Method for Comfort Noise Generation and Voice Activity Detection for use in Echo Cancellation System

Kirill Sakhnov
Dept. of Telecommunication Engineering
Czech Technical University in Prague
Prague, Czech Republic
sakhnkir@fel.cvut.cz

Boris Simak
Dept. of Telecommunication Engineering
Czech Technical University in Prague
Prague, Czech Republic
simak@fel.cvut.cz

Abstract— This paper relates to communications systems, and more particularly, to principles of comfort noise generation for echo cancellers in a bidirectional communications link. According to the invention, noise model parameters are computed during periods of speech inactivity (i.e., when only noise is present) and frozen during periods of speech activity. Prevailing noise model parameters are then used to generate high quality comfort noise which is substituted for actual noise whenever the actual noise is muted or attenuated by an echo suppressor. Since the comfort noise closely matches the actual background noise in terms of both character and level, far-end users perceive signal continuity and are not distracted by the artifacts introduced by conventional methods.

Keywords-comfort noise generation;voice activity detection;parametrical line predictive coding.

I. INTRODUCTION

In many communications systems, for example landline and wireless telephone systems, voice signals are often transmitted between two speakers via a bi-directional communication link. In such systems, speech of a near-end user is typically detected by a near-end microphone at one end of the communications link and then transmitted over the link to a far-end loudspeaker for reproduction and presentation to a far-end user. Conversely, speech of the far-end user is detected by a far-end microphone and then transmitted via the communications link to a near-end loudspeaker for reproduction and presentation to the near-end user. At either end of the communications link, loudspeaker output detected by a microphone may be transmitted back over the communications link, resulting in what may be unacceptably disruptive feedback, or echo, from a user perspective. In response to the above described challenges, it has been developed a wide variety of echo suppression mechanisms [1], [2], [3]. Problem situation occurs when an echo suppressor attenuates the entire speech signal. Besides attenuating the echo, the echo suppressor also attenuates any background noise and/or near-end speech which may be presented. In fact, the background noise can be suppressed to the point that the far-end user erroneously believe that the call has been disconnected when the echo suppressor is active. A lot of echo cancellers, however, do not insert any noise to replace the zero clipping of the echo suppressor. The result is a channel that suddenly sounds dead whenever the suppressor is active. To the far-end listener these sudden variations in the noise level on the channel causes an annoying effect, which impedes the conversation. The effect becomes even more pronounced and objectionable when network delays are present, such as in satellite communication networks. The zero clipping of the echo suppressor also causes a non-linear effect for vocoders. This also degrades their performance. The sudden transition in levels introduces high frequency components into the signal which vocoders can not handle. Therefore, there is a need for noise generation for use in echo cancellers to provide constant and continuous background noise to avoid perceptible variations in the noise characteristics.

II. BACKGROUND

To improve the quality of communication for the far-end user, up-to-date systems often add comfort noise to the output speech signal when the echo suppressor is active. For instance, some systems replace muted speech signals with the white noise produced by a pseudo-random number generator (PRNG), wherein a variance of the noise samples is set based on an estimate of the energy in the actual background noise [1]. Yet another solution is described in the U.S. patent application [2]. There a block of samples of the actual background noise is stored in memory, and the comfort noise is generated by outputting segments of successively stored samples beginning with random starting points within the block.

While the above described systems provide certain advantages, none provides the comfort noise which closely and consistently matches the actual environment noise in terms of both spectral content and magnitude. Further, the comfort noise generated by repeatedly

outputting segments of actual noise samples includes a significant periodic component and therefore often sounds as if it includes a distorted added tone. Thus, with conventional noise generation techniques, the far-end user perceives continual changes in the character and content of the transmitted background noise, as the comfort noise is selectively added or substituted only when the echo suppressor is active. Such changes in the perceived background noise can be annoying or even intolerable. For instance, with the relatively long delay in digital cellular phones, differences between actual background noise and modeled comfort noise are often perceived as whisper echoes.

