Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 52 (2010) 613–625
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

# Class-level spectral features for emotion recognition ☆

Dmitri Bitouk [a,*], Ragini Verma [a], Ani Nenkova [b]

[a] Department of Radiology, Section of Biomedical Image Analysis, University of Pennsylvania, 3600 Market Street, Suite 380, Philadelphia, PA 19104, United States
[b] Department of Computer and Information Science, University of Pennsylvania, 3330 Walnut Street, Philadelphia, PA 19104, United States

Received 24 November 2009; received in revised form 2 February 2010; accepted 9 February 2010

**Abstract**

The most common approaches to automatic emotion recognition rely on utterance-level prosodic features. Recent studies have shown that utterance-level statistics of segmental spectral features also contain rich information about expressivity and emotion. In our work we introduce a more fine-grained yet robust set of spectral features: statistics of Mel-Frequency Cepstral Coefficients computed over three phoneme type classes of interest – stressed vowels, unstressed vowels and consonants in the utterance. We investigate performance of our features in the task of speaker-independent emotion recognition using two publicly available datasets. Our experimental results clearly indicate that indeed both the richer set of spectral features and the differentiation between phoneme type classes are beneficial for the task. Classification accuracies are consistently higher for our features compared to prosodic or utterance-level spectral features. Combination of our phoneme class features with prosodic features leads to even further improvement. Given the large number of class-level spectral features, we expected feature selection will improve results even further, but none of several selection methods led to clear gains. Further analyses reveal that spectral features computed from consonant regions of the utterance contain more information about emotion than either stressed or unstressed vowel features. We also explore how emotion recognition accuracy depends on utterance length. We show that, while there is no significant dependence for utterance-level prosodic features, accuracy of emotion recognition using class-level spectral features increases with the utterance length.
© 2010 Elsevier B.V. All rights reserved.

*Keywords:* Emotions; Emotional speech classification; Spectral features

## 1. Introduction

Emotion content of spoken utterances is clearly encoded in the speech signal, but pinpointing the specific features that contribute to conveying emotion remains an open question. Descriptive studies in psychology and linguistics have mostly dealt with prosody, concerned with the question *how* an utterance is produced. They have identified a number of acoustic correlates of prosody indicative of given emotions. For example, happy speech has been found to be correlated with increased mean fundamental frequency ($F0$), increased mean voice intensity and higher variability of $F0$, while boredom is usually linked to decreased mean $F0$ and increased mean of the first formant frequency ($F1$) (Banse and Scherer, 1996). Following this tradition, most of the work on automatic recognition of emotion has made use of *utterance-level statistics* (mean, min, max, std) of prosodic features such as $F0$, formant frequencies and intensity (Dellaert et al., 1996; McGilloway et al., 2000). Others employed Hidden Markov Models (HMM) (Huang and Ma, 2006; Fernandez and Picard, 2003) to differentiate the type of emotion expressed in an utterance based the prosodic features in a sequence of frames, thus avoiding the need to compute utterance-level statistics.

☆ This paper is an expanded version of the work first presented in Interspeech 2009 (Bitouk et al., 2009). This paper includes additional material on feature selection and feature analysis, experiments linking utterance length and system performance, speaker-dependent recognition and expanded discussion of the related work.

* Corresponding author.
*E-mail address:* Dmitri.Bitouk@uphs.upenn.edu (D. Bitouk).

0167-6393/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2010.02.010

On the other hand, spectral features, based on the short-term power spectrum of sound, such as Linear Prediction Coefficients (LPC) and Mel-Frequency Cepstral Coefficients (MFCC), have received less attention in emotion recognition. While spectral features are harder to be intuitively correlated with affective state, they provide a more detailed description of speech signal and, thus, can potentially improve emotion recognition accuracy over prosodic features. However, spectral features, which are typically used in speech recognition, are segmental and convey information on both *what* is being said and *how* it is being said. Thus, the major challenge in using spectral information in emotion analysis is to define features in a way that does not depend on the specific phonetic content of an utterance, while preserving cues for emotion differentiation.

Most of the previous methods that do use spectral features ignore this challenge by modelling how emotion is encoded in speech independent of its phonetic content. Phoneme-level classification of emotion has received relatively little attention, barring only a few exceptions. For example the work of Lee et al. (2004) takes into account phonetic content of speech by training phoneme-dependent HMM for speaker-dependent emotion classification. Sethu et al. (2008) used phoneme-specific Gaussian Mixture Models (GMM) and demonstrated that emotion can be better differentiated by some phonemes than others. However, such phoneme-specific approach cannot be directly applied to emotion classification due to sparsity of phoneme occurance.

In this paper, we present novel spectral features for emotion recognition computed over phoneme type classes of interest: stressed vowels, unstressed vowels and consonants in the utterance. These larger classes are general enough and do not depend on specific phonetic composition of the utterance and thus abstract away from what is being said. Unlike previous approaches which used spectral features, our class-level spectral features are technically simple and exploit linguistic intuition rather than rely on sophisticated machine learning machinery.

We use the forced alignment between audio and the manual transcript to obtain the phoneme-level segmentation of the utterance and compute statistics of MFCC from parts of the utterance corresponding to the three phoneme classes. Compared to previous approaches which use utterance-level statistics of spectral features, the advantage of our approach is two-fold. Firstly, the use of phoneme classes reduces dependence of the extracted spectral features on the phonetic content of the utterance. Secondly, it captures better the intuition that emotional affect can be expressed to a greater extent in some phoneme classes than others, and, thus, increases the discriminating power of spectral features.

In our work we analyze performance of phoneme class spectral features in speaker-independent emotion classification of English and German speech using two publicly available datasets (Section 5). We demonstrate that class-level spectral features outperform both the traditional prosodic features and utterance-level statistics of MFCC. We test several feature selection algorithms in order to further improve emotion recognition performance of class-level spectral features and evaluate contributions from each phoneme class to emotion recognition accuracy (Section 7). Our results indicate that spectral features, computed from the consonant regions of the utterance, outperform features from both stressed and unstressed vowel regions. Since consonant regions mostly correspond to unvoiced speech segments which are not accounted for by prosodic features derived from pitch and intensity profiles, this result implies that class-level spectral features can provide complimentary information to both utterance-level prosodic and spectral features.

Finally, cross-corpus comparisons of emotion recognition motivated an analysis of the impact of utterance length on classification accuracy which, to the best of our knowledge, has not been addressed in the literature (Section 6). We investigate this dependence using synthetic emotional speech data constructed by concatenating short utterances from LDC dataset. We demonstrate that, while there is no significant dependence for utterance-level prosodic features, performance of class-level spectral features increases with utterance length, up to utterance length of 16 syllables. Further increases in utterance length do not seem to affect performance. It should be noted that these results are obtained using concatenated emotional speech and need to be cross-validated on naturally spoken emotional corpora when appropriate corpora become available.

