論文 / 著書情報
Article / Book Information

論題(和文)	
Title(English)	A Regression Approach to Emotion Estimation in Spontaneous Speech
著者(和文)	Wang Qiongqiong, 篠田 浩一
Authors(English)	Qiongqiong Wang, Koichi Shinoda
出典(和文)	日本音響学会2013年秋季研究発表会講演論文集, , , pp. 87-88
Citation(English)	, , , pp. 87-88
発行日 / Pub. date	2013, 9

A Regression Approach to Emotion Estimation in Spontaneous Speech

◎Qiongqiong Wang and Koichi Shinoda (Tokyo Institute of Technology) *

1 Introduction

The task of emotion recognition in speech is to recognize the emotional state of a speaker from his or her speech [1]. Most of the past research in this field has focused on recognition of acted or prototypic speech of categorized emotions. In this paper, we focus on emotion estimation in spontaneous speech.

Few researches have been done on spontaneous speech [2]. Existing corpora for spontaneous speech are limited and unbalanced. The characteristics of natural speech, mostly mild or neutral, makes it hard to classify human's continuous and mixed emotion into discrete categories. Annotations are often erroreous. Moreover, the performance degrades because of variability caused by other factors such as speaker variability, session variability etc.

To tackle these problems, we proposed a regression method to predict the three dimensional representation of emotion, Valence(V) - positive or negative, Activation(A) - the level of excitation, Dominance(D) - the apparent strength or weakness of the speaker [3]. We also proposed a factor analysis based speaker variability removal technique. We use Support Vector Regression (SVR) and k-NN regression as the regression method.

2 Regression Method

2.1 SVR

SVR uses the same principles as the SVM for classification, with only a few minor differences. In SVR, output is a real number. The model produced by SVR depends only on a subset of the training data, because the cost function for building the model ignores any training data close to the model prediction (within a threshold ϵ). We only use linear SVR in this research.

Linear SVR

$$y = \sum_{i=1}^{N} (\alpha_i - \alpha_i^*) \cdot \langle x_i, x \rangle + b$$

***k*-NN regression**

The k-NN algorithm can also be adapted for regression. In this research, we simply used the inverse of the Euclidean distance between test data and training data as the weight for the effect of this training data. For each test sample, we choose k training samples close to it and estimate its emotion degree E' as:

$$E' = \frac{\sum_{i=1}^{k} E_i / s_i}{\sum_{i=1}^{k} 1/s_i},$$

where E_i is the emotion degree of training speech sample i, s_i is the Euclidean distance between the test sample and its i^{th} nearest training sample.

3 Speaker Variability Removal Technique Based on Factor Analysis

Speaker variability has the effect to degrade emotion recognition. Therefore, besides PCA, we propose a method to find speaker variability based on factor analysis and then remove it from all variabilities. GMM supervectors are extracted from GMMs with F mixture components trained from d-dimensional feature vectors. A target GMM supervector M can be written as:

$$M = m + V\alpha + U\beta, \tag{1}$$

where $m \in R^{Fd}$ represents the UBM supervector, $V \in R^{Fd \times N_V}$ is a matrix of N_V 'eigenemotions', $U \in R^{Fd \times N_U}$ is a matrix of N_U 'eigenspeakers', α and β represent emotion and speaker factors.

The first step is to train speaker specific GMMs. A UBM is first built using a large neutral based corpus. Then, given all the training data for the target speaker, we use MAP adaption to adapt the pretrained UBM to obtain speaker-specific GMM models.

The second step is to generate speaker factors β for the target speaker. We compare each speaker's GMM supervector M to the UBM supervector m by ignoring emotion variability.

$$M = (m + V\alpha) + U\beta = (m + V\alpha) + \sum_{j=1}^{N_U} w_j E_j, \tag{2}$$

where $U = [E_1, E_2, ...]$, $\beta = [w_1, w_2, ...]^T$. Thus each utterance in both training data and testing data will get one speaker factor β.

