ACOUSTIC ECHO SUPPRESSION BASED ON SEPARATION OF STATIONARY AND NON-STATIONARY ECHO COMPONENTS

Fabian Kuech [1]*, Markus Kallinger*[1]*, Markus Schmidt*[1]*, Christof Faller*[2] *and Alexis Favrot* [2]

fabian.kuech@iis.fraunhofer.de
[1] Fraunhofer IIS, Am Wolfsmantel 33, 91058 Erlangen, Germany
[2] Illusonic LLC, Chemin de Trabandan 28 A, 1006 Lausanne, Switzerland

ABSTRACT

Hands-free telecommunication systems require acoustic echo control to cancel echoes that arise from a coupling between loudspeakers and microphones. Acoustic echo suppression (AES) represents a robust and efficient approach to cope with such echoes. Since AES applies a frequency selective attenuation of the microphone signal, it may also affect the near-end speech quality in case of non-ideal behavior of the AES. In this contribution we present a method to assure reliable suppression of echoes, while minimizing distortions of the near-end speech. The proposed approach is based on performing the suppression separately for stationary and non-stationary echo components. This allows for different optimization strategies for determining the corresponding echo suppression rules.

Index Terms— Acoustic Echo Cancellation, Acoustic Echo Suppression, Spectral Subtraction.

1. INTRODUCTION

In telecommunication systems acoustic echoes result from an acoustic feedback of the loudspeaker signal to the microphone. Echo signals represent a very distracting disturbance and can inhibit interactive, full-duplex communication. This is especially true for systems including large delay [1], as e.g., todays modern voice over IP (VoIP) telecommunication systems. Besides their annoyance, acoustic echoes can also cause howling due to instability effects in electro-acoustic feedback loops. Figure 1 illustrates the general set-up of the acoustic echo control problem. A conventional approach to cope

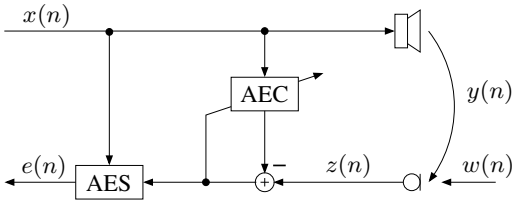

Fig. 1. Set-up of the acoustic echo cancellation problem.

with echoes is to place an acoustic echo canceler (AEC) in parallel to the propagation path of the echo signal. In the AEC, a digital replica of the echo signal is estimated which is then subtracted from the recorded microphone signal. In standard approaches, the acoustic echo path is modeled by a linear filter, and the AEC is implemented accordingly, i.e., as a linear adaptive FIR filter [1]. To model typical echo paths, FIR filters of lengths up to some hundreds of milliseconds are required which also implies high computational complexity.

In practice, however, the achievable echo attenuation of the AEC is often not sufficient due to, e.g., an undermodeling of the echo path or convergence problems of the adaptive filter. Commonly, these residual echoes are attenuated by using acoustic echo suppression (AES). AES is usually realized as a nonlinear post-processor which attenuates the residual echoes in a frequency selective way [2].

Recently, a number of AES approaches have been proposed that are similar to the aforementioned residual echo suppression, when completely discarding the AEC part [3, 4]. These approaches do not require the identification of the room impulse response, but only use a rough estimates of the echo power spectrum transfer function. Thus, these methods are computationally far less complex than conventional AECs, especially for high sampling rates and long room impulse responses. Moreover, these approaches provide robust suppression of echoes even in highly time variant acoustic environments.

The weakness of the AES systems [3, 4] is that their performance is not satisfactory in case of stationary components included in the loudspeaker signal. These approaches do not distinguish between stationary and non-stationary echo components. Consequently, noise-like echoes are suppressed as aggressively as speech echoes, although they are not perceived as annoying. Unfortunately, such aggressive suppression of stationary echo components leads to severe distortions of the near-end speech signal. In this paper we propose a method to mitigate near-end speech distortions, while still providing sufficient echo attenuation. This is achieved by performing the echo suppression separately for stationary and non-stationary echo components, where different attenuation strategies are applied.

This paper is organized as follows. In Section 2 we summarize the concept of AES. The proposed method for determining the desired echo suppression filters is presented in Section 3. Simulation results that illustrate the performance of the method and the differences compared to a standard single filter approach are discussed in Section 4.