## 2. Prior work

Although the main body of previous work on emotion recognition in speech uses suprasegmental prosodic features, segmental spectral features which are typically employed in automatic speech recognition have also been studied for the task. The most commonly used spectral features for emotion recognition are Mel-Frequency Cepstral Coefficients (MFCC) (Tabatabaei et al., 2007; Lee et al., 2004; Kwon et al., 2003; Neiberg et al., 2006; Schuller et al., 2005; Luengo et al., 2005; Hasegawa-Johnson et al., 2004; Vlasenko et al., 2007; Grimm et al., 2006; Schuller and Rigoll, 2006; Meng et al., 2007; Sato and Obuchi, 2007; Kim et al., 2007; Hu et al., 2007; Shafran et al., 2003; Shamia and Kamel, 2005; Vlasenko et al., 2008; Vondra and Vich, 2009; Wang and Guan, 2005). As in automatic speech recognition, MFCC are extracted using a 25 ms Hamming window at intervals of 10 ms and cover frequency range from 300 Hz to the Nyquist frequency.

In addition to MFCC, the log-energy as well as delta and acceleration coefficients (first and second derivatives) are also used as features. A low-frequency version of MFCC (Neiberg et al., 2006) which uses low-frequency filterbanks in 20–300 Hz range has been found not to provide emotion recognition performance gains. Other spectral fea-

ture types used for emotion recognition are Linear Prediction Cepstral Coefficients (LPC) (Nicholson et al., 2000; Pao et al., 2005), Log Frequency Power Coefficients (Nwe et al., 2003; Song et al., 2004) and Perceptual Linear Prediction (PLP) coefficients (Scherer et al., 2007; Ye et al., 2008). Prior approaches which used spectral features for emotion recognition in speech are summarized in Table 1.

The majority of spectral methods for emotion recognition make use of either *frame-level* or *utterance-level* features. Frame-level approaches model how emotion is encoded in speech using features sampled at small time intervals (typically 10–20 ms) and classify utterances using either HMMs or by combining predictions from all of the frames. On the other hand, utterance-level methods rely on computing statistical functionals of spectral features over the entire utterance.

### 2.1. Frame-level methods

HMMs have been applied with great success in automatic speech recognition to integrate frame-level information, and can be used similarly for emotion recognition as well. One group of HMM-based methods models all utterances using a fixed HMM topology independent of what is being said. In this case, each emotion is represented using its own HMM. Ergodic HMM topology is often used in order to accommodate varying utterance length.

Unlike automatic speech recognition in which HMM states usually correspond to sub-phoneme units, there is no clear interpretation for the states of emotion-level HMM which are employed as a mean to integrate frame-level information into a likelihood score for each emotion. For example, Nwe et al. (2003) trained speaker-dependent, four-state ergodic HMMs for each emotion. To classify a novel utterance into an emotion category, likelihood scores of the utterance features given each emotion were evaluated using the trained HMMs. The utterance is classified as expressing the emotion which yields the highest likelihood score. This method achieved 71% accuracy in classifying the six basic emotions in speaker-dependent settings, but its speaker-independent performance was not investigated. Song et al. (2004) proposed a straightforward extension of this approach to estimate three discrete emotion intensity levels by effectively treating each emotion's intensity levels as a separate category. Another group of HMM-based approaches aims to integrate emotion labels into automated speech recognition systems. This is typically accomplished by building HMMs with emotion-dependent states. Meng et al. (2007) proposed joint speech and emotion recognition by expanding the dictionary to include multiple versions of each word, one for each emotion. Emotion classification was then performed using majority voting between emotion labels in the hypothesis obtained using standard decoding algorithms. A similar emotion-dependent HMM approach was also used by Hasegawa-Johnson et al. (2004) to differentiate between confidence, puzzle and hesitation affective states in an intelligent tutoring application. Lee et al. (2004) train phoneme-dependent HMMs in order to take into account phonetic content of speech. During the training stage, emotion-dependent HMMs were constructed for each of the five phoneme classes – vowels, glides, nasals, stops and fricatives. In order to classify an utterance, its likelihood scores given each emotion were computed and the emotion with the maximum likelihood score was chosen as the decision. However, this approach was only tested for speaker-dependent emotion recognition using a proprietary database which consisted of recording from a single speaker.

Another popular approach to emotion recognition at frame-level is to ignore temporal information altogether and treat acoustic observations at each time frame as the values of independent, identically distributed random variables. Under this assumption, Gaussian Mixture Models (GMMs) are commonly used to model conditional distributions of acoustic features in the utterance given emotion categories. Neiberg et al. (2006) used GMMs trained on the extracted MFCC and pitch features to classify utterances into neutral, negative and positive emotion categories in call center and meeting datasets. Luengo et al. (2005) also employed GMMs to classify utterances from a single speaker database into the six basic emotions. A real-time systems for discriminating between angry and neutral speech was implemented in Kim et al. (2007) using GMMs for MFCC features in combination with a prosody-based classifier. Vondra and Vich (2009) applied GMMs to emotion recognition using a combined feature set obtained by concatenating MFCC and prosodic features. Hu et al. (2007) employed the GMM supervector approach in order to extract fixed-length feature vectors from utterances with variable durations. A GMM supervector consists of the estimated means of the mixtures in GMM. A mixture model was trained for each utterance and GMM supervectors were used as features for support vector machine classifiers. Frame-wise emotion classification based on vector quantization techniques was used by Sato and Obuchi (2007). In the training stage, a set of codewords was obtained for each emotion. In order to classify an input utterance, an emotion label was computed for each frame by finding the nearest emotion codeword. Finally, the whole utterance was classified using a majority voting scheme between frame-level emotion labels. It was demonstrated that such a simple frame-wise technique outperforms HMM-based methods. Vlasenko et al. (2007) integrated GMM log-likelihood score with commonly-used suprasegmental prosody-based emotion classifiers in order to investigate combination of features at different levels of granularity. Sethu et al. (2008) used phoneme-specific GMMs and demonstrated that emotion can be better differentiated by some phonemes than others. However, such phoneme-specific approach cannot be directly applied to emotion classification due to sparsity of phoneme occurance.

Table 1
Prior work on emotion recognition in speech using spectral features.

| Reference | Language | Spectral features | Granularity | Use of prosody | Classifier | Speaker independence |
|---|---|---|---|---|---|---|
| Neiberg et al. (2006) | Swedish, English | MFCC | Frame | ✓ | GMM | |
| Luengo et al. (2005) | Basque | MFCC | Frame | ✓ | HMM | |
| Hasegawa-Johnson et al. (2004) | English | MFCC | Frame | ✓ | HMM | |
| Meng et al. (2007) | German | MFCC | Frame | | HMM | |
| Sato and Obuchi (2007) | English | MFCC | Frame | | Voting | ✓ |
| Kim et al. (2007) | English | MFCC | Frame | ✓ | GMM | |
| Hu et al. (2007) | Mandarin | MFCC | Frame | | GMM, SVM | |
| Shafran et al. (2003) | English | MFCC | Frame | ✓ | HMM | |
| Vondra and Vich (2009) | German | MFCC | Frame | ✓ | GMM | ✓ |
| Pao et al. (2005) | Mandarin | LPC, MFCC ... | Frame | | HMM, kNN | ✓ |
| Nwe et al. (2003) | Burmese, Mandarin | LFPC | Frame | | HMM | |
| Song et al. (2004) | Mandarin | LFPC | Frame | | HMM | |
| Scherer et al. (2007) | German | PLP | Frame | | kNN | |
| Vlasenko et al. (2007) | German, English | MFCC | Frame | ✓ | GMM | ✓ |
| | | | Utterance | | SVM | |
| Sethu et al. (2008) | English | MFCC | Frame | ✓ | GMM | ✓ |
| Lee et al. (2004) | English | MFCC | Phoneme | | HMM | |
| Tabatabaei et al. (2007) | English | MFCC | Utterance | ✓ | SVM | |
| Kwon et al. (2003) | English, German | MFCC | Utterance | ✓ | SVM, LDA | ✓ |
| Schuller et al. (2005) | German | MFCC | Utterance | ✓ | SVM, AdaBoost | ✓ |
| Grimm et al. (2006) | English | MFCC | Utterance | ✓ | Fuzzy logic | |
| Wang and Guan (2005) | Multiple | MFCC | Utterance | ✓ | LDA | |
| Ye et al. (2008) | Mandarin | MFCC, PLP | Utterance | | SVM | |
| Schuller and Rigoll (2006) | German | MFCC | Segment | ✓ | Various | |
| Shamia and Kamel (2005) | English | MFCC | Segment | ✓ | kNN, SVM | |
| Nicholson et al. (2000) | Japanese | LPC | Segment | ✓ | NN | |
| **This paper** | English, German | MFCC | Phoneme | ✓ | SVM | ✓ |

