Emotional Speech Classification Based On Multi View Characterization

Ammar Mahdhaoui, Mohamed Chetouani
Université Pierre et Marie Curie
Institut des Systèmes Intelligents et de Robotique, CNRS UMR 7222
4 Place Jussieu 75252 Paris Cedex 05, France
Ammar.Mahdhaoui@isir.upmc.fr, Mohamed.Chetouani@upmc.fr

Abstract

Emotional speech classification is a key problem in social interaction analysis. Traditional emotional speech classification methods are completely supervised and require large amounts of labeled data. In addition, various feature sets are usually used to characterize the emotional speech signals. Therefore, we propose a new co-training algorithm based on multi-view features. More specifically, we adopt different features for the characterization of speech signals to form different views for classification, so as to extract as much discriminative information as possible. We then use the co-training algorithm to classify emotional speech with only few annotations. In this article, a dynamic weighted co-training algorithm is developed to combine different features (views) to predict the common class variable. Experiments prove the validity and effectiveness of this method compared to self-training algorithm.

1. Introduction

Non-verbal communication plays a major role during social interaction and carry information between the different speakers. In [9], different non-verbal behavioral cues have been defined: physical appearance, gestures and postures, face and eyes behaviors, vocal behavior, space and environment behaviors. The combination of different codes make it possible to convey various information such as emotion, intention but are also useful for managing interaction, and/or sending relational messages (dominance, persuasion, embarrassment, etc.). In this article, we focus on the analysis of a class of non-verbal behaviors which accompanies the verbal message termed as vocal behaviors in [9], more specifically emotional speech. Accordingly, it has become more important to automatize the annotation of face-to-face interaction databases to enable social interaction study. There have been many studies on emotional speech classification, both on feature extraction and on classification. However, there are still some limitations with the existing methods. One of these limitations is that most of these approaches are completely supervised, i.e., large amounts of labeled data are needed. While in a real-life emotional speech recognition, the manual annotation of data is very costly and time consuming. Therefore, semi-supervised techniques provide elements for combining both labeled and unlabeled data. In addition, if more than one feature sets is used, we simply concatenate the different features to form a long feature vector and we applied feature selection/reduction techniques. Therefore, it is desirable to study how to effectively combine these different feature sets so that more discriminative information can be extracted from the speech segments. We propose a semi-supervised algorithm based on multi-view characterisation which combines the classification results of different views to obtain a single estimate for each observation. The proposed algorithm is a novel form of co-training, which is more suitable for problems involving both classification and data fusion. Algorithmically, the proposed co-training algorithm is quite similar to other co-training methods available in the literature. However, a number of novel improvements, using different feature sets and dynamic weighting classifiers fusion, have been incorporated that makes the proposed algorithm more suitable for multi-view classification problems.

The paper is organized as follows. Section 2 presents the different feature extraction methods. Section 3 presents the details of the proposed method for semi-supervised classification of emotional speech with multi-view features. In Section 4, experimental results of the proposed method are given. In the last section, some concluding remarks and the direction of future works are presented.

1051-4651/10 $26.00 © 2010 IEEE
DOI 10.1109/ICPR.2010.1090

2. Supervised classification

2.1 Database

The emotional speech database used in this experiments contains 300 utterances which cover 2 different speech registers, infant-directed speech [2] and adult-directed speech, all sequences were extracted from the Pisa home movies video database [4]. There are 150 utterances in each class, the utterances are typically between 0.5s and 4s in length. The verbal interactions of the mothers have been carefully annotated by two psycholinguists on two categories (Cohen's kappa=0.82): infant-directed speech and adult-directed speech. We divided the database into two parts, unlabeled data U which contains 200 examples and labeled data L which contains 100 examples.

2.2 Emotional Speech Characterization

Feature extraction aims at computing an efficient representation of the speech signals. Temporal and frequential features are usually investigated in emotion recognition [3] [6]. In this study, 70 prosodic, 16 cepstral and 96 spectral features were extracted, which were shown to be the most efficients [3] [5]. In addition, we computed frame-level and utterance-level features that can be used in different modeling techniques to develop emotional speech detection system.

2.2.1 Frame-level features extraction

Frame-level features are extracted each 20 ms, so the number of the resulting feature vectors is variable and depends on the length of the utterance. Cepstral features such as MFCC are often successfully used in speech recognition. Short-term cepstral signatures of emotional speech are characterized by MFCC features.
Several studies have shown that fundamental frequency (F0) and energy features are very important to emotion recognition applications [3]. F0 and energy were estimated each 20 ms, and we computed, for each voiced segment, 3 statistics: mean, variance and range for both F0 and short-time energy resulting in a 6 dimensional vector.

