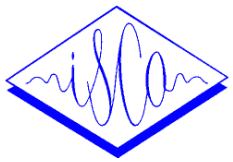


7th International Conference on Spoken Language Processing [ICSLP2002]
Denver, Colorado, USA
September 16-20, 2002

IMPROVED PERFORMANCE SPEECH CODEC FOR MOBILE COMMUNICATIONS

K. Humphreys and R. Lawlor

Department of Electronic Engineering
National University of Ireland, Maynooth
Ireland
kenneth.humphreys@eeng.may.ie

ABSTRACT

This paper presents the application of a Voice Gender Normalization algorithm to the GSM Speech Codec and describes the refinements that can be made to the Codec as a result. By reducing the dynamic range of the speech signals entering the Codec gender specific adaptations can be made to the Codec to improve its performance in terms of subjective sound quality or its transmitted bit rate.

1. INTRODUCTION

Any Speech Codec (encoder and decoder) employed in a mobile communications system must strive to remove as much redundant information as is possible from the speech waveform while encoding it in a robust manner. In addition the Codec is constrained to maintain a toll level of speech quality. Most Codecs employed in this field use Linear prediction based algorithms and perform their encoding and decoding using discrete mathematical models of the speech production mechanism resulting in a parametric representation of speech. To compress data as much as possible the parameters are rigorously quantized prior to transmission. The quantization stages and indeed the Codec itself try to accommodate the different dynamic ranges and distributions of parameters that occur for male and female speech. Limiting the type of input signal a Codec has to deal with and dedicating its resources to a smaller class of signal could improve the performance of a Codec. Specifically the performance could be improved by limiting the input signals to one gender only [1]. Furthermore efficient algorithms exist that are capable of normalizing the speech of one gender onto that of another gender (or even some interim gender neither male nor female) and then de-normalizing back to the original gender with little or no perceptual degradation in the quality of speech [2]. This paper presents the application of one such Voice Gender Normalization (VGN) algorithm to the speech Codec used in the GSM public mobile communications system, and presents the resulting modifications that can be made to the Codec for the purposes of lowering its bit rate while maintaining its current speech quality or improving quality at the current bit rate.

2. OVERVIEW OF VGN ALGORITHM

The term Voice Gender Normalization is used in this paper to mean the mapping of a particular class of speech waveforms onto another, so as to reduce the dynamic range of the speech signals presenting at the input to a Codec. The procedure is performed using an overlap-add algorithm known AOLA (Adaptive Overlap-Add) [3]. This is a time scale modification algorithm i.e. it is capable of lengthening or shortening the duration of a signal without altering its frequency content. The normalization of a speech signal is achieved by repeatedly using AOLA to scale the pitch (glottal) and formant (vocal tract) components of the signal by separate scaling factors denoted here as S_p and S_f respectively. The procedure for normalizing is outlined below.

- Time–scale modify the input speech by S_f (thus reducing the duration but preserving frequency content), then play back at S_f times the original sampling rate (thus duration is restored and both pitch and formant components are scaled by S_f)
- Use Linear Predictive Analysis to provide an approximate de-convolution of the formant and glottal components of the signal
- Time-scale modify the Linear Prediction Residual (representing glottal component only) by S_p / S_f (thus undoing the unwanted scaling of the pitch component by S_f and scaling it by the desired factor S_p)
- Using the Linear Prediction coefficients obtained in the second step, synthesis the speech and play back at S_p times the original sampling rate

This completes the normalization process. De-normalization is achieved by repeating the above process but with the scaling factors inverted.

Subjective listening tests showed that the back-to-back performance of the normalization and de-normalization processes is best when female speech is normalized to male and then de-normalized back to female. Normalization under these circumstances effectively constitutes a voice gender conversion. To perform such a conversion the scaling factors are chosen so as to correspond to the physiological ratios of the male and female speech production mechanisms. Given that the average adult male vocal tract is approximately 1.2 times the length of the average adult female vocal tract and the male vocal chords are on average 1.6 times the length of the female chords, the appropriate scaling factors are S_f=0.8 and S_p=0.6.

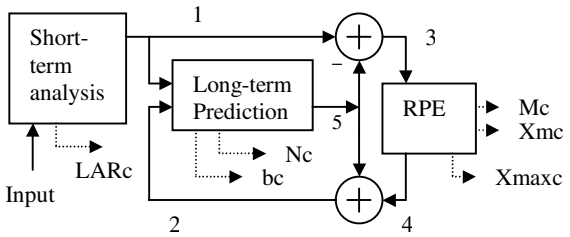
It is clear from the description of the VGN process above, that the use of Linear Prediction to provide separation of the glottal and formant components of the speech signal so as to allow independent scaling of the two is an integral part of the algorithm. As such it is likely that a Codec utilizing a Linear Prediction based algorithm will provide the greatest synergy when combining VGN with an existing Speech Codec.

