Speaker-Independent Emotion Recognition based on Feature Vector Classification

Jeong-Sik Park¹, Ji-Hwan Kim², Sang-Min Yoon¹, Yung-Hwan Oh¹

¹Computer Science Division, Korea Advanced Institute of Science and Technology, Daejeon, Korea
²Department of Computer Science and Engineering, Sogang University, Seoul, Korea
¹{dionpark, sangmin, yhoh}@speech.kaist.ac.kr, ²kimjihwan@sogang.ac.kr

Abstract

This paper proposes a new feature vector classification for speech emotion recognition. The conventional feature vector classification applied to speaker identification categorized feature vectors as overlapped and non-overlapped. This method discards all of the overlapped vectors in model training, while non-overlapped vectors are used to reconstruct corresponding speaker models. Although the conventional classification showed strong performance in speaker identification, it has limitations in constructing robust models when the number of overlapped vectors is significantly increased such as in emotion recognition. To overcome such a drawback, we propose a more sophisticated classification method which selects discriminative vectors among overlapped vectors and adds the vectors in model reconstruction. Our experimental results based on an LDC emotion corpus, our classification approach exhibited superior performance when compared to the conventional method.

Index Terms: feature vector classification, emotion recognition

1. Introduction

As speech recognition technology has advanced as an essential field of artificial intelligence, research on challenging issues such as emotion recognition or language identification has also been brought to the fore. Especially, emotion recognition is primarily being investigated for human-machine interaction and has broad range of applications including call center service, medical monitoring, etc. While various technical approaches for emotion recognition have been introduced, it is still difficult to guarantee satisfactory performance owing to certain factors. The performance of emotion recognition systems depends intensively on speaker or language characteristics [1]. Moreover, an increment in the number of emotions deteriorates recognition accuracy because acoustically similar characteristics between emotions make it hard to classify emotions due to the similarity between the models. For these reasons, it is a critical challenge to construct robust emotion models that minimize the acoustical similarity between emotions as well as the performance dependency according to speaker characteristics.

To overcome such issues, we apply a feature vector classification scheme to emotion recognition. A feature vector classification normally categorizes vectors in two groups: acoustically discriminative vectors between models and non-discriminative vectors. Then, models are reconstructed by using only discriminative vectors, based on the assumption that non-discriminative vectors are generated by non-speech segments such as silence and environmental noise, or speech segments whose corresponding models have acoustically similar characteristics with other models.

The conventional feature vector classification was applied to speaker identification, where feature vectors are categorized as non-overlapped vectors and overlapped vectors [2]. If the log-likelihood of a vector on its corresponding speaker model does not indicate the maximum value, compared to the log-likelihood on the other models, this vector is regarded as an overlapped vector and discarded in the reconstruction of the corresponding speaker model. The conventional method showed strong experimental results in speaker identification.

Emotion decisions for speech data are ambiguous in many cases (even when a person annotates the type of emotion). This domain-oriented ambiguity increases overlapped region between models substantially. For this reason, as the number of emotions increases, an amount of vectors will be discarded in the conventional method. This causes the sparseness of training vectors and the reconstructed emotion models are not trained sufficiently. Therefore, it is necessary to devise a more sophisticated feature vector classification method.

This paper is organized as follows. In Section 2, we will present the standard GMM-based emotion recognition. Then the conventional feature vector classification and our proposed classification method will be described in Section 3. In Section 4, experimental setups and results will be presented and discussed. Finally, we conclude this paper in Section 5.

2. GMM-based emotion recognition

The GMM-based classifier using short-term features has been suggested as being better suited for the task of classifying emotions than other static discriminative classifiers such as support vector machines and decision trees [3][4].

GMM-based emotion recognition is composed of training and test phase. The purpose of training is to construct a reliable GMM for each emotion. Once feature vectors are extracted from training data, maximum likelihood model parameters are estimated using the iterative expectation-maximization (EM) algorithm. During test phase, GMMs are used to identify test utterances. The log-likelihood on a GMM model \( \lambda_e \) \( (i=1,...,E) \), where \( E \) are the number of emotions) for a sequence of feature vectors \( X = \{ \tilde{x}_1, ..., \tilde{x}_T \} \), which are extracted from a test utterance, is computed as follows:

\[
\log p(X|\lambda_e) = \sum_{t=1}^{T} \log p(\tilde{x}_t|\lambda_e)
\]

Then, a model which has the maximum log-likelihood with a given test utterance is decided as a recognition result.

