STATIONARY-TONES INTERFERENCE CANCELLATION USING ADAPTIVE TRACKING

Ivan Tashev and Henrique S. Malvar

Microsoft Research, One Microsoft Way, Redmond, WA 98052, USA
{ivantash, malvar}@microsoft.com

ABSTRACT

It is usual in practice that recorded sounds are contaminated by stationary tones coming from power wiring (50/60 Hz or 400 Hz and their harmonics), frame or line frequencies from a nearby TV or monitor, computer fans, hard disk drives, etc. They are mostly stationary, but their removal using a stationary noise suppressor results in notch filtering, removing the speech content at those frequencies, because the SNRs are usually low. In this paper we propose fast, real-time algorithm for removing constant tones while keeping intact the speech content. We build and adaptively update a model of the constant tones, extrapolating it for subtraction from the next frame. In our evaluations, the proposed algorithm reduces the amplitudes of such stationary tones up to 15 dB, without introducing artifacts such as nonlinear distortions or musical noises. This algorithm is applicable as pre-processor before a classic gain-based stationary noise suppressor.

Index Terms **—** Signal Restoration, Speech Enhancement, Acoustic Signal Processing, Estimation, Extrapolation.

1. INTRODUCTION

Frequently speech recordings are contaminated by stationary tones. They usually come from power wiring, inadequate shielding or grounding of the microphone cables, or placement of the microphones near power lines or transformers. In those cases the interference frequency is 50/60 Hz or 400 Hz and their harmonics. Other kinds of stationary-tone interferences come form microphones positioned near TVs, monitors, or video cameras; the microphones can capture interference at frame or line frequencies acoustically from transformers or electronically from the cables. Yet another source of this kind of interferences are noises coming from the acoustical environment, such as fans, computer hard drives, and air conditioning. Because of nonlinearities and room reverberation, these signals behave mostly as random zero mean Gaussian noise, but usually there are still predictable components. The frequencies of the predictable portion of these noises vary depending of the fan or hard disk spindle rotating speed. The common property of all of these signals is that they are practically stationary. In their time-frequency representations they show up as horizontal lines with constant amplitude.

The most intuitive approach to solve this problem and to clean up the contaminated signal is to apply band pass filtering or notch filters tuned to the constant tones. These approaches remove speech signal components if the interfering frequencies are within the speech band. If the speech signal is contaminated by single-tone interference, then a notch filter works well and the missing frequency is usually inaudible. If the contaminating signal has multiple harmonics, then a set of notch filters or a comb filter may be needed to achieve significant filtering, and that can distort substantially the speech signal.

Classic noise suppressors assume the noise is stationary zero mean Gaussian process and build a statistical model of the noise as vector of variances per frequency bin. The stationary tones have probability density function (PDF) that is usually not Gaussian. Using a Gaussian PDF as a model of these signals and some of the known suppression rules (Wiener [1] or Ephraim and Malah [2], etc.) results in complete suppression of the speech signals in these frequency bins, i.e. the noise suppressor converts to a notch filter for these frequencies.

The problem of tracking frequencies in time-frequency representation is well studied. In [3] an ARCAP method is used (AR – autoregressive, CAP – Capone algorithm) to identify the spectral lines, followed by a Kalman filtering to track their movement. The method is illustrated with processing of avalanche signals. It is sensitive to noise and best results are obtained with forward-backward Kalman filter, which makes it inapplicable for real-time algorithms where low latency is desired. Improving the algorithm further [4] by adding trajectory smoothing with a Fraser filter still keeps the algorithm good for off-line processing only. The birth/dead time estimation of spectral lines is improved in this paper as well. In [5] a particle filter is used to perform optimal estimation in jump Markov systems for detection and tracking of spectral lines. The proposed time-varying

autoregressive (TVAR) estimator is evaluated with synthetic signals. It is computationally expensive and sensitive to the times of birth/death of spectral lines. In [6] image processing techniques are used to detect, model and remove spectral lines from time-frequency representation. All of these approaches solve problems that are more complex than necessary, and are mostly suitable for off-line processing of the contaminated signals.

