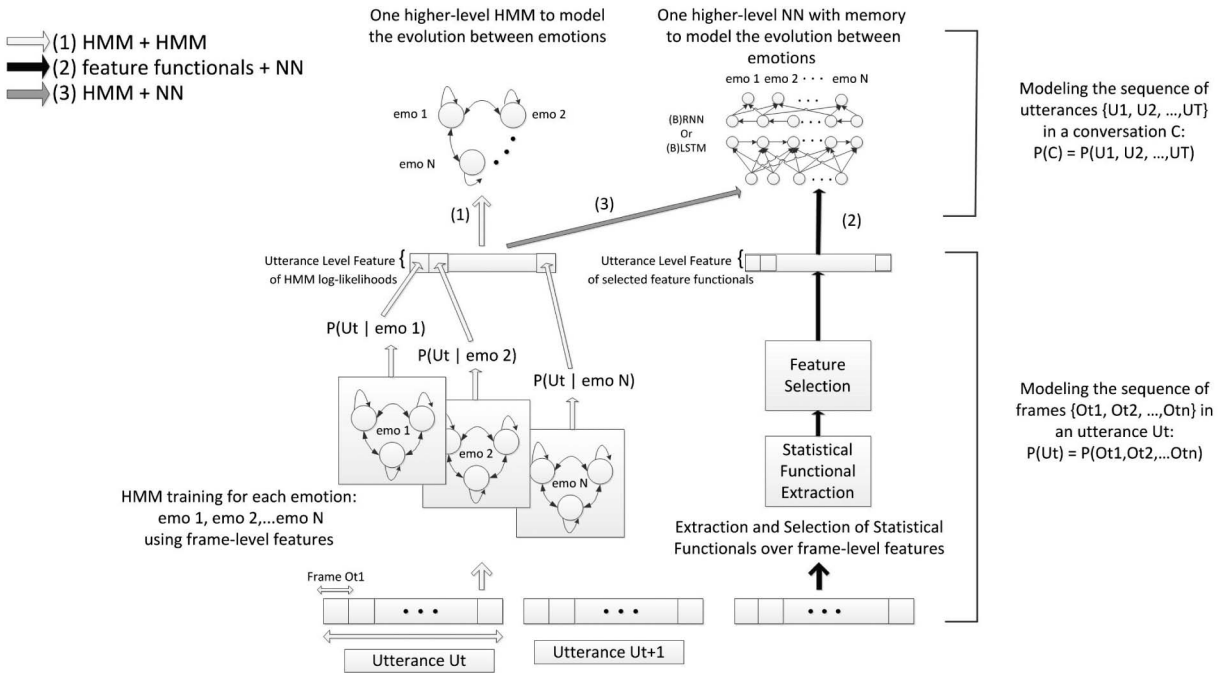


Fig. 1. A summary of our classification systems under the proposed hierarchical, context-sensitive framework. At the lower (utterance) level, modeling of emotional utterances U_t is performed through emotion-specific HMMs, as illustrated in the lower left part of the figure, or by computing statistical properties of each emotional class, as illustrated in the lower right part. At the higher level, which represents the conversation context, emotional flow between utterances of a conversation C is modeled by an HMM or a Neural Network (Unidirectional or Bidirectional RNN or BLSTM). The different combinations of the approaches at lower and higher level lead to the three systems that we describe in this work: 2-level HMM, (NNs) trained with feature functionals, and hybrid HMM/NN.

assume that an emotional utterance can be described by a single emotional label, e.g., a single level of activation, valence, or a single cluster in the valence-activation space. However, an emotional conversation may contain arbitrary emotional transitions between utterances and may consist of a variety of emotional manifestations. Therefore, utterance modeling captures the dynamics within emotional categories, while conversation modeling captures the dynamics across emotional categories.

Let us denote a conversation C as a sequence of utterances U_t, $t = 1, \ldots T$: $C = U_1, U_2, \ldots U_T$ and an utterance U_i as a sequence of low-level observations (frames) O_{tj}, $j = 1, \ldots \tau$: $U_t = O_{t1}, O_{t2}, \ldots O_{t\tau}$. In Fig. 1 we present a summary of our approaches for modeling utterance and conversation dynamics.

At the utterance level, we examine dynamic modeling by using fully connected HMMs, which captures feature statistics and underlying emotional characteristics in the audio-visual feature streams. The intuition for using fully connected HMMs is that there is no apparent left-to-right property in the dynamic evolution of the facial or vocal characteristics during emotional expression (as opposed to the evolution of phonemes during speech that is exploited in phoneme-specific left-to-right HMMs in ASR). The use of coupled instead of simple multistream HMMs enables us to model asynchrony between the audio-visual streams. Alternatively, we model the emotional utterance by estimating static, utterance-level, statistical features through the use of statistical functionals over the low-level frame sequence. Such an approach implicitly captures some of the observation dynamics while it makes fewer modeling assumptions compared to the HMM (no Markovian property, conditional independence, or synchronicity assumptions of the underlying audio-visual sequences). At the dialog level, we examine the use of HMM and discriminatively trained neural network classifiers (RNN, BLSTM). The latter make fewer assumptions on the underlying sequence of emotional utterances and may potentially capture more complex patterns of emotional flow.

Our approach of combining first and second layer HMMs for dialog and utterance modeling leads to a two-level structure, along the lines of multilevel HMMs [24] and Hierarchical HMMs [17]. Alternatively, we examine the performance of discriminatively trained neural network classifiers for conversational modeling when statistical functionals are extracted at the utterance level. In the hybrid HMM/BLSTM approach, we keep the probabilistic dynamic modeling of HMMs at the utterance level and use the emotion-specific HMM log likelihoods to form an utterance-level feature vector, which is the input of the BLSTM at the conversation level. In the next sections, we elaborate on these three approaches.

2.2 Hierarchical HMM Classifiers

The state of the art in sequence modeling, such as in ASR, utilizes the Markov chain framework to model temporal sequence context be it acoustic feature dependencies, phonetic symbol structure, local word structure, or even dialog states [20]. Here, we adopt a two-level HMM structure for modeling the sequence of audio-visual observations at both the utterance and the conversation level (context).

At the utterance level, the HMM classifiers are fully connected 3-state models, trained separately for each of N emotional categories. Thus, we have N models λ_i,

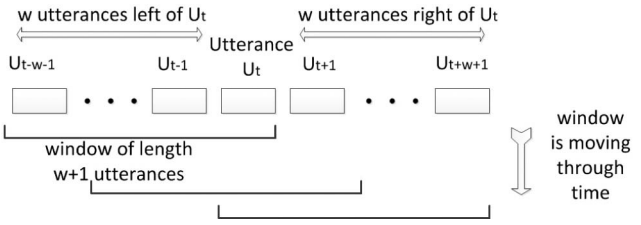


Fig. 2. Sequential Viterbi Decoding passes. Viterbi Decoding is performed in sequential subsequences (of length w + 1) of the entire utterance observation sequence. The labeling decision for utterance U_t at time t is affected by the labeling decisions of w past and w future utterances.

$i = 1, \ldots N$, denoted in Fig. 1 as emo_i. At the conversation level we have one fully connected HMM with N states modeling the N emotional categories. For each testing sequence of utterances, we estimate the most likely emotional category i for the current utterance U_t at time t by finding the HMM λ_i, with the maximum likelihood (score) $P(U_t|\lambda_i)$. These scores are utilized by the higher level HMM to represent the observation probabilities of the hidden emotional states, while the transition probabilities between emotional states $P(j|i)$ can be computed from the train set. The most probable sequence of emotional categories $Q = q_1, q_2, \ldots q_T$ for an observed conversation $C = U_1, U_2, \ldots U_T$ can be found using the Viterbi decoding (VD) algorithm over the conversation utterances. Therefore, when we make a decision about the emotional content of the current utterance, we take into account all past and future utterances of the utterance sequence.

We have also investigated the effect of a variable length of bidirectional context, which could range from one past or future utterance to all the utterances in the conversation. For this purpose, we have applied a modified Viterbi decoding over shorter sequential windows which could also be useful in real-time scenarios. Real-time systems may not be able to afford to wait for the whole conversation to end before making a decision about the emotional content of the utterances of the conversation. More specifically, we perform Viterbi decoding in overlapping windows of length $w + 1$ utterances that scan the whole sequence. The score s_t^i of each emotional class i, $i = 1 \ldots N$ in the Viterbi lattice is initialized by the likelihood $P(U_t|\lambda_i)$. Then, the sequence of scores is sequentially updated: We take into account the decisions of the previous Viterbi passes by incorporating them as weights c_t^i in the current Viterbi pass. The most probable emotion according to the previous pass gets the highest weight. An utterance U_t will be updated from $w + 1$ consecutive Viterbi passes, starting when the moving window begins at time $t - w - 1$ and ends at t, and until the moving window has reached time t as its starting point, as illustrated in Fig. 2. The score s_t^i of each emotional class i is finalized after the window has moved after time t, and through this update process w utterances left and right of U_t have been taken into account when making a decision about the label of U_t.

