Line Cepstral Quefrencies and Their Use for Acoustic Inventory Coding

Guntram Strecha, Matthias Eichner, Rüdiger Hoffmann

Laboratory of Acoustics and Speech Communication
Technische Universität Dresden, Germany
{guntram.strecha, matthias.eichner, ruediger.hoffmann}@ias.et.tu-dresden.de

Abstract

Line spectral frequencies (LSF) are widely used in the field of speech coding. Due to its properties, the LSF are qualified for the quantisation and the efficient compression of speech signals. In this paper we introduce the line cepstral quefrencies (LCQ). They are derived from the cepstrum in the same manner as the LSF are derived from linear predictive coding (LPC) features. We show that the combination of the pole-zero transfer function of the cepstrum with the properties of LSF offers advantages for speech coding. We apply the LCQ features to compress an acoustic inventory, which is used for a low resource speech synthesis. It is shown that the compression performance of the LCQ features is better than those of the LSF features in terms of the mean spectral distance to the original inventory.

Index Terms: low resource speech synthesis, inventory coding

1. Introduction

Line spectral frequencies are widely used features for the coding of speech signals. They have been shown to perform well in the quantisation and the compression of speech signals [1, 2]. However, LSF are based on linear predictor coding with an all-pole transfer function that models only formant peaks. Missing spectral information is shifted to the excitation signal and has to be coded carefully together with the phase information, which is related to the pitch information. For example, the Adaptive Multi-Rate Wideband (AMR-WB) codec uses about ninety percent of the total bit allocation for excitation parameters in the 23.85 kbit/s mode and still seventy percent in the 6.6 kbit/s mode [2].

In contrast to the LPC analysis the transfer function of the cepstrum models poles and zeros. They correspond to formant peaks and valleys. Furthermore, the cepstral model approximates the spectral envelope more accurate in the low order analysis. Because of these advantages we suggest to model LSF not based on LPC but on the cepstrum.

In the generalised model of LPC and cepstrum analysis described in [3], the authors apply the LSF transformation to the mel-generalised cepstrum coefficient. However, if the calculation of LPC and cepstrum is considered separately as shown in Fig. 1, we can refer to the result of the LSF transformation applied to the cepstrum as line cepstral quefrencies (LCQ).

The goal of our investigation is to reduce the memory consumption for low resource speech synthesis by compression of the acoustic inventory. Conventional speech codecs, as the AMR-NB/WB, effectively reduce the inventory size and offer efficient techniques for integrating acoustic synthesis into the codec as well as changing voice characteristics [4]. However, these code excited linear prediction (CELP) based codecs require a lot of memory to store the excitation signal parameters. If the excitation parameters are not transmitted to the decoder, the size of the coded acoustic inventory decreases noticeably. The best quality is obtained by shifting as much spectral information as possible from the excitation signal to the spectral model. As the pitch information of the speech signal is part of the (lost) excitation signal, the pitch marks have to be transmitted separately. In either case, i.e. independently of the used codec, the pitch marks are stored in the acoustic inventory. Therefore, the size of the inventory is not affected.

Section 2 gives a brief review of the LSF calculation, the calculation of the LCQ and the properties of both. Section 3.1 and 3.2 describes the coding and decoding of an acoustic inventory.

The results of the evaluation by means of measuring spectral distance are given in Sec. 3.3.

2. Properties of LSF and LCQ

2.1. Description of LSF

The transfer function of an all-pole digital filter, derived from a linear predictive analysis, is given by

\[ H(z) = \frac{1}{A_p(z)} = \frac{1}{1 + \sum_{k=1}^{p} a_k z^{-k}} \]  

If the filter is stable, i.e. the roots of \( A_p(z) \) located inside the unit circle, the LSF can be derived by decomposing \( A(z) \) to a set of two transfer functions, having even (\( Q'_{p+1} \)) and odd (\( P'_{p+1} \)) symmetry

\[ P'_{p+1}(z) = A_p(z) - z^{-(p+1)} A_p(z^{-1}) \]  
\[ Q'_{p+1}(z) = A_p(z) + z^{-(p+1)} A_p(z^{-1}) \]
Removing the two known roots at ±1 the order of \( P' \) and \( Q' \) can be reduced, i.e.