In this paper we propose computationally inexpensive real-time algorithm for stationary-tone interference removal, which is applicable as pre-processor to a conventional noise suppressor. It is based on adaptive building and updating of a model of the constant tones with consequent extrapolation and subtraction from the next audio frame.

2. MODELING AND EXTRAPOLATION

We assume the contaminating signal as stationary or pseudo-stationary, i.e. its spectral changes are much slower than those of the speech signal. All the processing is done in frequency domain, as in most of the audio processing systems today, which makes the proposed algorithm easily pluggable into an existing frequency-domain noise suppression system. We process each frequency bin separately, assuming they are statistically independent, which is not quite true in this case, but proper measures are taken to reflect the nature of correlated neighbor bins.

2.1. A model for the contaminating signal

We consider the contaminating signal as a linear combination of sinusoidal signals and noise:

$$z(t) = \sum_{i=1}^{L} A_i \sin(2\pi f_i t) + \mathbb{N}(0, \lambda) \qquad (1)$$

where L is the number of stationary tones, each with frequency f_i. Converting this signal to frequency domain yields the following model for the n-th audio frame:

$$Z_k^{(n)} = \sum_{i=1}^{L} W_T(k) * A_i e^{-j2\pi n T f_i} + \mathbb{N}(0, \lambda_N) \qquad (2)$$

where W_T is the Fourier image of the frame weighting function, T is the audio frame step, n is the frame number and k is the frequency bin.

We note the following:

- Due to the "smearing" of the spectral lines because of the weighting, bins neighboring the central bin (for each contaminating frequency) contain portions of the energy.
- These neighboring bins will rotate in the complex plane (phase shift) from frame to frame with the same speed, which can be different than the speed of the each bin's central frequency $e^{-j2\pi n T f_s / K}$.

These two aspects introduce additional complications in the extrapolation of the signal model for the next frame.

2.2. Extrapolating the contaminating signal

Assuming we have perfect estimation $\hat{Z}_k^{(n-1)}$ for frame (n-1), then the extrapolation for the n-th frame will be:

$$\hat{Z}_k^{(n)} = \hat{Z}_k^{(n-1)} \frac{\sum_{i=1}^{L} W_T(k) * A_i e^{-j2\pi n T f_i}}{\sum_{i=1}^{L} W_T(k) * A_i e^{-j2\pi (n-1) T f_i}}. \qquad (3)$$

The second term is a complex number that represents the "speed" of rotation of our complex model from frame to frame. As it was noted in 2.1 this "speed" can be different than the "speed" of the central frequency of the bin. Because $W_T(k)$ decays quickly with increasing k, we can assume that one frequency from the contaminating signal dominates in each frequency bin. In this case

$$\frac{\sum_{i=1}^{L} W_T(k) * A_i e^{-j2\pi n T f_i}}{\sum_{i=1}^{L} W_T(k) * A_i e^{-j2\pi (n-1) T f_i}} \approx e^{-j2\pi T f_I} + \mathbb{N}(0, \lambda_E). \qquad (4)$$

where f_I is the dominant frequency and $\mathbb{N}(0, \lambda_E)$ is an error term, modeled as zero mean Gaussian noise. As the dominant frequency is unknown the extrapolation can be presented as:

$$\hat{Z}_k^{(n)} = \hat{Z}_k^{(n-1)} \hat{Y}_k^{(n-1)} \qquad (5)$$

where $\hat{Z}_k^{(n-1)}$ is the contaminating signal estimation for frame (n-1), and $\hat{Y}_k^{(n-1)}$ is the rotating "speed" of the model towards the next frame. Both components have additive Gaussian noise with variances λ_N and λ_E correspondingly.

