Emotion recognition using neutral acoustic speech models

Carlos Busso¹, Emily Mower¹, Takayuki Hidehira ²

Affective Computing (CSCI534)
¹Electrical Engineering Department, ²Department of Computer Science
University of Southern California, Los Angeles, CA 90089
{busso, mower, hidehira}@usc.edu

Abstract

Since emotional speech can be regarded as a variation of neutral (non-emotional) speech, it is expected that a robust neutral speech model can be useful in contrasting different emotions expressed in speech. The study presented explores this idea by creating acoustic models trained with spectral features, measured from the emotionally-neutral TIMIT corpus. The performance is tested using two emotional speech databases: one recorded with a microphone (acted), and another recorded from a telephone application (spontaneous). This method obtained accuracy levels of up to 78% and 65%, respectively, can be achieved in the binary and category emotion discriminations. Raw Mel Filter Bank (MFB) output outperformed the conventional MFCC, with both broad-band and telephone-band speech. These results suggest that well-trained neutral acoustic models can be effectively used as a front-end for emotion recognition, and once trained with MFB, it may work well regardless of the channel characteristics.

Index Terms: Emotion recognition, Neutral speech, HMMs, Mel filter bank (MFB), TIMIT

1. Introduction

Detecting and utilizing non-lexical or paralinguistic cues from a user is one of the major challenges in the development of usable human-machine interfaces (HMI). Notable among these cues are the universal categorical emotional states (e.g., angry, happy, sad, etc.), prevalent in day-to-day scenarios. Knowing such emotional states should help to adjust system responses so that the users interacting with the system can be more engaged and have a more effective interaction with the system.

Identifying user’s emotion from the speech signal is desirable. The recording and feature extraction process from this modality is comparatively easier and simpler than that of other modalities, such as facial expression and body posture. Previous studies on the automatic categorization of emotional speech have shown accuracy between 50% and 85% depending on the task (e.g. number of emotion labels, number of speakers, size of database) [1]. A comprehensive review of the current approaches is given in [2].

However, such emotion categorization performance is largely specific to individual databases (usually off-line) and it is not plausible to easily generalize the results to different databases or on-line recognition tasks. This is due to inherent speaker-dependency in emotion expression, acoustic confusion among emotional categories, and differences in acoustic environments across recording sessions. It is also fairly difficult, if not infeasible, to collect enough emotional speech data so that one can train robust and universal acoustic models of individual emotions. This problem is further complicated when one considers that there exist more than a dozen emotional categories whose possible combinations can be used to differentiate affective states and attitudes [3].

As a possible way to circumvent the fundamental problem in emotional categorization based on speech acoustics, this study tests a novel idea of discriminating emotional speech against neutral (i.e., non-emotional) speech. That is, instead of training individual emotional models, we build a single, neutral speech model and use it for emotion evaluation either in the categorical approach or in the dimensional approach [3]. This idea is based on the assumption that emotional speech production is a variant of the non-emotional counterpart in the (measurable) feature space. It has been shown that the speech rate, the speech duration, the fundamental frequency (F0) and the RMS energy are simultaneously modulated to convey emotional information [4]. In the articulatory domain, it has been shown that the tongue tip, jaw and lip kinematics are different from neutral speech during expressive speech production [5, 6]. Hence, modeling the differential properties with respect to neutral speech is hypothesized to be advantageous. In addition, robust neutral-acoustic speech models can be built because there are many neutral speech corpora. This paper presents our first attempt to examine this idea.

In this preliminary report, the TIMIT database is used to train neutral acoustic models and two emotional speech databases are probed. Hidden Markov Models (HMMs) are trained with two different acoustic feature sets, Mel Filter Bank (MFB) and Mel-Frequency Cepstrum Coefficients (MFCCs), and their behaviors are examined in a broad phonetic-class recognition experiment setting based on recognition likelihood scores. Emotional discrimination performance by the two feature sets are also investigated and compared using a discriminant analysis. The results show that binary emotion recognition accuracies of up to 78% can be achieved using only these features.

Building upon these ideas, this paper presents a semi real-time system to detect emotional speech. Based on the accuracy of the system, we claim that this approach is suitable to detect emotional speech in real applications.

