Proc. IEEE Speech Coding Workshop (Porvoo, Finland), pp. 135-137, June 1999

AN IMPROVED BACKGROUND NOISE CODING MODE FOR VARIABLE RATE SPEECH CODERS

Khaled El-Maleh   Peter Kabal

Dept. Electrical & Computer Engineering
McGill University, Montreal, Quebec, Canada H3A 2A7
Voice: (514) 398-7130, Fax: (514) 398-4470
Email: {khaled,kabal}@tsp.ece.mcgill.ca

ABSTRACT

In this paper, we present a novel background noise coding scheme for variable rate speech coders. Existing approaches to noise coding at very low bit rates (i.e. below 1 kbps) fail to faithfully reproduce background noise resulting in a degradation of the overall perceptual quality. In our approach, classification of the noise type is used to select the type of excitation to be used at the receiver. To illustrate the benefits of our scheme, we have modified the noise coding mode of the CDMA enhanced variable rate codec (EVRC) to include the proposed class-dependent noise excitation model. Evaluation tests have shown that we have improved the overall quality with the proposed noise coding scheme without an increase in bit rate.

1 INTRODUCTION

A typical telephone conversation contains approximately 40% active speech bursts and 60% silence and non-speech sounds. Variable bit rate (VBR) coding can exploit non-speech activity to reduce the average bit rate and to increase system capacity [1]. Reduction of the bit rate in a VBR coder is achieved by using a lower rate coding mode for the non-speech periods. In noisy environments, the speech gaps are filled with background environment sounds such as those in a car, restaurant, street, and office.

The overall quality of a telephone conversation using VBR coding depends on three major components: speech coder, noise coder, and voice activity detector (VAD). In the present generation of VBR speech coders, a simple excitation-filter model is used for noise synthesis [1], [2], and [3]. Existing noise coding algorithms, with bit rates below 1 kbps, fail to naturally reproduce many of the acoustic background noise commonly encountered in mobile environments (i.e. restaurant, bus, shopping center, and airport). The change in the character of the noise between talk spurts and speech pauses is noticeable and can be annoying [4]. The ITU-T Study Group 12 (Question 17 "Noise aspects in evolving networks") is currently studying the effect of background noise in voice communications systems [5]. Moreover, some speech coders fail to handle background noise resulting in unnatural-sounding artifacts to the listener [6]. Recently, Beritelli [4] has proposed a multi-mode noise coding scheme with bit rates from 0–4.9 kbps as part of a CS-ACELP VBR coder.

In this paper, we present an improved background noise coding mode for variable bit rate speech coders. The novel part of the scheme is the use of noise classification and residual substitution to achieve natural-quality reconstruction of acoustic environment sounds at very low bit rates.

The paper is organized as follows. Section 2 reviews background noise coding based on linear prediction synthesis models. In Section 3, we present a novel background noise coding scheme using class-dependent residual substitution. Implementation and evaluation of the coding scheme are presented in Section 4. Finally, conclusions and future work are discussed in Section 5.

2 LINEAR PREDICTION-BASED NOISE CODING

A common scheme to encode background noise signals at very low bit rates (below 1 kbps) is to excite a low-order linear prediction (LP) synthesis filter with a white Gaussian noise (WGN) matching the background noise residual energy. The few bits available for each frame are allocated to quantize the LPC spectral parameters and the residual energy. The residual waveform is not encoded, instead a random number generator is used at the receiver. As an example, we show in Table 1 the bit allocation of the 800 bps noise coding mode in CDMA VBR coders (QCELP, Q13, and EVRC) [2], [7], and [8]. We have studied the performance of the above noise coding model for different types of background noise common in mobile environments (i.e. car, street, restaurant, bus, and office). From our evaluation, we found that the above model works well for some noises (i.e. car, computer fan, and traffic), but it fails to naturally reproduce other sounds like restaurant, bus, and shopping center noise (even with an unquantized LPC model). The main reason is that the WGN excitation is a poor model for the LP residual of some noises.

As we can not afford to directly encode the residual waveform, we have investigated different ways to parameterize the noise LP residual using only few bits. A slight improvement in quality was gained by using the residual time envelope to modulate the WGN excitation [3]. We have observed that some noises have non-flat residual amplitude spectra and thus using a white excitation is not proper. We represented the residual amplitude spectrum with a low order (i.e. 18) critical band gain vector and used a random phase spectrum. With this spectral excitation model, the quality improved slightly but with some muffling. We have identified that the residual Fourier phase contributes significantly to preserving naturalness in the synthesized noises. This agrees with the conclusion reported in [9] that the phase spectrum of the LP residual determines the temporal waveform and contributes significantly to the overall quality. Direct coding of residual phase is not possible with the very low-bit rate constraint.

