SPEAKER RECOGNITION AND VOICE MINING

Olakunle Ogunsuyi

Mentor: Srikanth Vishnubhotla & Dr. Carol Espy-Wilson

Reasearch Internship In Telecommunication Engineering (RITE)

MERIT 2006

INTRODUCTION

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 from the real ENRON speech corpus, the task is to identify conversations that have one or two speakers in common.

Comparing Enron & Switchboard

Switchboard 

- Fixed training and test data
- Prior target speaker models
- Clean audio
- Controlled collection
- Only 2 speakers

Enron 

- No defined training/test split
- No target speaker models
- Difficult acoustics
- Real-life database
- Sometimes many speakers

Two issues studied:

- Verifying speaker identification performance by training speaker models with and without segmentation, the segmentation being automatic and manual.
- Analyzing the performance of an algorithm for automatic detection of creakiness, a voice quality that will eventually be used as a parameter for speaker identification.

METHOD

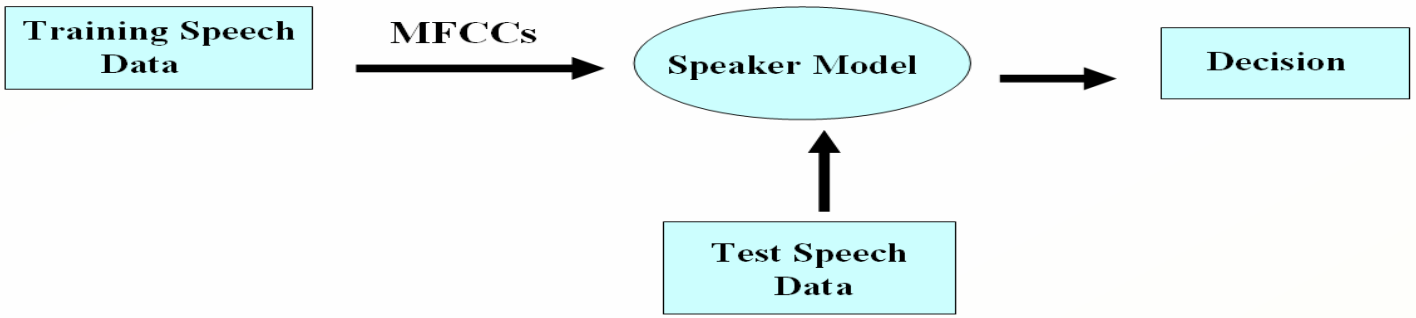
The Voice Mining Task

- Using the real ENRON database, select ten target speakers which are present in numerous conversations
- Ten conversations were selected with each of them having at least one of the target speakers present .
- Each conversation was manually segmented with the silence portion of all conversations discarded until four minutes of speech from the target speaker was obtained.

Training and Testing Speaker Model

- The Mel-frequency Cepstral Coefficients (MFCCs) were the speaker-dependent parameters extracted from the ten segmented speech samples to implicitly code the vocal tract and source information .

- Speaker models were created and trained from the features using Gaussian mixture models (GMM) which form a statistical representation of the speaker information/features



- Each test file was compared against the speaker model (based on a single conversation) and a Universal Background Model (or imposter model) constructed from disjoint training and testing data

- 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.