2. ACOUSTIC ECHO SUPPRESSION

In this section we briefly recall the general approach of acoustic echo suppression. Thereby, we basically follow the method as described in [4].

The microphone signal $z(n)$ is composed of the acoustic echo signal $y(n)$ that results from the feedback of the loudspeaker signal $x(n)$ and the near-end signal $w(n)$. Here, we assume that the room impulse response can be expressed as the combination of a direct propagation path, corresponding to a delay of d samples between the loudspeaker signal and the microphone signal, and a linear filter g_n which models the acoustic properties of the enclosure. Then, the

microphone signal $z(n)$ can be expressed by

$$z(n) = g_n * x(n-d) + w(n), \tag{1}$$

where $*$ denotes convolution. The short-time Fourier transform (STFT) domain representation of (1) is given by

$$Z(k,m) = G(k,m)X_d(k,m) + W(k,m), \tag{2}$$

where k is the block time index and m denotes the frequency index. $X_d(k,m)$ is defined as the the STFT-domain correspondence of the delayed loudspeaker signal $x(n-d)$.

The acoustic echo suppression is performed by modifying the magnitude of the STFT of the microphone signal $Z(k,m)$, while keeping its phase unchanged. This can be expressed by

$$E(k,m) = H(k,m)Z(k,m), \tag{3}$$

where $H(k,m)$ represents a real-valued, positive attenuation factor. In the following we refer to $H(k,m)$ as echo suppression filter (ESF). The optimum values for the ESF $H(k,m)$ can be derived by minimizing the contribution of the echo components to the output signal $E(k,m)$ in the mean square error (MSE) sense. It is reasonable to assume that the near-end signal $W(k,m)$ and the loudspeaker signal $X(k,m)$ are uncorrelated. Then, regarding the results derived in [5], we obtain

$$H_{\text{opt}}(k,m) = \frac{\mathrm{E}\{|Z(k,m)|^2\} - \mathrm{E}\{|Y(k,m)|^2\}}{\mathrm{E}\{|Z(k,m)|^2\}}, \tag{4}$$

where $\mathrm{E}\{\cdot\}$ denotes the expectation operator, and where

$$Y(k,m) = G(k,m)X_d(k,m) \tag{5}$$

denotes the STFT of the echo components in the microphone signal. Note that, in practice, both the echo power transfer function $|G(k,m)|^2$ and the delay d are not known and have to be replaced by corresponding estimates.

A practical approach for the computation of the echo suppression filter $H(k,m)$ is based on generalized, instantaneous versions of (4). In [4] it has, e.g., been proposed to use a parametric spectral subtraction approach analogously to [6]:

$$H(k,m) = \left(\frac{|Z(k,m)|^\alpha - \beta|\hat{Y}(k,m)|^\alpha}{|Z(k,m)|^\alpha} \right)^{\frac{1}{\gamma}}, \tag{6}$$

where α, β, and γ represent design parameters to control the echo suppression performance. Parameter β will be addressed explicitly in Section 3. For presentational convenience, we assume $\alpha = 2$ and $\gamma = 1$ throughout this paper, i.e., $H(k,m)$ can be considered as an estimate of $H_{\text{opt}}(k,m)$ according to (4). The estimate of the power spectrum of the echo signal is obtained by

$$|\hat{Y}(k,m)|^2 = |\hat{G}(k,m)|^2 |X_d(k,m)|^2, \tag{7}$$

where $|\hat{G}(k,m)|^2$ represents an estimate of the echo power transfer function $|G(k,m)|^2$. A method for estimating the echo power transfer function is, e.g., presented in [4], and an improved solution, which is robust against near-end noise, is proposed in [7].

The final echo suppression is based on a temporally smoothed version of $H(k,m)$ in order to avoid artifacts due to fast varying gain values and to keep the echo attenuation high for a longer period of time. Thus, (3) becomes

$$E(k,m) = H_{\text{sm}}(k,m)Z(k,m), \tag{8}$$

where $H_{\text{sm}}(k,m)$ represents a time-smoothed version of $H(k,m)$.