3. OVERVIEW OF SPEECH CODEC

3.1. Outline of Codec

In order to test the application of VGN in the area of mobile communications it is necessary to develop a concise model of an existing speech Codec. It is also desirable that this be Linear Prediction based. The GSM 06.10 Speech Codec was chosen as the test model on the basis that it widely used (in over 100 countries) and details of its implementation are available in the form of ETSI documents [4]. This Codec uses an algorithm called Long-Term Prediction – Regular Pulse Excitation (LTP-RPE). LTP-RPE is a member of the class of residual excited linear prediction systems and it uses both short and long-term prediction to produce a long-term residual signal that is then encoded using the RPE algorithm [5].

Referring to Figure 1, short-term analysis is performed by means of an 8th order, all zero, lattice filter, using the auto-correlation method of Linear Prediction. Short-term analysis produces two things; firstly, it produces 8 Log Area Ratios (denoted here as LARc (i)), which represent the filter parameters and describe the frequency response of the vocal tract over that 20ms frame; secondly, short-term analysis produces a short-term residual signal (STR) representing the glottal component of the speech waveform. The 8 LARc's form part of the transmitted bit stream while the STR is passed to the Long-Term Predictor (LTP). At the LTP delaying and weighting a section of the reconstructed STR generates a prediction of the STR. The delay and weight (denoted Nc and bc respectively) form the next part of the transmitted bit stream. The estimate or prediction is subtracted from the actual STR to form an error signal called the long-term residual, which is then encoded using the RPE algorithm. The parametric description of the long-term residual forms the remainder of the transmitted bit stream.



1 = short-term residual
2 = reconstructed short-term residual
3 = long-term residual
4 = reconstructed long-term residual
5 = prediction of short-term residual

Figure 1: Simplified diagram of GSM encoder (dashed lines indicate transmitted parameters).

3.2. Outline of transmitted parameters

Since the primary objective of this application is to improve the performance of a Speech Codec by lowering the number of bits needed to represent the parameters while maintaining quality (or improving quality at existing bit rate), it is useful here to summarize the parameters of the Codec and the cost of each parameter in terms of bits per frame.

Name	Description	No. of Bits	No. per frame	Bits per frame
LARc (1)	Log Area Ratio	6	1	6
LARc (2)	Log Area Ratio	6	1	6
LARc (3)	Log Area Ratio	5	1	5
LARc (4)	Log Area Ratio	5	1	5
LARc (5)	Log Area Ratio	4	1	4
LARc (6)	Log Area Ratio	4	1	4
LARc (7)	Log Area Ratio	3	1	3
LARc (8)	Log Area Ratio	3	1	3
Nc	LTP lag	7	4	28
bc	LTP gain	2	4	8
Mc	RPE index	2	4	8
Xmaxc	Block Amplitude	6	4	24
Xmc(1..13)	13 RPE pulses	39	4	156

Table 1: Parameters transmitted per frame.

This gives rise to a bit rate of 260 bits/frame x 50 frames/second = 13kbps.

It can be seen in Table 1, that the majority of the bits transmitted per second relate to the RPE encoding of the long-term residual signal. RPE works by decimating a block of 40 samples (5ms) of the long-term residual into 4 interleaved sequences each with 13 values. The interleaved sequence with the most energy is taken to be the best estimate of the 40-sample block, and those 13 values are normalized (using the block maximum, Xmaxc), quantized and transmitted along with the index of the first element of the sequence (Mc). The 40-sample block of the long-term residual is reconstructed by placing the 13 values back in their correct position and filling in the remaining values with zeros.

4. NORMALIZATION APLIED TO GSM CODEC

4.1. Implementation

The GSM Codec having been modeled using MATLAB™ was placed between the voice normalization and de-normalization

algorithms, as shown in Figure 2, so that female speech at the input to the system is normalized to male sounding speech, encoded and decoded, and then de-normalized back to female speech. Subjective listening tests performed on a small panel of listeners using speech samples from the TIMIT database showed no significant difference between the output from this system and the output from the normal back-to-back encoder and decoder.

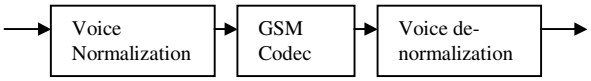


Figure 2: Normalization applied to GSM Codec.

