ON SPEECH QUALITY ASSESSMENT OF ARTIFICIAL BANDWIDTH EXTENSION

Patrick Bauer[1], *Cyril Guillaumé*[2], *Wouter Tirry*[2], *and Tim Fingscheidt*[1]

[1]Institute for Communications Technology, Technische Universität Braunschweig, Germany
[2]NXP Software, Leuven, Belgium

{bauer, fingscheidt}@ifn.ing.tu-bs.de, {cyril.guillaume, wouter.tirry}@nxp.com

ABSTRACT

During the transition to wideband speech telephony, artificial bandwidth extension (ABE) could help to preserve customer satisfaction by enhancing speech quality in case of narrowband (NB) calls. However, the assessment of speech quality for ABE systems is still an open question. In the literature, instrumental measures are often used to judge the quality of ABE solutions. When subjective listening tests are considered, they most often use a comparison category rating (CCR) scale and, more rarely, an absolute category rating (ACR) scale. This paper investigates the relevance of instrumental and subjective assessment methods for ABE systems. An ACR and a CCR test are organized. Their results are compared and discussed. Discrepancies between these two tests open the discussion for the design of a proper subjective listening test for ABE systems. Some instrumental measures are also evaluated. A poor correlation between these measures and the subjective results is observed.

Index Terms— speech quality assessment, bandwidth extension

1. INTRODUCTION

Despite upcoming mobile telephone speech services offering a wideband (WB) frequency range of $0.05 \ldots 7.0$ kHz, such as HD Voice [1, 2], most telephone calls are still narrowband (NB) providing a speech bandwidth < 4.0 kHz. In order to establish a WB call, both conversational partners need to have WB-capable terminals and be located in WB-capable cells of the same network provider. Otherwise, the call falls back into a NB mode, which will evoke customer dissatisfaction [3]. During the transition phase from NB to WB telephone speech services, artificial bandwidth extension (ABE) could serve as fall-back solution by recovering absent frequency components [4, 5]. Some ABE schemes perform not only an extension to the upper band, but also to low frequencies, however, this typically comes along with severe artifacts [6]. Furthermore, loudspeakers of small devices are hardly able to reproduce low frequencies in speakerphone mode anyway. This paper therefore focuses on those ABE approaches that only perform a high-band extension, such as [7].

Due to the lack of signal components in the high band, the intelligibility of NB telephone speech is severely reduced. Early experiments on meaningless syllables revealed a reduction of syllable articulation from 98 % to 90 %, when decreasing the upper cut-off frequency from 7 kHz to 3.4 kHz [8]. Particularly fricatives, such as /s/ and /z/, containing most of their energy parts at high frequencies, suffer from this [9]. Hence, most ABE approaches have problems to identify and extend them correctly [4]. Audible artifacts are the consequence [10]. By means of a phoneme-specific codebook design according to [11], the NB speech intelligibility of critical fricatives was considerably improved in [12]. Particularly hearing-impaired persons could benefit from that, as reported in [13, 14].

Bandwidth limitation also degrades speech quality. According to [15], WB speech reveals a mean opinion score (MOS) of 4.5, representing a speech quality between good and excellent, whereas NB speech attains only a fair speech quality of 3.2 MOS points. The ability of ABE to improve NB speech quality can be investigated by instrumental measures or subjective listening tests. For instrumental evaluation, distance measures as well as methods predicting the overall quality or quality dimensions are commonly applied, while subjective evaluations are usually obtained by listening-only and conversational tests [5, Sec. 6]. Instrumental assessments have the advantage of saving time and costs, but at the expense of an inaccurate prediction for ABE algorithms [16]. Möller et al. recently examined different ABE schemes via both instrumental and subjective speech quality assessment [17]. Among the instrumental measures, the ITU-T-recommended WB extension to perceptual evaluation of speech quality (WB-PESQ) [18, 19] attained the highest correlation with an absolute category rating (ACR) listening test [20, Annex B]. It even performed slightly better than its new successor perceptual objective listening quality assessment (POLQA) [21]. However, all instrumental measures revealed rank order problems [16]. Furthermore, none of the ABE methods was found to significantly improve NB speech quality. To our knowledge, the only *significant* speech quality improvement of ABE over NB obtained in an ACR test until now was reported in [22], showing a MOS gain of 0.25 points, however, no speech codecs were applied to the employed speech data.