A. Linear Predictive Coding

As linear predictive coding (LPC) is used to calculate parametric coefficients of the background noise model ib the proposed algorithm, the following description is presented. The common idea of LPC is to build a model of speech signal that is based on the strong correlation that exists between adjacent samples [4]. Instead of transferring the whole signal waveform only parameters of the LPC model are transferred. The algorithm on the opposite side of the telecommunication link rebuilds the model and generates the speech signal very similar to the original. In this way only the essential information of the sound is need to be transferred. It helps to reduce the bandwidth and to achieve higher transmission rates. First, the algorithm tries to predict the sample of an input signal s(n) based on several previous samples

$$\hat{s}(n) = \sum_{k=1}^{N} a_k \cdot s(n-k). \tag{1}$$

In (1) the sample $\hat{s}(n)$ is estimated as a linear combination of N previous samples of the input signal and autoregressive coefficients a_k. Equation (1) is called an autoregressive (AR) model. Parameter N corresponds to the degree of the AR model. The prediction becomes more correct with increasing number of samples. It should be mentioned that there is a sharp trade-off between complexity of the algorithm and its efficiency. Computation complexity will also increases with high values for the degree of the AR model. The LPC coefficients a_k are chosen in such a way that the squared error between the real input sample and its predicted value is minimized. Then, the predictive error $e(n)$ is calculated, as it follows

$$e(n) = s(n) - \sum_{k=1}^{N} a_k \cdot s(n-k). \tag{2}$$

By transferring the previous equation to the frequency plane with the z-transform the transfer function of the analyzing filter is obtained

$$E(z) = S(z) - \sum_{k=1}^{N} a_k \cdot S(z) \cdot z^{-k} \tag{3}$$

$$= S(z) \cdot \left(1 - -\sum_{k=1}^{N} a_k \cdot z^{-k}\right) = S(z) \cdot A(z).$$

The error signal $e(n)$ is presented as the product of the original input signal $S(z)$ and the transfer function $A(z)$. So as to generate the original signal it is enough to get an inverse transfer function

$$\frac{1}{A(z)} = \frac{1}{1 - \sum_{k=1}^{N} a_k \cdot z^{-k}} \tag{4}$$

and multiply it with the excitation signal. The excitation signal as well as the LPC algorithm used in the proposed comfort noise generation algorithm is specified in the following section.

B. Voice Activity Detection

This subsection describes the principle of the proposed voice activity detector (VAD). The implemented VAD is an energy-based detector. The energy of the input speech signal is calculated using the root mean square energy (RMSE), which is the square root of the average sum of the squares of the amplitude of the signal samples

$$E = \sqrt{s^T(n) \cdot s(n) \cdot N^{-1}}. \tag{5}$$

Here, $\mathbf{s}(n)$ is the vector containing N samples of the input speech signal. The VAD is based on the observation that the evolution of the estimated short-term energy exhibits distinct peaks and valleys. While peaks correspond to speech activity the valleys can be used to obtain the estimation of noise energy. It is necessary to store into the memory the estimation of minimum, $E_{min}(n)$ and maximum, $E_{max}(n)$ energy values. A detection threshold between speech and silence is calculated, as in

$$T(n) = k_1 \cdot E_{\max}(n) + k_2 \cdot E_{\min}(n), \tag{6}$$

where parameters k_1 and k_2 are used to interpolate the threshold value to an optimal performance. If the current estimated energy is under the threshold, the frame is marked as active. Otherwise, it is declared to be non-active. There is also a hangover time of four non-active frames to overcome sudden variations in the final decision. Since low energy anomalies can occur during classification procedure, there is prevention needed for this. The parameter E_{min} is slightly increased for each input frame

$$E_{\min}(n) = E_{\min}(n-1) \cdot \sigma(n). \tag{7}$$

Practical experiments show that the parameter $\sigma(n)$ for each frame can be calculated, as in

$$\sigma(n) = \sigma(n-1) \cdot 1{,}0001 . \quad (8)$$

It is also possible to introduce (6) using a single parameter $\lambda = k_2$. Then the threshold is

$$T(n) = (1-\lambda) \cdot E_{\max}(n) + \lambda \cdot E_{\min}(n) , \quad (9)$$

where λ is a scaling factor controlling estimation process. Voce detector performs reliably when λ is in the range of [0.950…0.999]. However, the values λ for different types of signals may be different and a priori information has still been necessary to set up λ properly. The equation

$$\lambda = (E_{\max}(n) - E_{\min}(n)) \cdot E_{\max}^{-1} \quad (9)$$

shows how to make the scaling factor to be independent and resistant to the variable background environment. Fig. 1 shows example of the speech signal, estimated energy and threshold curves obtained in Matlab environment using the above presented algorithm.

