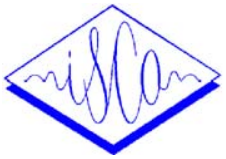

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

NATURAL-QUALITY BACKGROUND NOISE CODING USING RESIDUAL SUBSTITUTION

Khaled El-Maleh *Peter Kabal*

Dept. Electrical & Computer Engineering
McGill University, Montreal, Quebec, Canada H3A 2A7

10.21437/Eurospeech.1999-516

ABSTRACT

Existing approaches to background noise coding at very low bit rates (i.e., below 1 kbps) fail to reproduce the noise with natural quality, resulting in a degradation of the overall perceived quality. In this paper, we propose a novel scheme for natural-quality reduced-rate coding of background acoustic noise in voice communication systems. A better representation of the excitation signal in a noise-synthesis model is achieved by classifying the type of acoustic environment noise. Class-dependent residual substitution is used at the receive side to synthesize background noise that sounds similar to the background noise at the transmit side. The improvement in the quality of synthesized noise during speech gaps helps in preserving noise continuity between talk spurts and speech pauses, and enhances the perceived quality of a conversation.

1 INTRODUCTION

Reduction of the level of transmitted power is important in wireless voice communication systems to reduce co-channel interference and to prolong the battery life of portable units. Interference reduction in cellular systems allows for an increase in system capacity. In general, reducing the transmission bit rate translates into reduction of the power level of transmitted information. A typical telephone conversation contains approximately 40% active speech bursts and 60% silence and background environment noises. Background acoustic noise carries less information than speech and thus it is inefficient to use full rate speech coding for background noise.

One way to exploit the low voice activity is to use a voice activity detection (VAD) device that discriminates between voice and non-voice signals [1]. If voice is detected, then the full rate is used to encode speech signals. For non-speech signals (i.e. silence and background noise) a lower bit rate is used [2] [3].

In wireless communications systems with a discontinuous transmission mode, the transmitter is switched off during the absence of speech [4]. To fill the gaps between speech bursts, a synthetic "comfort" noise is generated at the receive-side using transmitted noise information. A periodic update of the noise statistics is transmitted using what are known as silence insertion descriptor (SID) frames [5]. The combination of voice activity detection, discontinuous transmission, and comfort noise insertion has been used by the Global System for Mobile Communications (GSM).

A silence compression scheme is important in bit-rate sensitive applications such as digital simultaneous voice and data systems, and voice over internet-protocol [6]. A silence compression device includes a VAD algorithm and a comfort noise generator. When speech is not present, the transmitter is idle except for periodic updates of noise information. A synthetic noise is substituted for the background noise.

Other wireless systems require a continuous mode of transmission for system synchronization and channel monitoring. During absence of speech, a lower rate coding mode is used to encode background noise. In Code Division Multiple Access (CDMA) wireless communication systems, variable bit rate (VBR) coding is used to reduce the average bit rate and to increase system capacity [7] [8].

In the present generation of background noise coding and comfort noise insertion algorithms, a simple excitation-filter model is used for noise synthesis. A signal is modelled as the output signal of a filter excited by a source signal. Existing schemes for background noise coding at very low bit rates (below 1 kbps) fail to reproduce background noise with natural quality during speech inactivity. However, when speech is present and being coding at the full rate, incidental noise will be coded along with the speech. The change in the character of the noise during speech activity and speech pauses is noticeable and can be annoying. This results in a degradation of the perceived quality of voice [9]. The ITU-T Study Group 12 (Question 17 "Noise aspects in evolving networks") is currently studying methods to mitigate the undesirable effect of background noise in voice communications systems [10].

In this paper, the quality of synthesized background noise during speech inactivity is improved by using a novel representation of the excitation signal to the noise synthesis model. Class-dependent residual substitution is proposed to achieve natural-quality reconstruction of acoustic environment sounds at very low bit rates.

2 LINEAR PREDICTION SYNTHESIS MODEL

A common signal modelling paradigm is one based on linear prediction (LP) methods. A signal is decomposed into two components: the LP residual signal and a set of parameters characterizing the spectral envelope. Ideally, passing the LP residual signal through the LPC synthesis filter generates the original signal. The spectral envelope is important perceptually as it preserves the spectral content of the signal. The amount of information left in the LP residual depends mainly on the predictor order, and the type of the input signal. The prediction filter decorrelates the input signal to produce the LP residual. The noise residual signal is often assumed to be statistically Gaussian with white spectral content.

In high-quality LP-based speech coding, a large portion of the overall bit rate is allocated to encode the LP residual. Kubin *et al.* [11] have shown that replacing the LP residual waveform of the unvoiced speech with a white

Gaussian noise, it is possible to reproduce unvoiced speech with high perceptual quality. However, for voiced speech, a WGN does not model the fine details embedded in the voiced residual.

