A Study of Speech Coding Parameters in Speech Recognition

Jari Turunen¹ & Damjan Vlaj²

1) Tampere University of Technology, Pori, Pohjoisranta 11, P.O.Box 300, FIN-28101 Pori, Finland
2) University of Maribor, Faculty of electrical engineering and computer science, Smetanova 17, Maribor, SI-2000, Slovenia

jari.turunen@pori.tut.fi, damjan.vlaj@uni-mb.si

Abstract

Speech recognition over different transmission channels will set demands to the parametric encoded/decoded speech. The effects of different types of noise have been studied a lot and the effects of the parameterization process in speech has been known to cause degradation in decoded speech when compared to the original speech. But does the encoding/decoding process modify the speech so much that it will cause degradation in the speech recognition result? If it does what may cause the speech recognition degradation? We have studied the effect of the parameterization and the causes of the nine different codec configurations to isolated word recognition.

1. Introduction

Speech coding over different communication channels has gradually become established into everyday use. Voice over IP services, multimedia and videoconferencing applications, GSM and immense efforts towards UMTS are now reality. Due to the naturalness of the speech-based user interface the need and the possibility to make different automated services is increasing together with the availability of different communication channels [1][2][3]. It is also a known fact that the quality of speech will degrade when it is being encoded/decoded in several phases over the transmission channel. But how about other speech related services, for example speech recognition applications? They have successfully served over the POTS (Plain Old Telephone System) and ISDN (Integrated Services Digital Network) networks over decades, but are they suitable to serve the speech recognition demands over other data networks and speech that is coded within several coding schemes?

Speech recognition applications in handheld computers and part of portable phones will be easier to make in the future than nowadays due to increasing processing power and sophisticated technology. But still it is more efficient to process some of the speech processing tasks in some central server than locally and this causes demands on both the transmission channel and the quality of the codec. On the other hand there are still a few alternatives available to handle this situation, for example the handheld device can encode speech or extract the necessary features from the speech (or do both) and send the pre-processed stream to the central server [4].

The degradation of the speech recognition accuracy has been studied widely, especially in noisy environments and speaker-dependent variations, for example in [5][6]. The transmission channel and the influence of the codec itself affect the recognition result and an extensive study on this has been made in [7].

In our experiments we want to find out what features are important for speech recognition over a speech encoding/decoding process when the LPC vocal tract model is excluded. The simulation method and measurement techniques are described below.

2. Methods

The baseline speech recognition system was built on an HP-Unix system. The HTK tool [8] was used for the hidden Markov model construction. The HMM training was done on the whole word models. For feature vector definition, we used 12 Mel-cepstral coefficients and the addition of energy logarithm and delta and acceleration coefficients (MFCC_E_D_A). Thus, we obtained the feature vector of the size of 39 coefficients. The grammar was set in the way, that only one digit could be recognized at the time (isolated word recognition). Our experiments were made on the studio database TIDIGITS [9]. The TIDIGITS database consists of 16-bit linear speech samples with 16 kHz sampling frequency. We used isolated digits both for training and testing. In the training data part, we used 112 speakers and in the testing 113 speakers, who pronounced two tokens of each of the eleven digits. This means totally 2464 words in the training (men 1210 and women 1254) and 2486 words in the testing (men 1232 and women 1254).

The training was done only on clean data. The schematic diagram of the system is shown in Figure 1. For the testing of the effects of different codecs on speech recognition we tested nine different configurations, shown in Table 1, in simulated clean transmission channel environments, i.e. there was no additional noise in any tests.

Figure 1: Schematic diagram of the simulation system.
The G.729 is a delay codec that operates with 16 kbit/s bitrate [12]. It also has a possibility to post-filter the output speech. The G.728 is bit rate adaptive filtering coding method. The G.727 encoding/decoding scheme was chose three alternatives described in Table 1. The G.723 is dual rate speech codec operating at 5.3 and 6.4 kbit/s bitrates described in [10]. There is a built-in filtering mechanism in this codec on the decoder side, and we chose the best quality available for this codec in order to test high-defined to operate at 16, 24, 32 and 40 kbit/s [11]. We chose the segmental signal-to-noise (SNRSEG) ratio:

\[
\text{SNR}_{\text{SEG}} = \frac{1}{M} \sum_{j=0}^{M-1} \log_{10} \left[ \frac{\sum_{n=m_j-N+1}^{m_j} s(n)^2}{\sum_{n=m_j-N+1}^{m_j} [s(n) - \hat{s}(n)]^2} \right]
\]

(1)

where \(s(n)\) is the original speech sample at time \(n\), and \(\hat{s}(n)\) is the encoded/decoded speech sample at time \(n\). \(M\) is the number of segments, \(m_j\) is the end of the current segment and \(N\) is the segment length. The second objective measurement is the segmental root-mean-square error (RMSE) method:

\[
\text{RMSE}_{\text{SEG}} = \frac{1}{M} \sum_{j=0}^{M-1} \left[ \sum_{n=m_j-N+1}^{m_j} (s(n) - \hat{s}(n))^2 \right] / N
\]