In this paper, *coded* speech combined with different ABE implementations is assessed simulating real telephony conditions. The instrumental assessment is based on WB-PESQ and POLQA. For subjective evaluation, the best ABE candidates resulting from an ACR test are further evaluated by a comparison category rating (CCR) test [20, Annex E]. The results obtained from absolute and comparative rating scales are discussed, as well as the correlation between instrumental and subjective speech quality assessment.

The remainder of this paper is organized as follows: Sec. 2 introduces the ABE systems being investigated. After having described the setup of the performed subjective listening tests in Sec. 3, Sec. 4 discusses their results and compares them with those obtained by instrumental measures. Finally, conclusions are drawn in Sec. 5.

2. ABE SYSTEMS UNDER TEST

Two ABE systems have been employed for speech quality assessment. On the one hand, Section 2.1 briefly describes a state-of-the-art ABE largely following [4]. On the other hand, a phonetically motivated ABE built on the basis of [11] is presented in Section 2.2.

Part of this work was funded by the German Research Foundation (DFG) under grants no. FI 1494/2-1, and FI 1494/4-1.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

2.1. Baseline ABE System

The baseline ABE approach [4] operates as follows. As proposed in [4, Sec. 5.3.5], the following static features are extracted from the NB input speech frames: A gradient index, local kurtosis, normed relative frame energy, spectral centroid, zero crossing rate, and 10 auto-correlation coefficients. To save complexity of the statistical model, a linear discriminant analysis (LDA) is performed. It reduces the dimension of the feature vectors from 15 to 5 in case of using static features only, and from 45 to 10 in case of using first and second order dynamic features in addition. The elements of the LDA-transformed feature vectors are largely mutually uncorrelated. Hence, Gaussian mixture models with diagonal covariance matrices are trained via the expectation maximization algorithm to model the state observation probability density functions. A first-order hidden Markov model (HMM) defines 32 states that are uniquely related to pre-trained cepstral codebook vectors representing spectral envelopes of the upper frequency band. The vector quantizer codebook is trained by the Linde-Buzo-Gray algorithm. Due to real-time requirements, state a-posteriori probabilities are computed via the forward recursion [4, Sec. 6.4.1]. They are used to weight the codebook vectors in order to obtain a minimum mean square error estimation of the upper frequency band. After WB power spectrum assembly and inverse discrete Fourier transform, 16-order WB linear prediction analysis and synthesis filters are obtained from a Levinson-Durbin recursion. They are applied to the interpolated NB speech frames to remove and synthesize the estimated spectral shape caused by the human vocal tract, respectively. In between, the resulting NB residual is extended by spectral translation with a fixed modulation frequency of 4 kHz to estimate the excitation signal at the human vocal cords [4, Sec. 3.3.2].

We decided to implement three baseline ABE versions that mainly vary in the frame structure and feature extraction:

- The first version denoted by *ABE1a* does without frame overlap by applying a rectangular window with a frame shift of 20 ms. Hence, there is no look-back or look-ahead information available. Furthermore, only static features are taken into account.
- The second version *ABE1b* slightly differs from *ABE1a* in the training process by applying a Blackman window for the computation of the cepstral codebook vectors via selective linear prediction (SLP), as proposed in [4, Sec. 4.1.2][1].
- In contrast to *ABE1a* and *ABE1b*, the third version *ABE1c* makes use of 50% symmetrically overlapping frames with a frame shift of 10 ms and Blackman windowing. Additionally, it computes first and second order dynamic features based on the static ones. Due to real-time requirements, however, the computation of the Δ- and $\Delta\Delta$-features considers only one frame algorithmic delay [12].

2.2. Phonetically Motivated ABE System

A second ABE scheme is based on the baseline ABE implementation *ABE1c* in Sec. 2.1. According to this, it uses frame overlap and dynamic features, however, it varies in other parts of the algorithm.