In existing noise coding and comfort noise systems, only the spectral parameters and the residual energy are quantized. The residual waveform is not encoded, instead a random noise excitation, matching the noise residual energy, is used at the receiver to excite the LPC synthesis filter. We have studied the LP residual of different background noise signals. We have observed that substituting the LP residual with a random excitation retains naturalness for some noises (i.e., car, computer fan), but it fails to reproduce with natural-quality other "structured" background noises such as babble, street and office noises.

2.1 Spectral Excitation Model

We have studied the spectral content of the LP residual of different types of background acoustic noises. We have observed that some noises have non-flat residual amplitude spectra and thus using a white excitation is not appropriate. To parameterize the residual amplitude spectrum, a set of critical-band gain values is computed. The residual Fourier phase is replaced by a randomly-generated phase. A slight improvement in quality was gained from using this spectral excitation model. Using the "true" LP residual phase with the coarse representation of the residual amplitude spectrum reproduced noises with high quality. We have identified that the residual Fourier phase contributes significantly to preserving naturalness in the synthesized noises. This agrees with the conclusion reported in [12] that the phase spectrum of the LP residual determines the temporal waveform and contributes significantly to the overall quality.

3 CLASS-DEPENDENT RESIDUAL SUBSTITUTION

With a limited bit-budget for each frame, it is not feasible to encode the LP residual phase. Moreover, with the variety of noise sources, modelling the noise residual phase is not an easy task. In listening to long segments of different types of acoustic noise signals, we have observed that there is a large amount of temporal perceptual redundancy. This suggests that segments of a given noise type have perceptually-similar sound texture. We have exploited this observation to propose the class-dependent residual substitution excitation model. In our approach, the LP residual of the background noise during speech gaps is replaced at the receiver by an excitation signal that maintains the perceptual character. This is achieved by using a noise classification module that identifies the type of the background noise. The excitation selection module uses the noise classification decision to output an excitation signal from the identified noise class.

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 street noise is encountered. Our experimental results have confirmed that

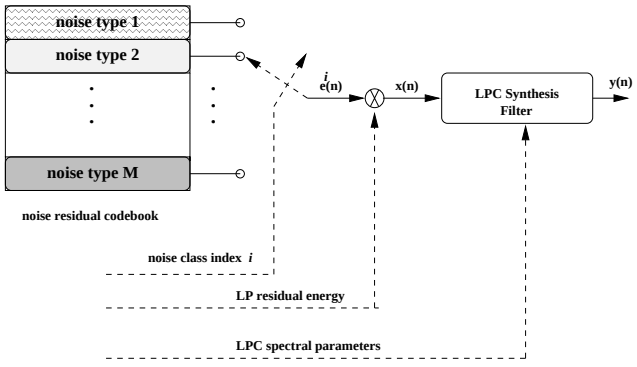

Fig. 1 Class-dependent Residual Substitution

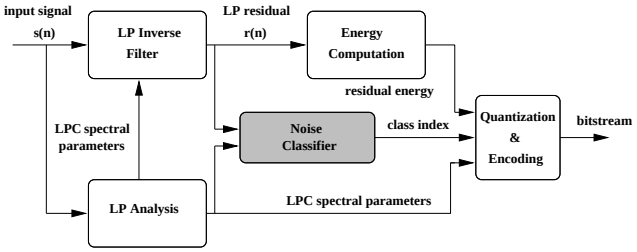

Fig. 2 Noise classification at the transmit side

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 Fig. 1, we show a block diagram of an LPC synthesis model with the proposed class-dependent residual substitution.

4 NOISE CLASSIFICATION

The first step in designing an M-class noise classifier is to define the M noise classes of interest. Then, a set of signal features is specified that in combination with a selected classification algorithm give good classification results. Training data from each noise class, in the form of labelled feature vectors, are used to train the classification algorithm. In the test phase, the classification rule maps an input feature vector to the closest class[1]. A set of noise features are used as input to the noise classifier. The classifier outputs a noise class index i that is transmitted to the receiver for class-dependent excitation selection. Classification at the transmitter can use any set of features from the input signal that discriminates between noise classes. Figure 2 shows the encoder part of an LP-based noise coder with a noise classification module.

We have experimented with classifying the background noise into a number of canonical types. 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[2]. Good classification results were obtained using a quadratic Gaussian classifier (QGC) with the line spectral frequencies (LSFs) as features. A sample of the results is shown in Table 1 in the form of classification matrix. The details of our work in designing and evaluating noise classification algorithms can be found in [13].

[1] To further improve the classification accuracy, the output decision can be smoothed using a decision-processing module.

[2] Such an accuracy is sufficient for our application.