### 2.2. Utterance-level methods

In contrast to frame-level approaches, utterance-level methods rely on extracting fixed-length feature vectors. Such features are usually composed of various statistics of acoustic parameters computed over the entire utterance. Commonly-used statistics are mean, standard deviation, skewness and extrema values. In utterance-level emotion recognition, statistics of spectral features are often combined with statistics of prosodic measures, and classification is performed using both sources of information. For example, Kwon et al. (2003) used statistics such as mean, standard deviation, range and skewness of pitch, energy and MFCC to recognize emotions using Support Vector Machine (SVM) classifiers. Similarly, 276 statistical functionals of pitch, energy, MFCC and voice quality measures along with linguistic features were used in Schuller et al. (2005). Instead of computing independent statistics, Ye et al. (2008) used covariance matrices of prosodic and spectral measures evaluated over the entire utterance. Since positive-definite covariance do not form a vector space, classification has to be performed using manifold learning methods. Schuller and Rigoll (2006) investigated levels of granularity finer than the entire utterance. In particular, they demonstrated that statistics of spectral and prosodic features computed over speech segments obtained by splitting utterances at fixed relative positions (such as halves and thirds) can improve recognition performance over the utterance-level features. Shamia and Kamel (2005) computed prosodic and spectral features for each voiced segment of the utterance and constructed segment-level emotion classifier. In order to classify an utterance, the posterior class probability was evaluated by combining posterior probabilities of each voiced segment in the utterance.

## 3. Databases

In our work we used two publicly available databases of emotional speech: an English emotional speech database from Linguistic Data Consortium (LDC) (2002) and Berlin database of German emotional speech (Burkhardt et al., 2005).

### 3.1. LDC emotional speech database

The LDC database contains recording of native English actors expressing the following 15 emotional states: *neutral, hot anger, cold anger, happy, sadness, disgust, panic, anxiety (fear), despair, elation, interest, shame, boredom, pride* and *contempt*. In addition, utterances in LDC database vary by the distance between the speaker and the listener: tet a tet, conversation and distant. In our experiments we consider only the utterances corresponding to "conversation" distance to the listener and the six basic emotions which include *anger* (*hot anger*), *fear, disgust, happy, sadness*

and *neutral*. There are 548 utterances from seven actors (three female/four male) corresponding to these six basic emotions from LDC database. Almost all of the extracted utterances are short, four-syllable utterances containing dates and numbers (e.g. "Nineteen hundred"). Note that emotion labels of utterances in LDC database were not validated by external subject assessment as done in the creation of other databases (such as Berlin Emotional Speech Database). Instead, each utterance is simply labeled as the intended emotion given to the speaker at the time of recording. As a result, some of the utterances might not be that good examples of the intended emotion and could even be perceived by listeners as expressing a different emotion altogether. Such characteristics of the corpus are likely to negatively affect emotion recognition accuracy.

### 3.2. Berlin emotional speech database

Berlin dataset contains emotional utterances produced by 10 German actors (five female/five male) reading one of 10 pre-selected sentences typical of everyday communication ("She will hand it in on Wednesday", "I just will discard this and then go for a drink with Karl", etc). The dataset provides examples of the following seven emotions: *anger, boredom, fear, disgust, joy (happy), sadness, neutral emotion*. Utterances corresponding to *boredom* were removed from the analysis and we focus on the six basic emotions that were also present in LDC data.

In comparison to LDC dataset, utterances in Berlin dataset are notably longer. The underlying sentences were designed to maximize the number of vowels. In addition, each of the recorded utterances was rated by 20 human subjects with respect to perceived naturalness. Subjects were also asked to classify each utterance as expressing one of the possible emotions. Utterances for which intended emotion recognition was low or which had low perceived naturalness were removed from the dataset. Due to these differences in corpus preparation, we expected to achieve higher emotion recognition rates on Berlin dataset than on LDC dataset (and this indeed was the case).

## 4. Features

In our work we compared and combined two types of features: traditional prosodic features and spectral features for three distinct phoneme classes. Prosodic features used in this paper are derived from pitch, intensity and first formant frequency profiles as well as voice quality measures. Our spectral features which are comprised of statistics of Mel-Frequency Cepstral Coefficients (MFCC).

Given an input utterance, the first step in our feature extraction algorithm is to obtain its phoneme-level segmentation. For LDC dataset, we used Viterbi forced alignment (Odell et al., 2002) between an utterance and its transcript to find the starting and ending time of each phoneme, as well as to detect presence of lexical stress for each of the vowels in the utterance. Forced alignment was performed using generic monophone HMM models of English trained on non-emotional speech. We used a pronunciation dictionary which contained multiple transcriptions of each word based on various pronunciation variants and stress positions. For each utterance, its transcript was expanded into a multiple-pronunciation recognition network using the dictionary. We used Viterbi decoding to order to find the most likely path through the network which yields starting and ending times of each phoneme in the utterance as well as the actual lexical stress. It should be noted that the obtained vowel stress is not fixed to a single dictionary pronunciation but depends on the observed acoustic evidence. For Berlin dataset, we did not have available German acoustic models, so we used the manual segmentations provided as a part of the dataset. Thus, the emotion recognition results on Berlin dataset presented in the paper cannot be considered fully automatic.

We grouped phonemes into three phoneme type classes of interest: stressed vowels, unstressed vowels and consonants. Class-level features were created by computing statistics of prosodic and spectral measurements from parts of the utterance corresponding to these classes. Such partition of phonemes into classes reduces dependence of our features on specific utterance content and, at the same time, provides robustness and avoids sparsity given that a single utterance contains only a small number of phonemes.

In order to analyze the usefulness of class-level spectral features and compare their performance with existing approaches, we computed four different sets of features, varying the type of features (spectral or prosodic) and the region of the utterance over which they were computed as shown in Fig. 1. The regions were either the entire utterance or local regions corresponding to phoneme classes. In the latter setting the features from each of the three phoneme classes were concatenated to form the feature vector descriptive of the entire utterance.