3. PROPOSED ECHO SUPPRESSION FILTER

The estimate of the echo power spectrum $|\hat{Y}(k,m)|^2$ according to (7) is usually not very accurate, since only a fraction of the length of the true echo path can be covered. To prevent that these inaccuracies result in residual echoes, the ESF in (8) is computed to attenuate the microphone signal aggressively such that no residual echo remains. This is, e.g., achieved by intentionally over-estimating the echo power spectrum in (8) and by a suitable time-smoothing of $H(k,m)$. When the loudspeaker signal contains stationary components such as noise, an ESF based on (6), (8) also aims at suppressing the corresponding echo components. However, due to the required aggressive tuning of the ESF, this may also lead to a significant impairment of the desired near-end speech signal. Additionally, artifacts like musical tones, as known from noise reduction [8], can be caused by inappropriate echo suppression filters. In this section, we present a method for designing the ESF such that a trade-off between echo suppression and near-end speech distortion is achieved that can be tuned to a given preference.

For the following discussions it will be useful to express the delayed loudspeaker signal $X_d(k,m)$ as the superposition of stationary signal components $X_{\text{s},d}(k,m)$ and non-stationary signal components $X_{\text{ns},d}(k,m)$:

$$X_d(k,m) = X_{\text{ns},d}(k,m) + X_{\text{s},d}(k,m). \tag{9}$$

In practice, the non-stationary signal components mainly correspond to the far-end speech signal, whereas the stationary component, e.g., consists of background noise picked up by the microphone at the far-end. Thus, it is reasonable to assume that both components are uncorrelated. Introducing the signal model (9) in (6), we obtain

$$H(k,m) = \Big(|Z(k,m)|^2 - \beta|\hat{G}(k,m)|^2 |X_{\text{ns},d}(k,m)|^2 - \beta|\hat{G}(k,m)|^2 |X_{\text{s},d}(k,m)|^2 \Big) / |Z(k,m)|^2, \tag{10}$$

where $\alpha = 2$ and $\gamma = 1$ has been used.

In case of the ESF according to (6), smoothing or any other post-processing of $H(k,m)$ cannot be performed differently for the different echo components. However, different signal characteristics cause different distortions or artifacts in the context of spectral subtraction. Therefore, it would be preferable to optimize the suppression of the stationary and non-stationary echo components separately. Non-stationary echoes resulting from speech need to be removed aggressively in order to prevent audible residual echoes. On the other hand, stationary echoes are not perceived as annoying as speech echoes. However, such components also contribute to howling effects due to feedback loops. Stationary echoes should therefore be removed less aggressively in order to avoid distortions in the desired near-end signal.

The echo suppression approach proposed in this paper is based on two different ESFs for stationary and non-stationary echo components, respectively. This is achieved by introducing a generalized version of (10) according to

$$H_{\text{g}}(k,m) = \min\left\{ H_{\text{ns}}(k,m),\ H_{\text{s}}(k,m) \right\}, \tag{11}$$

where the ESFs $H_{\text{ns}}(k,m)$ and $H_{\text{s}}(k,m)$ are defined as

$$H_{\text{ns}}(k,m) = \frac{|Z(k,m)|^2 - \beta_{\text{ns}}|\hat{G}(k,m)|^2 |X_{\text{ns},d}(k,m)|^2}{|Y(k,m)|^2}, \tag{12}$$

$$H_{\text{s}}(k,m) = \frac{|Z(k,m)|^2 - \beta_{\text{s}}|\hat{G}(k,m)|^2 |X_{\text{s},d}(k,m)|^2}{|Y(k,m)|^2}. \tag{13}$$

The advantage of the separate formulation of the ESFs is that it allows for different optimization strategies for the different ESFs $H_{\mathrm{ns}}(k,m)$ and $H_{\mathrm{s}}(k,m)$, respectively. To give an example, the choice of $\beta_{\mathrm{ns}} > \beta_{\mathrm{s}}$ makes the suppression of non-stationary echoes more aggressive than for stationary echoes. In practical implementations, the maximum attenuation introduced by $H_{\mathrm{ns}}(k,m)$ and $H_{\mathrm{s}}(k,m)$, respectively, is usually limited by introducing corresponding minimum values. Accounting for their different tasks, $H_{\mathrm{ns}}(k,m)$ should be allowed to have minimum values of -40 dB or even down to -60 dB, whereas -15 dB is already sufficient for suppressing stationary echo components.

