Front-End Signal Processing for Speech Recognition

MILAN RAMLJAK¹, MAJA STELLA², MATKO ŠARIĆ²
¹Ericsson Nikola Tesla
Poljička 39, HR-21000 Split
²FESB - University of Split
R. Boškovića 32, HR-21000 Split
CROATIA
milan.ramljak@ericsson.com

Abstract: - The evolution of computer technology, including operating systems and applications, resulted in designing intelligent machines that can recognize the spoken word and find out its meaning. Different front-end models have specific processing time required for calculating the same number of coefficients used for pattern recognition. During the years, it has been significantly improved, not only thanks to improvements in algorithms, but also with more processing power of nowadays computers. In this paper we analyze processing time and reconstructed speech quality of the three common front-end methods (Linear Predictive Coding - LPC, Mel-Frequency Cepstrum - MFC, Perceptual Linear Prediction - PLP) for calculating coefficients. Reconstructed speech quality is measured with Perceptual Evaluation of Speech Quality (PESQ) score. It is visible from our analysis that, if required, higher number of coefficients could be used without significant impact on processing time for MFC and PLP coefficients.

Key-Words: - speech recognition, front-end, LPC, MFC, PLP, signal processing, PESQ

1 Introduction

Speech recognition technology is constantly advancing and is becoming more present in everyday life, e.g. applications for mobile phones, cars, information systems, smart homes [1-4].

One of the most important components of the speech recognition system is the front-end. The general point for speech is that the sounds generated by a human are filtered by the shape of the vocal tract. This shape determines what sound is created. If we define the shape accurately, this should give us an accurate representation of the phoneme being produced. The shape of the vocal tract manifests itself in the envelope of the short time power spectrum, which is the basis for speech signal recognition. Besides the spectral envelope representation of speech signals, in practice there is a wide range of other possibilities like energy, level crossing rates, and zero crossing rates.

Many recognition systems use different models for front-end processing, like linear predictive coding (LPC) [5], Mel Frequency Cepstral Coefficients (MFCC) [6], Perceptual Linear Prediction (PLP) [7] and other. Besides speech recognition, predictive coding technologies are also an important factor in image processing systems. LPC provides a good approximation to the vocal tract spectral envelope creating simple relation between speech sound and spectral characteristics. The model is mathematically very precisely defined opening good perspective for implementation in hardware and software. This is especially important in case of real time signal processing, because of decrease of required processing and computational operations. The MFC model is successor of LPC and is the most widely used front-end in automated speech recognition. It employs auditory features like variable bandwidth filter bank and magnitude compression. Perceptual linear prediction, similar to LPC analysis, is based on the short-term spectrum of speech. In contrast to pure linear predictive analysis of speech, perceptual linear prediction (PLP) modifies the short-term spectrum of the speech by several psychoacoustical transformations in order to model a human auditory system more closely [8].

In this paper we analyze processing time and reconstructed speech quality of the three common front-end methods (LPC, MFC, PLP) for calculating coefficients. Reconstructed speech quality is measured with Perceptual Evaluation of Speech Quality (PESQ) score [9]. Although modern computers are continuously becoming faster, this computational time is still not negligible, as our results present, and the choice of the coefficients, besides in terms of recognition performance is also essential in terms of computational complexity, especially given the large number of users for
server-oriented speech recognition (as used today by major companies in its products – Google Now, Apple Siri, Nuance Communications Dragon Naturally speaking server).

The paper is organized as follows. In Section 2, common speech recognition front-ends are shortly described. In Section 3, we present their characteristics in terms of computational time and reconstructed speech quality. The paper is concluded with Section 4.

2 Common front-ends
Speech recognition front-ends are designed to transform speech signal into the appropriate characteristics that serve as input parameters for the recognition algorithm. These characteristics should emphasize the essential characteristics of speech signal, which are relatively independent of the speaker and channel conditions.