While the primary goal of this paper is to investigate the performance of the class-level spectral features alone and in

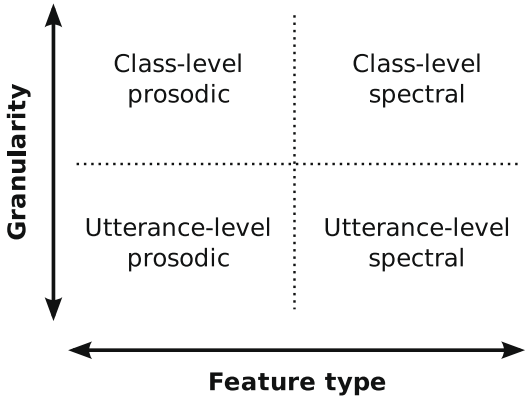


Fig. 1. We computed four different types of features by varying the type of features (prosodic or spectral) and the region of utterances where they are computed (utterance-level and class-level).

a combination with prosodic features, we used the additional feature sets as baseline benchmarks in emotion classification experiments as well as to gain insights on how phoneme-level analysis can improve emotion differentiation in speech. Below, we describe each of the feature sets in detail.

### 4.1. Utterance-level prosodic features

Previous approaches to emotion analysis in speech have used various statistics of the fundamental frequency ($F0$), formant frequencies and voice intensity profiles. Following prior work, we used Praat software (Boersma and Weenink, 2001) to estimate $F0$ and $F1$ contours. For each utterance, we normalized intensity, $F0$ and $F1$ contours by computing speaker-specific $z$-scores. In addition to the features derived from formant frequencies and voice intensity, we also extracted micro-prosodic measures of voice quality such as jitter (the short-term period-to-period fluctuation in fundamental frequency) and shimmer (the random short-term changes in the glottal pulse amplitude) as well as the relative duration of voiced segments which characterizes speech rhythm and the relative spectral energy above 500 Hz (HF500). We computed statistics over the entire utterance such as mean value, standard deviation, minimum and maximum of $F0$ and its first derivative, voice intensity and its derivative as well as of the first formant frequency ($F1$). In total, the set of utterance-level prosodic features contains 24 features:

- mean, std, min, max of $F0$ and $F0$ derivative
- mean, std, min, max of $F1$
- mean, std, min, max of voice intensity and its derivative
- jitter, shimmer, HF500
- relative duration of voiced segments

### 4.2. Class-level prosodic features

Instead of utterance-level statistics, class-level prosodic features use statistics of voice intensity and formants computed over utterance segments which correspond to stressed and unstressed vowel classes. We did not use the consonant class since formant frequencies are not defined for voiceless phonemes. Jitter, shimmer and HF500 were computed over the voiced part of the utterance. The set of class-level prosodic features consists of 44 individual features:

- mean, std, min, max of $F0$ and $F0$ derivative over stressed vowel region
- mean, std, min, max of $F0$ and $F0$ derivative over unstressed vowel region
- mean, std, min, max of $F1$ over stressed vowel region
- mean, std, min, max of $F1$ over unstressed vowel region
- mean, std, min, max of voice intensity and its derivative over stressed vowel region
- mean, std, min, max of voice intensity and its derivative over unstressed vowel region
- jitter, shimmer, HF500
- relative duration of voiced segments

### 4.3. Utterance-level spectral features

Utterance-level spectral features are mean values and standard deviations of MFCC computed over the entire utterance. For each utterance, we computed 13 MFCC (including log-energy) using a 25 ms Hamming window at intervals of 10 ms. For each utterance, we normalized MFCC trajectory by computing speaker-specific $z$-scores. In addition, we computed delta and acceleration coefficients as the first and second derivatives of MFCC using finite differences (26 features). The total number of utterance-level spectral features is 78 which includes means and standard deviations of MFCC as well as the delta and acceleration coefficients.

### 4.4. Class-level spectral features

Class-level spectral features model how emotion is encoded in speech at the phoneme-level. Using the phoneme-level segmentation of the utterance, we formed the spectral feature vector by concatenating class-conditional means and standard deviations of MFCC for each of stressed vowel, unstressed vowel and consonant classes. In addition, we computed the average duration of the above phoneme classes. In summary, the class-level spectral feature vector is 237 dimensional and consists of the following feature groups:

- mean and std of MFCC over stressed vowel region
- mean and std of MFCC over unstressed vowel region
- mean and std of MFCC over consonant region
- mean duration of stressed vowels
- mean duration of unstressed vowels
- mean duration of consonants

### 4.5. Combined features

In order to investigate performance of spectral features in combination with prosodic features, we created a combined feature set by concatenating the sets of class-level spectral and utterance-level prosodic features.[1] In total, the combined set consists of 261 features.

## 5. Emotion classification

In our experiments on emotion recognition, we used SVM classifiers with radial basis kernels constructed using LIBSVM library (Chang and Lin, 2001). Since the number

[1] Since utterance-level features are the most common prosodic features, we do not report any other combinations to avoid clutter in presentation and interpretation of the results.

of utterances per emotion class varied widely in both LDC and Berlin datasets, we used the Balanced ACcuracy (BAC) as a performance metric for emotion recognition experiments presented below. BAC is defined as the average over all emotion classes of recognition accuracy for each class:

$$\mathrm{BAC} = \frac{1}{K}\sum_{i=1}^{K}\frac{n_i}{N_i}, \quad (1)$$

where $K$ is the number of emotion classes, $N_i$ is the total number of utterances belonging to class $i$ and $n_i$ is the number of utterances in this class which were classified correctly. Unlike the standard classification accuracy defined as the total proportion of correctly classified instances, BAC is not sensitive to imbalance in distribution between emotion classes. For example, let us consider a binary classification of neutral emotion versus happiness in a dataset containing 90 neutral and 10 happy utterances. Predicting all utterances as the majority class (neutral) would correspond to the relative accuracy of classification of 90%, while BAC is equal to 50%.

In order to confirm stability and speaker independence of the obtained classifiers, testing was performed using Leave-One-Subject-Out (LOSO) paradigm such that the test set did not contain utterances from the speakers used in the training set. Classification experiments were performed in a round-robin manner by consecutively assigning each of the speakers to the test set and using utterances from the rest of the speakers in the database as the training set.[2] The optimal values of the SVM parameters for each fold were computed using a cross-validation procedure over the training set. We computed the overall BAC recognition accuracy by applying Eq. (1) to the set of predictions combined from all of the LOSO folds.

In the experiments presented below, we investigated performance of each of the four sets of features introduced in Section 4, plus that of the combination of utterance-level prosodic features and class-level spectral features (*combined*). It should be noted that, while a number of previous approaches described in Section 2 focused only on speaker-dependent emotion recognition, our experiments are on *speaker-independent* emotion recognition since our recognition experiments made use of utterances from the speakers which were unseen during classifier training.

### 5.1. Multi-class emotion recognition

In our first experiment, we considered the task of multi-class classification of the six basic emotions. The accuracy of speaker-independent, multi-class classification on LDC and Berlin datasets is shown in Table 2 for features of different types (prosodic and spectral) and granularity levels (utterance-level and class-level). The accuracies for the complete set of 15 emotions for LDC data is given to allow comparison with prior work.

