Shah *et al. EURASIP Journal on Audio, Speech, and Music Processing* (2015) 2015:4
DOI 10.1186/s13636-014-0049-y

EURASIP Journal on Audio, Speech, and Music Processing
a SpringerOpen Journal

RESEARCH Open Access

Within and cross-corpus speech emotion recognition using latent topic model-based features

Mohit Shah*, Chaitali Chakrabarti and Andreas Spanias

Abstract

Owing to the suprasegmental behavior of emotional speech, turn-level features have demonstrated a better success than frame-level features for recognition-related tasks. Conventionally, such features are obtained via a brute-force collection of statistics over frames, thereby losing important local information in the process which affects the performance. To overcome these limitations, a novel feature extraction approach using latent topic models (LTMs) is presented in this study. Speech is assumed to comprise of a mixture of emotion-specific topics, where the latter capture emotionally salient information from the co-occurrences of frame-level acoustic features and yield better descriptors. Specifically, a supervised replicated softmax model (sRSM), based on restricted Boltzmann machines and distributed representations, is proposed to learn naturally discriminative topics. The proposed features are evaluated for the recognition of categorical or continuous emotional attributes via within and cross-corpus experiments conducted over acted and spontaneous expressions. In a within-corpus scenario, sRSM outperforms competing LTMs, while obtaining a significant improvement of 16.75% over popular statistics-based turn-level features for valence-based classification, which is considered to be a difficult task using only speech. Further analyses with respect to the turn duration show that the improvement is even more significant, 35%, on longer turns (>6 s), which is highly desirable for current turn-based practices. In a cross-corpus scenario, two novel adaptation-based approaches, instance selection, and weight regularization are proposed to reduce the inherent bias due to varying annotation procedures and cultural perceptions across databases. Experimental results indicate a natural, yet less severe, deterioration in performance - only 2.6% and 2.7%, thereby highlighting the generalization ability of the proposed features.

Keywords: Speech emotion recognition; Topic models; Cross-corpus; Suprasegmental features

Introduction

Emotion conveys important information about a speaker's mood or personality, which makes it an ideal choice for improving human-machine interaction [1]. Speech-based emotion recognition is applicable to automatic speech recognition (ASR), spoken dialog systems (SDS) [2], automated call centers [3], education [1], entertainment [4], patient care and post-traumatic stress disorders [5]. Since emotions are highly specific to speakers, expression types, cultures, or context, determining the appropriate set of features that generalize well across such conditions is considered a difficult and challenging task.

*Correspondence: mohit.shah@asu.edu
School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe 85287, USA

Traditionally, speech is first segmented into turns based on the voice activity of a speaker, followed by an extraction of frame-level features. Segmental approaches operate directly on these frames using static Gaussian mixture models (GMM) [6], dynamic hidden Markov models (HMM) [7,8], or their variants [9]. Based on studies indicating the suprasegmental behavior of emotions [2,10], turn-level features have shown to significantly outperform HMM-based approaches. Typically, such features are obtained via statistics computed over the frame-level features [11-13]. Functions commonly include moments, extremes, percentiles, ranges, as well as the slope and error of linear regression. Classification/regression is then performed using various discriminative techniques such as k-nearest neighbor [14], linear discriminant or support vector classifiers [12], random forests [15], etc.

It is difficult to choose the right statistics *a priori*, hence a brute-force mechanism is typically used to extract 300 to 5,000 different features [11,16-18]. Often, feature selection is performed to obtain a reduced set. This selection is highly specific to each database, thus features selected for one database may be irrelevant on another set, leading to poor generalization. Furthermore, due to turn-based segmentation, turns are often long in duration depending on the speaker's activity. As the frame-level features are normalized over a turn to obtain these statistics, important local information, such as a short period of emotional burst, might get lost and consequently affect the performance.

To overcome these limitations, we propose a novel approach for extracting turn-level features using latent topic models (LTMs). LTMs are primarily used to extract topics that capture the various themes from a collection of text documents and their word occurrences. Extending this model to emotional speech, we posit that each turn (document) can be represented as a mixture of emotion-specific topics. The topics, in turn, capture emotionally salient information from the the co-occurrence behavior of frame-level acoustic features (words). Turns with similar emotions would exhibit a similar distribution over topics, thus allowing the latter to be used as high-level features for classification. Although such features do not model the temporal structure, local information is captured from the word occurrences, making them better suited for turns of longer durations as opposed to turn-level statistics. More importantly, the proposed approach offers a generative model-based explanation of how emotionally relevant features can be learnt from speech, thus overcoming the need for brute-force collection-based methods.

Besides natural language processing [19-21], LTMs have previously been used for human activity recognition [22], image annotation and segmentation [23], and image-based object recognition [24]. To the best of our knowledge, our earlier [25] and current works are the first to utilize LTMs for the purpose of learning emotionally relevant features from speech. In [25], we showed that topics derived via an unsupervised latent Dirichlet allocation (LDA) model outperformed HMMs over highly exaggerated emotions. Recently, replicated softmax models (RSM) [26], which are based on the idea of distributed representations, have shown to outperform LDA for text classification. In this paper, we use RSM as a building block and extend our earlier work in a significant manner while making notable contributions as follows:

- A supervised replicated softmax topic model (sRSM) to extract a set of naturally discriminative, turn-level features is proposed.
- A point-wise mutual information-based measure is proposed to qualitatively assess the relationship between topics, emotions, and acoustic features.
- Cross-corpus adaptation using two novel strategies, instance selection and weight regularization, is presented.

Our first contribution addresses the shortcomings of RSMs, which are primarily unsupervised and hence, the inferred topics are not naturally suited for discriminative tasks. We incorporate supervised learning by devising an sRSM - a feed-forward neural network with its initial weights obtained from an unsupervised RSM, followed by fine-tuning the weights using backpropagation. As opposed to random initialization, we use the RSM as a pre-training stage, which learns topics that initially capture properties of the underlying input only. Backpropagation allows us to slightly perturb and refine these topics with respect to the output labels, thereby, facilitating learning of features that are optimal for discriminative tasks. RBMs for learning discriminative features have also been proposed in [27,28], where a deep neural network-based generalized discriminant analysis (DNN-GerDA) was used to learn emotion-specific, turn-level features. The proposed sRSM is fundamentally different in the following key aspects: (i) DNN-GerDA employs the Fisher discriminant criterion, which maximizes the ratio of between-class variance to within-class variance, while sRSM directly minimizes the cross-entropy error, which is more appropriate for classification-related tasks [29,30]; (ii) DNN-GerDA assumes that the extracted features are drawn from Gaussian class-conditional distributions, while sRSM makes no assumptions regarding the statistical properties of the inferred topics; (iii) DNN-GerDA accepts arbitrarily distributed, real-valued observations as input, whereas, sRSM models discrete, count-like observations commonly found in text collections; and (iv) the discriminative features in [28] are learnt over turn-level statistics extracted via brute-force as opposed to the acoustic bag-of-words used in this work.

Our second contribution addresses the qualitative aspects by providing a physical interpretation of the topics in terms of high-level emotions and the frame-level acoustic words. Using a normalized point-wise mutual information-based measure between emotions and topics learnt in an unsupervised manner, we show that the former are nicely separated in the topic space - topics that co-occur frequently with one emotion, rarely, or never co-occur with other emotions. We further show that topics induce a natural grouping over acoustic features based on their energy distribution across frequency, which indirectly relates to the emotional state.

Our final contribution addresses the generalization ability via cross-corpus emotion recognition. Most

cross-corpus studies [31-33] do not account for the varying annotation procedures and cultural perceptions across corpora. These differences bias the classifier such that the decision boundary learnt over the training corpus is rendered sub-optimal for the test corpus. To reduce this bias and improve cross-corpus performance, we propose two novel adaptation-based approaches - (1) instance selection and (2) weight regularization. In the former approach, we identify instances from the training corpus that are wrongly classified according to the test corpus and train a new classifier after removing such misleading instances. In the latter approach, we modify the conventional *L2*-norm regularization, which penalizes large weights, to instead penalize the weights for being too different from reference weights estimated over the test corpus. Our instance selection approach differs from that of [34], where the selection criteria is based on the distance in the feature space as opposed to the classifier.

We perform experiments on two databases, USC IEMOCAP [35] and SEMAINE [36], with acted and spontaneous expressions, respectively. We evaluate four sets of features, obtained via unsupervised or supervised LDA and RSM, for speaker-independent binary and categorical and continuous prediction and show that sRSM provides the best unweighted average recall in each scenario. We further investigate the performance with respect to the turn duration and demonstrate that sRSM and in general, LTMs, are better suited to handle turns of longer durations (>6 s) than turn-level statistics. Specifically, for valence-based classification, which is considered to be a challenging problem using only speech [12,13,37], sRSM outperforms turn-level statistics by 16.75% over all the turns of SEMAINE, and a remarkable 35% over turns longer than 6 s. In a cross-corpus setting, adaptation using instance selection and weight regularization demonstrates a relative deterioration of only 2.6% and 2.7% respectively, further highlighting the generalization ability of the proposed features for speech emotion recognition.

The remainder of this paper is organized as follows. First, we provide a background on latent topic models and their extension for learning emotionally relevant features. This is followed by a description of the experiments and results for within and cross-corpus recognition. Finally, we provide the conclusions and directions for future work.

Latent topic models

Latent topic models are based on the assumption that a text document can be represented as a mixture of topics, and each topic can be represented as a mixture over a dictionary of words. For notation purposes, we describe a document d in a collection of D documents as a stream of N words $\mathbf{V} = [v_1, \ldots, v_N]$. Each word, v_n, is a K-dimensional unit vector, where $v_{nk} = 1$, if the n^{th} word belongs to the k^{th} dictionary element. Given the observed words of a document, the objective is to infer J latent topics $h_1, \ldots, h_J$ that maximize the joint likelihood of words and topics, i.e. $p(\mathbf{V}, h)$. Since thematically similar documents will exhibit nearly similar distributions over the latent topics, the latter can be used as intermediate or high-level features for subsequent classification. LDA and its supervised counterpart, sLDA, are quite well known and hence are briefly described here. RSM and our proposed extension to supervised RSM are covered in more detail.

