Evaluation of Speaker Identification using GSM data

Lerato Lerato, Daniel Mashao  
Department of Electrical Engineering, University of Cape Town  
Rondebosch, Cape Town, South Africa  
llerato@crg.ee.uct.ac.za  
daniel@eng.uct.ac.za

Abstract: Speaker identification is a process of identifying speakers using their voices. Speaker identification systems do not perform well on noisy speech or on speech that has been compressed and decompressed. Speaker identification systems have successfully worked for clean speech. The current work relates to speech that has been compressed and decompressed through GSM network.

The GSM network compresses speech data in order to increase the available bandwidth, this, however, removes speech features that may be needed to achieve perfect identification of the speaker. In this paper we research the impact of GSM data on speaker identification system and report preliminary results. We also look at possible methods of improving the speaker identification.

1. INTRODUCTION

Speaker recognition and speech coding are two fields of speech technology. Speaker recognition can be divided into speaker identification and speaker verification. Speaker identification identifies the talker without any a priori claim of identity while the latter verifies the speaker’s claimed identity.

There are many applications that can benefit from speaker recognition systems. Some of these include carrying out banking transactions over the telecommunication network that transmits voice, building access by voice identification, forensic investigations applications and subscriber voice identification at call centers.

Most of the speaker recognition applications are likely to use telecommunication channels. The process of voice transmission over the communication network is done using packet technology which exhibits robustness and gains in bandwidth. In packet technology the analogue speech signal is converted into a digital signal and encoded for transmission. The digital signal is more resistant to noise than the original analogue signal and because of compression requires less bandwidth for transmission. Lossy compression such as GSM is used. This however means that some speech features are destroyed by the compression and decompression. In speaker recognition the destroyed features may be needed for better identification or verification.

GSM 06.10 coder from Technical University of Berlin [1] is used in this study. This coder simulates a communication network compression and decompression of speech signals. GSM 6.10 coder compresses the signal at the bit rate of about 13 kbps [2]. GSM coders exist in three different types namely full rate (FR), enhanced full rate (EFR) and half rate (HR). These are discussed in section 4. The signals resulting from these coders are used as the database for identification experiment in this work.

The communication channel links the user and the SID system. Clean speech database (TIMIT) is transcoded by GSM 06.10 to form GSM database (GSM DATA). The transcoding process represents the communication channel. The identification results from GSM DATA form the baseline product of this study.

This paper first describes the speaker identification (SID) system followed by the brief explanation of the TIMIT database. GSM coders are consequently tabled following which the baseline experiments and their corresponding results are discussed. Finally, conclusions are made and possible ways of improving the speaker identification system are briefly stated.

2. SPEAKER IDENTIFICATION (SID) SYSTEM

Speaker identification identifies who the speaker is from the unidentified utterance with reference to the models obtained from the training data [3]. Figure 1 illustrates generic design of the SID system. This system comprises training phase (speaker enrollment) and testing phase (identification processes). The speaker enrollment is the training process where speech features are extracted in order to form models for the identification process. The identification process of the SID system uses the same feature extraction method as that of training. The feature extraction block is usually known as the front-end and the classifier is normally called the back-end. The back-end has two modes: (a) the training mode which executes the speaker modeling procedures and generates the speaker models to form the models database; (b) the testing mode which matches the features of the talker to those that exist in the database of speaker models until the talker with highest logarithmic likelihood is determined. Figure 2 shows the organization of the back-end of the SID.

The authors would like to thank Telkom SA, Siemens, National Research Foundation (NRF) and The Department of Trade and Industry (DTI) for supporting this research project.
Figure 1: The speaker identification (SID) system

2.1 Front-end of the SID system

Features are extracted so that a model of the speaker can be generated. There are several feature extraction methods that are used in speaker identification systems. Linear predictive coding coefficients (LPCC), mel-frequency cepstral coefficients (MFCC) [13], and auditory-based methods such as EIH are examples of front-end feature extraction methods. The SID for this work uses, parameterised feature set (PFS) [12] which utilises MFCC for its speech feature extraction process.

2.2 Back-end of the SID system

The back-end is the section (see fig. 2) where the SID system computes the likelihood of the incoming features with the speaker models and then produces relevant results. Various classifiers can be used at the back-end for training and testing the speakers. Hidden Markov models (HMM’s), support vector machine (SVM) and Gaussian mixture models (GMMs) are some examples of back-end methods. GMM’s were used at the back-end of the SID for this study. A GMM is used to model the speakers who are being trained for the system. During identification, the identity of the unknown speaker is determined using the maximum logarithmic likelihood decision rule [2].

A speaker’s model $\lambda$ is a function of the speakers’ features $\mathbf{X} = \{\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$ such that $\lambda = \mathbf{r}(\mu, \sum, w)$ where $\mu$ is mean vector, $\sum$ is the variance of the feature belonging to the mean vector; and $w$ is the weight of probability of $\mu$ to represent the speakers feature vectors.

Given an utterance from an unknown speaker $\mathbf{\hat{X}} = \{\hat{x}_0, \hat{x}_1, \hat{x}_2, \ldots, \hat{x}_T\}$, the speakers models whose likelihood of having generated $\mathbf{\hat{X}}$ is the highest or maximum is identified as the person who spoke the utterance. This is done using maximum log likelihood ($L$) estimates when $k = \max_j \{L(\mathbf{\hat{X}} | \lambda_j)\}$ where $j = [1, N]$.

This SID system is implemented in software and speech data that need to be tested is then processed according to the research specifications. This work uses GSM transcoded data.