The algorithm is described in Box 1 and the update function works as follows:

$$s_t^i \leftarrow \log P(U_t|\lambda_i)$$

$$s_t^i \leftarrow s_t^i + c_t^i$$

$$c_t^i = \begin{cases} \log(a) & \text{if } q_t = i \text{ from the previous pass} \\ \log(b) & \text{otherwise,} \end{cases}$$

$$where\ b < a\ and\ \sum_{i=1}^{N} c_t^i = 1, t\ \in window.$$

Box 1. Sequential Viterbi Decoding (seqVD)

1: place window of length w+1 at the beginning of the sequence of utterances
2: **repeat**
3: current_utterances ← utterances[window]
4: output_sequence[window] ← VD(current_utterances)
5: **update**(current_utterances) using output_sequence [window]
6: utterances[window] ← current_utterances
7: shift window forward one utterance
8: **until** the end of all the utterances

The sequential Viterbi algorithm of Box 1 contains parameter a, which is the weight from the previous pass, and parameter w, which is the window size. In our experiments these parameters are optimized across all folds using the Nelder Mead algorithm [25]. This sequential Viterbi decoding algorithm shares some similarities with the Viterbi decoding method proposed in [26]. However, in contrast to [26], where the sequential Viterbi passes would fix the decision of the initial observation within the current window, in our approach a Viterbi pass just places a weight on an observation utterance to be used for the next pass.

The approach described in this section constitutes a second layer of computation over the utterance-level, emotion-specific HMM classifiers, and models the transitions between them. This can be viewed as a higher level HMM, where hidden states correspond to emotions and the transitions describe the emotional evolution between utterances in a conversation (temporal context).

2.3 BLSTM and RNN Architectures

Classifiers such as neural networks are able to capture a certain amount of context by using cyclic connections. These so-called recurrent neural networks can in principle map from the entire history of previous inputs to each output. Yet, the analysis of the error flow in conventional recurrent neural nets led to the finding that long-range context is inaccessible to standard RNNs since the back-propagated error either blows up or decays over time (vanishing gradient problem [27]). One effective technique to address the problem of vanishing gradients for RNN is the Long Short-Term Memory architecture [28], which is able to store information in linear memory cells over a longer period of time. LSTM networks are able to overcome the vanishing gradient problem and can learn the optimal amount of contextual information relevant for the classification task. Thus, LSTM architectures seem to be well suited for modeling context between successive utterances for enhanced emotion recognition.

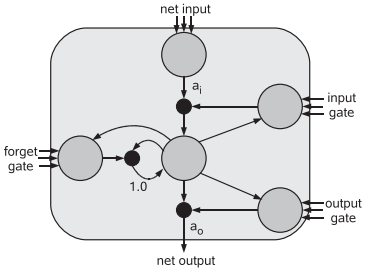


Fig. 3. LSTM memory block consisting of one memory cell: the input, output, and forget gates collect activations from inside and outside the block which control the cell through multiplicative units (depicted as small circles); input, output, and forget gates scale input, output, and internal states, respectively; a_i and a_o denote activation functions; the recurrent connection of fixed weight 1.0 maintains the internal state.

An LSTM layer is composed of recurrently connected memory blocks, each of which contains one or more memory cells, along with three multiplicative "gate" units: the input, output, and forget gates. The gates perform functions analogous to read, write, and reset operations. More specifically, the cell input is multiplied by the activation of the input gate, the cell output by that of the output gate, and the previous cell values by the forget gate (see Fig. 3). The overall effect is to allow the network to store and retrieve information over long periods of time. For example, as long as the input gate remains closed, the activation of the cell will not be overwritten by new inputs and can therefore be made available to the net much later in the sequence by opening the output gate.

Another problem with standard RNNs is that they have access to past but not to future context. This can be overcome by using bidirectional RNNs [29], where two separate recurrent hidden layers scan the input sequences in opposite directions. The two hidden layers are connected to the same output layer, which therefore has access to context information in both directions. The amount of context information that the network actually uses is learned during training and does not have to be specified beforehand. Fig. 4 shows the structure of a simple bidirectional network. Combining bidirectional networks with LSTM gives bidirectional LSTM (BLSTM) networks [15], which have been successfully used in various pattern recognition applications such as phoneme recognition [30] and emotion recognition from speech [31].

In this work, we use unidirectional and bidirectional LSTM networks trained with utterance-level features, specifically statistical functionals of audio-visual features. The LSTM networks consist of 128 memory blocks with one memory cell per block while the BLSTM networks consist of two LSTM layers with 128 memory blocks per input direction. The input layer has the same dimensionality as the feature vector and the output layer has the same dimensionality as the number of emotional classes we want to classify. For training we used the RNNLib toolbox, which is available for download at [32].

2.4 Combination of HMM and BLSTM Classifiers

We examine a combination of the HMM and BLSTM classifiers that takes advantage of both the explicit dynamic utterance modeling of the HMM framework and the ability

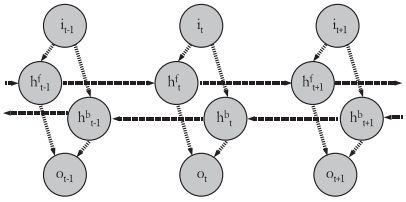


Fig. 4. Structure of a bidirectional network with input i, output o, and two hidden layers (h^f and h^b) for forward and backward processing.

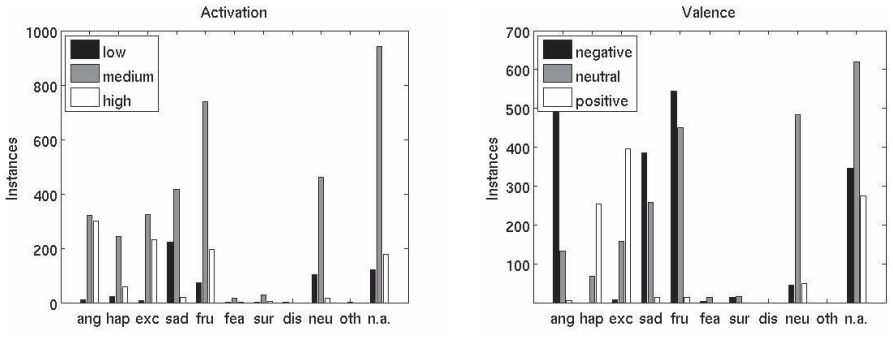
of the BLSTM to learn an arbitrarily long amount of bidirectional context. We utilize the BLSTM network as a second layer of computation over the HMM classifiers as an alternative to Viterbi decoding. This combination has the advantages of a two-layer classification structure; therefore, there is transparency as to the performance improvement we can gain from context modeling. Furthermore, the HMM+BLSTM combination may potentially capture more complex structure in the underlying emotional flow than the one captured by an HMM.

In our implementation, we collect the log-likelihoods $\log P(U_t|\lambda_i)$ for each utterance sequence U_t, $t = 1, \ldots T$, generated by the emotion-specific HMM models λ_i, $i = 1, \ldots N$, and we create an N-dimensional, utterance-level feature vector of log likelihoods. This is used as the input to the higher level BLSTM, as illustrated in method (3) of Fig. 1. Therefore, at each time t the BLSTM will have as input a feature vector containing the log likelihoods of each emotional category at that time. The BLSTMs are trained using the output log likelihoods produced on the training set utterances.

3 Database Description

The database used in this work is the Interactive Emotional Dyadic Motion Capture (IEMOCAP) database, which contains approximately 12 hours of audio-visual data from five mixed gender pairs of actors [22]. Each recorded session lasts approximately 5 minutes and consists of two actors interacting with each other in scenarios that encourage emotional expression. During each recording, both actors wore microphones and one of them had face MoCap markers. In this study, we examine the emotions expressed when the actors wear the markers, so that there is audio-visual information available.

The IEMOCAP database has been carefully designed to elicit complex emotional manifestations and to avoid caricatures by exploiting acting techniques such as improvisation. The importance of such acting techniques in collecting naturalistic, acted databases has been affirmed elsewhere in the emotion literature as a useful tool for studying emotions in controlled environments [33], [34]. Here, two acting styles were used: improvisation of scripts and improvisation of hypothetical scenarios. Each improvisation was designed to convey a general emotional theme; for example, a subject is sharing the news of her recent marriage (happiness/excitement), a subject is talking about the death of a close friend (sadness), a subject who just lost her valuable luggage at the airport learns that she will receive only a small refund (anger/frustration). The scripts are taken from theatrical plays and they are generally



(a) Activation (b) Valence

Fig. 5. Analysis of activation and valence classes in terms of categorical labels for all utterances of the database: anger (ang), happiness (hap), excitement (exc), sadness (sad), frustration (fru), fear (fea), surprise (sur), disgust (dis), neutral (neu), other (oth), and no agreement (n.a.). We notice that the categorical tags are generally consistent with the activation and valence tags.

characterized by a more complex emotional flow. The recordings, even those with a general theme, contain multiple emotional displays which vary in their intensity and clarity. The emotional content of each utterance is not predefined and it generally depends on the interpretation of the script/improvisation by the actors and the course of their interaction. The goal was to elicit emotional displays that resemble natural emotional expression and are generated through a suitable context.