(2)

where \(s(n)\) is the original speech sample at time \(n\), and \(\hat{s}(n)\) is the encoded/decoded speech sample at time \(n\). \(M\) is the number of segments, \(m_j\) is the end of the current segment and \(N\) is the segment length. For evaluating the objective spectral similarity the Itakura (ID) and Itakura-Saito distance (ISD) measures were used [15]:

\[
\text{ID}_{\text{SEG}} = \frac{1}{M} \sum_{j=0}^{M-1} \log \left( \beta^T (m_j) \tilde{R}_e(m_j) \beta(m_j) \right)
\]

(3)

\[
\text{ISD}_{\text{SEG}} = \frac{1}{M} \sum_{j=0}^{M-1} \left( \alpha^T (m_j) \tilde{R}_e(m_j) \alpha(m_j) \right)
\]

(4)

where \(\beta(m_j)\) is a linear prediction coefficient vector calculated from the encoded/decoded speech segment \(m_j\), and \(\tilde{R}_e(m_j)\) is its transpose, \(\alpha(m_j)\) is a linear prediction coefficient vector calculated from the original speech segment and \(\tilde{R}_e(m_j)\) is an autocorrelation matrix derived from the original speech segment. We used 10 linear prediction coefficients in the experiments.

The reason for using so many objective methods was to facilitate the making of conclusions. Another reason was that the SNR and RMS measuring methods do not make justice for all codecs that might be regarded as excellent when using for example the Mean Opinion Score (MOS) subjective evaluation method.

The test data was driven through different encoding/decoding schemes, described in Table 1, and then fed to the speech recognizer. In order to evaluate the speech recognition results several methods for encoded/decoded speech quality were used.

We used four different objective methods for evaluating the speech quality from different codec configurations. The first is segmental signal-to-noise (SNRSEG) ratio:

\[
\text{SNR}_{\text{SEG}} = \frac{1}{M} \sum_{j=0}^{M-1} \log_{10} \left[ \frac{\sum_{n=m_j-N+1}^{m_j} s(n)^2}{\sum_{n=m_j-N+1}^{m_j} [s(n) - \hat{s}(n)]^2} \right]
\]

The reason for using so many objective methods was to facilitate the making of conclusions. Another reason was that the SNR and RMS measuring methods do not make justice for all codecs that might be regarded as excellent when using for example the Mean Opinion Score (MOS) subjective evaluation method.

The segment size in all evaluation experiments was 256 samples. The segmental methods do not work very well in silence sections. The signal range varies from –32768 to 32767 (16 bit), so we experimentally put the silence threshold level to 64. The frame was regarded as silence frame if the energy of the frame was smaller than 256*64^2, otherwise the segment was accepted in the calculation. Some of the codecs will add a small delay to the decoded sound files. This phenomenon was overruled by sliding the processed sound file over the original one and simultaneously searching for the optimal place with the minimum MSE error criterion.

3. Results

The HMM recognizer was trained with TDIGITS clean speech, and in the testing phase the encoded/decoded test speech samples were fed to the recognizer. The results of the experiments are shown in Figure 2. For every codec configuration totally 2486 words, of which 1232 spoken by men and 1254 spoken by women, were encoded and decoded.
The baseline in Figure 2 means the baseline system tested with the clean training data, and the rest of the codec configurations are the same as in Table 1 respectively. In Table 2 the codec quality evaluation results are presented.

Table 2: The simulation results

<table>
<thead>
<tr>
<th>Codec</th>
<th>Offset (O)</th>
<th>SNR</th>
<th>RMS</th>
<th>ID</th>
<th>ISD</th>
</tr>
</thead>
<tbody>
<tr>
<td>G.723.1_53_NOPF</td>
<td>60</td>
<td>14.8</td>
<td>2498.6</td>
<td>0.04</td>
<td>0.12</td>
</tr>
<tr>
<td>G.723.1_53_PF</td>
<td>60</td>
<td>12.5</td>
<td>2973.0</td>
<td>0.05</td>
<td>0.14</td>
</tr>
<tr>
<td>G.723.1_NO_PF</td>
<td>60</td>
<td>16.3</td>
<td>2031.1</td>
<td>0.04</td>
<td>0.10</td>
</tr>
<tr>
<td>G.727</td>
<td>0</td>
<td>36.3</td>
<td>242.7</td>
<td>0.02</td>
<td>0.06</td>
</tr>
<tr>
<td>G.728_NOPF</td>
<td>1</td>
<td>1.4</td>
<td>9749.2</td>
<td>0.17</td>
<td>0.58</td>
</tr>
<tr>
<td>G.729</td>
<td>59</td>
<td>1.3</td>
<td>11058.2</td>
<td>0.15</td>
<td>0.79</td>
</tr>
<tr>
<td>GSM_6.10</td>
<td>0</td>
<td>11.0</td>
<td>2631.5</td>
<td>0.08</td>
<td>0.22</td>
</tr>
<tr>
<td>GSM.G.727</td>
<td>0</td>
<td>11.0</td>
<td>2644.1</td>
<td>0.08</td>
<td>0.22</td>
</tr>
<tr>
<td>GSM.G.727.GSM</td>
<td>0</td>
<td>8.5</td>
<td>3366.6</td>
<td>0.09</td>
<td>0.27</td>
</tr>
</tbody>
</table>

In Table 2, “O” (offset) means the average distance of the samples where a minimum distortion between the original and decoded speech samples was found. For example offset value 60 means that 60 samples have been taken out from the beginning of the encoded/decoded signal and then compared to the original signal. “SNR” means signal-to-noise ratio in decibels, “ID” means Itakura distance and “ISD” Itakura-Saito distance.