Latent Dirichlet allocation

LDA [21] can be depicted as a directed graphical model as shown in Figure 1a. According to LDA, the process for generating a document and its words is described as follows:

- Choose J topics $h \sim \text{Dirichlet}(\alpha)$
- For each word v_n in the turn -
  - Choose a topic $x_n \sim \text{Multinomial}(h)$
  - Choose a word $v_n \sim p(v_n|x_n, W)$

The topics h, sampled once for each document, are drawn from a Dirichlet distribution parametrized by α. x_n is a J-dimensional, unit-basis vector indicating which topic is active for the n^{th} word. The relationship between topics and words is defined by a matrix W of size $J \times K$. Each row of W, i.e. $W_{j\cdot}$, is a discrete distribution over K dictionary words. Inference in LDA involves the estimation of hidden topics h and x given the words v for each document, i.e. $p(h, x|\mathbf{V}, \alpha, W)$. Exact inference being intractable, an iterative variational approximation method is used in this study. A detailed description of this method can be found in [21]. The parameters α and W are learnt via the expectation-maximization (EM) algorithm.

In order to perform binary or multi-class, categorical emotion recognition, a softmax regression-based classifier is trained over the posterior topics h inferred from the training examples. The classifier parameters, θ, are estimated by minimizing the cross-entropy error with standard L2 regularization, as per Equation 1. Here, $1\{\cdot\}$ is the indicator function, S denotes the number of training examples, C denotes the number of classes, $t^{(s)}$ denotes the ground truth for example s, and λ denotes the regularization parameter. Iterative minimization is performed using minibatch stochastic gradient descent with a batch-size of 100 and a learning rate and momentum of 0.005 and 0.8, respectively.

$$L(\theta) = -\frac{1}{S}\left[\sum_{s=1}^{S}\sum_{c=1}^{C} 1\left\{t^{(s)} = c\right\} \log p\left(y^{(s)} = c|h^{(s)}, \theta\right)\right] + \frac{\lambda}{2}\|\theta\|_2^2 \quad (1)$$

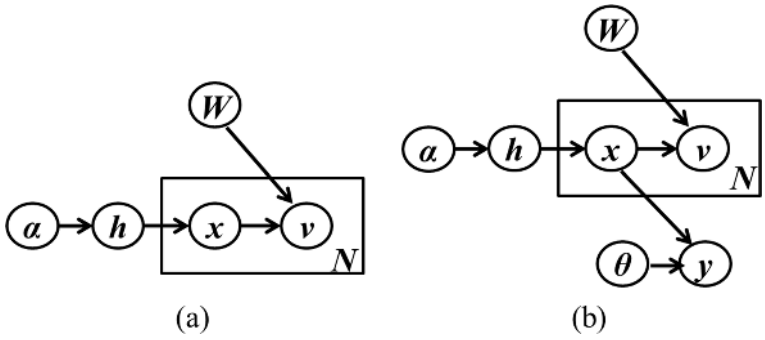

Figure 1 Graphical model representation of (a) LDA and (b) sLDA. Plates (rectangular) drawn around nodes indicate replication, which corresponds to the number of input observations or words in a document, i.e. *N*.

$$p\left(y^{(s)}=c|h^{(s)},\theta\right)=\frac{\exp\left(\theta_c^T h^{(s)}\right)}{\sum_{l=1}^{C}\exp\left(\theta_l^T h^{(s)}\right)} \tag{2}$$

The output, $y^{(s)}$, defined by the softmax function in Equation 2, returns the posterior probability for each class. The label is then predicted by evaluating $\text{argmax}\, p\left(y^{(s)}=c|h^{(s)},\theta\right)$. Similar expressions can be derived for the case of predicting real-valued outputs using linear regression.

Supervised LDA (sLDA) [38] differs from its unsupervised counterpart in the following aspect: an additional node y, the class label or output, is introduced as shown in Figure 1b. The output, dependent on the topic indicators x, is predicted according to Equation 3, where $\bar{x}=\sum_{n=1}^{N}x_n$ represents the empirical topic frequencies. The variable θ, which parametrizes the relationship between the topic indicators and the output label, is estimated in an iterative manner along with the topics. As a result, class-specific information is considered while learning topics, thereby leading to better discrimination in comparison to LDA. Variational approximation is used to infer the latent variables as described in [38]. The class label is predicted by evaluating $\text{argmax}\,\theta_m^T\bar{x}$. Equation 3 is specific to binary or categorical classification; a similar expression can be derived for regression.

$$y_m\sim\frac{\exp\left(\theta_m^T\bar{x}\right)}{\sum_{l}^{C}\exp\left(\theta_l^T\bar{x}\right)} \tag{3}$$

Replicated softmax model

RSM belongs to the family of undirected, energy-based models known as restricted Boltzmann machines (RBMs) [39]. The visible unit is modeled as a softmax unit instead of a Bernoulli variable as in RBM [26], which facilitates the modeling of occurrence or count-like observations. A graphical representation of this model is shown in Figure 2a. For a document with N words, the observation $\mathbf{V}$ is an $N\times K$ binary matrix, and $h\in\{0,1\}^J$ are the binary stochastic latent topics. The energy of this configuration is defined as per Equation 4, while the conditional probabilities of words and topics are defined as per Equations 5 and 6, respectively.

$$E(v,h)=-\sum_{n=1}^{N}\sum_{j=1}^{J}\sum_{k=1}^{K}W_{njk}h_j v_{nk}-\sum_{n=1}^{N}\sum_{k=1}^{K}v_{nk}a_{nk}-\sum_{j=1}^{J}h_j b_j \tag{4}$$

$$p(v_{nk}=1|h)=\frac{\exp\left(a_{nk}+\sum_{j=1}^{J}h_j W_{njk}\right)}{\sum_{q=1}^{K}\exp\left(a_{nq}+\sum_{j=1}^{J}h_j W_{njq}\right)} \tag{5}$$

$$p(h_j=1|v)=\sigma\left(b_j+\sum_{n=1}^{N}\sum_{k=1}^{K}v_{nk}W_{njk}\right) \tag{6}$$

Here, W_{njk} denotes the weight between visible unit n that takes on value k and hidden topic j; b_j is the bias of hidden topic j, and a_{nk} is the bias of visible unit n that takes on value k. $\sigma(x)=1/(1+\exp(-x))$ is the logistic function. Ignoring the sequence in which words arrive, if the k^{th} unit for each word v_n is forced to share its weight with the k^{th} unit of all the other words in the document, then W_{njk} can be written simply as W_{jk}. This process is depicted in Figure 2b. The weight-sharing property reduces the number of parameters to be learnt from $N\times J\times K$ to $J\times K$ and also allows the model to account for documents of different lengths. This property is essential for emotional speech since the duration of a turn is not fixed and varies depending on the speaker's activity. The energy of the configuration after weight-sharing is then defined using Equation 7.

$$E(\mathbf{V},h)=-\sum_{j=1}^{J}\sum_{k=1}^{K}W_{jk}h_j\hat{v}_k-\sum_{k=1}^{K}\hat{v}_k a_k-N\sum_{j=1}^{J}h_j b_j \tag{7}$$

Here, $\hat{v}_k=\sum_{n=1}^{N}v_{nk}$ denotes the frequency with which the k^{th} dictionary element appears in the turn. Unlike LDA, each word in an RSM is generated by multiple

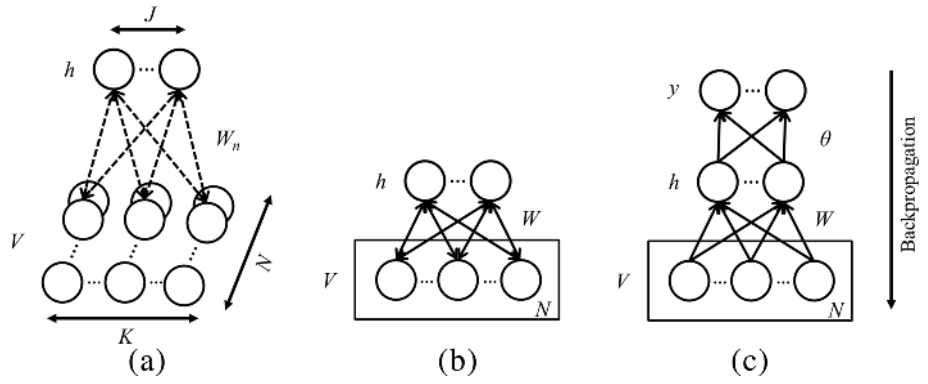

Figure 2 Graphical model representation of (a) RSM without weight-sharing, (b) RSM after weight-sharing, and (c) sRSM. In **(a)**, the weights are only shown for the n[th] word. Plates (rectangular) drawn around nodes in **(b)** and **(c)** indicate replication, which corresponds to the number of input observations or words in a document, i.e. *N*.

topics, leading to distributed and richer representations. Furthermore, the topics are inferred in a single pass via Equation 6, offering a significant reduction in computational complexity.

The parameters, $\{W, a, b\}$, are estimated by maximizing the log-likelihood of the observed words, i.e. $P(\mathbf{V}) = \sum_h \exp(-E(\mathbf{V}, h))/Z$, where Z is the normalizing constant. Exact learning is intractable, hence, we use an approximate technique called *contrastive divergence* (CD) algorithm [26,39]. The update rule for weights, according to CD, is given in Equation 8. Here, η is the learning rate and E_{model_T} represents the expectation with respect to the distribution after running a Gibbs chain for T steps. $T = \infty$ is equivalent to maximum likelihood learning. Further details on this technique and its convergence properties can be found in [40]. Update equations can be derived similarly for biases a and b.

$$\Delta W_{jk} = \eta \left(E_{\text{data}} \left[\hat{v}_k h_j \right] - E_{\text{model}_\text{T}} \left[\hat{v}_k h_j \right] \right) \tag{8}$$

In this work, RSM is trained for 200 epochs with a batchsize of 100 and a learning rate and momentum of 0.002 and 0.8, respectively. For CD, we found $T = 1$ to be sufficient for generating good features. Classification/regression is then performed over the inferred topics h using softmax or linear regression as outlined earlier for LDA, i.e. using Equations 1 and 2.

