GENDER IDENTIFICATION USING A GENERAL AUDIO CLASSIFIER

Hadi Harb, Liming Chen

Dept. Mathématiques Informatique, Ecole Centrale de Lyon, France
{Hadi.harb, liming.chen}@ec-lyon.fr

ABSTRACT

In the context of content-based multimedia indexing gender identification using speech signal is an important task. Existing techniques are dependent on the quality of the speech signal making them unsuitable for the video indexing problems. In this paper we introduce a novel gender identification approach based on a general audio classifier. The audio classifier models the audio signal by the first order spectrum's statistics in 1s windows and uses a set of neural networks as classifiers. The presented technique shows robustness to adverse audio compression and it is language independent. We show how practical considerations about the speech in audio-visual data, such as the continuity of speech, can further improve the classification results which attain 92%.

1. INTRODUCTION

Gender identification based on the voice of a speaker consists of detecting if a speech signal is uttered by a male or a female. Automatically detecting the gender of a speaker has several potential applications. In the context of Automatic Speech Recognition, gender dependent models are more accurate than gender independent ones. Hence, gender recognition is needed prior to the application of one gender dependent model. In the context of speaker recognition, gender detection can improve the performance by limiting the search space to speakers from the same gender. Also, in the context of content based multimedia indexing the speaker's gender is a cue used in the annotation. Therefore, automatic gender detection can be a tool in a content-based multimedia indexing system.

This paper describes a novel approach for voice-based gender identification for audio-visual content-based indexing. Several acoustic conditions exist in audio-visual data: compressed speech, telephone quality speech, noisy speech, speech over background music, studio quality speech, different languages, and so on. Clearly, in this context, a gender identification system must be able to process this variety of speech conditions with acceptable performance.

Two main approaches can be used for gender identification. One approach is to use gender dependent features, such as the pitch. The other approach is to use a general pattern recognition approach based on general speech features such as the Mel Frequency Cepstral Coefficients (MFCC).

The pitch information was used in [3]for the problem of gender identification. However, pitch estimation relies considerably on the speech quality. This drawback makes such an approach non-suitable for our problem of video indexing. Also, the reported results are based on 5s files which is not an image of the frame-based classification accuracy in a continuous speech signal.

[1]followed a general audio classifier approach using MFCC features and Gaussian Mixture Models (GMM) as a classifier. When applied to gender identification, the results are 73% of classification accuracy which is not promising.[4] used a combination of pitch-based approach and general audio classifier approach using GMM. The reported results are based on 7s files after silence removal.

The majority of the proposed techniques for gender identification are not suitable for the speech in audio-visual data since they are dependent on the quality of speech and they assume some preprocessing of the speech segments, such as silence removal, voiced speech detection, or phoneme recognition.

In this paper we propose a gender identification system based on a general audio classifier. The proposed technique doesn't assume any constraint on the speech quality or segment lengths, in contrary to the existing techniques, with a classification accuracy that attains 92%.

2. THE GENERAL AUDIO CLASSIFIER APPROACH

Gender identification is a general audio classification problem with two classes: male speech, and female speech. The potential general features that can be used to obtain feature vectors are, for instance, MFCC or Mel Frequency Spectral Coefficients (MFSC) features. However, those features are typically extracted every

10ms implying a great variability of the features in each class. Also, such features capture phoneme-like characteristics. It is desirable that the features capture characteristics that are not limited to phonemes or words. We propose the use of the long term structure of the audio spectrum as the features for gender identification. The long term structure has less variability within each class. We use the first order statistics of the signal's spectrum in relatively large windows (1s) of audio to capture the structure of the spectrum [6]. It is also preferable not to use the pitch in the creation of the feature vector since this increases the complexity. Also, pitch estimation algorithms are not sufficiently effective when noise or music is mixed with the speech signal.

We use neural networks as classifiers.

3. THE GENDER DETECTOR

The system we propose for robust gender identification is based on three main steps: the Feature Calculation, the Normalization/Statistics, and the Neural Networks.

3.1. Signal analysis

The signal analysis that we perform consists of extracting the Fast Fourier Transform (FFT) with a Hamming window of 30ms width and a 20ms overlap. The spectrum is further filtered conforming to the Mel Scale to obtain a vector of 20 Spectral coefficients every 10ms: the Mel Frequency Spectral Coefficients (MFSC). We do not use the Mel Frequency Cepstral Coefficients (MFCC) features since it was experimentally shown that the MFSC features perform slightly better.

