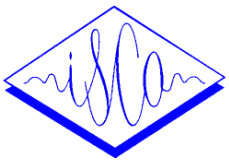

4th European Conference on
Speech Communication and Technology
EUROSPEECH '95
Madrid, Spain, September 18-21, 1995

VERY LOW-BITRATE SPEECH CODING USING PERCEPTUALLY-DERIVED SPECTRAL DATA

D. Tsoukalas, J. Mourjopoulos, and G. Kokkinakis
Wire Communications Laboratory
University of Patras, PATRAS 265 00
GREECE
email: tsoukala@wcl.ee.upatras.gr

ABSTRACT

A new family of very low bitrate speech coders employing models of human perception is presented. The coding methodology is based on non-linear modulation of a random broadband noise source with signals derived from speech, following two main strategies for representing coded speech: one using the minimum audible difference between the original and the modulated speech signals, and another using a minimum log-error criterion along with some perceptually derived harmonic information. Depending on the methodology and the degree of accuracy employed for coding several implementations are allowed starting from 1 kb/s. High intelligibility is achieved even for the lower bitrate implementations, although, some increase in the bitrate is required for high-quality speech.

1. INTRODUCTION

The problem of coding speech at very low-bitrates has been investigated over the last 20 years and significant results have been reported [1],[2] which allow efficient coding below 2400 b/s. Given, however, that intelligibility scores for such coders are in the order of 90 % [1],[3], this area remains largely open for further investigation. Of main interest are also quality and bitrate reduction improvements.

Low bitrate coding has a number of applications, starting from military communications, where, very low bitrates, usually below 1 kb/s, are required [3], to the cellular telephony where the high-quality speech requirement imposes an increase in the bitrate. Early efforts on speech coding were based on the vocoder techniques [4], while the majority of newer techniques, are based on parametric speech models such as LPC [5], which later evolved into Code Excited Linear Prediction (CELP) [6]. Apart from these techniques, a number of different approaches have been developed such as Sinusoidal Transform Coding (STC) [7], Multiband Excitation Coding (MBE) [8],[9] and methods based on HMMs [10]. Recently, however, new approaches have been developed, applied mainly to broadband audio coding, based on models of human perception [11]. Such approaches are attractive, since it is well known that the human auditory system performs significant data reduction based on the cochlea and other central neural system mechanisms.

Following such psychoacoustic-based approaches, a novel family of speech coding techniques is presented here. The basis of the proposed approach is the modulation of a random noise source using a parametric speech model, previously applied to speech enhancement [12]. According to the proposed approach, the noise signal is optimally modulated in specific frequency bands so that the audible difference between coded and original speech is minimized. An alternative approach, also presented, is based on optimal modulation by minimization of a log-error function in conjunction with sufficient perceptually-derived harmonic information extracted from the clean speech signal. The modulation parameters are used for speech reconstruction employing the same noise signal combined with the harmonic speech information. The first approach, is capable of producing intelligible speech at very low bitrates, while the second produces high-quality speech at somewhat increased bitrates.

2. PERCEPTUAL MODELS FOR PARAMETRIC FREQUENCY DOMAIN CODING

2.1. Speech Representation model

Let the speech signal be given by x(n), and a random broadband stationary noise signal be given by d(n). The Short-Time Fourier Transforms of these signals will be given by:

$$X_w(k,i)=\sum_{n=0}^{K-1} x(n+off_i)w(n)l_K^{kn}, \quad 0\le k\le K-1 \qquad (1)$$

$$D_w(k,i)=\sum_{n=0}^{K-1} d(n+off_i)w(n)l_K^{kn}, \quad 0\le k\le K-1 \qquad (2)$$

where $l_K^{kn}=e^{-j(2\pi kn/K)}$, $w(n)$ is a window function, K is the length of the Fourier transform, k and i are the frequency and time-domain indices, and, off_i is an offset, assuming that speech is transformed using overlapping time windows. The power spectra of these quantities will be given by the square modulus of the Fourier Transform, i.e.:

$$X_p(k,i)=|X_w(k,i)|^2, \quad 0\le k\le K-1 \qquad (3)$$

$$D_p(k,i)=|D_w(k,i)|^2, \quad 0\le k\le K-1 \qquad (4)$$

As was proposed in [12], and then used in [13], a robust model for speech representation is given by the non-linear parametric expression:

$$\hat{X}_p(k,i)=\frac{Y_p(k,i)}{\alpha(k,i)+Y_p(k,i)}Y_p(k,i),\quad 0\le k\le K-1 \qquad (5)$$

where, $Y_p(k,i)$ is a signal based on a combination of the speech and noise signals $X_p(k,i)$ and $D_p(k,i)$, and $\alpha(k,i)$ is a time-frequency dependent modulation parameter. Apparently, if estimation of $\alpha(k,i)$ is based on frequency components, then no data reduction can be performed, since one parameter per frequency bin and time window will be needed. Therefore, it is assumed that $\alpha(k,i)$ is constant over selective frequency bands.

