BIT-RATE REDUCTION STRATEGIES FOR NOISE SUPPRESSION WITH A REMOTE WIRELESS MICROPHONE

Nemanja Cvijanović

RWTH Aachen University
Aachen, Germany
nemanja.cvijanovic@rwth-aachen.de

Ousman Sadiq

University of Strathclyde
Glasgow, United Kingdom
ousman.sadiq@strath.ac.uk

Sriram Srinivasan

Philips Research
Eindhoven, The Netherlands
sriram.srinivasan@philips.com

ABSTRACT

In single-channel non-stationary noise reduction it is paramount that a good noise reference is available in a timely manner to maintain a high quality speech signal. Using a remote wireless microphone placed close to a noise source, a good estimate of the noise power spectral density (PSD) can be acquired. This estimate, however, needs to be transmitted to the primary microphone for noise reduction. As wireless transmission is power intensive, it is desirable to reduce the bit-rate while maintaining good performance. In this paper, we propose techniques such as quantizing, frequency bin clubbing and intermittent PSD transmission to reduce the transmission bit-rate, and investigate their impact on performance.

***Index Terms*—** Speech enhancement, noise reduction, wireless microphone, ad-hoc sensor network.

1. INTRODUCTION

Hands-free communication is an interesting application where speech enhancement is required to reduce the effects of a noisy environment. At home, users engage in conversations with family and friends using portable network devices that may not be held close to the user [1], which capture both speech and ambient noise. Technologies such as VoIP have reduced the cost of these services and the growing computational power of the devices encourages the development of techniques to support longer and more comfortable conversations. A method to increase the level of comfort in speech conversation is to better deal with the effects of highly non-stationary noise, for example, from music devices and kitchen appliances, and this can be quite challenging [2].

In [3], a directional remote wireless microphone (RWM) was placed near a noise source to better approximate the power spectral density (PSD) of the observed noise signal. The PSD captured by the RWM is transmitted to the primary microphone and is then used to construct a Wiener filter for noise reduction. Such a scheme has several benefits. In more conventional approaches such as adaptive noise canceling, accurate timing synchronization between the signal observed by the RWM and the primary microphone signal is necessary. By using PSDs we relax the requirement for timing synchronization. Another advantage of PSD transmission is the robustness against small sample rate deviation, since this only slightly expands or compresses the spectrum. Since the PSD is symmetric for real signals, only the positive frequencies are needed and this allows flexibility in reducing the number of bits compared to an adaptive noise canceler.

In a practical system, the PSD of the RWM signal needs to be transmitted wirelessly to the primary microphone. As wireless transmission is power intensive, it is desirable to reduce the bit-rate required. We propose strategies to reduce the transmission bit-rate, exploiting the fact that the transmitted PSD is only used to generate a Wiener filter at the primary microphone. These strategies include clubbing adjacent frequency bins of the PSD into bands, quantizing these bands using a specified number of bits, and intermittent transmission of these bands.

The remainder of this paper is organized as follows. In Section 2 the signal model is explained in the case of perfect network conditions, i.e., high bandwidth, perfect synchronization and no quantization. Section 3 discusses proposed bit-rate reduction strategies. Real audio recordings will be used to understand the impact of these strategies on performance and the results are presented in Section 4. Section 5 summarizes the conclusions.

2. SIGNAL MODEL

In the following, an additive noise model is used, where the primary microphone signal can be written as

$$p(k) = x(k) + n(k), \tag{1}$$

where k is the time index, $x(k)$ and $n(k)$ are the sampled clean speech signal and the noise signal, respectively, observed at the primary microphone. Expressed with PSDs in the frequency domain and assuming the speech and noise are uncorrelated, this model becomes:

$$P_p(\omega) = P_x(\omega) + P_n(\omega), \tag{2}$$

where P_x, P_n and P_p are the clean speech, noise and primary microphone signal PSDs, respectively, and ω is the frequency index.

The signal observed by the remote microphone is $r(k)$ and its PSD is $P_r(\omega)$. If the reverberation level is low and the directional remote microphone is much closer to the noise source than the primary microphone, then $P_r(\omega)$ is a good estimate of $P_n(\omega)$, except for a frequency-independent level difference between the two PSDs due to the distance between the microphones, i.e.,

$$P_n(\omega) = g_r P_r(\omega) \tag{3}$$

We consider the use of a Wiener filter for noise reduction. We note that $P_r(\omega)$ can be used as an estimate of the noise PSD in any single-channel speech enhancement method [2], and the specific choice is not relevant to the discussion here.