III. PROPOSAL OF CNG-VAD SYSTEM

Fig. 2 shows a functional block diagram of the comfort noise generating system. This is a method for forming a comfort noise according to characteristics of the near-end speech signal before the non-linear process has been performed by the echo suppressor, and for adding the comfort noise to the voice signal after the non-linear process has been performed by the suppressor. For this purpose, the comfort noise is generated in a parallel manner with the echo suppressor. The comfort noise generating block comprises a noise buffer, an LPC analyzing part together with a coefficient register inside and a synthesis filter for generating noise samples. The echo canceller dynamically models the echo path and attempts to cancel any echo contained in the incoming near-end signal. Then the echo suppressor processes the output signal coming from the echo canceller and provides residual echo suppression. More specifically, it executes the non-linear process in accordance with a level of the signal. It removes residual components, so that the echo signal is attenuated completely and does not return back to the far-end speaker. The energy-based VAD outputs a binary flag indicating the presence or absence of speech in the near-end signal. When the VAD indicates that no speech is present, i.e. only noise is present, the echo canceller output signal is connected via the switch to the input of the comfort noise generator, and the LPC analyzing part computes and updates a parametric noise model. However, when the VAD indicates that speech is present in the near-end signal, the switch is open and the noise model parameters are frozen. The synthesizing filter uses stored LPC coefficients to generate samples of the comfort noise. When the speech signal passes to the sample buffer during periods of no speech, the excitation signal is generated by randomly selecting samples from the sample buffer. Thus the excitation signal consists of white noise samples having power equal to that of the actual background noise. The signal buffer should be

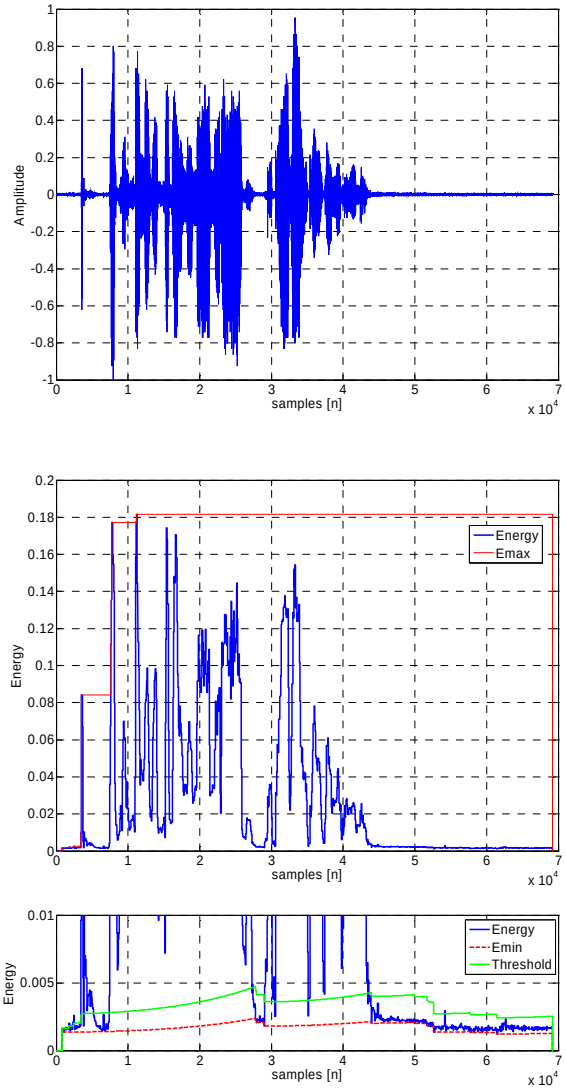

Figure 1. Example speech signal, estimated maximum and minimum energy and threshold curves.

long enough to provide continuous excitation. The LPC analyzing part estimates autoregressive coefficients using Itakura's algorithm [9]. They are first stored into the coefficient register and then transmitted to the synthesizing filter. The filter generates continuously samples of the comfort noise using the excitation signal from the noise buffer and the transfer function inversed to the one that has been estimated before. Finally the signal from the synthesizing filter is added to the signal coming out from the echo suppressor. An output signal S_{out} is formed and sent to the line.

IV. EXPERIMENTAL RESULTS

Following section presents results of experiments that were carried out to investigate the performance of the proposed algorithm on real speech signals. Simulations were made with the help of Matlab environment and audio visualization software GoldWave250. Real speech signals from far-end and near-end speakers were used as an input to the echo canceller. All signals were ten seconds in duration with a sampling rate of 8000 Hz ($8 \cdot 10^4$ samples). Fig. 3 shows the speech signal at the input of the echo canceller.

