De-Identification of Speech

Bhaskar Adavi
Carnegie Mellon University
Pittsburgh, PA 15213
badavi@andrew.cmu.edu

Bharat Kandoi
Carnegie Mellon University
Pittsburgh, PA 15213
bkandoi@andrew.cmu.edu

Prashant Malani
Carnegie Mellon University
Pittsburgh, PA 15213
malani@cmu.edu

Sahil Sapre
Carnegie Mellon University
Pittsburgh, PA 15213
sahilsapre@cmu.edu

Abstract

In this paper, we describe an efficient method of de-identification of speech such that the transformation from the source speech is furthest away from the source features, yet fully intelligible. We have designed a speaker ID system that is 91.8% accurate in identifying 20 utterances spoken by 30 speakers - 23 standard American newsreaders, 5 speakers from the CMU Arctic database, and 2 native Indian speakers. We then de-identify these voices using voice conversion such that the speaker ID system trained to correctly identify these speakers gets confused when presented with the de-identified voices as input.

1 Introduction

Telephonic dialogues are a popular means of information retrieval. However, they are also a hidden channel for the invasion of the end-users' privacy. Companies could maintain records of what the users said, without their consent or knowledge. For example, recorded conversations between doctors and patients could potentially reveal confidential information about the patients. One way of protecting people from this invasion of privacy is to de-identify the speaker's voice such that the speech still sounds natural and fully intelligible, yet does not reveal information about the identity of the speaker.

In order to truly test the efficacy of any de-identification mechanism, an accurate speaker ID system is required. Such a system should, once trained to correctly identify a set of speakers, be able to identify any new voice input from these speakers with high accuracy. Such a system should also maintain its accuracy when the set of speakers have similar voice characteristics, like pitch and accent.

The organization of this paper is as follows. In section 2, a brief summary of the various feature extraction methods employed in both the speaker ID system as well as voice transformation is discussed. In section 3, the working principle behind the speaker ID system is explained, along with the test results from simulations. In section 4, the Voice Conversion technique is presented, and in section 5 the method used to select the 'best' transformation for de-Identification is outlined. In section 6, a summary for the results of the de-identification algorithm are presented and section 7 contains concluding remarks.

2 Feature Extraction

2.1 Mel-frequency cepstral coefficients (MFCC)

The mel-frequency cepstrum (MFC) is a representation of the short term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency. Here, the frequency bands are equally spaced on the mel scale, which better approximates the human auditory response than linearly-spaced bands.

2.2 Fundamental Frequencies (F0)

Frequency measure that denotes the number of times the vocal folds open and close per second. The typical male F0 value is 120 Hz, while the value for females is 180 Hz

3 Speaker Identification System Design

The basic methodology behind the Speaker ID system is to create statistical models that most closely approximate the voices to be tested, and then compare inputs against these models to find the model that most closely resembles them. We have improved upon a basic MFCC-based speaker ID system [1], adding weights for clustering of MFCCs and for F0s.

3.1 Generation of Gaussian Mixture Models (GMMs)

From each input file that is used to train the system for a particular speaker, the MFCCs and F0s are used to create GMMs that best approximate these parameters. Using an Expectation Maximization algorithm, GMM parameters such as means, variances and weights are obtained which best represent the speakers voice, for each audio file.

3.2 Clustering of MFCCs

To further improve the accuracy of the speaker ID system, clustering is performed on the sets of MFCCs for each of the training audio files for a speaker using the k-means method. The mean of the centroids obtained from each file is computed and stored. The clustering method used was developed by Esfandiar Zavarehei.

The GMM models for the MFCCs and F0s, along with the results of clustering, form the code book used for testing.

3.3 Training Phase

To train the speaker ID, we use the spectral parameters of each of the training input files, and obtain a log likelihood estimate of how close a particular parameter is to each of the GMMs constructed for that parameter. This process is performed for both MFCCs and F0s. The mean and standard deviation of the log likelihood estimates is then measured for both MFCC and F0 values.

This process can be repeated for multiple speakers to create a bank of speaker models to test against.

3.4 Testing Phase

To test an arbitrary input file for a particular speaker, the extracted MFCCs and F0s of the test file are sent into the system. The MFCCs are compared against a speakers GMMs (calculated in section 3.1) and the mean is taken of the log likelihoods that results for each GMM. A score is then assigned to the average value based on how many standard deviations it is away from the mean calculated during training. The standard deviation used here is the one obtained during training. This process is repeated for the F0 values of the test input.