The Wiener filter is constructed using the PSDs of the remote and primary microphone signals:

$$H_t(\omega) = \frac{P_p(\omega) - \hat{g}_r P_r(\omega)}{P_p(\omega)}, \tag{4}$$

where $\hat{g}_r$ is the estimated time-varying, frequency-independent level factor. This gain factor is calculated a priori using the noise only signals. In practice the method specified in [3] can be used.

The performance of the Wiener filter depends strongly on the accuracy of the transmitted PSD of the remote microphone signal $P_r(f)$. In a realistic implementation it is useful to introduce a smoothed Wiener filter estimate.

$$H_t(\omega) = \alpha H_{t-1}(\omega) + \beta \cdot \max\left(\frac{P_p(\omega) - \hat{g_r} P_r(\omega)}{P_p(\omega)}, \varepsilon\right), \qquad (5)$$

where $\alpha + \beta = 1$ and ε is a small number that acts as a lower limit in the amount of attenuation. If $P_r(\omega)$ is transmitted frequently enough (every 20-32 ms), highly non-stationary noise types can be suppressed effectively.

3. BIT RATE REDUCTION

The main objective of this work is to reduce the bit-rate when transmitting the PSDs of the remote microphone signal, $P_r(\omega)$, to the primary microphone, and three bit-rate reduction methods are presented, bin clubbing, quantization and intermittent transmission of the noise PSD observed by the RWM. The effects of these strategies on performance are expected to be minimal since the remote PSD is only used to construct the Wiener filter and not to reconstruct the desired speech signal.

3.1. Bin clubbing

In the frequency domain, groups of adjacent bins can be represented by a single value, thus decreasing the number of PSD values required to be sent. We refer to this as bin clubbing.

Bin clubbing exploits the psychoacoustic nature of the human ear to group adjacent bins in the frequency domain into a certain number of critical bands. We consider two well known scales - the Bark scale [4] and the Equivalent Rectangular Bandwidth (ERB) scale [5].

The Bark scale defines 24 critical bands up to 16 kHz based on human perception and is defined by the following expression:

$$z(\omega) = 13 \cdot \text{atan}\left(\frac{0.76\omega}{1000}\right) + 3.5 \cdot \text{atan}\left(\left(\frac{\omega}{7500}\right)^2\right), \qquad (6)$$

where the output, $z(\omega)$, is the Bark level. Solving (6) for $z \in \mathbb{N}$ gives the frequency borders b_i^{B} for the 24 bands and the values for each band are calculated as

$$P_r^{\text{B}}(i) = \overline{P_r(\Omega^{\text{B}})}, \qquad (7)$$

where $\Omega^{\text{B}} = \omega \in [b_i^{\text{B}}, b_{i+1}^{\text{B}})$, and the overbar is used to represent the mean operation. In this work audio signals were sampled at 16 kHz and therefore only 22 Bark bands are used.

The ERB scale is a modification of the classical Bark scale and the defined critical bands are narrower. For audio signals that are sampled at 16 kHz, 34 bands are used. The value of each band, P_r^{E} is given by the following expression:

$$P_r^{\text{E}}(i) = \overline{P_r(\Omega^{\text{E}})} \qquad (8)$$

where $\Omega^{\text{E}} = \omega \in [b_i^{\text{E}}, b_{i+1}^{\text{E}})$ are the frequencies in each ERB band and $b_i^{\text{E}} = (10^{(i/21.4)} - 1)/0.00437$ defines the i^{th} border of the band.

Due to a decreased frequency resolution after bin clubbing, the noise reduction performance will suffer. The ERB scale results in a more accurate PSD description than the Bark scale due to the higher number of bands. The difference between both methods is illustrated by applying the clubbing to a speech PSD signal as shown in Figure 1. It can be seen that the ERB follows the speech PSD levels more closely. We consider both scales as they present different trade-offs between bit-rate and performance. Results in terms of SSNR and PESQ will be presented for both in Section 4.

