A NOVEL APPROACH TO VOCODERS: SCALED SPEECH CODER (SSC)

Dr. Osman Eroğul[1] and Dr. Hakkı Gökhan İlk[2]

[1]Gülhane Military Medical Academy, Biomedical and Clinical Engineering Centre,

[2]Ankara University, Department of Electronics Engineering, Ankara, Turkey

oerogul@obs.gata.edu.tr

ABSTRACT

A novel approach to speech coding is introduced in this paper for the coding of speech signals used in vocoders. While conventional methods code the original input speech signal, the proposed algorithm codes the time-scale modified signal instead. In this method, the transmitter speeds up original input speech signal by a factor of *beta* using waveform similarity overlap-add-method (WSOLA). This signal is then coded using any of the low-bit rate speech coding algorithms. At the receiver side, the coded signal is first decoded to obtain the time-scale modified signal. This signal is then slowed by a factor of *1/beta* in order to reconstruct the original signal. The advantage stems from the fact that the time-scale modified signal requires half the coding time (when *beta*=0.5) than that of the original input speech.

Before the realisation of the proposed system, the performances of three different time-scale modification (TSM) algorithms were evaluated through a series of subjective listening tests. Furthermore, it has been demonstrated that the time-scale modification of speech signals, before coding, reduces as much as half of the bit rate over conventional methods and produces high quality speech.

1. INTRODUCTION

In many applications it is desirable to transform a speech waveform into a signal which is more useful than the original. For example, in time-scale modification speech can be sped up in order to compress the words spoken into an allocated time interval or to quickly scan a passage. As an application, full-duplex link can be achieved over a single-channel radio system if both ends of the link operate in what is known as Time Division Duplex Mode. In this mode the channel is allocated half of the time to transmit information in each of the two directions. That is, the radio channel is divided into time slots of T/2 seconds, with each end transmitting in alternate time slots stated simply when one end is transmitting, the opposite end is receiving the information and vice versa. In this way a single radio channel can support information flow in both directions resulting in a virtual full-duplex link [1]. Alternatively, the articulation rate can be slowed down to make degraded speech more intelligible. For example, in phonocardiography, the heart sounds can be slowed down to improve physician's capability in recognition and discrimination of dissimilarities resulting from cardiac disorders.

In this paper a novel approach, called *Scaled Speech Coder (SSC)*, is proposed in order to reduce the bit rate required to transmit the speech signal. The proposed method is particularly useful for the existing low bit-rate speech coding algorithms such as Mixed Excitation Linear Prediction [2] (MELP) vocoder because time-scale modification is performed as a pre and post-process at the transmitter and receiver respectively.

Numerous methods in both frequency and time domains have been attempted for the modification of speech waveforms. The key requirement, however, is that qualities such as naturalness and intelligibility as well as speaker dependent features such as pitch and formant structure, be preserved. One frequency domain approach is based on the sinusoidal representation that explicitly estimates the amplitude and phase of the vocal cord excitation and vocal tract system function contributions to each sine wave [3]. This approach is called Sinusoidal Analysis/Synthesis Method, SASM. Another frequency domain approach manipulates an excitation by deconvolving the original speech with a vocal tract spectral envelope estimate [4]. Time expansion is achieved by doubling the unwrapped phase of the spectrum. This approach is called Speech Transformation Without Pitch Extraction, STWPE.

There are also time domain algorithms for the time-scale modification of speech signals. One important example is the synchronised overlap-and-add procedure [5] (SOLA) which uses a modified overlap-add procedure on the speech waveform. Another form of synchronisation is obtained by applying a time domain pitch-synchronised overlap-and-add technique (TD-PSOLA) to the original waveform [6]. With TD-PSOLA the overlap-and-add procedure is performed pitch synchronously on the segments that are, accordingly, excised in a pitch synchronous way from an original speech signal. A modified version of TD-PSOLA is waveform similarity method (WSOLA) which ensures sufficient signal continuity that exists in the speech signal [7].

WSOLA is a powerful time-scale modification technique that is based on the waveform similarity overlap-and-add procedure. It performs better than other OLA algorithms because it does not require a pitch estimate and ensures maximal similarity at the segment joints by using waveform similarity measures. Some examples of similarity measures that can be successfully applied to WSOLA algorithm are discussed in the next section.