\[
P_p(z) = \frac{P'(z)}{1 - z} \quad (4)
\]

\[
Q_p(z) = \frac{Q'(z)}{1 - z} \quad (5)
\]

The LSF are the \( p/2 \) angles of the roots of \( P_p(z) \) and \( Q_p(z) \), which are located in the range \( 0 \leq \omega_i \leq \pi \). Following [1] the LSF have the following properties:

- All roots of \( P_p(z) \) and \( Q_p(z) \) lie on the unit circle.
- The roots of \( Q_p(z) \) and \( P_p(z) \) alternate each other, so that \( 0 \leq \omega_0^Q < \omega_1^P < \omega_1^Q < \omega_2^P < \cdots \leq \pi \) with \( i = 0 \ldots p/2 \).

The distributions plots of LSF parameters from the analysis (order \( p = 5 \)) of the acoustic inventory (see section 3) are shown in Fig. 2.

![Distribution plots of LSF parameters of the acoustic inventory, described in Sec. 3, for analysis order \( p = 5 \).](image)

2.2. Description of LCQ

The complex cepstrum corresponding to \( s(n) \) is defined to be the stable sequence \( \hat{c}(n) \) whose \( z \)-transform is

\[
\hat{C}(z) = \log S(z) = \sum_{n=-\infty}^{\infty} \hat{c}(n)z^{-n} \quad (6)
\]

where \( S(z) \) is the \( z \)-transform of \( s(n) \). If the region of convergence of \( \log S(z) \) include the unit circle, then \( \hat{c}(n) \) is stable.

There exists various methods to get an estimation \( \hat{C}(z) \) of \( C(z) \). A common way is the calculation via the real cepstrum \( c_r(n) \) using a window function \( l(n) \)

\[
c(n) = c_r(n)l(n) \quad (7)
\]

\[
C_r(z) = \log |S(z)| = \sum_{n=-\infty}^{\infty} c_r(n)z^{-n} \quad (8)
\]

\[
l(n) = \begin{cases} 0 & n < 0 \\ 1 & n = 0 \\ 2 & 0 < n \leq p, \end{cases} \quad (9)
\]

where \( p \) is the order of \( c(n) \). According to [5] the transfer function

\[
H(z) = e^C_p(z) = e^{c_r(0)+2\sum_{n=1}^{p} c_r(n)z^{-n}} \quad (10)
\]

describes a minimum-phase system, due to the causality of \( c(n) \). If \( c(n) \) is a stable sequence, then the transformation of Eqs. (2-5) can be applied to \( C_p(z) \).

\[
P'_{p+1}(z) = C_p(z) - z^{-(p+1)}C_p(z^{-1}) \quad (11)
\]

\[
Q'_{p+1}(z) = C_p(z) + z^{-(p+1)}C_p(z^{-1}) \quad (12)
\]

\[
P_p(z) = \frac{P'_{p+1}(z)}{1 - z} \quad (13)
\]

\[
Q_p(z) = \frac{Q'_{p+1}(z)}{1 - z} \quad (14)
\]

Respectively, we denote the angles of the \( p/2 \) roots of \( P_p(z) \) and \( Q_p(z) \), which are located in the range \( 0 \leq \omega_i \leq \pi \), as line cepstral quefrencies (LCQ). If \( z \) of Eq. (8) is replaced by the all-pass system of Eq. (16) a transform of the frequency axis is performed and \( c(n) \) is called mel-cepstrum. The properties of LCQ are the same as that of LSF outlined above. A comparison of the distribution plots of LSF parameters from the cepstrum and mel-cepstrum (\( \lambda = 0.3 \)) is shown in Fig. 3.