3.2. Statistics

As we mentioned before, the basic features of our gender detector are the first order statistics of the signal's spectrum. The signal is segmented into non-overlapped windows of duration "T" (T is typically 1s). In each window the mean and the variance of the MFSC vectors are calculated. That is, each window "T" is modeled by a mean vector and a variance vector of MFSC features. The basic unit for training and testing is then the "T" window. Notice that in the majority of cases this resolution of 1s is sufficient.

3.3. The feature vector

One feature vector is extracted for each window "T". The feature vector is created by concatenating the 40 values of the mean and the variance of the MFSC in the "T" window. However, we need that the classifier captures the relation between the frequencies in the spectrum not the frequencies themselves. So, we normalize the mean values by their respective maximum and the same is done for the variance values. Hence, the classifier will capture the relations between the peak in the spectrum and the other frequency bands.

This normalization is extremely important for the system to be robust to loudness changes. Also, the fact of using a Neural Network as a classifier and using the sigmoid function as an activation function necessitates this kind of normalization of the feature vector. Generally optimal values in the feature vectors are in the [0-1] range. The Neural Network risks saturation if feature vectors contain values higher than 1. The saturation means that synaptic weights change very slowly when training the neural network, implying a very long training time.

3.3. The Neural Network as a classifier

Theoretically, any classifier can be used for the classification of the normalized feature vectors. However, the use of a Neural Network as a classifier is suitable for our problem.

A Neural Network is very fast in the classification once trained; which is important for the real time applications. Moreover, a Neural Network can theoretically model complex shapes in the feature space. Also, the training time (or the number of epochs) can be an indicator on the complexity of the decision boundary for a given training set. Beside the classical accuracy indicator, this is another important indicator to examine the efficiency of the selected features and their suitability for a given classification problem. Finally, the compact representation of Neural Networks facilitates potential hardware implementation of the classifier.

The Neural Network we have used is a Multi Layer Perceptron (MLP) with the error back-propagation training algorithm and the sigmoid function as an activation one.

4. PRACTICAL ISSUES

Although the evaluation of a gender identifier is generally based on the frame classification accuracy, real world applications imply some improvements over the frame-based accuracy.

There are mainly two considerations for real world systems:

4.1. Combining experts

In many cases different training data can be used to train different classifiers (experts). It is desirable that we can combine the results of different experts that are trained on different training data, Figure 1. Some improvements in the classification results can be obtained when a minimum correlation between the results of different experts exists.

It is also required that the combination of experts performs better than the average of single experts.

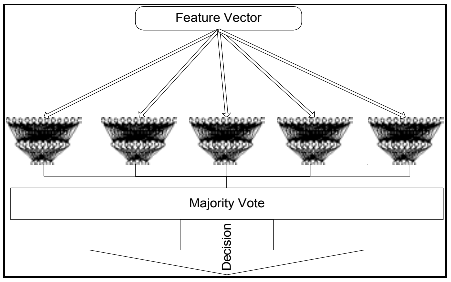

Figure 1 The multi-expert combination

Figure 2 shows the classification error for 8 experts and their combination. We can notice that the combination of experts (expert10) performs better than 7 of them and worse than the best expert. This means that such a combination is advantageous.

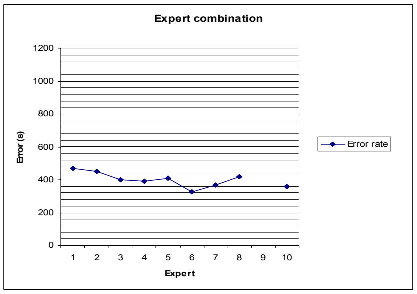

Figure 2 The classification error for several experts on the same test data. We can notice that the combination of experts (expert10) performs better than average.

4.2. Smoothing the classification results

The speech contained in audio-visual programs is continuous. This assumption leads us to incorporate the results of neighboring frames to smooth the classification results. We segment the speech signal using a metric-based approach and the KullBack-Leibler distance as measure of similarity between neighboring windows [5]. Each segment is assumed to contain one acoustic condition, in our case speech from the same gender. The labels of the frames in each segment are smoothed based on the average classification result for the entire segment. It is supposed that each segment contains speech from the same gender. However, we minimize the risk of mixing different genders in the same segment by decreasing the threshold used to segment the speech signal. Generally the classification results are slightly improved using such a smoothing technique, and in the worst cases the results do not change.

5. EXPERIMENTS

Several experiments were carried out to evaluate the proposed classifier for several classification conditions. In all the experiments we evaluated a single Expert gender identifier with and without smoothing. The database used to evaluate the system consists of recordings from four French radio stations and one English radio station. Training data, Train_F, was extracted from the recording of 1 French radio station; it consists of speech from news programs and meetings. The data from the other radio stations was used as the French test data, Test_F. Test_F data was also compressed with MPEG layer-3 coder at a 16Kbps rate to obtain Test_F_mp3. Furthermore, 1400s were selected from the English radio station containing telephone speech, outdoor speech and studio speech constituting the Test_E dataset. Table 1 shows the composition of the datasets used in the experiments.