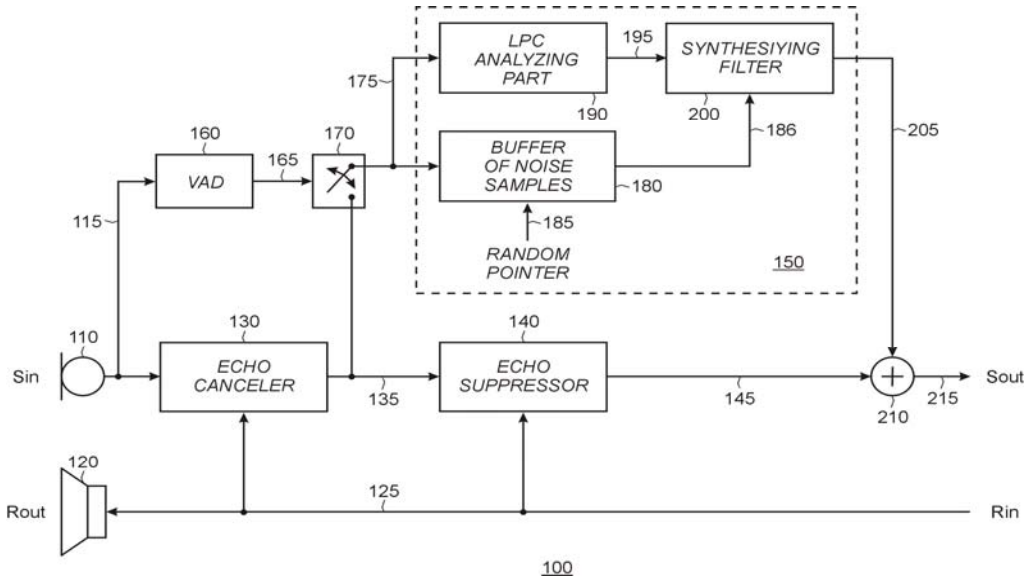

Figure 2. Comfort noise generating system.

It consists of the speech of the near-end speaker and the far-end echo. Fig. 4 shows the output signal from the echo canceller with unsuppressed residual echo. Fig. 5 contains the signal coming out from the echo suppressor. It could be seen that residual echo was suppressed together with the background noise. The result is a channel that suddenly sounds dead. The far-end listener may think that the call was disconnected. The proposed CNG-VAD algorithm was designed to prevent this. Fig. 6 shows the signal S_{out} at the output of the CNG-VAD system. The suppressed background noise is successfully replaced by the artificially generated comfort noise.

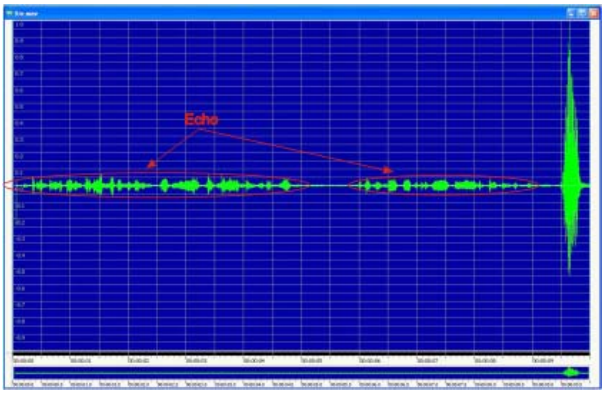

Figure 3. Speech signal at the input of the echo canceller.

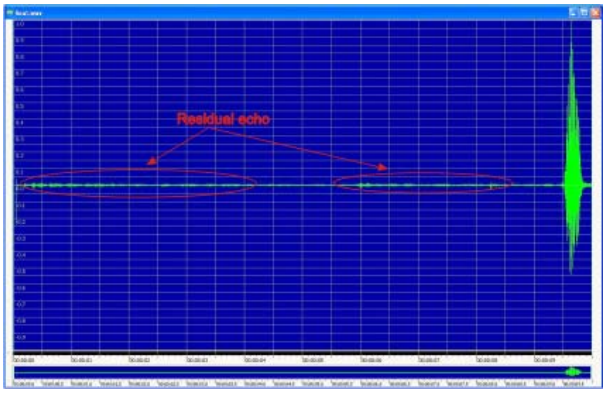

Figure 4. Output signal from the echo canceller.

