Emotion Recognition based on Phoneme Classes

Chul Min Lee, Serdar Yildirim, Murtaza Bulut, Abe Kazemzadeh
*Carlos Busso, *Zhigang Deng, Sungbok Lee, Shrikanth Narayanan*

Emotion Research Group –Speech Analysis and Interpretation Lab & Integrated Media Systems Center
Department of Electrical Engineering, *Department of Computer Science
USC Viterbi School of Engineering, University of Southern California
`http://sail.usc.edu`

Abstract

Recognizing human emotions/attitudes from speech cues has gained increased attention recently. Most previous work has focused primarily on suprasegmental prosodic features calculated at the utterance level for modeling against details at the segmental phoneme level. Based on the hypothesis that different emotions have varying effects on the properties of the different speech sounds, this paper investigates the usefulness of phoneme-level modeling for the classification of emotional states from speech. Hidden Markov models (HMM) based on short-term spectral features are used for this purpose using data obtained from a recording of an actress' expressing 4 different emotional states - anger, happiness, neutral, and sadness. We designed and compared two sets of HMM classifiers: a generic set of "emotional speech" HMMs (one for each emotion) and a set of broad phonetic-class based HMMs for each emotion type considered. Five broad phonetic classes were used to explore the effect of emotional coloring on different phoneme classes, and it was found that spectral properties of vowel sounds were the best indicator of emotions in terms of the classification performance. The experiments also showed that the better performance can be obtained by using phoneme-class classifiers than generic "emotional" HMM classifier and classifiers based on global prosodic features. To see the complementary effect of the prosodic and spectral features, the two classifiers were combined at the decision level. The improvement was 0.55% in absolute (0.7% relatively) compared with the result from phoneme-class based HMM classifier.

1. Introduction

In this paper, we investigate the classification of emotional information contained in human speech signals. This topic has been widely studied in psychology and linguistics, and significant progress has been made concerning what emotions are, and how the acoustic speech properties change for different emotional states. Recently, the problem of automatic emotion recognition has gained increased attention, especially because of the desire to develop natural and effective interfaces for human-machine communication applications [1][2]. A recent study has also included emotion as an indicator/signature of a speaker in the context of speech-based content mining applications [3].

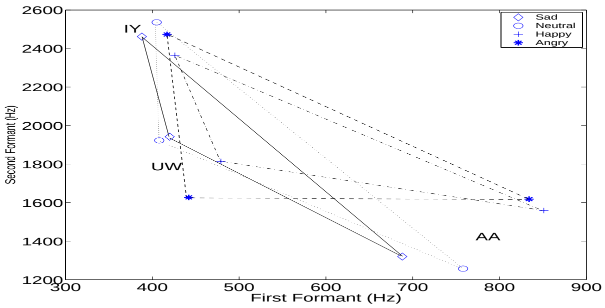

Figure 1: *Emotion dependencies on the vowel triangle: Based on measured first and second formant frequencies for the three vowels, /IY/, /UW/, and /AA/ for various emotional states. Notice the difference in the effects of emotion coloring across different vowels. For example, the first formant frequencies of /AA/ are more affected by emotional coloring, while the second formant frequencies of /IY/ are more affected.*

Despite the progress in understanding the mechanisms of emotions in human speech, progress in the development and design of emotion recognition systems for practical applications is still in its infancy. The reasons behind limited progress in developing an emotion recognition system include: (1) challenges in identifying what signal features are suitable and optimal to achieve reliable recognition; (2) variability arising from a number of, often confounding, sources; for example, variability in emotion can result from inter-speaker differences such as the variation in language and culture. Therefore, it is often difficult to extend and generalize the results obtained from a particular domain and database to other cases. The consensus among researchers has, however, shown that primary emotional states such as the "Big Six" - anger, happiness, sadness, disgust, fear, and surprise - share similarity across culture and language [4] and provide at least a good starting point to explore specific scientific questions in machine emotion recognition. The specific focus of this study is to explore the role of spectral features and phonetic segmental dependencies on emotion recognition.

With regards to the problem of speech signal features for emotion classification, most previous research has used suprasegmental/prosodic features as their acoustic cues. Such cues have been known to be an important indicator to emotional states [5], and thus used in the design of many emotion recognition systems [6][7][8][9]. The spectral information of speech is yet another important feature for representing emotional states, which has been found to be useful for emotion classification [5][10]. One recent study has also showed that there are variations across emotional states in the spectral features at the phoneme level, especially vowel sounds [11], where band-pass filtered Fourier spectra were investigated by self-organizing map. Our study explores this notion further in the context of automatic emotion recognition. Specifically, the central hypothesis of the current study is that different emotional categories affect different phonemes in distinct ways; hence, automatic emotion classification has to incorporate phoneme dependencies.