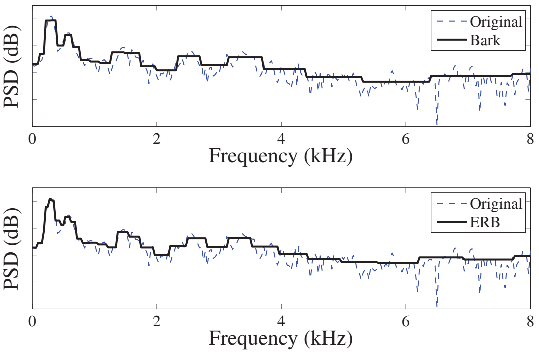

Fig. 1. Comparison of the bin clubbing of a speech segment using Bark (top) and ERB (bottom) scales.

3.2. Quantization

The overall bit-rate may be further reduced by decreasing the number of bits used to describe the bands of the noise PSD in the Bark or ERB scales. Since the transmitted PSD is only used for constructing the gain function in (5) and not reconstructing the enhanced signal, such a scheme can provide a good trade-off between bit-rate reduction and performance.

In this work all bands are assumed to be quantized with the same number of bits and logarithmic companding in the frequency domain is applied to reduce the dynamic range of the PSD values. Non-uniform bit allocation across bands can result in greater savings, and is the subject of future research. Scalar quantizer codebooks were trained with 3 different music types for each band. This resulted in different sized code libraries for the ERB and Bark methods. Each codebook contains 2^N codevalues, where N is the number of bits used for quantizing each band. The LBG splitting algorithm [6] was used for training.

Clearly, there is a trade-off between the performance of the noise reduction algorithm and amount of bit-rate reduction, which will be investigated in Section 4. Future work will investigate a DPCM based strategy for quantization.

3.3. Intermittent PSD transmission

A method to further reduce the required bit-rate is to transmit the PSDs only intermittently to the primary microphone, e.g., instead of sending the noise PSD every 16 ms it can be sent every $n \cdot 16$ ms, with a break period of $(n-1) \cdot 16$ ms, where $n \in \mathbb{N}$. At the primary microphone the latest PSD from the RWM is repeated during those break periods.

This approach is motivated by two observations. First, for small values of n, it may be assumed that the noise PSD varies slowly during the time window of interest, even for non-stationary noise signals. Secondly, there is an inherent smoothing present in the computation of the Wiener filter in (5), to reduce perceptual artifacts.

Naturally, the bit-rate reduction comes at the cost of decreased noise reduction, and the exact trade-off depends on the noisy type, and is investigated in Section 4, where music is considered as the interfering signal.

4. EXPERIMENTAL RESULTS

In this section experimental results based on recordings acquired using the primary and remote microphones are presented. The performance under certain bit-rates is examined for the different methods described in Section 3 to select a suitable technique that has minimal performance impacts while reducing the bit-rate requirements and to understand the trade-offs asociated with each method.

4.1. Set-up

Recordings used during the experiments were made in a moderately reverberant room ($T_{60} = 400$ ms). The distance between the primary microphone and the source of the desired speech signal was 75 cm, which is typical for a VoIP setup with a speaker and a computer. A directional RWM was used and it was located 10 cm away from the noise source. The sources of the desired speech and the noise were 350 cm apart. To enable objective performance evaluation and mixing at any desired SNR, the speech and noise signals were recorded separately at both microphones.

In this work 32 ms Hann-windowed signal segments were used, and the PSD estimate was calculated using a 50% overlap. This results in 512 samples per frame at a 16 kHz sampling rate. For the construction of the Wiener filter in (5), the noise floor was empirically set to $\varepsilon = 0.2$ while the smoothing parameters were set to be $\alpha = 0.3$ and $\beta = 0.7$.

To simulate a realistic scenario, music was chosen as the interfering signal for the experiments. Its highly non-stationary nature also makes it challenging for single-channel speech enhancement algorithms. Three music clips were used, rock, jazz and classical music. These were mixed with the clean speech signals, of duration 50 seconds, at input SNRs of 5 dB and 10 dB at the primary microphone.

4.2. Objective speech quality measures

The quality measures used for evaluation of the noise reduction performance were the segmental SNR (SSNR) and perceptual evaluation of speech quality (PESQ). The SNR was computed as

$$\mathrm{SNR} = 10\log_{10}\left(\frac{\sum_{k=1}^{K} x^2(k)}{\sum_{k=1}^{K}(x(k)-\hat{x}(k))^2}\right), \tag{9}$$

where $\hat{x}(k)$ represents the modified (noisy or enhanced) speech signal and K is the length of the utterance in samples. The SSNR was then computed as the average of the SNR for each utterance frame except for silent frames. The excluded silent frames were frames whose energy was 40 dB below the long-term average energy of the utterance [7]. For PESQ the evaluation was carried out according to [8], using the MATLAB code provided in [2]. Although PESQ was not originally developed for evaluating the performance of speech enhancement algorithms, it has been shown to have a good correlation to subjective quality, and including these results also serves to validate PESQ as a measure.