Figure 2: The SID processes in the back-end section

3. TIMIT DATABASE [4]

The TIMIT database has a total of 630 speakers, 438 of whom are male. The signal is at 16kHz sampling rate and 16 bit pulse code modulated (PCM). There are 10 utterances (sentences) for each speaker. Each utterance 3 seconds long on average. Eight sentences differ for each person while two sentences are the same for every speaker in the database. No conversational records were made in TIMIT; rather, all sentences were read.

4. GSM CODERS

GSM coders use compression and decompression algorithms on the speech signal transmitted over GSM network. GSM codec uses linear predictive coding (LPC) [5]. This is sensible because it reduces the bit rate of speech to be transmitted. LPC is essential since it mimics the vocal tract. GSM codec uses three types of coding algorithms which are full rate (FR), enhanced full rate (EFR) and half rate (HR). These coders use different compression and decompression techniques. Full rate uses regular pulse excitation - long term prediction (RPE-LTP) while EFR is run by algebraic code excited linear prediction (ACELP) and HR has CELP- vector sum excited linear prediction (CELP-VSELP) as its codec algorithm.

4.1 Full Rate (FR) coder

RPE-LTP is widely used in GSM networks and is one of the first coding algorithms to be implemented for GSM networks and also the most popular protocol in digital cellular phones [1]. GSM 6.10 RPE-LTP compresses and decompresses a speech signal of frames of 20ms. These are the frames of 160, 13-bit samples with a sampling frequency of 8kHz [6]. The bit rate is then about 13kbps. Linear predictive filters are used during the compression and decompression of the speech signal. These cause some
degradation of the voice signal which affects the performance of the speaker identification systems.

4.2. Half Rate (HR) and Enhanced Full Rate (EFR) coders

An HR coder is meant to deal with the ever-increasing demand of subscribers [7]. It uses a CELP-VSELP coder. It is a 5.6kbps coder and is similar to full rate under noise-free environment.

EFR codec uses algebraic code excited linear prediction (ACELP). Enhanced full rate is a new coding method and not all mobile phones are able to use it. It uses 12.2kbps for speech coding [7]. EFR is an improvement as compared to the FR and HR coders. It is meant to be used in the full rate channel. This evaluation study has not yet covered HR and EFR coders.

5. BASELINE EXPERIMENTS AND RESULTS

The SID system in this study performs efficiently when tested on clean speech as reflect by results from table 1. TIMIT speech is sent to the SID and the identification results are measured. This GSM transcoded database (GSM DATA) [2] is formed by sending all the TIMIT sentences through a GSM 6.10 coder. This coder requires an 8kbps sampled signal and so the 16 kHz sampled TIMIT speech is first down-sampled to 8kHz. GSM coder is applied to it forming the GSM database (GSM DATA). GSM DATA has all acoustic properties required by the SID but the sampling rate. The GSM DATA signal is then up-sampled (zero-padding) back to 16 kHz before it is processed for identification. This process is illustrated in Figure 3. Speaker identification is then carried out using the SID shown in figure 1 where GSM DATA now forms the SID system input.

<table>
<thead>
<tr>
<th>Number of talkers</th>
<th>gender</th>
<th>Identification (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>mixed</td>
<td>100</td>
</tr>
<tr>
<td>14</td>
<td>Female</td>
<td>100</td>
</tr>
<tr>
<td>24</td>
<td>Male</td>
<td>100</td>
</tr>
<tr>
<td>29</td>
<td>Mixed</td>
<td>100</td>
</tr>
<tr>
<td>38</td>
<td>Mixed</td>
<td>100</td>
</tr>
</tbody>
</table>

Table 1. Preliminary recognition results from the UCT SID.

The results shown in Table 1 were obtained from the process illustrated in Figure 2. The initial interest in these results was to observe the identification rate as the population of speakers is increased. Eight utterances were used to train the system while only two functioned as test sentences for each talker.

4.1 Preliminary results and discussions

Results shown in Table 1 were obtained from the process illustrated in Figure 2. The initial interest in these results was to observe the identification rate as the population of speakers is increased. Eight utterances were used to train the system while only two functioned as test sentences for each talker.

Figure 3. The process of changing TIMIT database into a GSM transcoded database (GSM DATA)

6. CONCLUSIONS

Preliminary results from section 4 reflect that GSM transcoded speech signal recognition is of low rate due to the degradation of speech during transmission. These baseline results form a foundation for future work on the improvement of the SID performance and robustness of communication network transmitted speech. The baseline experiments will still continue for testing the SID performance on accent, dialect and possibly language.

7. POSSIBLE WAYS FOR IMPROVEMENT OF THE SID PERFORMANCE

The baseline results found in this work leads to the second phase of this research. This work will now involve finding the best possible solution for improving the SID system performance on GSM coded speech. The baseline tests will continue on the other GSM coders (HR and EFR). Telephone speech database (NTIMIT) from NIST [8, 10] has been used by other researchers and may be tested as additional baseline experiment.

Once the UCT SID system performance on GSM data is fully known, several solutions are foreseen as ways to improve identification over the communication network.

These are a few:

- use of noise cancellation techniques;
- attempt to use inverse problem approaches to improve speaker recognition;
- explore more auditory-based feature extraction methods;
- use back-end methods (e.g. HMM, SVM, GMM) that other researchers in our group are already working on;
• combine some of the above methods; and
• continue searching for better and better improvement methods.

The first approach will be on noise cancellation techniques. This approach may also involve working at the front-end of the SID system. A compensation method will then be evaluated to find out if it improves the performance of the SID system. Yao K. et al.[11] have shown that the robustness of the MFCC feature extraction process can be improved by using noise compensation methods. Figure 4 shows the overall system with noise cancellation method included.

![Diagram of SID system with noise cancellation process](image)

Figure 4. The SID system with noise cancellation process

8. REFERENCES