2.2. Speech Coding using the Auditory Masking Threshold

Let $T_b(i)$, denote the Auditory Masking Threshold (AMT) [14] of the speech signal $X_p(k,i)$, where it is assumed that this function is constant and independent of frequency within a critical band b and a data-window [15]. The AMT defines a frequency domain pattern below which all frequency components are masked (i.e. are made inaudible), in the presence of a masker signal. This has successfully been used in audio coding applications [15],[16]. The audible spectra of the speech and noise signals can be now defined by using the *max*{} operator, i.e. by taking the maximum between the corresponding signal and the AMT. Such a function, defines the perceptually-significant spectrum, i.e. those components that contribute to audible signal information. The difference between audible speech and audible noise defines the audible difference between speech and "speech-like" modulated noise. Minimization of this function results in the following estimate for the parameter $\alpha(k,i)$ [12]:

$$\alpha(k,i)=\alpha_b(i)=D_{pb}+\frac{D_{pb}^2}{T_b(i)},\quad k_{lb}\le k\le k_{hb} \qquad (6)$$

where, it was assumed that $\alpha(k,i)$ is constant within frequency band b (bounded by k_{lb} and k_{hb}), and D_{pb} is the mean power spectrum of the noise signal in band b. According to this formula, the speech signal can be reconstructed using eq. (5), with $Y_p(k,i)=D_p(k,i)$, and the value of $\alpha_b(i)$ given by the above expression.

This finding suggests that efficient speech coding can be achieved by using as many parameters as the number of critical bands, namely the values of the AMT. Although, it was suggested to encode these values using 2750 b/s for a 16 kHz sampling rate frequency speech signal [12], it will be shown in the next section that the required bitrate can become as low as 1 kb/s.

2.3. Speech coding using a harmonic excitation signal

The previously described method is successful in producing intelligible speech, as will be shown in section 3. Nevertheless, it lacks the harmonic information which makes coded speech sound pitchless. To overcome this problem, a harmonic speech signal was defined in [13] by:

(a) Extracting all the tonal components of the speech signal per data-window using the methodology adopted in the ISO/IEC audio coding standard [16],
(b) Rejecting those tonal components below the AMT of the signal and below the Absolute Auditory Threshold [14],
(c) Employing an algorithm for keeping only harmonically-structured tonal components. In this way, sparse tonal components are also rejected.

This harmonic signal was then combined with the noise signal to produce $Y_p(k,i)$, of eq. (5). Combination is done in such a way, so that harmonic peaks of the speech signal appear 7-10 dB higher in the combined signal $Y_p(k,i)$. The parameters $\alpha_b(i)$, are then extracted by minimization of the log spectral distance [17] between coded (by eq. (5)) and clean speech. This procedure results in the following expression:

$$\alpha_b(i)=\left[\prod_{k=k_{lb}}^{k_{hb}}\alpha(k,i)\right]^{1/(k_{hb}-k_{lb}+1)},\quad 0\le b\le B-1 \qquad (7)$$

where $\alpha(k,i)$, follows directly from eq. (5) with $\hat{X}_p(k,i)=X_p(k,i)$.

3. EVALUATION

In this section three implementations of the proposed speech coders are evaluated. These coders were implemented as follows:

(a) CODER-A is a direct implementation of eq. (6) and (5), i.e. is based on the AMT of the speech signal. The modulation parameters $\alpha_b(i)$, were obtained for B=22 critical bands [14], due to the 16 kHz sampling rate of the speech signal.
(b) CODER-B was based on modulation of the harmonic excitation signal, i.e. eq. (7) and (5). Again, B=22 critical bands were employed.
(c) CODER-C follows a similar implementation to that of CODER-B, except that the spectrum was divided into B=33 bands, i.e. below 3.5 kHz two modulation parameters were used for every critical band, while one such parameter was used above 3.5 kHz.

The block diagrams for the proposed coders and decoders are shown in Fig. 1 and Fig. 2. Note that the harmonic information modules must be omitted in the case of CODER-A.

Parameter coding for the proposed coders results in variable bitrate implementations [13]. Therefore, the mean bitrate was measured using speech material

from more than 30 speakers from the SAM-A database [18]. Results are shown in TABLE I. From this table it can be observed that CODER-A requires about 1000 b/s, CODER-B requires about 1800 b/s, while CODER-C requires about 2700 b/s. However, for CODER-B and CODER-C, 400 b/s approximately are required for harmonic information.