It should be pointed out that, analogously to (8), the actual echo suppression is not performed by directly applying the echo removal filters $H_{\mathrm{ns}}(k,m)$ and $H_{\mathrm{s}}(k,m)$, but it is based on corresponding temporally smoothed versions instead. It also should be pointed out that the temporal smoothing parameters can be optimized separately for the suppression of non-stationary and stationary echo components, respectively.

In the following, we illustrate the different optimization strategies by a realistic example. By introducing the signal-to-echo-estimate-ratio $\mathrm{SER}_{\mathrm{ns}}$ for the non-stationary echo component

$$\mathrm{SER}_{\mathrm{ns}} = \frac{|Z(k,m)|^2}{\beta_{\mathrm{ns}}|\hat{G}(k,m)|^2|X_{\mathrm{ns},d}(k,m)|^2}, \tag{14}$$

and a corresponding definition for the stationary components $\mathrm{SER}_{\mathrm{s}}$, we can rewrite (12), and (13) according to

$$H_{\mathrm{ns}}(k,m) = 1 - {\mathrm{SER}_{\mathrm{ns}}}^{-1}, \tag{15}$$

$$H_{\mathrm{s}}(k,m) = 1 - {\mathrm{SER}_{\mathrm{s}}}^{-1}. \tag{16}$$

In Figure 2, $H_{\mathrm{ns}}(k,m)$ and $H_{\mathrm{s}}(k,m)$ are compared with each other as a function for the corresponding SER, where $\beta_{\mathrm{ns}} = 6$ and $\beta_{\mathrm{s}} = 2$ has been chosen. Furthermore, $H_{\mathrm{ns}}(k,m)$ has been limited to -40 dB, whereas the minimum value of $H_{\mathrm{s}}(k,m)$ has been set to -15 dB. As can be seen, $H_{\mathrm{ns}}(k,m)$ becomes small already for moderate

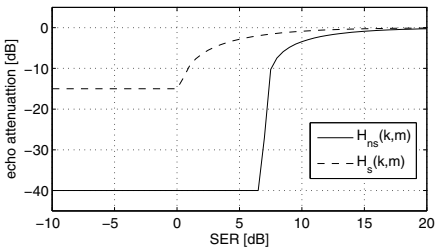

Fig. 2. Comparison of the characteristics of $H_{\mathrm{ns}}(k,m)$ and $H_{\mathrm{s}}(k,m)$ for typical parameter settings.

SER values in order to suppress undesired speech echoes reliably. Note that values of $\mathrm{SER} < 0$ dB are well possible due to the high overestimation factor β_{ns}. On the other hand, the attenuation by $H_{\mathrm{s}}(k,m)$ increases only slowly with increasing echo power and is clearly limited. The behavior of $H_{\mathrm{s}}(k,m)$ basically corresponds to the suppression rule applied in noise reduction for speech enhancement [1].

Let us now look at the estimation of the stationary and non-stationary signal components in the loudspeaker signal. The separation of stationary noise from non-stationary speech signals represents a well established signal processing task in speech enhancement [1, 6]. Thus, we can directly apply these methods for the discrimination of the different loudspeaker signal components. Following the approach of parametric spectral subtraction [6], an estimate of the power spectrum of the non-stationary signal component $|\hat{X}_{\mathrm{ns},d}(k,m)|^2$ is obtained by

$$\left|\hat{X}_{\mathrm{ns},d}(k,m)\right| = F(k,m)\,|X_d(k,m)|\,, \tag{17}$$

where the filter $F(k,m)$ is defined analogously to (6), i.e.,

$$F(k,m) = \frac{|X_d(k,m)|^2 - \beta_{\mathrm{x}}\left|\hat{X}_{\mathrm{s},d}(k,m)\right|^2}{|X_d(k,m)|^2}. \tag{18}$$

The power spectrum of the stationary loudspeaker signal component $|\hat{X}_{\mathrm{s},d}(k,m)|^2$ can be obtained using well known methods such as minimum statistics or spectral envelope smoothing techniques [1].

It should be pointed out that the separation described above is especially relevant in presence of near-end speech, where aggressive suppression of stationary echo component would impair the quality of the desired near-end signal. On the other hand, this is not crucial if only far-end speech is present, where aggressive suppression of echoes is required. The degree of separation can be easily controlled by the parameter β_x in (18), where, e.g., a value of $\beta_x = 0$ corresponds to the case for which no separation is performed at all. In other words, the parameter β_x can be used to tune the aggressiveness of the echo suppression approach to any given preference.