Table 2
Speaker-independent, multi-class emotion classification rates for six emotion task on LDC and Berlin datasets using prosodic and spectral features with different levels of granularity: utterance-level (UL) and class-level (CL). Classification rates for the complete set of 15 emotions for LDC data is given to allow comparison with prior work. Best performance is shown in bold.

| | LDC dataset six emotions (%) | Berlin dataset six emotions (%) | LDC dataset 15 emotions (%) |
|---|---|---|---|
| UL prosody | 23.1 | 68.1 | 17.0 |
| CL prosody | 27.7 | 68.6 | 17.4 |
| UL spectral | 33.5 | 67.0 | 24.4 |
| CL spectral | **44.5** | 75.9 | **30.7** |
| Combined | 43.7 | **78.2** | 29.7 |

Our results indicate that class-level spectral features perform better than other types of features for both LDC and Berlin datasets. Class-level spectral features also outperform the utterance-level prosodic features by absolute 21.4% in LDC and 7.8% in Berlin datasets. There are also noticeable improvements over commonly used utterance-level spectral features. For Berlin dataset, the best results are obtained when combination of the class-level spectral and utterance-level prosodic features is used. However, the combined features perform slightly worse than class-level spectral features in LDC dataset.

While Table 2 presents *speaker-independent* emotion recognition accuracy, the majority of previous work did not use LOSO paradigm and focused on recognizing emotions in *speaker-dependent* settings. For the sake of comparison, we computed *speaker-dependent* emotion recognition accuracy using prosodic and spectral features with different granularity levels (utterance- and class-level). Each dataset was randomly split into the training set which contained 70% of the total number of utterances and the test set which included remaining 30% of the utterances. The accuracy of *speaker-dependent* emotion recognition for LDC and Berlin datasets is shown in Table 3. While similarly to the *speaker-independent* case, class-level spectral features outperform other feature types, the overall recognition performance is significantly higher than for the *speaker-independent* case. For example, *speaker-dependent* performance of utterance-level prosodic features in LDC dataset is almost twice the accuracy of the *speaker-independent* recognition.

In order to compare performance of the class-level spectral features to the results of previous work on speaker-independent emotion classification (Yacoub et al., 2003; Huang and Ma, 2006), we conducted an experiment on classification of all 15 emotions in LDC dataset. The accuracy of 15-class classification is given in the last column of Table 2. Classification accuracy of 30.7% obtained using class-level spectral features is considerably higher than the prosody-based classification accuracy of 18% reported in Huang and Ma (2006) and 8.7% reported in Yacoub et al. (2003) on the same task. Note that the results might

[2] This in effect corresponds to 7-fold and 10-fold cross-validation for LDC and Berlin datasets respectively.

Table 3
Speaker-dependent multi-class emotion classification rates for six emotion task on LDC and Berlin datasets using prosodic and spectral features with different levels of granularity: utterance-level (UL) and class-level (CL).

| | LDC dataset six emotions (%) | Berlin dataset six emotions (%) |
|---|---|---|
| UL prosody | 46.3 | 71.9 |
| CL prosody | 46.6 | 72.3 |
| UL spectral | 48.5 | 72.3 |
| CL spectral | 63.2 | 82.3 |
| Combined | 62.6 | 84.8 |

not be directly comparable because it is unclear how the earlier studies accounted for imbalance between emotion classes or how cross-validation folds were formed.

### 5.2. One-versus-all emotion recognition

In the second experiment, we performed recognition of each of the six basic emotions versus the other five emotions. For example, one of the tasks was to recognize if an utterance conveys *sadness* versus some other emotion among *anger, fear, disgust, happy* and *neutral*. The balanced accuracy of one-versus-all classification on LDC and Berlin datasets is shown in Tables 4 and 5 for sets of features with different types (prosodic and spectral) and granularity levels (utterance-level and class-level). Recognition accuracy changes with respect to granularity for both prosodic and spectral features. Our results indicate that the class-level prosodic features do not provide any consistent improvement over the utterance-level features. This is not surprising since prosodic features are suprasegmental.

On the other hand, class-level spectral features provide a consistent performance improvement over the utterance-level spectral features in most of the cases with exception of recognition of *disgust* and *happiness* in LDC and *neutral* in Berlin database. For example, the absolute performance gain is as high as 13.9% for recognition of *disgust* in Berlin dataset.

*Class-level spectral* features also yield noticeably higher emotion recognition accuracy compared to *utterance-level prosodic* features for most of the emotions. For instance, the absolute improvements in recognition accuracy of *neutral* for LDC and *disgust* for Berlin datasets are 30.3% and 25.6% respectively. The only exceptions are recognition of *fear* and *happiness* in the Berlin dataset, where prosodic features lead to improvements over spectral features of 3.1% and 5.9% respectively.

Moreover, the combination of the class-level spectral and the utterance-level prosodic features yields even further improvements in some cases. In other cases, the combined set of features yields classification accuracy which is lower than accuracy of either utterance-level prosodic or class-level spectral features. We believe that this is due to high dimensionality of the combined feature set. We test several feature selection algorithms in Section 7 in order to improve the performance of both class-level spectral and combined features.

## 6. Utterance-length dependence

Multi-class emotion recognition results presented in Table 2 indicate that the overall accuracy of emotion recognition obtained on Berlin dataset is much higher than the one on LDC dataset. Besides differences in language and recording scenarios between the two datasets, better separation between emotions can be attributed to the fact that Berlin dataset contains longer utterances. To the best of our knowledge, effects of the utterance length on emotion recognition accuracy have not been explored in the literature.

Table 4
Accuracy of one-versus-all classification for LDC dataset using prosodic and spectral features with different levels of granularity: utterance-level (UL) and class-level (CL). Best performance is shown in bold.

| | Anger (%) | Fear (%) | Disgust (%) | Happy (%) | Sadness (%) | Neutral(%) |
|---|---|---|---|---|---|---|
| UL prosody | 63.6 | 55.9 | 51.6 | 56.7 | 53.2 | 53.5 |
| CL prosody | 67.7 | 51.4 | 51.2 | 61.4 | 53.3 | 59.0 |
| UL spectral | 66.2 | 50.7 | 53.9 | 58.8 | 55.8 | 66.8 |
| CL spectral | 71.3 | **60.9** | **51.6** | 57.6 | **60.4** | **83.8** |
| Combined | **71.9** | 58.6 | 48.4 | **59.2** | 59.4 | 79.8 |

Table 5
Accuracy of one-versus-all classification for Berlin dataset using prosodic and spectral features with different levels of granularity: utterance-level (UL) and class-level (CL). Best performance is shown in bold.

| | Anger (%) | Fear (%) | Disgust (%) | Happy (%) | Sadness (%) | Neutral(%) |
|---|---|---|---|---|---|---|
| UL prosody | 84.9 | **85.9** | 65.1 | **72.1** | 88.5 | 87.5 |
| CL prosody | 87.8 | 76.2 | 65.5 | 67.2 | 93.0 | 85.9 |
| UL spectral | 81.8 | 73.5 | 76.8 | 63.2 | 83.5 | 88.9 |
| CL spectral | 88.2 | 82.8 | **90.7** | 66.2 | **93.8** | 88.6 |
| Combined | **89.0** | 82.8 | 88.2 | 65.5 | 92.9 | **89.4** |

In order to investigate how emotion recognition performance depends on the utterance length, we constructed longer speech segments by concatenating utterances from LDC dataset. We built four additional synthetic datasets with progressively longer speech segments by containing together 2, 3, 4 and 5 randomly chosen utterances produced by the same actor. While utterances in LDC dataset contain four-syllable phrases, the four synthetic datasets contained speech segments with 8, 12, 16 and 20 syllables respectively. For example, in order to build a synthetic dataset consisting of eight-syllable segments, we randomly split the set of utterances produced by the same speaker into pairs and concatenated them. Since each utterance from the original dataset was used only once, the synthetic datasets contained fewer utterances than the original LDC dataset.