Supervised RSM

In order to devise an sRSM, we propose a fully connected, feed-forward neural network (FNN) as shown in Figure 2c. The input and hidden layer are the same as that of an RSM, while the topmost layer performs output prediction. For C-class, categorical recognition, the top layer is a softmax layer and the output is computed via Equation 2. Typically, an FNN is initialized with random weights before using backpropagation to perform fine-tuning. An sRSM differs from a conventional FNN in the following aspect - instead of random initialization, we use the weights obtained via CD learning of an RSM to initialize the network. Thus, the RSM is treated as a pre-training stage whose task is to learn initial weights that capture properties of the underlying input observations in an unsupervised manner. Backpropagation in an sRSM can then be viewed as slightly perturbing these weights to account for the output labels and as a result, learn features that are better suited for the discriminative task under consideration. For backpropagation in sRSM, we use the cross entropy error as the cost function for classification and the mean squared error (MSE) for linear regression. Stochastic gradient descent is used to update the parameters with a learning rate and momentum of 0.005 and 0.8, respectively.

The advantage of using pre-trained weights can be seen in Figure 3, which shows the classification error across epochs, averaged over 100 runs. The results are displayed for the task of arousal-based, binary classification on two databases, SEMAINE and USC IEMOCAP, which are described in more detail in subsequent sections. It is clearly evident that backpropagation over randomly initialized weights is prone to get stuck at a bad local optima and yield a higher classification error. On the contrary, pre-training using an RSM first models the observations in an unsupervised manner and finds a good starting point for the weights, which leads to a lower classification error.

It is important to note that the impact of pre-training is higher for smaller databases. Here, the SEMAINE and USC IEMOCAP databases consist of approximately 1,000 and 5,000 training examples, respectively. From Figure 3, we can observe that there is a slight decrease in the effectiveness of pre-training from SEMAINE to USC

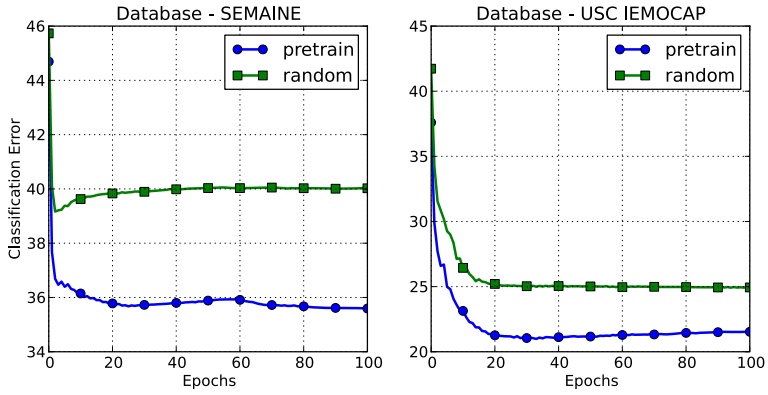

Figure 3 Effects of pre-training. A comparison of the classification error (%) between an sRSM with pre-trained and randomly initialized weights on SEMAINE and USC IEMOCAP. The higher error achieved in the latter case is due to the parameters getting stuck at a bad local optima. Pre-training overcomes this limitation and provides a better starting point using the input observations only.

IEMOCAP. To further investigate this behavior, we varied the number of training examples used for fine-tuning the sRSM. Using the same database, USC IEMOCAP, Figure 4 shows that the difference in classification error between pre-trained and randomly initialized sRSM gradually declines as the training examples increase from 500 to 5,000.

LTM-based features for emotion recognition

Numerous frame-level features of the prosodic, voice quality, spectral, and perceptual [41-43] kind have been used previously for emotion recognition. In this work, we only use energy, fundamental frequency (F0), and the first 12 Mel frequency cepstral coefficients (MFCCs) (ignoring the 0^{th} coefficient) as they have shown to be the most successful across different databases. The features are extracted using a frame and step size of 25 and 10 ms, respectively, i.e. a frame rate of 100 frames/s. The first- and second-order differences are appended to obtain a 42-D feature vector per frame. Energy and MFCCs are extracted using the HTK Toolkit [44], while the F0 estimates are extracted using the OpenEar Affect Recognition Toolkit [16]. Principal component analysis (PCA) is further applied to reduce the dimensionality to 13 features.

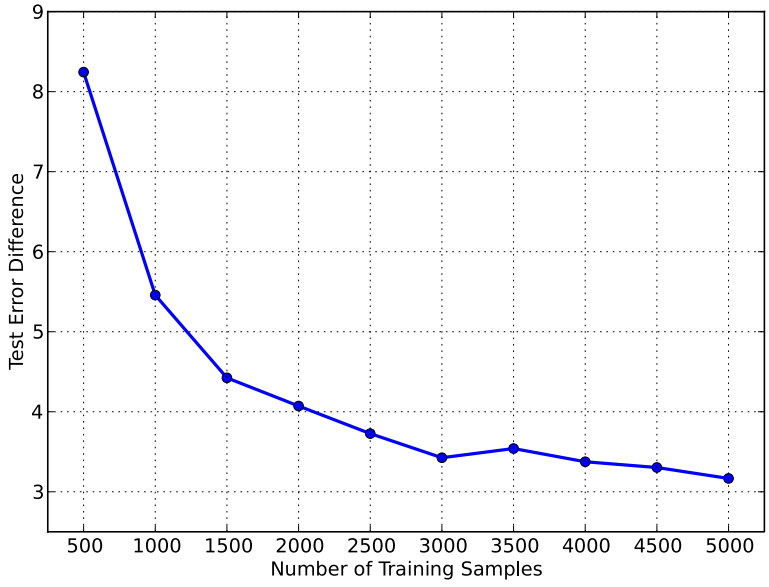

Figure 4 Impact of pre-training using training sets of different sizes. The difference in classification error (%) between an sRSM with pre-trained and randomly initialized weights on USC IEMOCAP. Note the gradual decline in the difference as the number of examples available for fine-tuning the sRSM is increased.

In order to learn topics over these frame-level features, we must first define the equivalent terms for documents and words. The definition of a document is straightforward; we consider each turn to be analogous to a text document. A turn can thus be viewed as a stream of multi-dimensional, real-valued frames, where the latter still need to be converted to discrete values or symbols. There are a number of methods for dictionary learning and encoding; we use a simple vector quantization-based approach to learn a dictionary of K candidate feature vectors. Each frame vector is then denoted by the index of the dictionary candidate it is closest to in terms of the Euclidean distance. Each turn is a stream of discrete words $\mathbf{V} = [\nu_1, \ldots, \nu_N]$, where N is the total number of words/frames in a turn. The topics, h, are then inferred using either unsupervised or supervised LDA or RSM, followed by classification or regression as described earlier.

It is worthwhile to investigate and provide a physical interpretation of topics in terms of their relationship with emotions and the underlying acoustic words. Ideally, one would expect to have as many topics as emotion categories, but, variations across speakers, spoken content, and mannerisms cause the number of topics, J, to usually lie between the number of emotions, C, and the dictionary size, K. Among these topics, only a few may convey emotion-specific information, while certain topics may be present across all emotions and can be considered uninformative or irrelevant. In order to identify such topics, we propose a normalized point-wise mutual information (PMI) measure between the individual topics h and emotions y. This measure allows us to quantify the discrepancy between the joint probability of h and y and their marginal distributions under the assumption of independence. The PMI is computed as per Equation 9, where $1 \leq j \leq J$ and $1 \leq c \leq C$. The values are further normalized to a range of $[-1, +1]$ as described in [45] using Equation 10. Here, -1 indicates never co-occurring, $+1$ indicates always occurring together, and 0 indicates independence. The most informative topics for each emotion are then identified and ranked by the decreasing order of their normalized PMI values.

$$pmi(h_j; y_c) = \log \frac{p(h_j, y_c)}{p(h_j)p(y_c)} \tag{9}$$

$$npmi(h_j; y_c) = \frac{pmi(h_j; y_c)}{-\log p(h_j, y_c)} \tag{10}$$

In Figure 5, we display the normalized PMI values for 64 topics, extracted using unsupervised LDA and RSM, for a single male speaker across the four emotions (neutral, sad, happy, angry) of USC IEMOCAP. We observe topics that exhibit a high co-occurrence with sadness to occur never or rarely with angry or happy emotions. Similar observations can be made for the vice versa case. Hence, even without using any label information while learning topics, we can observe that the emotions are nicely separated in the topic space, with sad and happy topics being the most easily distinguishable from each other. The only exception is neutral; very few topics show a high co-occurrence with only neutral emotions, with a majority of them also co-occurring with the other three emotions. Between LDA and RSM, we can observe that the topics obtained via LDA capture neutral emotions slightly better than RSM, i.e. the highly ranked topics for neutral emotions co-occur less with other emotions. On the other hand, RSM represents the remaining emotions such as sad, happy, and angry better than LDA.

Taking further advantage of the generative mechanism of topic models, we can also interpret the relationship between topics and the underlying acoustic words. We identify the most probable words, using the weight matrix W characterizing $p(\nu|h)$, under the highest ranked topic for each emotion. The spectrograms, reconstructed from the MFCCs, for the top 3 words are shown in Figure 6. For acoustic words grouped under sad or neutral topics, we observe that most of their energy is concentrated at lower frequencies (<2,000 Hz). In comparison, words grouped under happy or angry topics show that the energy is more spread out across frequency. Thus, it can be said that the individual topics induce a natural grouping over acoustic words primarily based on their energy distribution across different frequencies.