Saint-Arnaud *et al.* [10] define a sound texture as a signal that exhibit similar "perceptual" characteristics over time. From our experience with many types of environment sounds, we have observed that there is a large amount of "perceptual redundancy" over time. We have exploited this redundancy to devise a new excitation model for the LP residual of background noise.

Table 1 Bit allocation for a 20 ms noise frame in CDMA VBR coders

Parameter	QCELP	Q13	EVRC
LSF coefficients	10	10	8
Residual energy	6	6	8
Residual waveform	-	-	-
Total	16	16	16

3 CLASS-DEPENDENT NOISE CODING

In Figure 1, we show a block diagram of a LPC synthesis model with the proposed class-dependent residual substitution. In our approach, instead of using a white noise excitation to a LPC synthesis filter, we choose an excitation signal from a set of pre-stored excitation signals, one for each of a number of classes of noise types. We have observed that if a stored LP residual waveform of an appropriate type is used, the character of the noise is well preserved. An example of a class of noise is babble noise (a large number of simultaneous talkers). For example, if we save the residual from one instance of babble noise and use it for another, the output of the LPC synthesis filter is perceptually similar to the original. A different stored residual would be used if car noise is encountered. Our experimental results have confirmed that this class-dependent residual substitution can produce natural quality for a number of different noise types common in mobile environments (i.e. car, street, bus, and restaurant).

In our scheme, classification of environment acoustic noise is used to select the type of excitation for each frame. The type of residual to be used at the receiver can be transmitted as a short code or extracted during speech activity (i.e. using information from the hangover frames). A set of features are extracted from a noise frame and a classifier is used to decide the appropriate noise class.

4 IMPLEMENTATION & EVALUATION

We have modified the noise coding mode of the EVRC speech coder to include our class-dependent residual substitution scheme. We have replaced the pseudo-random noise

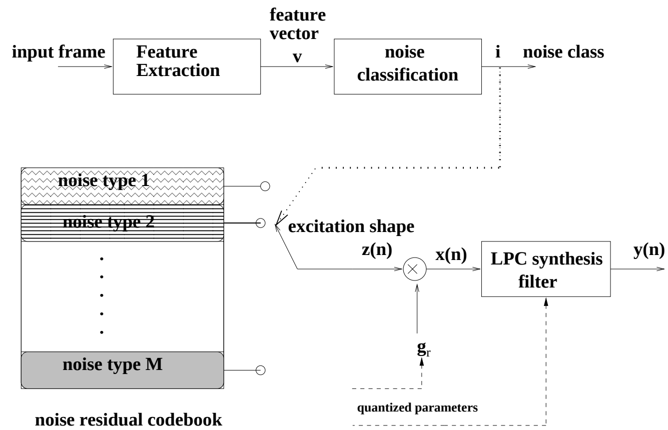

Fig. 1 Class-dependent Residual Substitution

generator with a codebook containing stored LP residual from M noise types. The residual codebook has a size of $M \times L$, where M is the number of noise types, and L is the length (in frames) of stored LP residual for each noise type. The length of the stored residual should be long enough to prevent any perceivable repetition of the noise. For our implementation, we have selected M to be 4 noise classes. To preserve the texture of the reconstructed noise, sequential residual substitution has to be done for each noise type. A frame counter for each noise class is used to select a segment of stored residual sequentially.

We have experimented with classifying the background noise into a number of canonical types. The details of our work in designing and evaluating noise classification algorithms can be found in [11]. A decision is made once every 20 ms to select the noise type. Classification accuracies of about 89% were obtained, with the accuracy depending on the noise class. The best classification results were obtained using a quadratic Gaussian classifier (QGC) with the line spectral frequencies (LSFs) as features. A sample of the results are shown in Table 2 in the form of classification matrix.

Using an unquantized LPC synthesis model with the class-dependent residual substitution excitation, we have obtained natural quality synthesis of different background noises. However, when we evaluated the EVRC coder with both the existing and the proposed noise coding schemes, only a slight improvement was perceived. To investigate the problem, we have studied the spectral quantizer of the EVRC noise coding mode. It uses two codebooks, each of size 16 codevectors and dimension 5. We have observed that even though 8 bits per frame are allocated for spectral quantization, only a few codevectors are used for all the noises we have tested. In Tables 3 and 4, we show the codevectors usage statistics in the two noise codebooks. We conclude that the noise spectral codebooks are not being used efficiently. We have designed a new 8-bit spectral vector quantizer[1] using background noise training data. With the new quantizer, we have obtained natural quality syn-

[1]The complexity is still lower than the spectral quantizer in the speech mode

thesis of the background noises we have considered. This improvement in the quality of coded background noise during non-speech activity helps in preserving noise continuity between talk spurts and speech pauses. Our scheme provides a significant improvement in quality over the original EVRC coder.

