INTERSPEECH 2009 BRIGHTON

Improving Emotion Recognition using Class-Level Spectral Features

Dmitri Bitouk[1], *Ani Nenkova*[2], *Ragini Verma*[1]

[1]Department of Radiology, Section of Biomedical Image Analysis, University of Pennsylvania
[2]Department of Computer and Information Science, University of Pennsylvania

Dmitri.Bitouk@uphs.upenn.edu, nenkova@seas.upenn.edu, Ragini.Verma@uphs.upenn.edu

10.21437/Interspeech.2009-582

Abstract

Traditional approaches to automatic emotion recognition from speech typically make use of utterance level prosodic features. Still, a great deal of useful information about expressivity and emotion can be gained from segmental spectral features, which provide a more detailed description of the speech signal, or from measurements from specific regions of the utterance, such as the stressed vowels. Here we introduce a novel set of spectral features for emotion recognition: statistics of Mel-Frequency Spectral 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 features or utterance-level spectral features. Combination of our phoneme class features with prosodic features leads to even further improvement.
Index Terms: emotion recognition

1. Introduction

Since emotions expressed in speech play an important role in communication, automatic recognition of emotions has attracted attention in a number of human-computer interaction applications such as voice response systems [1, 2] and computer-human tutoring systems [3].

Descriptive studies in psychology and linguistics have mostly been concerned with identifying differences in acoustic correlates of prosody in different emotions, capturing differences in *how* an utterance is produced. 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) [4]. 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 [5, 6]; others employed Hidden Markov Models [7, 2] 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.

Spectral features, based on the short-term power spectrum of sound such as Linear Prediction Coefficients (LPC) and Mel-Frequency Cepstral Coefficients (MFCC), provide a much more detailed description of the speech signal than 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. Because of this, 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. The majority of the approaches introduced so far ignore this challenge by using either utterance-level statistics of spectral features [8, 9, 10, 11], statistics over fixed-length utterance segments [12] or speaker-dependent Hidden Markov Models (HMMs) [13].

On the other hand, phoneme-level classification of emotion has received relatively little attention. The work of Lee et al. [14] takes into account phonetic content of speech for emotion classification by training phoneme-dependent HMM. We are unable to compare our results with their work since it concentrated on speaker-dependent emotion recognition using a proprietary dataset which consisted of utterances produced by only one speaker. Sethu et al. [15] used phoneme-specific Gaussian Markov models 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.

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 (Section 3). Compared to the 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 is more expressed over vowel segments of speech rather than consonants and thus potentially increases the discriminating power of spectral features.

We analyze the performance of phoneme class spectral features in speaker-independent emotion classification in English and German speech using two publicly available datasets (Section 2). We demonstrate that phoneme-level spectral features outperform both the traditional prosodic features and utterance-level statistics of MFCC (Section 4).

2. Databases

In our experiments, we used two publicly available databases of emotional speech: an English emotional speech database from Linguistic Data Consortium (LDC) [16] and Berlin database of German emotional speech [17].

2.1. LDC Emotional Speech Database

The LDC database contains 4628 utterances produced by 7 native English speakers (3 female/4 male) who were instructed to convey one of 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 this paper, we focus on the six basic emotions which include *anger* (merging hot and cold anger productions), *fear, disgust, happy, sadness* and *neutral*. The LDC dataset contains 2433 utterances corresponding to these six basic emotions. Almost all of the utterances in the LDC datasets are short 4-syllable utterances containing dates and numbers.

2.2. German Emotional Speech Database

The Berlin dataset contains emotional utterances elicited from 10 German actors (5 female/5 male) reading one of 10 preselected 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 comprises of the following 7 emotions: *emotion, anger, boredom, fear, disgust, joy (happy), sadness*. In our experiments, we did not use utterances corresponding to *boredom*.

In comparison to the LDC dataset, utterances in the Berlin dataset are notably longer and were chosen to maximize the number of vowels. In addition, each of the recorded utterances were rated by 20 human subjects with respect to perceived naturalness. Subjects were also asked to classify which of the emotion categories correspond to each utterance. Utterances for which intended emotion recognition was low or which had low perceived naturalness were removed from the dataset. Due to these differences, we expected to achieve higher emotion recognition rates on the Berlin dataset than on the LDC dataset.

3. Features

For the LDC dataset, we used forced alignment [18] 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 acoustic models of English trained on non-emotional speech. We did not have available German acoustic models for the Berlin dataset, so we used the manual segmentations provided as a part of the dataset. We grouped phonemes into stressed vowel, unstressed vowel and consonant classes in order to both minimize dependence of our features on specific utterance content and avoid 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. 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 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.

3.1. 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 [19] 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 derivative, voice intensity and its derivative as well as of 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

3.2. 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

3.3. 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 25ms Hamming window at intervals of 10ms. 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.

3.4. 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

3.5. 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. In total, the combined set consists of 261 features.

