Automatic noise substitution in natural audio signals

Janto Skowronek, Steven van de Par

Philips Research, Digital Signal Processing Group, 5656 AA Eindhoven, Netherlands, Email: janto.skowronek@philips.com

Introduction

One key concept in audio coding is to remove signal redundancies during encoding by transforming the signal into a more bit-rate efficient representation. A difficult signal type for some audio coders are noisy or noise-like audio signals, for instance unvoiced parts of speech. Due to their high entropy, a waveform description of these noisy signals typically requires a high bit-rate.

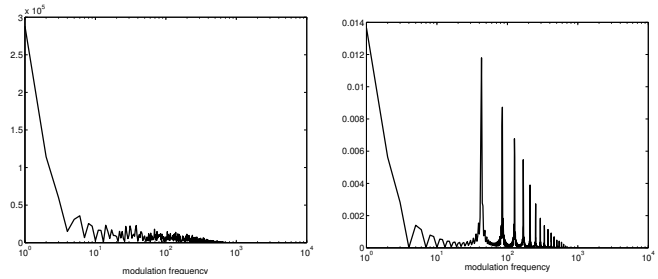
In contrast, the human auditory system is not able to distinguish between two realizations of noise, provided that the temporal and spectral envelopes of both noises are the same. That means, an identification of the spectro-temporal envelopes of those noisy audio signals will allow a substitution with other noise without any perceptual change. Such an inaudible noise substitution will then enable a coding technique that only encodes the parameters of the spectro-temporal envelope and not the noisy signal part itself.

The idea of this paper is to implement an algorithm that automatically identifies and substitutes noisy signal parts in different spectro-temporal regions. We applied a perceptual model in order to get a decision variable, that determines whether such a noise substitution will be audible or not. Based on that decision a substitution algorithm was developed and different nearly stationary audio signals were processed. Finally we used these signals in a subjective test in order to evaluate the quality of the signal modifications.

Model and decision variable

In order to derive a suitable perceptual decision variable, we need a model that allows to distinguish between noisy and deterministic parts of an perceived audio signal. Dau et al. [1] developed a modulation filter-bank model, which has been successful in describing the masking behaviors of different noise types. This model evaluates the modulation spectra of the signal within the auditory system. It is promising for our purpose because the modulation spectra for noisy and tonal signals seem to be clearly different (see Fig. 1).

Since we were focussing on stationary signals, we used a less complex version of Dau's model omitting its adaptation loops and internal noise (cf. [2]) and replacing its modulation filter-bank by an analysis stage of the modulation spectrum. The input signal is sent through 41 ERB-scaled gamma-tone filters, which we used for a basilar-membrane modelling. Each filter output is half-wave rectified, low-pass filtered and down sampled in order to approximate the processing in the inner hair-cells. Dau's model uses a modulation filter-bank for describing the desired modulation properties, but we decided to parameterize the modulation spectrum in a simplified way. Our algorithm computes the FFT-spectrum of the hair-cell output and integrates the spectral amplitudes under a set of Hanning windows having the same bandwidths as Dau's modulation filters. From this model, we obtained an internal representation $I_l(c)$: We get for each basilar-membrane filter l a set I of coefficients (index c), which parameterize the modulation spectrum.



(a) (b)

Figure 1: Modulation spectra of a noisy (a) and a tonal (b) signal at the output of the hair-cell simulation for one basilar-membrane filter.

The decision, whether a noise substitution is allowed or not is based on a comparison between the internal representation of the original signal $R_l\,(c)$ and the substituted signal $I_l\,(c)$. For that we compute a distance measure per basilar-membrane filter l

$$\rho_l = \sum_c \frac{\left(I_l\,(c) - R_l\,(c)\right)^2}{\|R_l\,(c)\,\|^2} \tag{1}$$

and compare its sum $\rho = \sum_l \rho_l$ over all basilar-membrane filters with a fixed criterion ρ_{crit} that we specified during the development.

Algorithm

The algorithm is divided into an analysis and a synthesis part. The analysis is based on an iterative trial and error method:

1. Take a segment of the audio signal in time (time frame n) and frequency (ERB-scaled rectangular filter l).
2. Replace this patch with noise, having the same power as the signal patch.

3. Calculate the internal representations of original and modified signal and compute ρ.
4. Derive decision: $\rho \leq \rho_{crit} \Rightarrow$ *yes* otherwise *no*.
5. Store this decision into a decision matrix $OK\,(n, l)$ and go to next signal patch.

