Acoustic feature analysis in speech emotion primitives estimation

Conference Paper · January 2010
Source: DBLP

CITATIONS
23

READS
53

3 authors:

Dongrui Wu
DataNova
101 PUBLICATIONS 3,980 CITATIONS
SEE PROFILE

Thomas D Parsons
University of North Texas
222 PUBLICATIONS 3,405 CITATIONS
SEE PROFILE

Shrikanth S Narayanan
University of Southern California
877 PUBLICATIONS 14,223 CITATIONS
SEE PROFILE

Some of the authors of this publication are also working on these related projects:

FUZZIEEE 2017 - Special Session on Fuzzy Systems for Physiological and Affective Computing - Call For Papers View project

Speechlinks View project

All content following this page was uploaded by Dongrui Wu on 18 July 2017.

The user has requested enhancement of the downloaded file.
Acoustic Feature Analysis in Speech Emotion Primitives Estimation

Dongrui Wu\textsuperscript{1,2}, Thomas D. Parsons\textsuperscript{1}, Shrikanth S. Narayanan\textsuperscript{2}

\textsuperscript{1}University of Southern California, Institute for Creative Technologies, Marina del Rey, CA, USA
\textsuperscript{2}Signal Analysis and Interpretation Lab, University of Southern California, Los Angeles, CA, USA
dongruiw@usc.edu, tparsons@ict.usc.edu, shri@sipi.usc.edu

Abstract

We recently proposed a family of robust linear and nonlinear estimation techniques for recognizing the three emotion primitives—valence, activation, and dominance—from speech. These were based on both local and global speech duration, energy, MFCC and pitch features. This paper aims to study the relative importance of these four categories of acoustic features in this emotion estimation context. Three measures are considered: the number of features from each category when all features are used in selection, the mean absolute error (MAE) when each category is used separately, and the MAE when a category is excluded from feature selection. We find that the relative importance is in the order of $\text{MFCC} > \text{Energy} \approx \text{Pitch} > \text{Duration}$. Additionally, estimator fusion almost always improves performance, and locally weighted fusion always outperforms average fusion regardless of the number of features used.

Index Terms: Emotion estimation, 3D emotion space, speech analysis, estimator fusion, support vector regression, robust regression, locally linear reconstruction, locally weighted fusion

1. Introduction

Emotions may be recognized from many different information sources, e.g., speech [1–3], facial expressions [4, 5], physiological signals [6, 7], or their multimodal combination [8, 9]. In this paper we focus on emotion recognition from speech signals.

A majority of research on speech emotion recognition classifies emotions into a small number of categories [1, 10, 11]. Emotion psychology research [12–14] has shown that emotions can also be represented as points in a multi-dimensional space, i.e., emotions can be quantified as continuous numbers instead of categorical values. One of the most frequently used emotion spaces consists of three primitives [14, 15] of Valence, Activation, and Dominance. This 3D representation is easier to implement than categories because, though computing with words [16, 17] is possible, computers are better at dealing with numbers for deriving inferences and decision making. This 3D representation of emotions is used in this paper.

We [3] have recently introduced three elementary models and two fusion approaches for estimating speech emotions in 3D space using the VAM corpus [18]. Four categories of acoustic features widely adopted in the literature (duration, energy, MFCC, and pitch) were used. The three elementary models had comparable performance with the state-of-the-art results [2, 19], and the two fusion models outperformed them. This paper aims to provide more insights on features that are more important for speech emotion estimation in the 3D space using the newly proposed estimators. Particularly, it studies the relative importance of the four categories of acoustic features using three different measures.

Feature importance is a key topic in speech emotion recognition. It has been investigated in a number of papers [11, 20–24]. However, all of them study the importance of acoustic features in classifying emotions into categories. In contrast, this paper investigates feature importance in estimating the continuous values of emotions in 3D space, a topic that has not been widely addressed. Our novel contributions are as follows:

1. We introduce three measures to evaluate the relative importance of the four categories of acoustic features in emotion estimation.
2. We use three elementary estimators in each of the three measures, and also two fusion models in the latter two measures.

The rest of this paper is organized as follows: Section 2 briefly introduces our experiment setup. Section 3 presents our experimental results. Finally, Section 4 draws conclusions and proposes some future works.

2. Experiment Setup

The VAM corpus [18] was used in this paper. It contains spontaneous speech with authentic emotions recorded from guests in a German TV talk-show Vera am Mittag (Vera at Noon in English). There are 947 emotional utterances from 47 speakers (11m/36f). Each utterance was evaluated by 6–17 human listeners in the 3D space of valence, activation and dominance. Then, their evaluations in each dimension were aggregated to obtain a number in [-1, 1]. So, the emotion of each utterance is represented by a 3D vector, which serves as our reference.

46 acoustic features, which are listed in the second column of Table 1, were extracted. They are the same as those used in [2, 3] on the same corpus and cover four major categories:

- **Duration features** (5): mean and standard deviation of the duration of voiced and unvoiced segments, ratio between the duration of unvoiced and voiced segments.
- **Energy features** (6): energy mean, standard deviation, maximum, 25% and 75% quantiles, and the inter-quartile distance.
- **MFCC features** (26): mean and standard deviation of 13 Mel frequency cepstral coefficients (MFCC).
- **Pitch features** (9): f0 mean, standard deviation, median, minimum, maximum, range, 25% and 75% quantiles, and the inter-quartile distance.

