USING NOISE REDUCTION IN MODE SELECTION AND PITCH SEARCH

Lasse Laaksonen and Anssi Rämö

Media Laboratory, Nokia Research Center, Tampere, Finland
lasse.j.laaksonen@nokia.com, anssi.ramo@nokia.com

Abstract
We present a novel method of exploiting noise reduction prior to mode selection in classification-based speech coding. Certain parameter estimation and mode selection is performed on a denoised signal to ideally remove the non-speech components before signal classification, although the original input is coded. This method is employed in a recent ITU-T standard, G.718, which provides state-of-the-art performance for narrowband and wideband speech between 8 and 32 kbps. We also present the pitch tracking algorithm of G.718. Open-loop pitch lags are estimated from the denoised signal.

1. Introduction
ITU-T Study Group 16 started an initial study of variable bit rate speech coding in 1999. The work evolved into Question 9 (Q9/16), which aimed to standardise an embedded audio codec for a wide range of applications such as packetised voice, high quality audio/video conferencing, 3rd generation and future wireless systems, and multimedia streaming. In March 2007, the most promising technology [1, 2], jointly developed by Ericsson, Motorola, Nokia, Texas Instruments, and VoiceAge, was selected to form a baseline for an optimisation and characterisation phase, scheduled to end in April 2008. France Telecom, Huawei, Matsushita/Panasonic, and Qualcomm also contributed in the open collaboration towards the final standard.

G.718 (previously known as G.EV-VBR) is an embedded codec that comprises 5 layers (L1-L5) and covers bit rates from 8 kbps to 32 kbps. The core layer (L1) is based on the VMR-WB speech coding standard [3] and utilises several coding modes for different input signals. The core layer implements a Gaussian excitation or adaptive and algebraic (ACELP) excitations. Layers L2 through L5 code the coding error from the core layer: L2 provides an additional innovation codebook and the higher layers further enhance the coding by means of modified discrete cosine transform (MDCT) coding. In addition, L3 provides robustness features that significantly improve the codec’s frame erasure performance.

The codec accepts both wideband (WB) and narrowband (NB) inputs sampled at 16 kHz and 8 or 16 kHz, respectively. In case of 16-kHz sampled NB input, the bandwidth of the input signal is detected algorithmically. The NB processing (of 8-kHz signals) is limited to the lowest two layers with bit rates of 8 and 12 kbps, while WB processing is available at all rates. Further extensions will allow stereo (and mono) input sampled at 16 and 32 kHz and provide wideband and super wideband stereo (and mono) output.

One of the key features of the G.718 standard is embedded scalability. This means that the higher layer bit stream can be discarded during transmission without affecting the decoding of the lower layers. The embedded scalability and the wide range of bit rates place certain requirements for high quality processing of different signals. While background signals can be considered as part of the signal conveying ambient information at higher bit rates, they tend to significantly degrade the speech coding performance at lower bit rates. In G.718, a noise reduction algorithm [4] is included as a command line option for the two lowest bit rates. It has an internal sampling frequency of 12.8 kHz and is therefore suited only for the lower layers.

As the result of above considerations, a novel way of exploiting noise reduction where the core layer mode selection and estimation of certain parameters is carried out on a denoised signal (while still coding the noisy signal and maintaining its characteristics especially at higher bit rates) was developed and adopted for G.718. Figure 1 illustrates the overall effect of this method on the accuracy of mode selection for noisy speech when compared to the mode selection for the same speech signal with no added noise.

Reliable estimation of the fundamental frequency or pitch of the input speech signal is an
important feature in a speech codec. A new pitch estimator was implemented in G.718 in order to improve its pitch estimation performance. The estimator exploits pitch tracking in two separate tracks.

Guilmin et al. [5] have studied the influence of noise reduction on speech parameter estimation in the context of low bit rate parametric speech codecs. The study confirms the positive effect of noise reduction on pitch detection accuracy. This aspect is exploited in G.718, as the open loop pitch lag estimates and gains are estimated from the denoised signal.

Figure 1. Mode selection accuracy with and without noise reduction.

The paper is organized as follows. In Section 2 we present an overview of the G.718 pitch estimation technology. In Section 3 we describe the mode selection used in the codec. Section 4 discusses the effects of background noise on the mode selection performance, and in Section 5 we present the improvements achieved with noise reduction in the mode selection process as shown in Figure 1. Section 6 will then provide the conclusion.

2. Pitch tracking

Open-loop pitch analysis is performed to obtain a smooth pitch evolution contour and to confine the closed-loop pitch search to a smaller number of relevant candidate lags. The smoothness of the pitch evolution contour improves the coding efficiency of the ACELP waveform, and it helps to better preserve the output quality in case of channel errors. Frame erasure concealment techniques typically try to maintain the energy of the synthesised signal in quasi-periodic segments.

The performance of such methods especially during voiced speech is therefore sensitive to the accuracy of the pitch lag.

The G.718 pitch tracking is done once each 20-ms frame on the perceptually weighted signal, downsampled to 6.4 kHz in order to lower the complexity of the analysis. The algorithm produces three pitch lag estimates that correspond to each half (10 ms) of the present frame and the lookahead.

The pitch lag range is divided into sections such that for any pitch lag the section does not contain also its multiple. In addition, the length of the correlated vectors depends on the pitch lag search range in order to reduce the number of erroneously selected multiples or fractions of the correct pitch lag. In G.718, the algorithm evaluates two sets of these sections, which are summarized in Table 1. The lengths of the correlated vectors are also shown.

The sets have been selected in such a way that the sections in one set overlap the range boundaries between sections in the other set. For example, the section boundary between sections 1 and 2 in the 1st set is covered by section 2 in the 2nd set. This overlapping of the sections is introduced to eliminate rare cases, where the evolving of pitch contour across a section boundary results in erroneous selection of a pitch lag and thus a suboptimal contour for the frame. Note in Table 1 that section 2 of the 2nd set uses a different length for the correlated vectors for the lookahead (62 samples) than for other half frames (80 samples), which is due to the number of available lookahead samples. In addition, section 3 is not searched for the lookahead in either set for the same reason. Instead the values of the second half frame are reused in the lookahead.

The correlations $C(l)$ are calculated in each section of both sets in each half frame and lookahead as given by

$$C(l) = \sum_{n=0}^{L_{sec}} s(n)s(n-l)$$

where $s(n)$ is the weighted and decimated speech signal, $L_{sec}$ is the length of the vectors in the section, and $l$ is the pitch lag. The maximum correlation is then searched for each of the sections, and the section maxima are normalised

$$C_{score}(l) = \frac{C(l)}{\left(\sum_{n=0}^{L_{sec}} s^2(n)s^2(n-l)\right)^{\frac{1}{2}}}$$

(2)
The normalised section maxima are further reinforced in three levels. The first step is to favour pitch lags around the pitch lag extrapolated from past frame pitch lag estimates. The second level of reinforcing then favours smaller pitch lags provided that the normalised correlation is strong also around their multiples.

<table>
<thead>
<tr>
<th>Set</th>
<th>Section</th>
<th>Pitch lag range</th>
<th>Length of vectors (L_m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>10 – 16</td>
<td>40</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>17 – 31</td>
<td>40</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>32 – 61</td>
<td>62</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>62 – 115</td>
<td>115</td>
</tr>
</tbody>
</table>

Table 1. Pitch tracking sections in G.718

While the first two steps are very similar to the pitch tracking algorithm in [3], the third and final step differs more considerably mostly due to having two sets of sections in the pitch lag search. It consists of reinforcing those lags that have similar strong lags in the other half frames and the lookahead. This process starts with selecting the highest normalised correlation and the corresponding pitch lag over all sections for both sets in each half frame and the lookahead. A total of six or eight pitch lags are thus selected depending on whether section 0 is searched in the current frame. These pre-selected pitch lags are then compared with all the pitch lags corresponding to the highest normalised correlation values in all sections \( j \) of both sets \( m \) in the other two half frame(s) and/or lookahead, \( k \).

In case the tested pitch lags exhibit coherence with the pre-selected pitch lags, the corresponding correlation values are reinforced. The reinforcing further favours pitch lag values that are from the same pitch lag search range (i.e., same section of same set) as given by

\[
\bar{C}_{n,m}(\hat{\ell}_k^n) = \bar{C}_{n,m}(\hat{\ell}_k^n)\left[1 + \alpha(1 - \delta_{pit}/14)\right],
\]

where \( \delta_{pit} \) is the absolute difference between the tested pitch value and the pre-selected pitch value, and \( \alpha \) is a reinforcement factor upper-bounded to 0.40 (\( j = p \) and \( m = n \)), or to 0.25 (\( j \neq p \) or \( m \neq n \)), where \( p \) is the pitch search section in set \( n \) of the pre-selected pitch lag value.

Finally, the best track is selected by finding the highest correlation value in each half frame across both sets.

3. Signal classification and mode selection

The core layer of G.718 utilises signal classification based encoding with four distinct signal classes. Unvoiced coding mode is optimised for unvoiced speech, Voiced coding mode is optimised for quasi-periodic frames with smooth pitch evolution, Transition mode is for encoding frames following voicing onsets in order to minimise error propagation in presence of frame erasures, and Generic coding mode is used for all remaining frames. The codec also contains a scheme for discontinuous transmission (DTX) with comfort noise generation (CNG).

Figure 2 presents a flow chart of the classification logic. First, frames to be encoded using Unvoiced coding are selected. Additionally, when DTX is not used, Unvoiced coding is used for the inactive frames. In Unvoiced coding, the excitation is selected from a Gaussian codebook and no adaptive codebook is used.

Quasi-periodic frames that exhibit a smooth pitch evolution contour are encoded with Voiced coding mode. Voiced coding is similar to Generic mode, but due to the smooth pitch evolution, more bits can be allocated for coding the algebraic
codebook excitation. The smoother the pitch contours from the pitch tracking algorithm are, the more frames we can encode using the Voiced coding mode.

The Transition coding mode limits the use of past frame information and thus helps in preserving the codec performance in case of frame erasures. On the other hand, its usage is limited to only the most critical frames to ensure the negative effect on clean channel performance is tolerable. In Transition coding, a fixed codebook is used instead of the adaptive codebook in the subframe containing the glottal impulse of the first pitch period. The adaptive codebook is omitted altogether in the subframes before this replacement. The remaining frames are encoded with Generic coding mode, which implements an Algebraic CELP (ACELP).

4. Effects of background noise

Background noise conditions often present a challenge in speech coding, specifically at low bit rates. The issue of background noise can be tackled in many ways, e.g., training material for codebooks typically includes a considerable amount of speech with different types of background noise. On the other hand, noise reduction methods may be employed.

In G.718, an optional noise reduction is available for the two lowest bit rates. It is however not utilised by default and the coding framework considers background noise as part of the signal especially at higher bit rates. The background noise component should therefore not be completely removed from the output signal.

In Table 2 it can be seen that Unvoiced coding is often used instead of Generic coding and vice versa. This behaviour is, however, perhaps the least important from the point of view of perceptual quality, as these misclassifications typically occur in low energy frames. In frames with higher energy, they can degrade the codec performance. Frames that the voice activity detector (VAD) evaluates as silence (or non-active speech) either for the clean signal or for the noisy signal naturally also exhibit low energy.

Table 2. Classification (%) of noisy frames (columns) compared to clean frames (rows).

<table>
<thead>
<tr>
<th></th>
<th>Silence</th>
<th>Unvoiced</th>
<th>Voiced</th>
<th>Generic</th>
<th>Trans.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Silence</td>
<td>95.55</td>
<td>2.00</td>
<td>0.03</td>
<td>2.30</td>
<td>0.12</td>
</tr>
<tr>
<td>UV</td>
<td>5.58</td>
<td>64.89</td>
<td>0.06</td>
<td>29.24</td>
<td>0.24</td>
</tr>
<tr>
<td>Voiced</td>
<td>0.24</td>
<td>0.40</td>
<td>87.95</td>
<td>8.42</td>
<td>2.99</td>
</tr>
<tr>
<td>Gen.</td>
<td>5.12</td>
<td>15.39</td>
<td>1.55</td>
<td>75.73</td>
<td>2.20</td>
</tr>
<tr>
<td>Trans.</td>
<td>0.11</td>
<td>0.14</td>
<td>2.19</td>
<td>30.15</td>
<td>67.41</td>
</tr>
</tbody>
</table>

In Table 2 it can be seen that Unvoiced coding is often used instead of Generic coding and vice versa. This behaviour is, however, perhaps the least important from the point of view of perceptual quality, as these misclassifications typically occur in low energy frames. In frames with higher energy, they can degrade the codec performance. Frames that the voice activity detector (VAD) evaluates as silence (or non-active speech) either for the clean signal or for the noisy signal naturally also exhibit low energy.

5. Effects of noise reduction

In order to reduce the adverse effects of background noise on both the perceptual quality of the coded speech and the overall frame error
robustness of the codec, a novel approach was developed to improve the mode selection and the estimation of certain parameters by making the codec less dependant on background noise. The approach proceeds from the assumption that the clean signal classification is usually optimal also for the noisy signal, which is supported by informal expert listening. This approach utilises the noise reduction method of [4]. While the source coding is based on a signal with no noise reduction, a denoised signal is used in open-loop pitch analysis and signal classification. In G.718, noise reduction is employed at the level of 14 dB.

Table 3 presents the noisy mode selection for the same speech signal as used in Table 2. In this case a noise reduction at 14 dB was applied to the signal prior to open-loop pitch estimation, signal classification and mode selection. It can be seen that the mode selection accuracy compared to clean speech classification is improved, although it still differs considerably from the clean speech classification. This behaviour is expected, as the denoised signal characteristics, in comparison to noisy signal characteristics, should more closely match that of the clean signal while the actual signal waveform should still be very different from the clean signal both due to the remaining background noise and to artefacts introduced by the noise reduction. In more detail, the mode selection match improves for silence, Unvoiced coding mode, Voiced coding mode, and Transition mode. On the other hand, there is a slight drop in accuracy for the Generic coding mode.

