Classified Comfort Noise Generation for Efficient Voice Transmission

Yasheng Qian, Wei-Shou Hsu, Peter Kabal *

Department of Electrical and Computer Engineering
McGill University, Montreal, Canada H3A 2A7
yasheng@tsp.ece.mcgill.ca, whsu@tsp.ece.mcgill.ca, kabal@ece.mcgill.ca

10.21437/Interspeech.2006-63

Abstract

Comfort noise insertion during speech pause has been applied to Voice-over-IP and wireless networks for increasing bandwidth efficiency. We present two classified comfort noise generation (CCNG) schemes using Gaussian Mixture classifiers (GMM-C). Our first scheme employs a classified prototype background noise codebook with the prototype noise waveform chosen using a GMM-C. The second scheme utilizes a classified enhanced excitation codebook. The new CCNG algorithms provide better comfort noise during speech pauses and a smaller misclassification rate. We have retrofitted the scheme into existing speech transmission system, such as ITU-T G.711/Appendix II and G.723.1/Annex A. The perceived quality of a voice conversation of the novel system has been noticeably enhanced for car and babble noise. For the G.711 system, a large improvement is obtained for car noise while the largest amelioration is for babble noise in the G.723.1 system.

Index Terms: Comfort Noise, Gaussian Mixture classifier, classified prototype codebook, enhanced classified excitation codebook, soft-decision Gaussian mixture classifier.

1 Introduction

Voice conversation in telephone networks exhibits speech gaps which are not silence, but are usually filled with background noise. These occupy around 60% of talking time on the average. A speech gap is illustrated in in Fig. 1. Methods to exploit those silence periods was first introduced to increase the capacity of the transatlantic cables about fifty years ago. With the rapid evolution of the wireless cellular communication systems and packet based telephone networks, many current telecommunications networks take advantage of speech pauses to increase transmission efficiency. Some wireless systems cease transmission during speech gaps. This is known as discontinuous transmission (DTX). For Voice-over-Internet (VoIP) networks, discontinuous transmission reduces the number of speech packets by about 40%. In CDMA wireless systems, during speech pauses transmission occurs at a lower bit rate (and lower power). This reduced power increases channel capacity by reducing co-channel interference during speech pauses.

Present communications networks are converging classical telephony and packet-based VoIP networks for cost reduction and greater service flexibility. All recent VoIP and mobile networks standards have exploited speech gaps.

Although a background noise contains no speech information, the transmission of the background noise during speech gaps has been shown to be critical for naturalness of a conversation. If the noise during speech pauses abruptly disappears at the decoder, it can be very annoying and unpleasant to listeners [1]. Such effects may even reduce the intelligibility of the speech. The International Telecommunication Union (ITU-T) [2] uses the term "comfort" noise to denote the the reconstructed environment noise during speech gaps. Figure 1 illustrates the substitution of a comfort noise segment into a speech gap.

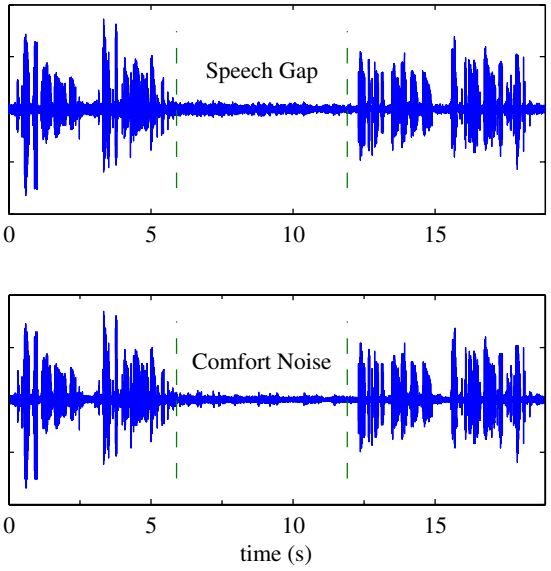

Fig. 1 Background noise and comfort noise in voice conversation: original speech and speech gap with background noise at the encoder (top); coded speech and comfort noise at the decoder (bottom).

