High-level Feature Weighted GMM Network for Audio Stream Classification

Rongqing Huang and John H.L. Hansen

Robust Speech Processing Group
Center for Spoken Language Research
University of Colorado, Boulder, CO, USA
{huangr,jhlh}@cslr.colorado.edu

Abstract

The problem of unsupervised audio classification continues to be a challenging research problem which significantly impacts ASR and Spoken Document Retrieval (SDR) performance. This paper addresses novel advances in audio classification for speech recognition. A new algorithm is proposed for audio classification, which is based on Weighted GMM Network (WGN). Two new high-level features: VSF (Variance of the Spectrum Flux) and VZCR (Variance of the Zero-Crossing Rate) are used to pre-classify the audio and supply weights to the output probabilities of the GMM networks. The classification is then implemented using weighted GMM networks. Evaluations on a standard data set — DARPA Hub4 Broadcast News 1997 evaluation data, shows that the WGN classification algorithm achieves over a 50% improvement versus the GMM network baseline algorithm. The WGN also obtains very satisfactory results on the more diverse and challenging NGSW (National Gallery of the Spoken Word [8]) corpus. Classification based on segmentation method is also explored.

1. Introduction

Most audio classification techniques focus on two different aspects: one is the particular feature employed, the other is the statistical model used. Speech and non-speech (music, songs, environmental sounds, etc.) segments have different distribution characteristics in both the time and frequency domains, so feature-based classification is generally an effective method. A number of studies have focused on alternative feature selection or development. For example, Zhang and Kuo [12] considered the Energy Function, Average Zero-Crossing Rate, Fundamental Frequency and Spectral Peak Tracks as their features; Lu, et al. [7] and Li [6] both considered 6 features for classification. The success of feature-based methods depends mainly on the discriminative power of the features, and methods are implemented either in a complex threshold-dependent scheme [7, 12], or with some pattern classification method (Euclidean Distance, Nearest Neighbor [11], Nearest Feature Line [6], etc.). Model-based classification methods have also been popular recently. Hain, et al. [3] trained four GMMs to classify Broadcast News (BN) data into 4 classes. Ajmera, et al. [1] compared the GMM with a MLP (Multilayer Perceptron) and reported that they were comparable; Scheirer, et al. [9] compared the GMM with MAP, KNN, and K-d models. The distinction between feature-based methods and model-based methods can be obscure, since many researchers consider both aspects to obtain the best performance improvement. For example, Ajmera, et al. [1] applied two posterior probability based features: entropy and dynamism for GMM and MLP classifiers; while Scheirer, et al. [9] compared thirteen features with four classifiers.

The features used in feature-based methods can be considered high-level features, and are represented in the time domain (Zero-Crossing Rate, Energy, etc.), or in the frequency domain (Subband Power, Low Short-Time Energy Ratio, etc.), and are typically not suitable for training a statistical model. As an alternative, low-level features such as the spectral based MFCC features are decorrelated and highly independent across the feature vector. However, low-level features encode phoneme level information, which can be inappropriate for speech/non-speech classification. Based on concepts discussed so far, features used in the feature-based methods and features used in the model-based methods are quite different. As such, most researchers treat the feature-based methods and model-based methods seperately and do not consider them in an integrated manner.

In this paper, a novel classification algorithm is proposed that combines the feature and model in a compact way that results in very efficient audio classification. First, two new high-level features, VSF (Variance of the Spectrum Flux), and VZCR (Variance of the Zero-Crossing Rate) are proposed; next, they are applied for pre-classification of input audio streams. The pre-classification will produce weights which are supplied to the output probabilities of a GMM network, and then the final classification is implemented using Weighted GMM Network.
Network (WGN). This algorithm combines the feature-based and model-based methods in a compact way rather than in a separate way, and achieves satisfactory results for audio streams from a range of acoustic scenarios.

2. Audio Classification in the WGN

Here, two high-level features are designed. One is in the time domain, the other is in the frequency domain. Features from different domains might better reflect diverse aspects of the audio structure.

2.1. VSF

The first feature is the Variance of the Spectrum Flux (VSF). Fig. 1 shows a flow diagram of the VSF feature extraction process for speech/non-speech classification.