Since a user of the system never hears the normalized speech its likeness to the target gender is unimportant so long as the de-normalized speech sounds the same as the input speech.

4.2. Results

Having first determined by subjective means that voice normalization can be successfully applied to the Codec without degrading the quality of the output speech, it remains to develop objective measures of the effects of normalization on the Codec parameters to determine what savings can be made by the application.

4.2.1. Short-term Analysis

The parameters produced by short-term analysis are the 8 LARc's describing the vocal tract frequency response over one frame. In order to gain an insight into the effects of normalization on these parameters, female speech signals consisting of several hundred frames were processed, first by the Codec alone and then by normalization algorithm followed by the Codec. By plotting the results in histogram form the distribution of LARc's over their quantization range can be observed. Figure 3 compares the histograms of the first LARc for a female speech signal taken from a Siemens Codec Test CD. Figure 3 (A) shows the parameter's unaltered distribution through its quantization range and Figure 3 (B) shows how the distribution has been affected by normalizing the speech prior to it entering the Codec.

LARc (1) is quantized to a range of 64 levels from –32 to +31 and for normal female speech occupies a significant part of that range. However after normalization has been applied LARc (1) occupies less than half its normal quantization range, meaning that it could be quantized as a 5-bit number rather than the usual 6 bits. This represents a saving of 1 bit/frame or 50 bits/second.

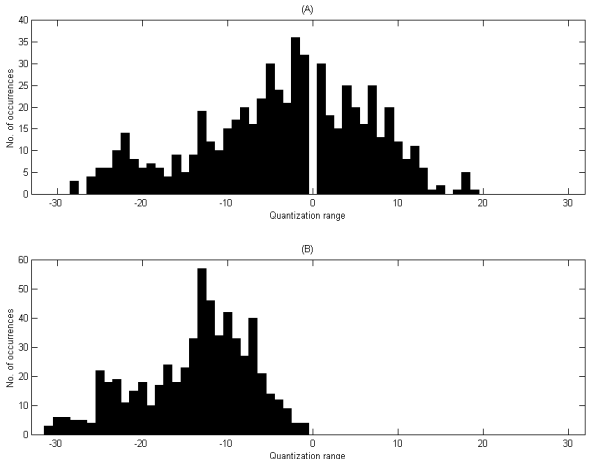


Figure 3: Histogram of LARc (1) for 600 frames of female speech.

Figure 4 displays the result from the same test carried out on the second LARc. Figure 4 (A) is the histogram of LARc (2) for the normal female speech signal and Figure 4 (B) is the histogram of LARc (2) when the speech is normalized prior to encoding. After normalization LARc (2) occupies less than half of its normal quantization range, meaning that this too could be represented by a 5-bit number rather than its current 6 bits per frame. This also constitutes a 50 bit/second saving in the transmitted data.

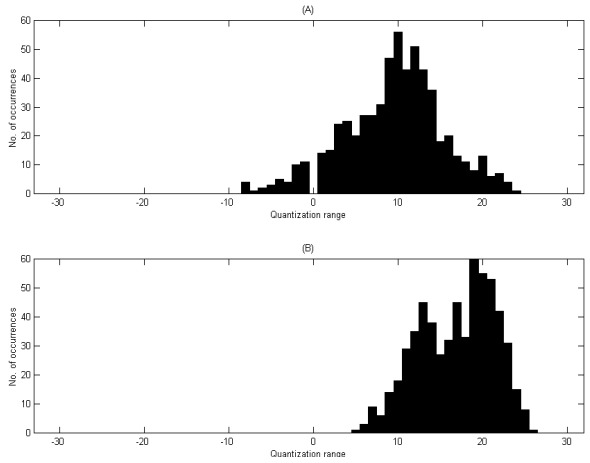


Figure 4: Histogram of LARc (2) for 600 frames of female speech.

Each of the 8 LARc's have a different dynamic range and distribution density and so are quantized to progressively smaller ranges. As the number of levels in the quantization range decreases, so too does the effect of the normalization on their distribution. LARc (3) and LARc (4) are quantized to a 32 level range but normalization does not affect the distribution of their values enough to allow a reduction in their quantization range. This is also true of the remaining LARc's.

