How Sparse Can We Make The Auditory Representation Of Speech?

Christian Feldbauer, Gernot Kubin

Signal Processing and Speech Communication Laboratory
Graz University of Technology, Austria
feldbauer@inw.tugraz.at, kubin@inw.tugraz.at

Abstract

In this work we deal with a speech coder which is based on a model of the human peripheral auditory system and uses a neural auditory representation as its code. This representation consists of multi-channel sparse pulse trains but is still highly overcomplete and therefore not efficient in terms of its data compression capability.

The emphasis of this paper is on answering the question ‘How sparse can we make the auditory representation?’, i.e., on finding a bound for the number of pulses which can be omitted without degrading the quality of the reconstructed speech signal.

For this purpose we incorporate a second auditory model which allows to decide whether a pulse is needed or not based on the excitation pattern caused by the signal when a single pulse is resynthesized. This model accounts for both simultaneous and temporal masking. We also propose a method for compensating for the loss of energy due to the elimination of pulses which makes possible spectral distortion inaudible. Results show that about 74% of the pulses can be omitted without degrading the quality of the reconstructed speech signal.

1. Introduction

In [1] a speech coder was proposed which resembles the pathway of signal transduction in a human cochlea. The model is simple and does not account for the nonlinear, active behavior of human hearing such as the dependency on stimulus level. The basilar membrane (BM) is modeled by a gammatone filterbank, the inner hair cell (IHC) by static nonlinearities, and an ensemble of neurons connected to one IHC by a peak-picking procedure (i.e., non-uniform sampling at the maxima locations).

The output of this model which is hereafter referred to as auditory representation is a set of pulse trains and consists largely of zeros.

Our decoder reconstructs the speech signal from the auditory representation in an extremely simple way because it does not need an iterative reconstruction algorithm, nevertheless it achieves a high reconstruction quality.

While the auditory representation is sparse, it still contains more pulses than the original signal samples (typically three times more). The high number of pulses is not caused by the model’s inaccuracy, the redundancy is rather natural for neural responses (about 30,000 afferent fibers are attached to about 3,000 IHCs [2] in a human cochlea). It is even present when all physiological effects responsible for masking have been considered carefully because then two stimuli, masker and probe, produce the same perceived neural response as the masker alone.

However, one advantage of coding an accurate auditory representation is that a quantizer or bit allocation scheme based on the straightforward sum of squared errors is able to predict the degree of perceived degradation because the transformation of the signal into the perceptual domain also transforms a complex distortion criterion into a simple one. While the pulse amplitudes preserved by the peak picker can be quantized very coarsely [1, 3], the pulse positions are crucial. Thus coding such a representation needs an exceptionally high amount of side information. One step further towards efficient coding is to reduce the amount of side information by reducing the number of pulses. This can be achieved by exploiting the quantized nature of perception, e.g. consideration of masking effects or just noticeable differences.

In [3] masking has already been included to eliminate pulses. Separate models, one for simultaneous and one for temporal masking are used. Their model for simultaneous masking is taken from MPEG-1 and is based on a frame-by-frame prediction of a masking threshold. For temporal forward masking (or postmasking), a model is incorporated which assumes an exponentially decaying masking threshold.

In our work we propose a pulse-based excitation pattern model for both simultaneous and temporal masking. Furthermore, we show that a pulse is also able to mask preceding pulses considerably (i.e., backward masking or premasking) since we carefully account for the synthesis procedure of the decoder.

In section 2 we briefly review the speech coder proposed in [1] and describe the reconstruction process. Section 3 develops a universal framework which enables us to examine the importance of parts of the auditory representation. For that purpose we carefully incorporate the final listener and consider the problem from a transmultiplexer point of view. In section 4 we propose the new masking model and section 5 deals with an amplitude correction method to compensate for spectral distortions.

