INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Blind Recovery of Perceptual Models in Distributed Speech and Audio Coding

Tom Bäckström, Florin Ghido and Johannes Fischer

International Audio Laboratories Erlangen, Friedrich-Alexander University (FAU), Germany
tom.backstrom@audiolabs-erlangen.de

Abstract

A central part of speech and audio codecs are their perceptual models, which describe the relative perceptual importance of errors in different elements of the signal representation. In practice, the perceptual models consists of signal-dependent weighting factors which are used in quantization of each element. For optimal performance, we would like to use the same perceptual model at the decoder. While the perceptual model is signal-dependent, however, it is not known in advance at the decoder, whereby audio codecs generally transmit this model explicitly, at the cost of increased bit-consumption. In this work we present an alternative method which recovers the perceptual model at the decoder from the transmitted signal without any side-information. The approach will be especially useful in distributed sensor-networks and the Internet of things, where the added cost on bit-consumption from transmitting a perceptual model increases with the number of sensors.

Index Terms: speech analysis, auditory perception, envelope modelling, internet of things, distributed sensor networks

1. Introduction

The era of Internet of Things (IoT) is approaching, whereby the next generation of speech and audio coders should embrace it. The design goals of IoT-systems however fit poorly with the classic design of speech and audio coders, whereby a larger redesign of the coders is required. Primarily, whereas state-of-the-art speech and audio coder such as AMR-WB, EVS, USAC and AAC consist of intelligent and complex encoders and relatively simple decoders [1, 2, 3, 4], since IoT should support distributed low-complexity sensor-nodes, we would like the encoders to be simple. Secondly, since sensor-nodes are in encoding the same source signal, application of the same quantization at each sensor-node would represent over-coding and potentially a serious loss in efficiency. Especially, since the perceptual model should be more or less the same at every node, transmitting it from every node is almost pure over-coding.

Conventional speech and audio coding methods consist of three parts; i) a perceptual model which specifies the relative impact of errors in different parameters of the codec, ii) a source model which describes the range and likelihood of different inputs and iii) an entropy coder which utilizes the source model to minimize perceptual distortion [5]. Further, the perceptual model can be applied in either of two ways: i) Signal parameters can be weighted according to the perceptual model, such that all parameters can then be quantized with the same accuracy. The perceptual model must then be transmitted to the decoder such that the weighting can be undone. ii) The perceptual model can alternatively be applied as an evaluation model, such that the synthesis output of different quantizations are compared, weighted by the perceptual model, in an analysis-by-synthesis iteration. Though here we do not need to transmit the perceptual model, this approach has the disadvantage that quantization cells shapes are not regularly shaped which reduces coding efficiency. More importantly, however, to find the optimal quantization, we must use a computationally complex brute-force search of different quantizations.

Since the analysis-by-synthesis approach thus leads to a computationally complex encoder, it is not a viable alternative for IoT. We must therefore device a method which has access to the perceptual model also at the decoder. However, as noted above, explicit transmission of the perceptual model (or equivalently, an envelope model of the signal spectrum), is not desirable because it lowers coding efficiency.

The proposed approach consists of two parts; distributed quantization of the input signal using random projections and 1 bit quantization similar to [6], and implicit transmission of the perceptual model. By quantization of random projections, we make sure that each transmitted bit encodes a unique piece of information and we thus avoid over-coding. The perceptual model is generated independently at each sensor-node and we transmit the quantized perceptually weighted signal. Note that perceptual weighting makes the signal more flat, but that the basic shape is retained. We can therefore inversely deduce what the original envelope must have been, even from the perceptually weighted signal.

Though distributed source coding is a well-studied subject (e.g. [7, 8]) and it has been applied in other applications such as video [9], only a few a have worked on distributed audio coding (e.g. [10, 11, 12, 13]), and none of them however address the over-coding problem with respect to perceptual and envelope models. The scalable coding approach in [14] comes the closest, but even that model includes envelope coding with scale factors. The multiple description coding approach does also have some similarities, but instead of distributed coding it has been applied only to packet loss concealment [15, 16].

