INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Super-Wideband Spectral Envelope Modeling for Speech Coding

Guillaume Fuchs[1], Chamran Ashour[1], Tom Bäckström[2]*

[1] Fraunhofer Institut für Integrierte Schaltungen (IIS), Erlangen, Germany
[2] Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland
guillaume.fuchs@iis.fraunhofer.de

Abstract

Significant improvements in the quality of speech coders have been achieved by widening the coded frequency range from narrowband to wideband. However, existing speech coders still employ a limited band source-filter model extended by parametric coding of the higher band. In the present work, a super-wideband source-filter model running at 32 kHz is considered and especially its spectral magnitude envelope modeling. To match super-wideband operating mode, we adapted and compared two methods; Linear Predictive Coding (LPC) and Distribution Quantization (DQ). LPC uses autoregressive modeling, while DQ quantifies the energy ratios between different parts of the spectrum. Parameters of both methods were quantized with a multi-stage vector quantization. Objective and subjective evaluations indicate that both methods used in a super-wideband source-filter coding scheme offer the same quality range, making them an attractive alternative to conventional speech coders that require additional bandwidth extension.

Index Terms: speech coding, spectral envelope modeling, LPC

1. Introduction

Extension of the audio bandwidth in telephony has shown to be an important feature for the user experience [1]. Narrowband speech (NB, up to 4 kHz audio bandwidth), traditionally transmitted in communication, has been extended to wideband (WB, up to 8 kHz) in digital communications by different standards [2, 3]. This trend continued with the more recent introduction of super-wideband (SWB, up to 16 kHz) speech coders [4, 5]. Nonetheless, existing SWB coding solutions are all built on a dual-band system, coding the lower band separately from the upper band, according to a source-filter model of speech production and pure perceptual considerations using a parametric representation, respectively.

The purpose of this work is to study the feasibility of extending Code-Excited Linear Prediction (CELP), the most common coding scheme based on the source-filter model, to SWB speech, and to avoid the need to split the processing into two bands. One of the most important aspects in this regard is to extend spectral envelope modeling to SWB. The spectral envelope is generally modeled in speech coding by Linear Predictive Coding (LPC) parameters characterizing the vocal tract and acting as an autoregressive estimation of the speech. Although applying LPC to NB and WB speech is well studied in the literature [6], very little or no consideration has been given in the past to the SWB case.

The present paper considers two methods for modeling the spectral envelope: LPC and Distribution Quantization (DQ), an efficient representation of scaling factors in the frequency domain. The latter method was originally designed for perceptual audio coding operating in the frequency domain for quantization noise shaping [7, 8]. Both methods are associated with the same quantization technique, a state-of-the-art Multi-stage Vector Quantization (MSVQ), and evaluated within the same SWB CELP, extension of AMR-WB [3] operating at 32 kHz sampling frequency.

Results of objective and subjective tests indicate that both spectral envelope modeling methods allow extending CELP to SWB with a quality that is competitive with conventional and state-of-the-art dual-band coding systems combining WB CELP and parametric bandwidth extension, as standardized in MPEG-D USAC [9]. The new SWB system has the advantage of having a lower algorithmic delay and a lower structural complexity since it does not require additional resampling steps or filter banks.

The paper is organized as follows. In the next section, we will present how LPC and the vector quantization of its coefficients was extended to SWB. Section 3 is dedicated to the DQ method while section 4 introduces the extension of CELP to SWB used as a framework for evaluating the two envelope modeling methods. Results of evaluations are given in section 5 before concluding in section 6.

2. Linear Predictive Coding

Linear Predictive Coding (LPC) is a key component of CELP. We explain in the following how it can be extended to SWB speech.

2.1. Extension to SWB

The widening of the audio bandwidth has two major impacts on the design of LPC; the order of the prediction and the design of the high frequency pre-emphasis filter usually applied before the linear prediction analysis. As the number of samples increases with a higher sampling frequency, the prediction order must be adjusted accordingly. The LPC order of 10 usually adopted for NB was increased to 16 for WB and should probably be even higher for SWB. However, the order has a significant impact on bit-rate required for transmitting the LPC coefficients.

