SPEAKER RECOGNITION AND VOICE MINING

Olakunle Ogunsuyi, Dr. Carol Y. Espy-Wilson, Sandeep Manocha and Srikanth Vishnubhotla

INTRODUCTION

This work investigated issues related to speaker recognition, with emphasis on speech detection in multi-speaker conversations from the real ENRON speech corpus. Two tasks were performed - the first was to verify speaker identification performance by training speaker models with and without segmentation, the segmentation being automatic and manual. The second was to analyze the performance of an algorithm for automatic detection of creaky voice quality. Research has shown that a small number of well chosen acoustic parameters that capture voice quality can greatly enhance speaker recognition, and thus a creakiness detection algorithm can prove useful for such applications.

Voice mining is an extension of the speaker identification task, and involves speaker detection in a set of multi-speaker conversations. Given a database of telephone conversations, the task is to identify conversation that have a speaker in common. Segmentation is needed to overcome the problem of two speakers talking at the same time and also rapid speaker interchange which hinders accuracy of speaker dependent feature extraction. Voice mining techniques can be implemented without segmentation, or with automatic or manual segmentation. The ENRON database used for the experiments had the problems of poor audio quality and background noise. Thus, automatic segmentation is not very effective and manual segmentation is considered.

METHOD

The Voice Mining Task

- Using the real ENRON database, ten target speakers which were all present in numerous conversations were selected. Ten conversations were selected with each of them having at least one of the target speakers present in them.
- Each conversation was manually segmented with the silence portion of all conversations discarded until two minutes of speech from the target speaker was obtained. Another two minutes of speech was obtained by manual segmentation for each of the ten target speakers to give four minutes of speech for training the speaker models.

Training and Testing Speaker Model

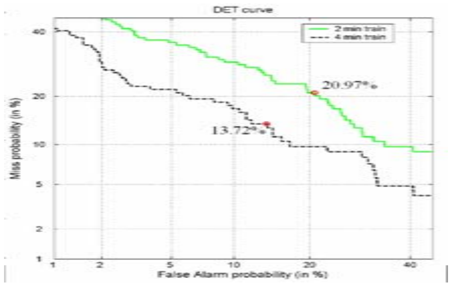
- Speaker dependent features/parameters were extracted from each of the ten segmented speech samples. The Mel-frequency Cepstral Coefficients, which implicitly code the vocal tract and source information, were used.
- Ten speaker models were created and trained using the two minute and four minute speech data. Gaussian mixture models (GMM) were used to form a statistical representation of the speaker information/features. A Universal Background Model (UBM), also called an imposter model, was constructed from a large amount of the ENRON speech database that is disjoint from the training and testing data. Each test file was compared against the speaker model (based on a single conversation) and the UBM.
- A score was computed for all conversations. A higher score implies that the training and test conversation have a target speaker in common .

Automatic Detection of Creakiness

- Acoustic Parameters attempt to explicitly capture the source information and different vocal tract configurations for speaker ID applications. These parameters have a better performance compared to using MFCCs in speech feature extraction. To evaluate the effect of adding creakiness voice quality as one of these parameters, a creakiness detection algorithm is being developed in the lab. This algorithm was executed on 50 speech files of 35 seconds duration. The accuracy of the algorithm was analysed by comparing the creakiness detection profile of each speech file with the perceptual voice quality of the speech file. In addition, pitch information is used since creakiness often involves a very low pitch.

RESULTS

DET Curve Showing the Effect of Increasing amount of Training Data



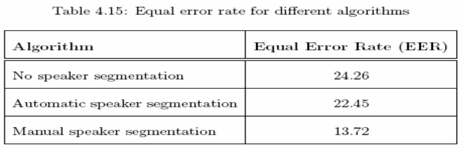
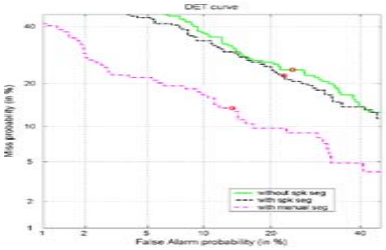


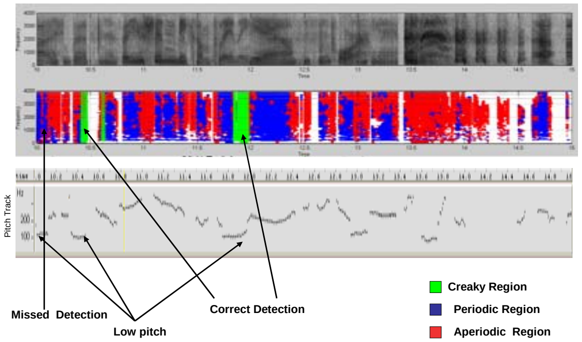
Table 4.15: Equal error rate for different algorithms

Algorithm	Equal Error Rate (EER)
No speaker segmentation	24.26
Automatic speaker segmentation	22.45
Manual speaker segmentation	13.72

DET Curve Showing The Effect of Manual Segmentation



Creakiness Detection Algorithm Output with Pitch Information



CONCLUSION

The two minute training data speaker model with manual segmentation outperformed the automatic segmentation method slightly, giving an equal error rate (ERR) of 20.97%. Using four minutes for the speaker model with manual segmentation distinctively outperformed the automatic segmentation method, giving an EER of 13.72%. These results show the importance in having pure data for training the speaker model, and that increasing the amount of training data further improves the speaker model yielding improved performance.

For the creakiness detection, the algorithm performed better on female speech data than on male speech data. It was discovered that the pitch information of the speech data substantially helps in detecting creaky regions. Therefore, it is suggested that low pitch should be added as one of the conditions for creakiness