2. Distributed Source Coding

The main topic of the current paper is perceptual modelling in a distributed coding system. However, for practical experiments we need to apply quantization of the signal. The topic of distributed source coding is well-understood [7, 8], whereby we will here not dwell deeply into its theory. Instead, we apply a simple quantization scheme inspired by the 1 bit quantization method which has been used in compressive sensing systems [6]. This approach provides a realistic platform where we can study the main focus of this paper, that is, perceptual modelling in a distributed system. A more detailed discussion of quantization is thus left for further study.

The objective of the quantizer is to allow quantization at

International Audio Laboratories Erlangen is a joint institute of Fraunhofer IIS and FAU.

http://dx.doi.org/10.21437/Interspeech.2016-27

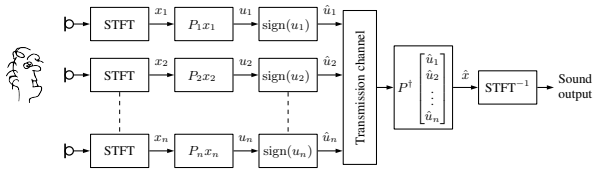

Figure 1: Flow-diagram of proposed distributed audio coding system, excluding perceptual modelling.

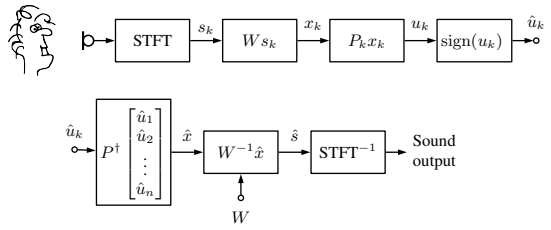

Figure 2: Flow-diagram of the kth sensor-node with perceptual weighting as well as the corresponding decoder.

independent sensors, such that we can ensure that each transmitted bit improves quality, without communication between sensor-nodes. In the extreme, we could have a sensor send only one bit and still be able to use that single bit to improve quality.

The proposed quantization scheme is based on random projections of a real-valued representation of the signal spectrum and transmitting the sign of each dimension. In other words, let x be the real-valued $N \times 1$ vector containing the spectrum of the input signal, and P a $K \times N$ random matrix whose columns are normalized to unit length. We will then transform x by $u = Px$ and quantize the sign of each component of u, that is, the quantization is $\hat{u} = \text{sign}(u)$, which can be transmitted losslessly with K bits. The reconstruction, an approximation of x can readily calculated by

$$\hat{x} = P^\dagger \hat{u} = P^\dagger \,\text{sign}(Px), \tag{1}$$

where $P^\dagger$ is the pseudo-inverse of P. As long as the seed for the pseudo-random columns of P is known at the decoder, the decoder can thus decode the signal from $\hat{u}$ only. In the case of multiple sensor-nodes, the input signal x is assumed to be the same or noisy versions of the same signal, but each sensor has its own random matrix P_k. At the decoder the random matrices can be collated to a single large matrix $P = [P_1, P_2 \dots]$ whereby Eq. 1 remains unchanged.

It is well-known that if $K \ll N$, then P is approximately orthonormal, $P^T P \approx I$ and the quantization is near-optimal. In our case K is not necessarily smaller than N, whereby the orthonormality becomes less accurate. Using the transpose instead of the pseudo-inverse decreases algorithmic complexity and coding efficiency, but does not impose a limitation to our experiments on perceptual modelling, since every transmitted bit still improves the accuracy of the output signal.

We expect that a source model would then be applied on the decoder side and that such a model would increase the accuracy of the reconstruction. Since our focus is on perceptual modelling, it is however not necessary to implement a source model, since its effect can be simulated by increasing the accuracy by sending more bits.

The flow-diagram of the overall system (excluding perceptual model) is illustrated in Fig. 1.