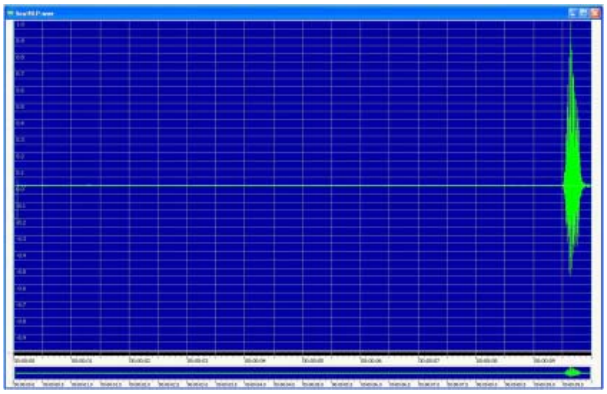

Figure 5. Output signal from the echo suppressor.

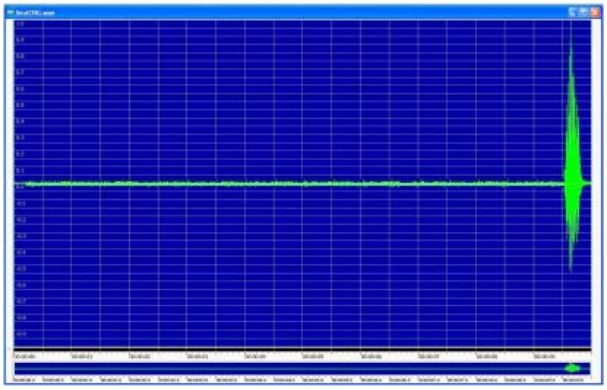

Figure 6. Output signal from the comfort noise generator.

V. CONCLUSIONS

This article is a forecast on comfort noise generating approach to insert synthesized noise instead of clipped speech segments during echo suppression procedure. An alternative method and apparatus for comfort noise generation is introduced. The presented background noise model is based on a set of noise model parameters which are in turn based on measurements of actual background noise in the echo suppression system. The Itakura's LPC algorithm is used for parametrically modeling background noise. As a result, the comfort noise closely matches the actual background noise in terms of both character and level. It does not sound artificially. Consequently, the far-end user perceives signal continuity and is not distracted by the artifacts introduced by conventional methods. The alternative energy-based voice activity detector is also introduced. The expounded algorithm is universal and easily can be integrated into most voice activity detectors used by vocoders and other speech enhancement systems.

VI. ACKNOWLEDGMENT

Research described in the paper was supported by the Ministry of Education, Youth and Sports of the Czech Republic by the research program MSM6840770014 and the CTU grand No.OHK3-108/10.

REFERENCES

[1] J. A. Rasmusson, "Method and apparatus for echo reduction in a hands-free cellular radio cimmunication system," WO/1996/022651, 1995.

[2] E. D. Rosemburg, J. A. Rasmusson, "Method and apparatus for improved echo suppresion in communication systems," WO/1999/035814, 1999.

[3] E. D. Rosemburg, "Echo canceller for use in communications system," US Patent 6 185 300, 2001.

[4] P. Sovka, P. Pollak, Selected digital signal processing methods, Prague, Czech Republic, 2003.

[5] E. D. Rosemburg, L. S. Bioebaum, C. N. S. Guruparan, "Method and apparatus for prividing comfort noise in communication systems," US Patent 6 163 608, 2000.

[6] S. Gupta, P. K. Gupta, B. Kepley, "Comfort noise generator for echo cancellers," US Patent 5 949 888, 1999.

[7] J. A. Stephens, D. L. Barron, S. S. You, "Low-complexity comfort noise generator," US Patent 7 243 065, 2007.

[8] H. Torrel, Voice activity detection in the Tiger platform M. S. Thesis, Linkoping, Sweden, 2006.

[9] P. Venkatesha, R. Sangwan, A. Jamadagni, "Comparison of voice activity detection for VoIP," proc. of the Seventh International Symposium on Computers and Communications – ISCC 2002, Taormina, Italy, pp. 530-532, 2002.

[10] P. Pollak, P. Sovka, J. Uhlir, "Noise system for a car," proc. of the Third European Conference on Speech, Communication and Technology – EUROSPEECH'93, Berlin, Germany, pp. 1073-1076, Sept. 1993.

[11] P. Renevev, A. Drygailo, "Entropy based voice activity detection in very noise conditions," proc. of the Seventh European Conference on Speech Communication and technology – EUROSPEECH 2001, Aalborg, Denmark, pp.1883-1886, 2001.

[12] A. Kindoz, A. M. Kondoz, Digital Speech; Coding for Low Bit Rate Communication Systems. John Wiley & Sons, Inc., New York, NY, 2004.