The dyadic sessions were later manually segmented into utterances, where consecutive utterances of a speaker may or may not belong to the same turn. We examine the sequence of utterances of a certain speaker during a recording, and we make no distinction between utterances that are separated by one or more utterances of the other speaker and utterances that belong to the same turn. The emotional content of each utterance was annotated by human annotators in categorical labels (annotators had to choose between the following emotions: angry, happy, excited, sad, frustrated, fearful, surprised, disgusted, neutral, and "other-please specify") and in dimensional descriptions of valence and activation. Valence describes how positive versus negative and activation how calm versus excited is the expressed emotion. Value 1 denotes very low activation and very negative valence and 5 denotes very high activation and very positive valence. Those properties are rated on scales 1-5 and are averaged across two annotators (or, for some few utterances, across three annotators).

The dimensional tags seem to provide more general descriptions of an emotional expression. For example, a large part of the utterances of the database (approximately 17 percent) do not have a categorical label with a majority-vote agreement of the evaluators. These utterances may be displaying subtle or ambiguous emotions that are hard to classify with a single categorical label. However, annotators seem to have an acceptable level of agreement in the dimensional labels of such utterances. An analysis of the IEMOCAP dimensional labels shows that for the data labeled by two evaluators, in 94 percent of the utterances evaluators agreed or were one point apart in their rating of valence, and in 85 percent of the utterances evaluators agreed or were one point apart in their rating of activation.

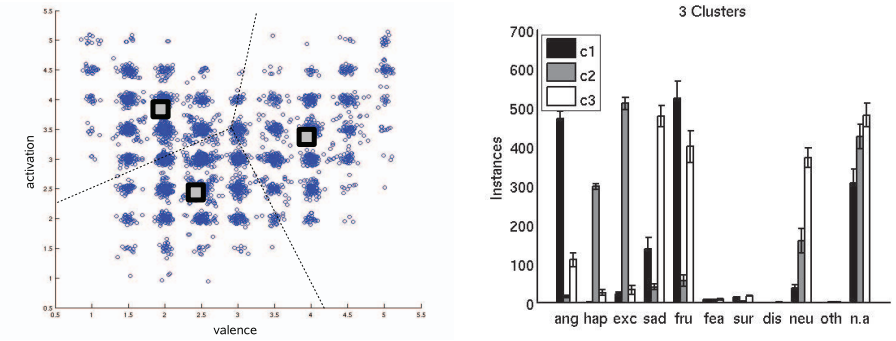
Moreover, the use of dimensional labeling seems particularly suitable for analyzing temporal emotional context, which is the main focus of this paper, since it enables us to have a label for every utterance in the sequence of consecutive utterances that make up a conversation. Here, this label is derived by averaging the decisions of the annotators. Considering categorical labels would introduce "gaps" in our observation label sequence for utterances with no evaluator agreement or utterances that are labeled with rarely occurring emotions, like fear and disgust. Training emotional models for rare emotions would not be possible because of lack of data, while the treatment of a "no agreement" class does not seem straightforward. Although both categorical and dimensional representations have their merits, in the context of this work we focus our analysis on the dimensional representation, and we use the categorical labels to validate and interpret our results, as explained in the following sections.

4 Emotions and Emotion Transitions

4.1 Valence and Activation

The first emotion classification task that we consider in this work consists of the classification of three levels of valence and activation: Level 1 contains ratings in the range [1, 2], level 2 contains ratings in the range (2, 4), and level 3 contains ratings in the range [4, 5]. These levels intuitively correspond to low, medium, and high activation, and to negative, neutral, and positive valence, respectively. The class sizes are not balanced since medium values of labels are more common than extreme values. The choice of three levels instead of five, which was the initial resolution of the valence and activation ratings, was made to ensure that a sufficient amount of data are available in each class for emotional model training. For example, for activation 2 percent of the averaged annotator labels (87 utterances) has a rating of less or equal to 1.5, which corresponds to the cases of very low activation. When we used the 3-level scale, the low activation instances are 11 percent of the data or a total of 557 utterances.

The emotional information provided by the dimensional tags is still related to categorical emotions in a meaningful way. To examine this, we analyzed the available categorical tags of the utterances that belong to each of the dimensional classes. In Fig. 5 we show how the utterances of each class break down into categorical tags. Specifically the categorical tags that we are considering in the IEMOCAP corpus are:



(a) 3 clusters in the valence-activation space (fold1)

(b) 3 clusters

Fig. 6. Analysis of classes in the three cluster task in terms of categorical labels. The bars and the error bars correspond to the mean and standard deviation computed across the 10 folds. We notice that the data-driven clusters tend to contain different categorical emotional manifestations according to their position in the emotional space. Specifically, clusters 1, 2, and 3 roughly contain categorical emotions of "anger or frustration," "happiness or excitement," and "neutrality or sadness or frustration," respectively.

Angry, Happy, Excited, Sad, Frustrated, Fearful, Surprised, Disgusted, Neutral, Other, and No Agreement (n.a.). The annotators of the categorical and dimensional tags of an utterance are usually not the same. Specifically, in Fig. 5a we show how many utterances from each activation class fall into each categorical emotional tag. Similarly, we construct the bar graph for valence, which is presented in Fig. 5b.

For both valence and activation assessment, the categorical labels generally agree with the dimensional tags according to what is known in the emotion literature about the position of categorical emotions in the valence-activation space [35]. Overall, the resulting bar graphs are intuitive. For example, in the valence plot in Fig. 5, we notice that utterances that are annotated as having negative valence are also generally perceived to express "negative" emotions such as anger, sadness, and frustration, while utterances with positive valence are generally perceived to express "positive" emotions, such as happiness and excitement. An interesting observation can be made regarding frustration (an emotion hard to classify since it could resemble anger, sadness, or neutrality) where we can see that the valence assignment for frustration observations is almost equally divided between negative and neutral.

4.2 Clusters in the Emotion Space

We also examine the joint classification of the emotional dimensions by building three and four clusters in the valence-activation emotional space. The motivation for clustering the valence-activation space is to build classifiers that provide richer and more complete emotional information by combining valence and activation information. We apply data-driven clustering through K-means to automatically select clusters that fit the distribution of the emotional manifestations of our database in the emotional space (similar approaches are also followed in [23], [9]). The ground truth of every utterance is assigned to one of the clusters using the minimum euclidean distance between its annotation and the cluster midpoints.

When abstracting our emotion classes into clusters of the valence-activation space, we also study their relation to the corresponding categorical annotations and investigate which categorical emotional manifestations tend to fall into each cluster. Specifically, we examine how each cluster breaks down in terms of categorical labels. For example, in Figs. 6a and 6b we illustrate the three clusters in the emotional space and the histogram of the categorical emotional tags of the utterances they contain. Note that the utterances that belong to each cluster depend on the training set of each fold. Thus, the bar graphs in Fig. 6b represent the mean over the 10 folds of our experiment (see also the experimental setup in Section 6.1) and the error bars represent the standard deviation over the 10 folds. The plot of Fig. 6a corresponds to the first fold of our experiment, but the differences across folds are relatively small (the average standard deviation of the cluster centroid coordinates across the 10 folds is as low as 0.05).

Looking at Fig. 6b we notice that cluster c1 contains large portions of utterances tagged as angry or frustrated, cluster c2 contains utterances tagged as happy or excited, and cluster c3 utterances tagged as sad or neutral or frustrated. Therefore, we could think of the three clusters as roughly containing emotions of "anger/frustration," "happiness/excitement," and "sadness/neutrality/frustration." This agrees with Fig. 6a, where the positions of the three clusters are in areas of the valence-activation space that are generally expected to contain emotional manifestations of "anger," "happiness," and "sadness or neutrality." Similar observations can be made for the four cluster classification task, where, by examining the corresponding bar graph of categorical tags, we notice that the four clusters roughly contain emotions of "happiness/excitement," "sadness/frustration," "anger/frustration," and "neutrality" (the plot and bar graph are omitted for lack of space).