4.2.2. Long-Term Prediction

The purpose of long-term prediction is to estimate the pitch component of the speech signal, and to produce a long-term residual signal or error signal that contains only information about the speech waveform that is not easily predictable. Referring again to Figure 1., prediction is made by delaying the reconstructed STR signal and multiplying it by a gain term. As such long-term prediction produces 2 parameters, a cross-correlation lag term (Nc) and a gain factor (bc). These are calculated and transmitted 4 times per frame. As might be expected normalization has little discernable affect on the gain term. The effect of normalization on the lag term however is not immediately apparent. Figure 5 is a histogram of the correlation lag term for a female speech sample from the TIMIT database. Figure 5 (A) depicts the histogram of the Nc parameter while (B) depicts the histogram of the Nc parameter after the signal has been normalized. It can be seen in Figure 5 (A) that many of the delays for normal female speech fall right on the lower limit of the delay range, suggesting that the range may not be adequate to represent all female pitch frequencies. This would not normally be a problem as in the event of the pitch frequency not being accommodated by the delay range; twice the pitch frequency would have the next highest correlation value and the lag corresponding to that would be used to generate the prediction of the STR. For a quasi-periodic signal this would still be a reasonable estimate.

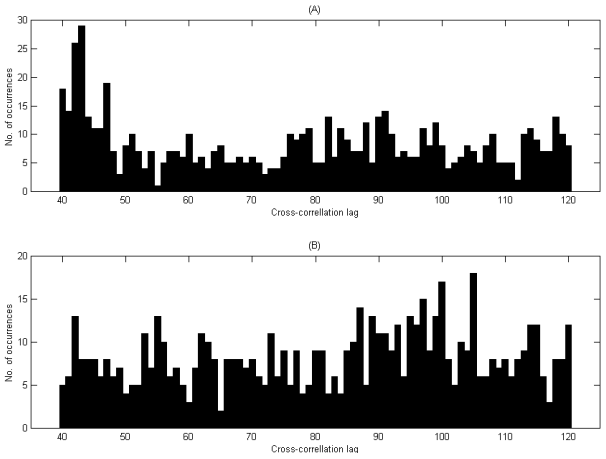


Figure 5: Histogram of correlation lag term.

Normalization however shifts the distribution of the lags down towards the male region (which is adequately represented by the range) as can be seen in Figure 5 (B). This means that the prediction of the short-term residual is more accurate, which in turn means that there is less energy in the long-term residual (LTR), since LTR = STR – predicted STR.

4.2.3. Regular Pulse Excitation

As was discussed above the spreading of the correlation lag term over more of its range in the case normalized female speech would suggest the production of a long-term residual signal with less energy in it. This is indeed the case. When the long-term residual of a female speech signal is compared to the residual of the same signal after it has been normalized a dramatic reduction can be seen in the overall amplitude of the normalized long-term residual. The reduction itself is speaker dependent but is normally of the order of two thirds. This is extremely significant as the RPE encoded parameters of the long-term residual signal make up over 70% of the bit stream (188 of the 260 bits transmitted per frame). Any small improvement in their representation would give rise to an extremely significant reduction in the overall bit rate.

5. CONCLUSIONS

Presented here is an overview of a voice gender normalization algorithm capable of mapping female speech to speech with male characteristics, which is then encoded on a GSM speech Codec. As a result of the Codec only having to process male speech it is possible to refine its operation. Specifically 100 less bits per second can be used to encode the first two filter parameters. Normalization also significantly reduces the amount of energy in the long-term residual signal. This is an important result as parameterization of the long-term residual signal forms the bulk of the transmitted bit stream, and a signal with less energy can be represented with fewer bits or alternatively reproduced more accurately using the same number of bits. Overall the savings made could be used to refine the Codec and increase the quality of encoding it provides or the saving could be passed directly onto the communications channel, either reducing the bit rate or allowing for extra error checking. Further work is needed to adapt the RPE algorithm to efficiently encode the reduced energy long-term residual signal and to incorporate the VGN algorithm into the Codec to take advantage of both algorithms requiring a linear prediction residual.

6. REFERENCES

[1] Marston, D. F., "Gender Adapted Speech Coding," *IEEE International Conference on Acoustics, Speech and Signal Processing*, 357-360, 1998.

[2] Lawlor, R., "Audio Time-Scale and Frequency-Scale Modification," Ph.D. Thesis, University College Dublin, 2000.

[3] Lawlor, B., Fagan, A., "A Novel Efficient Algorithm for Audio Time-Scale Modification," *Irish Signals and Systems Conference* 1999, NUI Galway, Ireland.

[4] GSM-06.10 "Full Rate Speech Transcoding," V. 8.1, 1999.

[5] Kroon, P., Deprettere, E. F., Sluyter, R. J., "Regular-Pulse Excitation – A novel Approach to Effective and Efficient Multipulse Coding of Speech," *IEEE Trans. ASSP*-34, 1054-1063, 1986.