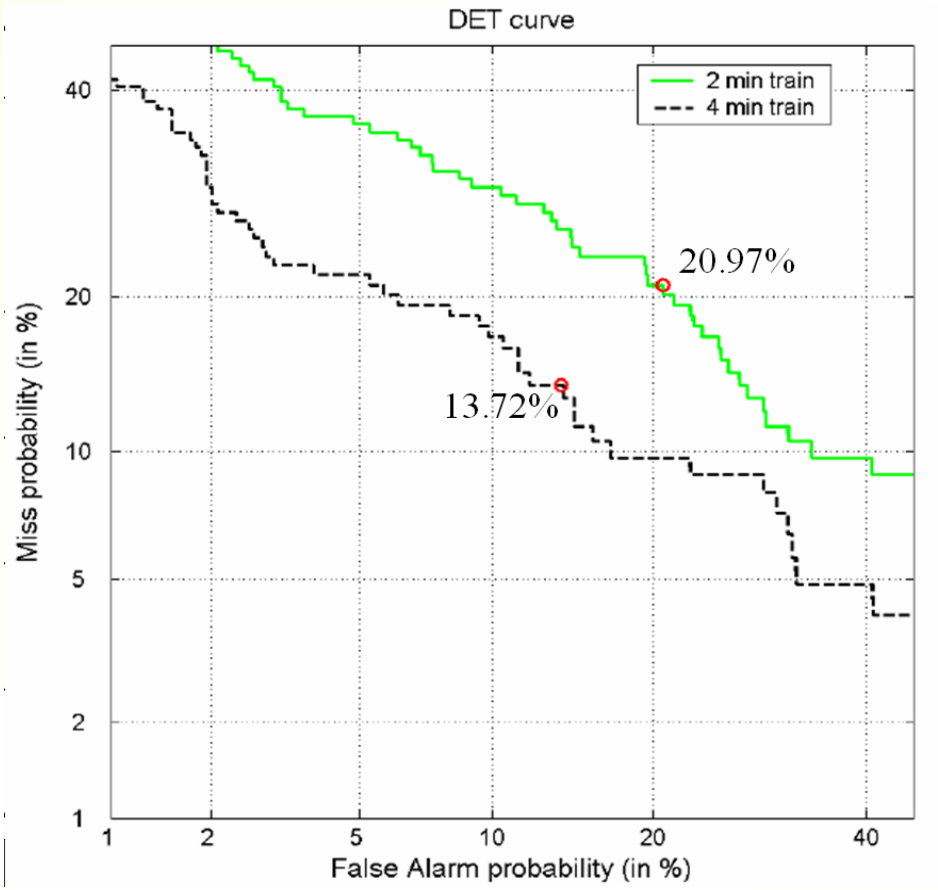
• These parameters have a better performance compared to using MFCCs in speech feature extraction.

• To evaluate the effect of adding creakiness voice quality as one of these parameters, a creakiness detection algorithm is being developed in SCL.

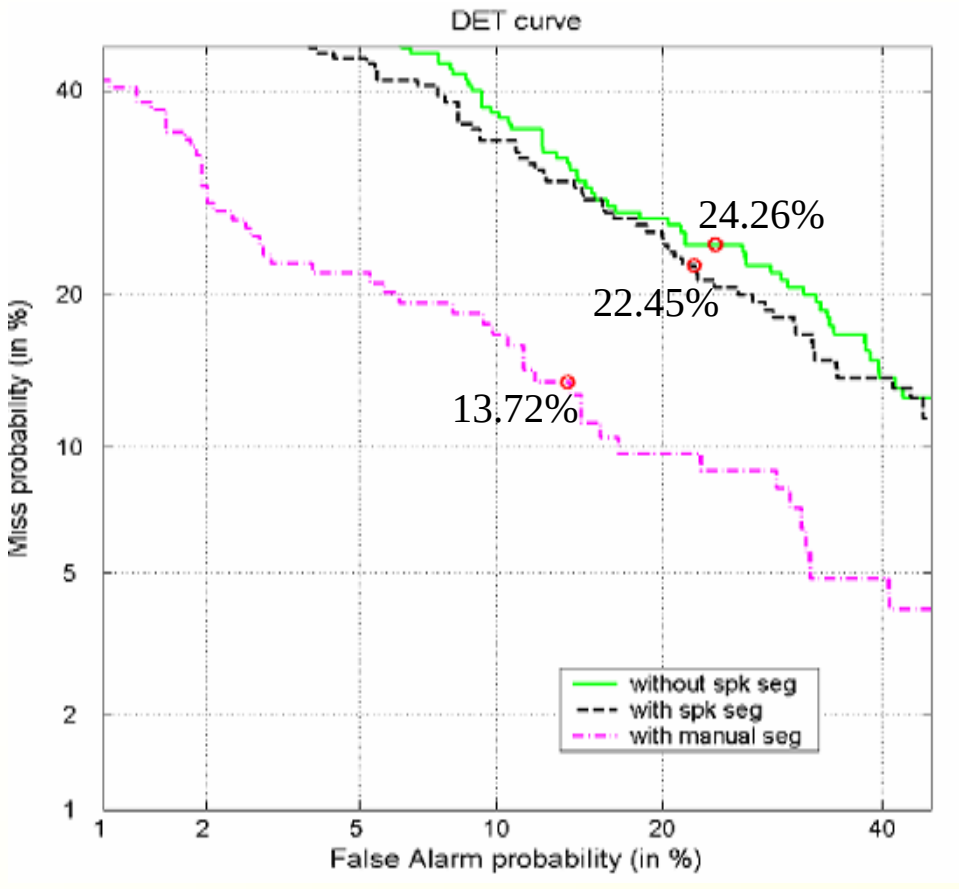
• This algorithm was executed on 50 speech files of 35 seconds duration. The accuracy of the algorithm was analyzed by comparing the creakiness detection profile of each speech file with the perceptual voice quality of the speech file