Evident from both, the normalized PMI values and the most likely words for individual topics, there is a strong overlap between happy-angry and neutral-sad emotions. The inability to visualize distinguishable characteristics across the valence axis is a well-known limitation of speech, which is more reactive to changes along the arousal dimension. Our experiments in the subsequent section will highlight the importance of LTMs, which allow us to represent a turn as a mixture of multiple topics, towards classification, especially along the valence dimension.

Within-corpus emotion recognition

In this section, we describe the experiments for evaluating the proposed features in a within-corpus setting, where the train and test examples belong to the same corpus.

Databases

The USC IEMOCAP [35] corpus was built by asking five pairs of male-female actors to elicit emotions either by reading from a script or via improvisation in a conversational setting. There are a total of 151 dialogs which, after

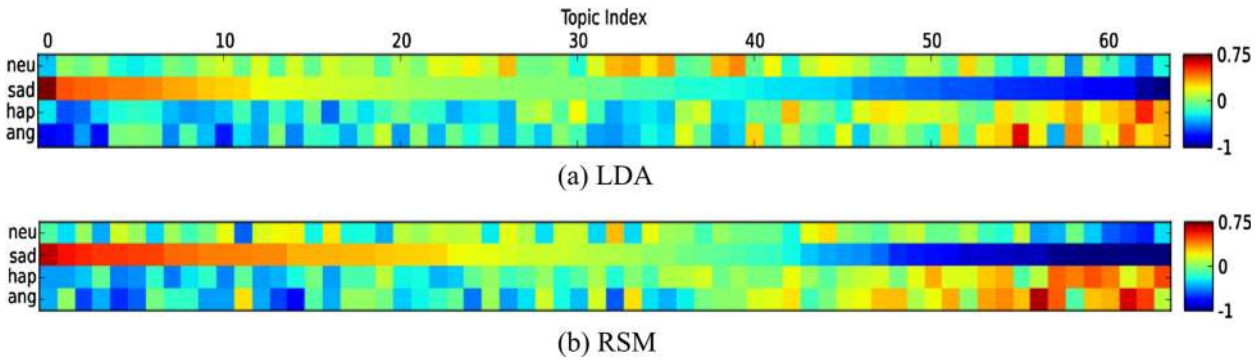

Figure 5 Qualitative interpretation of emotions and topics using (a) LDA and (b) RSM. The normalized PMI between 64 topics and four emotions reveals that individual topics capture emotion-specific information. Topics obtained via LDA capture neutral emotions slightly better than RSM, whereas, the latter achieves better separability in case of the remaining emotions - sad, happy, and angry.

turn-based segmentation, yield a total of 10,039 turns. At least three evaluators assigned a categorical attribute to each turn. Attributes include neutral, sad, happy, excited, angry, frustrated, surprised, disgust, fear, and unknown. Following the selection criteria outlined in [8,11,46], only those turns for which a majority consensus was reached among the evaluators are considered in this study. Of these, turns labeled as neutral, sad, happy, excited, and angry are selected, while the remaining attributes are not considered as they are under-represented. Happy and excited are treated as the same emotion and merged as one class. This results in a total of 5,531 turns distributed across ten speakers and four emotions: neutral (1,708), sad (1,084), happy (1,636), and angry (1,103).

The SEMAINE [36] corpus consists of spontaneously expressed emotions elicited via interaction with human operators enacting characters with pre-defined emotional traits. The corpus consists of a total of 95 sessions recorded from 24 speakers, of which, 82 sessions include a force-aligned transcript necessary for turn-based segmentation. Each turn is annotated at a frame rate of 20 ms by atleast two evaluators with real-valued arousal and valence attributes. For each attribute, the average value across all evaluators is calculated per frame, followed by an average over all the frames in a turn to yield a single value. For regression tasks, this value is treated as the ground truth. While, for binary classification, the global mean calculated over all the turns is used to threshold and obtain a 0 (low arousal/negative valence) or 1 (high arousal/positive valence) binary label for each turn. Following the specifications provided in the 2011 Audio-Visual Emotion Challenge [18], turns are partitioned into three speaker-disjoint sets: train (1,185), development (960), and test (673).

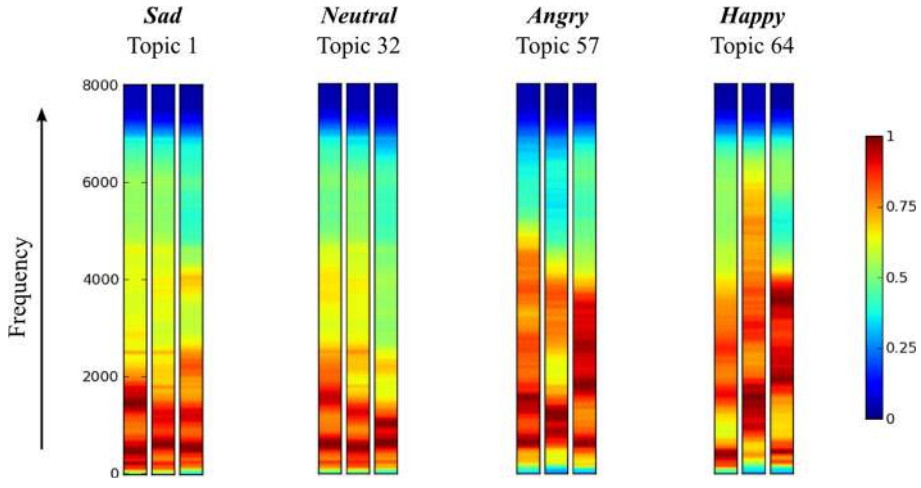

Figure 6 The top 3 probable acoustic words for the highest ranked emotion-specific topics obtained using RSM. The words are visualized via spectrograms reconstructed from the MFCCs. Individual topics induce a natural grouping over acoustic words based on their underlying distribution of energy across frequency.

Baseline and metrics

In addition to comparison between different LTMs (LDA, sLDA, RSM, and sRSM), we consider two baseline approaches - IS09 and VQ. The former comprises of a set of 384 brute-force-based statistical features and is the popular choice among researchers as evident from its use in the 2009 InterSpeech and 2011/2012 AVEC challenges [17,18,47]. These features were extracted using the openEAR Affect Recognition toolkit [16]. Classification is performed via a linear kernel SVM/SVR trained using the WEKA toolkit [48]. Results for the second baseline approach, VQ, are obtained via a linear kernel SVM/SVR trained directly over the acoustic word occurrences. This allows us to assess the relative gain offered by LTMs.

We use the weighted average (WA) and unweighted average (UA) recall metrics, as defined in [17], to evaluate binary or categorical recognition performance. The former is defined as the average classification accuracy, while the latter is defined as the average of the class-wise accuracies. These metrics are calculated using Equation 11.

$$\text{WA} = 100 \cdot \frac{m}{M} \qquad \text{UA} = 100 \cdot \frac{1}{C} \sum_{c=1}^{C} \frac{m_c}{M_c} \tag{11}$$

Here, m denotes the number of examples correctly classified, and M denotes the total number of examples. Similarly, m_c and M_c denote the number of examples correctly classified and total number of examples, respectively, for a specific class c. Since UA recall is more appropriate for unbalanced datasets, statistical significance over the baseline is determined using a one-tailed test (difference of proportions) over the UA recall values. Unless mentioned otherwise, the confidence level is higher than 95%. Finally, the correlation coefficient (COR) is used to evaluate performance for regression tasks.

USC IEMOCAP

To ensure speaker independence, we performed experiments using a leave-one-speaker-out (LOSO) strategy, resulting in a 10-fold process corresponding to the ten speakers. For each fold, we first normalized the frame-level features of the test partition to have the same mean as that of the training partition. We further learnt dictionaries of multiple sizes, $K \in \{64, 128, 256, 512\}$, while the number of topics $J \in [K/4, K/2]$. The optimal values of K and J are highly dependent on the training examples, thus vary for each fold. Cross-validation over the training set was used to determine their optimal values.

The WA and UA recall values are presented in Table 1. Classification over acoustic words, i.e. VQ, yields a reasonable performance of 52.18%. LDA and RSM show a relative increase of 6.31% and 11.51%, respectively, over VQ. Supervised learning leads to further improvements as demonstrated by the better recall obtained using sLDA and sRSM over their respective unsupervised counterparts.

Compared to IS09, LDA shows a marginal improvement; however, RSM and sRSM demonstrate significant improvements with a recall of 57.92% and 59.03%, respectively. Relatively, RSM and sRSM offer gains of 5% and 7% over IS09, respectively. Compared to earlier works, our method clearly outperforms the UA recall of 50.69% achieved using HMMs in Metallinou et al. [8]. It also compares well with the previous best result of Lee et al. [46], where a recall of 58.46% was achieved using the IS09 feature set combined with a hierarchical decision-tree-based classifier. Since the train and test partitions in these works differ slightly from our experiments, an exact comparison is not feasible.

Further inferences can be drawn from the class-wise accuracies provided in Table 1. We observe that speech-based methods are best at recognizing sadness while achieving similar performances for happy and anger. In case of neutral emotions, sRSM shows a 56.38% accuracy, which is better than the previous best of 54.54% obtained by Lee et al. [46] and 35.23% of Metallinou et al. [8]. The difficulty in recognizing neutral emotions is also evident from the inter-evaluator agreement - of the 1,708 neutral turns in this set, evaluators were in complete agreement for only 340 turns, i.e. 19.9%. Whereas, across all emotions, human evaluators were in complete agreement with each other for only 2,040 out of 5,531 turns (36.88%). Given such ambiguity inherent in perceiving even acted expressions, the overall improvement in recall achieved here is of significant value.

Compared to IS09, we observe that RSM/sRSM performs slightly worse at recognizing anger. This can be attributed to the different frame-level features used across the two approaches. In our work, bag-of-words features are constructed from F0, energy, and MFCCs. In contrast, the IS09 feature set uses two additional frame-level features: zero crossing rate (ZCR) and harmonic-to-noise (HNR) ratio. In [46], the same feature set combined with a tree-based classification scheme provided a similar recall for anger as IS09. This further suggests that the higher recall achieved on anger is mainly due to the type of features and not dependent on the the type of classifier.