4.3 Emotional Grammars

Analysis of the emotional dialogs of our database reveals that certain emotional transitions are more probable than others, indicating that the underlying emotional flow follows certain typical patterns. In the two-level HMM approach, we have assumed that the underlying emotional states form a Markov chain and applied an HMM at the conversation level to model emotion dynamics. Equivalently, we could view this modeling as a Probabilistic Regular Grammar (PRG) describing emotional transitions,

TABLE 1
Emotional Transition Bigrams for the Valence, Activation, 3, and 4 Cluster Classification Tasks

valence				activation				3clusters				4clusters				
	Neg	Neu	Pos		Low	Med	High		c1	c2	c3		c1	c2	c3	c4
Neg	0.72	0.27	0.01	Low	0.34	0.63	0.03	c1	0.66	0.06	0.28	c1	0.71	0.06	0.03	0.20
Neu	0.23	0.65	0.12	Med	0.09	0.76	0.15	c2	0.06	0.78	0.16	c2	0.04	0.56	0.21	0.18
Pos	0.02	0.26	0.72	High	0.04	0.48	0.48	c3	0.24	0.12	0.64	c3	0.03	0.29	0.58	0.09
												c4	0.19	0.25	0.08	0.49

For the three cluster case, the most frequent categorical emotions per cluster are : $c1 =$ *"ang/fru,"* $c2 =$ *"hap/exc,"* $c3 =$ *"neu/sad." For the four cluster case the most frequent emotions per cluster are:* $c1 =$ *"hap/exc,"* $c2 =$ *"sad/fru,"* $c3 =$ *"ang/fru" and* $c4 =$ *"neu."*

using the equivalence between PRGs and HMMs [36]. We can define a PRG with initial state S^1 as

$$S^i \rightarrow w^j S^k \; or \; S^i \rightarrow w^j,$$

where S^i represents an emotional state, w^j represents an emotional observation, and the right arrow represents an emotional transition. This modeling assumes that given an internal emotional state S^i, a person emits an audiovisual emotional expression w^j, e.g., tone of voice and facial expression, and transitions to an internal emotional state S^k. In this work, we only compute bigram emotional utterance probabilities which give us a rough description of the emotional flow between utterances. We utilize these emotional grammars to gain insights about typical evolution patterns in emotional expression and apply this knowledge to inform emotion recognition tasks. This approach could be extended to higher order transition models, which would capture more detailed emotional patterns. Alternatively, we also consider models that make fewer assumptions on the underlying sequence and potentially have richer representation power, such as neural networks (here, BLSTM).

In Table 1 we present the transition probabilities for the valence, activation, 3, and 4 cluster HMMs. The transition probabilities have been approximated by counting all transitions between consecutive utterances of the database. For the cluster tasks where the cluster tags of each utterance may change according to the fold we present the average transition probabilities across all folds. To test the statistical significance of the transitions of Table 1, we performed lag sequential analysis [37]. For all classification tasks and all emotional transitions we find that the observed transitions are statistically significantly different from the expected transitions if emotion at utterance t was independent of the previous emotion at utterance $t-1$ (p-value < 0.001, [37, Chapter 7.4]). Similar conclusions are made when we compute table-wise statistics as described in [37, Chapter 7.6], specifically Pearson and likelihood-ratio chi-square values, which support the hypothesis that the rows and columns of each transition matrix are dependent (p-value < 0.001). These statistical tests were performed using the GSEQ5 Toolbox [38].

The valence HMM states contain large diagonal self-transition probabilities, suggesting that interlocutors tend to preserve their valence states locally over time. In contrast, low and high activation states tend to be mostly isolated phenomena since interlocutors tend to transition to the medium activation state and preserve that state. This indicates that emotional states of negative, neutral, or positive valence tend to last longer compared to high or low-activated emotional states, which seem to be transient. Indeed, out of all the low-activated emotional manifestations of our database, 96 percent last at most three consecutive utterances, and out of all the high-activated emotional manifestations of our database, 90 percent last at most three consecutive utterances. The corresponding proportions for the manifestations of negative, positive, and neutral valence are around 75 percent, which means that a significant proportion of such emotions spans over more than three utterances.

For the 3 cluster bigrams, frequent transitions happen between the "anger/frustration" and the "neutrality/sadness" clusters, as well as the "happiness/excitement" and the "neutrality/sadness" clusters, while transitions between "anger/frustration" and "happiness/excitement" clusters are very rare. This indicates that interlocutors generally transition between a neutral state and an emotional state (of positive or negative valence) but not directly between the extreme valence states. For the 4 cluster HMM, we notice that the "happy/excited" cluster is the one with the highest self-transition probability. Frequent transitions happen between "anger/frustration" and "sadness/frustration," and between "neutrality" and most other states. These suggest that interlocutors preserve their positive or negative valence while changing their activation levels, i.e., transitioning between sadness, frustration, and anger. Neutrality appears to be an intermediate state when transitioning between emotions of opposite valence.

The above observations concerning emotional transitions depend on the structure of our database. Even though the database design may not cover the full range of human emotional interactions, one could argue that the conclusions presented here could prove useful for processing human-machine interactions where the variety and complexity of emotions and emotional transitions are often limited compared to the general possibilities in interpersonal human interactions.

5 Feature Extraction and Fusion

5.1 Audio-Visual Frame-Level Feature Extraction

The IEMOCAP data contain detailed MoCap facial marker coordinates. The positions of the facial markers can be seen in Fig. 7. The markers were normalized for head rotation and translation and the nose marker tip is defined as the local coordinate center of each frame. In total, information from 46 facial markers is used, namely, their (x, y, z) coordinates. This results in a 138-dimensional facial representation, which tends to be redundant because it

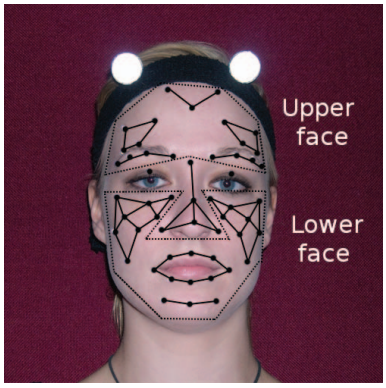


Fig. 7. Positions of the MOCAP face markers and separation of the face into lower and upper facial regions.

does not exploit the correlations of neighboring marker movements and the structure of the human face.

In order to obtain a lower dimensional representation of the facial marker information, we use Principal Feature Analysis (PFA) [39]. This method performs Principal Component Analysis (PCA) as a first step and selects features so as to minimize the correlations between them. In contrast to PCA, PFA selects features in the original feature space (here marker coordinates) instead of linear combinations of features, which makes selection results interpretable. We select 30 features since the PCA transformation explains more than 95 percent of the total variability. To these we append the first derivatives, resulting in a 60-dimensional representation. The facial features are normalized per speaker to smooth out individual facial characteristics that are unrelated to emotional expressions. Our speaker normalization approach consists of finding a mapping from the individual average face to the general average face. This is achieved by shifting the mean value of each marker coordinate of each subject to the mean value of that marker coordinate across all subjects. The feature selection and normalization framework is described in our previous work [40].

In addition, we extract a variety of features from the speech waveform: 12 MFCC coefficients, 27 Mel Frequency Band coefficients (MFB), pitch, and energy values. We also compute their first derivatives. All the audio features are computed using the Praat software [41] and are normalized using z-standardization. The audio and visual features are extracted at the same frame rate of 25 ms, with a 50 ms window. The utterance-level audio HMMs were trained using the 27 MFBs, pitch and energy along with their first derivatives, while the visual HMMs were trained using the 30 PFA features with their first derivatives. For the audio-visual HMMs and coupled-HMMs we used both these voice and face features, fused at feature or at model-level.

5.2 Utterance-Level Statistics of Audio-Visual Features

We use a set of 23 utterance-level statistical functionals that are computed from the low-level acoustic and visual features (see Table 2). Thus, we obtain $142 \times 23 = 3{,}266$ utterance-level features. All functionals were calculated using the openSMILE toolkit [42].

In order to reduce the size of the resulting feature space, we conduct a cyclic Correlation-based Feature Subset Selection (CFS) using the training set of each fold. The main idea of CFS is that useful feature subsets should contain features that are highly correlated with the target class while being uncorrelated with each other [43], [44]. Note that we deliberately decided for a filter-based feature selection method since a wrapper-based technique would have biased the resulting feature set with respect to compatibility to a specific classifier.

Applying CFS to the 3,266D feature space results in an automatic selection of between 66 and 224 features, depending on the classification task and the fold. For the valence classification task, on average $84 \pm 1.1\%$ of the selected features are facial features, whereas for classification of the degree of activation, only $44 \pm 1.8\%$ of the features selected via CFS are facial features. This underscores the fact that visual features tend to be well suited for determining valence while acoustic features reveal the degree of activation and agrees with the unimodal classification results that are presented in the results section. For a detailed analysis of the selected features see Table 3.