We calculated speaker-independent emotion recognition accuracy of the six basic emotions for each of the synthetic datasets. Each of the datasets was split into the training and test sets using the LOSO procedure described in Section 5. Fig. 2 shows recognition accuracy as a function of the number of syllables in the utterances. While the performance of utterance-level prosodic features does not noticeably change with the utterance length, accuracy of the class-level spectral features increases as longer utterances are used. We would like to point out that, since our results are obtained on synthetic datasets, these predictions will not necessarily apply to naturally spoken emotional utterances. It would be interesting to further investigate how emotion recognition performance depends on utterance length using emotion speech corpora which is rich with utterances of different durations in order to cross-validate our findings.

## 7. Feature selection

The high dimensionality of class-level spectral feature vectors as well as the presence of highly correlated features in the combined set of prosodic and spectral features can negatively affect performance of machine learning algorithms such as, for example, SVM classifiers used in this paper. Moreover, multi-class emotion classification results on LDC dataset (Table 2) indicate that the combination of prosodic and spectral features performs worse than spectral features used alone which might be due to presence of irrelevant or highly correlated features. Emotion classification accuracy of class-level spectral feature and the combined features can be improved by using feature selection algorithms which aim to find a lower-dimensional subset of features which yields better classification performance.

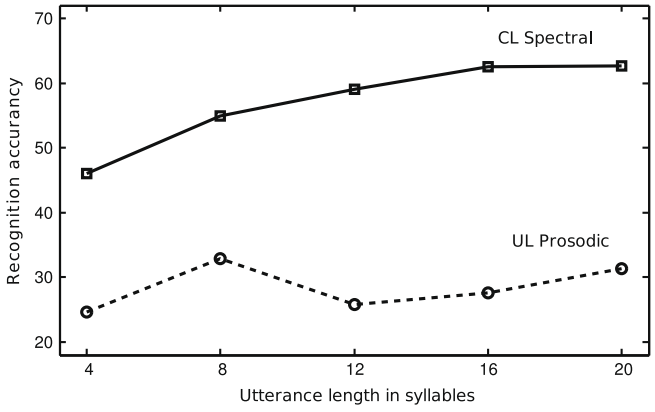


Fig. 2. Dependence of emotion recognition accuracy on utterance length.

In this section, we apply feature selection algorithms in component-wise and group-wise settings. In the component-wise case, our goal is to select an optimal set of individual features by performing a greedy search over the set of all possible feature combinations. On the other hand, group-wise selection aims to find the best combination of feature subgroups such as prosodic and spectral features defined over each phoneme class. While, in principle, component-based feature selection should achieve better classification accuracy, group-wise selection and ranking can help us to understand how different types of features contribute to emotion differentiation.

### 7.1. Component-wise feature selection

Component-wise feature selection methods rely on searching through all possible subsets of features and fall into either wrapper or filter categories based on the type of criteria used to evaluate subsets of features. While wrapper approaches search for a subset of features by maximizing accuracy of a classifier on a hold-out subset of the training data, filter methods perform selection of features as a pre-processing step independent of any particular classification approach.

In this paper, we used wrapper methods which maximize the accuracy of linear SVM classifiers. We also used filter methods such as *subset evaluation* (Hall and Smith, 1997) based on correlation measures and *information gain ratio* (Hall and Smith, 1998). Since exhaustive evaluation of all possible subsets of features is computationally prohibitive, we employed greedy search algorithms. *Greedy stepwise* search starts with the full set of features and iteratively removes individual features until the objective criterion can no longer be improved. *Rank search* uses a ranking of features based on the gain ratio metric in order to evaluate feature subset of increasing size which are constructed by iterative addition of the best ranked features. In the case of *information gain ratio* selection criterion, we did not use any search algorithm and simply selected features with positive information gain ratio values.

For each LOSO fold, all utterances from one of the speakers in the training set were sequentially assigned to the hold-out set used for evaluation and the rest of utterances were used to train linear SVM classifiers for wrapper-based feature selection. This process was repeated for each speaker in the training set and selected features from each iteration were combined into a single set of selected

features. It should be noted that different sets of features were selected in different LOSO folds. In order to compare different feature selection algorithms, we calculated their balanced accuracy over LDC and Berlin datasets by combining classifier predictions from individual LOSO folds. We applied feature selection algorithms to utterance-level prosodic, class-level spectral and combined sets of features. Tables 6 and 7 compare emotion recognition of different wrapper and filter feature selection algorithms used in this paper. Wrapper selection utilizing greedy stepwise search improves the accuracy of utterance-level prosodic and class-level spectral features in LDC dataset by 1.0% and 2.5% respectively. However, none of the feature selection algorithms provide any noticeable improvement for the combination of prosodic and spectral features in LDC dataset. On the other hand, the best results in Berlin dataset are achieved by filter selection based on the information gain ratio yielding modest improvements of 1.8% and 0.3% for utterance-level prosodic and class-level spectral features, and 0.9% improvement for their combination. Despite the high dimensionality of class-level spectral features, none of the feature selection methods tested in this paper lead to clear performance gains for either Berlin or LDC datasets.

### 7.2. Group-wise feature selection and ranking

While component-wise feature selection searches through all possible combinations of individual features, the list of selected features is difficult to analyze. Instead, one may be interested in investigating combinations and contributions from features defined over different phonetic groups rather than looking at individual feature components such as filterbank responses. In this section, we perform group-wise feature selection and ranking by focusing on subgroups of class-level spectral features. We split the set of class-level spectral features into consonant, stressed vowel and unstressed vowel subgroups. Then the combined features consist of four subgroups which include consonant, stressed vowel and unstressed vowel spectral subgroups as well as the group of prosodic features.

Our goal is to find a combination of feature subgroups which maximizes emotion classification accuracy. For the small number of groups in this case, such a combination can be found by performing a brute force search among all possible subgroup combinations. Group-wise selection follows wrapper feature selection procedure described in Section 7.1. However, instead of using a greedy search to find an optimal combination of subgroups, we evaluate classifier performance on all possible subgroup combinations and choose the one which yields the best classification accuracy. Tables 8 and 9 show accuracy of recognizing the six basic emotions using group-wise feature selection for LDC and Berlin datasets. While group-wise selection improves classification accuracy of class-level spectral features by 1.6% and 1.2% in LDC and Berlin datasets respectively, feature subgroups selected from the combined set of utterance-level prosodic and class-level spectral features in Berlin dataset performs worse than the entire combined set.

Table 10 shows how often each subgroup was selected in all of the LOSO folds in LDC and Berlin datasets. For example, prosodic subgroup was selected in one out of seven LOSO folds in LDC and 10 out of 10 folds in Berlin dataset. On the other hand, spectral features derived from the stressed vowel regions were always selected in LDC and only in five out of 10 folds in Berlin datasets. Spectral features derived from unstressed vowels were chosen in three out of seven folds in LDC and nine out of 10 folds in Berlin dataset. While selection frequency of prosodic and vowel spectral features varied between both datasets, spectral features derived from consonant regions were chosen for all of LOSO fold in both datasets. We believe that this is due to the fact that consonant spectral features are always complementary to prosodic and vowel spectral subgroups.