In the ABE training process, the cepstral vectors are computed by means of a modified SLP technique that has also access to the preprocessed NB telephone speech signal in addition to the WB speech signal. In contrast to [4, Sec. 4.1.2], this provides a better match to the base-band gain factor computation in [4, Sec. 6.1.2]. Inspired by the phoneme-specific codebook training in [11, Sec. 4.1.2], 8 HMM states are purely trained on the phonemes /s/ and /z/, while the remaining 16 HMM states are dedicated to the other speech sounds. Different from [11], in order to find the 8 most individual representatives of /s/ and /z/, the cepstral distance between the mean and the 64 preliminary centroids is computed by omitting the zeroth cepstral coefficients. Thus, the spectral shape is taken more into account as compared to the absolute energy. Due to the overrepresentation of /s/ and /z/ in our codebook related to their normal appearance within speech, the HMM tends to stay in the /s/- and /z/-states provoking temporally smeared offsets. We therefore slightly increased the pre-trained joint state probabilities for transitions from these states to the remaining ones. To allow in general more transitions from one state to another, the main diagonal of the joint state probability matrix was further decreased. Moreover, zeros in the joint state probability matrix due to insufficient training data were smoothed out.

In the ABE test process, we applied a spectral folding technique for the extension of the excitation signal, as proposed in [13]. The modulation by means of a cosine function sampled at integer multiples of the Nyquist frequency produces a factor $(-1)^n$ alternating with sample index n [4, Sec. 3.3.1]. The resulting upper frequency band of the residual is attenuated by two weights. On the one hand, a fixed weight controls the upper-band energy of the ABE. On the other hand, artifacts during noisy speech pauses are reduced by means of an adaptive weight driven by a three-state speech pause detection (SPD) according to [23] (in order to make soft instead of hard decisions, the SPD has been slightly modified). Furthermore, the extension of speech sounds that tend to be degraded by ABE is adaptively suppressed via a pre-trained phonetic classifier. To further control the upper cut-off frequency of the ABE, a lowpass post-filter can be applied in steps of about 0.5 kHz.

We finally selected the following ABE versions by varying the overall aggressiveness of the extension via the fixed upper-band attenuation weight and the upper cut-off frequency of the post-filter:

- The first and most aggressive ABE version *ABE2A* includes 3 dB attenuation and a cut-off frequency of about 6.8 kHz.
- The second ABE version *ABE2B* involves an attenuation of 6 dB and about 6.3 kHz cut-off frequency.
- The third and most conservative ABE version *ABE2C* implies 9 dB attenuation and a cut-off frequency of about 5.8 kHz.

3. SUBJECTIVE LISTENING TESTS

For speech quality assessment of the ABE systems in Sec. 2 we performed both ACR and CCR listening tests. Sec. 3.1 first explains the preprocessing of the employed speech data. The ACR and CCR test setups are then described in Sec. 3.2 and 3.3, respectively.

3.1. Data Preprocessing

The employed speech data was taken from two female and two male German speakers of the well-known, multi-lingual NTT-AT database for telephonometry [24]. Each speaker provided four utterances of about 8 sec. For each utterance, 22 variants were preprocessed: Seven NB conditions, six ABE conditions, and nine WB conditions.

The NB conditions were derived from the original, 16 kHz-sampled speech data by bandpass-filtering to a range of about 0.2 ... 3.5 kHz via the MSIN highpass and FLAT1 lowpass filters, as specified by the ITU-T in [25], scaling to an active speech level of -26 dBov according to [26], and decimation to a sampling rate

[1] Please note that we decided to consistently use Blackman windowing for the base-band gain factor computation in [4, Sec. 6.1.2], too.

of 8 kHz. Subsequently, six of the NB conditions were generated by using the modulated noise reference unit (MNRU) [27] with speech to modulated noise power ratios of 6 dB, 12 dB, 18 dB, 24 dB, 30 dB, and ∞ dB (clean). The last NB condition was obtained by applying the adaptive multirate narrowband (AMR-NB) speech codec [28] at the bitrate 12.2 kbps. All NB conditions were finally interpolated to 16 kHz corresponding to the sampling rate of the ABE and WB conditions.