TABLE 2
Statistical Functionals Used for Turnwise Processing

group	functionals
extremes	position of maximum, position of minimum
regression	linear regression coefficients 1 and 2, quadratic mean of linear regression error, quadratic regression coefficients 1, 2, and 3, quadratic mean of quadratic regression error
means	arithmetic mean
percentiles	quartiles 1, 2, and 3, interquartile ranges 1-2, 2-3, and 1-3, 1 %-percentile, 99 %-percentile, percentile range
others	number of non-zero values, standard deviation, skewness, kurtosis

5.3 Audio-Visual Feature Fusion

For the utterance-level HMM approaches where frame-level features are used, we apply multistream HMM classifiers (here denoted simply as HMMs). These assign different importance weights to the audio and visual modalities and assume synchronicity between them. When modeling the dynamics of high-level attributes such as emotional descriptors of a whole utterance, allowing asynchrony in the dynamic evolution of the underlying audio-visual cues could be beneficial. We also apply model-level fusion through the use of coupled Hidden Markov Models (c-HMMs), which allow this type of asynchrony, and have been widely used in the literature [45], [46]. All models are

TABLE 3
Distribution of the Features Selected via CFS for the Classification of Valence (VAL) and Activation (ACT) as well as for the Discrimination of 3 and 4 Clusters in Emotional Space (see Section 4.2)

feature group	VAL	ACT	3 clusters	4 clusters
pitch	5 %	4 %	3 %	4 %
energy	0 %	1 %	1 %	1 %
MFCC	4 %	21 %	11 %	11 %
MFB	7 %	30 %	18 %	19 %
lower face (Fig.7)	63 %	32 %	50 %	49 %
upper face (Fig.7)	21 %	12 %	17 %	16 %

trained using the HTK Toolkit [21]. HTK offers the functionality for defining and training a multistream HMM but it does not explicitly allow for coupling of multiple single stream HMMs. However, following the analysis presented in [46], [47], we can implement c-HMMs in HTK using a product HMM structure.

6 Experiments and Results

6.1 Experimental Setup

Our experiments are organized in a cyclic leave-one-speaker-out cross validation. The feature extraction PCA transformations (for the face features) and the feature z-normalization constants are computed based on the respective training set of each fold. The mean and standard deviation of the number of test and training utterances across the folds is 498 ± 60 and $4{,}475 \pm 61$, respectively. For each fold, we compute the F1-measure, which is the harmonic mean of unweighted precision and recall, as our primary performance measure. As a secondary measure we also report unweighted recall (unweighted accuracy). The presented recognition results are the subject-independent averages over the 10 folds and the corresponding standard deviation.

In the next sections, we present the results of the context-sensitive neural network and HMM frameworks for the various classification tasks. We trained 3-state ergodic HMMs and c-HMMs with observation probability distributions modeled by Gaussian mixture models. The stream weights and the number of mixtures per state (varying from 4 to 32 mixtures) have been experimentally optimized on a validation set. For the context-sensitive HMM approaches, the bigrams and the BLSTMs of each fold have been computed or trained on the corresponding training set. The LSTM networks consist of 128 memory blocks with one memory cell per block, while the BLSTM networks consist of two LSTM layers with 128 memory blocks per input direction. To improve generalization, we add zero mean Gaussian noise with standard deviation 0.6 to the input statistical features during training. The BLSTM networks that are trained to process the HMM outputs (HMM+BLSTM) consist of 32 memory blocks per input direction.

6.2 Context-Free versus Context-Sensitive Classifiers

In this section, we compare the classification performance between emotion-specific HMMs or c-HMMs that do not make use of context information and our proposed context-sensitive frameworks: hierarchical HMMs (or c-HMMs) using sequential Viterbi Decoding with bidirectional window of $w+1$ utterances (HMM+HMM, c-HMM+HMM), BLSTM trained with utterance-level feature functionals (BLSTM) and hybrid HMM/BLSTM classifiers (HMM+BLSTM, c-HMM+BLSTM). Table 4 shows the performances for discriminating three levels of valence and activation, as well as classification into three and four clusters in the valence-activation space. To test the statistical significance of the differences in average F1 performance between (c-)HMM, (c-)HMM+HMM, (c-)HMM+BLSTM, and BLSTM classifiers, we conducted repeated measures ANOVA at the subject level with Bonferonni adjustment for post-hoc tests, using SPSS [48].

TABLE 4
Comparing Context-Free and Context-Sensitive Classifiers for Discriminating Three Levels of Valence and Activation, and Three and Four Clusters in the Valence-Activation Space, Using Face (f) and Voice (v) Features: Mean and Standard Deviation of F1-Measure and Unweighted Accuracy across the 10 Folds (10 Speakers)

classifier	features	F1	Acc.(uw)
valence			
HMM	v	49.85 ± 3.18	49.99 ± 3.63
HMM	f	58.85 ± 3.86	60.98 ± 4.96
HMM	v+f	60.79 ± 2.53	62.50 ± 3.39
cHMM	v+f	60.42 ± 3.59	61.75 ± 4.66
HMM+HMM(w=2)	v+f	62.02 ± 2.25	63.16 ± 3.18
HMM+BLSTM	v+f	63.97 ± 3.03	62.78 ± 6.43
BLSTM	v+f	65.12 ± 5.13	64.67 ± 6.48
activation			
HMM	v	57.54 ± 3.33	61.92 ± 4.88
HMM	f	49.04 ± 4.40	51.36 ± 4.14
HMM	v+f	57.56 ± 4.27	60.00 ± 4.45
cHMM	v+f	57.39 ± 3.25	61.29 ± 5.16
HMM+HMM(w=4)	v+f	57.71 ± 4.23	60.02 ± 4.54
HMM+BLSTM	v+f	53.41 ± 5.99	46.93 ± 5.69
BLSTM	v+f	54.90 ± 5.02	52.28 ± 5.37
3 clusters			
HMM	v+f	67.33 ± 5.15	66.18 ± 6.69
cHMM	v+f	68.45 ± 3.38	67.95 ± 3.18
cHMM+HMM(w=4)	v+f	70.36 ± 3.48	69.76 ± 3.09
cHMM+BLSTM	v+f	68.09 ± 4.16	68.02 ± 4.72
BLSTM	v+f	72.35 ± 5.10	71.83 ± 5.46
4 clusters			
HMM	v+f	56.54 ± 4.29	56.64 ± 5.90
cHMM	v+f	57.28 ± 3.65	57.87 ± 4.33
cHMM+HMM(w=4)	v+f	58.65 ± 3.80	58.89 ± 4.59
cHMM+BLSTM	v+f	58.21 ± 5.24	57.94 ± 5.89
BLSTM	v+f	62.80 ± 6.69	61.96 ± 7.02

Concerning the valence results, we observe that facial features are much more effective in classifying valence than voice features, which agrees with our previous findings [16]. Regarding the audio-visual classifiers, BLSTM achieves the highest average F1 measure, while the emotion-specific HMMs benefit from the use of long-range context, either through higher level Viterbi Decoding (HMM+HMM) or through the use of a higher level BLSTM (HMM+BLSTM). Statistical significance tests reveal that the average HMM+BLSTM F1 measure is significantly higher than that of the HMM at the 0.05 level. HMM+HMM performance was not found significantly higher than HMM performance, but it has a p value very close to the threshold ($p = 0.055$). Similarly, the comparison of BLSTM and HMM gives a p value of 0.06.

For the activation task, incorporating visual cues does not improve performance significantly, indicating that audio cues are more informative that visual cues. This agrees with previous results in the emotion literature [49], [50]. Overall, we notice that taking temporal context into account does not benefit activation classification performance for HMMs (HMM+HMM), and the BLSTM and HMM/BLSTM classifiers on average perform worse than the context-free HMMs. This could be attributed to the isolated nature of the extreme activation instances, as we have observed in Section 4.3. High and low activation events are isolated between repeated instances of medium

TABLE 5
Confusion Matrices of the HMM, the Hierarchical HMM, the Hybrid HMM/BLSTM and the BLSTM Classifiers for the Activation and Three Cluster Classification Tasks

activation

HMM, F1 = 57.56

	Low	Med	High
Low	**53.32**	41.47	5.21
Med	14.66	**62.61**	22.73
High	0.93	30.93	**68.14**

HMM+HMM, F1 = 57.71

	Low	Med	High
Low	**53.32**	41.47	05.21
Med	14.40	**62.96**	22.64
High	0.93	31.13	**67.94**

HMM+BLSTM, F1 = 53.41

	Low	Med	High
Low	**7.18**	92.46	0.36
Med	1.28	**92.31**	6.41
High	0	58.66	**41.34**

BLSTM, F1 = 54.90

	Low	Med	High
Low	**28.42**	69.27	2.31
Med	6.75	**78.65**	14.60
High	0.51	49.13	**50.36**

3clusters

cHMM, F1 = 68.45

	c1	c2	c3
c1	**70.78**	13.96	15.26
c2	16.39	**67.56**	16.05
c3	22.83	11.86	**65.31**

cHMM+HMM, F1 = 70.36

	c1	c2	c3
c1	**71.82**	13.24	14.94
c2	15.98	**69.15**	14.88
c3	22.83	10.55	**66.61**

cHMM+BLSTM, F1 = 68.09

	c1	c2	c3
c1	**70.12**	10.70	19.18
c2	18.60	**65.22**	16.18
c3	21.11	9.82	**69.07**

BLSTM, F1 = 72.35

	c1	c2	c3
c1	**72.20**	8.73	19.07
c2	16.77	**67.75**	15.48
c3	14.72	9.69	**75.59**

For the three cluster case, the correspondence between emotions and clusters is: $c1 =$ *"ang/fru,"* $c2 =$ *"hap/exc,"* $c3 =$ *"neu/sad."*

activation; therefore the use of context-sensitive methods like (VD) or BLSTM tends to underestimate their probability of occurrence and ends in misclassifying them as medium activation events (oversmoothing). Especially, the long-range context modeling of the BLSTM does not seem to capture well the activation evolution and degrades activation classification. For example, when looking at the confusion matrices of Table 5, we notice that when using the hybrid HMM/BLSTM or BLSTM, the performance of the high and low activation classes greatly decreases compared to the simple HMM.