Table 6
Multi-class emotion classification rates with feature selection for six emotion recognition in LDC datasets.

| | W/o selection (%) | Rank search SVM wrapper (%) | Rank search subset eval. (%) | Greedy stepwise SVM wrapper (%) | Info gain ratio (%) |
|---|---|---|---|---|---|
| UL prosody | 23.1 | 23.8 | 23.3 | 24.1 | 23.9 |
| CL spectral | 44.5 | 44.8 | 46.2 | 47.0 | 46.6 |
| Combined | 43.7 | 41.7 | 42.3 | 39.1 | 43.8 |

Table 7
Multi-class emotion classification rates with feature selection for six emotion recognition in Berlin dataset.

| | W/o selection (%) | Rank search SVM wrapper (%) | Rank search subset eval. (%) | Greedy stepwise SVM wrapper (%) | Info gain ratio (%) |
|---|---|---|---|---|---|
| UL prosody | 68.1 | 67.7 | 68.6 | 69.4 | 69.8 |
| CL spectral | 75.9 | 75.4 | 75.1 | 76.0 | 76.2 |
| Combined | 78.2 | 78.5 | 81.3 | 78.2 | 79.1 |

Table 8
Group-wise feature selection for LDC dataset.

| | W/o selection (%) | Group-wise selection (%) |
|---|---|---|
| CL spectral | 44.5 | 46.1 |
| Combined | 43.7 | 44.2 |

Table 9
Group-wise feature selection for Berlin dataset.

| | W/o selection (%) | Group-wise selection (%) |
|---|---|---|
| CL spectral | 75.9 | 77.1 |
| Combined | 78.2 | 76.7 |

Table 10
Group occurance frequencies during group-wise feature selection on LDC and Berlin datasets.

| | LDC (%) | Berlin (%) |
|---|---|---|
| Prosodic | 14.3 | 90.0 |
| Spectral consonants | 100.0 | 100.0 |
| Spectral stressed vowels | 100.0 | 50.0 |
| Spectral unstressed vowels | 42.9 | 90.0 |

Indeed, consonant features mostly correspond to unvoiced portions of the utterance, while prosodic and vowel spectral features are derived from the voiced parts.

In order to evaluate contributions of different feature types to emotion differentiation, we ranked feature subgroups based on their emotion classification accuracy when used alone. Table 11 displays rankings of stressed vowel, unstressed vowel and consonant spectral features as well as prosodic features on LDC and Berlin datasets. Relative rankings of feature subgroups differ between LDC and Berlin datasets. For example, stressed vowel spectral features yield the highest emotion recognition accuracy in LDC dataset. However, the best performing subgroup for Berlin dataset corresponds to consonant spectral features. While utterance-level prosodic features are the second best subgroup in Berlin dataset, they yield the worst recognition accuracy in LDC dataset. Surprisingly, consonant spectral features are ranked as one of the best performing features in both datasets. Consonant features outperform the second best prosodic features by 4.5% in Berlin dataset and lag behind the best performing stressed vowel features by 0.7% in LDC dataset.

## 8. Discussion

In this paper, we introduced a novel set of spectral features for emotion recognition which uses class-level statistics of MFCC. We compared performance of the class-level spectral features with traditional utterance-level prosodic and spectral features in emotion recognition on publicly available LDC and Berlin datasets. While previous work on spectral features for emotion recognition used utterance-level statistics, our results indicate that representing how emotion is encoded in spectral domain at the phoneme-level improves classification accuracy. We demonstrated that the class-level spectral features outperform both prosodic and utterance-level spectral features in multi-class emotion recognition.

Table 11
Classification rates for each of the feature groups.

| | LDC (%) | Berlin (%) |
|---|---|---|
| Prosodic | 23.1 | 68.1 |
| Spectral consonants | 40.2 | 72.50 |
| Spectral stressed vowels | 40.9 | 62.6 |
| Spectral unstressed vowels | 29.0 | 64.3 |

While the class-level features introduced in this paper allow to model emotion at the level of granularity finer than utterance-level features, this comes at the expense of the increased feature space dimensionality. Moreover, not all of the class-level features are necessarily relevant to discriminating between different emotions. The high dimensionality of the feature space, presence of the large number of irrelevant or highly correlated features is known to hurt performance of machine learning algorithms such as SVM classifiers used in this paper. For example, our results on multi-class emotion recognition (Table 2) show that the high-dimensional combination of utterance-level prosodic and class-level spectral features performs worse than class-level features used alone. In attempt to alleviate this problem and improve classification accuracy of class-level spectral and combined features, we tested several feature selection algorithms to automatically find a smaller subset of features yielding better emotion recognition performance. Given the large number of class-level spectral features, we expected feature selection will improve results even further, but none of the selection methods considered in this paper lead to clear gains. While feature selection slightly improved performance of class-level spectral features, feature selection applied to the set of combined features did not yield any noticeable improvement.

Class-level spectral features consist of three subgroups corresponding to phoneme classes of interest: stressed vowels, unstressed vowels and consonants. We explored relative contributions from each of these phoneme classes to emotion differentiation and observed that the consonant spectral features outperform spectral features derived from stressed and unstressed vowel regions. Moreover, consonant features alone outperform prosodic features in Berlin dataset and only lag behind by 0.7% in LDC dataset. Traditional prosodic features, derived from pitch and intensity profiles, only use measurements from the voiced segments of speech which mostly correspond to vowel regions of the utterance. On the other hand, consonant spectral features describe unvoiced regions which are not accounted for by prosodic features. This can explain why class-level spectral features outperform the traditional prosodic measures.

We observed that the overall accuracy of emotion recognition obtained on Berlin dataset is much higher than the one on LDC dataset. Besides differences in language and

recording scenarios between the two datasets, better separation between emotions can be attributed to the fact that Berlin dataset contains longer utterances. To the best of our knowledge, dependence of emotion recognition accuracy on the utterance length has not been explored in the literature. We investigated this dependence by constructing synthetic speech segment of increasing length using utterances from LDC database. Our experiments demonstrate that accuracy of class-level spectral features increases with utterance length reaching the asymptote at utterance length of 16 syllables. Since these findings are obtained using synthetic data, dependence of emotion recognition accuracy on utterance length needs to be further cross-validated on naturally spoken emotional corpora.

There are several aspects of feature extraction for emotion recognition that need to be explored in the future research. Firstly, the three phoneme type classes of interest used in this paper represent only one choice for defining the level of granularity for spectral features. It would be interesting to investigate finer phoneme classes or even use phoneme-level features for emotion recognition. However, there is a trade-off between number of classes of interest and amount of training data available for classifier training. Consonant, stressed and unstressed vowel classes used in this paper provide a good balance by utilizing phoneme-specific measurements while avoiding data sparseness. Using finer classes of interest such as, for instance, individual phonemes, would require more data to be available for classifier training. In addition, emotion recognition would have to be treated as a sparse classification problem since it is possible to obtain measurements from only a few phonemes in a typical utterance. Secondly, normalization of spectral features in order to accommodate inter-speaker difference has not been fully addressed in the literature. While this paper uses a simple method based on speaker-specific $z$-scores to normalize MFCC trajectories, we believe that more sophisticated speaker normalization and adaptation techniques akin to MLLR adaptation (Legetter and Woodland, 1996) which has been proven useful in automatic speech recognition can further improve speaker-independent emotion recognition accuracy.

## Acknowledgements

This work is supported by NIH Grant R01MHO73174. The authors would like to thank Dr. Jiahong Yuan for providing us with the code and English acoustic models for forced alignment.