The Spectrum Flux (SF) is the Ordinary Euclidean Norm of the delta spectrum magnitude, which is calculated as:

\[
SF = \| S_i - S_{i-1} \|_2 = \frac{1}{N} \left( \sum_{k=0}^{N-1} (S_i(k) - S_{i-1}(k))^2 \right) \quad (1)
\]

where \( S_i \) is the spectrum magnitude vector of frame \( i \), which is defined as:

\[
S_i(k) = \left\{ \sum_{n=0}^{N-1} s(n)w(n) \exp \left\{ \frac{-2\pi kn}{N} \right\} \right\}, \quad (2)
\]

where \( s(n) \) is the audio data, \( k, n \in [0, N-1] \), and \( w(n) \) is the window function.

Actually, the SF itself does not reflect significant differences between speech and non-speech. It is observed that speech alternates between transient, non-periodic speech and short-time stationary, periodic speech due to the phoneme transitions. On the other hand, music and environmental sounds could be periodic or monotonic and have more constant rates of change versus that seen in speech. This means the variance of the SF of speech should be larger than that for music or most environmental sounds. To explore this idea, the following experiment is designed. A group of 8 speech clips and 8 music clips are selected, where each is 5 seconds in duration. The clips are concatenated in an alternating fashion (speech, music, speech, music,...). At the end of this 80-second stream, a 10-second speech clip is concatenated followed by a 19-second environmental sound clip. Therefore, the resulting audio stream is 109 seconds long in duration. Using this 109-second stream, the SF and the VSF are calculated, then speech/non-speech classification is performed using the VSF feature. The SF is calculated on a frame-basis using Eq. 1, where the frame size is 20 ms with a skip rate of 10 ms. The VSF is calculated using the SF over a window of 20 sequential frames. Therefore, in one second of audio, there will be 5 combination decision frames, each resulting in a 0/1 decision for non-speech/speech. If the sum of the 0/1 decisions is less than 3 over 1 second period, the audio segment is classified as non-speech, otherwise it is classified as speech. Fig. 1 shows this decision process. Fig. 2 shows the classification results for the 109-second audio stream. The results show that only two seconds of music was mislabeled as speech from the 109-second passage.

\[\text{SPEECH} \quad 1>\quad \text{NON-SPEECH} \quad 1<\]

\[\text{1 Second has}\quad \text{VSF Points}\quad \text{VSF to VSF Conversion}\quad \text{20 SF to VSF Conversion}\quad \text{Spectrum Flux}\]

**Figure 1:** Speech/Non-speech Classification with VSF

\[\text{SPEECH} \quad \geq 3\quad \text{NON-SPEECH} \quad \leq 3\]

\[\text{SPEECH} \quad \text{VSF to}\quad \text{VSF Conversion}\quad \text{Spectrum Flux}\]

**Figure 2:** Speech/Non-speech labeling based on VSF of the 109-second test audio file: (a) SF value per frame (each frame is 0.01 sec.); (b) VSF value per combo frame (each combo frame is 0.2 sec.); (c) Final labeling of speech/non-speech per second. 1:speech; 0:non-speech.

2.2. VZCR

The second feature is based on the Zero-Crossing Rate (ZCR), which is a commonly used high-level feature in classification. ZCR is the number of zero-crossings within a frame in the time domain[5]. Since music and environmental sounds are more periodic or monotonic than speech, their ZCR will be more constant with less fluctuations. This should mean that the variance of the ZCR of speech should be larger than that for music and environmental sounds. Therefore, the same approach can be applied as that used for VSF to calculate the Variance of the ZCR (VZCR). Replacing the steps in the dashed box in Fig. 1 with the ZCR calculation, the audio
stream can be classified into speech/non-speech using the VZCR.

2.3. Classification in the WGN

Next, a new classification algorithm is proposed that combines feature-based methods and model-based methods in an efficient way. Here, the high-level features VSF and VZCR proposed in the previous section are used. The GMM Network (GN) is selected, since it is an efficient model-based classification method that has been considered by others [2, 3]. For the system formulation, three sets of GNs are trained: speech and non-speech; female and male; female broadband, female narrowband, male broadband, male narrowband and non-speech. The training data is the DARPA Hub4 Broadcast News 1996 training set. For the speech model, the training data is taken from the seven BN focus conditions [10]: F0–F5 and FX. For the non-speech model, the training data is selected from the gaps between the BN speech segments. The training data for narrowband models is from the F2 condition, and data from the other 6 focus conditions is used for broadband models. A 39-dimensional MFCC feature set is used for the models. The GMMs have between 96-256 mixture components depending on training data size and contain diagonal covariance matrices. There are no restrictions on the transition between GMMs/states, so all GMMs are connected to each other, forming a GMM Network. The Viterbi algorithm is then implemented to obtain the best classification results.