For the three cluster task, we notice that context-sensitive classifiers, such as c-HMM+HMM and BLSTM, on average perform higher than the simple HMMs and c-HMMs, and that the BLSTM classifier achieves the highest average F1 measure. The average F1 measure of cHMM+HMM was found significantly higher that of c-HMM at the 0.05 level. Similarly, for the four cluster task context-sensitive classifiers tend to outperform simple HMMs and cHMMs in terms of average F1, although these differences are not statistically significant at the 0.05 level.

In Table 5 we present the confusion matrices of the (c-)HMM, the hierarchical (c-)HMM, the hybrid (c-)HMM/BLSTM, and the BLSTM classifiers for the three cluster and activation classification tasks in order to give a description of the confusion between classes for a task where context is beneficial (three cluster) versus a task where context is not beneficial (activation).

Overall, our results suggest that incorporating context is beneficial for the valence and the three and four cluster classification. The BLSTM classifier generally achieves the highest classification performance, although performance across folds has a relatively high variance. The hierarchical HMM and hybrid HMM/BLSTM classifiers perform similarly in general, and lower than the BLSTM in terms of average F1 measure, although they tend to have more consistent performance across subjects (smaller variance). Regarding the hierarchical HMM approach, we notice that a small amount of bidirectional context (e.g., $w = 4$) can give a performance increase. We have omitted the results of the HMM+HMM architecture where Viterbi Decoding is used over the total observation sequence. For all our classification tasks, the results are very similar to the ones obtained through sequential VD with small window sizes, which suggests that it is possible to increase recognition performance even when a small amount of bidirectional context is used. These observations are encouraging and suggest that this algorithm could be applied in practical scenarios where an emotion recognition system might not be able to afford to wait for the conversation to end in order to perform recognition, while it might be acceptable to wait a few utterances before making a decision.

Note that there are significant variations in the performance and rankings across different folds for all classification approaches, as indicated by the variances of Table 4 and the results of the statistical significance tests. These suggest that no approach is clearly superior for all speakers. Our insight is that these variations could result from speaker-dependent characteristics of emotional expression, i.e., some speakers may be more overtly expressive than others or may make different expressive use of the audio and visual modalities.

To the best of our knowledge, there are no published works that report classification results of dimensional labels using the IEMOCAP database in a way that would allow direct comparison with our results. The most relevant past works are [9] and [51]; however for both cases our experimental setup is more generic. In [9], the authors perform speech-based classification of dimensional labels; however, the train and test sets are randomly split into 15 cross validations. Our subject-independent setting is considerably more challenging for classification. In [51], the authors perform speech-based classification of valence and activation only for utterances with categorical labels of angry, happy, neutral, and sad. Furthermore, the authors remove utterances that seem to have conflicting categorical and dimensional labels, which simplifies the problem as it removes potentially ambiguous emotional manifestations.

6.3 Context-Sensitive Neural Network Classifiers

In this section, we compare the recognition performances of various Neural Network classifiers which take into account different amount of unidirectional and bidirectional context. The results are presented in Table 6. (B)LSTM architectures achieve a higher average F1 measure compared to (B)RNN architectures, which indicates the merit of learning a longer range of temporal context for emotion recognition tasks. Also, bidirectional neural networks, such as BLSTMs and BRNNs, outperform their respective unidirectional counterparts, such as LSTM and RNN,

TABLE 6
Comparing Context-Sensitive Neural Network Classifiers for Discriminating Three Levels of Valence and Activation, and 3 and 4 Clusters in Valence-Activation Space Using Face (f) and Voice (v) Features: Mean and Standard Deviation of F1-Measure and Unweighted Accuracy across the 10 Folds (10 Speakers)

classifier	features	F1	Acc.(uw)
		valence	
RNN	v+f	63.34 ± 4.58	62.92 ± 6.00
BRNN	v+f	64.10 ± 5.05	63.68 ± 6.64
LSTM	v+f	63.71 ± 4.86	63.76 ± 5.95
BLSTM	v+f	65.12 ± 5.13	64.67 ± 6.48
		activation	
RNN	v+f	52.78 ± 5.21	48.54 ± 5.59
BRNN	v+f	53.93 ± 4.12	49.98 ± 4.62
LSTM	v+f	53.65 ± 4.97	50.35 ± 5.83
BLSTM	v+f	54.90 ± 5.02	52.28 ± 5.37
		3 clusters	
RNN	v+f	69.59 ± 5.75	69.34 ± 5.95
BRNN	v+f	69.94 ± 5.65	69.76 ± 6.00
LSTM	v+f	70.34 ± 5.85	69.53 ± 6.61
BLSTM	v+f	72.35 ± 5.10	71.83 ± 5.46
		4 clusters	
RNN	v+f	58.30 ± 6.63	57.29 ± 7.28
BRNN	v+f	60.10 ± 5.96	59.14 ± 6.72
LSTM	v+f	61.93 ± 5.96	61.02 ± 6.15
BLSTM	v+f	62.80 ± 6.69	61.96 ± 7.02

which suggests the importance of bidirectional context for these architectures. The performance differences between these context-sensitive NNs, although not statistically significant, indicate a consistent trend in performance across all classification tasks, with BLSTM being the highest performing classifier.

6.4 Context Learning of the BLSTM Architectures

To investigate the importance of presenting training and test utterances in the right order during BLSTM network training and decoding, we repeated all BLSTM classification experiments using randomly shuffled data, i.e., we processed the utterances of a given conversation in arbitrary order so that the network is not able to make use of meaningful context information. As can be seen in Table 7, this downgrades recognition performance. To test the statistical significance of this result, we performed paired t-tests and found that the differences in average F1 measures are statistically significant at the 0.05 level for all classification tasks, except for the case of activation. These results suggest that the high performance of the BLSTM classifiers is to a large extent due to their ability to effectively learn an adequate amount of relevant emotional context from past and future observations. They can also be interpreted as further evidence that learning to incorporate temporal context information is important for human emotion modeling.

7 Conclusion and Future Work

In this work, we have described and analyzed context-sensitive frameworks for emotion recognition, i.e., frameworks that take into account temporal emotional context when making a decision about the emotion of an utterance. These methods, which utilize powerful and popular

TABLE 7
Recognition Performances (Percent) of BLSTM Networks When Training on the Original Sequence of Utterances Compared to When the Utterances Are Randomly Shuffled: Mean and Standard Deviation of F1-Measure and Unweighted Accuracy across the 10 Folds (10 Speakers)

classifier	features	F1	Acc.(uw)
		valence	
BLSTM	v+f	65.12 ± 5.13	64.67 ± 6.48
BLSTM(shuffled)	v+f	59.71 ± 4.51	58.98 ± 5.14
		activation	
BLSTM	v+f	54.90 ± 5.02	52.28 ± 5.37
BLSTM(shuffled)	v+f	52.10 ± 6.86	46.35 ± 6.78
		3 clusters	
BLSTM	v+f	72.35 ± 5.10	71.83 ± 5.46
BLSTM(shuffled)	v+f	67.86 ± 5.08	66.61 ± 4.95
		4 clusters	
BLSTM	v+f	62.80 ± 6.69	61.96 ± 7.02
BLSTM(shuffled)	v+f	59.27 ± 6.40	57.93 ± 6.88

classifiers such as HMMs and BLSTMs, could be viewed under the common framework of a hierarchical, multimodal approach, which models the observation flow both at the utterance level (within an emotion) and at the conversation level (between emotions). The different classifiers that can be chosen for each level reflect different modeling assumptions on the underlying sequences and account for different system requirements. Our emotion classification experiments indicate that taking into account temporal context tends to improve emotion classification performance. Overall, context-sensitive approaches outperform methods that do not consider context for the recognition of valence states and emotional clusters in the valence-activation space, in terms of average F1 measure. However, the relatively large performance variability between subjects suggests that no method is clearly superior for all subjects. Additionally, the use of context from both past and future seems beneficial, as suggested by the slightly higher performance of bidirectional neural networks (BLSTM, BRNN) compared to their unidirectional counterparts. Even the use of a small amount of context around the current observation, e.g., from the use of the sequential VD algorithm with a small window of $w+1$ utterances, leads to performance improvement, which is an encouraging result for designing context sensitive frameworks with performance close to real time. The only emotion classification task that does not benefit significantly from context is activation, possibly because of the isolated nature of the extreme activation events, which makes this structure difficult to model.