2.2.2 Utterance-level features extraction

Utterance-level features refer to features extracted per whole utterance, so we have one feature vector by utterance. In addition to pitch estimations per frame, we also measured some more global higher-level pitch features to capture the fluctuations and variability of fundamental frequency. We computed 32 statistical features in order to model the dynamic variations of the bark spectral perceptive representation. Three other features are also extracted from the pitch contour and the loudness contour, by using histograms and considering the maximum, the bin index of the maximum and the center value of the corresponding bin. These 3 features are relevant for pitch and energy contours modelling.
To extract spectral bark features, for a given spectral model we performed the analysis on successive time frames along each utterance. We then extracted a set of statistical features from these representations. These features can be applied either along the time axis or along the frequency axis . We also consider the average of energy of the bands (a perceptive Long Term Average Spectrum) and extract statistical features from it. As a result, 32 statistical features are used and applied a) along time axis (Bark TL), b) along frequency axes (Bark SL), c) on the average perceptive spectrum (Bark MV) to obtain a first set of 32 features.
Finally, we have 9 kind of feature vectors with different dimensions and are presented in Table 1.

Table 1. Feature extraction

	Features	Dimension
1	MFCC	16 MFCCs
2	Pitch+Energy-6	6 statistics on pitch and energy
3	Pitch-35	35 statistics on pitch
4	Energy-35	35 statistics on energy
5	Pitch+energy-70	70 statistics on pitch and energy
6	Bark TL+SL+MV	96 statistics
7	Bark TL	32 statistics
8	Bark SL	32 statistics
9	Bark MV	32 statistics

2.3 Classification

After the feature extraction stage, the classification task can be achieved using standard machine learning methods. In this study, four different classifiers: gaussian mixture model (GMM), neural network (MLP), support vector machines (SVM) and k-nearest neighbor (k-NN) classifiers, were investigated. Each classifier considers the selected features as the most efficient for the two-classes discrimination problem. The best combination between the classifiers and the different features sets (views) and their performances, using 100 examples for training and 200 for test, are presented in Table 2.

Table 2. Different views and their performance on supervised learning

Classifiers	Combination	Accuracy (%)
h1	GMM trained with MFCC	73.5
h2	GMM trained with Pitch+Energy-6	59.5
h3	k-NN trained with Pitch-35	55
h4	k-NN trained with Energy-35	68.5
h5	SVM trained with Pitch+energy-70	65.5
h6	GMM trained with Bark TL+SL+MV	61
h7	MLP trained with Bark TL	58.5
h8	GMM trained with Bark SL	65
h9	MLP trained with Bark MV	64

3. Semi-Supervised classification

3.1 Co-training

Co-training algorithm [1] and related multi-view learning methods [8] assume that various classifiers are trained over multiple feature views of the same labeled examples. These classifiers are encouraged to make the same prediction on any unlabeled example. The co-training algorithm proposed in [1] is a prominent achievement in semi-supervised learning. It initially defines two classifiers on distinct attribute views of few labeled data. Either of the views is required to be conditionally independent to the other and sufficient for learning a classification system. Then iteratively, each classifier's predictions on unlabeled examples are selected to increase the training data set. This co-training algorithm and its variations [10] have been applied in many application areas because of their theoretical justifications and experimental success. The originality of our work is to use multi features views and dynamic weighting classifiers fusion to minimize the classification error.

3.2 Co-Training Algorithm Based On Multi View Characterization

In this work, we propose a co-training procedure that iteratively trains a base classifier within each view and then combines the classification results to obtain a single estimate for each observation. The proposed algorithm is a novel form of co-training, which is more suitable for problems involving both classification and data fusion. This algorithm is designed to improve the performance of a learning machine with both few labelled utterances and large amounts of cheap unlabelled utterances. Let L be the labelled utterances, U the unlabelled utterances and V_i be different feature views. The algorithm works as described in Table 3. First, to initialize the algorithm we found the best feature set for each classifier as presented in Table 2. Secondly, we set all of the initial weights equally so that $\omega_k = 1/v$; v is the number of view (9 in our case). Thirdly, while the unlabeled database U not empty we loop on:

- Classification: classify all the unlabeled utterances, the class of each utterance is obtained using a decision function, in our case we compute the maximum likelihood, otherwise we can use other decision functions.

- Update the labeled and unlabeled databases: first we take U_1 the utterances from U classified on Class 1 and U_2 classified on Class 2, after that we calculate the classification confidence for each utterance that we called $margin$.

$$margin(x, y) = \frac{y \sum_{k=1}^{v} \omega_k \times h_k(x)}{\sum_{k=1}^{v} \omega_k} \quad (1)$$

  Where x is the feature view to be classified on the class y, ω_k is the weight of the classifier h_k and v is the number of views. The $margin$ value is included in the interval [-1,1] and is positive only if x is correctly classified. This number can be interpreted as a measure of confidence as it is done for SVM [7]. Then we take T_j from U_j the utterances which has classified on $Class_j$ with a probability upper to the mean value of classification confidence ($margin$) of the $Class_j$.

- Update weights: finally, we update the weights of each view as described in Table 3. The new weight of each classifier is proportional to its contribution in the final classification. In other words, weights of efficient classifiers will be increased.