3. Perceptual Modelling

Speech and audio codecs are based on efficient modelling of human auditory perception. The objective is to obtain such a weighting of quantization errors that optimization of the signal-to-noise ratio in the weighted domain gives the perceptually best possible quality. Audio codecs operate generally in the spectral domain, where the spectrum of an input frame s can be perceptually weighted with a diagonal matrix W such that the weighted spectrum $x = Ws$ can be quantized $\hat{x} = \lfloor Ws \rceil$, where $\lfloor \cdot \rceil$ denotes quantization. At the decoder, we can reconstruct the spectrum with the inverse operation $\hat{s} = W^{-1}\hat{x}$. The flow-diagram of perceptual quantization is depicted in Fig. 2.

Specifically, the perceptual weighting model consists of two parts; i) A fixed part corresponding to the limits of perception at different frequency bands. Perceptual models such as Bark- and ERB-scales model the density of frequencies such that the warped axis has uniform perceptual accuracy [17]. However, since our objective is to measure error energy on the warped scale, we can equivalently scale the magnitude of spectral components such that the computationally complex warping operation can be avoided [18]. This operation is also similar to the pre-emphasis operation applied in speech codecs [1, 2, 3]. Since this part of the weighting is fixed, it does not need to be explicitly transmitted, we can apply it at the encoder and directly reverse it at the decoder.

ii) The signal-adaptive part of the perceptual model corresponds to the frequency-masking properties of perception. Namely, high-energy components of the signal will mask lower energy components and thus render them inaudible, if the two are sufficiently close in frequency [5]. The shape of the frequency-masking curve is thus equal to the shape of the signal envelope, but with a smaller magnitude.

If $|x|$ is the magnitude spectrum of the input signal, we can obtain its spectral envelope by $y = A\Lambda A^T|x|$, where matrix A is a filterbank such as in Fig. 3(a). In difference to the common MFCC-type filterbanks [19], we use asymmetric Hann-window type windows with an overlap extending from the kth filter to the $k-2$ and $k+2$ filters (see Fig. 3(a)). The diagonal matrix Λ contains normalization factors for each band such that we obtain unit-gain. We can here use a Mel-, Bark- or ERB-scale as desired and with a suitable number of bands. At a sampling rate of 12.8 kHz, we used a Mel-filterbank with 20 bands. An alternative to the MFCC-type filterbank matrix would be to use spreading by filtering, whereby A becomes a convolution matrix. Since filtering operations are well-understood digital signal processing methods, we can readily find their inverses as well.

The perceptual weighting factors model the frequency masking effect, which in turn corresponds to spreading and scaling of energy over frequencies [20, 5]. The envelope model matrix A already achieves the effect of spreading, whereby we still need to model scaling of energy.

The energy scaling corresponds to compression of the signal, which reduces the magnitude-range of the envelope (see Fig. 3(b)). Hence, if we multiply the spectrum s with the perceptual weighting matrix W, we obtain a spectrum $x = Ws$ which has a reduced range (see Fig. 3(c)). Perceptual weighting thus *reduces* the range or flattens the spectrum, but it does *not* produce a spectrum with an entirely flat envelope. The range of the envelope is reduced, whereby a part of its range is retained, and we can use that remaining range to recover the original signal following the expanded envelope.