4. Emotion Classification and Results

In our experiments on emotion recognition, we used SVM classifiers with radial basis kernels constructed using LIBSVM library [20]. For each of the experiments presented below, we split the data into training and test sets. 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. We computed recognition accuracy as the proportion of utterances which were classified correctly in all of the folds.

Since the number of utterances per emotion varied widely, we balanced the data using random pruning such that it contained an equal number of utterances for each of the emotion classes. We computed the optimal values of the SVM parameters using a cross-validation procedure over the training set. In the experiments presented below, we investigated performance of each of the four sets of features introduced in Section 3, 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 [13, 5] 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.

Binary classification. In our first experiment on one-versus-all classification, we performed recognition of each of the 6 basic emotions versus the other 5 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 accuracy of one-versus-all classification on LDC and Berlin datasets is shown in Tables 1 and 2 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 which conforms with the suprasegmental nature of prosody.

	Anger	Fear	Disgust	Happy	Sadness	Neutral
UL Prosody	65.0%	57.2%	55.2%	71.3%	57.6%	68.4%
CL Prosody	63.8%	58.5%	53.5%	71.2%	60.1%	67.8%
UL Spectral	58.4%	59.7%	60.5%	66.6%	50.9%	63.9%
CL Spectral	66.1%	63.1%	**66.6%**	69.5%	60.5%	67.4%
Combined	**68.1%**	**64.7%**	64.6%	**71.5%**	**60.8%**	**69.7%**

Table 1: *Accuracy of one-versus-all classification for LDC dataset. Best performance is shown in bold.*

	Anger	Fear	Disgust	Happy	Sadness	Neutral
UL Prosody	88.6%	**88.5%**	74.1%	75.4%	93.8%	90.4%
CL Prosody	87.6%	77.0%	69.9%	74.0%	95.1%	90.8%
UL Spectral	85.2%	79.1%	83.6%	73.1%	92.3%	91.1%
CL Spectral	88.2%	84.2%	**89.1%**	77.8%	97.3%	92.1%
Combined	**91.1%**	84.4 %	88.6%	**78.6%**	**97.3%**	**93.0%**

Table 2: *Accuracy of one-versus-all classification for the Berlin dataset. Best performance is shown in bold.*

On the other hand, class-level spectral features provide a consistent performance improvement over the utterance-level spectral features. For example, the absolute performance gain is as high as 9.6% in recognition of *sadness* in the LDC dataset.

Class-level spectral features also yield notably higher emotion recognition accuracy compared to *utterance-level prosodic* features for most emotions. For instance, the absolute improvement in recognition accuracy of *disgust* is 11.4% for the LDC and 15.0% for the Berlin datasets. The exception are for recognition of *happy* and *neutral* in the LDC dataset, where prosodic features lead to 1% better results and *anger* and *fear* in the Berlin datasets for which prosodic features lead to improvements over spectral features of 0.4% and 4.3% respectively.

Moreover, the combination of the class-level spectral and the utterance-level prosodic features yields even further improvements. In some cases, the combined set of features yields classification accuracy lower compared to accuracy of either the utterance-level prosodic or the class-level spectral features. We believe that this is due to high dimensionality of the combined feature set and can remedied by employing feature selection algorithms which are beyond the scope of this paper.

Multi-class recognition. In the second experiment, we considered the task of multi-class classification of the six basic emotions. In this setting class-level spectral features are clearly the best. Classification accuracy for LDC and Berlin datasets is shown in the second and third columns of Table 3. Class-level spectral features also outperform the utterance-level prosodic features by absolute 7.7% in the LDC and 7.1% in the Berlin datasets. The improvements over utterance-level spectral features are even greater. For both datasets, the best results are obtained when combination of the class-level spectral and utterance-level prosodic features is used.

Comparison with prior work. Finally, in order to compare performance of the class-level spectral features to the results of previous work on emotion classification [21, 7], we conducted an experiment on classification of all 15 emotions in the LDC dataset. The accuracy of 15-class classification is given in the last column of Table 5. Classification accuracy of 23.9% obtained using combination of utterance-level prosodic and class-level spectral features is considerably higher than the prosody-based classification accuracy of 18% reported in [7] and 8.7%

	LDC Dataset 6 emotions	Berlin Dataset 6 emotions	LDC dataset 15 emotions
UL Prosody	28.7%	71.0%	15.7%
CL Prosody	29.7%	68.7%	15.6%
UL Spectral	26.6%	70.1%	15.3%
CL Spectral	36.4%	77.1%	23.7%
Combined	**37.6%**	**80.4%**	**23.9%**

Table 3: *Multi-class emotion classification rates for 6 emotion task on LDC and Berlin datasets, and 15 emotion task on LDC dataset. Best performance is shown in bold.*

reported in [21] on the same task. Note though that the results might not be directly comparable because it is unclear whether the LDC dataset was balanced for the experiments in these earlier studies.