To train the diagonal covariance GMMs, low-level features such as MFCCs must be employed. Although high-level features cannot be used in model training directly, it is possible to integrate them as a weighting process of the output probabilities of the GMMs. High-level features can classify audio into speech versus non-speech as previously discussed. The weights of the output probabilities of the GMMs are set as follows: if the current segment is classified as speech, then the output probability of the speech GMM for that segment is multiplied by a weight larger than one, and unchanged for the non-speech model probability; if the segment is classified as non-speech, then that output probability of the non-speech GMM is multiplied by a weight larger than one, and unchanged for the non-speech model probability. If the segment is classified as non-speech, then the output probability of the non-speech GMM for that segment is multiplied by a weight larger than one, and unchanged for the speech model probability. This algorithm can improve classification accuracy significantly, since it combines the strengths of both high-level features and models in an efficient way.

3. Experiments

For the experiments, the evaluation data is DARPA BN data and NGSW data. This test data includes: (i) different structures of audio: interviews, reports, debates, etc.; (ii) various recording equipments: microphone, telephone, Edison cylinder disks, TV/radio; and (iii) various background noise: music, audience laughing, clapping, automobile noise (road, wind, turn signals), etc.

3.1. WGN Classification Evaluation

The DARPA Hub4 BN 1997 evaluation data is used in the classification evaluation. The GN algorithm will represent the baseline algorithm.

Table 1: Speech/Non-speech Classification in the WGN

<table>
<thead>
<tr>
<th>Scheme</th>
<th>Precision Speech</th>
<th>Recall Speech</th>
<th>Precision Non-Sp</th>
<th>Recall Non-Sp</th>
<th>Total Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>GN</td>
<td>98.08%</td>
<td>94.86%</td>
<td>52.81%</td>
<td>94.9%</td>
<td>93.4%</td>
</tr>
<tr>
<td>VSF</td>
<td>98.74%</td>
<td>92.21%</td>
<td>90.09%</td>
<td>91.5%</td>
<td>92.9%</td>
</tr>
<tr>
<td>VSF-WGN</td>
<td>98.22%</td>
<td>93.40%</td>
<td>94.43%</td>
<td>96.7%</td>
<td>92.9%</td>
</tr>
<tr>
<td>VZCR</td>
<td>98.80%</td>
<td>83.63%</td>
<td>94.09%</td>
<td>83.5%</td>
<td>96.7%</td>
</tr>
<tr>
<td>VSF-VZCR-WGN</td>
<td>98.22%</td>
<td>98.48%</td>
<td>54.44%</td>
<td>96.9%</td>
<td>96.9%</td>
</tr>
</tbody>
</table>

Table 1 demonstrates that although the classification performance of high-level features (i.e., VSF and VZCR) is less accurate than that of GN, they can still contribute to improving GN classification. The VSF weighted GN can outperform both GN and VSF algorithms; The VZCR weighted GN can also outperform both GN and VZCR algorithms. Finally, the combined VSF and VZCR weighted GN outperforms the baseline GN at all levels (precision and recall in both speech and non-speech parts) and it improves the total frame accuracy from 93.4% to 96.9%. The relative improvement in error reduction is over 50%. Table 1 also shows that if VSF and VZCR are combined in the weighted GN, it only achieves slightly better results than the single VSF weighted GN classifier. This suggests that VSF and VZCR contribution for classification do overlap.

VSF and VZCR features can help the GN significantly in classifying speech and non-speech, however, they cannot distinguish female from male speech, or broadband from narrowband speech. Therefore, if further audio classification is needed, it is necessary to find other high-level features or turn to other methods. During the experiments, we observed that performing segmentation first will help in the classification process. Since speech blocks from segmentation are much more homogeneous (same speaker in a consistent acoustic environment) than any pre-defined processing windows for classification, it is reasonable to expect the classification result based on segmentation to be better than classification without segmentation. Table 2 demonstrates such improvement by applying \( T^2 \)–BIC [4, 13] segmentation.