The AMR-NB coded files sampled at 8 kHz served as input for our ABE versions in Sec. 2 (i.e., *ABE1a-c* and *ABE2A-C*) to create the six ABE conditions. In contrast, the WB conditions were derived from the original, 16 kHz-sampled speech data by transmitter-sided P.341 filtering to a range of about $0.05 \ldots 7.0$ kHz according to [29] and scaling to an active speech level of -26 dBov [26]. The adaptive multirate wideband (AMR-WB) speech codec [30] was applied at the bitrates 8.85 kbps, 12.65 kbps, and 23.85 kbps for three WB conditions. The remaining six WB conditions were generated by using MNRU [27] with speech to modulated noise power ratios of ∞ dB (clean), 45 dB, 35 dB, 25 dB, 15 dB, and 5 dB.

3.2. ACR Test Setup

We performed a formal ACR listening test largely following [20, Annex B], being supported by 7 female and 17 male naïve German listeners without known hearing impairment. The subjects were acquired in student courses and compensated for their participation with a service charge. After having split them into three listening panels, each listening panel was further divided into two groups of four listeners. Using a standard laptop PC with an external `Roland UA-1010` sound card the test files were diotically presented to each group in a quiet room over four `Philips SHP-8900` equalized headphones at a sound pressure level of 73 dB.

One utterance per speaker was spent for a preliminary familiarization phase that comprised 16 files in total. The remaining utterances were used in the main test, which provided 88 files. For each listening panel a disjoint set of different test files was presented in random order. The files were randomly selected under the following constraint: Each set should contain all conditions for two female and two male varying utterances, so that every file was selected once over the whole listening test. Note that the speech codecs and ABE versions were applied in order to simulate different telephone speech conditions, whereas the MNRU conditions (and some AMR bitrates) mainly served as reference anchors to exploit the dynamic range of the absolute rating scale in MOS from 1 (bad) to 5 (excellent).

3.3. CCR Test Setup

After having evaluated the ACR results, the best candidate of each ABE system was compared to the AMR-NB condition by means of a formal CCR listening test largely following [20, Annex E]. For this purpose, 7 female and 9 male normal-hearing German non-experts were acquired in the same way as in the ACR test. We organized eight sessions of two listeners providing them with both condition pairs of all utterances and speakers in randomized A/B-B/A-orders. The resulting 128 files per subject were diotically presented in a quiet room via a standard laptop PC with an external `RME Fireface 400` sound card over an equalized `Philips SHP-8900` headphone at a sound pressure level that was individually calibrated during a preliminary familiarization phase. The listeners were allowed to repeat single samples of a condition pair before rating the quality of the second compared to the first sample in terms of comparison MOS (CMOS) between -3 (much worse) and $+3$ (much better).

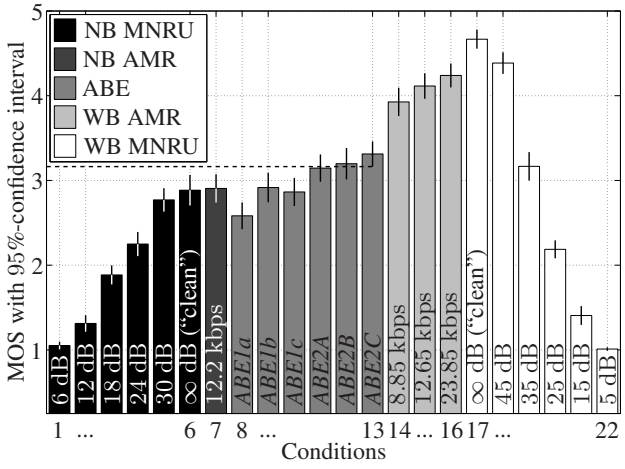

Fig. 1. Subjective results for the ACR listening test in terms of MOS averaged over 24 subjects given a 95%-confidence interval.

4. EXPERIMENTAL RESULTS

In addition to the ACR and CCR listening test results given in Sec. 4.1, Sec. 4.2 presents an instrumental evaluation based on WB-PESQ and POLQA. The subjective results as well as their correlation with the instrumental measures are further discussed in Sec. 4.3.

4.1. Subjective Speech Quality Assessment

The overall MOS results of the subjective ACR listening test are depicted in Fig. 1 for all conditions with 95%-confidence intervals. Obviously, the speech quality of the NB and WB MNRU reference anchors is reduced consistently by decreasing the speech to modulated noise power ratio. As expected, the "clean" WB condition 17 attains the highest MOS. The WB AMR conditions are rated somewhat lower and perform better for higher bitrates. Note, however, that the WB AMR conditions are scored somewhat higher than expected (especially for the lowest bitrate) and that no significant difference was observed between the clean NB and NB AMR conditions.