In AMR-WB, a high frequency pre-emphasis filter was introduced for reducing the spectral tilt due to the wide dynamic range between low and high frequencies of wideband speech. That is achieved by a first-order high-pass filter applied before the linear prediction analysis:

$$p(z) = 1 - \mu z^{-1}, \tag{1}$$

where μ is a fixed constant, called pre-emphasis factor, set to 0.68 in AMR-WB for speech sampled at 12.8 kHz. For estimating the pre-emphasis factor and the LPC order in SWB, the prediction gain was optimized between the signals before and after the pre-emphasis filter followed by the LPC analysis filter.

* Now with Ericsson Research, Stockholm, Sweden

http://dx.doi.org/10.21437/Interspeech.2019-1620

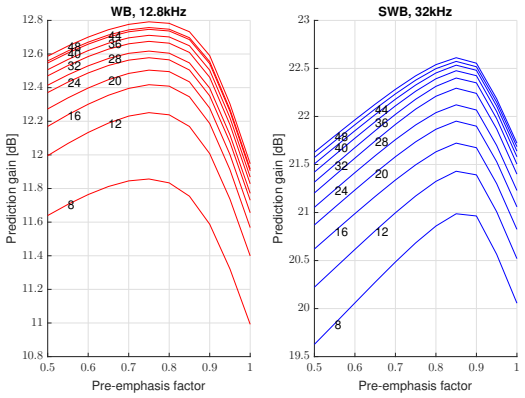

Figure 1: *Short-term prediction gain in function of LPC order and pre-emphasis factor for WB and SWB.*

The prediction gain is computed over a 30 min file of concatenated clean and noisy speech items. Figure 1 shows that the gain increases with the prediction order, but saturates at a certain point. By analogy with WB where 16 is considered sufficient, we chose an order of 24 for SWB. We chose the pre-emphasis factor of 0.85 which maximizes the prediction gain in SWB.

2.2. Spectral envelope parameters coding

In this work the Moving Average Multi-Stage Vector Quantizer (MA-MSVQ) was adopted since it is one of the most efficient techniques for coding the Linear Spectral Frequencies (LSF) representing the LPC coefficients. The multi-stage structure allows reducing complexity and memory requirements compared to a single stage quantizer which becomes complexity-wise impractical at high bit-rates. The structural constraint of MSVQ comes at the cost of some loss in coding efficiency, which can be greatly alleviated by optimizations.

The M-best search is one of the most efficient optimizations for MSVQ [10]. It follows M best paths from the first stage to the last stage for finding the optimum. The training can also benefit from a joint optimization of the stages instead of being iterative. The training and search procedures are further enhanced by taking into account the sensitivity of the LSFs and weighting them while optimizing error measures, such as weighted mean-squared error (WMSE) [11].

The Moving Average (MA) part of MA-MSVQ is a prediction of the current frame parameters by the previously quantized parameters, which exploits the similarities between successive frames. In our work the prediction factor is constant over time and was set to $1/3$.

The spectral distortion (SD) is a common measure of LPC quantization performance and is defined as follows:

$$\mathrm{SD} = \sqrt{\int_{\omega=0}^{\pi} 10\log_{10}\left(\left[\frac{P(\omega)}{\hat{P}(\omega)}\right]^2\right)}, \qquad (2)$$

where $P(\omega)$ and $\hat{P}(\omega)$ are the linear prediction power spectra before and after quantization, respectively. It is generally accepted that transparent quality is achieved if the average SD is less than 2 dB with less than 2 % outliers above 2 dB and no outliers above 4 dB [12]. Though, this empirical law can be relaxed for low bit-rates, where lower accuracy can be acceptable. Moreover, a perceptual weighting of SD was found to better predict listener preference [13] and is defined by

$$\mathrm{SD}_w = \sqrt{\frac{1}{W_0}\int_{\omega=0}^{\pi} W_b(\omega)10\log_{10}\left(\left[\frac{P(\omega)}{\hat{P}(\omega)}\right]^2\right)}, \qquad (3)$$

where W_0 is a normalization constant, and the Bark weighting $W_B(\omega)$ is defined as a function of sampling rate f_s by:

$$W_b(\omega) = \frac{1}{25 + 75\left(1 + 1.4\left(\frac{\omega f_s}{2\pi\cdot 1000}\right)\right)}. \qquad (4)$$

To code the 16 LSFs in WB, MA-MSVQ requires about 31 bits which is reflected in Table 1. Two codebooks have been trained, one to minimize SD, its performance is then reported by SD, and the other to minimize SD_w.