The aim of the Comfort Noise Generator (CNG) algorithm is to create a noise that matches the actual background noise, but which can be generated with a reduced rate relative to active speech. CNG systems do not need to exactly reproduce the background noise waveform. Dif-

*This work was supported by an NSERC grant in Canada

ferent CNG algorithms offer different trade-offs of quality and simplicity. Many CNG algorithms use a white noise excitation as an input to a linear prediction (LP) synthesis filter. Based on linear prediction speech production model, the more the information we have about the background noise — excitation and the LP spectrum — the better the CNG quality but the higher the transmission rate.

There are two approaches to taking advantage of speech pauses. In discontinuous transmission, transmission ceases for speech pauses, except for the occasional silence insertion description (SID) frame. These SID frames describe the spectral envelope and energy of the background noise. Variable bit-rate (VBR) speech coding schemes have attempted to implement CNG by employing simplified (reduced rate) coding schemes during inactive speech segments (e.g., [3], [4]), continuously.

K. El-Maleh and P. Kabal [5] have proposed a classified excitation codebook to be used as the excitation signal in the LP model. Because the excitation codebook contains a richer texture for babble noise, for instance, it results in more natural comfort noise. At the same time, only a few additional bits are required to specify the class index.

We present two classified comfort noise generation (CCNG) schemes, as an extension to their previous work. The first scheme employs a Gaussian Mixture classifier (GMM-C) at the transmitter or at the receiver and a prototype codebook at the receiver. The GMM-C reduces the average misclassification rate by 16%, compared to the previous Quadratic Gaussian Classifier (QG-C) [5]. A classified prototype codebook replaces the LP synthesis approach used in the earlier work. The LP parameters are used to select the codebook entry. This scheme is able to reproduce more harmonics, and low and high frequencies than a Gaussian or an excitation codebook approach. Since only the energy and the class of background noise need be transmitted, the coded bits for CNG are reduced by up to 82% compared to the ITU-T G.711 CNG standard.

The second scheme uses a soft-decision Gaussian mixture classifier and an enhanced classified excitation codebook at the receiver side to generate the comfort noise. The codebook is specially designed to improve the quality of background babble noise, while measures are taken to ensure good performance for noises not included in the codebook. Because the system can be implemented entirely at the receiver side, no modifications to the transmitter are required. The novel CNG algorithms provide better background noise reproduction during speech pauses, a lower misclassification rate, and requires a small transmission bandwidth for the CNG information.

2 Maximum likelihood classifier with Gaussian Mixture Model (MLCGM)

The spectral characteristics of background noise can be well represented by a spectrum corresponding to linear prediction (LP) filter. The filter can be parameterized by the direct LP coefficients or their transformations, such as line spectral frequencies (LSF), reflection coefficients, and cepstral coefficients. We favour the use of the LSF representation as the features for noise classification, because of its good separability for different noise classes. A 10th order of LP analysis is carried out for each 80 samples (10 ms). An unsymmetrical window of length 200 is used for analysis. The window uses half of a Hamming window over 170 samples of the present and the past data and a half-cosine window is provided for a 30 sample lookahead. The LP parameters are represented by 10 LSF coefficients.

The LSF histograms of five class noises show noticeable differences from a bell-shaped Gaussian probability density. The histograms for the 5th LSF and the 6th LSF of street noise are shown in Fig. 2. These distributions show a significant deviation from Gaussianity. The histograms and kurtosis values of the LSF features help convince us that we require a more general modelling of their probability density functions.