From an unsupervised RSM to an sRSM, we can observe a slight deterioration towards recognizing sadness. This effect is well explained by the imbalance across different classes in the training examples. The cross-entropy error loss function aims to maximize the average classification accuracy, i.e. WA recall. Owing to the higher number of neutral and happy utterances, the topics learnt during the fine-tuning stage of sRSM are slightly biased towards these emotions. In order to address this issue, one can restrict each class to have the same number of examples

Table 1 Recall values (%) for categorical recognition on USC IEMOCAP

Metric	IS09	VQ	LDA	sLDA	RSM	sRSM
Neu	53.62	39.64	51.70	52.46	49.88	56.38
Sad	62.45	71.31	66.79	67.34	74.72	70.39
Hap	47.00	54.10	50.37	51.53	54.77	54.76
Ang	57.57	49.95	52.04	52.95	52.31	54.58
WA	54.17	52.18	54.33	55.20	56.68	*58.29*
UA	55.14	51.94	$55.22^{\dagger}$	$56.07^{\dagger}$	$57.92^{*\dagger}$	$*59.03^{*\dagger}$

Symbols $*$ and † indicate statistical significance over IS09 and VQ, respectively.

during training, i.e. a balanced dataset. Alternatively, one can also modify the loss function in Equations 1 to 12, where the training examples are individually weighted, $w^{(s)}$.

$$L(\theta) = -\frac{1}{S}\left[\sum_{s=1}^{S}\sum_{c=1}^{C} w^{(s)} 1\left\{t^{(s)} = c\right\} \log p\left(y^{(s)} = c|h^{(s)}, \theta\right)\right] + \frac{\lambda}{2}\|\theta\|_2^2 \tag{12}$$

SEMAINE

Cross-validation was not required for this database, since the training, development, and test partitions, as specified in [18,47], do not overlap in the speakers. The frame-level features are normalized as per the method outlined earlier for USC IEMOCAP. The development set was used to select the model parameters, i.e. the dictionary size, $K \in \{64, 128, 256\}$, and number of topics, $J \in [K/4, K/2]$. Results are reported for both partitions, development, and test.

The results for arousal-based, binary classification and regression are shown in Table 2. As observed for the USC IEMOCAP database, LTMs outperform simple VQ-based features. Specifically, LDA and RSM achieve relative gains of 1.5% and 13.7%, respectively. Topics learnt in a supervised manner, as expected, lead to even further improvements; 7.6% and 14.6% for sLDA and sRSM, respectively. Compared to IS09, the proposed features demonstrate a significant improvement on the development set. Whereas on the test set, LDA and sLDA perform worse than IS09. RSM and sRSM are marginally better with relative gains of 0.7% and 1.4%, respectively. In case of regression, however, LTMs outperform IS09 on both the sets. Once again, sRSM yields the best performance with a COR of 0.384 and 0.444, compared to 0.238 and 0.288 using IS09, on the development and test sets, respectively.

Table 3 shows a comparison for the case of valence-based, binary classification and regression. Compared to VQ, LDA and RSM demonstrate an improvement of 8.8% and 10.3%, respectively. Once again, supervised learning via sLDA or sRSM improves upon its unsupervised counterparts. Unlike arousal, LTM-based features comprehensively outperform IS09 features. The latter, in this case, performs slightly worse than chance. Again, the best recall is obtained using sRSM - relative gains of 12.1% and 16.75% over IS09 on the development and test sets, respectively. In case of regression, sRSM obtains a COR of 0.349 and 0.171 on the development and test sets respectively, which is clearly better than 0.191 and 0.007 obtained using IS09.

The results obtained on SEMAINE are comparable to earlier works; in [49], a WA recall of 64.98% (arousal) and 63.51% (valence) was achieved using SVM and AdaBoost over statistics-based features. While, in [50], an UA recall of 65.7% (arousal) and 65.4% (valence) was achieved using a bag of HMMs approach. In each of these studies, the

Table 2 Results for arousal-based classification and regression on SEMAINE

Metric	IS09	VQ	LDA	sLDA	RSM	sRSM
			Development set			
WA	60.73	60.72	63.85	65.31	66.04	*66.35*
UA	61.08	60.81	64.03	$65.39^{\dagger}$	$66.02^{*\dagger}$	$*66.38^{*\dagger}$
COR	0.238	0.325	0.350	0.364	0.357	*0.384*
			Test set			
WA	67.16	66.86	67.90	71.03	71.47	*72.66*
UA	63.46	56.17	57.05	60.49	$63.90^{\dagger}$	$*64.38^{\dagger}$
COR	0.288	0.255	0.312	0.322	0.430	*0.444*

Classification results are expressed in percentage (%). Symbols $*$ and † indicate statistical significance over IS09 and VQ, respectively.

Table 3 Results for valence-based classification and regression on SEMAINE

Metric	IS09	VQ	LDA	sLDA	RSM	sRSM
			Development set			
WA	59.61	58.33	63.03	64.10	65.50	*66.45*
UA	57.64	56.51	58.86	61.96†	63.53*†	*64.62**†
COR	0.191	0.191	0.330	0.332	0.327	*0.349*
			Test set			
WA	49.93	51.56	55.13	57.21	56.32	*57.80*
UA	49.68	51.54	56.12*	57.63*†	56.88*†	*58.00**†
COR	0.007	0.045	0.128	0.154	0.127	*0.171*

Classification results are expressed in percentage (%). Symbols $*$ and † indicate statistical significance over IS09 and VQ, respectively.

features were extracted over individual spoken words as opposed to turns, hence, a direct comparison with our approach is not feasible.

Based on the above experimental results on USC IEMOCAP and SEMAINE, we can make the following observations. Firstly, LTMs learn simplified, yet better, representations over acoustic words and their co-occurrences as demonstrated by the clearly higher recall obtained over VQ. Secondly, the performance difference between LDA and RSM can be attributed to the latter's distributed representations. In LDA, each word in a turn is assigned to a single topic, while, in RSM, each word is modeled by multiple topics. This allows each topic in the latter to define elementary features and their combination to give rise to more complex and richer representations. Combined with a lower complexity of inference, RSM-based approaches are more suited for tasks involving real-time recognition. Thirdly, learning topics in a supervised manner is highly beneficial, as evident from the improvements obtained using sRSM over competing LTMs on both databases. Finally, except for arousal-based binary classification on the test set of SEMAINE, each LTM outperforms turn-level statistics, i.e. IS09. These improvements are significantly higher for regression and valence-based classification over the spontaneous expressions of SEMAINE, suggesting that the co-occurrence information captured by the topics is highly representative of the underlying emotional content.

Effect of turn duration

As a result of the turn-based segmentation procedure, the duration of a turn varies depending on the speaker's activity. Turns are often long and and may consist of multiple emotions expressed in varying degrees and no seemingly regular structure. Consider, for example, neutral speech with occasional bursts of emotional activity. Here, we describe the experiments conducted to examine the behavior of features with respect to the turn duration. To this extent, we split all the turns in three categories based on their duration: $<$1.5 s, 1.5 to 6 s, and $>$6 s. The UA recall over each category is used to compare the behavior of IS09 and sRSM-based features.

For USC IEMOCAP, there are 408, 3,832, and 1,291 turns in each category, respectively. The class-wise accuracy across the four emotion categories and their average is shown in Figure 7. The relative improvement from the shortest to the longest duration is 7.87% for IS09 while 14.38% for sRSM. The absolute difference in UA recall between sRSM and IS09 for turns less than 1.5 s is $-$0.04%, whereas the difference for turns longer than 6 s is 6.4%. Emotions such as sad, happy, and angry are recognized with a higher accuracy as the duration increases, yet their accuracy is surprisingly low for shorter duration turns. This is probably due to the unavailability of enough sad/happy/angry examples with shorter durations. For instance, of the 408 turns with duration less than 1.5 s, 47% are neutral.

The decline in recall rate of neutral speech, for either feature set, as the duration increases is particularly interesting, since a similar trend is not evident from the ground truth labels provided by human evaluators. The percentage of turns for which there is complete agreement for the three duration categories shows an increasing trend - 15.10%, 18.63%, and 28.42%. Figure 8 shows the average posterior probability for all the misclassified, neutral utterances in USC IEMOCAP across the three duration categories. We can observe that these utterances tend to be misclassified as either happy or sad. Yet, the neutral content is captured as a secondary or minor emotion with slightly lower probability estimates. One may consider the emotional profile (EP) framework presented in [11] as a possible solution, which allows us to account for major-minor emotions in order to resolve such ambiguity. This framework is independent of the type of features or classifiers used and can be easily combined with the proposed approach.

For SEMAINE, there are 350, 373, and 237 turns in each duration category for the development set and 347, 219, and 107 for the test set. The results for arousal and valence classification are shown in Tables 4 and 5, respectively. For

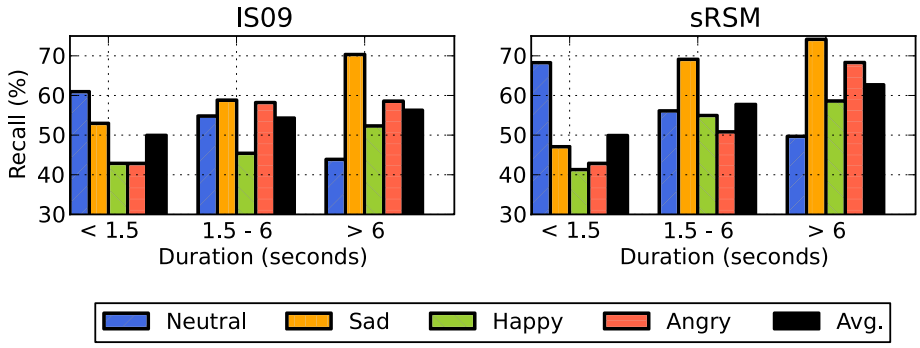

Figure 7 Effect of turn duration. A comparison between the recall obtained using IS09 and sRSM for turns of different durations in USC IEMOCAP. Note the increase and relatively better performance of sRSM as the duration of a turn increases from less than 1.5 s to greater than 6 s.

arousal, IS09 and sRSM are quite similar as one outperforms the other, either on the development set or the test set. On the other hand, sRSM achieves a significant gain over IS09 for valence discrimination - 26.9% and 35% on the development and test set, respectively.