Table 1 The evaluative datasets' durations

dataset	duration (s)
Train_F	2200
Test_F	1800
Test_F_mp3	1800
Test_E	1400

5.1. Classification accuracy

In this experiment the data form Train_F was used to train the system, and the data from Test_F was used to evaluate the classifier. The amount of training data was changed in order to observe the effect of increasing the training data on the classification accuracy. It is expected that by increasing the amount of training data the classification accuracy would increase. However, the experimental results show that if a classifier was trained on 160 s of speech the error rate is about 22% and this error rate is about 11% if the classifier was trained on 2200s of speech.

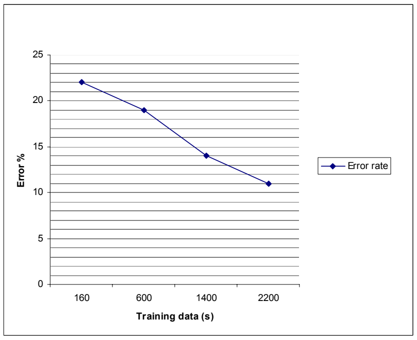

Figure 3 The error rate as a function of the duration of the training data

The overall frame-based classification results shown in Table 2 are 89%. Those results are comparable to the reported results in[4] although in [4] the results are based on 7seconds files and some preprocessing (silence removal) was made to the testing and training data. A comparison to the results reported in [1] (73%) shows the effectiveness of the proposed approach over a straightforward general audio classifier approach based on MFCC features and Gaussian Mixture Models. One can notice that when the smoothing is performed, Section 4.2., the classification results are improved.

Table 2 Classification accuracy for radio data with and without smoothing

	Male Accuracy	Female Accuracy	Total Accuracy
Smoothing	94.70 %	88.75 %	91.72 %
Frame-based	90.00 %	88.10 %	89.05 %

We have carried out another experiment to observe the effectiveness of the proposed approach for language change and when the test data contains telephone and outdoor speech. The system was trained on 2200s from Train_F and tested on 1400s from Test_E data. The results shown in Table 3 demonstrate that the proposed approach is language independent though the performance is slightly degraded. Notice that the system was faced to language and channel changes.

Table 3 Classification accuracy for English data containing telephone and outdoor speech with and without smoothing

	Male Accuracy	Female Accuracy	Total Accuracy
Smoothing	93.14 %	89.14 %	91.14 %
Frame-based	88.28 %	85.00 %	86.64 %

The last experiment that we have carried out was to test the effectiveness of the gender identifier when the speech is compressed at low compression rates. The Test_F_mp3 was used as the test set and Train_F was used for training. As shown in Table 4 the proposed approach is robust to low compression ratios.

Table 4 Classification accuracy for radio data compressed at 16Kbps with MPEG3 coder with and without smoothing

	Male Accuracy	Female Accuracy	Total Accuracy
Smoothing	94.22 %	88.00 %	91.11 %
Frame-based	89.11 %	87.23 %	88.17 %

6. CONCLUSION

This paper presented a voice-based gender identification system using a general audio classifier. The audio classifier models the audio signal by the spectrum's first order statistics in 1s windows and uses a set of neural networks as classifiers. The system was tested on adverse conditions of compression, channel mismatch and language change. The results of 89% of frame based accuracy are satisfying and consist of a clear improvement over a straightforward general audio classifier approach based on MFCC features and Gaussian Mixture Models. It was also shown how smoothing the classification results can further improve the accuracy of the system up to 91.72%.

11. REFERENCES

[1]. Tzanetakis G., Cook P. Musical genre classification of audio signals *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, July 2002

[2]. Simon Haykin, "Neural Networks A Comprehensive Foundation", Macmillan College Publishing Company,1994.

[3]. Parris E. S., Carey M. J., Language Independent Gender Identification, *Proceedings of IEEE ICASSP*, pp 685-688, 1996

[4]. Slomka S., Sridharan S., Automatic Gender Identification Optimised For Language Independence, *Proceeding of IEEE TENCON-Speech and Image Technologies for Computing and Telecommunications* pp 145-148, 1997

[5]. M. Seigler, U. Jain, B. Raj, R. Stern, Automatic segmentation, classification, and clustering of Broadcast news audio*, Proc. Of the DARPA speech recognition workshop*, February 1997.

[6]. Hadi Harb, Liming Chen, "Segmentation et classification du son", patent pending nf 02 08 548, July 2002