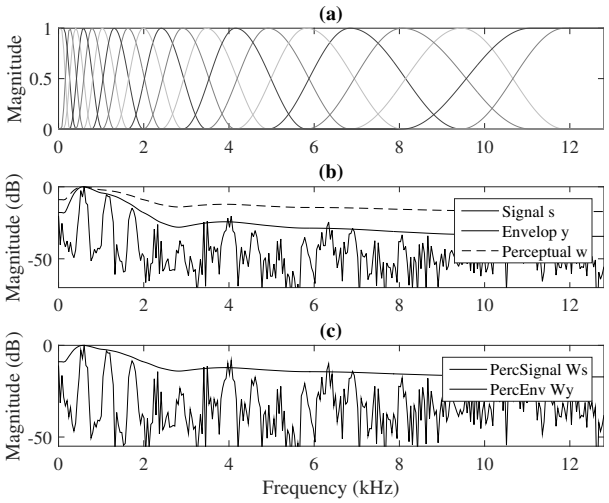

Figure 3: Illustration of (a) the columns of a filterbank matrix A on the Mel-scale with 20 bands, (b) the spectrum of a signal s, its envelope y and corresponding perceptual model w, and (c) the perceptually weighted spectrum Ws and the corresponding weighted signal envelope Wy.

The range-reduction or compression function $w = f(y)$ for the envelope y (where vector w gives the diagonal elements of W), can be applied for example as a sample-wise exponent $f(y) = y^p$ with $0 < p < 1$. While it is possible to use any function which compresses the range of y, exponentiation has the benefit that it leads to simple analytic expression in envelope reconstruction at the decoder.

The encoder algorithm is then:

1. Calculate the envelope of the magnitude spectrum.
2. Compress the envelope to obtain the perceptual weighting model.
3. Apply weighting on spectrum $x = Ws$.
4. Quantize and transmit weighted spectrum $\mathrm{sign}(Px)$.

This algorithm is applied independently at every sensor-node.

4. Reconstruction of the Perceptual Model

At the decoder side, we can recover an estimate $\hat{x}$ of the perceptual signal x (see Eq. 1) by $\hat{x} = P^{\dagger}\,\mathrm{sign}(Px)$. The main task is thus to recover an estimate $\hat{s}$ of the original signal s from the quantized perceptual signal $\hat{x}$. We know that $x = Ws$ whereby $\hat{x} \approx x = Ws \approx W\hat{s}$, furthermore, $w = f(s)$ and W depends on w. Therefore, if we have an estimate of w, we can estimate $\hat{s}$, whereby we can estimate w, and iterate until convergence. This is thus an Expectation Maximization -type algorithm.

Formally, we have the algorithm:

1. Get an initial guess of w_0 by, for example, $w_0 = A\Lambda A^T\hat{x}$ and set W_0 appropriately.
2. Repeat from $k = 0$ until converged
   (a) Calculate $\hat{s}_k = W_k^{-1}\hat{x}$
   (b) Calculate $w_{k+1} = f(\hat{s}_k)$ and set W_{k+1} appropriately.
   (c) Increase k.

The last values $\hat{s}_k$ and W_k are our final estimates of $\hat{s}$ and $\hat{W}$. Typically, less than 20 iterations are required for convergence.

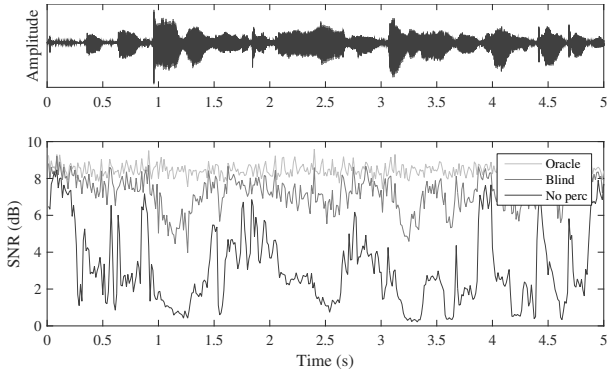

Figure 4: A speech wave-file and the perceptual SNR of the proposed quantization using 3 kbits/frame, with the oracle perceptual model, the blindly estimated perceptual model and as a reference, quantization without a perceptual model.

5. Experiments

To evaluate the performance of each part of the proposed system, we performed the following experiments. We compared three versions of the input audio; the quantized and reconstructed signal, 1) without and 2) with perceptual modelling such that the perceptual is known at the decoder, as well as 3) the perceptually quantized signal where the reconstruction is performed with the blindly estimated perceptual model.