4. SIMULATION RESULTS

In the following we present simulation results for an AES scenario to illustrate the performance of the proposed method. The evaluation is based on comparing the single ESF that results for the standard approach according to [3, 4], and the corresponding ESFs that are obtained for the stationary and non-stationary signal components.

In the simulations, the loudspeaker signal is composed of speech and additive colored noise with a signal-to-noise ratio (SNR) of 20 dB. The echo signal has been obtained by convolving the loudspeaker signal with a room impulse response that has been measured in an office room. The near-end signal is composed of a speech signal and additive colored noise with an SNR of 30 dB. The conversation sequence can basically be divided into three periods of equal length: In the first period only the far-end talker is active, whereas in the second, only near-end speech is present. The last period represents a double-talk situation, i.e., both, the far-end and the near-end speaker are talking simultaneously.

The spectrogram of the microphone signal is depicted in Fig. 3(a). In Fig. 3(b), the spectrogram of the overall loudspeaker signal $|X(k,m)|^2$ is shown and Figs. 3(c) and 3(d) show the corresponding estimates of the non-stationary and stationary components, $|X_{\mathrm{ns}}(k,m)|^2$ and $|X_{\mathrm{s}}(k,m)|^2$, respectively. The estimation of $|X_{\mathrm{s}}(k,m)|^2$ has been performed based on variable envelope smoothing as discussed in [1]. For the separation according to (17), (18) a fixed value of $\beta_{\mathrm{x}} = 2$ has been used.

Fig. 4(b) shows the ESF $H_{\mathrm{ns}}(k,m)$ for the non-stationary echo components according to (12) that has been obtained for the considered conversation sequence. To assure sufficient echo suppression, a large over-estimation factor $\beta_{\mathrm{ns}} = 6$ has been chosen. Additionally, the temporal smoothing of the ESF has been performed such that the echo attenuation is kept sufficiently high during far-end speech activity. As already indicated by the characteristics of $H_{\mathrm{ns}}(k,m)$ in Fig. 2, the attenuation basically toggles between 0 dB and the pre-defined minimum value of -40 dB: whenever there is speech included in

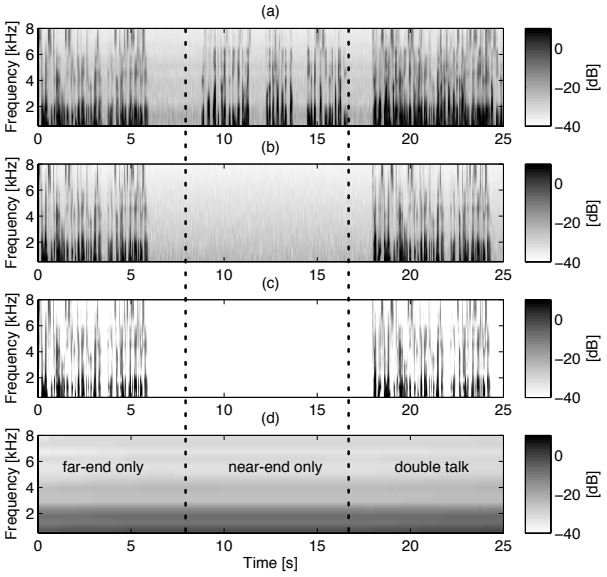

Fig. 3. Illustration of the spectrograms of the microphone signal (a), the loudspeaker signal (b), and the corresponding estimates of the non-stationary and stationary loudspeaker signal components.

the loudspeaker signal, the echo suppression is high, whereas during speech pauses, almost no attenuation is performed. As can be seen from Fig. 4 (c), the ESF $H_{\mathrm{s}}(k,m)$ according to (12) basically shows a reversed behavior. Due to dominant speech echoes included in the microphone signal, $H_{\mathrm{s}}(k,m)$ equals to 1 in case of far-end speech. The same obviously also applies in case of an active near-end talker. However, in periods of significant levels of stationary echo components in the microphone signal, the ESF takes its minimum value of -15 dB in order to avoid howling effects due to instable electroacoustic feedback loops. Note that the temporal smoothing of $H_{\mathrm{s}}(k,m)$ is performed such that near-end speech distortions and artifacts such as musical tones are kept as low as possible. The actual ESF used for the echo suppression according to (3) is obtained as the minimum of both ESFs according to (11).