Experimental results on both, USC IEMOCAP and SEMAINE, indicate that sRSM, and, in general, LTMs are better suited to handle turns of longer durations. The extraction of turn-level statistics loses important local information, such as bursts of emotional activity, as the frame-level features are normalized over the turn. LTMs, in spite of generating turn-level descriptors, capture some local information from the word occurrences. The necessity for retaining such information is particularly relevant for valence-based discrimination, where sRSM demonstrates a significantly better recall over all turns and an even further improvement over turns longer than 6 s.

Cross-corpus emotion recognition

Speaker-independent, within-corpus evaluations are useful for preliminary validation of an approach. However, real-world scenarios involve cases where the data does not belong to the same domain as the one used for training the system. For example, changes in elicitation techniques (acted vs. spontaneous), language, culture, accent, etc. are quite common. Cross-corpus evaluations, in such cases, can provide a more reliable measure of how well the approach generalizes across such differences.

The same databases, USC IEMOCAP and SEMAINE, are used in our experiments. We first preprocess the data to compensate for differences in recording conditions. Various methods, such as z-normalization [31] or min-max normalization [32], have been applied at the speaker and corpus level for this purpose. We follow a corpus normalization approach, where we normalize the frame-level features of the training and test corpus to have the same mean. Accordingly, if M_{train} and M_{test} are the respective mean vectors of the training and test corpus, then each frame of the test corpus is multiplied by $M_{\text{train}}/M_{\text{test}}$. After normalization, we obtain acoustic words and topics according to the dictionary and topic models learnt over the training corpus. Secondly, to

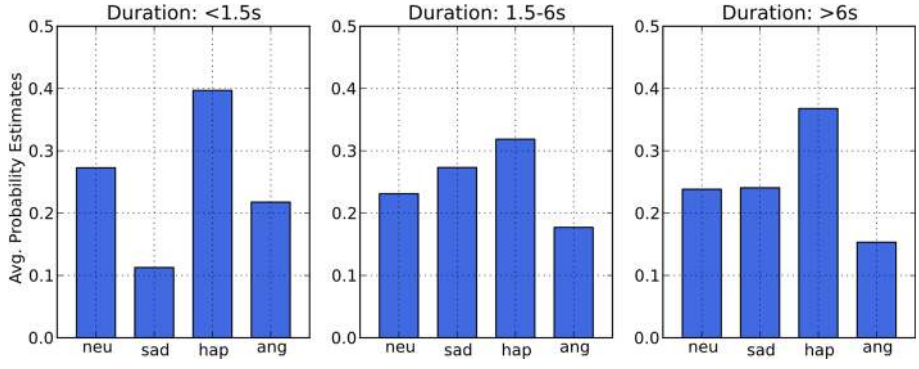

Figure 8 Error analysis of neutral utterances across different duration categories using sRSM. The average posterior probability estimates for all misclassified neutral utterances in USC IEMOCAP. Although utterances tend to get misclassified as happy or sad, the neutral content is captured as the secondary emotion with slightly lower probability estimates. This further indicates the presence of multiple emotions and ambiguity as the turns become longer in duration.

Table 4 Effect of turn duration for arousal-based classification on SEMAINE

Features	< 1.5	1.5 to 6	>6
		Development set	
IS09	54.78	51.41	62.85
sRSM	70.32*	63.46*	65.72
		Test set	
IS09	54.93	60.65	76.31
sRSM	64.92*	56.52	73.01

Results are expressed in percentage (%). Symbol $*$ indicates statistical significance over IS09.

ensure a valid comparison, the labels must be the same across each corpus. As described earlier, the turns of USC IEMOCAP were labeled categorically, while those of SEMAINE were labeled with binary arousal/valence attributes. The four categories of USC IEMOCAP are converted to binary, arousal (low - neutral/sad, high - happy/angry), and valence (negative - sad/angry, positive - neutral/happy) attributes.

We denote the topics extracted over all turns of the training and test corpus as $\mathbf{h}^{\text{trn}}$ and $\mathbf{h}^{\text{tst}}$, respectively. The rows of $\mathbf{h}$ correspond to turns while the columns to topics. For the test corpus, we further split $\mathbf{h}^{\text{tst}}$ into two disjoint partitions - (1) $\mathbf{h}^{\text{tst},l}$, a set of turns with labels $y^{\text{tst},l}$; and (2) $\mathbf{h}^{\text{tst},u}$, the set of unlabeled turns for which we wish to predict labels. When SEMAINE is designated as the test corpus, $h^{\text{tst},l}$ corresponds to the features extracted from the training set of SEMAINE, while $h^{\text{tst},u}$ corresponds to the test set. Alternatively, when USC IEMOCAP is designated as the test corpus, $h^{\text{tst},l}$ corresponds to a set comprising of nine out of ten speakers, while $h^{\text{tst},u}$ corresponds to the remaining speaker. This process is repeated for each of the ten speakers, resulting in a 10-fold process.

According to conventional cross-corpus experiments conducted earlier [31,51], the test corpus is evaluated using the parameters, $\theta^{\star}$, of the classifier learnt over the training corpus as per Equation 13, where L is the cost function. These works do not account for the fact that emotions are perceived differently across geographical regions or cultures causing the annotations to be biased to their respective databases. In other words, even if the definition of labels are same across corpora, there is a significant difference between $p(y^{\text{trn}}|\mathbf{h}^{\text{trn}})$ and $p(y^{\text{tst}}|\mathbf{h}^{\text{tst}})$. Hence, the decision boundary learnt over the training corpus is no longer optimal for the test corpus. The results obtained in this case also indicate the joint performance loss due to both, the features and the classifier.

Table 5 Effect of turn duration for valence-based classification on SEMAINE

Features	<1.5	1.5 to 6	>6
		Development set	
IS09	53.56	52.53	52.74
sRSM	58.26	67.93*	66.95*
		Test set	
IS09	50.50	53.09	43.96
sRSM	57.75	55.29	59.36*

Results are expressed in percentage (%). Symbol $*$ indicates statistical significance over IS09.

$$\theta^{\star} = \underset{\theta}{\text{argmin}} \frac{1}{N} \sum_{i=1}^{N} L\left(y_i^{\text{trn}}, h_i^{\text{trn}}; \theta\right) + \frac{\lambda}{2}\|\theta\|_2^2 \tag{13}$$

In order to improve the cross-corpus performance and determine the generalization of solely the topic features, we propose two approaches to compensate for this bias. In each of these approaches, we will assume that a few labeled turns from the test corpus are available, i.e. $\mathbf{h}^{\text{tst},l}$, and that we can learn parameters $\hat{\theta}$ characterizing $p\left(y^{\text{tst},l}|\mathbf{h}^{\text{tst},l}\right)$. Using $\hat{\theta}$ as a guide, we then learn new parameters from the training corpus such that the decision boundary changes to reflect the distribution of the test corpus.

Instance selection

In this approach, we identify instances in the training corpus that are not modeled well according to $p\left(y^{\text{tst},l}|\mathbf{h}^{\text{tst},l}, \hat{\theta}\right)$. Such instances can be viewed as misleading or confusing, hence removing them would serve to bring $p\left(y^{\text{tst}}|\mathbf{h}^{\text{tst}}, \theta\right)$ closer to $\left(y^{\text{tst},l}|\mathbf{h}^{\text{tst},l}, \hat{\theta}\right)$. Accordingly, we first evaluate $\mathbf{h}^{\text{trn}}$ on $\hat{\theta}$. We then select the top k or all instances that are correctly classified and assign a large weight to these instances, while we assign a smaller weight to the wrongly classified instances. The new parameters $\theta^{\star}$ are now estimated via Equation 14, where α_i indicates the weight assigned to each instance. We follow a simple procedure to set the weights: $\alpha_i = 1$ if correct, else $\alpha_i = 0$.

$$\theta^{\star} = \underset{\theta}{\text{argmin}} \frac{1}{N} \sum_{i=1}^{N} \alpha_i L\left(y_i^{\text{trn}}, h_i^{\text{trn}}; \theta\right) + \frac{\lambda}{2}\|\theta\|_2^2 \tag{14}$$

Weight regularization

The differences between $p\left(y^{\text{trn}}|\mathbf{h}^{\text{trn}}, \theta\right)$ and $p\left(y^{\text{tst},l}|\mathbf{h}^{\text{tst},l}, \hat{\theta}\right)$ can alternatively be explained by the difference in their weights θ and $\hat{\theta}$. In traditional $L2$-norm regularization, i.e. Equations 13 and 14, we penalize the weights from becoming too large. If instead, we penalize the difference, $\|\hat{\theta} - \theta\|_2^2$, from being large, then we will

effectively learn weights $\theta \rightarrow \hat{\theta}$. Parameters $\theta^{\star}$, in this case, are learnt as per Equation 15.

$$\theta^{\star} = \underset{\theta}{\operatorname{argmin}} \frac{1}{N} \sum_{i=1}^{N} L\left(y_i^{\text{trn}}, h_i^{\text{trn}}; \theta\right) + \frac{\lambda}{2} \|\hat{\theta} - \theta\|_2^2 \quad (15)$$

Results and discussion

The best within-corpus (WC) results are obtained as per our experiments described in the earlier section. In addition to different methods used to elicit emotions, USC IEMOCAP and SEMAINE are also recorded and annotated using subjects belonging to different cultures. The former comprises of American speakers, while the latter comprises of speakers from eight countries across Europe. These factors affect the performance such that the cross-corpus recall will, in general, be lower than under a within-corpus setting [31].