This paper is organized as follows. Section 2 describes the proposed approach. Section 3 analyzes the likelihood scores from emotional and non-emotional speech obtained with the neutral models. Section 4 provides a discriminant analysis of the likelihood scores. Section 5 describes the implementation of this system. Finally, Section 6 gives a discussion and future direction of this work.
The proposed emotional segmentation approach is a two step process, as described in Figure 1. In the first step, neutral models are built to measure the degree of similarity between the input speech and the reference neutral speech. The output of this block is a fitness measure of the input speech. In the second step, these measures are used as features to infer whether the input speech is emotional or neutral. The primary focus of this paper is placed on the first block.

The popular read-speech TIMIT database was used as a reference for neutral speech [11]. This database contains 4620 sentences for the training set, and 1680 for the testing set, collected from 460 speakers. The nature, size, and the inter-speaker variability make this database suitable to train the proposed “neutral-speech” models. Two corpora are used as emotional databases. The first one is the EMA database [5], in which three subjects read 10 sentences five times portraying four emotional states: sadness, anger, happiness and neutrality. Although this database contains articulatory information, only the acoustic signals were analyzed. The EMA data have been perceptually evaluated and inter-evaluator agreement is 81.9% [12].

The second corpus was collected from a call center application [13]. It provides spontaneous speech of different speakers from a real human-machine application. The data was labeled as negative or non-negative (neutral). Only the sentences with high agreement between the raters were considered (1027 neutral, 338 negative). This database is referred to from here on as call center database (CCD). More details of these two emotional databases can be found in [5, 13], respectively.

The sampling rate of the TIMIT database is 16KHz, while the sampling rate of the CCD corpus is 8KHz (telephone speech). To compensate this mismatch, the TIMIT database was downsampled to 8KHz to train the reference neutral models used to assess the CCD corpus. In contrast, for the EMA corpus the broad band TIMIT data was used for training, since its speech files were also recorded at 16KHz.

### 3. Analysis of the likelihood scores

#### 3.1. MFB-based neutral speech models

After the MFB-based neutral models were built with the TIMIT training set, the likelihood scores of the emotional testing corpora were computed. Figures 2 and 3 show plots with the mean and standard deviation of the likelihood scores for the broad phonetic categories, obtained with the EMA and CCD databases, respectively. For reference, the likelihood scores for the TIMIT testing set were also plotted.

These figures reveal that the mean and the variance of the likelihood score for emotional speech differ from the results observed in neutral speech, especially for emotion with high level of arousal such as anger and happiness. We also observed that some broad phonetic classes present stronger differences than others. For example, front vowels present distinctive emotional modulations. In contrast, the likelihood scores for nasal sounds are similar across emotional category (see Fig. 4-c), suggesting that during articulation there are not enough degrees of freedom to convey emotional modulation. These results agree with our previous work, which indicated that emotional modulation is not displayed uniformly across speech sounds [4, 5].
However, it has been shown that with speech features the neutral and sad speech mainly differ in the valence and prosodic features [4, 12]. One possible explanation is that emotional categories with other acoustic speech features (spectral trends were observed in our previous work between these emotional classes) are denoted by $j\in\{B, F, D, L, N, T, C\}$; and the emotional information conveyed in the likelihood scores and which corresponds to classification (Fig. 1). We analyze the emotional modulation is observed, these figures show that the differences between the likelihood scores for emotional categories are not as strong as the differences obtained with the MFB-based models (note the values in the vertical axes). Interestingly, MFCCs are calculated from MFB, by applying the Discrete Cosine Transform (DCT) over the Mel log-amplitudes. This post-processing step seems to blur the acoustic differences between emotional and neutral speech.

4. Discriminant analysis

In this section, we describe the second step of the framework, which corresponds to classification (Fig. 1). We analyze the emotional information conveyed in the likelihood scores and discuss whether they can be used to segment emotional speech automatically.