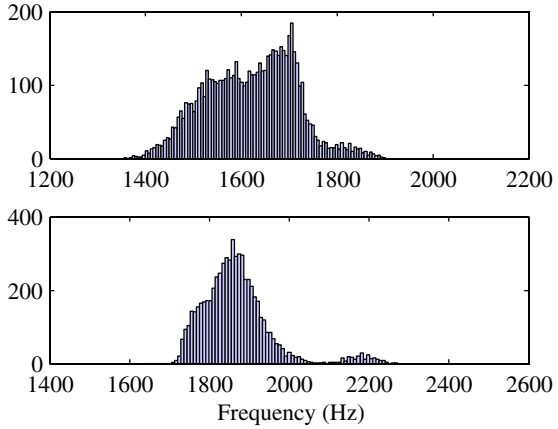

Fig. 2 The histogram of the 5th LSF (top) and the 6th LSF (bottom) for street noise.

The Gaussian Mixture Mode (GMM) probability density function (pdf) is a weighted sum of M D-dimensional joint Gaussian density distributions. For noise class k, the GMM is

$$p_{ZK}(\boldsymbol{z}|\boldsymbol{\alpha}_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \sum_{i=1}^{M} \alpha_{ik} b_{ik}(\boldsymbol{z}|\boldsymbol{\mu}_{ik}, \boldsymbol{\Sigma}_{ik}). \quad (1)$$

where M is the number of individual Gaussian components in the mixture, the α_{ik}, $i = 1, \ldots, M$ are the ith (positive) mixture weights of the component pdfs, and Z is a D-dimensional random vector, representing the 10 LSFs. Each density is a D-variate Gaussian pdf of the form,

$$b_{ik}(\boldsymbol{z}|\mu_{ik}, \boldsymbol{\Sigma}_{ik}) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}_{ik}|^{1/2}} \exp(-\frac{1}{2}(\boldsymbol{z} - \boldsymbol{\mu}_{ik})^T \boldsymbol{\Sigma}_{ik}^{-1}(\boldsymbol{z} - \boldsymbol{\mu}_{ik})). \quad (2)$$

with a mean vector $\boldsymbol{\mu}_{ik}$ and a covariance matrix $\boldsymbol{\Sigma}_{ik}$ for the ith component. The Gaussian Mixture pdf is defined by the mean matrix, the covariance matrices and the mixture weights for the Gaussian components.

The parameter set $\{\boldsymbol{\alpha}_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}$ can be iteratively determined using the expectation-maximization (EM) algorithm employing the the maximum likelihood criterion [7] for a set of training data. The training data are separately taken from babble, white, car, factory and street noise files. The number of mixtures M is equal to 4 for our implementation. The covariance matrices, $\boldsymbol{\Sigma}_{ik}$, are assumed to be diagonal for simplicity.

In the maximum likelihood classifier using a GMM, N_k (number of noise classes) GMM pdf values are calculated by Eq. (1) for each input LSF vector. The class $k_{\max}$ is chosen as the class whose GMM pdf gives the maximum likelihood for a given input vector $\boldsymbol{z}$,

$$k_{\max} = \arg\max_k \{p_{Zk}(\boldsymbol{z}|\boldsymbol{\alpha}_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}. \quad (3)$$

Five classes of noises each with 6,250 frames were used to test the accuracy of the GMM-C. The GMM-C determines their class as given by Eq. (3). The classification results are listed in a classification matrix Table 1. The type of the input noise is written at the head of the table. The percentage of the class decisions for the input noise are given in the columns. The accuracy rates for classification are presented along the the diagonal. All other points are the mis-classification rates. The average error rate is 4.96%, which is a reduction of about 16% compared to a classifier with $M = 1$, i.e. the QG-C scheme. The results show that GMM-C works extremely well for car, white, and factory noise. We note that the somewhat heterogenous noise types (babble, factory and street), have larger fractions of mis-classifications, but are never mistaken for the homogeneous car and white noises.

Table 1 Classification matrix of GMM-C

	Babble	White	Car	Factory	Street
Babble	90.8	0	0	2.6	9.6
White	0	100	0	0	0
Car	0	0	100	0	0
Factory	4.0	0	0	95.8	1.8
Street	5.2	0	0	1.6	88.6

3 Classified prototype codebook

A prototype codebook consists of N_k codewords, which are constructed from N_k classes of canonical noise. The classified CNG is created by the kth codeword multiplied by a gain factor g , as disclosed in Fig. 3. The index k of the noise class can be decoded from a transmitted bit-stream $c_k(m)$ coded at the encoder or generated by an GMM-C at the decoder. The gain factor g is produced by an encoder or a decoder, as the class index k.