2.1 LPC
The basic idea behind the LPC model is that a speech sample at time \( n \), \( s(n) \), can be approximated as a linear combination of the past \( p \) speech samples:

\[
s(n) \approx a_1 s(n-1) + a_2 s(n-2) + \ldots + a_p s(n-p) \tag{1}
\]

where the coefficients \( a_1, a_2, \ldots, a_p \) are assumed constant in analysis frame. This assumption is good for most speech signals with low resonance. LPC analysis produces \( N \) complex poles, where \( N \) is the order of predictor. This would imply that specific order of predictor should be defined according to specific application of this signal analysis method.

\[\begin{align*}
\text{u}(n) & \xrightarrow{X} \text{H}(z) \xrightarrow{S(n)} \\
\end{align*}\]

Fig. 1. Linear predictive coding diagram – adapted from [10]

\[
H(z) = \frac{S(z)}{G(z)} = \frac{1}{1 - \sum_{i=1}^{p} a_i z^{-i}} A(z) \tag{2}
\]

Fig. 1. represents the LPC model, defined with equation 2. Many formulations in definition of appropriate \( p \)th order LPC model depend on mathematical convenience, statistical properties and stability of system.

Mathematical convenience is most important for LPC implementation in software and hardware automated recognition systems. This is narrowly connected with required mathematical operations for LPC calculations. Whole process should be parameterized and calibrated as much as possible to perform effective speech analysis. LPC with higher order can describe speech signal much better, but requires additional mathematical operations.

2.2 MFC
Mel Frequency Cepstral Coefficients (MFCCs) are a feature widely used in automatic speech and speaker recognition. The difference between the cepstrum and the mel-frequency cepstrum is that in the MFC, the frequency bands are equally spaced on the mel scale, which approximates the human auditory system's response more closely than the linearly-spaced frequency bands used in the normal cepstrum. MFCC values are not very robust in the presence of additive noise, and so it is common to normalize their values in speech recognition systems to lessen the influence of noise.

MFCCs are commonly derived in following steps: [11]
1. Take the Fourier transform of (a windowed excerpt of) a signal.
2. Map the powers of the spectrum obtained above onto the mel scale, using triangular overlapping windows.
3. Take the logs of the powers at each of the mel frequencies.
4. Take the discrete cosine transform of the list of mel log powers, as if it were a signal.
5. The MFCCs are the amplitudes of the resulting spectrum.

There can be variations on this process, for example, differences in the shape or spacing of the windows used to map the scale [12]. MFC coefficients are very popular and are the most widely used coefficients today, in the early 2000s ETSI defined a standardized MFCC algorithm to be used in mobile phones [13].

2.3 PLP
PLP coefficients are designed to model a human auditory system mode closely than previous methods. The calculation is similar to MFC coefficients, but they use critical bands, equal loudness curve and intensity-loudness power law. Calculation is performed in the following steps [14]:
- computation of short-term speech spectrum
- nonlinear frequency transformation and critical-band spectral resolution
- critical bands adjustments to the curves of equal loudness
- weighted spectral summation of power spectrum samples
- enforcing the intensity-loudness power-law
- all-pole spectrum approximation
- transformation of the PLP-coefficients to the PLP-cepstral representation.

3 Front-end processing characteristics and results

The experiments were performed with the signal processing tool written in Matlab. The calculation of processing time required to obtain specific number of coefficients for these three methods is measured in Matlab with core i5 processor and 512 MB of RAM.

First, LPC spectrogram example is shown for the same input signal and different LPC orders.

<table>
<thead>
<tr>
<th>LPC order</th>
<th>Processing time</th>
<th>PESQ MOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>5th</td>
<td>1.336002 s</td>
<td>1.212</td>
</tr>
<tr>
<td>13th</td>
<td>1.535249 s</td>
<td>2.347</td>
</tr>
<tr>
<td>40th</td>
<td>3.336561 s</td>
<td>2.451</td>
</tr>
</tbody>
</table>

Table 1. LPC processing time and reconstructed speech quality

By increasing the order of LPC, more information from spectral diagrams can be obtained, but on the other hand it requires more processing time. However, this does not necessarily mean better recognition performance and 13 coefficients are most commonly used in speech recognition. Table 1. shows real processing times and PESQ scores measured for three different orders of LPC. For input speech signal with duration of 2.5 s, and using 13<sup>th</sup> order of LPC measured time is 1.535 s which implies that for each second of input signal, approximately 600ms of processing time is required. It should also be noted that this is only one part of speech recognition system, additional processing time is also required for back-end processing.