3. Feature vector classification for emotion recognition

In the standard GMM approach, the interrelation between models critically affects the final decision. Especially, several factors such as acoustically similar characteristics between emotions or environmental noises may generate overlaps of
emotion models and thus induce decision errors. Usually, the more emotions are included in recognition system, the more amounts of overlaps exist. Hence, it is considerably important to mitigate the overlap effects.

3.1. Conventional feature vector classification for speaker identification

A feature vector classification method was proposed to reduce the overlap effects in speaker identification [2]. The method classifies training vectors into two categories: overlapped and non-overlapped vectors. Once the standard GMMs are built, each training vector is verified whether it indicates maximum log-likelihood on the corresponding model. If a vector indicates its maximum log-likelihood on one of other models, it is regarded as an overlapped vector. Otherwise, it is regarded as a non-overlapped vector. This approach assumes that overlapped vectors are generated by non-speech segments such as silence and environmental noise, or speech segments whose corresponding speaker models have acoustically similar characteristics with other speaker models. After classification, speaker models are reconstructed using only non-overlapped vectors. Figure 1 shows the procedure in speaker model training based on the conventional feature vector classification. The conventional method showed good experimental results in speaker identification.

3.2. Problems of conventional classification in emotion recognition

In emotion recognition, emotion decisions for speech data are ambiguous in many cases. This domain-oriented ambiguity increases overlapped region between models substantially. We carried out an experiment to measure the composition ratio of overlapped vectors, in order to verify whether the conventional feature vector classification is applicable to emotion recognition. Using emotional speech utterances, we constructed ten emotion GMMs, each of which has one state and one mixture based on the type of corresponding emotion data. And next, we investigated the number of overlapped vectors for five different emotions in two cases: ‘five models’ and ‘ten models’. In the former, five models represent five different emotions (boredom, anger, happy, neutral, and sadness) and in the latter, another five emotion models (anxiety, cold anger, panic, shame, and pride) are included.

Figure 2 represents the composition ratio of overlapped vectors when the conventional classification method is applied to emotion recognition. This result explains that an increment in the number of emotion models increases the number of overlapped vectors considerably. For example, more than 60% of feature vectors from ‘boredom’ and ‘neutral’ emotion were regarded as overlapped vectors. In other words, only less than 40% of feature vectors are used in model training. This causes the sparseness of training vectors and deteriorates the recognition accuracy. The higher the number of emotions included, the more vectors are discarded in GMM reconstruction. As a result, reconstructed emotion models are not trained sufficiently. Therefore, it is necessary to devise a more sophisticated feature vector classification method.

3.3. Advanced feature vector classification for emotion recognition

Several previous studies on emotion recognition show that an emotion tends to be recognized as another specific emotion, when it is misrecognized [3][5]. Such a tendency can be considered in relation to overlapped vectors. It is already mentioned that a large amount of vectors are regarded as overlapped vector in emotion recognition, due to the domain-oriented ambiguity. We believe that a substantial amount of overlapped vectors are caused by acoustically similar characteristics between two models, which have such tendencies in the corresponding pairs of misrecognition. Thus, we assume that such overlapped vectors retain their discriminative information.

In the conventional method, only non-overlapped vectors are treated as discriminative vectors and all of the overlapped vectors are discarded. In our method, addressed as ‘advanced feature vector classification’, we carefully expand the region of discriminative vectors to include the overlapped vectors which preserve discriminative information according to our assumption.

Figure 3 illustrates the procedures for our advanced feature vector classification. Firstly, we build each emotion model based on the standard GMM approach. Next, every feature vector is verified whether it is correctly recognized as a corresponding emotion, based on the log-likelihood estimated for each emotion model. If the vector is closer to its corresponding model than other models, we designate this vector as a discriminative vector as well as a non-overlapped vector. Steps proceed to classify the overlapped vectors. As mentioned above, some of the overlapped vectors still preserve discriminative information, if their corresponding models have acoustically similar characteristics to a specific emotion model. We designate such vectors as discriminative vectors of another type, while other overlapped vectors are regarded as non-discriminative vectors. Finally, vectors classified as discriminative vectors are applied to emotion model reconstruction; otherwise, non-discriminative vectors are used to construct overlap models.