The MFCCs of the test input are then clustered to find a centroid. The distance of this centroid is then measured from the mean of the centroids for the model, which was calculated in section 3.2.

Assuming the MFCCs, F0s and centroidal distances are orthogonal to each other, a Euclidean distance measure is obtained for each speaker model. The system chooses the model with the least score as its conclusion about the identity of the test voice.

4 Voice Transformation

The methodology used to de-identify speech is to transform the input voice such that it is as far as possible from the original voice, and successfully confuses the speaker ID system into mistaking it for another.

Subsequently, the Festvox transformation tools developed by the speech group at the Language Technologies Institute, Carnegie Mellon University, are used to perform voice transformations. This software creates mappings to convert the input voice to a specified output voice based on the joint probabilities of the two voices, while also factoring in global variance parameters.

Models are constructed to convert CMU's arctic database of voices to two native Indian speakers' voices, both male. The test utterances of 30 speakers (23 standard American newsreaders, 5 from the CMU arctic database, and the 2 native Indian speakers) through these models, and discard the unintelligible outputs.

On empirical analysis, it is observed that a male voice passed through a male-male transformation is usually intelligible. Similarly, a female voice passed through a female-male transformation is usually intelligible.

5 Transformation Selection for De-Identification

The best transformation is defined as the most de-identifiable yet fully comprehensible transformation of the input utterance.

By clustering the non-transformed original speaker's utterances using a variation of the k-means algorithm, a transformed utterance is chosen whose clusters are furthest away from it. We start with two clusters formed uniformly distributed about the mean of the input data, and successively split high population clusters, killing the low population clusters.

Initially, for a known speaker, it is empirically determined which transformations are intelligible, and only their clusters are used to find the most de-identifiable transformation.

The datasets of MFCCs and F0s for transformed voices were analyzed to obtain some trend or metric to gauge the intelligibility. After thorough experimentation, it was found that the variance of the MFCC coefficients for each frame had smooth transitions for intelligible voices. However, considerable transients were noticed for un-intelligible voices. Therefore, some preliminary conclusions can be drawn regarding a possible correlation between the variance of the MFCCs and the intelligibility of a given voice. We used a summed derivative along the frames axis to eliminate a subset of all those transformations that were unintelligible.

6 Results

For 20 non-transformed utterances of 30 speakers, of which 23 are standard American newsreaders, the speaker ID system proved accurate 91.83% of the time. For de-identified transformed voices, the speaker ID system gave an accuracy of only 4.5%, and thus was sufficiently confused. Some preliminary results linking MFCC variance to intelligibility were also established.

Table 1: Accuracy of Speaker ID on non-transformed voices

Speakers	Accuracy of Speaker ID on non-transformed voices
23 standard American voices	90.43%
5 CMU Arctic voices	98%
2 native Indian speakers	92.5%
Total Accuracy	91.83%

Table 2: Accuracy of Speaker ID on transformed voices

Speakers	Accuracy of Speaker ID on transformed voices
5 CMU Arctic voices	4.72%
2 native Indian speakers	60%
Total Accuracy	12%

7 Future Work

Establishing concrete statistical methods to ascertain the intelligibility of a voice is still an area of active research, and the results correlating MFCC variance to intelligibility should be probed further. The next step in de-identification would be to maintain the distribution of ethnicities and gender for a given set of speakers while successfully de-identifying all of them. Finally, it would also be worthwhile to research the ability to establish lexicographic rules based on the speakers' word choices that might prevent his/her ethnicity from being identified.

8 Acknowledgements

- Prof. Alan W. Black for his ideas, suggestions and support.
- Esfandiar Zavarehei (Brunel University) for the MATLAB code for K-means with splitting clusters.
- Anil Alexander, Andrzej Drygajlo (Swiss Federal Institute of Technology, Laussane) for ‘Speaker Identification: A Demonstration using MATLAB’.
- Malcolm Slaney for his mfcc.m function from his Auditory Toolbox.

References

[1] Toda, T., Black, A., and Tokuda, K. (2007) "Voice Conversion Based on Maximum-Likelihood Estimation of Spectral Parameter Trajectory", *IEEE Transactions on Audio, Speech and Language Processing, 15(8),* pp 2222-2236.