Fig. 2. Spectrograms for different LPC orders

Fig. 2 shows three spectral diagrams of input speech signal. First part visualizes original spectrogram of the spoken test sequence. Graph can be easily used to identify spoken words phonetically but for automatic speech recognition purposes some user specific voice characteristics should be decreased. To accomplish this, the order of LPC can be reduced. We experimented with few different orders of LPC, measuring required computation time.

<table>
<thead>
<tr>
<th>MFC order</th>
<th>Processing time</th>
<th>PESQ MOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>5th</td>
<td>0.458499 s</td>
<td>2.342</td>
</tr>
<tr>
<td>13th</td>
<td>0.495099 s</td>
<td>2.344</td>
</tr>
<tr>
<td>30th</td>
<td>0.536869 s</td>
<td>2.412</td>
</tr>
</tbody>
</table>

Table 2 – MFC processing time and reconstructed speech quality

Increasing the number of coefficients used in MFC spectrogram analysis, more precise representation of spoken words is obtained.
According to Table 2, it is easy to see that the processing time required for MFC coefficients calculation is not so different in relation to its number. This is one of the desirable features of this speech analysis method.

For the case of PLP analysis, spectrograms of the same test sentence are given in Fig. 4, and processing time and PESQ scores are given in Table 3. Results are similar to MFC with slightly higher processing times and PESQ scores.

<table>
<thead>
<tr>
<th>PLP order</th>
<th>Processing time</th>
<th>PESQ MOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>5th</td>
<td>0.522713 s</td>
<td>2.102</td>
</tr>
<tr>
<td>13th</td>
<td>0.617283 s</td>
<td>2.434</td>
</tr>
<tr>
<td>20th</td>
<td>0.632076 s</td>
<td>2.503</td>
</tr>
</tbody>
</table>

Table 3 – PLP processing time and reconstructed speech quality

Fig. 4. Spectrograms for different orders of PLP

Comparing all the results it can be easily seen that processing time for calculating the coefficients is varying according to order of the model, but the influence of the order is much lower for MFC and PLP coefficients. The intelligibility of reconstructed speech, calculated with perceptual evaluation of speech signal (PESQ), is increasing with the order of the model, but the processing time is higher, especially for LPC coefficients. This ratio is always important, especially if the speech recognition system has limited processing capabilities. Another interesting thing that we can find out from measurement tables is that in more complex front-ends such as PLP where some knowledge from psychoacoustics is applied to make it closer to human speech recognition, we can obtain better reconstructed speech quality values.

4 Conclusion

Speech recognition technology is constantly advancing and is becoming more present in everyday life. During the years, it has been significantly improved, not only thanks to improvements in algorithms, but also with more processing power of nowadays computers. In this paper we analyzed processing time and reconstructed speech quality of the three common front-end methods (LPC, MFC, PLP) for calculating the coefficients. As a measure of reconstructed speech quality, PESQ score was used. For PLP coefficients, where some knowledge from psychoacoustics is applied to make it closer to human speech recognition, our results showed that better reconstructed speech quality values can be obtained. Our analysis also showed that, if required, higher number of coefficients could be used without significant impact on processing time for MFC and PLP coefficients.

Although modern computers are continuously becoming faster, this computational time is still not negligible, and the choice of the coefficients, besides in terms of recognition performance is also essential in terms of computational complexity, especially given the large number of users for server-oriented speech recognition (as used today by major companies in its products – Google Now, Apple Siri, Nuance Communications Dragon Naturally speaking server).

References:


