Transcoding of Narrowband to Wideband Speech

Christian H. Ritz
University of Wollongong, critz@uow.edu.au

Nick Harders
University of Wollongong

Joseph Hermann
University of Wollongong

Matthew J. Baker
University of Wollongong, matthewb@uow.edu.au

Publication Details

Research Online is the open access institutional repository for the University of Wollongong. For further information contact the UOW Library: research-pubs@uow.edu.au
Transcoding of Narrowband to Wideband Speech

Abstract
Transcoding is required to facilitate the communication of compressed speech between networks that have adopted opposing speech coding standards. The traditional transcoding technique of tandem conversion by decoding from the old standard and then re-encoding with the new standard suffers from unacceptable delay and complexity. For real time applications, delay and complexity can be reduced by performing transcoding in the bit stream domain. This paper describes techniques for transcoding between narrowband and wideband speech coding standards. In particular, an examination of the performance of bit stream mapping approaches to transcoding from the ITU-T G.729 narrowband speech coder to the ITU-T G.722.2 wideband speech coder is presented. Results for the proposed transcoder compared with a tandem transcoder indicate significant reductions in computational complexity however speech quality results less satisfactory. It is concluded that an ideal transcoder must consider the interaction of all speech parameters to ensure satisfactory speech quality.

Keywords
Transcoding, Narrowband, Wideband, Speech

Disciplines
Physical Sciences and Mathematics

Publication Details

This conference paper is available at Research Online: http://ro.uow.edu.au/infopapers/1625
ABSTRACT

Transcoding is required to facilitate the communication of compressed speech between networks that have adopted opposing speech coding standards. The traditional transcoding technique of tandem conversion by decoding from the old standard and then re-encoding with the new standard suffers from unacceptable delay and complexity. For real time applications, delay and complexity can be reduced by performing transcoding in the bit stream domain. This paper describes techniques for transcoding between narrowband and wideband speech coding standards. In particular, an examination of the performance of bit stream mapping approaches to transcoding from the ITU-T G.729 narrowband speech coder to the ITU-T G.722.2 wideband speech coder is presented. Results for the proposed transcoder compared with a tandem transcoder indicate significant reductions in computational complexity however speech quality results less satisfactory. It is concluded that an ideal transcoder must consider the interaction of all speech parameters to ensure satisfactory speech quality.

1. INTRODUCTION

A variety of speech coding standards have been defined and adopted for various telecommunications applications such as fixed and mobile telephony. Each standard uniquely defines how to represent the speech signal using a set of parameters that are quantised to form a bitstream. Emerging speech applications require interoperability between networks and applications which may use different speech coding standards. Such communication requires conversion of the bitstream from one standard to another, which is commonly known as transcoding [1].

One approach to transcoding is tandem conversion, illustrated in Figure 1(a). In this approach, bitstream, $b_A$ of coder A is decoded to synthesised speech, $s'(n)$ and then re-encoded with coder B to bitstream $b_B$. However, the delay and complexity associated with the decode/re-encode stage is unacceptable for real-time applications, such as telephony [1]. An alternative is bit stream mapping, illustrated in Figure 1(b). Bitstream $b_A$ of coder A is directly mapped to bitstream $b_B$ of coder B without full decoding and re-encoding, thus reducing the delay and complexity associated with tandem conversion [1].

Figure 1. (a) Tandem transcoder. (b) Bitstream mapping transcoder.

Existing bitstream mapping approaches to transcoding, including [1]-[4] have focused on standards defined for narrowband speech, which has a bandwidth of up to 4 kHz. For 3rd and future generation mobile networks and other Internet applications, wideband speech, with a bandwidth of up to 8 kHz, is preferred. Hence, emerging speech coding technologies will require transcoding between narrowband and wideband speech coding standards and is the focus of this paper. In particular, this paper will describe transcoding between the narrowband speech coding standard ITU-T G.729 [5] to the wideband speech coding standard ITU-T G.722.2 [6]. Both these standards are predominant techniques for Internet telephony.

Section 2 will provide an overview of both coders. The transcoding techniques used for the speech coding parameters are described in Sections 3 to 6. Section 7 presents and discusses speech quality and computational complexity results for these techniques, with conclusions described in Section 8.

2. OVERVIEW OF THE CODERS

Both the G.729 and the G.722.2 speech coders are based on the Algebraic Code Excited Linear Prediction (ACELP) [5] technique, with the differences highlighted below.

2.1. G.729

The G.729 speech coder is defined for narrowband speech sampled at 4 kHz. Linear Prediction Coding (LPC) coefficients are derived using frames of 10 ms while pitch, excitation and gain parameters are extracted for sub-frames of 5 ms. The coder operates at 8 kbps. These parameters are quantised using the bit allocation shown in Table 1.
The G.722.2 is a multi-rate wideband speech coder defined for wideband speech sampled at 16 kHz. The coder operates at bit rates from 6.60 kbps to 23.85 kbps. For this work, the coder is chosen to operate at 8.85 kbps, (closest to the G.729 coding rate used here), and this coder quantises parameters using bit allocations shown in Table 1. The coder separates the speech into two sub-bands: 50 - 6.4 kHz and 6.4 - 7 kHz. The lower sub-band (re-sampled to 12.8 kHz) is coded using ACELP while the upper sub-band is represented using noise models that are generated from the lower sub-band. For the lower sub-band, LPC coefficients are derived for 20 ms frames while pitch, excitation and gain parameters are derived for 5 ms sub-frames. The coder also derives a Voice Activity Detection (VAD) flag for each frame.

3. LPC PARAMETER TRANSCODING

This section elaborates on the LPC parameter representation and quantisation used by both coders and describes codebook mapping approaches proposed for LPC parameter transcoding.

3.1. Comparison of LPC coefficient representation and quantisation

For G.729, 10th order LPC coefficients are represented using 10th order Line Spectral Frequency (LSFs) while for G.722.2, 16th order LPC coefficients are represented as 16th order Immittance Spectral Frequencies (ISFs). For the both coders, the LSF (or ISF) for the current frame is predicted from the LSF (or ISF) from the previous frame and the resulting prediction residual is quantised to the number of bits specified in Table 1 using a combination of multistage and split VQ [7]. Further details are provided in [5][6]. Due the use of predictive VQ by both coders, prediction errors will be uncorrelated with the speech spectral envelope. Hence prediction residuals are decoded to LSF and ISF vectors and transcoding is performed in that domain.

3.2. Transcoding the LPC parameters via codebook mapping

For transcoding of the LPC parameters, a codebook mapping approach is proposed. Such an approach is motivated by the bandwidth expansion techniques for narrowband speech, described in detail in [8]. In [8], codebooks are designed which contain representations of narrowband LPC spectra and their corresponding wideband LPC spectra. In this paper, we propose a similar technique, whereby codebooks are designed containing representations of the narrowband G.729 LSF vectors and their corresponding wideband G.722.2 ISF vectors. Such a scheme is illustrated in Figure 2.

Figure 2. Codebook mapping for transcoding of G.729 LSFs to G.722.2 ISFs.

In Figure 2, an input G.729 LSF vector is compared with those in the first LPC transcoder codebook to find the best match using a mean squared error search technique. The corresponding ISF in the second LPC transcoder codebook is then chosen as the transcoded ISF. The final step is to quantise this transcoded ISF using the standard techniques defined for G.722.2, resulting in the LPC bitstream for this coder.

3.3. Design of the LPC Transcoder Codebooks

The design of the LPC transcoder codebook is similar to that used in the codebook mapping approach to bandwidth extension [8]. A training database of LSF and corresponding ISF vectors is formed. A VQ codebook is designed for the LSF vectors using the standard Generalized Lloyd Algorithm (GLA) [7] and the ISF codebook is designed using the following algorithm:

- Quantise the LSF training vectors using the designed codebook.
- Partition the ISF training vectors into groups for which the corresponding LSF vector has the same quantised codeword.
- Average all ISF vectors within each partition to form the codewords of the ISF codebook.

The training database used in this work was obtained by encoding approximately 30 minutes of speech using the standard LPC techniques defined for the G.729 and G.722.2 coders, respectively. The performance of the trained codebooks can be

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Bits per frame</th>
</tr>
</thead>
<tbody>
<tr>
<td>G.729</td>
<td>G.722.2</td>
</tr>
<tr>
<td>LPCs</td>
<td>18</td>
</tr>
<tr>
<td>VAD Flag</td>
<td>0</td>
</tr>
<tr>
<td>Pitch (period)</td>
<td>13</td>
</tr>
<tr>
<td>Pitch (parity bit)</td>
<td>1</td>
</tr>
<tr>
<td>Excitation Signal</td>
<td>34</td>
</tr>
<tr>
<td>Gains</td>
<td>14</td>
</tr>
<tr>
<td>Total</td>
<td>80</td>
</tr>
</tbody>
</table>

Table 1. Bit allocation for the G.729 and G.722.2 speech coders.
measured using the Spectral Distortion (SD) \([9]\) (defined in (1)) resulting from quantising the ISF vectors using the designed codebooks of different sizes.

\[
SD = \frac{1}{K} \sum_{k=1}^{K} \left[ 20 \log_{10} \left( \frac{G_c A_j(\omega_k)}{\tilde{A}_j(\omega_k)} \right) \right]^2 d\omega
\]

where, \(G_c = \sum_{k=1}^{K} \tilde{A}_j(\omega_k) d\omega\)

In (1), \(\omega_k\) is the \(k\)th frequency out of the total set of \(K\) frequencies over which the \(j\)th original and transcoded magnitude spectra \(A_j\) and \(\tilde{A}_j\), respectively, are evaluated and \(G_c\) is used to scale the original spectra so that only the distortion in the envelope shape is evaluated, as suggested in \([8]\).

Figure 3 shows the SD results when transcoding a database of G.729 LSF vectors to G.722.2 ISF vectors using different sized codebooks. These vectors were derived for approximately 2 minutes of speech that is different from the training database. To investigate the performance over different frequency ranges, the SD is measured separately for the 0 to 4 kHz and the 4 kHz to 6.4 kHz frequency ranges.

Figure 3. SD versus bitrate resulting from ISF quantisation. Lowband: 0 to 4 kHz. Highband: 4 kHz to 6.4 kHz.

In Figure 3, the spectral distortion of the low frequency region decreases as the bit rate increases. Conversely, the SD of the high frequency region shows little change for the codebooks tested. These results indicate that the clustering of the wideband ISFs based on narrowband LSFs is justified for those representing the narrowband (0 to 4 kHz) region but not necessarily for the high frequency (4 to 6.4 kHz) region of the LPC spectral envelope. These results agree with existing work in bandwidth extension of narrowband speech, which has demonstrated that there is only minimal correlation between low and high frequency regions of LPC magnitude spectra \([8]\).

The results also indicate that the SD for the low frequency region will further reduce by increasing the size of the LPC transcoder codebooks. However, larger codebooks require increased search complexity. Hence, in this work, a 24 bit codebook was chosen to provide a good tradeoff between SD and search complexity. To further minimise search complexity, this codebook was implemented as a multistage codebook \([7]\), with three 8 bit stages.

### 3.4. Improved LPC transcoding by interpolation

To improve the performance of the codebook mapping approach, an interpolative technique is proposed, similar to that described in \([8]\) for narrowband to wideband LPC spectra mapping. In this approach, the \(K\) ISF vectors corresponding to the \(K\) closest matching LSF vectors are averaged to form a new ISF vector, as described in (5).

\[
y' = \frac{1}{N} \sum_{k=1}^{K} y_k
\]

In (5), \(y'\) represents the average ISF vectors, correspond to the \(K\) nearest matching ISF vectors, \(y_k\). To measure the performance of the interpolative ISF technique, the SD was measured using the 24 bit codebook described in Section 6.5 for various interpolation factors, \(K\). These results are shown in Figure 4.

Figure 4. SD versus interpolation factor, \(K\), for a 24 bit LPC transcoder codebook.

Figure 4 shows that little change in the SD results beyond an interpolation factor of 4 for both coders and so was chosen in this work.

### 4. PITCH AND VAD TRANSCODING

Both coders represent the pitch period using a value in samples. Absolute pitch period values are used for odd numbered sub-frames while differential pitch values are used for even numbered sub-frames. In both coders, pitch is calculated and quantised using the same sub-frame size and bit allocation. Hence, a G.722.2 pitch can be obtained from a G.729 pitch value by multiplying by the ratio of the sampling rates (in kHz) and is given by expression (2).

\[
T_{G.722.2} = \frac{12.8}{4} T_{G.729} = 1.6 T_{G.729}
\]

\(T_{G.722.2}\) pitch can be obtained from a G.729 pitch value by multiplying by the ratio of the sampling rates (in kHz) and is given by expression (2).
In (2), T_{G.729} and T_{G.722.2} are the pitch periods (in samples) for the G.729 and G.722.2 speech coders. In addition, some scaling has to be performed to account for the slightly different pitch ranges used in both coders (1.67 ms to 18.5 ms in G.729 versus 2.03 ms to 18.6 ms in G.722.2).

The Voice Activity Detector (VAD) flag is used to indicate bitrate reduction during non-speech activity and is only incorporated into the G.722.2 speech coder. Hence, the VAD flag was set to 1 for all transcoded frames.

5. EXCITATION PARAMETER TRANSCODING

The excitation signal for each of the coders is represented by four separate pulses whose amplitude is represented by a single sign bit and whose location is quantised to one of a set of locations specified in the fixed codebook. For G.729, 8 locations for tracks 1 to 3 and 16 locations for track 4 are specified requiring 3 bits and 4 bits for these tracks, respectively, making a total of 17 bits per subframe. For G.722.2, 16 locations are specified for each track hence requiring 4 bits per track, making a total of 20 bits per subframe.

The locations specified in the fixed codebooks of each coder differ by the ratio of the sampling rates. By examining the fixed codebooks of each coder (see [7-8]), direct conversion using this factor will only map track 1 accurately between each coder, with the location within other tracks requiring rounding. However, rounding of pulse locations will not guarantee a pulse from a given track within the G.729 fixed codebook is mapped to the same track in G.722.2 fixed codebook. For example, pulse position 3 in track 2 of the G.729 fixed codebook is 6, the closest rounded value following conversion by 1.6 is 10, which is a location within track 3 of G.722.2.

By comparing the rounding errors associated with the conversion using this factor, it was found that the mapping algorithm of Table 2 resulted in least location errors.

<table>
<thead>
<tr>
<th>G.729 Track</th>
<th>G.722.2 Track</th>
</tr>
</thead>
<tbody>
<tr>
<td>Location 0-7 of G.729 Track 4</td>
<td>Location 8-15 of G.729 Track 4</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 2. Best matching G.729 and G.722.2 excitation tracks.

In Table 2, G.729 tracks 2 and 4 are mapped differently depending on whether pulse 4 is located within positions 0 to 7 or 8 to 15 of track 4 to ensure minimal errors (due to rounding) in excitation mapping.

6. GAIN PARAMETER TRANSCODING

For both coders, the fixed (excitation) gain for the current frame is predicted from the fixed codebook gain of the previous gain. The resulting prediction coefficient is combined with the adaptive (pitch) gain and these are quantised together using vector quantisation.

6.1. Gain codebook mapping by nearest match

The G.729 coder uses a two-stage codebook with sizes of 3 bits and 4 bits for stage 1 and 2, respectively. The 8.85 kbps G.722.2 coder uses a single 6 bit codebook.

For transcoding, the gains were decoded using the relevant codebooks and a direct mapping approach investigated. In this approach, a table is formed that indicates, for each of the possible 128 G.729 gain vectors, a corresponding 6-bit index in the G.722.2 gain codebook. This table was created using a training procedure that minimises the mean squared error distortion described in (3) to find the best matching G.722.2 gain as described in (4).

\[ d(g_{729}, g_{722.2}) = 0.5 \left( \| g_{729, p} - g_{722.2, p} \| + \left\| g_{729, e} - g_{722.2, e} \right\| \right) \]  

(3)

\[ g_{tr}(j) = \min_i \left[ d(g_{729}(i), g_{722.2}(j)) \right] \]  

(4)

In (3), \([g_{729, p}, g_{729, e}]\) and \([g_{722.2, p}, g_{722.2, e}]\) are the G.729 and G.722.2 gain vectors, respectively, where subscripts \(p\) and \(e\) denote the pitch and excitation gain, respectively.

Informal listening tests found the resulting speech to be generally of poor quality when using the initial table lookup. Examination of speech waveforms found much of the distortion caused by clipping of the speech as a result of incorrect gain values. This was a consequence of the joint quantisation of both gains failing to ensure that the individual gain errors are minimised. Hence, an accurately mapped pitch gain may lead to a large error in the excitation gain and vice-versa.

6.2. Gain codebook mapping by most frequent match

To further investigate the correlation between the quantised gains for both coders, Figure 5 shows the gain codebook indices generated when coding 30 minutes of narrowband speech using the G.729 coder and the G.722.2 coder applied to an upsamped (to 16 kHz) version of the same speech.
As can be seen from Figure 5, the majority of indices chosen from the G.729 gain codebook, map to a wide range of possible indices within the G.722.2 gain codebook index as determined from the results of Figure 5. To minimise occasional spikes in the excitation gain (hence causing speech clipping), a simple smoothing technique was applied, whereby changes in the excitation gain between frames was limited. Informal listening tests found that the new codebook combined with gain smoothing produced speech of similar or better quality compared with the codebook mapping approach describe in Section 6.1.

An alternative approach adopted here is to form a table that maps the index from the G.729 gain codebook to the most frequent matching G.722.2 gain codebook index as determined from the results of Figure 5. To analyse the performance of the transcoding techniques developed in Sections 3 to 6, PESQ results were obtained for speech synthesised from G.722.2 bitstreams where only a single parameter was transcoded. When transcoding only a single parameter, the other parameters were represented using the G.722.2 bitstreams that would have been generated following a full encode of the original speech signal. These results are shown in Table 3.

<table>
<thead>
<tr>
<th>Synthesised Speech</th>
<th>PESQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>G.722.2 @ 8.85 kbps</td>
<td>3.7</td>
</tr>
<tr>
<td>G.729 @ 8 kbps</td>
<td>3.6</td>
</tr>
<tr>
<td>Tandem transcoded</td>
<td>3.4</td>
</tr>
<tr>
<td>Complete transcoded</td>
<td>1.8</td>
</tr>
<tr>
<td>Pitch transcoded only</td>
<td>2.9</td>
</tr>
<tr>
<td>LPCs transcoded only</td>
<td>3.0</td>
</tr>
<tr>
<td>Gain transcoded only</td>
<td>2.9</td>
</tr>
<tr>
<td>VAD transcoded only</td>
<td>4.5</td>
</tr>
<tr>
<td>Excitation transcoded only</td>
<td>2.4</td>
</tr>
</tbody>
</table>

Table 3. PESQ scores for various speech files.

In Table 3, results for G.729 and G.722.2 were obtained using original 8 kHz and 16 kHz sampled speech, respectively, as the reference files. The results for transcoding were obtained by using speech synthesised using the G.722.2 coder as the reference files; this was chosen as it is expected that this is the maximum quality that could be achieved when transcoding these two coders.

Table 3 shows that tandem transcoded speech has superior quality to the bit stream transcoded speech. When transcoding a single parameter, results are significantly better results than results obtained when transcoding all parameters using the proposed technique, however still inferior to results obtained for tandem transcoding.

When transcoding pitch, the LPCs or gain, the resulting PESQ is similar (2.9 or 3.0) compared with 1.8 when all parameters are transcoded. The worse result for transcoding a single parameter is for the excitation. The high result for transcoding VAD is due to the use of a G.722.2 synthesised speech files as reference files for PESQ analysis. Hence, a PESQ of 4.5 indicates that there is
virtually no loss in subjective quality when transcoding the VAD flag.

The PESQ results can be explained by analysing the techniques and results presented in Sections 3 to 6. While the pitch transcoding technique of Section 4 resulted in minimal errors during voiced speech, errors during unvoiced speech leading to distortions in these regions. One technique for improving pitch transcoding could be to utilise a smoothing technique to minimise occasional pitch errors.

Section 6 showed that the gain parameters derived for both coders display little correlation. This could be due to both coders using analysis by synthesis techniques, which compare original and reconstructed speech when quantising excitation and gain parameters. A better approach may be to perform gain transcoding in the excitation or speech domain, as suggested in [1] for G.729 to IS-641 transcoding.

The results presented in Section 3 for LPC parameter transcoding indicate significant distortion compared with the generally accepted spectral distortion limit of 1 dB to ensure minimal loss in subjective speech quality when quantising narrowband LPC spectra [10]. An improvement in LPC parameter transcoding could be obtained by adopting more sophisticated techniques similar to those used in bandwidth extension of narrowband speech, such as those suggested in [8].

7.2. Computational Complexity

An analysis of the computational complexity was performed by measuring the average CPU computation time. Bitstreams were derived for a 2 minute speech file using G.229 and converted to a G.723.2 bitstream using tandem conversion and the proposed transcoder, where each parameter is transcoded using the bit stream mapping approaches described in Sections 3 to 6. This was repeated for 20 trials and the average results per second of speech are shown in Table 4.

From Table 4, it can be seen that the proposed transcoder introduces almost 10 times less delay than a tandem conversion. It should be noted these are comparative results only and absolute delays would be dependent on the actual hardware implementation.

<table>
<thead>
<tr>
<th>Method</th>
<th>Delay per second (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tandem</td>
<td>262.5</td>
</tr>
<tr>
<td>Proposed</td>
<td>26.92</td>
</tr>
</tbody>
</table>

Table 4. Comparison of computational complexity

8. CONCLUSION

This paper has described a codebook mapping approach for the transcoding of G.729 bitstreams to G.722.2 bitstreams. Each of the pitch, gain, excitation and LPC parameters were treated separately during transcoding. Results for PESQ scores show that the proposed transcoding technique produces speech of inferior quality to speech produced by tandem conversion.

From this work it can be concluded that a G.729 to G.722.2 transcoder that considers the individual parameters only during parameter conversion will not produce speech of satisfactory quality. It is proposed that a better technique would be to consider the interaction of each of the parameters on the overall speech quality during transcoding.

REFERENCES