Table 2: Audio Classification Without/With Segmentation

<table>
<thead>
<tr>
<th>Discrimination Type</th>
<th>Frame Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female / Male</td>
<td>86.7% / 92.9%</td>
</tr>
<tr>
<td>Female-Broadband / Male-Broadband / Female / Male-Narrowband / Male-Narrowband / Non-Speech</td>
<td>81.0% / 82.1%</td>
</tr>
</tbody>
</table>

Since there are non-speech parts, the VSF or VZCR weighted GN can still be applied in the five-state classification framework (Female-Broadband / Male-Broadband / Female-Narrowband / Male-Narrowband / Non-Speech). From Table 3, the VSF or VZCR weighted GN algo-
algorithm can improve performance over classification based on segmentation algorithm.

Table 3: Five-State Classification in a WGN

<table>
<thead>
<tr>
<th>Scheme</th>
<th>Frame Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>GMM Network (GN)</td>
<td>81.0%</td>
</tr>
<tr>
<td>Based on Segmentation</td>
<td>82.1%</td>
</tr>
<tr>
<td>VZCR-WGN</td>
<td>83.4%</td>
</tr>
<tr>
<td>VSF-WGN</td>
<td>83.6%</td>
</tr>
<tr>
<td>VSF+VZCR-WGN</td>
<td>83.6%</td>
</tr>
</tbody>
</table>

It would be reasonable to question why classification with the WGN should be better than that based on segmentation. It is suggested that in the segmentation phase, missing the actual break points has a pronounced impact on classification, since it causes non-homogeneous audio segments to be combined together, corrupting the statistical analysis and therefore classification errors will occur. Furthermore, GMM networks can make decisions based on short duration audio blocks, so long homogeneous segments from segmentation will not provide much benefit. However, this additional data can still improve classification. Re-stated, if the non-speech parts in the audio stream need to be classified, the WGN (i.e., VSF or VZCR weighted) can be applied, otherwise classification based on segmentation is an acceptable method.

3.2. NGSW Data Evaluation

Two sets of audio materials from the NGSW corpus are selected. The first consists of audio samples from 7 decades (1940s—2000s), where each clip is a typical audio representation of topic, recording media/equipment and speaker content for that period. The second set is an audio stream from the 1960s which is more topic specific.

Table 4: NGSW Data Classification in WGN

<table>
<thead>
<tr>
<th>Scheme</th>
<th>Frame Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>7-decades</td>
<td></td>
</tr>
<tr>
<td>GMM Network (GN)</td>
<td>73.3%</td>
</tr>
<tr>
<td>1960s</td>
<td></td>
</tr>
<tr>
<td>VZCR only</td>
<td>79.6%</td>
</tr>
<tr>
<td>VZCR-WGN</td>
<td>86.5%</td>
</tr>
</tbody>
</table>

Here, VZCR is used as the high-level feature for GMM network weighting. Table 4 clearly shows the effectiveness of the WGN algorithm. Most clips in the 7-decades data were recorded outdoors, thus containing diverse and varying levels of noise. Some speech was misclassified as noise, and some audience noise was misclassified as speech, though overall frame accuracy was 86.5%. The 1960s data consists of indoor broadcast news, which is less noisy than the 7-decades data. Since the GMMs were trained with BN data, the 1960s test data should match the models well. However, the 7-decades data does not match the system models at all in terms of topic, speaker, and recording environment/equipment. Such differences can be reflected from the frame accuracy of GN classification of 96.0% versus 73.3%. However, the high-level feature shows the consistency in classification with an 82.2% versus 79.6% performance rate for clean or noisy conditions. From this observation, we conclude that GN classification performance is more sensitive to the data. If the test data is quite different from the training data, some form of model adaptation should be applied to the GMMs. Secondly, the high-level features are robust to the acoustic environments, however, their classification power is limited, and that by combining their effects with GN (i.e., WGN) results in an effective overall classification method.

4. Conclusions

This study has considered advances in unsupervised audio classification for speech recognition. Two new high-level features, VSF and VZCR were proposed for audio classification and a novel classification algorithm: Weighted GMM Network (WGN), was presented. VSF and VZCR were shown to be robust and effective for speech/non-speech classification. WGN combines a feature-based method and model-based method in a compact and reliable way. It improves the frame accuracy from 93.4% to 96.9% over traditional GMM Network (GN) and outperforms the baseline system at ALL levels. WGN also achieves satisfactory performance in the diverse NGSW corpus.

5. References