Since the prototype codebook delivers typical environment noise which may contain a richer harmonic texture, low and high frequencies, the CCNG is well matched to background noise spectrum. Meanwhile, only one byte of noise energy for silence insertion descriptor is needed to be transmitted. No noise spectrum parameters are encoded.

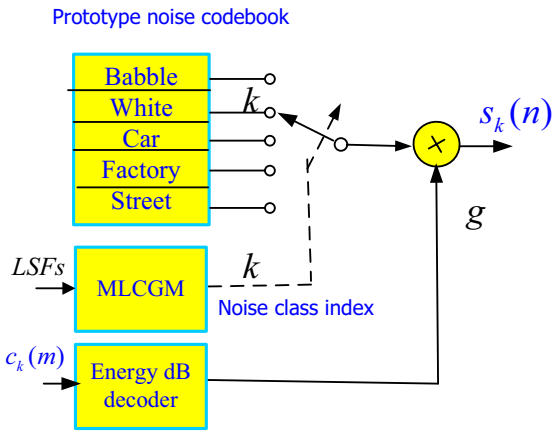

Fig. 3 The classified CNG generation

4 Enhanced Excitation Codebook

Since the classified excitation codebook can not include all kinds of background noise residuals, we employ a mixed soft-classified excitation from a classified excitation book. The system utilizes receiver-side noise classification and noise dependent excitation substitution to enhance the quality of comfort noise. Three noise classes — babble, white, and car — are employed, and Gaussian mixture models are used to model the posterior probability distributions of the line spectral frequencies of each type of noise. A soft decision classification scheme is used to determine the contribution (class weight) for each noise class. If the posterior probabilities of the babble and car noise classes are low, their contributions to the excitation are decreased and replaced by white noise.

It is assumed that the LP spectral envelope is coded in the SID information; therefore, the posterior probabilities can be calculated from each SID frame, and a decision on the noise class can then be made using the probabilities. To reduce the effects of misclassification and to smooth the transitions between different excitations, the excitation class weights obtained from the two most recent SID frames are averaged to give the current weight coefficients. With the noise classification done at the receiver side, no modification needs to be made to the encoder and no extra bits are required.

One of the key improvements of the new scheme is the modification of the babble excitation codebook entry. In the original system, the excitation codebooks are obtained by applying frame-by-frame linear prediction analysis to the noise signals, with the frame length chosen as in a typical speech processing application, which is short enough for speech to remain stationary during each frame. With an excitation taken from real babble noise, the CNG result for babble noise is improved, but still sounds somewhat synthetic and artificial.

Instead of using short frames to generate the excitation codebook entries, the new scheme employs a longer frame for LP analysis. As babble noise is highly non-stationary, its spectrum can change within a long frame, and the spectral envelope thus obtained is an average for the entire frame. As a result, the excitation undergoes less whitening, and some of the babble noise remains audible even after being passed through the LP synthesis filters of different noises. This gives the feeling of real background babble instead of a synthesized one.

One problem with the enhanced babble noise is that it also makes other noises sound like babble when misclassification occurs. To reduce the effects of this problem, it is noted that babble is less stationary than other types of noise. Assuming the transmitter sends spectral updates by detecting significant spectral changes, as many DTX algorithms do, background babble noise would require a higher frequency of SID frames than other noises. Therefore, a moving threshold is set for the maximum value for the weight coefficient of babble excitatin. Starting at 1.0, the threshold is decreased every time a null frame is received and increased when an SID frame is received. The threshold is restricted to lie between 0.5 and 1. When coded speech is encountered, the threshold is moved back to one.

