University of Sheffield Engineering Symposium (USES)
ISSN (Online) XXXX-XXXX
Volume I, Issue I (XXXXX 2014), PP.

Emotion Recognition from the Speech Signal by Effective Combination of Generative and Discriminative Models

Erfan Loweimi, Mortaza Doulaty, Jon Barker and Thomas Hain

Speech and Hearing Research Group (SPandH), Department of Computer Science, Engineering Faculty, University of Sheffield

Abstract

In this paper, we propose an effective way for combining the discriminative and generative models for emotion recognition from speech signal. Finding an efficient feature extraction algorithm which captures just the main attribute(s) pertinent to the task and filters out the other aspects of the data turns out to be very challenging, if not impossible. We propose an interface between the front-end and the back-end in order to compensate for the shortcoming of the parameterization block in suppressing the irrelevant dimensions of the signal. This interface is a generative model, which performs remarkable dimensionality reduction, allows for extraction of a long-term feature, and also paves the way for better classification of the data through a discriminative model. This method leads to a 7.6% absolute performance improvement in comparison with the baseline system and results in 87.6% accuracy in emotion recognition task. Human performance on the same database is reportedly 84.3%.

Keywords Discriminative model; Emotion recognition; Front-end; Generative model; Speech signal

1. Introduction

Speech is the most natural method of human communication. It reflects many aspects of us and this turns it into a complicated signal, which together with its lingual content, encodes environmental and speaker-dependent information like identity, emotional state, accent, dialect and age. These components of speech are combined through a complicated process and disentangling this signal into the aforementioned underlying dimensions is a challenging task from both signal processing and machine learning points of view.

When the aim is to design a system for capturing one of these components, all the other elements become noise. If these are not suppressed effectively, the system would only perform well under certain conditions. For example, if the goal of a system is to recognize the emotion, it should do the job regardless of the gender, speaker ID, lingual content and the background sounds. If such attributes are not filtered out, the system would be biased toward a particular situation which matches the training data. This is not desired in practice, as a reliable system is expected to generalize well to scenarios that are relatively different from the training circumstances.

Typically, pattern recognition systems consist of two main blocks, namely the front-end and back-end [1]. The front-end is tasked with extracting a representation of the data in which the task-pertinent attributes are preserved/enhanced and the irrelevant/misleading aspects of the data are filtered/weakened. This requires a particular data filtering in a very high-level domain where each attribute occupies a particular subspace. So, front-end ideally should do *information filtering* in the *information space* and it turns out to be highly difficult. The reason backs to the fact that such information space is categorically abstract and subjective. Therefore, a mathematical underpinning of a mapping that takes the data from low-level quantitative world to that high-level qualitative domain is extremely complicated (if not impossible).

In this work, we aim at enhancing the feature extraction process with an interface that contributes toward conducting information filtering. This interface is a generative model that tries to learn a rough task-dependent representation of the data. Such coupling of the generative and discriminative models results in further dimensionality reduction as well as up to a 7.6% performance elevation in comparison with the baseline GMM-based classification system. It leads to 87.6% accuracy which is higher than the reported human performance (84.3%) on the same task and database [2].

2. METHOD

Figure 1 shows the main parts of the proposed method. First, each speech waveform is converted into a feature matrix. MFCC is utilized in this phase. Then, the class data is pooled for training a GMM for each class. The utilized database includes 7 (acted) emotional states (classes), namely Anger, Boredom, Disgust, Fear, Happiness, Sadness and Neutral. Number of components (*M*) was set to 25 and the GMMs were trained through 5 iterations of the Expectation-Maximization algorithm [1]. After training the GMMs, the posterior of all the components of the models for each speech's frame is computed and averaged over all frames. Each GMM returns *M* elements and after concatenating the outputs of all the GMMs a fixed-length super feature vector containing 7 times *M* elements is built for each signal.

Such super feature vector has three main advantages. First, it is no longer a general feature like MFCC but is statistically optimized for the task at hand. Also, this approach provides an effective framework for capturing the long-term properties of the speech like the emotion. From a statistical standpoint, unlike the lingual content, which changes on a short-term basis, the speaker-dependent attributes are fairly stationary during the utterance. As a result, it is more sensible to steer the front-end toward extracting features that reflect the long-term properties of the speech in tasks like emotion recognition and the proposed approach allows for that. Thirdly, it allows for further dimensionality reduction and fixed-length representation. In fact, instead of representing the speech via a matrix with *feature_vector_length* (typically 39) times *#frames* (#: number of) elements, the signal is represented with *#components* of the GMM times *#classes* which is more compact.

For classification at the back-end, we have employed the SVM with RBF (radial basis function) kernel, which is a powerful discriminative model. It should be noted that the SVM could not handle variable length patterns efficiently [3] and removing the interface block degrades its performance.

3. Experimental results

The GMMs and SVM were trained using Scikit-learn package [4]. For evaluating the system's throughput, 5-fold cross validation [1] was used and the performance was investigated based on Accuracy (Acc), recall rate (Rec), precision (Pre) and F-measure (F-m) [1]. As seen in Table 1, the proposed method clearly outperforms the baseline system that is a GMM-based classifier.

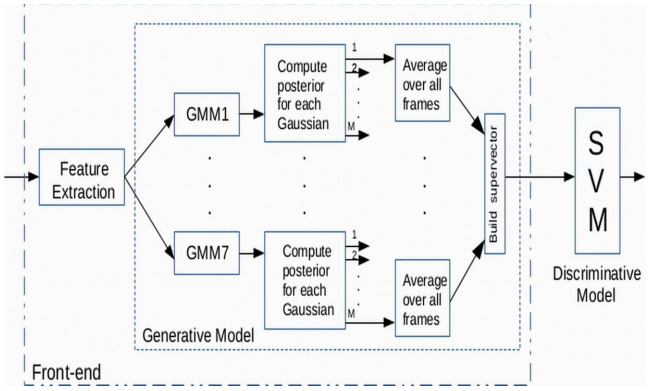

Figure 1. Workflow of the proposed method.

Table 1. Evaluation of the performance of the proposed emotion recognition system.

	Train				Test			
	Acc	Rec	Pre	F-m	Acc	Rec	Pre	F-m
GMM	99.9	99.9	99.9	99.9	80.0	78.5	82.0	79.4
Prop-osed	100	100	100	100	87.6	87.0	88.2	87.4

4. CONCLUSIONS

In this paper, we proposed an interface block between the front-end and back-end based on the GMM generative model. This block allows us to filter out unwanted attributes, extract long-term features from the speech, and achieve further dimensionality reduction. This paves way for efficient classification through discriminative model (SVM) by providing a fixed-length representation for this signal. It resulted in a 7.6% absolute performance improvement.

REFERENCES

[1] Murphy KP. Machine learning: a probabilistic perspective. Cambridge, MA; 2012.

[2] Burkhardt F, Paeschke A, Rolfes M, Sendlmeier W, Weiss B. A database of german emotional speech. In: in Proceedings of Interspeech, Lissabon; 2005. p. 1517–1520.

[3] Saz O, Doulaty M, Hain T. Background-Tracking Acoustic Features for Genre Identification of Broadcast Shows. In: Proceedings of Spoken Language Technology (SLT) Workshop. South Lake Tahoe NV, USA; 2014. p. 118–123.

[4] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011;12:2825–2830.