![Distribution plots of LCQ parameters of the acoustic inventory, described in Sec. 3, calculated from cepstrum (\( \lambda = 0 \)) and mel-cepstrum (\( \lambda = 0.3 \)) for analysis order \( p = 5 \).](image)

3. Acoustic inventory coding using LCQ

To show the suitability of the LCQ parameters an acoustic inventory was coded and afterwards decoded. We measure the loss of quality due to the compression by analysing the spectral difference to the original inventory. The inventory consists of 1176 German diphones spoken by a female person. At a sampling rate of 16 kHz and an accuracy of 16 bit the size of the PCM inventory is 4.4 Mbyte.

2874
3.1. Inventory coding

Similar to the AMR encoding procedure outlined in [6] the speech segments of the diphone inventory are processed separately. Each of the segments \( j \) is split into \( K_j \) frames consisting of \( I_k \) speech periods. The distances \( I_R \) of consecutive frames are multiples of the current pitch period. The cepstral coefficients \( c_\lambda(n) \) are calculated from the windowed \( k \)-th frame using the unbiased estimator of log spectrum (UELS) [7]. Especially for a sampling rate of 32 kHz the important spectral properties are located at lower frequencies. To get higher accuracy at this frequency range a spectral warping can be applied. Therefore a filter is used, which realizes a transformation of the frequency axis according to warping factor \( \lambda \). In case of \( \lambda = 0.45 \) the frequency transformation approximates the mel-scale, for \( \lambda = 0.6 \) the Bark-scale (sampling rate 16 kHz).

\[
\tilde{C}_k(\tilde{z}) = \sum_{n=0}^{N_0-1} \tilde{c}_k(n) \tilde{z}^{-n} \tag{15}
\]

with \( \tilde{z}^{-1} = \frac{z^{-1} - \lambda}{1 - \lambda z^{-1}} \) and \( N_0 \leq N \). \( \tag{16} \)

The LCQ transformation is applied to \( \tilde{C}_k \). A non-uniform scalar quantisation of the LCQ parameters of the whole inventory, obtained via the LBG training algorithm [8], is used for inventory compression.

3.2. Inventory decoding

Once the inventory is coded offline, the acoustic synthesis of a TTS system has to decode the required segments of the inventory at run-time. The LCQ parameters were back-transformed to the (mel-)icpstrum and a cepstral synthesis filter generates the speech signal using an excitation signal for the input. The excitation signal of voiced sections differs from the one of unvoiced sections mainly at the phases \( \varphi_n \) of the \( n \)-th pitch period \( f_n \). Setting \( \varphi_n = 0 \), \( \forall n \) a pulse train occurs, assigning equally distributed random values \( 0 \leq \varphi_n \leq 2 \pi \) produces white noise. Exciting the minimum phase system with a pulse train, corresponding to the target \( F_0 \) contour, results in short responses, depending on the number \( N_0 \) of cepstral coefficients. The energy is concentrated at the beginning of each synthesised period. This results in a strong voiced synthetic speech signal. The level of voicing could be seen as the randomness of phases \( \varphi_n \) between the pitch periods. Enhancing the excitation signal at voiced parts with noise at higher frequencies reduces the graviely of the voice. For each voiced period of the excitation signal the enhancement is applied to the phases \( \varphi_n \):

\[
\varphi_n = \begin{cases} 
0 & n = 0 \\
\text{rand}(r(n)) & 0 < n \leq \frac{N}{2} \\
-\varphi_{n-1} & \frac{N}{2} < n \leq N 
\end{cases} \tag{17}
\]

\[
r(n) = \frac{2 \pi}{1 + e^{-\alpha \left( \frac{p}{N_\text{p}} - 1 \right)}} , 
0 < n \leq \frac{N}{2}, \tag{18}
\]

where \( N \) is the order of the inverse Discrete Fourier Transformation (equals to the length of the excitation period) and \( \text{rand} \) generates uniform distributed random values in the range \((0, r(n))\). Examples of \( r(n) \) with different values of \( \alpha \) and \( \beta \) are plotted in Fig. 4.