5 Performance evaluation

We integrate the above-mentioned two CCNG schemes into a packet-based multimedia communication system (ITU-T G.711 Appendic II and G.723.1 Annex A). Subjective evaluation was carried out by the Degradation Category Rating (DCR) method defined in ITU-T Recommendation P.800. to assess the quality of G.711 and G.723.1 CNG and the CCNG, since the DCR method is more sensitive than Absolute Category Rating (ACR) method. The reference stimuli, as A in A-B pairs, are fully coded and decoded by G.711 or G.723.1 standard. The G.711 or G.723 CNG and CCNG encoded and decoded stimuli as B, constitute A-B pairs, respectively.

The DCR is defined on a 5-point scale as follows: 5 – Degradation is inaudible; 4 – Degradation is audible but not annoying; 3 – Degradation is slightly annoying; 2 – Degradation is annoying; 1 – Degradation is very annoying. The DCR testing of the G.711 CNG and CCNG with a classified protype codebook shows that the average CCNG DCR (3.64) improvement for car noise is 1.28. The DCR (4.50, 4.83, 4.98) improvements for babble, factory and street are 1.23, 1.16 and 0.65, respectively. The average G.723.1 CCNG DCR (3.46) improvement for babble noise is about 0.85. The G.723.1 CCNG demonstrates the highest enhancement for babble noise. As for other background noises, the improvements are less.

The enhanced excitation codebook CCNG scheme has been evaluated with informal tests. It shows significantly improved quality compared with the CCNG scheme with a classified protype codebook. This method produces comfort noise with improved quality, especially with unintelligible talk in the background, compared the ITU-T G.723.1 standard comfort noise [6].

6 Conclusions

We have developed a Gaussian Mixture classifier (GMC) which outperforms the Quadratic Gaussian Classifier (QGC) in terms of classification accuracy. We have incorporated the GMC into an improved background noise coding system using residual substitution. It delivers high quality CNG. We have further explored two classified comfort noise generation (CCNG) schemes. The first one employs a classified prototype background noise codebook with a GMC. The subjective DCR tests show that the proposed system is noticeably better than the existing ITU-T G.711 Appendix II in the car noise case and more natural than the ITU-T G.723.1 Appendix A in the babble noise case. The second scheme utilizes an enhanced residual codebook that improves the quality of background babble noise and also works well for other types of noises. Furthermore, the classification is done at the receiver side and no modifications are needed at the transmitter. However, the misclassification impairments and the remedies are to be further investigated.

References

[1] H. W. Gierlich and F. Kettler, "Background Noise Transmission and Comfort Noise Insertion: The Influence of Signal Processing on 'Speech'-Quality in Complex Transmission Scenarios", *Proc. Int. Workshop Acoustic Echo, Noise Control*, Sept. 2001.

[2] ITU-T, "Pulse Code Modulation (PCM) of Voice Frequency, Appendix II: A Comfort Noise Payload Definition for ITU-T G.711 Use in Packet-Based Multimedia Communication Systems", ITU-T Recommendation G.711/Appendix II, Feb. 2000.

[3] G. Ruggeri, F. Beritelli and S. Casale, "Hybrid Multi-Mode/Multi-Rate CS-ACELP Speech Coding for Adaptive Voice over IP",*Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Processing (Salt Lake City, Utah)*, Vol. II SP-P4.3, May, 2001.

[4] 3GPP2, "Enhanced Variable Rate Codec, Speech Service Option 3 for Wideband Spread Spectrum Digital Systems", Interim Standard TIA-IS-127, Jan. 1996.

[5] K. El-Maleh and P. Kabal, "Method and Apparatus for Providing Background Acoustic Noise During a Discontinued/Reduced Rate Transmission Mode of a Voice Transmission System," US Patent 6,782,361, Aug. 2004.

[6] ITU-T, " Dual rate speech coder for multimedia communications transmitting at 5.3 and 6.3 kbit/s, Annex A: Silence compression scheme", ITU-T Recommendation G.723.1/Annex A, Nov. 1996.

[7] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm", *J. Royal Statistical Soc.*, Series B, vol. 39, pp. 1–38, 1977.