According to our results, neural network architectures and, specifically, (B)LSTM networks trained with utterance level feature functionals achieve a higher average performance than HMM classification schemes. This could be attributed to their discriminative training, fewer modeling assumptions, and their ability to capture long-range, bidirectional temporal patterns of the input feature streams and output activations. BLSTM networks can learn an adequate amount of relevant emotional context around the current observation, during the training stage. When such context is not present, for example, when we randomly shuffle the utterances of a conversation, the performance of

the BLSTM classifiers significantly decreases. However, (B)LSTM and (B)RNN classifiers seem to have difficulties handling emotional expression variability between subjects; therefore their performance may vary significantly across people. HMM classification frameworks and hybrid HMM/BLSTM frameworks on average perform lower than neural networks, but generally achieve more consistent classification results across subjects. They provide a structured approach for modeling and classifying sequences at multiple levels, they have more trasparency as to what amount of context is used, and they are generally flexible. For example, HMM+HMM classifiers can be modified to use limited context so as to suit possible requirements of real-time emotion recognition systems.

Analysis of the emotional flow in the conversations of our database indicates an underlying structure in typical emotional evolution. For example, valence states seem to last longer than activation states of high or low activation, which are more transient. Also, some emotional transitions are more frequent than others, i.e., a transition between neutrality and an emotional state is much more likely than a transition between two emotional states of opposing valence (happy from/to angry). Simple first order transition probabilities provide a rough description of the emotional flow and lead to modeling emotional utterances in a conversation using an HMM. This can be seen equivalently as a PRG for emotional utterances. One could perform more complex modeling of a conversation and look for equivalent grammars with more representation power than a PRG. Although such conclusions depend on the design of our database and may not cover the full range of human emotional interactions, one could argue that they contain useful information about typical structure of emotional flow.

In this work, we have focused on context in the sense of past and future observations that are informative of the current observation. However, a broader definition of context could include general situational understanding which would give us prior information as to which emotional states are more likely to occur and when. In the future, we plan to focus on methods to incorporate higher level context in emotion recognition systems and exploit information from both interlocutors in the dyadic interaction. The emotional states of interlocutors in dyadic interactions are often related and influence each other; therefore, past and future context from the other speaker is expected to provide relevant information [9]. Also, our analysis assumes that each conversation is manually presegmented into utterances and that the emotional state within each utterance can be described by a single emotional label. It is important to examine alternative data-driven segmentations of a conversation into emotionally homogenous chunks that might allow us to look into emotional transitions at scales finer than the utterance level. Finally, in the future we would like to examine the performance of our methods on real life, spontaneous data sets.

REFERENCES

[1] R.E. Kaliouby, P. Robinson, and S. Keates, "Temporal Context and the Recognition of Emotion from Facial Expression," *Proc. HCI Int'l Conf.,* June 2003.

[2] J.M. Carroll and J.A. Russell, "Do Facial Expressions Signal Specific Emotions? Judging Emotion from the Face in Context," *J. Personality and Social Psychology,* vol. 70, pp. 205-218, 1996.

[3] H.R. Knudsen and L.H. Muzekari, "The Effects of Verbal Statements of Context on Facial Expressions of Emotion," *J. Nonverbal Behavior,* vol. 7, pp. 202-212, 1983.

[4] T. Masuda, P.C. Ellsworth, B. Mesquita, J. Leu, S. Tanida, and E. Van de Veerdonk, "Placing the Face in Context: Cultural Differences in the Perception of Facial Emotion," *J. Personality and Social Psychology,* vol. 94, pp. 365-381, 2008.

[5] A. Mehrabian, "Communication without Words," *Psychology Today,* vol. 2, pp. 53-56, 1968.

[6] B. de Gelder and J. Vroomen, "The Perception of Emotions by Ear and by Eye," *Cognition and Emotion,* vol. 14, pp. 289-311, May 2000.

[7] K. Oatley and J.M. Jenkins, *Understanding Emotions.* Blackwell Publishers Ltd, 1996.

[8] A.K. Dey and G.D. Abowd, "Towards a Better Understanding of Context and Context-Awareness," *Proc. First Int'l Symp. Handheld and Ubiquitous Computing,* 1999.

[9] C.-C. Lee, C. Busso, S. Lee, and S. Narayanan, "Modeling Mutual Influence of Interlocutor Emotion States in Dyadic Spoken Interactions," *Proc. 10th Ann. Conf. Int'l Speech Comm.,* 2009.

[10] C.M. Lee and S.S. Narayanan, "Toward Detecting Emotions in Spoken Dialogs," *IEEE Trans. Speech and Audio Processing,* vol. 13, no. 2, pp. 293-303, Mar. 2005.

[11] J. Liscombe, G. Riccardi, and D. Hakkani-Tur, "Using Context to Improve Emotion Detection in Spoken Dialog Systems," *Proc. Conf. Interspeech Comm.,* 2005.

[12] C. Conati, "Probabilistic Assessment of Users Emotions in Educational Games," *Applied Artificial Intelligence,* vol. 16, pp. 555-575, 2002.

[13] I. Cearreta, J.M. Lopez, and N. Garay-Vitoria, "Modelling Multimodal Context-Aware Affective Interaction," *Proc. Doctoral Consortium Second Int'l. Conf. Affective Computing and Intelligent Interaction,* 2007.

[14] G. McIntyre, "Towards Affective Sensing," *Proc. 12th Int'l Conf. Human-Computer Interaction,* 2007.

[15] A. Graves, S. Fernandez, and J. Schmidhuber, "Bidirectional LSTM Networks for Improved Phoneme Classification and Recognition," *Proc. 15th Int'l Conf. Artificial Neural Networks,* vol. 18, pp. 602-610, 2005.

[16] M. Wöllmer, A. Metallinou, F. Eyben, B. Schuller, and S. Narayanan, "Context-Sensitive Multimodal Emotion Recognition from Speech and Facial Expression Using Bidirectional LSTM Modeling," *Proc. 11th Ann. Conf. Int'l Speech Comm. Assoc.,* 2010.

[17] S. Fine, Y. Singer, and N. Tishby, "The Hierarchical Hidden Markov Model: Analysis and Applications," *Machine Learning,* vol. 32, pp. 41-62, 1998.

[18] A. Graves, S. Fernandez, M. Liwicki, H. Bunke, and J. Schmidhuber, "Unconstrained Online Handwriting Recognition with Recurrent Neural Networks," *Advances in Neural Information Processing Systems,* vol. 20, pp. 1-8, 2008.

[19] A. McCabe and J. Trevathan, "Handwritten Signature Verifcation Using Complementary Statistical Models," *J. Computers,* vol. 4, pp. 670-680, 2009.

[20] M.J.F. Gales and S.J. Young, "The Application of Hidden Markov Models in Speech Recognition," *Foundations and Trends in Signal Processing,* vol. 1, pp. 195-304, 2008.

[21] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book.* Entropic Cambridge Research Laboratory, 2006.

[22] C. Busso, M. Bulut, C-C Lee, A. Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee, and S. Narayanan, "IEMOCAP: Interactive Emotional Dyadic Motion Capture Database," *Language Resources and Evaluation,* vol. 42, pp. 335-359, 2008.

[23] M. Wöllmer, F. Eyben, B. Schuller, E. Douglas-Cowie, and R. Cowie, "Data-Driven Clustering in Emotional Space for Affect Recognition Using Discriminatively Trained LSTM Networks," *Proc. 10th Ann. Conf. Int'l Speech Comm. Assoc.,* pp. 1595-1598, 2009.

[24] I. Cohen, A. Garg, and T.S. Huang, "Emotion Recognition from Facial Expressions Using Multilevel HMM," *Proc. Neural Information Processing Systems,* 2000.

[25] J.A. Nelder and R. Mead, "A Simplex Method for Function Minimization," *Computer J.,* vol. 7, pp. 308-313, 1965.