3. CANCELLATION AND MODEL UPDATE

With speech signal $s(t)$ presented, (1) takes the form of

$$x(t) = s(t) + \sum_{i=1}^{L} A_i \sin(2\pi f_i t) + \mathbb{N}(0, \lambda). \qquad (6)$$

If the speech signal spectrum is $S_k^{(n)}$, the representation in frequency domain of the n-th frame is

$$X_k^{(n)} = S_k^{(n)} + \sum_{i=1}^{L} W_T(k) * A_i e^{-j2\pi n T f_i} + \mathbb{N}(0, \lambda_N). \qquad (7)$$

3.1. Contaminating signal cancellation

In this case our estimation of the speech signal is

$$\hat{S}_k^{(n)} = X_k^{(n)} - \hat{Z}_k^{(n)}, \qquad (8)$$

i.e. we just subtract the contaminating signal, estimated according to (5). The speech signal estimation contains the captured noise $\mathbb{N}(0, \lambda_N)$ and the cancellation adds additional noise component $\sim \mathbb{N}(0, \lambda_E)$ due to the approximations in the model and estimation errors.

3.2. Updating the model

In parallel with the contaminating signal cancellation, we should constantly update the contaminating signal model, which for each frequency bin consists of four elements: $\hat{Z}(k)$, $\hat{Y}(k)$, $\lambda_N(k)$, and $\lambda_E(k)$ (from which only the first two are involved in the constant tones cancellation process). The contaminating signal model is updated as follows:

$$\hat{Z}_k^{(n)} = (1-\alpha)\hat{Z}_k^{(n-1)} + \alpha\left(p_k^{(n)} X_k^{(n)} + \left(1-p_k^{(n)}\right)\hat{Z}_k^{(n-1)}\right), \quad (9)$$

where $\alpha = \frac{T}{\tau_Z}$, τ_Z is the adaptation time constant, and $p_k^{(n)}$ is the probability that we have only contaminating signal in $X_k^{(n)}$, i.e. the probability of speech absence. It can be provided by a voice activity detector (VAD), which produces per-bin probability estimation of speech presence.

The additive noise variance is updated as follows:

$$\lambda_N^{(n)} = (1-\alpha)\lambda_N^{(n-1)} + \alpha\left(p_k^{(n)}\delta_k^{(n)} + \left(1-p_k^{(n)}\right)\lambda_N^{(n-1)}\right). \quad (10)$$

where $\delta_k^{(n)} = \left\| X_k^{(n)} - \hat{Z}_k^{(n)} \right\|^2$.

The rotating speed estimation is updated in the same way:

$$\hat{Y}_k^{(n)} = (1-\beta)\hat{Y}_k^{(n-1)} + \beta\left(p_k Y_{mom}^{(n)}(k) + (1-p_k)\hat{Y}_k^{(n-1)}\right) \quad (11)$$

where $Y_{mom}^{(n)}(k) = \frac{Y_k}{\|Y_k\| + \varepsilon}$ is the normalized momentary rotation speed estimation $Y_k = \frac{X_k^{(n)}}{X_k^{(n-1)} + \varepsilon}$ for the current frame, ε is a small number, $\beta = \frac{T}{\tau_Y}$, and τ_Y is the adaptation time constant.

4. EXPERIMENTAL RESULTS

For frequency domain transformation we used the modulated complex lapped transform (MCLT), which is similar to a windowed DFT filter bank and has been successfully used for many frequency-domain processing algorithms that require efficient signal reconstruction [7]. In our experiments we used 16 kHz sampling rate, 16-bit precision, and a block size of 320 samples (40 ms frame duration, 20 ms frame step). The algorithm was implemented first in Matlab, with necessary measures taken to prevent overflow and the models and probabilities were limited inside reasonable boundaries. For computing the speech absence probability we used the VAD described in [8], which provides speech presence probability per frequency bin. With T = 20 ms we used update time constants $\tau_Z = 0.08s$ and $\tau_Y = 0.08s$. They were adjusted using synthetic contamination tones, consisting of several sinusoidal signals and additive white noise. The frequencies of the sinusoidal signals were chosen to vary from the center frequency of a subband (say 1012.5 Hz) to the middle point between the centers of two subbands (say 1000.0 Hz). An important part of the algorithm verification were tests with white noise only (to verify that we do not suppress it – no predictable component) and with clean speech – to verify that adaptation time constants are large enough to avoid canceling of the pitch and its harmonics from the speech signal. The initialization of the models was $\hat{Z}(k)$ with the first frame, $\hat{Y}(k) = e^{-j2\pi kTf_s/K}$. Brief experiments showed that using of VAD can be avoided assuming $p_k^{(n)} \equiv 1$, but that would require more careful tuning of the adaptation time constants to avoid introducing distortions into the speech signal.