The baseline ABE versions are not able to exceed the NB AMR performance, however, the use of windowing is found to be important, especially for SLP in the training process (*ABE1b*, Sec. 2.1). The second ABE system points out significant improvements, particularly *ABE2C* with a 0.4 point MOS gain, bridging one third of the gap observed between a NB (condition 7) and a WB (condition 15) call. Reducing the aggressiveness of the extension according to Sec. 2.2 turns out to be beneficial.

Note that the results are very similar, when taking the MOS for each listening panel, headphone, listener, and speaker. A statistical analysis by using t-tests according to [31, Annex C] demonstrates that all ABE conditions are found to be "not worse than" the NB AMR condition except for *ABE1a*, given 95%- and 99%-confidence levels. Nevertheless, none of the baseline ABE conditions meets the requirements of being "better than" the NB AMR condition. However, the conditions *ABE2A-C* all proof to be better at a confidence level of 95% and 99%.

The NB AMR condition was finally compared to the ACR winner conditions of both ABE systems, i.e., *ABE1b* and *ABE2C*, in a subjective CCR listening test. Tab. 1 shows the overall CMOS results with 95%-confidence intervals. *ABE2C* attains a CMOS of 1.17, so it demonstrates a significantly better speech quality than NB AMR and the highest score among all ABE conditions in both listening

Conditions	CMOS
NB AMR vs. *ABE1b*	0.79 (±0.11)
NB AMR vs. *ABE2C*	1.17 (±0.09)

Table 1. Subjective results for the CCR listening test in terms of CMOS averaged over 16 subjects given a 95%-confidence interval.

tests. *ABE1b* also reveals a significant speech quality gain over NB AMR of 0.79 CMOS in the CCR test, in contrast to the ACR results.

4.2. Instrumental Speech Quality Assessment

In addition to the subjective listening tests, we performed a speech quality assessment based on instrumental measures. The WB-PESQ standard and its new successor POLQA (in superwideband mode) were therefore employed according to [19] and [21], respectively. Both instrumental measures use an absolute rating scale corresponding to the ACR listening test, which cannot be directly related to the comparative rating scale of the CCR listening test.

Tab. 2 presents the overall instrumental results in terms of MOS listening quality objective (MOS-LQO) compared to the MOS listening quality subjective (MOS-LQS) results of the ACR listening test with 95%-confidence intervals. Obviously, the instrumental assessment provides the same rank order as the subjective assessment for NB MNRU conditions 1-6, WB AMR conditions 14-16 and WB MNRU conditions 17-22. However, the absolute values differ by up to one MOS point, as it is exemplarily the case for WB-PESQ in condition 6 or POLQA in condition 19.

Taking into account all ABE conditions 8-13 including the NB AMR condition 7, the rank orders of the instrumental assessment significantly deviate from each other and particularly from the one of the subjective ACR test. Furthermore, the absolute MOS range for these conditions varies clearly among the assessment methods. Actually, POLQA provides the smallest range. In contrast to WB-PESQ and the subjective assessment, it rates the speech quality of all ABE conditions and the NB AMR condition quite high, i.e., above 3.2 MOS. WB-PESQ captures the absolute values of the subjective ratings better.

4.3. Discussion

Both instrumental measures perform quite well on the AMR and MNRU conditions, at least in predicting consistent rank orders. However, the subjective ABE ratings are hardly predicted reliably. The following Pearson correlations according to [18] confirm these observations. On the one hand, WB-PESQ and POLQA reveal overall correlations with the subjective assessment of $r = 0.92\,(p < 0.01)$ and $r = 0.96\,(p < 0.01)$, respectively. On the other hand, the correlations decrease to $r = 0.82\,(p < 0.05)$ for WB-PESQ and $r = 0.75$ for POLQA, when taking into account only the ABE conditions 8-13. The corresponding root mean square errors of WB-PESQ (0.14 MOS points) and POLQA (0.48 MOS points) suggest that WB-PESQ seems to work even better for ABE than its successor POLQA. Unfortunately, none of the employed instrumental measures is fully capable of replacing subjective speech quality assessments for ABE systems.