Table 1 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**

4.1 RECEIVE-SIDE NOISE CLASSIFICATION

To minimize the occurrence of speech clipping resulting from classification of speech as background noise, VAD algorithms often include a hangover mechanism that delays the transition from speech to silence. A hangover period of few frames (i.e., 3–10) is commonly used [14]. In most cases, the hangover frames contain background noise. These noise frames are encoded using the full-rate of the speech coder. Using the coded background noise contained in the hangover frames, it is possible to do noise classification at the receiver side. This avoids the need to transmit noise classification bits. Moreover, the silence insertion description (SID) frames can be also used to perform noise classification at the receiver. Consider the DTX scheme used by the GSM enhanced full-rate speech coder [4]. An SID noise frame contains quantized parameters using the full-rate coder. The excitation waveform bits are replaced by a 95-bit codeword to help the receiver to identify SID frames. Classification features can be extracted from the quantized signal parameters and used as input to the classification module. For example, the quantized LSFs can be used as input features for the noise classification module.

5 NOISE RESIDUAL CODEBOOK

The noise residual codebook is populated with *prototype* LP residual waveforms from the M noise classes. 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 perceived repetition. To preserve the perceptual texture of the reconstructed noise, the excitation signal is constructed from sequential residual samples.

An alternative way to update the content of the noise residual codebook at the receive side, is to use the excitation signal of the hangover frames. The hangover frames are encoded with the full-rate of the speech coder, with a good reproduction of the LP residual at the transmit side. After classifying a hangover frame to one of the M noise classes, its excitation signal is used to update the excitation codevector of the corresponding noise class.

6 EVALUATION EXPERIMENTS

We have performed several concept-validation experiments to assess the improvement in quality using the proposed class-dependent residual substitution scheme.

Our experimental setup consists of a conventional linear prediction analysis-synthesis system. A $10th$ order LP analysis is performed every 20 ms using the autocorrelation method. A Hamming window of length 240 samples is used. The LP coefficients are calculated using the Levinson-Durbin algorithm and then bandwidth expanded using a factor $\gamma = 0.994$. The input noise signal is filtered through the LP inverse filter, controlled by the LPC spectral parameters, to produce the LP residual signal. The residual waveform is replaced by an LP residual from a similar noise class, with the same energy content. The *new* LP residual excites the unquantized LPC synthesis filter to produce a reconstructed noise signal. Listening tests confirm that substituting the residual of one noise class with an appropriate residual, preserves the perceptual texture of the input background noise.

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. We have replaced the pseudo-random noise generator with a codebook containing stored LP residual from M noise types. For our implementation, we have selected M to be 4 noise classes (car, street, babble, and factory). Evaluation tests have shown that we have improved the overall quality with the proposed noise coding scheme without an increase in bit rate.

In the GSM discontinuous transmission system, in a cycle of 24 noise frames, the first frame is transmitted using the full-rate coder, with zero bits for the remaining frames. At the receiver side, interpolation is used to substitute the parameters of the untransmitted frames. A randomly-generated excitation is used to replace the residual for all the frames in the cycle. The comfort noise generated using this approach sounds different from the background noise at the transmit side. The difference in quality is caused by discarding the residual waveform, and the infrequent transmission of spectral parameters.

We have simulated the GSM discontinuous transmission mode using a "controlled" frame loss model. In a cycle of L noise frames, we keep the spectral and the energy parameters of the first frame and discard the next $L-1$ frames. The LP residual of all the frames in the cycle are substituted with an LP residual from a similar noise class. The spectral and energy parameters are interpolated using [5],

$$p(n+i) = (1 - \frac{i}{L})p(n-L) + \frac{i}{L}p(n),$$

where $p(n+i)$ is the parameter of frame $n+i$ (for $i = 0, 1, ..., L-1$), $p(n)$ is the parameter of the first frame in the current cycle, and $p(n-L)$ is the parameter for the first frame in the second latest cycle.

We have experimented with different choices of the cycle length L. Listening quality tests confirm that using class-dependent residual substitution and interpolated spectral envelope reproduce background noise with natural quality, even for a large frame loss rate (i.e., $L = 50$ frames). Class-dependent comfort noise insertion schemes, using the proposed noise excitation model, can enhance the quality of voice communication using GSM-based wireless systems.