2. A physiologically motivated subband coder

The first stage of our coder simulates the vibration pattern along the BM and is implemented by a non-decimated, 20-channel, FIR gammatone filterbank with center frequencies equally spaced on an ERB-rate scale [4] between 100Hz and 3600Hz for speech signals sampled at 8kHz. The ERB-rate scale is closely related to the frequency-position mapping which means that the filterbank output represents the vibrations of 20 approximately equally spaced points along the membrane. At these locations, IHCs are situated which are modeled by static nonlinearities, namely a half-wave rectifier and a power-law amplitude compressor.

It is well known that neurons most probably fire when the stimulus has a local maximum (‘phase-locking effect’) which can be observed for frequencies below 4kHz. A single neuron generally does not fire more often than 250 times per second
and is by itself not able to preserve the time structure of higher-frequency components. But since in a human cochlea about 10 afferent neurons are attached to a single IHC, the ensemble is able to maintain it and so, we model such an ensemble rather than single neurons. This is easily done by a peak-picking procedure which adaptively subsamples the IHC output at local maxima and sets all other samples to zero.

Thus, the output of the coder consists of 20 non-uniformly subsampled subbands. It contains on the average about three times more pulses than samples in the original signal. This oversampled representation needs further processing to get a compact code (see the following sections).

For the reconstruction procedure, after undoing the power-law compression by power-law expansion, we compensate for the loss of energy caused by the peak-picking procedure by introducing a gain factor in each channel for magnitude correction (see [1, 5] for more details). Then, actual signal reconstruction is achieved by bandpass filtering the pulse trains and summing them in the synthesis filterbank. This filterbank uses time-reversed versions of the impulse responses of the gammatone analysis filterbank, allowing summation without phase distortion. The remaining magnitude ripple of about 4dB can be reduced by a linear-phase equalizer filter at the expense of additional delay.

3. Exploiting the quantized nature of perception

To reduce the amount of data we want to omit pulses. For this purpose, we should ask how much a single pulse of the coder’s output contributes to the final, actually perceived neural code. Then, we should be able to decide whether the pulse is important and, therefore, needed or not.

In Fig. 1 the entire signal flow graph is illustrated which consists of the coder, the decoder, and the final receiver—in our case a human listener. The aim is to provide the listener the same perception as the original signal would do. Therefore, we introduce a second auditory model to represent the human listener in the flow graph of Fig. 1 which can perceptually be more accurate than the one used in the coder (if a more accurate model would be used for the coder, signal reconstruction could be difficult or even impossible). In this way we introduce a domain where perceptual differences can be detected more accurately (even if the code is not an auditory representation). This is similar to MPEG audio coders, where the auditory model operates in a parallel signal path. In contrast to MPEG, our auditory model, which represents the final listener, uses the reconstructed signal as its input. Thus, we are able to directly compare the outputs of the listener model for the signal synthesized from a reduced auditory code and for the signal synthesized from the complete auditory code.

If we consider only the decoder of Fig. 1 with the pulse code as input and the human listener with the actually perceived code as output, we get a multi-input, multi-output (MIMO) system generally known as transmultiplexer. This is illustrated in Fig. 2. This approach simplifies the optimization of the pulse code generated by the coder in terms of minimizing the number of pulses needed at the input of the decoder (or synthesizer) to produce a proper perception.

On the right hand side of Fig. 2 the neural response (predicted by the same auditory model as used in the coder) is plotted when only one pulse from channel five is synthesized. It obviously consists of many pulses. Simply speaking, all additional pulses are for free and need not to be encoded if they are also present at the input. Based on this observation, a simple method to eliminate pulses is proposed in the following section.

The present problem is similar to finding the optimum multi-pulse excitation signal in linear prediction coding (MPELP [6]) but extended to a multi-channel signal and with implicit perceptual weighting.

4. Masking determined by an isolated-pulse BM excitation model

Since we just like to eliminate pulses and we do not consider the quantization of pulse locations here (i.e., the temporal fine structure of the auditory representation will be preserved) an auditory excitation pattern model is sufficient for the second auditory model. The output of an excitation pattern model is usually the set of envelopes (or powers [4, 7]) of the bandpass signals from an auditory filterbank. For the experiments, we use again the linear gammatone filterbank with 20 channels used in the coder for modeling the BM but, additionally, we consider a weighting of the channels according to an inverse 100-phon contour [7] which models the transfer function of the outer and the middle ear (OME).