Table 1: *Spectral distortion (SD) and perceptual weighed spectral distortion (SD_w) for quantizing 16 LSFs in WB*

Methods	**Mean**	**Outliers%**	
	dB	$2-4$ dB %	>4 dB %
31 bits SD	1.31	2.04	0.00
31 bits SD_W	1.12	3.92	0.16

To code the 24 LSFs in SWB, different codebook sizes were evaluated as shown in Table 2. Allocating 39 bits seems to be sufficient according to the SD measure since increasing the quantizer size by one additional bit improves SD by only 0.01 dB. For SD_w, 39 bits in SWB leads to an average distortion comparable to 31 bits in WB at the expense of higher number of outliers between 2 and 4 dB. For quality evaluations in Section 5 the 39 bits codebook is considered.

Table 2: *Spectral distortion (SD) and perceptual weighed spectral distortion (SD_w) for quantizing 24 LSFs in SWB*

Methods	**Mean**	**Outliers%**	
	dB	$2-4$ dB %	>4 dB %
35 bits SD	1.52	7.37	0.27
38 bits SD	1.39	3.35	0.00
39 bits SD	1.35	2.32	0.00
40 bits SD	1.34	2.31	0.00
39 bits SD_w	1.18	8.04	0.28

3. Scale factors and DQ

The spectral envelope can alternatively be described directly in the frequency domain by scale factors discretizing the frequency axis and quantizing the magnitude spectrum. Scale factors have great flexibility usually at the expense of a high correlation between the factors, which increases the bit demand. For instance the discretization of the frequencies is flexible enough to follow any scale from a linear to a perceptual motivated scale like Mel-scale.

On the other hand, Distribution Quantization (DQ) is a recently proposed technique for efficiently representing scale factors in quasi decorrelated parameters, which ease the subsequent coding while keeping their great flexibility [7, 8]. The main principle is to divide the spectrum in an hierarchical manner following split points. To decide the position of the split

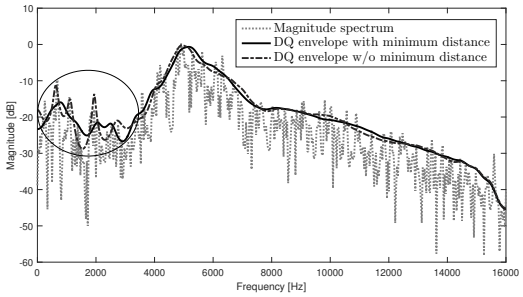

Figure 2: *Magnitude spectrum with its spectral envelope modeled by DQ with and without minimum distance of 430 Hz.*

Table 3: *Spectral distortion (SD) and perceptual weighed spectral distortion (SD_w) for quantizing 16 and 24 DQ energy ratios*

Methods	**Mean**	**Outliers %**	
	dB	$2-4$ dB %	>4 dB %
39 bits DQ16 SD	2.26	47.63	6.30
39 bits DQ16 SD_w	2.19	44.16	6.08
39 bits DQ24 SD	2.17	44.15	3.25
39 bits DQ24 SD_w	1.90	32.65	2.05

points different approaches have been investigated. In [7] the spectrum is split into segments of equal mass while in [8] split points represent the maximum variance of the signal. For each split point, an energy ratio is computed and coded. The first energy ratio reflects the coarsest distribution of the energy between the low and high frequencies while the subsequent energy ratios refine the details of the spectral envelope by modeling more local energy distribution. In our work and similar to [8], the energy ratio of each split point is coded. However, to keep the computational cost low, we use predefined split points.

In order to have a psychoacoustically relevant envelope, the Mel-scale was chosen to position on the frequency axis the split points. However, the Mel-scale resolution is too fine in low frequencies to correctly model the spectral envelope which can lead to modeling first harmonics of speech, especially in the case of female voices, which is undesired. To overcome this problem, a minimum distance of 430 Hz between the split points was required. Figure 2 illustrates the problem and how the minimum distance can solve it.

The number of energy ratios to transmit needs as in LPC to be defined in case of SWB speech. The objective is to obtain a spectral envelope as accurate as possible with minimum number of parameters. For a fair comparison between the two spectral envelope modeling methods, we decided to adopt the same quantization technique, namely MA-MSVQ, with the same number of bits. Given that 39 bits is allocated, the optimal number of DQ parameters was estimated by analyzing SD and with the objective evaluation method POLQA [14]. In Table 3, spectral distortions show that DQ using 16 and 24 parameters performs in the same range, which is also confirmed by the POLQA scores in Figure 4.