Figures 2, 3, 5, and 6 show that the means and the standard deviations of the likelihood scores differ from the values obtained with neutral speech. In this experiment, the average of these measures at the sentence level for each broad phonetic class was used as features for emotion recognition. If the emotional classes are denoted by $C$ and the features for the phonetic class $i_j \in \{F, B, D, L, N, T, C\}$ are denoted by $L_h_{i_j}$, the classification problem can be formulated as:
Table 2: Discriminant analysis of likelihood scores for CCD database (Neg=negative, Neu=neutral)

<table>
<thead>
<tr>
<th>Ground truth</th>
<th>MFB-based models</th>
<th>MFCC-based models</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Neg</td>
<td>Neu</td>
</tr>
<tr>
<td>Sad</td>
<td>0.42</td>
<td>0.16</td>
</tr>
<tr>
<td>Ang</td>
<td>0.58</td>
<td>0.84</td>
</tr>
</tbody>
</table>

\[
P(C|\text{Obs}) = \prod_{j=1}^{N} P(C|Lk_i)
\]

(3)

In our previous work, an accuracy of 66.9% was achieved for a similar task by using many acoustic features [12]. These two experiments reveal that with this approach neutral speech can be accurately separated from emotional speech when there is a high level of arousal (i.e., happiness and anger). For the MFCC-based neutral models, the performance decreases measurably. These results agree with the analysis presented in Section 3.

For the CCD corpus, the emotional classes considered in the experiment were negative versus neutral speech. The results for the MFB-based neutral models were 42% for the negative class, and 84% for neutral class (63% is the average). The performance for MFCC-based models was slightly lower than with MFB: 38% for the negative class, and 85% for neutral class (61.5% is the average). As a reference, our previous work has reached approximately 74% accuracy for the same task, by using many different acoustic speech features, including prosodic features [13]. One reason that the performance is worse in the CCD data than in the EMA data is that most of the sentences are short with only one word (median duration is approximately 1.5 seconds). This issue affects the accuracy of the features extracted from the likelihood scores.

5. Implementation

The system was implemented using HTK. HTK has a real-time option, which was the basis for the first approach. This function performs recognition on speech recorded from a microphone. As described in Section 2, the input speech must be normalized to
decrease the effects of the differences between the recording conditions in the model database and the input speech. However, this function does not allow for real-time normalization. Absent this normalization, the output likelihood values are very low, detracting from the classification performance.

This problem was addressed through the inclusion of a sequential program that normalizes the acoustic signal, extracts the features, recognizes the broad phonemes classes, and then uses the likelihood scores to classify emotions. Since the sentence needs to be recorded before processing, this approach is semi real-time (less than 1 second of delay after the sentence is recorded).

To train the classifier, we recorded a small corpus, using TIMIT database. To compute scale factor the same equipment used for the semi real-time system. We focused only on non-negative emotions with a high level of arousal (happy and angry) since, as previously discussed, this system performs well only on arousal (not valence) classification tasks. In total, the corpus contains 22 angry sentences, 21 happy sentences and 26 neutral sentences spoken by a single female subject.

Given the input feature vector $X$ and the recognized phonetic class $ph \in \{F, B, D, L, T, C\}$, the classifier computes the probability of the two classes $P_{ph}^{emo}$ (probability that input speech is emotional) and $P_{neu}^{ph}$ (probability that input speech is neutral) as follows:

$$P_{ph}^{emo} = \exp \left( -\frac{1}{2} (X - \mu_{emo})^T \Sigma_{ph}^{-1} (X - \mu_{emo}) \right)$$

$$P_{neu}^{ph} = \exp \left( -\frac{1}{2} (X - \mu_{neu})^T \Sigma_{ph}^{-1} (X - \mu_{neu}) \right)$$

where $\Sigma_{ph}$, $\mu_{emo}^{ph}$ and $\mu_{neu}^{ph}$ are the parameters of the classifiers. $P_{ph}^{emo}$ and $P_{ph}^{neu}$, are normalized such that they sum to one.

After this calculation, each phoneme in each sentence is described by a set of probabilities ($P_{ph}^{emo}$ and $P_{ph}^{neu}$). To determine the sentence level classification, the probabilities for each phoneme class in a single sentence are added. This provides a more robust sentence-level classifier. The resulting probabilities, $P_{emo}$ and $P_{neu}$, are normalized such that they sum to one. The sentences is assigned to class represented by the highest probability. For example, if $P_{emo}$ is greater than $P_{neu}$, the sentence is classified as an emotional utterance.

The flow of this system is controlled by a simple perl script, which first calls the initialization step and then performs recognition until the user ends the script.

Figure 7: Diagram of the semi real-time emotion recognition system.