As test material we used random speech samples from the NTT-AT dataset [21]. The input signals were resampled to 12.8 kHz, the STFT was implemented with discrete cosine transform to obtain a real-valued spectrum and we used an envelope model with 20 bands, distributed according to the Mel-scale [20, 5]. As a first approximation of the perceptual model, we used the range-reduction function of $f(y) = y^p$ with $p = 0.5$. This perceptual model was chosen merely as a way to demonstrate the performance of blind reconstruction, and should not be considered as a tuned end-product. The performance of the envelope model as well as the perceptual model were already illustrated in Fig. 3.

First, we will estimate the perceptual SNR for the quantization proposed without (SNR_O) and with blind reconstruction (SNR_B) of the perceptual model, respectively,

$$\mathrm{SNR}_O = \frac{\|x\|}{\|x - \hat{x}\|}, \quad \text{and} \quad \mathrm{SNR}_B = \frac{\|x\|}{\|x - W\hat{W}^{-1}\hat{x}\|}. \quad (2)$$

Figure 4 illustrates the perceptual SNR for a speech file quantized with the different methods ($K = 3000$). It is clear that when the perceptual model is known (oracle approach), the SNR is close to 8.4 dB. Blind reconstruction of the perceptual model clearly decreases quality (Blind) especially for voiced phonemes. However, the SNR of the system without a perceptual model (No perc) is more than twice worse than with blind recovery.

To further quantify the advantage of blind reconstruction instead of no perceptual modelling, we measured the mean SNR with different bit-rates K (see Fig. 5). The blind recovery and no-perceptual-model approaches are on average 1.1 dB and 5.8 dB worse than the oracle approach. Clearly SNR improves with bit-rate, though the no-perceptual-model case improves slower than with a perceptual model. Moreover, with increasing SNR, the blind recovery approaches the quality of the oracle approach asymptotically.

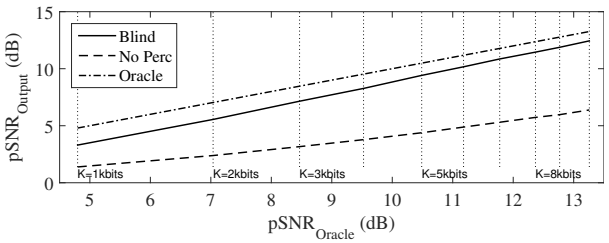

Figure 5: Performance of blind reconstruction of perceptual model as a function of the quantization SNR (solid line), in comparison to quantization without perceptual model (dashed line) and original perceptual model (dash-dot line). Vertical dotted lines indicate the number of bits per frame used.

Finally, to evaluate subjective quality, we performed a MUSHRA listening test with 8 listeners and 6 randomly chosen items from the NTT-AT dataset. The signal was quantized with 3 kbits/frame, which is a relatively low number given that we do not use any source modelling, whereby output SNR is also relatively low. This scenario was chosen to demonstrate a problematic condition and performance is expected to improve significantly at higher bit-rates as well as when applying a source model. From the differential MUSHRA scores in Fig. 6, we can see that for all items, perceptual modelling improves quality with both the oracle and blind estimation, by 29.9 and 22.3 points on average respectively. The statistical significance of the differences were confirmed with Student's t-test at $p > 99\%$.

6. Discussion

The proposed 1 bit quantization and coding scheme has several interesting consequences and properties. First, for analyzing *quantization* properties, note that each column of P is a projection to a 1-dimensional sub-space of the N-dimensional space of vector x. By encoding the sign of *one* projection, we thus split the N-dimensional space into two parts. By repeatedly encoding signs of Px, we thus keep splitting the N-dimensional space into ever smaller quantization cells. Since P is a random matrix, its columns are approximately orthogonal to each other, whereby the quantization cells remain near-optimal.