7 RESIDUAL MIXTURE SUBSTITUTION

We present in Fig. 3, a general model for the class-dependent residual substitution scheme. The LP residual of the background noise at the transmit side is replaced at the receiver by an excitation mixture signal $e(n)$. The LPC excitation signal $e(n)$ is modelled as a linear mixture of M excitation signals from the M noise classes, given as:

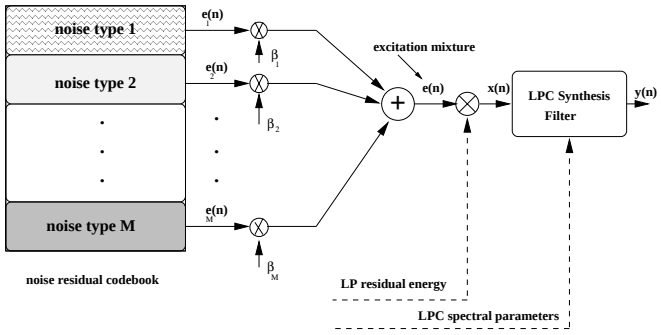

Fig. 3 Residual Mixture Substitution

$$e(n) = \sum_{i=1}^{M} \beta_i e_i(n),$$

where $e_i(n)$ is an excitation signal from the i^{th} noise class, and $\beta_i(n)$ is the i^{th} mixing coefficient, taking a value between 0 and 1, with $\sum_{i=1}^{M} \beta_i = 1$.

The mixing coefficients quantify the contribution of the excitation of each noise class to the excitation mixture. These mixing weights can be either sent to the receiver or determined at the receive side. A soft-decision classification module can be used to output M decision outputs with values ranging from 0 to 1. These soft-decision values can be transmitted to the receiver to be used as mixing coefficients.

The excitation model of Fig. 1 is a special case of the mixture excitation model. For example, if the the mixing vector is zero except for the j^{th} component, i.e., $\beta = [00..10..00]$, then we get an excitation signal from the j^{th} noise class, $e^j(n)$.

The excitation mixture model can be used as a means to reduce the effect of classification errors in the hard-decision model shown in Fig. 1. During a sudden transition from one class to another, a graceful blending of the excitation signals from the two noise classes is done.

In an acoustic noise environment, the background noise is often a mixture of acoustic signals from different noise sources. For example, in a public bus environment, the background noise consists of engine, babble, traffic and other ambient noises. A noise mixture model with soft-decision classification can be used to naturally synthesize such noises with a mixture of excitation signals from each noise type. A key issue in using the excitation mixture model is the estimation of the coefficients of the mixing model. The mixing values should reflect the energy-contribution of each noise type to the sound mixture.

8 CONCLUSION

We have proposed a novel scheme for natural-quality coding of background acoustic noise at very low bit rates. Using class-dependent reproduction of background noise during voice inactivity produces synthesized noise that sounds similar to the background noise during voice activity. The improvement in noise synthesis projects a much-enhanced overall noise environment to the listener, and improves the overall perceived quality of a voice communication system. We are currently continuing work on this promising noise coding technique, and we are investigating methods to estimate the coefficients of the excitation mixture model.

REFERENCES

[1] K. El-Maleh and P. Kabal, "Comparison of voice activity detection algorithms for wireless personal communications systems," *Proc. IEEE Canadian Conference on Electrical and Computer Engineering* (St. John's, Nfld), pp. 470–473, May 1997.

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

[3] A. Das *et al.*, "Multimode variable bit rate speech coding: An efficient paradigm for high-quality low-rate representation of speech signal," *Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Processing* (Phoenix, AZ), pp. 2307–2310, Mar. 1999.

[4] ETSI TC-SMG, GSM 06.62 Version 6.0.0 Release 1997, *Digital Cellular Telecommunications System (Phase 2+); Discontinuous Transmission (DTX) for Enhanced Full Rate (EFR) Speech Traffic Channels*, 1997.

[5] ETSI TC-SMG, GSM 06.81 Version 6.0.0 Release 1997, *Digital Cellular Telecommunications System (Phase 2+); Comfort Noise Aspects for Enhanced Full Rate (EFR) Speech Traffic Channels*, 1997.

[6] A. Benyassine *et al.*, "ITU-T Recommendation G.729 Annex B: A silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data applications," *IEEE Communications Magazine*, vol. 35, pp. 64–73, Sept. 1997.

[7] TIA/EIA/IS-733, *High Rate Speech Service Option for Wideband Spread Spectrum Communications Systems*, Feb. 1996.

[8] TIA/EIA/IS-127, *Enhanced Variable Rate Codec, Speech Service Option 3 for Wideband Spread Spectrum Digital Systems*, Jan. 1996.

[9] 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.

[10] 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.

[11] G. Kubin, B. Atal, and W. B. Kleijn, "Performance of noise excitation for unvoiced speech," *Proc. IEEE Workshop on Speech Coding for Telecommunications* (Sainte-Adele, Quebec), pp. 35–36, Oct. 1993.

[12] 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.

[13] 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), pp. 237–240, Mar. 1999.

[14] ETSI TC-SMG, GSM 06.82 Version 6.0.0 Release 1997, *Digital Cellular Telecommunications System (Phase 2+); Voice Activity Detector (VAD) for Enhanced Full Rate (EFR) Speech Traffic Channels*, 1997.