It is also interesting to note that the number of outliers is higher for DQ than for LPC. This can be explained by the interpolation employed to obtain the spectral envelope. From the energy ratios defined at the discrete split points on the frequency axis, a smoothed spectral envelope is derived using a spline interpolation as in [7]. However, the spectral sensitivity of the energy ratios to quantization errors is not taken into account when designing the quantizer, unlike what is done in the case of LPC with LSFs. It results that DQ requires a higher bit budget than LPC for the same number of coded parameters. Assigning 39 bits to the quantizer may then not be sufficient to correctly code 24 energy ratios. Considering memory requirements and complexity on top of that, we chose DQ16 for the comparison with LPC.

4. Super-Wideband CELP

State-of-the-art speech coders are mainly based on Coded-Excited Linear Prediction (CELP). It uses linear prediction for modeling short-term correlations and a long-term prediction for the fundamental frequency. The residual of the predictions are quantized and coded with an innovative codebook by perceptually optimizing the decoded synthesis signal in an Analysis-by-Synthesis (AbS) loop.

Extending CELP to a sampling rate of 32 kHz requires redesigning the LPC and the spectral envelope modeling, which was considered in the previous sections. Since CELP is based on linear prediction, the coded spectral envelope needs in case of DQ to be translated to LPC coefficients. For this, the spectral envelope defined in frequency domain is transformed in time domain and used as a short-term autocorrelation function for the Levinson-Durbin recursion. It was found that an order of 40 is usually sufficient for retaining most of the envelope details.

The perceptual weighting filter is, in CELP, derived from the non-coded LPC coefficients $A(z)$ and has the form:

$$W(z) = \frac{A(z/\gamma_1)}{1-\beta_1 z^{-1}}, \tag{5}$$

where in case of AMR-WB β_1 is equal to the pre-emphasis factor $\mu = 0.68$ and $\gamma_1 = 0.92$. Since the perceptual weighting filter of AMR-WB was proven to be efficient in the WB case, we aim to get a similar shaping in the limited band 0 to 6.4 kHz with a newly designed SWB perceptual weighting filter. Since $A(z)$ is not available for DQ, an additional LP analysis is performed independently of the envelope coding with a different order. In this respect, β_1 was fixed to the pre-emphasis factor, and the SWB parameters γ_1 and the order of $A(z)$ were optimized jointly by minimizing SD between the WB and SWB perceptual shaping in the frequency range 0 to 6.4 kHz. Optimization results are shown in Figure 3 where an order of 40 and $\gamma_1 = 0.97$ were found to be optimal.

The parameter coding of the long-term prediction remained unchanged from AMR-WB, only the fractional delay for the pitch lag was adjusted for the new sampling rate.

The extension of the innovative codebook for coding the residual of the predictions at 32 kHz was done in two steps. First, the framing was reduced from 20 ms to 16 ms, thus reducing the need to increase the dimension of the codevectors at the expense of more frequent transmission of coded parameters. In addition, the new innovative codebook was designed by upsampling by a factor of 2 the original AMR-WB codevectors: non-zero pulses are placed at even positions while odd samples are zeroed without the need of changing the indexing and the coding. This results in mirroring of low frequencies in high frequencies, which is a common non-linear operation in parametric bandwidth extension such as in AMR-WB+ [15]. This can be considered as an implicit bandwidth extension, with the difference that it is done after a short- and a long-term predic-

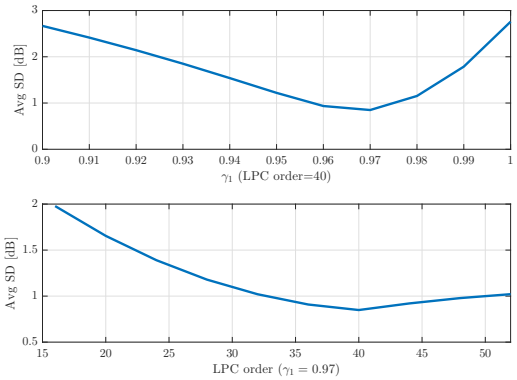

Figure 3: *SD between perceptual weighting filter of WB and SWB in 0 -6.4 kHz band.*