To motivate the problem, consider Figure 1. It illustrates the vowel triangle for vowels /IY/, /UW/, and /AA/, a plot of their first and second formant frequencies (F1 and F2) calculated from the database of this study (see next section for details) for 4 different emotional states. We can clearly observe distinct constellations for different emotional states supporting our hypothesis that emotions can have different effects on the various phonemes. The goal of this study, hence, is to explicitly model the spectral information at a local (segmental) level for categorizing emotions. We use Mel-frequency cepstral coefficients (MFCCs) as a local spectral feature. They have been widely used in speech recognition because of superior performance over other features, and for providing a high-level approximation of human auditory perception. Incidentally, these cepstrum-related spectral features have also been found to be useful in the classification of stress in speech [12].

In this study, the classification of 4 different emotional states - anger, happiness, neutral, and sadness - was implemented in the framework of HMM classifiers, which help model dynamic changes in the emotional state dependent features in an utterance. Previous studies in emotion recognition have used HMM classifiers based on prosodic features to capture the dynamics of expressed emotions [13][14]. However, potentially each phoneme experiences different coloring for each emotion type and hence has to be considered separately during classification as suggested by [11][12]. Here, we trained five broad phoneme classes - vowel, stop, glide, nasal, and fricative sounds - to investigate the contribution of different emotional states to each phoneme class. For comparison, we also designed a classifier with global prosodic features (utterance level statistics) using support vector machine [15], an approach which has provided some of the most promising results in the current research of emotion.

This paper is organized as follows. Section 2 describes the speech database we used, Section 3 explains the procedure for the training and testing of HMM classifiers. Experimental results are in section 4 and section 5 concludes the paper.

2. Speech Database

The primary data we used in this study were obtained from a recording by a semi-professional female actress. In the recording session, the actress was asked to read the same sentences with 4 emotional states, i.e., anger, happiness, neutral, and sadness. By doing this, we can reduce the variability due to the semantic content of sentences. The recording was made in a quiet room using a close talking SHURE microphone at the sampling rate of 48 kHz. The data acquisition was performed in conjunction with facial expression recording for multimodal emotion recognition. Finally, we collected 880 utterances (250 utterances for anger, 151 for happiness, 216 for neutral, and 263 for sadness). The differences in the number of utterances for each emotion is due to the errors that occurred in the recording session - some portions of the data were corrupted because of the software problem we used in the data collection, and thus could not be further processed. In the experiments, the speech data were downsampled to 16 kHz. A detailed analysis on acoustic correlates of emotional states is given in a companion paper [16].

3. HMM-based Classification of Emotions

We used HMM-based classifiers to identify the emotional state of spoken utterances. First, we created generic "emotional" HMM models of emotions where the observations in the emitting state were modeled by 12 mixture Gaussians (determined empirically with the test set in the data corpus). In addition, the recognition system included a non-speech HMM with similar topology to handle the non-speech portions in the utterances. As mentioned earlier standard Mel-frequency cepstral coefficients (MFCC) were used as the spectral features for emotion recognition. first 13 coefficients were calculated, and delta and acceleration coefficients were added to make the 39-dim feature vectors using 25 ms Hamming windowed samples every 10 ms frame.

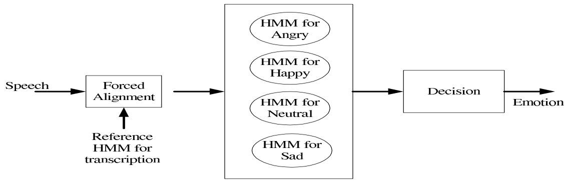

Figure 2: *Emotion recognition system*

Next, we created phoneme class-based HMM classifiers on the MFCC features for emotion recognition. 46 context-independent monophones (derived from TIMIT database) were abstracted into five broad phoneme classes - vowel, glide, nasal, stop, and fricative sounds. Each phoneme class HMM model had 3 emitting states and each state is modeled by 16 Gaussian mixtures.