Figure 7 shows a diagram of the system. This system includes two main blocks: initialization and recognition. In the initialization step, the users are asked to speak in a neutral tone of voice. The system records 15 seconds of audio, and computes the average energy ($E_{input}$). This value is compared with the previously computed energy of the TIMIT database. The system then computes the scaling factor $s$. This factor serves to modify the amplitude of the input signal to improve the match between the neutral input data and the neutral training data. This variable is estimated as:

$$s = \sqrt{\frac{E_{ref}}{E_{input}}}$$

where $E_{ref}$ (2.7608e-4) is the average energy over the entire TIMIT database.

In the recognition step, 10 seconds of speech are recorded. The amplitude of the collected speech is multiplied by $s$ after which the MFB outputs are computed (HCopy). HVite is then used to recognize the broad phonemic classes. This function outputs both the recognized phoneme and the likelihood of the assigned class. The average and standard deviation of these scores are used for binary emotion classification, as discussed in Section 2. The likelihood scores for the nasal sounds do not present significant emotional differentiation (Figures 2 and 3) and were not considered in this semi real-time system.

The linear discriminant classifier (LDC) was used for emotional classification (Section 4). This technique assumes that the samples from the classes are Gaussian distributed with similar covariance matrices but different means.

To train the classifier, we recorded a small corpus, using the same equipment used for the semi real-time system. We focused on prosodic features such as pitch and energy, which have been shown to convey emotional information. A challenging question is how to normalize those features to remove inter-subject, inter-gender, and inter-recording differences, preserving inter-emotional discrimination.

Although this framework addresses binary emotion classification, the proposed scheme can be used as a first step, in a more sophisticated emotion recognition system. After detecting emotional speech, a second level classifier can be used to achieve a finer emotional description of the speech.

6. Discussion and Conclusions

This paper presents a new approach to classifying emotional versus non-emotional speech by using neutral reference models. The results show that this approach can achieve accuracy up to 78% in the binary emotional classification task across four emotion categories. These results suggest that well-trained neutral acoustic models can be effectively used as a front-end for emotion recognition. Interestingly, the models trained with conventional MFCCs are found to perform worse than the models with the original MFBs for both emotional databases, suggesting that MFB-based models will achieve better performance regardless of the speech and recording characteristics.

These ideas were used to implement a semi real-time binary emotion recognition system. Although a small single-user corpus was used to train the system, the accuracy of the system is notably high across emotions and speakers. These results suggest that this system can reliably detect emotional speech in a real application.

The proposed approach can be enhanced and expanded in several directions. First, the proposed framework can be applied to prosodic features such as pitch and energy, which have been shown to convey emotional information. A challenging question is how to normalize those features to remove inter-subject, inter-gender, and inter-recording differences, preserving inter-emotional discrimination.

Although this framework addresses binary emotion classification, the proposed scheme can be used as a first step, in a more sophisticated emotion recognition system. After detecting emotional speech, a second level classifier can be used to achieve a finer emotional description of the speech.
The results indicated that some broad phonetic classes present more emotional differentiation than others. In general, vowels seem to have more degrees of freedom to convey emotional information than phonemes such as nasal sounds. These observations can be used for recognition by weighting the features extracted from the likelihood scores according to the observed emotional modulation conveyed in them.

Notice that the detected phoneme classes are the ones that maximize the likelihood in the Viterbi decoding path. If the transcript of the sentences is known, which may not be true for all real time applications, the phoneme recognition accuracy could also be used as a fitness measure for emotional discrimination. The hypothesis here is that the performance will be higher for neutral speech, compared to emotional speech. This approach could be extremely useful to automatically find emotional speech portions in larger scripted emotional corpora.

Figures 2, 3, 5, and 6 reveal that likelihood scores for the neutral set in the emotional corpus is close to the results for the neutral references. To reduce the mismatches between the neutral corpora even further, the acoustic models in the HMMs can be adapted to the neutral set in the emotional speech. These are some topics that we are currently pursuing.

Another approach to reduce the mismatches is to consider a better references corpus. The TIMIT database was recorded under perfect acoustic condition, and the subjects were asked to read the sentences. This setting may differ from spontaneous neutral speech recorded in a more realistic scenario. We are currently working in these interesting areas. We believe that building upon this approach, a robust emotion recognition system can be designed, which will be able to classify emotions not only in controlled experimental environment, but also in real time applications.

7. References