In a single node-system we could naturally design quantization approaches which are more efficient. However, in a distributed system it is more complicated – we need a simple method to prevent nodes from encoding the same information, that is, we need to avoid over-coding while retaining a low algorithmic complexity. The proposed quantization is most likely not an optimal quantization, but it is very simple and provides near-optimal performance.

Secondly, observe that we have not, in this preliminary study, employed *source coding* methods due to space limitations. It is however well-known that such modelling can be used to improve coding efficiency significantly. Source modelling can be applied at the decoder side by modelling the probability distribution of speech and audio signals (e.g. [22]). Source modelling is possible, since we can treat the quantized signal as a noisy observation of the "true" signal, whereby, by applying a prior distribution of the source, we can apply maximum likelihood optimization (or similar) to approximate the "true" signal. Since this optimization is applied in the network or at the decoder, the computational load is kept away from the sensor-nodes and the sensor-nodes can remain low-power.

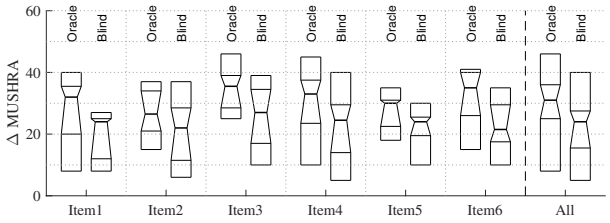

Figure 6: Box-plot of the differential MUSHRA scores of a subjective listening test, with oracle and blind perceptual modelling compared to no perceptual modelling. Positive values indicate improvement over no perceptual model.

Thirdly, from a *privacy* perspective, the random projection method can be designed to be a highly efficient encryption. If an eavesdropper does not know the seed for the random matrix, then the data will seem entirely random and meaningless. Assuming that the random seed is communicated in a secure manner, then only the encoder and the intended receiver can decrypt a message. This approach is in contrast to approaches such as [12, 13], where communication between nodes is intentionally employed. While such collaboration between nodes can be used to improve perceptual SNR, privacy is more difficult to guarantee. Even when assuming that sensor-nodes are operating over a secure network, it can take only one compromised node to gain access to all communications. In the proposed approach, in contrast, if an eavesdropper gains access to one sensor node, it compromises the data of that node only, since nodes can and should use different seeds. To limit the transmission power of sensor nodes, we can however allow that nodes relay packets, since packets remain readable by the intended recipient only and privacy is thus not compromised.

7. Conclusions

We have proposed a method for distributed audio coding, and provided a proof-of-concept for a mandatory part of that codec, the blind reconstruction of its perceptual model at the decoder side. The method is based on a 1 bit quantization idea, where on the encoder side, the perceptually weighted input signal is projected to random sub-spaces, and where the sign of each dimension is then transmitted. The decoder can invert the quantization with a pseudo-inverse, or similar, to obtain the quantized perceptually weighted signal.

The main part of the proposed method is then reconstruction of an estimate of the original signal, when we have access only to the perceptually weighted signal. The approach is based on an estimation-maximization (EM) algorithm, where we iteratively alternate between estimating the perceptual model and the original signal.

Our experiments demonstrate that 1) the 1 bit quantizer works as expected, 2) an approximation of the perceptual envelope can be blindly reconstructed with an average loss of 1.1 dB in perceptual SNR and approximately 7.6 MUSHRA points in comparison to an oracle.

The proposed distributed speech and audio coding algorithm is thus a viable approach for applications for the internet of things. It offers scalable performance for any number of sensor nodes and level of power consumption. Moreover, the algorithm is secure by design, since privacy of the communication channel can be guaranteed by encrypted communication of the random seed.

8. References

[1] *TS 26.445, EVS Codec Detailed Algorithmic Description; 3GPP Technical Specification (Release 12)*, 3GPP, 2014.

[2] *TS 26.190, Adaptive Multi-Rate (AMR-WB) speech codec*, 3GPP, 2007.

[3] ISO/IEC 23003–3:2012, "MPEG-D (MPEG audio technologies), Part 3: Unified speech and audio coding," 2012.