2. THE WSOLA ALGORITHM

The basic synthesis equation used by the WSOLA algorithm is:

$$y(n) = \frac{\sum_k v(n - L_k).x(n + \tau^{-1}(L_k) + \Delta_k - L_k)}{\sum_k v(n - L_k)}$$

where *v(n)* is the square of a windowing function, L_k represents the consecutive window positions, i.e. the synthesis instants, and $\tau^{-1}(L_k)$ represents an analysis instant. WSOLA seeks to find a segment that will overlap-add with the previous segment, which lies within the prescribed tolerance interval around the synthesis instant. The position of the best segment, *m*, is determined by finding the value $\Delta = \Delta_m$ lying within a tolerance region $[-\Delta_{max,...,}\Delta_{max}]$ around the analysis instant and maximises the cross-correlation coefficients between the previous segment and the segment under consideration.

By choosing regularly spaced synthesis instants $L_k = k.L$ and a symmetric window such that $\sum_k v(n - kL) = 1$,synthesis equation simplifies to

$$y(n) = \sum_k v(n - kL).x(n + \tau^{-1}(kL) - kL + \Delta_k)\ .$$

N representing the window length, some examples of similarity measures that can be applied successfully are:

i) A cross-correlation coefficient

$$c_c(m,\delta) = \sum_{n=0}^{N-1} x(n + \tau^{-1}((m-1)L) + \Delta_{m-1} + L).x(n + \tau^{-1}(mL) + \delta)$$

ii) A normalised cross-correlation coefficient

$$c_n(m,\delta) = \frac{c_c(m,\delta)}{\left(\sum_{n=0}^{N-1} x^2(n + \tau^{-1}(mL) + \delta)\right)^{1/2}}$$

iii) A cross-AMDF coefficient

$$\sum_{n=0}^{N-1}\left|x(n + \tau^{-1}((m-1)L) + \Delta_{m-1} + L) - x(n + \tau^{-1}(mL) + \delta)\right|\ .$$

All these three similarity measures have been applied to the WSOLA algorithm [7]. Our subjective listening tests indicated that normalised cross-correlation coefficient based WSOLA algorithm performs better than the other two. Although the non-linear square root operation is complicated and therefore requires more processing time, the subjective output speech quality is superior. Informal subjective listening tests have been conducted in order to evaluate the performances of the three different time-scale modification algorithms. Normalised cross-correlation coefficient based WSOLA technique was compared with two frequency domain approaches of SASM and STWPE algorithms. The details and results of the listening tests conducted are given in the next section.

3. LISTENING TEST RESULTS

3.1. Test Procedure for TSM Algorithms

Three different test procedures were used in order to evaluate the performances of the three different TSM algorithms. In order to evaluate the intelligibility of the reconstructed speech diagnostic rhyme test (DRT) was used in the first test. To assess the speech quality, mean opinion score (MOS) test was used while the degradation mean opinion score (DMOS) test was used in order to measure the degradation in the quality of the reconstructed speech with respect to the reference.

The DRT uses a corpus of words, 232 words in 116 rhyming pairs. In a given instance, one word of the pair is presented and the listener is asked to determine which word was spoken. The two words of each pair, for instance "Bob", "Gob", differ only in one attribute of the first consonant. So a correct response from the listener indicates that the speech processing system under examination preserves that attribute. Source speech samples for the DRT were obtained from three male and three female speakers [8]. In all cases the sampling frequency was 8kHz. Six different filesets, containing the same words but spoken by different speakers were used. A fileset corresponds to a single speaker. Speech sample duration in the filesets is about 5 minutes.

In MOS procedure, one sentence is presented on each trial and the listener is asked to rate the sample according to the absolute scale, ranging between 1 and 5. The quality scale ranges from "bad" for grade 1 to "excellent" for grade 5. The drawback of this procedure is that, ceiling and floor effects may obscure real differences in performance. To overcome this limitation, the degradation mean opinion score test was also used.

In DMOS procedure two samples are presented in each trial, a reference sentence and a test sentence. Listeners are asked to rate the quality of the second sample relative to the quality of the first. The quality scale ranges from "very much poor quality" for grade 1 to "the same or better quality" for grade 5.

Source speech files for the MOS and the DMOS tests were obtained from six males and six females. These sources had sampling frequencies of 8 kHz [9]. The MOS and the DMOS use Harward type-sentences. Two sentences, one spoken by a male and the other by a female, are used in each sample separated by a short silence. Sample durations are between 6 to 9 seconds.

Twenty-eight subjects were accessed for this study. Listeners were drawn from the Communications Research Centre (CRC), Ottawa, Canada, and from ordinary people. Many of these subjects drawn from the CRC are familiar with the testing procedure and know much about speech technology. On the other hand, subjects drawn from ordinary people have no experience with speech evaluation.