Above mechanism classifying feature vectors into three categories can be described using numerical expression as follows. Note that there are E emotion standard GMMs, $\lambda_{e}$ (where $e=1, ..., E$), and $x_{r,t}$ (where $r=1, ..., T_{e}$) is one of $T_{e}$ feature vectors used to construct a standard GMM $\lambda_{e}$.

After constructing the standard GMMs, we obtain N-best log-likelihood results of each $x_{r,t}$ from E emotion models. Let us denote $R_{e}(x_{r,t})$ as the emotion model index (ranging from
for each emotion. One is a reconstructed emotion model built after feature vectors are classified, we construct two models.

3.4. Emotion recognition based on advanced feature vector classification

After feature vectors are classified, we construct two models for each emotion. One is a reconstructed emotion model built from discriminative vectors, and the other is an overlap model built from non-discriminative vectors. We use overlap models as garbage models to exclude non-discriminative vectors included in a test utterance.

In the test phase, every feature vector extracted from test utterances is classified into two categories based on reconstructed emotion models and overlap models. If a vector indicates maximum log-likelihood on one of the overlap models, the vector is regarded as a non-discriminative vector and discarded in recognition process. Otherwise, the vector is a discriminative vector which preserves corresponding emotion characteristics. Next, we estimate maximum log-likelihood using a set of discriminative feature vectors and decide an emotion \( \hat{e} \), as follows.

\[
\hat{e} = \arg\max_{e=1,...,E} \log p(D|\vec{x}_e), \quad e=1,...,E
\]

where \( \log p(D|\vec{x}_e) = \frac{1}{F} \sum_{f=1}^{F} \log p(d_f|\vec{x}_e) \) and \( \vec{x}_e \) denotes the reconstructed GMM for \( e \)-th emotion. \( D \) is a set of feature vectors regarded as the discriminative vector for a given test utterance and \( F \) is the number of vectors in \( D \).

4. Experiments

We performed emotion recognition experiments on ten kinds of emotional speech data obtained from Emotional Prosody Speech of LDC [7]. The corpus consists of speech recorded by professional actors trying to express emotions while reading short phrases of dates and numbers. Ten different types of emotions are anxiety, boredom, cold anger, despair, happy, neutral, panic, sadness, shame, and pride. Speech data were recorded by seven speakers in clean environments. We used speech data from four speakers in model training and three speakers in test. We used the pitch, log energy, zero crossing rate and 12 dimensional MFCCs, as a feature vector. All vectors were computed within frames of 40ms with a Hamming window shifted by 10ms. By experiments, we confirmed 40ms is a minimum duration to estimate reliable emotion characteristics, especially including pitch information. This duration was also used in [8].

4.1. Experimental results

To verify the efficiency of our classification method, we investigated recognition accuracy changed by the number of emotions. We organized a set of three emotions consisting of anger, neutral, and sadness. Boredom and happy were added for a set of five emotions. Figure 4 shows the error rates according to the number of emotions. The higher the number of emotions is, the greater the error rates. The Advanced Feature Vector Classification (AFVC) demonstrated superior performance compared to the Conventional Feature Vector Classification (CFVC) over all experimental sets. On a set of five emotions, our method presented 15.7% and 11.5% of

![Figure 3: Advanced feature vector classification](image)

![Figure 4: Error rates (%) according to the number of emotions](image)
relative improvement over the standard GMM approach and CFVC, respectively. It is interesting to observe that on a set of ten emotions, the reconstructed GMM with CFVC gave even lower accuracy than the standard GMM. Such a result was caused by undesirable feature vector classification which discarded so many training vectors that reconstructed emotion models were not trained sufficiently. On the other hand, our method provides a more efficient classification in the selection of discriminative vectors and its use in model reconstruction.

We investigated the composition ratio of non-discriminative feature vectors classified based on CFVC and AFVC. We carried out the experiment over ten emotion models with the same procedures described in Section 3.2. As presented in Figure 5, on each emotion, almost a half of vectors regarded as overlapped vectors in CFVC were decided as discriminative vectors in AFVC. Therefore, our approach maintains a sufficient number of vectors in emotion model reconstruction.