<table>
<thead>
<tr>
<th>Silence</th>
<th>Unvoiced</th>
<th>Voiced</th>
<th>Generic</th>
<th>Trans.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Silence</td>
<td>96.27</td>
<td>2.56</td>
<td>0.01</td>
<td>1.09</td>
</tr>
<tr>
<td>UV</td>
<td>5.60</td>
<td>83.69</td>
<td>0.03</td>
<td>10.62</td>
</tr>
<tr>
<td>Voiced</td>
<td>0.23</td>
<td>0.52</td>
<td>91.33</td>
<td>5.74</td>
</tr>
<tr>
<td>Gen.</td>
<td>5.01</td>
<td>17.40</td>
<td>2.38</td>
<td>73.16</td>
</tr>
<tr>
<td>Trans.</td>
<td>0.08</td>
<td>0.14</td>
<td>2.81</td>
<td>24.24</td>
</tr>
</tbody>
</table>

Table 3. Classification (%) of noise reduced frames (columns) compared to clean (rows).

Perhaps the single most obvious advantage achieved in classification through noise reduction is the partial correction of the Transition mode misalignment problem observed in noisy signal classifications. When noise reduction is applied, the mode selection algorithm more often allocates the Transition mode right after the voicing onset, as designed. This better maintains the frame erasure robustness in case of high background noise levels. The Voiced mode classification is also improved.

Another change in the classification accuracy is observed in the low-energy frames. For the clean speech signal, these frames are predominantly classified as either Unvoiced coding or Generic coding frames. Many frames are also detected as silence. In case of background noise, the mode selection often changes from Unvoiced to Generic or vice versa. Noise reduction tends to significantly improve the mode selection accuracy for the low-energy Unvoiced frames, while the accuracy of the Generic classification experiences a slight drop as more of these frames are now classified as Unvoiced and also Voiced. In perceptual terms these two phenomena are the least important changes, since they can mostly be credited to low-energy frames where the effects of the coding mode selection are often inaudible.

An interesting question about the use of noise reduction for signal classification concerns the optimal amount of noise reduction: How heavy noise reduction is sufficient, and how much noise reduction should be applied on the signal to optimize its characteristics in terms of signal classification accuracy? In order to answer this question, we have carried out an extensive test covering street noise, car noise, and office noise at levels between -40 dB and 0 dB with 5 dB steps. All the combinations were tested with noise reduction set at 0, 4, 8, 11, 14, 17, and 20 dB.

Figure 3 presents the overall effect of street noise levels (between -30 dB and -10 dB) and noise reduction levels on classification accuracy. Figure 4 illustrates the reduction level effect in this experiment for all frame types at a street noise level of -15 dB and Figure 5 presents it similarly for car noise.
It was found out that street noise and car noise affect the mode selection in a similar manner, although the behaviour of Unvoiced and Generic coding modes noted above is more clear with car noise (as seen in Figure 5). It was also observed that the classification accuracy is generally somewhat higher for car noise than street noise, which is mostly due to car noise being the more stationary of these noise types.

Office noise, on the other hand, proved more difficult for the noise reduction and following mode selection. It can be concluded that there was no considerable improvement or degradation in case of office noise regardless of the noise reduction level at noise levels between -40 dB and -15 dB. While some improvement was observed at higher noise levels, it was not as clearly manifested as for street noise and car noise cases.

As can be seen in Figure 3, the noise reduction level of 14 dB, which is used in G.718, is effective and close to the saturation. It should be noted that this is also the default operation point for the noise reduction as used in the VMR-WB standard.

6. Conclusion

We have presented a new approach for open-loop pitch estimation, and signal classification and mode selection for noisy signals utilising noise reduction prior to these steps.

The presented method of noise-reduced mode selection, exploited in the new ITU-T G.718 embedded scalable speech and audio coding standard, is particularly useful for scalable coding, where it may not be desirable to apply noise reduction to the coded signal but instead treat the background noise as part of the signal. Basing the mode selection of the speech codec on the denoised signal improves the algorithm performance by considering only the speech signal in the classification also in case of background noise. It has been observed in the G.718 framework that this approach especially increases the codec’s robustness against frame erasures in background noise by improved allocation of the Transition coding mode. The new open-loop pitch estimator also enhances the coding of quasi-periodic signal segments by providing more stable pitch estimates.

7. References