Let E be the emotion of interest and $\mathbf{O}$ be the observed sequence of feature vectors. Let $\mathbf{S}$ represent a sequence of phoneme class in a given utterance. First, the speech data is forced aligned using phonetic HMM models trained from a large database (TIMIT). We then apply the phoneme-class emotion models, $\lambda_1, \ldots, \lambda_N$, in the same order of the phoneme classes as they occur in an utterance to decode the speech data for each emotion. The decoding is performed using the standard Viterbi algorithm to determine the maximum likelihood segmentations of the phoneme classes, i.e.,

$$\begin{aligned} P(\mathbf{O}|\mathbf{S}, E) \quad &= \max_{t_1, t_2, \ldots, t_T} P(O_1^{t_1}|S_1, E) P(O_{t_1+1}^{t_2}|S_2, E) \\ &\quad \ldots P(O_{t_{N-1}+1}^{t_N}|S_N, E) \qquad (1) \end{aligned}$$

where the observation vector O is given by

$$\mathbf{O} = \{O_1^{t_1}, O_{t_1+1}^{t_2}, \ldots, O_{t_{N-1}+1}^{t_N}\} \qquad (2)$$

and $t_1, t_2, \ldots, t_N$ are the end frame number of each segment. This procedure was performed for each emotion type in parallel.

Given a decoded phoneme class, S, and observation vector, $\mathbf{O}$, we formulated a decision rule based on maximum likelihood as follows:

$$E* = \arg\max_i P(\mathbf{O}|E_i, \mathbf{S}) \qquad (3)$$

where $E*$ denotes the expressed emotion of an utterance and i represents the number of emotions of interest. Figure 2 shows a blockdiagram for this emotion recognition system.

4. Experimental Results

4.1. Human Evaluation Results

For evaluation, we performed human listening tests with 4 untrained, English-speaking listeners. The confusion matrix of the subjective human evaluation is given in Table 1. A total of 100 utterances (25 from each emotion) were randomly selected from the database and played to all the listeners. The listeners were asked to identify the emotional state of utterances as one of the 4 emotions or "Other", which was given as a choice when the utterance did not seem to belong to the 4 specified categories.

	Ang	Hap	Neut	Sad	Other
Ang	82	2	3	1	12
Hap	12	56	7	6	19
Neut	8	1	74	14	3
Sad	5	1	20	61	13

Table 1: Confusion matrix from subjective human evaluation. Columns represent the emotion selected for utterances from the emotion input of each row. Ang stands for anger, Hap for happiness, Neut for neutral, and Sad for sadness. The total number of utterances for each emotion was 100.

The results of the evaluation were moderate: 68.3% of the utterances were correctly identified, 77.3% when "Other" option was excluded. One study reported 80% correction rate in human evaluation test for 7 emotions in the data recorded from actors [14]. The results in Table 1 showed that most errors occurred in two different sets of emotions: one is the confusion between anger and happiness and the other is between sadness and neutral emotion. From the human evaluation test, happiness is the most difficult category to discriminate in our database, which had the lowest accuracy rate. The number of "Other" choice, i.e., the number of undecided utterances, showed that the emotion of happiness had the difficulty to distinguish from other emotional categories. This trend was observed throughout the following classification experiments.

4.2. Support Vector Machine Classifier with Global Prosodic Features

To provide comparisons with other recent studies, we also designed a classifier using prosodic features. The classifier used here was a support vector machine classifier (SVC) with 2nd order polynomial kernel functions [15]. SVC was used in emotion recognition in the previous study and showed better performance than other statistical classification methods such as linear discriminant classification or nearest neighborhood classification [17]. Similarly, SVC performed the best in our data, and thus used in this experiment. The prosodic features were pitch-related features and speech rate. The F0 contours were calculated using the `ESPS` pitch tracker `get_f0` and the duration of each phoneme was obtained by segmentation information from the speech recognizer so as to obtain the speech rate. From the F0 contour, the mean, standard deviation, maximum and minimum values were calculated at the utterance level. Also we included F0 range, which was defined as the difference between mean and the 90%-quartile of the F0 contour. We also calculated the slopes of the F0 contour by using a moving window along the contour, from which the mean and maximum values of the slopes were calculated. Speech rate was computed as the reciprocal of the number of words per second.

The classification result for SVC with prosodic features is shown in Table 4 in terms of accuracy rate. The accuracy rate was 55.68%. Note that the prosodic features used in this experiment mainly come from F0. The performance could be further improved including much wider range of prosodic features [8].

4.3. Hidden Markov Model Classifiers

In the experiments with HMM classifiers, the speech data were divided into training and test data. The training data had 704 utterances and test data had 176 utterances. The training and testing of HMM classifiers for both generic "emotional" HMM and phoneme-class dependent HMMs were performed using the Hidden Markov Toolkit (HTK) [18]. The spectral features for HMM classifiers were MFCCs including delta and acceleration coefficients to make 39 dimensional feature vectors.