Evaluation of the proposed coders concerning intelligibility and quality was performed using several subjective tests. CODER-A was evaluated using the Diagnostic Rhyme Test [19], with both English Language speech data (E-DRT) and Greek Language speech data (G-DRT), and the Semantically Unpredictable Sentences (SUS) test [20] with Greek Language speech data. CODER-B and CODER-C were evaluated using the DRT with Greek Language speech data and also the Mean Opinion Score (MOS) test. Results for these test are shown in TABLE II and III.

As can be observed in these tables, CODER-A is capable of producing intelligibility scores of up to 85% for G-DRT, while this score is lower for the E-DRT (72%) and the SUS test (76%). The rest of the coders are capable of more than 92% (CODER-B) and 94% (CODER-C). The quality of these coders was also assessed and it was found that MOS scores of up to 3 can be achieved (for CODER-C).

It may be concluded, that the proposed speech reconstruction methodologies for speech coding are effective at low and medium bitrates. In fact, it was found that even higher-quality speech could be reconstructed at even higher bitrates (i.e., above 3kb/s).

TABLE I
BITRATES FOR THE PROPOSED CODERS

	CODER-A			CODER-B			CODER-C		
	Male	Female	Mean	Male	Female	Mean	Male	Female	Mean
Modulation data	980.4	1040.9	1010.7	1352.9	1495.4	1424.2	2211.3	2399.7	2305.5
Harmonic data	-	-	-	481.9	347.2	414.6	481.9	347.2	414.6
Total	980.4	1040.9	1010.7	1834.8	1842.6	1838.8	2693.2	2747.5	2720.1

TABLE II
DRT AND SUS TEST SCORES AND STANDARD ERROR (S.E.) FOR CODER-A

	E-DRT	G-DRT	SUS
Score (%)	72.22	85	76.36
S.E.	9.3	10.5	11.8

TABLE III
DRT AND MOS TEST SCORES AND STANDARD ERROR (S.E.) FOR CODER-B AND CODER-C

	CODER-B			CODER-C		
	Male	Female	Total	Male	Female	Total
DRT Score (%)	91.1	93.7	92.4	93.7	94.9	94.3
DRT S.E.	1.815	0.619	0.994	1.408	0.331	0.717
MOS Score	2.75	2.67	2.71	3.13	2.88	3.01
MOS S.E.	0.144	0.220	0.119	0.073	0.145	0.092

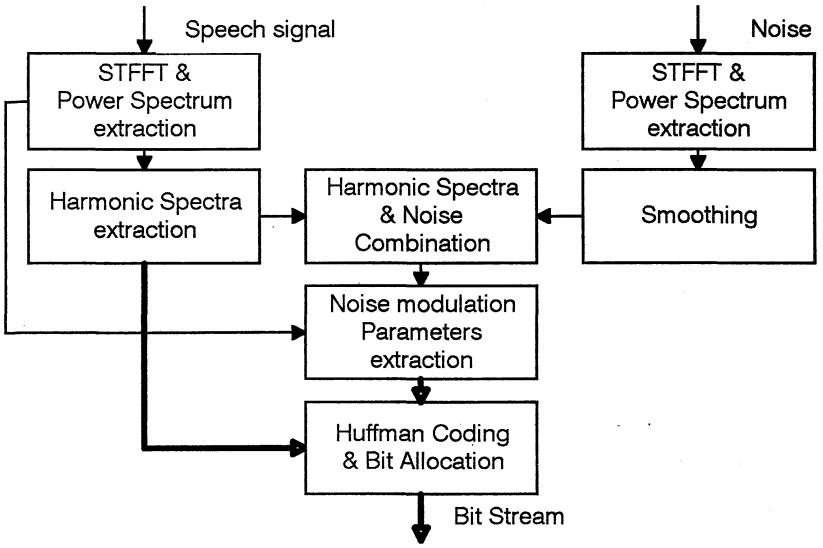

Fig.1. Block diagram for the proposed coder.

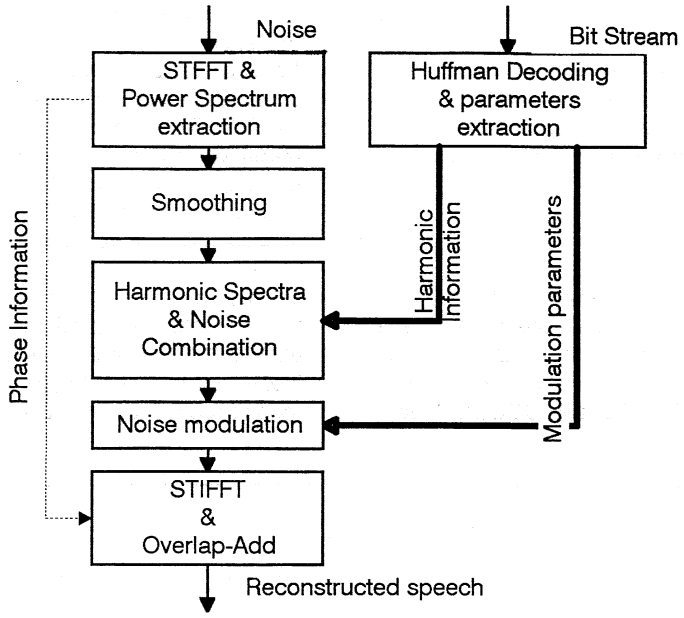

Fig. 2. Block diagram for the decoder.