5. Discussion and Future Work

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 utterance-level prosodic descriptors in multi-class emotion recognition. Our experiments also indicate that a combination of prosodic and spectral features can improve the performance of the utterance-level prosodic features used alone.

There are several aspects of feature extraction for emotion recognition that need to be explored in the future research. Firstly, we observed that the overall accuracy of emotion recognition obtained on the 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 the Berlin dataset contains longer utterances. It would be interesting to investigate dependence of emotion recognition accuracy on the utterance length. To the best of our knowledge, this issue has not been explored in the literature.

Secondly, the overall best results were those based on combination of spectral and prosodic features. We expect that the results from the combination would be even greater with more careful feature selection. Such expectation is justified given the high dimensionality of the feature space and the presence of highly correlated features in the combined set, which can hurt performance of machine learning algorithms such as SVM classifiers used in this paper. A way to overcome this limitation which we plan to address in the future work is to employ feature selection algorithms [22] that produce a smaller set of features yielding better recognition performance.

6. 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.

7. References

[1] V. Petrushin, "Emotion in speech: recognition and application to call centers," in *Proceedings of Artificial Neural Networks in Engineering*, 1999, pp. 7–10.

[2] R. Fernandez and R. Picard, "Modeling drivers' speech under stress," *Speech Communication*, pp. 145–159, 2003.

[3] D. J. Litman and K. Forbes-Riley, "Predicting student emotions in computer-human tutoring dialogues," in *Proceedings of Association for Computational Linguistics*, 2004, pp. 352–359.

[4] R. Banse and K. Scherer, "Acoustic profiles in vocal emotion expression," *Journal of Personality and Social Psychology*, vol. 70, no. 3, pp. 614–636, 1996.

[5] F. Dellaert, T. Polzin, and A. Waibel, "Recognizing emotion in speech," in *Proceedings of the CMC*, 1996, pp. 1970–1973.

[6] S. McGilloway, S. Cowie, E. Douglas-Cowie, S. Gielen, M. Westerdijk, and S. Stroeve, "Aproaching automatic recognition of emotion from voice: a rough benchmark," in *ISCA Workshop on Speech and Emotion*, 2000, pp. 200–205.

[7] R. Huang and C. Ma, "Toward a speaker-independent real-time affect detection system," in *International Conference on Pattern Recognition*, 2006, pp. 1204–1207.

[8] T. Tabatabaei, S. Krishnan, and A. Guergachi, "Emotion recognition using novel speech signal features," in *IEEE International Symposium on Circuits and Systems*, 2007, pp. 345–348.

[9] D. Neiberg, K. Elenius, and K. Laskowski, "Emotion recognition in spontaneous speech using gmms," in *Interspeech*, 2006.

[10] O. W. Kwon, K. Chan, J. Hao, and T. Lee, "Emotion recognition by speech signals," *Proceedings of 8th European Conference on Speech Communication and Technology*, pp. 125–128, 2003.

[11] B. Schuller, R. Mller, M. Lang, and G. Rigoll, "Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles," in *Interspeech*, 2005, pp. 805–809.

[12] J. Nicholson, K. Takahashi, and R. Nakatsu, "Emotion recognition in speech using neural networks," *Neural Computing and Applications*, 2000.

[13] T. Nwe, S. Foo, and L. D. Silva, "Speech emotion recognition using hidden Markov models," *Speech Communication*, vol. 41, no. 4, pp. 603–623, 2003.

[14] C. Lee, S. Yildirim, M. Bulut, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, and S. Narayanan, "Emotion recognition based on phoneme classes," in *Interspeech*, 2004, pp. 205–211.

[15] V. Sethu, E. Ambikairaja, and J. Epps, "Phonetic and speaker variations in automatic emotion classification," in *Interspeech*, 2008, pp. 617–620.

[16] Linguistic Data Consortium, "Emotional prosody speech and transcripts," LDC Catalog No.: LDC2002S28, University of Pennsylvania.

[17] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A Database of German Emotional Speech," in *Interspeech*, 2005, pp. 1–4.

[18] J. Odell, D. Ollason, P. Woodland, S. Young, and J. Jansen, *The HTK Book*. Cambridge University Press, 2002.

[19] P. Boersma and D. Weenink, "Praat, a system for doing phonetics by computer," *Glot International*, pp. 341–345, 2001.

[20] C.-C. Chang and C.-J. Lin, *LIBSVM: a library for support vector machines*, 2001, software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[21] S. Yacoub, S. Simske, X. Lin, and J. Burns, "Recognition of emotions in interactive voice response systems," *Proceedings of Eurospeech*, pp. 729–732, 2003.

[22] A. Jain and D. Zongker, "Feature selection: Evaluation, application, and small sample performance," *IEEE Trans. on PAMI*, vol. 19, no. 2, pp. 153–158, 1997.