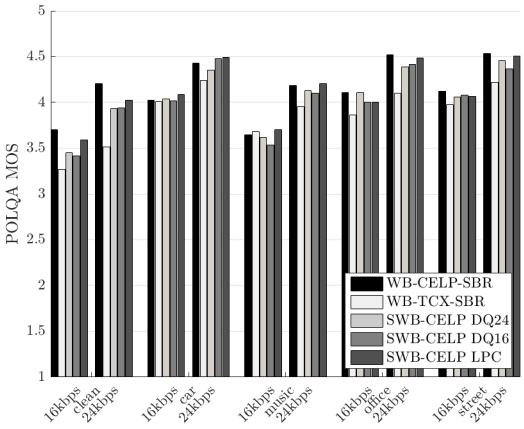

Figure 4: *POLQA scores for clean and noisy speech.*

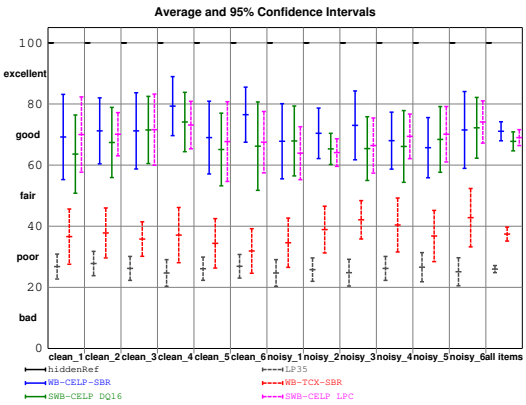

Figure 5: *MUSHRA results for clean speech and noisy speech.*

tion, with the effect of shaping the extended frequency content by both a spectral envelope and a harmonic structure.

The excitation gains codebook is also taken from AMR-WB so that in the end the new SWB CELP shares the majority of its coding tools with AMR-WB except for its spectral envelope coding.

5. Evaluations

The two SWB spectral envelope modeling approaches were evaluated within the proposed CELP extended for SWB speech both objectively and subjectively. The two methods were compared with a traditional WB-CELP combined with a bandwitdh extension for the higher band. For this purpose the speech coding mode of MPEG-D USAC was considered. It consists of the combination of a WB speech coder based on AMR-WB extended with the parametric bandwidth extension tool SBR introduced in HeAAC [16]. As a benchmark, the MDCT-based coding mode TCX of MPEG-D USAC [9] using low overlapping windows is also under test, even though this mode is not specifically designed for speech.

The tool for Perceptual Objective Listening Quality Assessment (POLQA [14]) was used for the objective assessment. 108 pair sentences from different languages were assessed in clean and noisy conditions, for which different background noises or music were mixed at different target SNR, namely 15 and 20dB. Figure 4 shows that the new SWB system performs slightly worse than the conventional WB-CELP combined with SBR for clean speech, for both spectral envelope modeling approaches. However, the gap is significantly reduced for noisy speech, where the new system can be considered on par with the state-of-the-art. It is not surprising that WB-TCX-SBR is the least efficient. Amongst the two spectral envelope coding schemes, LPC tends to perform the best.

To verify the objective POLQA scores, we also performed a formal subjective listening test at 24 kbps following the MUSHRA methodology [17]. 12 test items were tested and covered two categories: clean speech and speech mixed with background noise or music at different SNR levels ranging from 25 to 15 dB. The clean speech items come from the EBU SQAM disc [18] and consists of pairs of sentences uttered by male and female speakers in French, German and English. 10 expert listeners took part in the test. In addition to the conditions tested previously, a hidden reference (HR) and a 3.5 kHz low-pass anchor (LP35) were included. Results are given in Figure 5 and confirm the trend observed through POLQA scores that the new SWB system tends to be less efficient than the conventional WB-CELP system for clean speech while being on par for noisy speech. It also confirms the trend that LPC behaves better than DQ in SWB-CELP.

6. Conclusion and outlook

In this paper, we propose to widen the bandwidth of two spectral envelope modeling techniques for extending the source-filter model to SWB speech. In this context it has been demonstrated that CELP can be operated at 32 kHz while being competitive against traditional systems combining a WB-CELP and a parametric bandwidth extension. We further observed that the DQ approach is penalized by the low bit-rate constraint imposed by the adoption of the same vector quantization scheme optimized for LPC. While LPC with its LSF representation has a low spectral sensitivity to quantization errors, which is a substantial advantage at low bit rates, DQ requires a higher bit budget for the same number of parameters. Finally, the proposed SWB-CELP is in yet an early stage of development and various coding tools could still benefit from optimizations, which could be a topic for future investigations.