4. Experimentations

The classification accuracy of the co-training algorithm using multi-view feature sets with different number of annotations is presented in figure 1. It can be seen that our method can achieve a good result in infant-directed speech discrimination.
To further illustrate the advantage of the proposed method, especially in case of very few number of annotations, we compare our method with the self-training method with a single-view. We test the basic self-training algorithm, which replaces multiple classifiers in the co-training procedure with the best classifier that employs the most efficient feature, in our case GMM with MFCC (h_1). Figure 1 shows a comparison between our co-training method and self-training method.

Table 3. The Co-Training algorithm

Given:
a set L of m Labeled examples $(x_1, y_1), ..., (x_m, y_m)$ with labels $y_i = \{1, 2\}$
a set U of n Unlabeled examples $u_1, ..., u_n$
v = number of view (classifier)
Initialization:
ω_k (weights of classifier)= $1/v$ for all the view
While U not empty
A. Classify all the example of the test database:
Do for $i = 1, 2, ..., size(U)$
1. Use L to train each classifier h_k
2. Classify all examples of U by each h_k
3. Calculate the probability of classification for each example from U, $p(i)(j) = \sum_{k=1}^{v} \omega_k \times h_k(x_i, j)$
4. $Labels(x_i) = decision(p(x_i, j))$
End for
B. Update the training (L) and test (U) databases:
U_1 and U_2 the ensemble of example classified successively on $Class_1$ and $Class_2$
Do for $i = 1, 2, ..., size(U_j)$
$margin(x_i, j) = \frac{\sum_{k=1}^{v} \omega_k \times h_k(x_i, j)}{\sum_{k=1}^{v} \omega_k}$
End for
$\alpha_j = mean(margin(x_i, j))$
Take T_j from U_j the examples which has classified on C_j with a probability upper to α_j.
$T = T_1 + T_2$
Add T to L and remove it from U
C. Update weights: $\omega_k = \frac{\sum_{i=1}^{size(T)} h_k(x_i)}{\sum_{k=1}^{v} \sum_{i=1}^{size(T)} h_k(x_i)}$
End While

It can be seen that our method outperforms the self-training method, 75.5% vs 63% with 100 labeled utterances. In addition, the proposed co-training method gives a satisfactory result in the case of very few annotations, 61.5% with 10 labeled utterances.

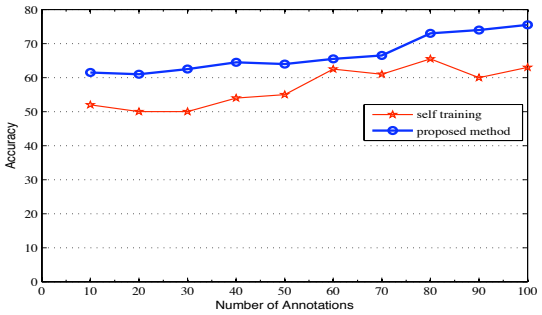

Figure 1. Classification accuracy with different number of annotations.

5. Conclusion

In this article, a co-training algorithm is presented. The main contribution of this work is a framework allowing the combination of multi-features and the penalization dynamically each classifier or views by calculating iteratively the classification confidence. Experimental results demonstrate the efficiency of this method. However, several issues have to be investigated in the future. The first one is to test this method on larger emotional speech databases. The second is to investigate the complementarities of the different views by analyzing the weights evolution of each classifier and the third one is to test other semi-supervised algorithms, especially algorithms using multi-view features.

References

[1] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In COLT, 1998.

[2] A. Fernald, P. Kuhl, Acoustic determinants of infant preference for Motherese speech. Infant Behavior and Development, 10, 279-293, 1987.

[3] K.P. Truong, D.A. van Leeuwen, Automatic discrimination between laughter and speech. Speech Communication 49 (2007) 144158

[4] Maestro S, et al. How young children treat objects and people: an empirical study of the first year of life in autism. Child psychiatry and Human Development 35 (4).

[5] Mahdhaoui A, Chetouani M, Zong C. Motherese detection based on segmental and supra-segmental features. In: IAPR international conference on pattern recognition, ICPR 2008; 2008.

[6] M. Shami, W. Verhelst, An evaluation of the robustness of existing supervised machine learning approaches to the classification of emotions in speech. Speech Communication 49 (2007) 201212

[7] Robert E. Schapire, and Y. Singer, Improved Boosting Algorithms Using Confidence-rated Predictions, Machine Learning. Springer Netherlands, 0885-6125 (Print) 1573-0565 (Online), Volume 37, Number 3 / dcembre 1999, 297-336

[8] U. Brefeld, T. Gaertner, T. Scheffer, and S. Wrobel. Efficient co-regularized least squares regression. In ICML06, Pittsburgh, USA, 2006.

[9] Vinciarelli A, Pantic M, Bourlard H, Pentland A. Social signals, their function, and automatic analysis: a survey. In ICMI08. 2008. p. 618.

[10] S. Goldman and Y. Zhou, Enhancing supervised learning with unlabeled data, Proc. of International Conference on Machine Learning, pp. 327334, 2000.