For the sake of simplicity and to save computational load, our masking model treats all pulses independently, i.e., it does not accumulate an overall excitation pattern produced by more than one pulse. Instead, we store 20 unit excitation patterns for the case of the 20-channel subband coder. In Fig. 3 two dif-
different unit excitation patterns are plotted, on the left the BM excitation of a unit impulse synthesized from channel five and on the right from channel 15. These patterns are computed once by convolving the impulse response of the corresponding channel’s synthesis filter with all OME-weighted impulse responses of the second model’s analysis filterbank and generating the Hilbert envelopes. We refer to the pattern produced by a pulse from channel ch as $E_{ch}[n,k]$ where $n$ is time and $k$ is the channel index. Additionally we call the maximum excitation $\tilde{E}_{ch} = \max_{n,k} E_{ch}[n,k]$ which, in our case where the impulse responses of synthesis and analysis filters are time-reversed versions of each other, occurs in channel $k = ch$ delayed by the length of the impulse response minus one. Since all used impulse responses have the same length $L$ (667 samples), the peak can always be found at $\tilde{E}_{ch} = E_{ch}[L-1, ch]$. The total length of a pattern is $2L - 1$.

Using these patterns, which span a spatio-temporal BM area, we are able to account for both simultaneous masking and temporal masking at the same time. Furthermore, since these patterns attack smoothly in a similar way as they decay, forward as well as backward masking is considered.

We consider every pulse individually as a masker and all pulses in its neighborhood, again pulse by pulse, as probes.

Let us denote the multi-channel pulse trains by $x[n,k]$ and we require that these signals are ready to be synthesized in the synthesis filterbank, i.e., steps such as power-law expansion have already been performed. Now, consider the pulse at $n = n_M$ and $k = k_M$ as the masker. All other pulses between $n = n_M - L + 1$ and $n = n_M + L - 1$ in all channels are probes. Let us choose the pulse at $n = n_P$ (which is in this range) and $k = k_P$ for the current probe. We define the criterion for masking as

$$x[n_P,k_P] \cdot \tilde{E}_{k_P} < r \cdot x[n_M,k_M] \cdot E_{k_M}[n_P-n_M+L-1,k_P]$$  \hspace{1cm} (1)

where $r$ is a constant between 0 and 1 which controls the impact of the criterion. If the criterion is fulfilled we can delete the current probe and continue to test the next one. If $r = 0$ then the criterion is never fulfilled and no pulses are omitted. With $r = 1$ a maximum number of pulses will be deleted.

The sequence for selecting a pulse to become the next masker plays an important role. Theoretically, an optimum pulse configuration should exist which produces a proper neural response with a minimum number of pulses to be synthesized, but it would be very difficult to find this configuration. A computationally simple possibility which finds a solution close to the optimum is to sort all pulses according to their amplitudes and start with the highest one. In a real-time coder this can be accomplished on a block-by-block basis introducing a small delay.

The influence of the impact factor $r$ on the reconstruction quality was determined experimentally in a listening test (two experienced listeners, A/B comparison) and it turned out that if $r \leq 0.88$ the quality remains perfect without any noticeable distortions. This value corresponds to a difference between masker and probe of about 1dB. With $r = 0.88$ a reduction in the number of pulses by 63% on average is possible.

If $r$ is increased further, an audible spectral distortion arises but no other kind of artifacts. In the next section we propose a simple method to correct the amplitudes of the remaining pulses to compensate for this spectral distortion.