	Angry	Happy	Neut	Sad
Ang	47	0	1	1
Hap	19	6	0	1
Neut	0	0	28	13
Sad	1	0	26	33

Table 2: Confusion matrix of generic "emotional" HMM classifier. The results are from the test data, which has 176 utterances.

	Angry	Happy	Neut	Sad
Ang	48	0	0	1
Hap	17	9	0	0
Neut	0	0	26	15
Sad	0	0	10	50

Table 3: Confusion matrix for phoneme-class dependent HMM classifier. The results are from the test data, which has 176 utterances.

In the training of phoneme-class dependent HMM classifiers, we bootstrapped the initial HMM models using the TIMIT data corpus because of insufficient data for the relevant emotion categories in our data. We then adapted the mean and variance in the HMM models for the phoneme classes of each emotional state using maximum likelihood linear regression (MLLR) method [18]. For testing, forced alignment was performed over the utterance, and then likelihoods from the emotion based models were compared to determine the emotional state which maximized the likelihood. The confusion matrices for both generic "emotional" HMM and phoneme-class dependent HMM classifiers are given in Tables 2 and 3. The confusion matrix results show that there are large confusions between anger and happiness, and between neutral and sadness. This trend is similar to that observed in the results of human evaluation. Note that in our data corpus, anger is the most salient emotional state, which has the highest accuracy rate compared with other emotional states.

The classification results in terms of accuracy rate are given in Table 4. The best result was obtained from the phoneme-class dependent HMM classifier with MFCC features compared with both the SVC with prosodic features and the generic "emotional" HMM classifier. This supports our original hypothesis that phoneme-level modeling provides better discriminability for emotion recognition.

We also calculated the accuracy rate in each specific phoneme class. In this case, the decision was made by comparing the average frame log likelihoods.

$$E* = \arg\max_{i} \frac{1}{M} \sum_{n=1}^{M} \log P(\mathbf{O}_n | E_i, \mathbf{S}) \qquad (4)$$

where M is the number of frames for a given phoneme class in the utterance and S represents a sequence of given phoneme class, e.g., vowel. The classification results from each phoneme class show that vowel sounds are good indicators for emotion recognition. As people express emotion in speech, the articulation and vocalization mechanism for the emotional states change. Differences in emotion perception for different sounds indicate a complex interplay between the underlying speech articulations related to linguistic and emotional expression. Vowel productions, characterized by open vocal tracts and the less constrained articulation, not surprisingly show the greatest effects of emotion coloring. Furthermore, the relatively less constricted low vowels such as /AA/ show greater effects than do high vow-

Classification Method		Accuracy (%)
SVC with prosodic features		55.68
generic “emotional” HMM		64.77
Phoneme-class dependent HMM	every phoneme class	75.57
	vowel only	72.16
	glide only	54.86
	nasal only	47.43
	stop only	44.89
	fricative only	55.11
Combination of prosody and phoneme-class classifier		76.12

Table 4: Classification accuracy for different classifiers. The results were obtained from the test data (176 utterances).

els like /IY/ (refer to 1). On the other hand, non-continuant stop sounds seem to carry the least emotional information.

We also calculated the accuracy rate by combining the prosodic feature based SVC classifier and phoneme-class based HMM classifier and the result are shown in Table 4. The combination is performed by averaging the outputs from the classifiers at the decision level; i.e., the average of the posterior probabilities of emotion given an utterance. The improvement is modest (0.55% better than the HMM classifier in absolute value and 0.7% relative). The reason for this was that the likelihood values of emotions in the HMM classifier had large differences compared with those in the global prosodic feature classifier because they are accumulated along the frames in the utterance, and thus the decision made by HMM classifier was dominated over that from prosodic feature classifier.

5. Conclusions

This paper investigated emotion recognition from speech signals using phoneme-class dependent HMM classifiers with short-term spectral features. The results showed that spectral features play a significant role in emotion recognition. Because the shape of vocal tract can potentially change under different emotional states, the spectral characteristics of speech differ for various emotions even when people speak the same sentence [12]. We should however note that the extent of such vocal tract changes depend on the type of the speech sound being articulated – the degree of control and constraint involved in their shaping – and hence the extent of the potential emotion coloring. The less constrained the articulation, the more the effect emotions will have, as borne out by the classification results. Evidence for this speculation requires direct articulatory information, a topic for future work.