Table 2 Classification matrix: Gaussian classifier

	Babble %	Car %	Bus %	Factory %	Street %
Babble	**79.8**	0.0	12.8	2.0	5.4
Car	0.0	**99.6**	0.2	0.2	0.0
Bus	8.8	0.0	**85.2**	2.2	3.8
Factory	1.0	0.0	5.6	**93.2**	0.2
Street	1.8	0.0	24.8	2.0	**71.4**

Table 3 Statistics of codevectors usage: noise LSF CB1

CB entry	Babble %	Bus %	Car %	Street %
1	11.60	12.60	63.67	19.10
2	54.90	66.50	36.32	24.75
3	0.07	0.15	0.00	0.17
4	33.00	19.63	0.01	53.20
5	0.10	0.46	0.00	1.02
6	0.07	0.07	0.00	0.11
7	0.01	0.03	0.00	0.00
8	0.14	0.04	0.00	0.35
9	0.00	0.02	0.00	0.00
10	0.07	0.14	0.00	0.39
11	0.03	0.18	0.00	0.76
12	0.01	0.19	0.00	0.03
13	0.00	0.00	0.00	0.00
14	0.00	0.01	0.00	0.11
15	0.00	0.00	0.00	0.01
16	0.01	0.01	0.00	0.00

Table 4 Statistics of codevectors usage: noise LSF CB2

CB entry	Babble %	Bus %	Car %	Street %
1	0.08	0.01	0.00	0.02
2	1.33	0.20	0.03	0.27
3	10.32	10.27	10.51	13.44
4	3.56	5.15	5.20	5.94
5	0.74	1.12	0.00	2.22
6	18.64	23.80	7.70	26.40
7	0.09	0.06	0.00	0.10
8	49.36	39.48	73.40	37.29
9	7.75	13.30	3.13	12.88
10	4.02	2.70	0.03	0.48
11	0.18	0.07	0.00	0.10
12	0.33	0.36	0.00	0.10
13	1.63	1.69	0.00	0.34
14	1.68	1.14	0.00	0.08
15	0.10	0.05	0.00	0.18
16	0.19	0.68	0.00	0.13

5 CONCLUSION

We have presented a novel very low bit rate background noise coding mode for variable rate speech coders. Class-dependent residual substitution is used at the receiver to naturally reproduce background noise. We have integrated our noise coding scheme within the EVRC coder to give improved overall quality. We are currently continuing work on this promising noise coding technique and we are developing a more general residual substitution algorithm using a noise mixture model with canonical noise classes.

REFERENCES

[1] A. Das, E. Paksoy, and A. Gersho, "Multimode and varaible-rate speech coding," *Speech Coding and Synthesis*, pp. 257–288, Eds. W. B. Kleijn and K. K. Paliwal, Elsiver, 1995.

[2] A. DeJaco, W. Gardner, P. Jacobs, and C. Lee, "QCELP:the north american CDMA digital cellular variable rate speech coding standard," *Proc. IEEE Workshop on Speech Coding for Telecommunications* (Sainte-Adele,Quebec), pp. 5–6, Oct. 1993.

[3] P. Kroon and M. Recchione, "A low-complexity toll-quality varaible rate coder for CDMA digital cellular," *Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Processing* (Detroit, MI), pp. 5–8, May 1995.

[4] F. Beritelli, "A modified CS-ACELP algorithm for variable-rate speech coding robust in noisy environments," *IEEE Signal Processing Letters*, vol. 6, pp. 31–34, Feb. 1999.

[5] ITU-T, Geneva, *COM 12-1-E- List and wording of questions allocated to Study Group 12 for study during the 1997–2000 study period*, Feb. 1997.

[6] T. Wigren, A. Bergstrom, S. Harrysson, F. Jansson, and H. Nilsson, "Improvements of background sound coding in linear predictive speech coders," *Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Processing* (Detroit, MI), pp. 25–29, May 1995.

[7] Qualcomm Document, *High Rate Speech Service Option for Wideband Spread Spectrum Communications Systems*, Feb. 1996.

[8] TIA Document, PN-3292, *Enhanced Variable Rate Codec, Speech Service Option 3 for Wideband Spread Spectrum Digital Systems*, Jan. 1996.

[9] C. Ma and D. O'Shaughnessy, "A perceptual study of source coding of Fourier phase and amplitude of the linear predictive coding residual of vowel sounds," *J. Acoust. Soc. Am.*, vol. 95, pp. 2231–2239, Apr. 1994.

[10] N. Saint-Arnaud and K. Popat, "Analysis and synthesis of sound textures," *Proc. AJCAI Workshop on Computational Auditory Scene Analysis* (Montreal, Quebec), pp. 125–131, Aug. 1995.

[11] K. El-Maleh, A. Samouelian, and P. Kabal, "Frame-level noise classification in mobile environments," *Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Processing* (Phoenix, AZ), Mar. 1999.