5. Elimination of spectral distortion

The following method of pulse amplitude correction is also based on a BM excitation model, i.e., we calculate envelopes within all auditory channels and compare the excitation generated by the original pulse trains with the one produced by the reduced pulse trains. For the envelope generation, we lowpass-filter the pulse trains using linear-phase FIR filters whose impulse responses are already contained in the unit excitation patterns $E_{ch}[n,k]$. For the auditory channel $k$, we use the impulse response $l_k[n] = E_k[n,k]$. Again, we call the original pulse trains $x[n,k]$ and refer to the reduced pulse trains as $\tilde{x}[n,k]$. Then we can generate the envelopes by

$$e_k[n] = x[n,k] \ast l_k[n]$$ \hspace{1cm} (2)

and

$$\tilde{e}_k[n] = \tilde{x}[n,k] \ast l_k[n].$$ \hspace{1cm} (3)

Now, the amplitude of the pulse $\tilde{x}[n_0,k_0]$ is corrected by multiplication with the ratio of the original to the reduced envelope at the proper location

$$\tilde{y}[n_0,k_0] = \tilde{x}[n_0,k_0] \frac{e_{k_0}[n_0 - L + 1]}{\tilde{e}_{k_0}[n_0 - L + 1]}$$ \hspace{1cm} (4)

since the group delay of all lowpass filters is $L - 1$. If the ratio is computed only for actual pulse locations, divisions by zero are impossible to occur. Using this amplitude correction method we are able to increase the impact factor $r$ even up to 1 while maintaining perfect resynthesis quality.

6. Simulation results

For the 8kHz speech signal ‘The juice of lemons makes fine punch’ spoken by a female speaker, the total number and the number of pulses found in the individual auditory channels are visualized in Fig. 4 before applying the masking criterion of section 4 with $r = 1$ and afterwards. We observe a reduction in the overall number of pulses from 76,752 to 19,723 what corresponds to an elimination of about 74% of the original pulses. If we consider the time unit of a sample we now need only 0.78 pulses per sample instead of 3.05. It also becomes apparent that in higher-frequency channels much more pulses can be omitted than at lower frequencies.

Fig. 5 shows a segment of the speech example mentioned above and compares the original and the reconstructed signal. In addition to the masking criterion with $r = 1$ the amplitude correction method of section 5 has been applied. The averaged segmental signal-to-noise ratio is only 12.2dB but the error is not audible.

The upper plot of Fig. 6 shows the auditory representation of the first 100 samples of the signal segment shown in Fig. 5 generated by the perceptual subband coder. In the lower plot the pulse representation is shown after the masking criterion with $r = 1$, which removes 74% of the pulses, has been applied.

To be able to fairly compare our method with [3] we ran a simulation with a wideband speech signal (i.e., 16kHz sampling-rate) and with a comparable setup of the coder: 21 channels ranging from 50Hz to 7,000Hz. With $r = 1$ and enabled amplitude correction we again eliminated 74% of the pulses what corresponds to merely 0.66 pulses per sample for the reduced representation and we achieve excellent reconstruction quality. In [3] 1.26 pulses per sample are necessary.

As already became apparent in Fig. 4, exploiting the masking model is much more efficient at higher frequencies. Thus,
we expect to get even better results when signals at higher sampling rate are processed, e.g. audio at 44.1kHz.

7. Discussion and conclusions

Testing an auditory model by resynthesis was proposed in [8]. But since the inverse of an auditory model is not unique, too much degrees of freedom exist for that the accuracy of the model could be evaluated on the basis of the reconstruction quality. Whereas the framework we propose in this paper, which regards the problem as a transmultiplexer, allows the evaluation of auditory models on that basis at least as far as prediction of masking is concerned.

We have proposed a new masking model which is based on the BM excitation when a single pulse of the representation is resynthesized. It accounts for simultaneous and temporal masking at the same time. We have also proposed a simple method for correcting the pulse amplitudes which compensates for spectral distortions when too many pulses are omitted.

Currently, we are able to omit 74% of the pulses in the auditory representation while maintaining transparency of the coded signal. Compared to the results in [3], we need only around a half of their number of pulses per time unit. This result demonstrates the enormous potential of the use of a multi-channel pulse representation, i.e., the auditory representation of speech and audio signals for high-quality coding applications.

8. References


