Comparing CDMA and GSM speech coding on speaker identification performance

Chipo Masawi, Daniel Mashao
Department of Electrical Engineering, university of Cape Town
Rondebosch, Cape Town, South Africa
mswich001@mail.uct.ac.za
daniel@eng.uct.ac.za

Abstract:
This paper compares the impact of two mobile transmission networks on speech technology. The two networks are GSM which is popular mostly in Europe and Africa and the CDMA networks are popular in North America and some countries in Asia. These two speech codecs used in this paper are GSM Full-rate RPE-LTP codec and the CDMA QCELP codec. In the experiments carried out, clean speech from the TIMIT database was passed through the codecs to obtain two trans-coded (coded and decoded) databases. Speech from these databases was then applied on a speaker identification system. The effect of the two speech coders on speaker identification performance was then assessed and compared.

1. Introduction

Speaker recognition can be defined as the process of recognizing a speaker based on information gathered from their speech. It can be classified into speaker identification and speaker verification. Speaker identification determines who is talking from a set of known speakers. Verification, on the other hand, determines whether a speaker is really who they claim to be by providing a simple accept or reject answer [1]. Speaker recognition has numerous applications which range from voice dialing to telephone banking. With the rapid increase in the use of wireless communication systems that is being experienced today, many more of these applications will be utilized over mobile cellular networks.

As the demand for mobile cellular communications continues to grow, two standards for cellular communications, CDMA (Code Division Multiple Access) and GSM (Global System for Mobile Communications), have now emerged as major competitors in the mobile industry. Both GSM and CDMA networks utilize speech coding algorithms (i.e. RPE-LTP and QCELP respectively) to compress and decompress speech data for transmission. The purpose of these coding schemes is to minimize transmission bit rates and reduce the required bandwidth in digital voice communication systems. Representing a speech signal with a fewer number of bits, results in a loss some of its features [2], [3]. It has been shown that the speech coding process introduces some distortion to the speech signal when it is reconstructed at the receiver’s end[4]. This distortion significantly affects the performance of a speaker recognition system[4]. The mobile networks are designed with the aim of making a conversation between people possible. This can be achieved even when some information such as a speakers identity is removed from the speech signal. Many speaker recognition systems, however, depends on the ‘discarded’ information, that is lost through compression. This paper looks at the difference in the degradation of speaker identification performance between the two speech coders used in GSM and CDMA networks.

GSM systems utilize a technique known as RPE-LTP, Regular Pulse Excited - Long Term Prediction to compress and decompress a speech signal at a bit rate of 13kbps[5] for the full-rate version. The half-rate version can use an even lower bit rate but the speech quality suffers even more. The GSM 06.10 coder is freely obtained from the Technical University of Berlin[5] and it is the version used in the reported experiments. It is available through the toast program. This algorithm is a simulation of the full rate (GSM FR) speech coder used in actual GSM networks. The CDMA systems utilize the variable rate QCELP, Qualcomm Code Excited Linear Predictive, speech codecs. The QCELP coder used in this study is TIA IS-96 [6], a simulation of the QCELP 8kbps variable rate coder. Although there are now three more advanced variable rate coders for CDMA systems, this older technology was chosen because it is freely available.

The GSM and QCELP codecs that were used in this study are briefly described in sections 2 and 3 respectively. The SIT system used is discussed in section 4. SIT experimental work and results are presented in section 5.
2. GSM Speech Coder

There are four types of GSM speech-coding algorithms. These are referred to as: Full-Rate (FR), Half Rate (HR), Enhanced Full Rate (EFR), Adaptive Multi-Rate (AMR) GSM coders. The speech coder that is used for comparison in this study is the Full Rate GSM coder. The following section discusses the GSM FR codec which was used in this study.

2.1 GSM FR

The FR coder is a Regular Pulse Excited Long Term Predictor - Linear Predictive Coder (RPE-LTP) [5]. At the encoder, a speech signal is split into frames that are 20ms long. Each frame, which consists of 160 speech samples, becomes 260 bits of data. Since 260 bits are encoded every 20 milliseconds, the overall bit rate of the coder is 13kbps. At the decoder, the 260 encoded bits are mapped back to 160 reconstructed speech samples [5]. The reason 260 bits are used is that there is limited error correction that is included in the data.