But also the subjective ratings reveal an inconsistency: *ABE1b* is found to be of the same speech quality as NB AMR in the ACR test, whereas it turns out to be significantly better than NB AMR in the CCR test. According to [32], the CCR results provide a higher sensitivity. Due to the direct comparison of NB AMR and ABE conditions in the CCR test, the listeners are assumed to be biased towards the extended bandwidth in spite of the ABE artifacts, whereas in the mixed-bandwidth ACR test, this bias is reduced and the weights of the perceptual dimensions related to the *bandwidth* and *artifacts* are more balanced in the subjective scores. Moreover, it may be extrapolated from these results that any comparison test will be influenced by such a bias. From an end-user point of view, ACR tests could be considered to better represent scenarios, where a WB call is followed by a separate NB call, or vice versa. In contrast, a handover call, which switches between WB and NB modes offering a direct comparison, may be better represented by CCR tests.

Conditions	WB-PESQ [MOS-LQO]	POLQA [MOS-LQO]	ACR test [MOS-LQS]
1	1.34 (±0.11)	1.28 (±0.12)	1.05 (±0.04)
2	1.78 (±0.22)	1.48 (±0.17)	1.31 (±0.10)
3	2.42 (±0.27)	1.81 (±0.28)	1.89 (±0.11)
4	3.00 (±0.25)	2.35 (±0.35)	2.25 (±0.14)
5	3.40 (±0.24)	3.07 (±0.25)	2.77 (±0.14)
6	3.97 (±0.11)	3.63 (±0.08)	2.89 (±0.18)
7	3.08 (±0.24)	3.42 (±0.15)	2.91 (±0.17)
8	2.63 (±0.08)	3.20 (±0.11)	2.58 (±0.16)
9	3.17 (±0.15)	3.53 (±0.12)	2.92 (±0.17)
10	2.93 (±0.13)	3.49 (±0.13)	2.86 (±0.17)
11	2.96 (±0.12)	3.44 (±0.14)	3.15 (±0.16)
12	3.12 (±0.13)	3.52 (±0.14)	3.20 (±0.18)
13	3.22 (±0.16)	3.53 (±0.14)	3.31 (±0.15)
14	3.14 (±0.21)	3.64 (±0.18)	3.93 (±0.17)
15	3.71 (±0.17)	4.11 (±0.15)	4.11 (±0.15)
16	4.03 (±0.14)	4.42 (±0.15)	4.24 (±0.14)
17	4.64 (±0.00)	4.68 (±0.03)	4.67 (±0.11)
18	4.47 (±0.03)	4.62 (±0.06)	4.39 (±0.13)
19	3.83 (±0.09)	4.19 (±0.19)	3.17 (±0.17)
20	2.64 (±0.11)	2.23 (±0.20)	2.19 (±0.11)
21	1.50 (±0.08)	1.29 (±0.09)	1.41 (±0.11)
22	1.10 (±0.01)	1.17 (±0.06)	1.01 (±0.02)

Table 2. Instrumentally assessed speech quality using WB-PESQ and POLQA compared to the subjective ACR test results from Fig. 1.

5. CONCLUSIONS

This paper investigates the speech quality assessment of artificial bandwidth extension (ABE). Besides an absolute category rating (ACR) and a comparison category rating (CCR) listening test, WB-PESQ and POLQA serve as instrumental measures. Although one ABE solution (*ABE2C*) consistently attains the highest score among all ABE conditions in both subjective tests outperforming NB AMR, another ABE solution (*ABE1b*) is found to be of the same speech quality as NB AMR in the ACR test, whereas it is judged to be significantly better in the CCR test. Due to the direct comparison between different bandwidths, CCR listeners are assumed to be biased towards higher bandwidths in spite of accompanying artifacts and provide more sensitive results. Thus, CCR tests may be rather suited for handover calls switching between WB and NB modes, and ACR tests for successive WB and NB calls. The instrumental measures turn out to poorly correlate with the ACR test results, so they are not recommended to be used for ABE assessment.