## References

Banse, R., Scherer, K., 1996. Acoustic profiles in vocal emotion expression. J. Personality Social Psychology 70 (3), 614–636.

Bitouk, D., Nenkova, A., Verma, R., 2009. Improving emotion recognition using class-level spectral features. In: Proc. Interspeech 2009.

Boersma, P., Weenink, D., 2001. Praat, a system for doing phonetics by computer. Glot Internat., 341–345.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., Weiss, B., 2005. A database of german emotional speech. In: Proc. Interspeech 2005, pp. 1–4.

Chang, C.-C., Lin, C.-J., 2001. LIBSVM: a library for support vector machines. Software available at %3chttp://www.csie.ntu.edu.tw/cjlin/libsvm%3e.

Dellaert, F., Polzin, T., Waibel, A., 1996. Recognizing emotion in speech. In: Proc. ICSLP 1996, pp. 1970–1973.

Fernandez, R., Picard, R., 2003. Modeling drivers' speech under stress. Speech Comm., 145–159.

Grimm, M., Mower, E., Kroschel, K., Narayanan, S., 2006. Combining categorical and primitives-based emotion recognition. In: Proc. 14th European Signal Processing Conference.

Hall, M., Smith, L., 1997. Feature subset selection: a correlation based filter approach. In: Proc. Internat. Conf. on Neural Information Processing and Intelligent Information Systems, pp. 855–858.

Hall, M., Smith, L., 1998. Practical feature subset selection for machine learning. In: Proc. 21st Australasian Computer Science Conference, pp. 181–191.

Hasegawa-Johnson, M., Levinson, S., Zhang, T., 2004. Children's emotion recognition in an intelligent tutoring scenario. In: Proc. Interspeech 2004.

Hu, H., Xu, M.-X., Wu, W., 2007. Gmm supervector based svm with spectral features for speech emotion recognition. In: Proc. ICASSP 2007, Vol. 4, pp. 413–416.

Huang, R., Ma, C., 2006. Toward a speaker-independent real-time affect detection system. In: Proc. Internat. Conf. on Pattern Recognition, pp. 1204–1207.

Kim, S., Georgiou, P., Lee, S., Narayanan, S., 2007. Real-time emotion detection system using speech: multi-modal fusion of different time-scale features. In: Proc. IEEE 9th Workshop on Multimedia Signal Processing.

Kwon, O.W., Chan, K., Hao, J., Lee, T., 2003. Emotion recognition by speech signals. In: Proc. 8th Eur. Conf. on Speech Communication and Technology, pp. 125–128.

Lee, C., Yildirim, S., Bulut, M., Kazemzadeh, A., Busso, C., Deng, Z., Lee, S., Narayanan, S., 2004. Emotion recognition based on phoneme classes. In: Proc. Interspeech 2004, pp. 205–211.

Legetter, C., Woodland, P., 1996. Mean and variance adaptation within the mllr framework. Comput. Speech Lang. 10, 249–264.

Linguistic Data Consortium, 2002. Emotional prosody speech and transcripts. LDC Catalog No.: LDC2002S28, University of Pennsylvania.

Luengo, I., Navas, E., Hernez, I., Snchez, J., 2005. Automatic emotion recognition using prosodic parameters. In: Proc. Interspeech 2005, pp. 493–496.

McGilloway, S., Cowie, S., Douglas-Cowie, E., Gielen, S., Westerdijk, M., Stroeve, S., 2000. Approaching automatic recognition of emotion from voice: a rough benchmark. In: Proc. ISCA Workshop on Speech and Emotion 2000, pp. 200–205.

Meng, H., Pittermann, J., Pittermann, A., Minker, W., 2007. Combined speech-emotion recognition for spoken human-computer interfaces. In: Proc. IEEE Internat. Conf. on Signal Processing and Communications, pp. 1179–1182.

Neiberg, D., Elenius, K., Laskowski, K., 2006. Emotion recognition in spontaneous speech using gmms. In: Proc. Interspeech 2006, pp. 809–812.

Nicholson, J., Takahashi, K., Nakatsu, R., 2000. Emotion recognition in speech using neural networks. Neural Comput. Appl. 9, 290–296.

Nwe, T., Foo, S., Silva, L.D., 2003. Speech emotion recognition using hidden Markov models. Speech Comm. 41 (4), 603–623.

Odell, J., Ollason, D., Woodland, P., Young, S., Jansen, J., 2002. The HTK Book. Cambridge University Press.

Pao, T., Chen, Y., Yeh, J., Liao, W., 2005. Detecting emotions in mandarin speech. Comput. Linguistics Chin. Lang. 10 (3).

Sato, N., Obuchi, Y., 2007. Emotion recognition using mel-frequency cepstral coefficients. Inf. Media Technol. 2 (3), 835–848.

Scherer, S., Schwenker, F., Palm, G., 2007. Classifier fusion for emotion recognition from speech. In: Proc. Internat. Conf. on Intelligent Environments, pp. 152–155.

Schuller, B., Mller, R., Lang, M., Rigoll, G., 2005. Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles. In: Proc. Interspeech 2005, pp. 805–809.

Schuller, B., Rigoll, G., 2006. Timing levels in segment-based speech emotion recognition. In: Proc. Interspeech 2006, pp. 1818–1821.

Sethu, V., Ambikairaja, E., Epps, J., 2008. Phonetic and speaker variations in automatic emotion classification. In: Proc. Interspeech 2008, pp. 617–620.

Shafran, I., Riley, M., Mohri, M., 2003. Voice signatures. In: Proc. IEEE Automatic Speech Recognition and Understanding Workshop, pp. 31–36.

Shamia, M., Kamel, M., 2005. Segment-based approach to the recognition of emotions in speech. In: Proc. IEEE Internat. Conf. on Multimedia and Expo 2005.

Song, M., Chen, C., Bu, J., You, M., 2004. Speech emotion recognition and intensity estimation. In: Computational Science and its Applications, pp. 406–413.

Tabatabaei, T., Krishnan, S., Guergachi, A., 2007. Emotion recognition using novel speech signal features. In: Proc. IEEE Internat. Symp. on Circuits and Systems, pp. 345–348.

Vlasenko, B., Schuller, B., Wendemuth, A., Rigol, G., 2008. On the influence of phonetic content variation for acoustic emotion recognition. In: Perception in Multimodal Dialogue Systems, pp. 217–220.

Vlasenko, B., Schuller, B., Wendemuth, A., Rigoll, G., 2007. Frame vs. turn-level: emotion recognition from speech considering static and dynamic processing. In: Affective Computing and Intelligent Interaction, pp. 139–147.

Vondra, M., Vich, R., 2009. Recognition of emotions in german speech using gaussian mixture models. In: Multimodal Signals: Cognitive and Algorithmic Issues, pp. 256–263.

Wang, Y., Guan, L., 2005. Recognizing human emotion from audiovisual information. In: Proc. ICASSP 2005, pp. 1125–1128.

Yacoub, S., Simske, S., Lin, X., Burns, J., 2003. Recognition of emotions in interactive voice response systems. In: Proc. Eurospeech 2003, pp. 729–732.

Ye, C., Liu, J., Chen, C., Song, M., Bu, J., 2008. Speech emotion classification on a riemannian manifold. In: Advances in Multimedia Information Processing – PCM 2008, pp. 61–69.