To emphasise the effects of algorithms on speech signals, quiet speech samples were first compressed by a factor of

0.5 and then expanded by a factor of 2 in order to recover the original signal. Later, the reconstructed signals using the WSOLA algorithm, the STWPE algorithm, and the SASM algorithm were recorded on audio tapes.

In the DMOS test a reference sample processed through the WSOLA algorithm was presented first on each trial followed by the identical sample processed through the other two TSM algorithms (STWPE, SASM). All subjects judged each of the TSM algorithms with different speakers.

3.2 Experimental Results for TSM Algorithms

In the DRT, it was found out that the WSOLA algorithm almost preserves the intelligibility of the reconstructed speech. The mean value and standard deviation of each algorithm's score for the DRT are given in Table-I.

Table 1. The DRT results

TEST	**DRT**		
ALGORITHM	WSOLA	STWPE	SASM
STND. DEV.	0.84	0.62	0.74
AVERAGE %	92	92	90

The MOS test demonstrated that WSOLA algorithm preserves the quality of the modified speech over the other two algorithms. The DMOS test also demonstrated that the WSOLA algorithm degrades the quality of the modified speech gracefully. The average score for each algorithm is given in Table-II. Subjective evaluation test results show that the WSOLA increases the intelligibility of the reconstructed speech signals while almost preserving the quality of the modified signal.

Table 2. MOS and DMOS test results

Test	**Algorithm**	**Average**
MOS	WSOLA	4.03
	STWPE	1.08
	SASM	2.80
DMOS Ref: WSOLA	STWPE	1.69
	SASM	2.75

4. MIXED EXCITATION LINEAR PREDICTION VOCODER

The MELP vocoder is based on the traditional Linear Predictive (LPC) parametric model, but also includes mixed excitation. The backbone of the Scaled Speech Coder (SSC) is the MELP vocoder, which has three main features.

1. *A reliable pitch estimate.* The MELP coder is based on representing short-term voiced speech as the summations of sinusoids. Therefore the new Federal Standard heavily relies on a robust pitch estimate. The pitch estimation procedure involves integer, fractional and final pitch calculation, pitch doubling check and average pitch update algorithms. Furthermore, all extracted speech parameters are interpolated pitch synchronously in the decoder.
2. *Parameter interpolation.* The extracted speech parameters via encoder, i.e. Fourier Magnitudes and pitch of the excitation signal and gain, Line Spectral Frequencies (LSFs), jitter and filter coefficients of the shaping filters are interpolated, during synthesis, in order to ensure smooth evaluation in the characteristics of the synthesised speech [2].
3. *Mixed Excitation.* The harmonic (voiced) and noise (unvoiced) speech components are synthesised separately. In order to separate the harmonic and noise components, five frequency bands are defined and each band is declared as Voiced/Unvoiced. The synthesised pulse and noise excitation are then filtered and summed to form the mixed excitation.

5. SCALED SPEECH CODER (SSC)

Obtaining high quality decoded speech at low bit rates (<4.0 kb/s) has been the focal point of considerable research activity in the last decade. The US Department of Defence Digital Voice Processing Consortium (DoD-DVPC) selected Mixed Excitation Linear Predictive Vocoder (MELP) [2] as the recommended new Federal Standard. This new standard provides equal or improved performance over the 4.8 kb/s CELP coder at only 2.4 kb/s.

In fact an increasing demand for digital speech coding applications made further development of low bit rate speech coding systems inevitable. In this paper, a novel approach to reduce the bit rate required to transmit the speech signal is proposed. Scaled Speech Coder (SSC) is particularly useful for the existing low bit-rate speech coding algorithms, such as MELP, Prototype Waveform Interpolation [10] and Sinusoidal Transform Coders [11], because time-scale modification is performed as a pre and post-process at the transmitter and receiver respectively. Figure 1 illustrates the overall block diagram of the SSC.