Three elementary estimators—robust regression (RR), support vector regression (SVR), and locally linear reconstruction (LLR), and two estimator fusion approaches—average fusion
(AF) and locally weighted fusion (LWF), were constructed, as shown in Fig. 1. They are exactly the same as those reported in our previous study [3]. The three elementary estimators were chosen because they are diverse and complementary, in the sense that they cover both local model (LLR) and global models (RR and SVR), and both linear models (RR and LLR) and nonlinear model (SVR). So, we expect better performance when they are fused properly. AF takes the average of the three elementary estimators. LWF computes the weighted average of RR, SVR and LLR, where the weights are adaptive and dependent on the local performance of the elementary estimators.

For feature selection, we first ranked the features using the iterative sequential backward selection method [25], and then performed cross-validation according to the rank of the features to select the best subset. We started with all 46 features and gradually removed the worst feature until the 10-fold cross-validation performance stopped improving.

3. Experimental Results

The mean absolute error (MAE) between the estimates, \( \hat{y}_n \), and the human evaluations, \( y_n \), i.e., \( MAE = \frac{1}{n} \sum_{n=1}^{n} |\hat{y}_n - y_n| \), was used in performance evaluation. It has also been used in [2,3,19]. Three measures based on MAE were employed in this paper:

1. Measure1: The number of features from each category when all four categories are used in feature selection.
2. Measure2: The MAE when each category of features are used separately in feature selection.
3. Measure3: The MAE when a category is excluded from feature selection (the remaining three categories are used together).

Experimental results on the three measures are presented next.

3.1. Experimental Results for Measure1

For Measure1, we used the number of features as an importance indicator and assume that the more features are selected from a certain category, the more important that category is. The features used by RR, SVR and LLR for estimating the three emotion primitives are shown in Table 1. Observe that:

1. Generally all four categories of features contributed in emotion primitives estimation. This is consistent with Bathliner et al.’s [21] observation when these categories were used in emotion classification.
2. The optimal feature subsets for the three elementary estimators were different for each emotion primitive.
3. For RR, the importance of the four categories was in the order of MFCC > Pitch > Energy > Duration.

![Table 1](image-to-url)

**Table 1:** The features used by RR, SVR and LLR for estimating the three emotion primitives. The rank of a feature in each category is determined by the number of times it appeared in the nine cases (shown in the last column); so, generally a higher rank indicates higher importance.
3.2. Experimental Results for Measure2

For Measure2, each category of features was used separately in feature selection. So, by comparing the MAEs of the models obtained from different categories, the importance of the four categories is directly evaluated. A comparison of the performances is shown in Fig. 2. Note that here a smaller MAE means a better performance and hence higher importance. Observe that:

1. For Valence, the importance of the four categories was in the order of MFCC > Pitch > Energy > Duration.
2. For Activation and Dominance, the importance of the four categories was in the order of MFCC > Energy > Pitch > Duration.
3. SVR always had the smallest MAE among the three elementary estimators, and LLR often had the largest.
4. When the number of features was small (i.e., when duration, energy, and pitch were used separately), LLR had much worse performance than RR and SVR, and fusion did not necessarily reduce the MAE.
5. LWF always outperformed AF, regardless of the number of features used.

In summary, the importance of the four categories was in the order of MFCC > Energy ≈ Pitch > Duration, which is consistent with our finding in Measure1.

4. Conclusions and Future Works

In this paper, we have compared the relative importance of four categories of acoustic features in speech emotion estimation,
using three elementary estimators and two fusion models. The main findings are:

1. For all three emotion primitives and all five models, MFCC features were always the most important. This finding is consistent with Vogt and Andre’s observation [20] in emotion classification using spontaneous speech, and also Schuller and Rigoll’s observation [24] in emotion recognition using both frame-level and supra-segmental features.

2. Often Duration features were the least important.

3. Pitch and Energy features were moderately important; however, there was no evidence that one was more important than the other.

4. In summary, the importance of the four categories was in the order of $\text{MFCC} > \text{Energy} \approx \text{Pitch} > \text{Duration}$. This pattern is consistent with Mower et al.’s findings [23], where the same 46 acoustic features were applied to two datasets in English and information gain was used in feature selection for emotion classification.

5. Generally, estimator fusion can achieve better performance than the elementary estimators. Between the two fusion approaches, LWF always outperformed AF, regardless of how many features were used.

We need to point out that only 46 acoustic features were used in this study, and the four categories had different numbers of features, i.e., MFCC, Energy, Pitch and Duration had 26, 6, 9, and 5 features, respectively. It is interesting to investigate whether or not extracting more and different features in these four categories will change their relative importance and improve the estimation performance. For example, the 137 acoustic features used by Grimm and Kroschel [19], and the 3304 acoustic features used by Batliner et al. [21], can be considered. Additionally, Busso et al. [26] showed that Mel Filter Bank (MFB) features seem more important than MFCC features in emotional speech analysis. It is interesting to augment MFB features in our study. Finally, we can extend this study to databases in different languages and study whether the pattern found in this paper is universal.

5. References