4. Conclusions

Our purpose was to find the features that may cause differences in speech recognition results. Several codecs were tested in the simulation. Most of the codecs belong to the ITU-T H.323 recommendation [16] which is the basis of modern videoconferencing systems. Most of them use a linear predictive coding parameter estimation (all pole model). This mechanism is almost the same between all codecs excluding the G.727 that does not use the all pole -model but the pole-zero –model.

The HMM-based speech recognizer uses the mel-cepstral coefficient based features, which adapt human characteristics in the speech recognition process. The cepstral coefficients have a relationship with the linear predictive coding coefficients, which model the spectrum envelope of the human vocal tract. It is important that the vocal tract is modeled and transmitted correctly in low-bit rate vocoders. In our case the recognition was not performed from the parameters, but it was first decoded and then recognized.

The encoding/decoding process emphasizes the effect of residual signal processing and the fact how accurately the residual signal is modeled, packed and quantized before transmission and how these values will affect the recognition result?

The results are good within isolated word recognition rate limits and these results cannot be compared to continuous speech recognition results. Although the overall recognition result is over 97 % it is interesting to see the relationship between different codec configurations and to find the effect of the different configurations.

When studying for example G.723.1 5.3 kbit/s with and without postfiltering, all the values in Table 2 and Figure 2 prove that the postfiltering loses some critical information, although it might be smoother to hear the postfiltered speech in a subjective evaluation scale. The postfiltering is performed to decoded speech and this effect is not caused by the encoding process.

Another comparison can be made with GSM6.10 and G.723.1 6.4 kbit/s codecs. The recognition rate is almost the same although the residual modeling approach is totally different. The results in Table 2 will defend the G.723.1 codec, because for example the Itakura and Itakura-Saito distances are almost half better for G.723.1 than for GSM6.10, although the bitrates are opposite. This means that both codecbook and data reduction methods for residual modeling will support an appropriate basis for transferring the necessary information to the recognition engine.

G.728 compared to G.727 is quite straightforward because both of the codecs will send only the residual signal of the filtered speech. Both codecs have separate local inverse filter estimators in both encoder and decoder ends. G.728 uses 50th order backward all-pole estimator and G.727 has 2nd order pole estimator and 6th order zero estimator. Both of G.728 and G.727 codecs are low-delay codecs, but G.727 uses almost 3 times more bits per second than G.728. This and the size of the vocal tract estimator (the bigger the estimator is the slower the changes are in it) may cause the ~2 % difference in the recognition result.

Another pair comparison, G.729 and G.729, emphasizes the significance of the vocal tract. Both of the codecs got the worst results measured with all metrics when compared to the other codecs. The G.729 got an even poorer result than G.729 but the recognition result is far better than the G.728 recognition result. This can be explained with the fact that G.729 send the linear prediction coefficients in the data stream and G.728 does not send. It is possible that the metrics we used do not sufficiently emphasize the features that the speech recognition engine uses, although the Itakura and Itakura-Saito distance metrics should be close to the metrics that the recognizer uses. On the other hand we are measuring the differences between the original and encoded/decoded signal while the recognizer is measuring the differences between the unknown signal features and averaged features of signals used in the training process combined with the statistical based decision logic.

The tandem coding shows the usual phenomenon, degradation, as expected. The speech recognition degradation is almost linearly related to the number of consecutively connected codecs.
When concluding the simulation, it seems that if the human vocal tract is modeled accurately in the encoding/decoding processes the speech recognition results are fairly good. The linear predictive parameters are not essential if the human vocal tract can be modeled in other ways, as in the case of the G.727 codec. The residual modeling and quantization is not so important, at least not in the case of the cepstral based recognizer.

When thinking about the multimedia communications and services over IP-networks that utilize speech interfaces directly, the low bit rate codecs can be used to transfer the critical information at least in isolated word recognition situations. The effect of noise is obvious and it cannot be forgotten.

It is possible that the processing power of handheld computers and portable phones etc. is nowadays sufficient to extract directly the necessary features from the speech. Unfortunately this may be too complicated to implement in practice on a large scale, so it seems to be easier to implement the service with certain decoders and a generic PCM type receiver, than to give separate feature extraction software to all users. Constantly evolving technology, standardization and engineering will give a solution to these problems in the future.

5. Acknowledgements

We want to thank professor Pekka Loula, from Tampere University of Technology, and the research group of professor Zdravko Kacic, from University of Maribor, for co-operation and valuable advice.

6. References