4.1. Evaluation criteria

As main evaluation criterion we choose the improvement in the signal-to-noise ratio (SNR). We believe that this is a better evaluation parameter than the suppression of the contaminating signal, because it reflects the added noise from estimation errors. All evaluation recordings were done with additional reference channel – signal from a close talk microphone. The reference close talk microphone was used only for signal/pause classification of the frames for better SNR estimation. The SNR is estimated as the proportion of the average energy of the signal and noise frames. A secondary criterion was listening evaluation of the quality of the estimated speech signal.

4.2. Results and discussion

Once we achieved good performance with synthetic signals, we proceeded to evaluations with real signals. They were recorded in normal office noise and reverberation conditions: noise floor of ~50 dB SPL and T_{60}=290 ms. The microphone was positioned 1.5 m from a human speaker, wearing a headset. The evaluation set of recordings consisted of white noise, clean speech from the close talk microphone, speech with office noise, speech with office noise and loud buzzers of two types:

1. high frequency ~2600 Hz, three harmonics;
2. low frequency ~300 Hz, twenty harmonics.

They were positioned 2 meters from the microphone. The suppression results are shown in Table 1. There is no suppression for white noise and clean speech, as expected. Listening tests confirmed absence of audible distortions in the clean speech signal. For office noise (three computers with their fans and hard disk drives, air conditioning) the algorithm improves the SNR with almost 3 dB, removing the predictable components from the noise. The proposed algorithm suppresses the signals from the two buzzers up to 15 dB. On Figure 1 we show the waveforms and the spectrograms of the input and output signals for Buzzer 1 case.

Table 1. SNR improvement for various conditions.

Recording	Input			Output			Impro-vement
	Signal	Noise	SNR	Signal	Noise	SNR	
White noise		-13.43			-13.43		0.00
Clean speech	-25.37	-60.44	35.07	-25.38	-60.75	35.37	0.30
Office noise	-34.55	-44.62	10.07	-35.02	-47.98	12.96	2.89
Buzzer 1	-21.42	-21.69	0.27	-23.19	-39.35	16.16	15.89
Buzzer 2	-18.56	-20.52	1.96	-24.21	-39.96	15.75	13.79

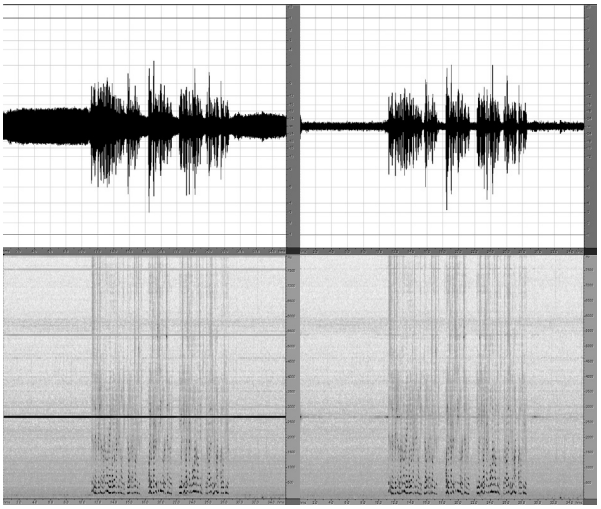

Figure 1. Processing example: speech in office conditions, contaminated with Buzzer 1. Left: input; right: output. Top: waveforms, 3600 ms segment; bottom: spectrograms – frequency range 0-8 kHz, dynamic range 96 dB.