The single temporally smoothed version of the ESF that is obtained from a joint processing of both stationary and non-stationary echo components according to [3, 4] is shown in Fig. 4 (a). The smoothing has been performed equivalently to the smoothing of $H_{\mathrm{ns}}(k,m)$ in order to provide sufficient echo attenuation. As can be seen, the stationary components included in the loudspeaker signal result in a maximum attenuation for a much larger area in the time-frequency plane. While this is beneficial with respect to echo attenuation in case of far-end speech only, this behavior is not desired during the period of double-talk. As the areas of high attenuation in Fig. 4 (a) indicate, the distortion of the near-end speech signal is much higher than in case of the separate ESFs. It can also be noted that during periods of near-end talk only (or noise only), the variation of the values of the single ESF is very high, leading to perceivable and annoying artifacts such as musical tones [8] . This behavior can also be avoided for the combined ESF, since it is mainly determined by its stationary part $H_{\mathrm{ns}}(k,m)$.

5. CONCLUSION

In this contribution we present a method to acoustic echo suppression that performs a separate processing of stationary and non-stationary echo components. This allows for an explicit optimization of the corresponding echo suppression filters with respect to the perceived quality of the desired near-end signal. The simulation results confirm that during both single near-end talk and double-talk situations, distortions of the microphone signal can be significantly reduced compared to standard AES approaches.

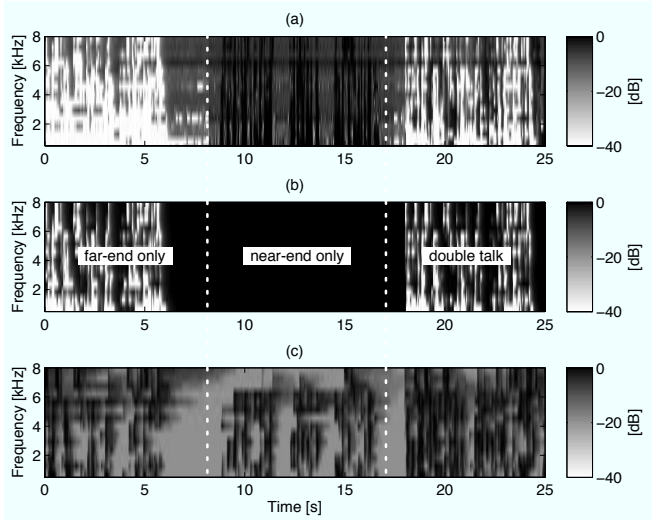

Fig. 4. Illustration of different ESFs obtained for the simulations. (a): joint processing of all loudspeaker signal components [3,4]. (b): $H_{\mathrm{ns}}(k,m)$ according to (12). (c): $H_{\mathrm{s}}(k,m)$ according to (13).

6. REFERENCES

[1] G. Schmidt and E. Hänsler, *Acoustic echo and noise control: a practical approach*, Hoboken: Wiley, 2004.

[2] S. Gustafsson, R. Martin, and P. Vary, "Combined acoustic echo control and noise reduction for hands-free telephony," *Signal Processing*, vol. 64, no. 1, pp. 21–32, 1998.

[3] C. Faller and J. Chen, "Suppressing acoustic echo in a sampled auditory envelope space," *IEEE Trans. on Speech and Audio Proc.*, vol. 13, no. 5, pp. 1048–1062, Sept. 2005.

[4] C. Faller and C. Tournery, "Estimating the delay and coloration effect of the acoustic echo path for low complexity echo suppression," in *Proc. Intl. Works. on Acoust. Echo and Noise Control (IWAENC)*, Eindhoven, Sept. 2005.

[5] S. Haykin, *Adaptive Filter Theory*, New Jersey: Prentice Hall, 1996.

[6] W. Etter and G. S. Moschytz, "Noise reduction by noise-adaptive spectral magnitude expansion," *J. Audio Eng. Soc.*, vol. 42, pp. 341–349, May 1994.

[7] A. Favrot *et. al.*, "Acoustic echo control based on temporal fluctuations of short-time spectra," in *Proc. Intl. Works. on Acoust. Echo and Noise Control (IWAENC)*, Seattle, Sept. 2008.

[8] O. Cappé, "Elimination of the musical noise phenomenon with the ephrain and malah noise suppressor," *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 2, pp. 345–349, April 1994.