7. References

[1] J. Schnitzler and P. Vary, "Trends and perspectives in wideband speech coding," *Signal Processing*, vol. 80, pp. 2267–2281, 11 2000.

[2] *7kHz Audio Audio-Coding within 64kbit/s ITU-T Recommendation G.722*, ITU-T G.722. [Online]. Available: https://www.itu.int/rec/T-REC-G.722-198811-S/en

[3] B. Bessette, R. Salami, R. Lefebvre, M. Jelinek, J. Rotola-Pukkila, J. Vainio, H. Mikkola, and K. Jarvinen, "The adaptive multirate wideband speech codec (amr-wb)," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 8, pp. 620–636, Nov 2002.

[4] M. Neuendorf, P. Gournay, M. Multrus, J. Lecomte, B. Bessette, R. Geiger, S. Bayer, G. Fuchs, J. Hilpert, N. Rettelbach, R. Salami, G. Schuller, R. Lefebvre, and B. Grill, "Unified speech and audio coding scheme for high quality at low bitrates," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, April 2009, pp. 1–4.

[5] M. Dietz, M. Multrus, V. Eksler, V. Malenovsky, E. Norvell, H. Pobloth, L. Miao, Z. Wang, L. Laaksonen, A. Vasilache, Y. Kamamoto, K. Kikuiri, S. Ragot, J. Faure, H. Ehara, V. Rajendran, V. Atti, H. Sung, E. Oh, H. Yuan, and C. Zhu, "Overview of the evs codec architecture," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp. 5698–5702.

[6] D. O'Shaughnessy, "Linear predictive coding," *IEEE Potentials*, vol. 7, no. 1, pp. 29–32, Feb 1988.

[7] T. Jähnel, T. Bäckström, and B. Schubert, "Envelope modeling for speech and audio processing using distribution quantization," in *2015 23rd European Signal Processing Conference (EUSIPCO)*, Aug 2015, pp. 584–588.

[8] S. Korse, T. Jähnel, and T. Bäckström, "Entropy coding of spectral envelopes for speech and audio coding using distribution quantization," Sep 2016, pp. 2543–2547.

[9] *Information technology – MPEG audio technologies – Part 3: Unified speech and audio coding*, ISO/IEC 23 003-3:2012. [Online]. Available: https://www.iso.org/obp/ui/#iso:std:iso-iec:23003:-3:ed-1:v1:en

[10] W. P. Le Blanc, S. A. Mahmoud, and V. Cuperman, *Joint Design of Multi-Stage VQ Codebooks for LSP Quantization with Applications to 4 kbit/s Speech Coding*. Boston, MA: Springer US, 1993, pp. 101–109.

[11] W. R. Gardner and B. D. Rao, "Theoretical analysis of the high-rate vector quantization of lpc parameters," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 5, pp. 367–381, Sep. 1995.

[12] K. K. Paliwal and B. S. Atal, "Efficient vector quantization of lpc parameters at 24 bits/frame," *IEEE Transactions on Speech and Audio Processing*, vol. 1, no. 1, pp. 3–14, Jan 1993.

[13] A. McCree and J. C. De Martin, "A 1.7 kb/s melp coder with improved analysis and quantization," in *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98 (Cat. No.98CH36181)*, vol. 2, May 1998, pp. 593–596 vol.2.

[14] *Perceptual objective listening quality prediction, Recommendation ITU T P.863*, ITU-T, series P P.863. [Online]. Available: https://www.itu.int/rec/T-REC-P.863-201803-I/en

[15] J. Makinen, B. Bessette, S. Bruhn, P. Ojala, R. Salami, and A. Taleb, "Amr-wb+: a new audio coding standard for 3rd generation mobile audio services," in *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, vol. 2, March 2005, pp. ii/1109–ii/1112 Vol. 2.

[16] M. Dietz, L. Liljeryd, K. Kjorling, and O. Kunz, "Spectral band replication, a novel approach in audio coding," in *Audio Engineering Society Convention 112*, Apr 2002. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=11328

[17] *Method for the subjective assessment of intermediate quality level of audio systems, Recommendation ITU-R BS.1534-3*, ITU-R BS.1534-3. [Online]. Available: https://www.itu.int/rec/T-REC-P.863-201803-I/en

[18] *EBU SQAM CD, Sound Quality Assessment Material recordings for subjective tests*, EBU Tech 3000 series. [Online]. Available: https://tech.ebu.ch/publications/sqamcd