The third step is to remove speaker variability. We assume the first k eigenspeakers have the most

---

* 自発音声における感情推定のための回帰的手法
ワンチョンチョン, 篠田 浩一 (東工大)

speaker variability, so we remove the first k dimensions and get the new vectors which have less speaker variability.

$$\hat{M} = V\alpha = (M - m) - \sum_{j=1}^{k} w_j E_j. \qquad (3)$$

4 Experiment

The spontaneous speech corpus we used in our research is the VAM German Database from a German talk-show. Each utterance was annotated by multiple human labelers in 3 dimensions V, A, and D in the normalized range of $[-1, +1]$. 10-fold cross validation was used according to utterances and speakers respectively. 10,084 utterances from GLOBALPHONE-German, with training data from VAM were used in UBM training.

We use 39 dimensional MFCC features, including static, first and second derivatives of MFCCs and normalized energy parameter. Then, using these features, we train a GMM with 64 components and construct a GMM supervector. To this supervector, we add 18 dimensional statistics of fundamental frequency (F0) and its 1st derivative: mean, median, maximum, minimum, variance, 25% quantile, 75% quantile, difference between max and min and difference between the quartiles. We compared several systems combining different features and estimators.

1. **Features:** (a) utterance GMM supervector in utterance-based validation (b) projection vectors from utterance-based PCA of M (c) projection vectors combined with F0 features (d) normalized utterance GMM supervector in speaker-based validation (e) without the first principal components of speaker variability (f) without the first and the second principal components of speaker variability

2. **Estimaors:** (A) SVR-linear (B)k-NN, where $k = 1 : 30$.

4.1 Results

The results are shown in Fig 4.1. The correlation coefficients are moderate to high ($0.46<r<0.79$). SVR is best when using projection vector with F0 features for $V(0.44)$, $A(0.79)$, $D(0.76)$. k-NN gets the best correlation for $V(0.46, k = 14)$, $A(0.70, k = 25)$, $D(0.66, k = 20)$; Both k-NN and SVR system outperformed the previous work [4] in V, and got similar correlation in A and D.

When using SVR estimator, correlations of all primitives get improved after PCA is applied (0.4 on average), which indicates PCA helps removing some non-emotion information. After F0 features are added, every primitive gets the best performance (0.44, 0.79, 0.76), 0.03 more on average. When speaker factor analysis is applied, removing principal components of speaker variability decreases correlation of all primitives slightly, which is opposite to our expectation. The reason should be that the principal components we removed still include some emotional information.

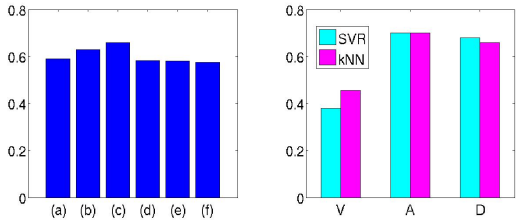

Fig. 1 Left: Average correlations with SVR using (a)∼(f); Right: Correlation of k-NN and SVR systems using GMM supervector.

k-NN gets better correlation in V, the same in A, and better in D than SVR.

5 Conclusions

The system we proposed with PCA did help removing non-emotion variabilities in emotion regression in speech, which outperformed the previous work on VAM database. However, the technique we have demonstrated for removing speaker factor degrades the performance slightly. In the future, we would like to apply joint factor analysis technique and neural network for regression.

References

[1] T. Polzin, A. Waibel, F. Dellaert, In *Proc. ICSLP*, Vol. 3. pp. 1970-1973, 1996.

[2] D. Sauter, F. Eisner, A. Calder, S. Scott, *The Quarterly journal of Experimental Psychology*, pp. 2251-2272, 2010.

[3] R. Kehrein, In *Proceeding of Speech Prosody*, pp. 423-426, 2002.

[4] M. Grimm, K. Troschel, E. Mower, S. Narayanan, *Speech Communication*, vol. 49, pp. 787-800, 2007.