The UA recall obtained using a conventional cross-corpus strategy without adaptation for different LTMs and train/test scenarios are shown in Figure 9. For LDA, sLDA, RSM, and sRSM, the average deterioration across all the scenarios is 11.1% ± 3.3%, 6.6% ± 3.1%, 8.7% ± 3.3%, and 5.2% ± 3.4%, respectively. The recall values for cross-corpus using instance selection are presented in Figure 10. In this case, the average deterioration across all the scenarios is 8.1% ± 1.7%, 5.1% ± 1.7%, 5.3% ± 2.4%, and 2.6% ± 1.8 for LDA, sLDA, RSM, and sRSM, respectively. Similar results for cross-corpus using weight regularization are shown in Figure 11. The average deterioration across all the scenarios is 7.5% ± 4.1%, 5.8% ± 3.6%, 5.3% ± 3.9%, and 2.7% ± 2.3% for LDA, sLDA, RSM, and sRSM, respectively.

The improvements demonstrated by either adaptation strategy over a conventional approach confirm the existence of a classifier-specific bias due to varying perceptions across corpora. Adaptation successfully reduces this bias by using parameters ($\hat{\theta}$) as a reference during learning. Between the two approaches, the mean deterioration is almost similar for both instance selection and weight regularization; however, the latter has a comparatively larger standard deviation, thus making the former a more suitable approach. We further performed experiments to combine the two approaches, but we did not obtain any improvements.

When the spontaneous expressions of SEMAINE are evaluated over the acted expressions of USC IEMOCAP, the relative deterioration using sRSM with instance selection and weight regularization is 1.0% ± 0.5% and 1.1% ± 0.8%, respectively, compared to 4.2% ± 1.1% and 4.4% ± 1.9%, respectively, for the vice versa case. This can mainly be attributed to the number of examples available for training the classifier; USC IEMOCAP is approximately five times larger than SEMAINE. Between arousal and valence-based classification, we observe that the deterioration is more severe for the latter case. Using sRSM with weight regularization and instance selection, a relative deterioration of 1.8% ± 1.3% and 1.4% ± 1.1%, respectively, is obtained for arousal. Whereas, a relative deterioration of 3.5% ± 1.9% and 4.2% ± 2.3%, respectively, is obtained for valence. Again, the inherent limitations of speech coupled with the differing perceptions of valence across cultures and geographical regions possibly account for this loss. This phenomenon was also observed in a previous cross-corpus study conducted over different databases [31]. There are no previous reports of cross-corpus studies over the two databases used in this study.

Between different LTMs, the supervised LTMs outperform their unsupervised counterparts as observed in a within-corpus setting. sRSM, once again, achieves the least deterioration across all train/test scenarios and adaptation approaches. In case of LDA and RSM, the topics learnt initially over the training corpus remain unchanged

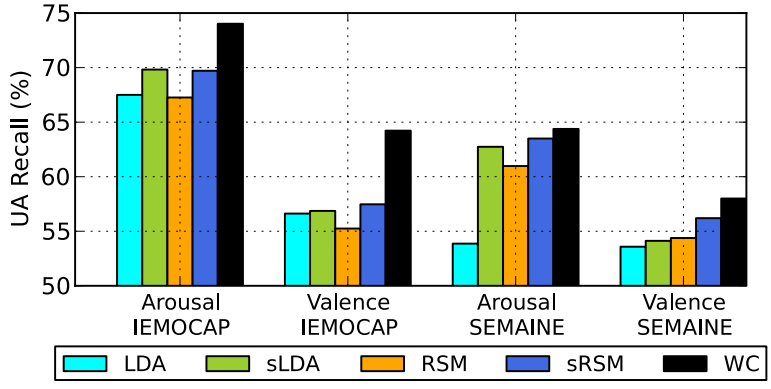

Figure 9 Cross-corpus recall without adaptation. Horizontal axis indicates the classification task and test corpus. The figure shows a detailed comparison between four different LTMs along with the best within-corpus (WC) recall in each case.

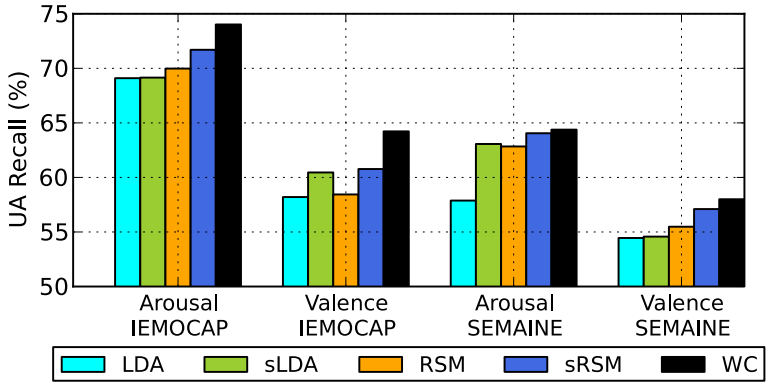

Figure 10 Cross-corpus recall with instance selection. Horizontal axis indicates the classification task and test corpus. The figure shows a detailed comparison between four different LTMs along with the best within-corpus (WC) recall in each case. There is a significant improvement in performance as opposed to a conventional approach without instance selection, i.e. Figure 9.

and only the top-level classifier is modified. Although the results show a relatively higher deterioration compared to sRSM or sLDA, they single out the performance loss due to the topic-based features alone and are indeed promising.

Conclusions

In this work, we proposed a novel approach for the extraction of turn-level features using latent topic models for speech emotion recognition. This is the first work to draw similarities between text documents and emotional speech; we showed that the latter can be viewed as a mixture of multiple emotion-specific topics, where the topics capture salient information from the co-occurrence patterns of frame-level features. We considered two fundamentally different models, LDA and RSM, and their supervised counterparts for the purpose of generating topic-based features. Specifically, sRSM, which treats the RSM as a pre-training stage followed by fine-tuning via backpropagation, was proposed to learn features that are optimal for discriminative tasks.

The proposed features were evaluated on different types of emotional expressions and output representations, outperforming state-of-the-art methods in each case. On the acted emotions of USC IEMOCAP, sRSM obtained a relative improvement of 7% compared to turn-level statistics collected via brute force. Whereas on the spontaneous expressions of SEMAINE, sRSM obtained an improvement of 16.75% for valence-based classification, which is quite significant considering the well-known difficulty of valence discrimination using only speech information. With respect to the turn duration, we showed that sRSM and in general, LTMs, are better suited for longer turns (>6 s), which is highly desirable for current turn-based practices. The improvement over turn-level statistics for valence-based classification is particularly

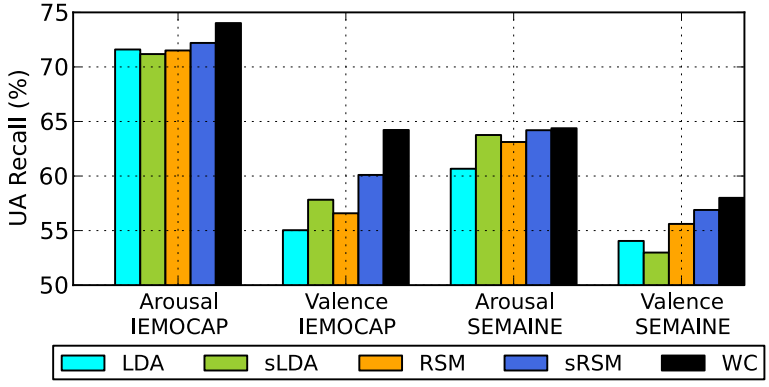

Figure 11 Cross-corpus recall with weight regularization. Horizontal axis indicates the classification task and test corpus. The figure shows a detailed comparison between four different LTMs along with the best within-corpus (WC) recall in each case. There is a significant improvement in performance as opposed to a conventional approach with standard L2 regularization, i.e. Figure 9.

significant - 26% and 35% on the development and test sets of SEMAINE, respectively.

In a cross-corpus setting, we showed that classifiers are inherently biased because of the annotation procedures and cultural perceptions specific to each corpus, which leads to poor generalization. To compensate for this bias and improve cross-corpus performance, we proposed two novel adaptation-based approaches. Compared to the best within-corpus performance, sRSM showed the least relative deterioration of only 2.6% and 2.7% using instance selection and weight regularization, respectively. This further highlights that the proposed features can efficiently generalize across different accents, speakers, and elicitation types (acted vs. spontaneous).

Qualitative aspects of the features were investigated using a normalized point-wise mutual information measure between topics and emotions. Our analyses revealed the emotions to be naturally and well separated in the topic space. Topics ranked higher for one emotion received a lower rank for other emotions, further demonstrating that the co-occurrence information captured by topics is strongly related to the underlying emotion, thus offering a novel, generative-model-based interpretation of how emotions influence the observed speech characteristics.

Finally, we comment on the flexibility of the proposed approach. Although energy, F0, and MFCCs were used as frame-level features in this work, words and topics can be derived from other frame-level features or modalities and be combined to decrease the confusion between happy-angry and neutral-sad emotions and lead to further improvements. Similarly, a simple logistic/softmax regression classifier can be replaced by more sophisticated classifiers [15] or alternative tree-based schemes [46] to achieve even better discrimination.

Competing interests
The authors declare that they have no competing interests.

Acknowledgements
This work was supported in part by a grant from NSF CSR 0910699 and the SenSIP center.