4.3. Results

The effects on performance of using the Bark and the ERB scales for bin clubbing are presented in Table 2 and Table 1, in terms of the improvement in SSNR and PESQ for input SNRs of 5 dB and 10 dB. These show the average improvement compared to the noisy input and the reference gains when no clubbing is applied to the remote PSD. The results are averaged over three different music types (rock, classic and jazz). The bit proportion in brackets can be used to compute the possible savings in bit-rate when using one of the bin clubbing methods. It is apparent from the tables that while there is a significant savings in bit-rate, there is only a marginal performance loss compared to transmitting the individual frequency bins. Due to the higher number of frequency bands and a more detailed representation of the remote PSD, the ERB method results in better performance compared to the Bark scale, but at a higher bit-rate. Note that due to the different gain calculation method and the different noise types used, the unclubbed results differ from [3].

Method (bit proportion)	10 dB	5 dB
No Clubbing (1)	4.86	5.86
ERB (0.13)	4.54	5.58
Bark (0.09)	4.44	5.53

Table 1. SSNR increase (in dB) compared to the noisy input for bin clubbing at two different input SNRs at the primary microphone. Results are averaged over three different types of music used as interfering signals. The bit-rate reduction relative to the reference (no clubbing) is provided in parentheses.

Method (bit proportion)	10 dB	5 dB
No Clubbing (1)	0.34	0.33
ERB (0.13)	0.26	0.23
Bark (0.09)	0.24	0.21

Table 2. PESQ increase compared to the noisy input for bin clubbing at two different input SNRs at the primary microphone. Three different types of music were used as interfering signals and the results are averaged over them.

The effects of quantization on the noise reduction performance are presented in Figure 2 and Figure 3 in terms of SSNR and PESQ improvement, respectively, for increasing bit-rates. Using N bits to quantize each band results in a bit-rate of $N \times N_b \times T$ bits per second, where N_b is the number of bands and T is the frame rate in seconds. The figures also show the SSNR and PESQ improvements for noise reduction with the unquantized ERB and Bark bands, which serve as an upper bound on performance.

First, it can be seen that performance converges towards the upper bound as the bit-rate increases, and that the curves flatten out for relatively low bit-rates. It is interesting to note that at lower bit-rates using the Bark bands results in better performance than using the ERB bands. As the Bark scale has fewer bands, for a given bit-rate, more bits are available to describe the Bark bands than the ERB bands. At the lower rates, an accurate description of the bands appears to have a larger impact on performance than a finer frequency resolution. A choice between the two scales may be made depending on the available bit-rate.

In the simulation to evaluate the intermittent PSD transmission scheme as described in Section 3.3, to exclude the effect of the particular frames being dropped in an utterance, the the dropped frame was varied in each run of the simulation. Results averaged over the

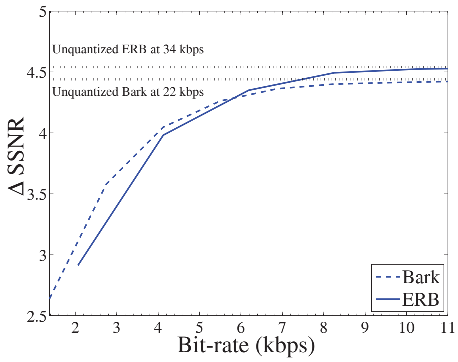

Fig. 2. SSNR increase (in dB) after noise reduction compared to the noisy input for quantized ERB and Bark scale bin clubbing for an increasing number of quantization bits at 10 dB input SNR at the primary microphone.