Table 1 and Table 2 show the confusion matrix of 5-class emotion recognition for CFVC and AFVC, respectively. In this recognition experiment, the average recognition accuracy was reported as 56.49% and 60.99% on the reconstructed GMM with CFVC and AFVC, respectively. The performance was improved for all types of emotions, and especially for ‘happy’ and ‘anger’ (in terms of relative improvement).

Based on these experimental results, we conclude that our proposed method solves two problems of the conventional method in emotion recognition: the sparseness of training vectors and insufficient model training. The proposed method can improve the recognition accuracy when it is applied to a domain which has substantial domain-oriented ambiguity, such as emotion recognition.

Table 1. Confusion matrix of 5-class emotion recognition based on CFVC

<table>
<thead>
<tr>
<th></th>
<th>boredom</th>
<th>anger</th>
<th>happy</th>
<th>neutral</th>
<th>sadness</th>
</tr>
</thead>
<tbody>
<tr>
<td>boredom</td>
<td>0.42</td>
<td>0.1</td>
<td>0.11</td>
<td>0.21</td>
<td>0.15</td>
</tr>
<tr>
<td>anger</td>
<td>0.03</td>
<td>0.78</td>
<td>0.12</td>
<td>0.06</td>
<td>0.01</td>
</tr>
<tr>
<td>happy</td>
<td>0.03</td>
<td>0.18</td>
<td>0.69</td>
<td>0.04</td>
<td>0.06</td>
</tr>
<tr>
<td>neutral</td>
<td>0.29</td>
<td>0.13</td>
<td>0.09</td>
<td>0.36</td>
<td>0.12</td>
</tr>
<tr>
<td>sadness</td>
<td>0.18</td>
<td>0.04</td>
<td>0.05</td>
<td>0.16</td>
<td>0.57</td>
</tr>
</tbody>
</table>

Table 2. Confusion matrix of 5-class emotion recognition based on AFVC

<table>
<thead>
<tr>
<th></th>
<th>boredom</th>
<th>anger</th>
<th>happy</th>
<th>neutral</th>
<th>sadness</th>
</tr>
</thead>
<tbody>
<tr>
<td>boredom</td>
<td>0.46</td>
<td>0.09</td>
<td>0.11</td>
<td>0.19</td>
<td>0.15</td>
</tr>
<tr>
<td>anger</td>
<td>0.02</td>
<td>0.83</td>
<td>0.10</td>
<td>0.04</td>
<td>0.01</td>
</tr>
<tr>
<td>happy</td>
<td>0.01</td>
<td>0.19</td>
<td>0.75</td>
<td>0.02</td>
<td>0.03</td>
</tr>
<tr>
<td>neutral</td>
<td>0.25</td>
<td>0.13</td>
<td>0.10</td>
<td>0.42</td>
<td>0.10</td>
</tr>
<tr>
<td>sadness</td>
<td>0.19</td>
<td>0.03</td>
<td>0.05</td>
<td>0.15</td>
<td>0.58</td>
</tr>
</tbody>
</table>

The experimental results also support that our classification approach yields strong performance in speaker independent emotion recognition system. In the speaker independent system, minimizing individual speaking styles is required in model training, since training data spoken by a speaker with extraordinary emotion characteristics may not guarantee reliable emotion models. Our classification method effectively discards such unexpected data, which must be decided as non-discriminative feature vectors.

5. Conclusions

This paper proposed an advanced feature vector classification for speaker independent emotion recognition. The conventional feature vector classification, which was applied to speaker identification, uses only the non-overlapped vectors in model training, while discarding all of the overlapped vectors. For this reason, it has limitations in constructing robust models when the number of overlapped vectors is significantly increased such as in emotion recognition. To overcome such a drawback, we proposed a more sophisticated classification method. We assume that an amount of overlapped vectors still preserve discriminative information. We select such discriminative feature vectors among overlapped vectors and add them in model reconstruction. We performed emotion recognition experiments on Emotional Prosody Speech Corpus (LDC). Our approach yielded superior performance when compared to the conventional classification method as well as the standard GMM approach. We will investigate the performance changed by the number of Gaussian mixtures and verify our method on other emotion corpus.

6. Acknowledgements

This work was partially supported by the Defense Acquisition Program Administration and the Agency for Defense Development under contract.

7. References