[26] M. Pilu, "Video Stabilization as a Variational Problem and Numerical Solution with the Viterbi Method," *Proc. IEEE CS Conf. Vision and Pattern Recognition,* 2004.
[27] S. Hochreiter, Y. Bengio, P. Frasconi, and J. Schmidhuber, "Gradient Flow in Recurrent Nets: The Difficulty of Learning Long-Term Dependencies," *A Field Guide to Dynamical Recurrent Neural Networks,* S.C. Kremer and J.F. Kolen, eds., IEEE Press, 2001.
[28] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation,* vol. 9, no. 8, pp. 1735-1780, 1997.
[29] M. Schuster and K.K. Paliwal, "Bidirectional Recurrent Neural Networks," *IEEE Trans. Signal Processing,* vol. 45, no. 11, pp. 2673-2681, Nov. 1997.
[30] A. Graves and J. Schmidhuber, "Framewise Phoneme Classification with Bidirectional LSTM and Other Neural Network Architectures," *Neural Networks,* vol. 18, nos. 5/6, pp. 602-610, June 2005.
[31] M. Wöllmer, B. Schuller, F. Eyben, and G. Rigoll, "Combining Long Short-Term Memory and Dynamic Bayesian Networks for Incremental Emotion-Sensitive Artificial Listening," *IEEE J. Selected Topics in Signal Processing,* vol. 4, no. 5, pp. 867-881, Oct. 2010.
[32] A. Graves, "RNNLib Toolbox," http://sourceforge.net/projects/rnnl/, 2012.
[33] T. Bänziger and K.R. Scherer, "Using Actor Portrayals to Systematically Study Multimodal Emotion Expression: The GEMEP Corpus," *Proc. Second Int'l Conf. Affective Computing and Intelligent Interaction,* 2007.
[34] F. Enos and J. Hirschberg, "A Framework for Eliciting Emotional Speech: Capitalizing on the Actors Process," *Proc. First Int'l Workshop Emotion: Corpora for Research on Emotion and Affect (Int'l Conf. Language Resources and Evaluation),* 2006.
[35] R. Cowie, E. Douglas-Cowie, B. Apolloni, J. Taylor, A. Romano, and W. Fellenz, "What a Neural Net Needs to know About Emotion Words," *Computational Intelligence and Applications,* N. Mastorakis, ed., pp. 109-114, Word Scientific Eng. Soc., 1999.
[36] C.D. Manning and H. Schutze, *Foundations of Statistical Natural Language Processing,* chapter 11. The MIT Press, 1999.
[37] R. Bakeman and J.M. Gottman, *Observing Interaction: An Introduction to Sequential Analysis,* second ed. Cambridge Univ. Press, 1997.
[38] R. Bakeman and V. Quera, *Analyzing Interaction: Sequential Analysis with SDIS and GSEQ.* Cambridge Univ. Press, 1995.
[39] I. Cohen, Q.T. Xiang, S. Zhou, X. Sean, Z. Thomas, and T.S. Huang, "Feature Selection Using Principal Feature Analysis," *Proc. 15th Int'l Conf. Multimedia,* 2002.
[40] A. Metallinou, C. Busso, S. Lee, and S. Narayanan, "Visual Emotion Recognition Using Compact Facial Representations and Viseme Information," *Proc. Int'l Conf. Acoustics, Speech, and Signal Processing,* pp. 2474-2477, 2010.
[41] P. Boersma, "Praat, a System for Doing Phonetics by Computer," *Glot Int'l,* vol. 5, nos. 9/10, pp. 341-345, 2001.
[42] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile—The Munich Versatile and Fast Open-Source Audio Feature Extractor," *Proc. ACM Multimedia,* 2010.
[43] M.A. Hall, "Correlation-Based Feature Selection for Machine Learning," PhD thesis, Univ. of Waikato, 1999.
[44] I.H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques,* second ed. Morgan Kaufmann, 2005.
[45] A.V. Nefian, L. Liang, X. Pi, L. Xiaoxiang, C. Mao, and K. Murphy, "A Coupled HMM for Audio-Visual Speech Recognition," *Proc. Int'l Conf. Acoustics, Speech, and Signal Processing,* pp. 2013-2016. 2002,
[46] M. Brand, N. Oliver, and A. Pentland, "Coupled Hidden Markov Models for Complex Action Recognition," *Proc. IEEE CS Conf. Vision and Pattern Recognition,* 1997.
[47] G. Gravier, G. Potamianos, and C. Neti, "Asynchrony Modeling for Audio-Visual Speech Recognition," *Proc. Second Int'l Conf. Human Language Technology Research,* pp. 1-6, 2002.
[48] *SPSS Base 10.0 for Windows User's Guide,* SPSS Incorporated, 1999.
[49] J.A. Russell, J.-A. Bachorowski, and J.-M. Fernndez-Dols, "Facial and Vocal Expressions of Emotion," *Ann. Rev. of Psychology,* vol. 54, pp. 329-349, Feb. 2003.
[50] A. Metallinou, S. Lee, and S. Narayanan, "Decision Level Combination of Multiple Modalities for Recognition and Analysis of Emotional Expression," *Proc. Int'l Conf. Acoustics, Speech, and Signal Processing,* pp. 2462-2465, 2010.
[51] H.Perez Espinosa, C.A. Reyes Garca, and L.V. Pineda, "Acoustic Feature Selection and Classification of Emotions in Speech Using a 3D Continuous Emotion Model," *Proc. IEEE Ninth Int'l Conf. Automatic Face and Gesture Recognition,* 2011.

Angeliki Metallinou received the diploma in electrical and computer engineering from the National Technical University of Athens, Greece, in 2007, and the master's degree in electrical engineering in 2009 from the University of Southern California (USC), where she is currently working toward the PhD degree. Since Fall 2007 she has been a member of the Signal Analysis and Interpretation Lab (SAIL) at USC. Her research interests include multimodal signal processing, machine learning, statistical modeling, and affective computing. She is a student member of the IEEE.

Martin Wöllmer received the diploma in electrical engineering and information technology from the Technische Universität München (TUM), where his current research and teaching activity includes the subject areas of pattern recognition and speech processing. He works as a research assistant at TUM. His focus lies on affective computing and conversational speech recognition. He is a member of the IEEE.

Athanasios Katsamanis received the diploma in electrical and computer engineering (with highest honors) and the PhD degree from the National Technical University of Athens, Athens, Greece, in 2003 and 2009, respectively. He is currently a postdoctoral research associate at the School of Electrical Engineering in the University of Southern California, a member of the Signal Analysis and Interpretation Laboratory. His research interests lie in the area of speech and multimodal signal analysis and processing for modeling of human behavior, as well as image, acoustic, and articulatory data processing for speech production modeling. He is a member of the IEEE.

Florian Eyben received the diploma in information technology from the Institute for Human-Machine Communication at the Technische Universität München (TUM). He works for the Institute for Human-Machine Communication at TUM. Teaching activities of his comprise Pattern Recognition and Speech and Language processing. His research interests include large scale hierarchical audio feature extraction and evaluation as well as automatic emotion recognition. He is a student member of the IEEE.

Björn Schuller received the diploma in 1999 and the doctoral degree in 2006, both in electrical engineering and information technology from the the Technische Universität München (TUM) where he is now tenured as a senior researcher and lecturer in pattern recognition and speech processing. From 2009 to 2010 he was with the CNRS-LIMSI Spoken Language Processing Group in Orsay, France, and a visiting scientist in the Imperial College London's Department of Computing in London. He is a member of the ACM, HUMAINE Association, IEEE, and ISCA and (co)authored two books and more than 200 peer reviewed publications leading to more than 2,000 citations—his current H-index equals 23. He serves as a member and secretary of the steering committee, associate, and guest editor of the *IEEE Transactions on Affective Computing*, a guest editor and reviewer for more than 30 leading journals and multiple conferences in the field, and challenge organizer, including the INTERSPEECH 2009 Emotion, 2010 Paralinguistic, and 2011 Speaker State Challenges and the first Audio/Visual Emotion Challenge, and chairman and program committee member of numerous international workshops and conferences. He is a member of the IEEE.

Shrikanth Narayanan is the Andrew J. Viterbi professor of Engineering at the University of Southern California (USC), and holds appointments as a professor of electrical engineering, computer science, linguistics, and psychology and as the founding director of the Ming Hsieh Institute. Prior to USC he was with AT&T Bell Labs and AT&T Research from 1995-2000. At USC he directs the Signal Analysis and Interpretation Laboratory. He is also an editor for the *Computer Speech and Language Journal* and an associate editor for the *IEEE Transactions on Multimedia*, *IEEE Transactions on Affective Computing*, and the *Journal of the Acoustical Society of America*. He is a recipient of a number of honors, including Best Paper awards from the IEEE Signal Processing society in 2005 (with Alex Potamianos) and in 2009 (with Chul Min Lee) and selection as an IEEE Signal Processing Society Distinguished Lecturer for 2010-2011. Papers with his students have won awards at ICSLP '02, ICASSP '05, MMSP '06, MMSP '07, and DCOSS '09, InterSpeech 2009-Emotion Challenge, Interspeech 2010, InterSpeech 2011-Speaker State Challenge. He has published more than 450 papers and has had 10 US patents granted. He is a fellow of the IEEE, the Acoustical Society of America, and the American Association for the Advancement of Science (AAAS).