[4] M. Bosi, K. Brandenburg, S. Quackenbush, L. Fielder, K. Akagiri, H. Fuchs, and M. Dietz, "ISO/IEC MPEG-2 advanced audio coding," *Journal of the Audio engineering society*, vol. 45, no. 10, pp. 789–814, 1997.

[5] M. Bosi and R. E. Goldberg, *Introduction to Digital Audio Coding and Standards*. Dordrecht, The Netherlands: Kluwer Academic Publishers, 2003.

[6] P. T. Boufounos and R. G. Baraniuk, "1-bit compressive sensing," in *Information Sciences and Systems, 2008. CISS 2008. 42nd Annual Conference on*. IEEE, 2008, pp. 16–21.

[7] Z. Xiong, A. D. Liveris, and S. Cheng, "Distributed source coding for sensor networks," *IEEE Signal Process. Mag.*, vol. 21, no. 5, pp. 80–94, 2004.

[8] Z. Xiong, A. D. Liveris, and Y. Yang, "Distributed source coding," *Handbook on Array Processing and Sensor Networks*, pp. 609–643, 2009.

[9] B. Girod, A. M. Aaron, S. Rane, and D. Rebollo-Monedero, "Distributed video coding," *Proc. IEEE*, vol. 93, no. 1, pp. 71–83, 2005.

[10] A. Majumdar, K. Ramchandran, and L. Kozintsev, "Distributed coding for wireless audio sensors," in *Applications of Signal Processing to Audio and Acoustics, 2003 IEEE Workshop on*. IEEE, 2003, pp. 209–212.

[11] H. Dong, J. Lu, and Y. Sun, "Distributed audio coding in wireless sensor networks," in *Computational Intelligence and Security, 2006 International Conference on*, vol. 2. IEEE, 2006, pp. 1695–1699.

[12] A. Zahedi, J. Østergaard, S. H. Jensen, P. Naylor, and S. Bech, "Coding and enhancement in wireless acoustic sensor networks," in *Data Compression Conference (DCC), 2015*. IEEE, 2015, pp. 293–302.

[13] A. Zahedi, J. Østergaard, S. H. Jensen, S. Bech, and P. Naylor, "Audio coding in wireless acoustic sensor networks," *Signal Processing*, vol. 107, pp. 141–152, 2015.

[14] J. Li, J. Johnston, and W. Y. Chan, "Perceptual, scalable audio compression," Nov. 16 2010, US Patent 7,835,904.

[15] G. Kubin and W. B. Kleijn, "Multiple-description coding (MDC) of speech with an invertible auditory model," in *Speech Coding, IEEE Workshop on*, 1999, pp. 81–83.

[16] V. K. Goyal, "Multiple description coding: Compression meets the network," *IEEE Signal Process. Mag.*, vol. 18, no. 5, pp. 74–93, 2001.

[17] J. O. Smith III and J. S. Abel, "Bark and ERB bilinear transforms," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 6, pp. 697–708, 1999.

[18] T. Bäckström, "Vandermonde factorization of Toeplitz matrices and applications in filtering and warping," *IEEE Trans. Signal Process.*, vol. 61, no. 24, pp. 6257–6263, Dec. 2013.

[19] F. Zheng, G. Zhang, and Z. Song, "Comparison of different implementations of MFCC," *Journal of Computer Science and Technology*, vol. 16, no. 6, pp. 582–589, 2001.

[20] H. Fastl and E. Zwicker, *Psychoacoustics: Facts and models*. Springer, 2006, vol. 22.

[21] NTT-AT, "Super wideband stereo speech database," http://www.ntt-at.com/product/widebandspeech, accessed: 09.09.2014. [Online]. Available: http://www.ntt-at.com/product/widebandspeech

[22] S. Korse, T. Jähnel, and T. Bäckström, "Entropy coding of spectral envelopes for speech and audio coding using distribution quantization," in *Proc. Interspeech*, 2016.