With Eq. (17) and (18) the excitation signal \( e(n) \) is calculated as:

\[
e(n) = F^{-1} \left\{ \tilde{e}(\varphi_n) \right\} , 
0 \leq n \leq N = N_\text{p}, \tag{19}
\]

Figure 4: Function \( r(n) \), which declares the range of the uniform distributed random values \( \varphi_n \).

with \( N_\text{p} \) the length of excitation period given from the target \( F_0 \) contour. Voiceless periods are generated in the same way, setting \( \alpha \) to a high and \( \beta \) to a negative value. To reduce the calculation time one can set \( N \) to a multiple of the power of 2 and use the inverse Fast Fourier Transformation (FFT). Accordingly, the excitation signal has to be truncated respective zero-padded to fit the target period length \( N_\text{p} \).

3.3. Evaluation

An objective measurement in time domain is not reasonable, because of the lost original phases of the pitch periods. For the objective evaluation of the proposed coding method we use the mean spectral distance [1, pp. 102ff] between the cepstral smoothed spectrum of the original PCM inventories and the LCQ coded/decoded inventories (s. Tab. 1 and 2).

\[
\text{SD} = \frac{1}{M} \sum_{m=1}^{M} \text{sd}_m \tag{20}
\]

\[
\text{sd}_m^2 = \frac{2}{N} \sum_{n=0}^{N/2-1} \left| \log S_m(\omega_n) - \log \tilde{S}_m(\omega_n) \right|^2 \tag{21}
\]

Since the pitch marks of the inventory are known, the analysis frame consists of two pitch periods and consecutive frames were shifted by one period. The number of cepstral coefficients, used

Table 1: Mean spectral distances SD of 16 kHz-LCQ coded inventories at various values for the LCQ order \( p \), the warping factor \( \lambda \), the frame length \( I_\text{f} \) (in periods), the frame rate \( I_R \) (in periods), the no. of bits used for quantised LCQ parameters and the size of the inventory in kbyte and in % of the uncoded inventory.

<table>
<thead>
<tr>
<th>( p )</th>
<th>( \lambda )</th>
<th>( I_\text{f} )</th>
<th>( I_R )</th>
<th>bits / update</th>
<th>inventory size / kbyte</th>
<th>%</th>
<th>SD dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>40</td>
<td>0.15</td>
<td>2</td>
<td>2</td>
<td>202</td>
<td>390</td>
<td>8.59</td>
<td>5.52</td>
</tr>
<tr>
<td>40</td>
<td>0.15</td>
<td>2</td>
<td>2</td>
<td>162</td>
<td>317</td>
<td>6.99</td>
<td>5.58</td>
</tr>
<tr>
<td>34</td>
<td>0.15</td>
<td>2</td>
<td>2</td>
<td>138</td>
<td>276</td>
<td>6.08</td>
<td>5.70</td>
</tr>
<tr>
<td>34</td>
<td>0.15</td>
<td>2</td>
<td>2</td>
<td>104</td>
<td>217</td>
<td>4.77</td>
<td>5.90</td>
</tr>
<tr>
<td>34</td>
<td>0.15</td>
<td>3</td>
<td>3</td>
<td>154</td>
<td>316</td>
<td>3.43</td>
<td>6.22</td>
</tr>
<tr>
<td>28</td>
<td>0.15</td>
<td>3</td>
<td>3</td>
<td>86</td>
<td>135</td>
<td>2.97</td>
<td>6.37</td>
</tr>
<tr>
<td>28</td>
<td>0.15</td>
<td>4</td>
<td>4</td>
<td>86</td>
<td>109</td>
<td>2.41</td>
<td>6.72</td>
</tr>
<tr>
<td>22</td>
<td>0.15</td>
<td>4</td>
<td>4</td>
<td>68</td>
<td>95</td>
<td>2.09</td>
<td>7.14</td>
</tr>
<tr>
<td>28</td>
<td>0.15</td>
<td>4</td>
<td>4</td>
<td>58</td>
<td>86</td>
<td>1.89</td>
<td>7.42</td>
</tr>
<tr>
<td>22</td>
<td>0.15</td>
<td>4</td>
<td>4</td>
<td>46</td>
<td>76</td>
<td>1.68</td>
<td>7.82</td>
</tr>
</tbody>
</table>
for cepstral smoothing, were set to 40 (16 kHz) and 20 (8 kHz), according to half of the mean base frequency of the speaker. The sizes of the compressed inventories depend on the number of bits used for the quantised LCQ parameters and the update rate, which equals to the frame rate $I_R$. The mean spectral distances of the 16 kHz-LCQ coded inventory were compared to that of the LSF, AMR-WB and SPEEX coded/decoded inventories (s. Fig. 5). The AMR-WB restores frequencies only up to 7 kHz. For a fair comparison we reduced the frequency range, from which the spectral distances were measured, to 7 kHz. This is done by setting $N$ in Eq. (21) to $7/8$ of the FFT order. The LSF coding of the inventory follows the same procedure as that of the LCQ coding. The LSF coefficients are derived from the LPC coefficients, which were calculated from the windowed analysis frames by the Levinson-Durbin recursion. As Fig. 5 shows, the use of LCQ features for inventory coding performs better for similar compression rates than the use of LSF features. The comparison between the mean spectral distances of the