In summary, the advantage of HMM classifiers over other statistical classifiers, such as SVC in this work, is that they can model dynamic changes of acoustic features in given emotional state. By designing phoneme class HMM classifiers, we can investigate the effect of different emotional coloring on each phoneme class. The results in Table 4 showed that the vowel sounds were shown to be an important indicator of emotions in terms of classification accuracy. The combined decision of the prosodic feature-based classifier and the phoneme-class spectral feature HMM classifier at the decision level showed only modest improvement due to the different dynamic ranges of the likelihoods. Both prosodic and spectral features play significant roles in emotion recognition and we need to explore the way to effectively and complementarily combine those information sources to further improve the classification performance.

There are several issues that need to be further explored. First, prosodic features such as pitch, energy, and duration play an important role in emotion expression, and they should be efficiently combined with spectral features in order to improve the performance of the emotion recognition system. Another interesting direction is the design of a multimodal emotion recognition system. The database used in this work also provides video information - facial expression - in addition to speech, which can be used to further improve the classification performance. For example, happiness and anger were the emotional state pair most confused in the confusion matrix results; however, a preliminary result showed that the facial expression of those emotions were significantly different and thus overcame the errors that occurred in the speech-only classification. These are topics of ongoing investigation.

6. References

[1] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor, “Emotion recognition in human-computer interaction,” *IEEE Sig. Proc. Mag.*, vol. 18(1), pp. 32–80, Jan 2001.

[2] C. M. Lee and S. Narayanan, “Towards detecting emotions in spoken dialogs,” *IEEE Trans. on Speech & Audio Processing*, in press, 2004.

[3] I. Shafran, M. Riley, and M. Mohri, “Voice signatures,” in *ASRU ’03*, 2003.

[4] K. Scherer, “A cross-cultural inverstigation of emotion inferences from voice and speech: Implications for speech technology,” in *Proc. ICSLP 2000*, Beijing, China, Oct 2000.

[5] R. Banse and K. Scherer, “Acoustic profiles in vocal emotion expression,” *J. Personality Social Psych.*, vol. 70(3), pp. 614–636, 1996.

[6] J. Ang, R. Dhillon, A. Krupski, E. Shriberg, and A. Stolcke, “Prosody-based automatic detection of annoyance and frustration in human-computer dialog,” in *Proc. ICSLP 2002*, Denver, CO, Sep 2002.

[7] R. Tato, R. Santos, R. Kompe, and J. Pardo, “Emotional space improves emotion recognition,” in *Proc. ICSLP 2002*, Denver, Co, Sep 2002.

[8] C. M. Lee, S. Narayanan, and R. Pieraccini, “Recognition of negative emotions from the speech signal,” in *Proc. Automatic Speech Recognition and Understanding*, Dec 2001.

[9] D. Litman and K. Forbes, “Recognizing emotions from student speech in tutoring dialogues,” in *ASRU ’03*, 2003.

[10] K. R. Scherer, “Vocal communication of emotion: A review of research paradigms,” *Speech Comm.*, vol. 40.

[11] L. Leinonen and T. Hiltunen, “Expression of emotional-motivational connotations with a one-word utterance,” *J. Acoust. Soc. Am.*, vol. 102(3), pp. 1853–1863, Sep 1997.

[12] J. Hansen and B. Womack, “Feature analysis and neural network based classification of speech under stress,” *IEEE Trans. on Speech & Audio Processing*, vol. 4(4), pp. 307–313, Jul 1996.

[13] B.-S. Kang, C.-H. Han, S.-T. Lee, D.-H. Youn, and C. Lee, “Speaker dependent emotion recognition using speech signals,” in *Proc. ICSLP 2000*, Beijing, China, Oct 2002.

[14] A. Nogueiras, A. Moreno, A. Bonafonte, and J. B. Marino, “Speech emotion recognition using hidden markov models,” in *Eurospeech’01*, 2001.

[15] C. Burges, “A tutorial on support vector machines for pattern recognition,” *Dat Mining and Know. Disc.*, vol. 2(2), pp. 1–47, 1998.

[16] S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, and S. Narayanan, “An acoustic study of emotions expressed in speech,” submitted to ICSLP’04.

[17] C. M. Lee, S. Narayanan, and R. Pieraccini, “Classifying emotions in human-machine spoken dialogs,” in *ICME’02*, 2002.

[18] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *the HTK Book 3.2*, 2002.