The changes in the magnitude of the buzzer in the input signal are due to people moving in the room and changing the reverberation conditions. These changes cause some residuals in the cleaned signal, because the system needs time to adapt. Figure 2 contains the output signal from recording where the buzzer was turned on at sixth second and shows the adaptation speed of the proposed algorithm. For less than four seconds the algorithm builds the contaminating signal models and converges to 99% of the final suppression. The adaptation speed depends mainly on the time constants in the VAD, the proposed algorithm itself should converge in less than 0.24 sec.

5. CONCLUDING REMARKS

In this paper we proposed computationally efficient real-time algorithm for removing stationary interfering tones from audio signals. Such tones can appear in the captured signal acoustically or electronically from the power lines, transformers, fans, hard disk drives, TV cameras and monitors. The removal is based on cancellation of the stationary tones with an adaptively updated model.

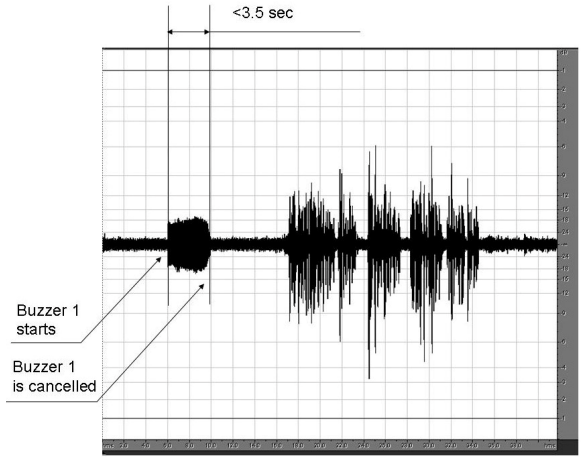

Figure 2. Adaptation speed of the proposed algorithm. Buzzer 1 is turned on at second 6 and completely suppressed at second 9.5.

Our algorithm removes the contaminating tones without introducing artifacts such as musical noise or nonlinear distortions. It can be used as pre-processor to a classic gain-based noise suppressor. Because the algorithm reduces the predictable stationary part of the noise by up to 15 dB, it allows the classic noise suppressor to apply less noise reduction, which leads to less distortion and musical noise.

6. REFERENCES

[1] N. Wiener, Extrapolation, Interpolation, and Smoothing of Stationary Time Series: With Engineering Applications, Principles of Electrical Engineering Series. MIT Press, Cambridge, MA, 1949.

[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. ASSP-33, no. 2, pp. 443–445, Apr. 1985.

[3] W. Roguet, N. Martin, A. Chehikian, "Tracking of Frequency in a Time-Frequency representation," *Proc. of IEEE Int. Symp. on TFTS*, pp. 341-344, 1996.

[4] M. Davy, B. Leprette, C. Doncarli, N. Martin, "Tracking of Spectral Lines in ARCAP Time-Frequency Representation," *Proc. of EUSIPCO*, Rhodes Island, Greece, 1998.

[5] C. Andrieu, M. Davy, A. Doucet, "Efficient Particle Filtering for Jump Markov Systems. Application to Time-Varying Autoregressions," *IEEE Trans. of Signal Processing*, Vol. 51, No. 7, July 2003.

[6] B. Andia, "Restoration of Speech Signals Contaminated by Stationary Tones Using an Image Perspective," *Proc. of IEEE ICASSP*, Toulouse, France, May 2006.

[7] H. S. Malvar, "A Modulated Complex Lapped Transform and Its Applications to Audio Processing," *Proc. of IEEE ICASSP*, Phoenix, March 1999.

[8] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection", *Signal Processing Letters*, vol. 6, pp. 1-3, Jan. 1999.