In the synthesis part, only those signal patches (n, l) with a *yes* entry in the decision matrix OK are substituted.

Since a few problems occurred during the development, several details were modified in order to improve the performance. Here we briefly discuss two of them.

Noise substitutions in low frequencies lead to internal representations, which are not clearly distinguishable from the representation of the original signal. The reason is that the frequency resolution of the algorithm ($\approx$ 43 Hz) lies in the range of the bandwidths of the lower ERB-scaled filters. And thus the noise substitution and model computation is based on only one frequency bin. That leads to only one degree of freedom for the noise substitution – the phase – since the power of the noise is fixed to the level of the original signal. That means that rather unreliable decisions are made about the noise substitution. As a consequence we decided to choose the lowest filter as $l_{min} = 18$ (bandwidth: 172 Hz $\approx$ 4 bins).

While testing the algorithm with broadband noise as input signal, not all time-frequency patches of the original noise signal were identified as noise. We found out that this holds for the case when one signal patch was substituted for the analysis but not for the case when more adjacent frames were simultaneously substituted. Our explanation is that the model needs a longer period in order to discriminate between tonal and noisy audio parts. Therefore we modified our algorithm: The decision variable for one time patch n bases on a noise substitution in the current and in its eight neighbouring frames. The classification is only performed for the central frame.

Evaluation

Using informal listening tests we found a decision criterion that allowed noise substitutions without causing too heavy audible distortions. Dependent on the used excerpts, the algorithm substituted between 40% and 80% of the signal's time-frequency tiles with noise.

For a formal evaluation of the algorithm's performance we conducted a listening experiment (one audibility test, one quality test) with eight subjects. Here we discuss the results from the quality test, where we presented twenty different 300 ms long excerpts. These nearly stationary excerpts were the same that we used for the development, including harmonic and noise-like spectro-temporal structures. In addition, we coded seven of them with the Fraunhofer MPEG 1 layer 3 coder at a constant bit-rate of 96 kbit/s (mono). Each excerpt was presented twice: first the original, second the noise-substituted or MPEG-coded one. The subjects had to judge the quality of distortions using a discrete five point scale: Imperceptible, Perceptible but not annoying, Slightly annoying, Annoying, Very annoying. Figure 2 shows the mean results over all subjects for the noise-substituted (solid line) and the MPEG-coded (dotted line) excerpts. The judgments for the noise substituted excerpts as well as for the MPEG-coded excerpts lie between *imperceptible* and *slightly annoying*. Thus the quality of the noise substitution algorithm is comparable to standard codecs working with a bit-rate of 96 kbit/s (mono) and it can be regarded as a good quality.

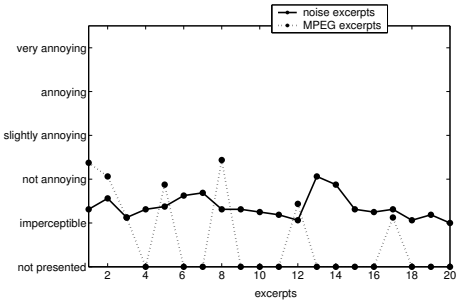


Figure 2: Qualility results averaged over all subjects. Not all excerpts were MPEG-coded and therefore not presented in the tests.

Conclusions and future work

An algorithm was presented in this paper, which substitutes spectro-temporal signal patches with noise with a minimum of audible signal distortions. The evaluation of the algorithm shows quality results comparable to standard audio coders.

The algorithm was developed and evaluated with nearly stationary signals. No conclusions can be drawn how this algorithm works for non-stationary signals. In addition this is an iterative algorithm, therefore the processing of signals needs a considerable amount of time.

In future work, the algorithm has first to be tested on non-stationary audio signals. Basing on these results, modifications of the algorithm (e.g. the used model) may be necessary. Finally the computation time has to be optimized by exchanging the iterative comparison based method with a direct classification method.

References

[1] Dau, T., Kollmeier, B., Kohlrausch, A. *Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers.* J.Acoust. Soc. Am., Vol.102, 2892-2905, 1997

[2] Dau, T., Püschel, D., Kohlrausch, A. *A quantitative model of the "effective" signal processing in the auditory system. I. Model structure.* J.Acoust. Soc. Am., Vol.99, 3615-3622, 1996