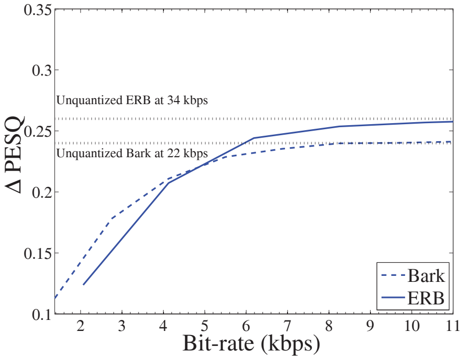

Fig. 3. PESQ increase after noise reduction compared to the noisy input for quantized ERB and Bark scale bin clubbing for an increasing number of quantization bits at 10 dB input SNR at the primary microphone.

different music types, for ERB bands at 6.2 kbps (corresponding to 3 bits per band at the chosen frame rate) are shown in Figure 4. The bands for the missing frames were replaced by the most recently received ones at the primary microphone.

A negligible decrease in performance of our noise reduction scheme was experienced in the case where the bands are sent every other frame. A bit-rate reduction of 50%, which corresponds to 3.1 kbps, thus results for only a marginal decrease in performance. This is due to the 50% overlap in the processing. Longer transmission breaks result in a steeper decline in performance, due to the non-stationarity of the interferer, music in our case. It is interesting to note, however, that there is a positive improvement in SSNR and PESQ even for relatively long transmission breaks. These results also indicate how the system copes with packet loss. Assuming that a lost packet is equivalent to a transmission break, random single packet errors have no noticeable effect on our scheme, while burst errors have a larger impact.

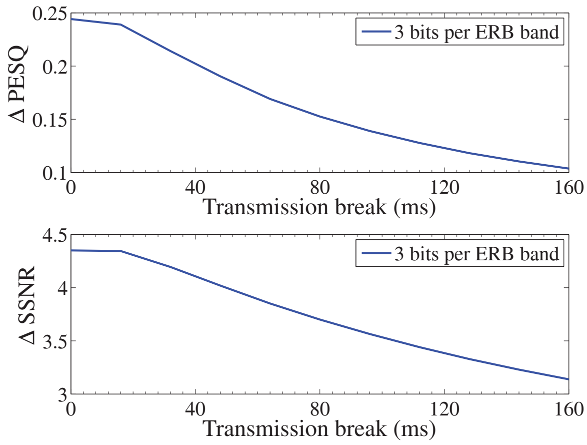

Fig. 4. Effects of intermittent PSD transmission on noise reduction performance, as the size of the transmission break is increased. Results are for the ERB scale, with each band quantized using 3 bits. Results are averaged over 3 different music types at 10 dB input SNR at the primary microphone.

5. CONCLUSION

Transmitting the power spectral density (PSD) estimate of the signal observed by a directional remote wireless microphone (RWM) located close to a noise source improves the noise reduction performance of a Wiener filter. In a practical scenario, it is desirable to reduce the bandwidth requirement of this PSD transmission. Bit-rate reduction strategies were proposed and evaluated in this paper. These strategies include clubbing adjacent frequency bins of the PSD into bands, quantizing these bands using a specified number of bits, and intermittent transmission of these bands. By combining these strategies, significant bit savings can be achieved with limited impact on performance. For example, a bit-rate of 3.1 kbps gives a PESQ increase of 0.24 and an SSNR increase of 4.35 dB, for 10 dB input SNR at the primary microphone for highly non-stationary interferences such as music.

6. ACKNOWLEDGMENT

The authors would like to thank Patrick Kechichian for his valuable comments.

7. REFERENCES

[1] A. Härmä, "Ambient human-to-human communication," in *Handbook of Ambient Intelligence and Smart Environments*, pp. 795–823. 2010.

[2] P. C. Loizou, *Speech Enhancement: Theory and Practice (Signal Processing and Communications)*, CRC, 1 edition, June 2007.

[3] S. Srinivasan, "Using a remote wireless microphone for speech enhancement in non-stationary noise," in *Proc. IEEE Int Acoustics, Speech and Signal Processing (ICASSP) Conf*, 2011, pp. 5088–5091.

[4] H. Fastl and E. Zwicker, *Psychoacoustics: facts and models*, Springer, 3. edition, 2007.

[5] J. O. Smith III and J. S. Abel, "Bark and ERB bilinear transforms," vol. 7, no. 6, pp. 697–708, 1999.

[6] Y. Linde, A. Buzo, and R. Gray, "An algorithm for vector quantizer design," vol. 28, no. 1, pp. 84–95, 1980.

[7] N. S. Jayant and P. Noll, *Digital coding of waveforms*, Prentice-Hall, New Jersey, 1984.

[8] "ITU-T recommendation P.862. Perceptual evaluation of speech quality (PESQ): an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," Feb. 2001.