RESULTS

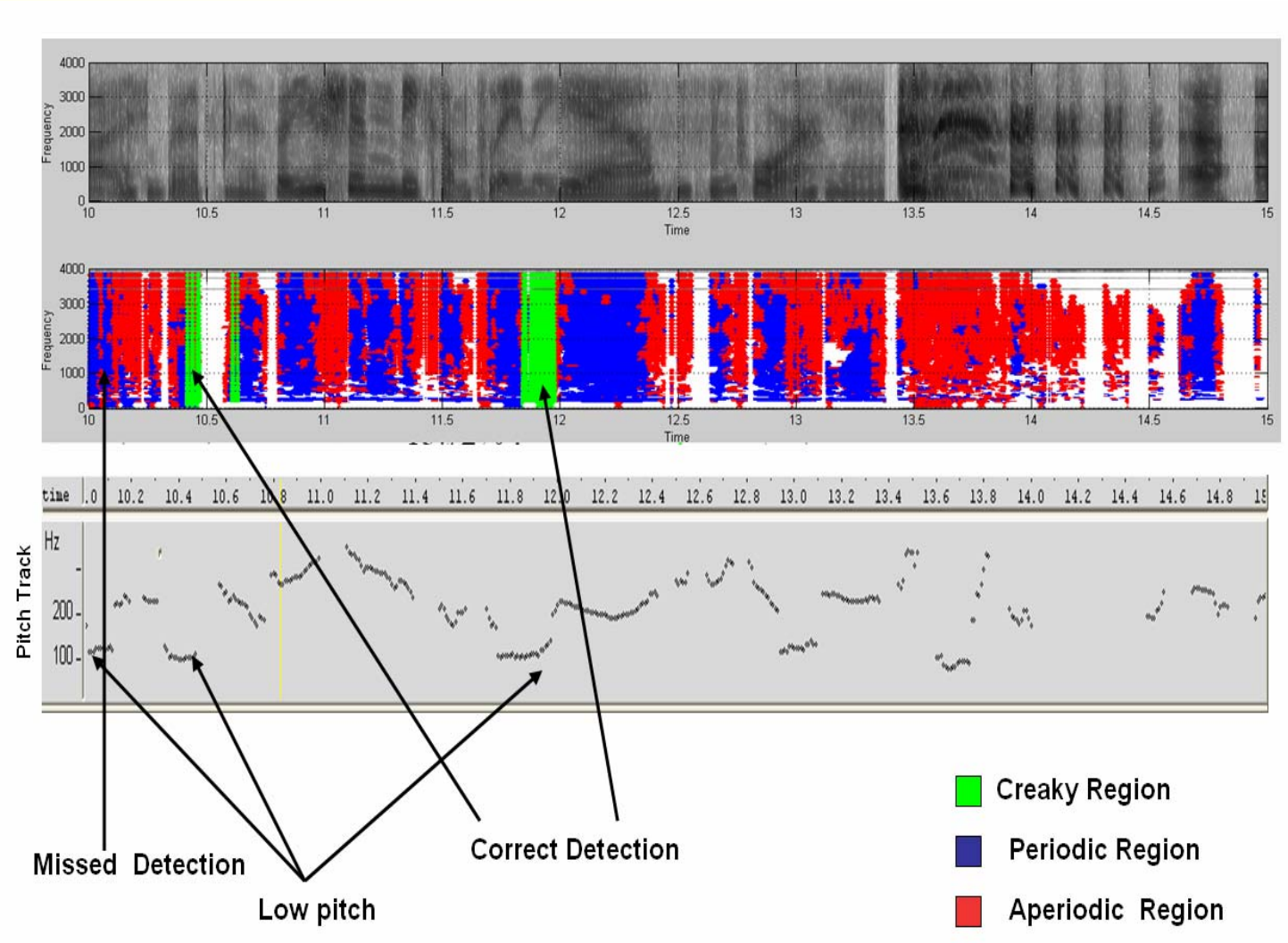
DET Curve Showing the Effect of Increasing amount of Training Data



DET Curve Showing The Effect of Manual Segmentation



Creakiness Detection Algorithm Output with Pitch Information



CONCLUSION

- Using four minutes for the speaker model with manual segmentation distinctively outperformed the automatic segmentation method.

- 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.

- 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