4. CONCLUSIONS

A new family of speech coding techniques was presented. The techniques are based on non-linear modulation of a random broadband noise source. Optimum modulation can be achieved by either minimization of the audible difference between original and coded speech or by a log-error minimization procedure using a harmonic excitation signal. Due to parametric coding, variable bitrate implementations of the proposed coders can be achieved starting from 1 kb/s where intelligibility scores of up to 85% can be achieved, while at higher bitrates intelligibility scores better than 94% can be achieved (for 2700 b/s). At such high bitrates, the high-quality speech was also confirmed by the MOS test.

REFERENCES

[1] S. Spanias, "Speech Coding: A Tutorial Review", *Proc. IEEE*, vol. 82, pp. 1541-1582, Oct. 1994

[2] A. Gersho, "Advances in Speech and Audio Compression," *Proc. IEEE,* vol. 82, pp. 900-918, June 1994

[3] T. E. Tremain, et al, "Evaluation of Low Rate Speech Coders for HF," *in Proc. IEEE ICASSP,* pp. II163-166, Apr. 1993

[4] R. Steele, and L. Cassel, "Dynamic encoding as applied to a channel vocoder," *J. Acoust. Soc. Amer.,* vol. 35, pp. 789, 1963

[5] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE,* vol. 63, pp. 651-580, Apr. 1975

[6] M. R. Schroeder, and B. Atal, "Code-Excited Linear Prediction (CELP): High Quality speech at very low bit rates," *in Proc. IEEE ICASSP,* pp. 937, Apr. 1985

[7] R. McAulay, and T. Quatieri, "Speech analysis based on a sinusoidal representation," *IEEE Trans. ASSP,* vol. 24, pp. 744, Aug. 1986

[8] D. Griffin, and J. Lim, "Multiband Excitation Vocoder," *IEEE Trans. ASSP,* vol. 36, p. 1223, Aug. 1988

[9] J. Hardwick, and J. Lim, "The application of the IMBE speech coder to mobile communications," *in Proc. IEEE ICASSP,* pp. 249-252, Apr. 1991

[10] E. P. Farges, and M. A. Clements, "Hidden Markov Models applied to very low bit rate speech coding," *in Proc. IEEE ICASSP,* pp. 433-436, 1986

[11] N. Jayant, J. Johnston, and R. Safranek, "Signal Compression Based on Models of Human Perception," *Proc. IEEE,* vol. 81, p. 1385, Oct. 1993

[12] D. Tsoukalas, J. Mourjopoulos, and G. Kokkinakis, "Speech Enhancement based on Audible Noise Suppression (ANS)," submitted for publication.

[13] D. Tsoukalas, J. Mourjopoulos, and G. Kokkinakis, "Low Bitrate Speech Coding by Perceptually-Optimized Noise Excitation Modulation," submitted for publication.

[14] E. Zwicker and H. Fastl, *Psychoacoustics, Facts and Models*, Springer-Verlag Berlin Heidelberg 1990

[15] J. D. Johnston, "Transform Coding of Audio Signal using Perceptual Noise Criteria", *IEEE Jour. Select. Areas in Commun.,* vol. 6, pp. 314-323, Feb. 1988

[16] ISO/IEC IS11172-3, "Coding of moving pictures and associated audio for digital storage media at up to 1.5 Mbits/s - Audio Part", Nov. 1992

[17] R. M. Gray, et al, "Distortion measures for speech processing", *IEEE Trans. ASSP,* vol. ASSP-28, pp. 367-376, Aug. 1980

[18] Speech Technology Assessment in Multilingual Applications (SAM-A), *ESPRIT Project 6819,* 1992

[19] S. Meister, "The Diagnostic Rhyme Test (DRT): An Air Force Implementation", RADC-TR-78-129, AD-A060917, 1978

[20] Benoit C., and M. Grice, "A manual for the SUS test: a unified methodology for multilingual text-to-speech synthesis assessment at the sentence level", ESPRIT PROJECT 2589 (SAM), Ref. No. SAM-ICP-UCL-001, April 1991