Received: 9 June 2014 Accepted: 8 September 2014
Published online: 25 January 2015

References

1. R Cowie, RR Cornelius, Describing the emotional states that are expressed in speech. Speech Commun. **40**(1), 5–32 (2003)
2. CM Lee, SS Narayanan, Toward detecting emotions in spoken dialogs. IEEE Trans. Speech Audio Process. **13**(2), 293–303 (2005)
3. L Vidrascu, L Devillers, in *Proceedings of INTERSPEECH*. Detection of real-life emotions in call centers (ISCA, Lisbon, 2005), pp. 1841–1844
4. S Steidl, Automatic classification of emotion-related user states in spontaneous children's speech (2009)
5. S Narayanan, PG Georgiou, Behavioral signal processing: Deriving human behavioral informatics from speech and language. Proc. IEEE. **101**(5), 1203–1233 (2013)
6. J Přibil, Přibilová A, Evaluation of influence of spectral and prosodic features on GMM, classification of Czech and Slovak emotional speech. EURASIP J. Audio Speech Music Process. **2013**(1), 1–22 (2013)
7. TL Nwe, SW Foo, LC De Silva, Speech emotion recognition using hidden Markov models. Speech Commun. **41**(4), 603–623 (2003)
8. A Metallinou, S Lee, S Narayanan, in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*. Decision level combination of multiple modalities for recognition and analysis of emotional expression (IEEE, Dallas, 2010), pp. 2462–2465
9. El Ayadi MM, MS Kamel, F Karray, in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing,* vol. 4. Speech emotion recognition using Gaussian mixture vector autoregressive models (IEEE, Honolulu, 2007), pp. 954–957
10. CE Williams, KN Stevens, Emotions and speech: Some acoustical correlates. J. Acoust. Soc. Am. **52**(4B), 1238–1250 (2005)
11. E Mower, MJ Mataric, S Narayanan, A framework for automatic human emotion classification using emotion profiles. IEEE Trans. Audio Speech Lang. Process. **19**(5), 1057–1070 (2011)
12. B Schuller, A Batliner, S Steidl, D Seppi, Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. Speech Commun. **53**(9), 1062–1087 (2011)
13. C Oflazoglu, S Yildirim, Recognizing emotion from Turkish speech using acoustic features. EURASIP J. Audio Speech Music Process. **2013**(1), 1–11 (2013)
14. O-W Kwon, K Chan, J Hao, T-W Lee, in *Proceedings of INTERSPEECH*. Emotion recognition by speech signals (ISCA, Geneva, 2003), pp. 125–128
15. B Schuller, A Batliner, D Seppi, S Steidl, T Vogt, J Wagner, L Devillers, L Vidrascu, N Amir, L Kessous, V Aharonson, in *Proceedings of INTERSPEECH*. The relevance of feature type for the automatic classification of emotional user states: low level descriptors and functionals (ISCA, Antwerp, 2007), pp. 2253–2256
16. F Eyben, M Wollmer, B Schuller. International Conference on Affective Computing and Intelligent Interaction and Workshops (IEEE, Amsterdam, 2009), pp. 1–6
17. B Schuller, S Steidl, A Batliner, in *Proceedings of INTERSPEECH*. The INTERSPEECH, 2009 emotion challenge (ISCA, Brighton, 2009), pp. 312–315
18. B Schuller, M Valstar, F Eyben, G McKeown, R Cowie, M Pantic, in *Proceedings of Affective Comput. Intell. Interaction,* vol. 6975. Avec 2011–the first international audio/visual emotion challenge (IEEE, Memphis, 2011), pp. 415–424
19. SC Deerwester, ST Dumais, TK Landauer, GW Furnas, RA Harshman, Indexing by latent semantic analysis. JASIS. **41**(6), 391–407 (1990)
20. T Hofmann, in *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*. Probabilistic latent semantic indexing (ACM, Berkeley, 1999), pp. 50–57
21. DM Blei, Ng, AY, MI Jordan, Latent Dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
22. T Huynh, M Fritz, B Schiele, in *Proceedings of the 10th International Conference on Ubiquitous Computing*. Discovery of activity patterns using topic models (ACM, Seoul, 2008), pp. 10–19
23. N Srivastava, R Salakhutdinov, in *Proceedings of Adv. Neural Inf. Process. Syst,* vol. 15. Multimodal learning with deep Boltzmann machines (NIPS, Lake Tahoe, 2012), pp. 2231–2239
24. D Liu, T Chen, in *IEEE International Conference on Computer Vision*. Unsupervised image categorization and object localization using topic models and correspondences between images (IEEE, Rio de Janeiro, 2007), pp. 1–7
25. M Shah, L Miao, C Chakrabarti, A Spanias, in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*. A speech emotion recognition framework based on latent Dirichlet allocation: Algorithms and FPGA implementation (IEEE, Vancouver, 2013), pp. 2553–2556
26. GE Hinton, R Salakhutdinov, in *Proceedings of Adv. Neural Inf. Process. Syst,* vol. 1. Replicated softmax: an undirected topic model (NIPS, Lake Tahoe, 2009), pp. 1607–1614
27. A Stuhlsatz, J Lippel, T Zielke, Feature extraction with deep neural networks by a generalized discriminant analysis. IEEE Trans. Neural Netw. Learn. Syst. **23**(4), 596–608 (2012)
28. A Stuhlsatz, C Meyer, F Eyben, T Zielke, G Meier, B Schuller, in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*.

Deep neural networks for acoustic emotion recognition: Raising the benchmarks (IEEE, Prague, 2011), pp. 5688–5691
29. S Press, S Wilson, Choosing between logistic regression and discriminant analysis. J. Am. Stat. Assoc. **73**(364), 699–705 (1978)
30. M Pohar, M Blas, S Turk, Comparison of logistic regression and linear discriminant analysis: a simulation study. Metodolski Zvezki. **1**(1), 143–161 (2004)
31. B Schuller, B Vlasenko, F Eyben, M Wollmer, A Stuhlsatz, A Wendemuth, G Rigoll, Cross-corpus acoustic emotion recognition: variances and strategies. IEEE Trans. Affect. Comput. **1**(2), 119–131 (2010)
32. D Neiberg, P Laukka, HA Elfenbein, in *Proceedings of INTERSPEECH*. Intra-, inter-, and cross-cultural classification of vocal affect (ISCA, Florence, 2011), pp. 1581–1584
33. F Eyben, A Batliner, B Schuller, D Seppi, S Steidl, in *Proceedings of the 3rd International Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect*. Cross-corpus classification of realistic emotions some pilot experiments (LREC, Valetta, 2010), pp. 77–82
34. B Schuller, Z Zhang, F Weninger, G Rigoll, in *Proceedings of the 2011 Afeka-AVIOS Speech Processing Conference*. Selecting training data for cross-corpus speech emotion recognition: Prototypicality vs. generalization (ACLP, Tel Aviv, Israel, 2011)
35. C Busso, M Bulut, C-C Lee, A Kazemzadeh, E Mower, S Kim, JN Chang, S Lee, SS Narayanan, IEMOCAP: Interactive emotional dyadic motion capture database. Lang. Resour. Eval. **42**(4), 335–359 (2008)
36. G McKeown, M Valstar, R Cowie, M Pantic, M Schroder, The SEMAINE database: annotated multimodal records of emotionally colored conversations between a person and a limited agent. IEEE Trans. Affect. Comput. **3**(1), 5–17 (2012)
37. M El Ayadi, MS Kamel, F Karray, Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recognit. **44**(3), 572–587 (2011)
38. C Wang, D Blei, F.-F Li, in *IEEE Conference on Computer Vision and Pattern Recognition*. Simultaneous image classification and annotation (IEEE, Miami, 2009), pp. 1903–1910
39. GE Dahl, D Yu, L Deng, A Acero, Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. IEEE Trans. Audio Speech Lang. Process. **20**(1), 30–42 (2012)
40. MA Carreira-Perpinan, GE Hinton, in *Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics*. On contrastive divergence learning (Society for Artificial Intelligence and Statistics NP Barbados, 2005), pp. 33–40
41. T Painter, A Spanias, Perceptual coding of digital audio. Proc. IEEE. **88**(4), 451–515 (2000)
42. A Spanias, T Painter, V Atti, *Audio Signal Processing and Coding*. (John Wiley & Sons, Hoboken, 2006)
43. MC Sezgin, B Gunsel, GK Kurt, Perceptual audio features for emotion detection. EURASIP J. Audio Speech Music Process. **2012**(1), 1–21 (2012)
44. S Young, G Evermann, D Kershaw, G Moore, J Odell, D Ollason, V Valtchev, P Woodland, *The HTK Book*, vol.2. (Entropic Cambridge Research Laboratory, Cambridge, 1997)
45. G Bouma, in *Proceedings of GSCL*. Normalized (pointwise) mutual information in collocation extraction (GSCL, Potsdam, 2009), pp. 31–40
46. C C-Lee, E Mower, C Busso, S Lee, S Narayanan, Emotion recognition using a hierarchical binary decision tree approach. Speech Commun. **53**(9), 1162–1171 (2011)
47. B Schuller, M Valster, F Eyben, R Cowie, M Pantic, in *Proceedings of the 14th ACM International Conference on Multimodal Interaction*. Avec 2012: the continuous audio/visual emotion challenge (ACM, Santa Monica, 2012), pp. 449–456
48. M Hall, E Frank, G Holmes, B Pfahringer, P Reutemann, IH Witten, The WEKA data mining software: an update. ACM SIGKDD Explorations Newsletter. **11**(1), 10–18 (2009)
49. S Pan, J Tao, Y Li, in *Proceedings of Affect. Comput. Intell. Interaction*, vol. 6975. The CASIA audio emotion recognition method for audio/visual emotion challenge 2011 (IEEE, Memphis, 2011), pp. 388–395
50. M Glodek, S Tschechne, G Layher, M Schels, T Brosch, S Scherer, M Kächele, M Schmidt, H Neumann, G Palm, F Schwenker, in *Proceedings of Affect. Comput. Intell. Interaction*, vol. 6975. Multiple classifier systems for the classification of audio-visual emotional states (IEEE, Memphis, 2011), pp. 359–368
51. L Devillers, C Vaudable, C Chastagnol, in *Proceedings of INTERSPEECH*. Real-life emotion-related states detection in call centers: a cross-corpora study (ISCA, Makuhari, 2010), pp. 2350–2353