8 kHz LCQ, LSF, AMR-NB and SPEEX coded/decoded inventories are plotted in Fig. 6. The results of the tested parameter sets for the proposed coding scheme are summarised in Tab. 2.

### Table 2: Mean spectral distances SD of 8 kHz-LCQ coded inventories at various parameter combinations (s. Tab. 1).

<table>
<thead>
<tr>
<th>$P$</th>
<th>$\lambda$</th>
<th>$I_N$</th>
<th>$I_R$</th>
<th>bits / update</th>
<th>inventory size kbyte</th>
<th>%</th>
<th>SD dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.0</td>
<td>2</td>
<td>2</td>
<td>102</td>
<td>216</td>
<td>9.44</td>
<td>5.49</td>
</tr>
<tr>
<td>20</td>
<td>0.0</td>
<td>2</td>
<td>2</td>
<td>82</td>
<td>180</td>
<td>7.84</td>
<td>5.54</td>
</tr>
<tr>
<td>20</td>
<td>0.0</td>
<td>2</td>
<td>2</td>
<td>62</td>
<td>144</td>
<td>6.30</td>
<td>5.75</td>
</tr>
<tr>
<td>17</td>
<td>0.0</td>
<td>2</td>
<td>2</td>
<td>54</td>
<td>131</td>
<td>5.70</td>
<td>5.87</td>
</tr>
<tr>
<td>17</td>
<td>0.0</td>
<td>3</td>
<td>3</td>
<td>54</td>
<td>99</td>
<td>4.32</td>
<td>6.24</td>
</tr>
<tr>
<td>17</td>
<td>0.0</td>
<td>4</td>
<td>4</td>
<td>62</td>
<td>90</td>
<td>3.92</td>
<td>6.44</td>
</tr>
<tr>
<td>17</td>
<td>0.0</td>
<td>3</td>
<td>3</td>
<td>36</td>
<td>79</td>
<td>3.43</td>
<td>6.72</td>
</tr>
<tr>
<td>17</td>
<td>0.0</td>
<td>4</td>
<td>4</td>
<td>36</td>
<td>68</td>
<td>2.97</td>
<td>7.11</td>
</tr>
<tr>
<td>14</td>
<td>0.0</td>
<td>4</td>
<td>4</td>
<td>30</td>
<td>63</td>
<td>2.75</td>
<td>7.70</td>
</tr>
<tr>
<td>11</td>
<td>0.0</td>
<td>4</td>
<td>4</td>
<td>24</td>
<td>58</td>
<td>2.55</td>
<td>8.17</td>
</tr>
</tbody>
</table>

4. Conclusion

We introduced the line cepstral frequencies that are derived from the (mel-)cepstrum in the same manner as the line spectral frequencies from the LPC coefficients. The compression performance of the LCQ parameters was studied by coding an acoustic inventory, that is used in the acoustic part of a TTS system. Applying the proposed coding technique shows better performance of LCQs compared to the LSF parameters in terms of the spectral distance measure. Subjective evaluations are planned in the future to test whether the achieved reduction of spectral distances do result in better speech quality or not.

5. References