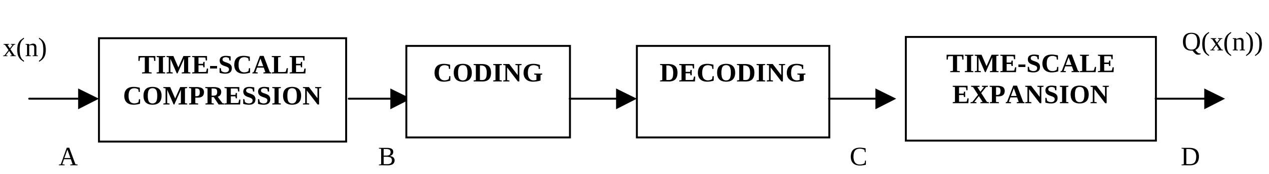

Figure 1. The overall block diagram of the *Scaled Speech Coder (SSC)*

The time-scale compressed Signal B, obtained using WSOLA, (preserving the pitch and formant structure of the original input speech) is then coded using MELP algorithm at 2.4 kb/s bit rate. At the receiver, the signal is first decoded according to MELP decoder algorithm. The decoded Signal C is then slowed down in order to obtain the reconstructed speech Signal D. The proposed scaled speech coder (SSC), produces high communication quality speech at half the bit rate of the conventional low bit-rate methods.

Although SSC incorporates Mixed Excitation Linear Predictive Vocoder for coding purposes, any low bit-rate speech coding algorithm is suitable for the proposed Scaled Speech Coder. It is evident, from our experiments and literature, that WSOLA algorithm is more suitable for time-scale purposes than other approaches. In the same way, MELP algorithm gives equal or better performance at 2.4 kb/s than other low bit rate speech coding applications. The performance of the MELP vocoder with other standard coders is assessed in Reference [12] thoroughly.

6. SCALED SPEECH CODER IMPLEMANTATION AND SIMULATION RESULTS

The proposed SSC has been computer simulated and its performance was evaluated using sentences spoken by a male and a female [9]. Figure 2 illustrates the original speech signal spoken by an English male speaker. At the transmitter, this signal (Signal A) is sped up by a factor of 0.5 using the WSOLA algorithm to obtain Signal B in Figure 3. This signal is then coded using the MELP vocoder at 2.4 kb/s standard bit rate. At the receiver, the compressed and coded signal is then decoded to obtain the reconstructed compressed Signal C, which is illustrated in Figure 4. Please note that the duration of Signal C is equivalent to Signal B and it is half of the duration of Signal A. Finally, Signal C is slowed by a factor of 2 to reconstruct the original speech Signal A. Figure 5 shows reconstructed Signal D of the original speech signal given in Figure 2.

It is possible to obtain different variants of the Scaled Speech Coder in conjunction with other TSM algorithms and/or other low bit-rate speech coding applications. The time-scale modified Signal B requires half the coding time than that of the original input speech Signal A, which means that it reduces the half of the bit rate. This bit rate can be further reduced using the desired time-scale modification factor.

Another useful approach, in a Variable Bit Rate Coding system, is to use a Voice Activity Detector (VAD) to discriminate silence from speech activity [13]. In this way, the average bit rate could be reduced further by compressing the voiced regions only.

Informal listening test results for the SSC, clearly indicated that the proposed system is capable of delivering high communications quality. The synthesised speech quality degrades gracefully and it is similar to that of MELP vocoder output at 2.4 kb/s. Although an increase in tonal artefacts is observed, the output speech sounds natural and speaker identification is preserved. This is mainly because the pitch and formant structure is well preserved by the WSOLA algorithm.

7. CONCLUSIONS

A novel approach to speech coding was introduced in this paper for the coding of speech signals used in vocoders. While the conventional methods code the original input speech signal, the proposed SSC algorithm codes the time-scale modified speech signal instead. WSOLA algorithm was used for time-scale modification while the new Federal US Standard MELP coder was employed for coding purposes.

Listening tests and comparisons between time-scale modification algorithms were also conducted in order to choose the best time-scale modification method. It has also been found out that a normalised cross-correlation coefficient obtained via a non-linear process performs better than other similarity measures.

The proposed scaled speech coder (SSC) produces high communication quality speech at only half the bit rate of the MELP vocoder. In addition, SSC does not require modifications on the low bit-rate speech coding systems. Time scale modification algorithms appear as cascaded blocks in order to compress and expand speech signals in the encoder and decoder respectively.

ACKNOWLEDGEMENTS

The authors wish to thank Dr. Nur Serinken of the communications Research Centre (CRC), Ottowa, Canada for his support in this study.

REFERENCES

[1] N. Serinken, B. Gagnon and O. Eroğul, "*Full-Duplex Speech for HF Radio Systems*", IEE HF Radio Systems and Techniques Conference Publications, No.411, 1997, pp.281-284.