3. CDMA speech coder

CDMA systems use 8 kbps and 13kbps QCELP-Qualcomm Code Excited Linear Prediction variable rate speech coders [6]. They are based on a speech coding technique known as CELP-Code Excited Linear Prediction. The 8kbps vocoder, which is used in this study, is an early QCELP version that was used in CDMA systems.

QCELP coders used in CDMA systems encode and decode speech signals at variable rates. The QCELP algorithm adjusts the encoded data rates based on the speech signal energy, background noise and other speech characteristics [7]. This leads to a decrease in the average data rate of the compressed speech while maintaining good voice quality.

The original speech is partitioned into 20 ms frames by the coder. Each frame is taken as an input to the encoder and a transmission packet of compressed data is produced as the output. Each speech frame is encoded at one of four rates: full, 1/2, 1/4, 1/8. These correspond to following rates: 9.6, 4.8, 2.4 and 1.2 kbps.

4. Speaker Identification (SID) system

Speaker identification is the process of identifying who spoke a particular utterance given a set of known utterances [1]. SID systems are composed of 2 distinct parts. These are shown in fig 1.

The feature extraction block is known as the front end. The classification block, where the testing and training phases occur, is known as the back-end. In the SID system used for this work, features are extracted from a speech signal using PFS (parameterized feature sets). Feature vectors extracted from a speech signal at the front ends are then passed to the back-end or classifier. At this stage, a model of the speaker is built and stored in a database during the training phase. During the testing phase, the unknown speaker’s features are matched with the reference models in the database and a decision is made to select the best match and identify (ID) the speaker.

6. Experimental Work and results

6.1 Experimental set up

Speech in the TIMIT database is sampled at 16 kHz [2] but the GSM and QCELP coders work on speech input sampled at 8 kHz [8]. Therefore, in order for the speech codec simulations to successfully code and decode speech, every utterance in TIMIT database was down-sampled to 8 kHz. This speech was then coded and decoded by the two speech coders to obtain two transcoded databases. These will be referred to as the GSM database and QCELP database. The speech database required for the SID system had to be sampled at 16 kHz [2] but the GSM and QCELP databases created contained speech sampled at 8 kHz. The speech files were therefore up-sampled (zero padding i.e. inserting zeros between every speech sample in a file and subsequent low pass filtering), to 16 kHz to form the GSM and QCELP databases that were appropriate for speaker identification tests.

Trans-coded speech, from the GSM and QCELP databases that were created, was applied as an input to the speaker identification system described in section 4. This process is illustrated in fig 2.
The speaker identification experiments were carried out using 38 speakers (24 males and 14 females) from the New England dialect (DR1) from the TIMIT database. The work carried out was to observe the effect on the SID performance using GSM and QCELP trans-coded speech on different population sizes. All speaker identification work was carried out under matched conditions. This means that both training and testing were carried out with speech from the two trans-coded databases. Eight utterances from each speaker were used for training while only 2 were used for testing.

6.2 Results and discussions

Table 1 below illustrates the effects of an increasing population on identification rates for both GSM and QCELP trans-coded data.

**Table 1:** Comparing speaker identification performance on different population sizes using GSM and QCELP trans-coded speech.

<table>
<thead>
<tr>
<th>Number of speakers</th>
<th>Gender</th>
<th>Identification (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>GSM Database</td>
</tr>
<tr>
<td>5</td>
<td>Female</td>
<td>100</td>
</tr>
<tr>
<td>14</td>
<td>Female</td>
<td>92.8</td>
</tr>
<tr>
<td>22</td>
<td>Mixed</td>
<td>90.9</td>
</tr>
<tr>
<td>24</td>
<td>Male</td>
<td>87.5</td>
</tr>
<tr>
<td>38</td>
<td>Mixed</td>
<td>92.1</td>
</tr>
</tbody>
</table>

The results shown in table 1 show that applying QCELP trans-coded speech as the input to a SID system produces slightly better identification performance than GSM trans-coded speech. This trend is also observed as the population size increases.

The difference in the identification rate between the two trans-coded databases varied between 0% (for 7 speakers) to the largest difference of 7.2% (for 14 speakers).

7. Conclusions and further work

The objective of this study was to compare the performance of GSM and QCELP coded speech applied to a speaker identification system. This was accomplished and the results from the experimental work presented in section 6, concluded that the identification rates of GSM trans-coded speech were on average slightly lower than those of QCELP speech.

Possible future work should include comparing the speaker recognition performance:

- On a larger population of speakers
- With the latest speech coders currently being utilised in CDMA and GSM systems.

8. References


186-336