[2] L.M. Supplee, R.P. Cohn, J.S. Collura and A.V. McCree, "*MELP: The New Federal Standard at 2400 BPS"*, ICASSP 97, pp. 1591-1594.

[3] T.F. Quatieri and R.J. McAulay, "*Speech transformation Based on a Sinusoidal Representation"*, Technical Report, Lincoln Laboratuvary MIT, Lexington, Massachusetts, 1986, pp. 1-56

[4] S. Seneff, "*System to Independently Modify Excitation and/or Spectrum of Speech Waveform Without Explicit Pitch Extraction"*, IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. ASSP-30, No. 4, 1982, pp. 566-578.

[5] S. Roucos and A. M. Wilgus , "*High Quality Time-Scale Modification for Speech"*, IEEE Int. Conf. Acoust.,Speech, Signal Process., ICASSP-85, pp. 493-496.

[6] E. Moulines and F. Chanpentier, "*Pitch-Synchronous Waveform Processing Techniques for text-to-Speech Synthesis Using Diphones*", Speech Communication, Vol. 9, 1990, pp. 453-467.

[7] W. Verhelst and M. Roelands *"An Overlap-add Technique Based on Waveform Similarity (WSOLA) For High Quality Time-Scale Modification of Speech."*, ICASSP-93, Vol. 2, pp. 554-557.

[8] Speech Corpus Used For the Diagnostics Rhyme Test (DRT), Communications Research Centre (CRC), Ottowa, Canada, 1997.

[9] NIST Speech Acoustic-Phonetic Continuous Speech Corpus, TIMIT Disc1-1.1, US Department of Commerce, National Institute of Standards and Technology, Gaithersburg, MD.

[10] W.B. Kleijn, "*Continuous Representations in Linear Predictive Coding*", ICASSP 91, pp. 201-204

[11] R.J. McAulay and T.F. Quatieri, "*Low Rate Speech Coding Based on Sinusoidal Model*", Advances in Speech Signal Processing edited by S. Furui and M.M. Sondhi, Marcel Dekker, 1991, pp. 165-207.

[12] M.A. Kohler, "*A Comparison of the New 2400 bps MELP Federal Stand with Other Standard Coders*", ICASSP 97, pp. 1541-1544

[13] E. Aksu, E. Ertan, H. İlk, H. Karcı, Ö. Karpat, T. Kolçak, L. Şendur, M. Demirekler, E. Çetin, *"Implementation of a Variable-Bit Rate MELP Vocoder on TMS320C548"*, The Second European DSP Education and Research Conference, 1998, pp. 67-71.

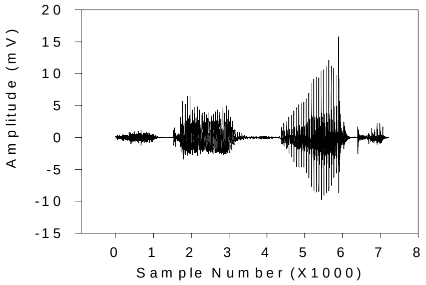

Figure 2. Original Signal A

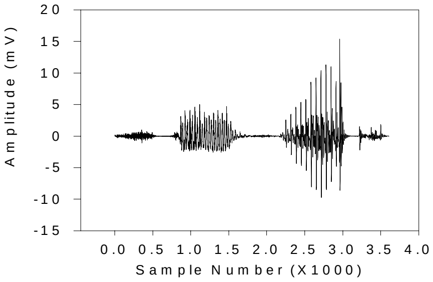

Figure 3 Compressed version of the original signal by a factor of 0.5, Signal B

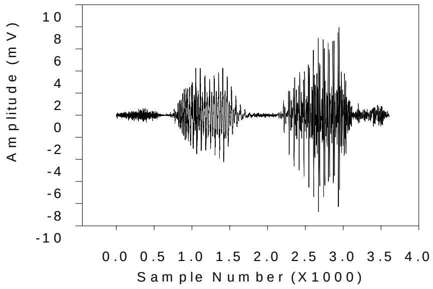

Figure 4 Decoded Signal C

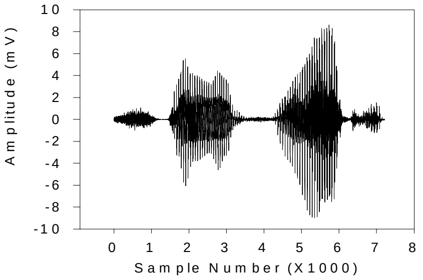

Figure 5. Reconstructed speech signal, after time expansion of the decoded Signal C, Signal D