6. REFERENCES

[1] S. Ferraz de Campos Neto and K. Jarvinen, "Wideband Speech Coding Standards and Wireless Services [Guest Editoral]," *IEEE Communications Magazine*, vol. 44, no. 5, pp. 56–57, May 2006.

[2] T. Fingscheidt, "The Silent Speech Bandwidth Revolution in Mobile Telephony," IEEE Speech and Language Processing Technical Committee Newsletter, Aug. 2012.

[3] S. Möller, M. Wältermann, B. Lewcio, N. Kirschnick, and P. Vidales, "Speech Quality While Roaming in Next Generation Networks," in *Proc. of IEEE International Conference on Communications*, Dresden, Germany, June 2009.

[4] P. Jax, *Enhancement of Bandlimited Speech Signals: Algorithms and Theoretical Bounds*, Ph.D. thesis, vol. 15 of P. Vary (ed.), Aachener Beiträge zu digitalen Nachrichtensystemen, 2002.

[5] H. Pulakka, *Development and Evaluation of Artificial Bandwidth Extension Methods for Narrowband Telephone Speech*, Ph.D. thesis, School of Electrical Engineering, Aalto University, 2013.

[6] H. Pulakka, U. Remes, S. Yrttiaho, K. Palomäki, M. Kurimo, and P. Alku, "Bandwidth Extension of Telephone Speech to Low Frequencies Using Sinusoidal Synthesis and a Gaussian Mixture Model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 8, pp. 2219–2231, Oct. 2012.

[7] H. Pulakka and P. Alku, "Bandwidth Extension of Telephone Speech Using a Neural Network and a Filter Bank Implementation for Highband Mel Spectrum," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2170–2183, Sept. 2011.

[8] N.R. French and J.C. Steinberg, "Factors Governing the Intelligibility of Speech Sounds," *Journal of the Acoustical Society of America*, vol. 19, no. 1, pp. 90–119, Jan. 1947.

[9] P. Bauer, T. Fingscheidt, and M. Lieb, "Phonetic Analysis and Redesign Perspectives of Artificial Speech Bandwidth Extension," in *Proc. of Conference on Electronic Speech Signal Processing*, Frankfurt a.M., Germany, Sept. 2008.

[10] M. Nilsson and W. B. Kleijn, "Avoiding Over-Estimation in Bandwidth Extension of Telephony Speech," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Salt Lake City, UT, U.S.A., May 2001, vol. 2, pp. 869–872.

[11] P. Bauer and T. Fingscheidt, "A Statistical Framework for Artificial Bandwidth Extension Exploiting Speech Waveform and Phonetic Transcription," in *Proc. of European Signal Processing Conference*, Glasgow, Scotland, Aug. 2009, pp. 1839–1843.

[12] P. Bauer, M.-A. Jung, J. Qi, and T. Fingscheidt, "On Improving Speech Intelligibility in Automotive Hands-Free Systems," in *Proc. of IEEE International Symposium on Consumer Electronics*, Braunschweig, Germany, June 2010.

[13] P. Bauer, R.-L. Fischer, M. Bellanova, H. Puder, and T. Fingscheidt, "On Improving Telephone Speech Intelligibility for Hearing Impaired Persons," in *Proc. of 10th ITG Conference on Speech Communication*, Braunschweig, Germany, Sept. 2012, pp. 275–278.

[14] P. Bauer, J. Jones, and T. Fingscheidt, "Impact of Hearing Impairment on Fricative Intelligibility for Artificially Bandwidth-Extended Telephone Speech in Noise," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, Vancouver, BC, Canada, May 2013.

[15] W. Krebber, *Sprachübertragungsqualität von Fernsprech-Handapparaten*, Ph.D. thesis (in German), VDI Fortschrittsberichte, Reihe 10, Nr. 357, 1995.

[16] T. Fingscheidt and P. Bauer, "A Phonetic Reference Paradigm for Instrumental Speech Quality Assessment of Artificial Speech Bandwidth Extension," in *Proc. of 4th International Workshop on Perceptual Quality of Systems*, Vienna, Austria, Sept. 2013.

[17] S. Möller, E. Kelaidi, F. Köster, N. Côté, P. Bauer, T. Fingscheidt, T. Schlien, H. Pulakka, and P. Alku, "Speech Quality Prediction for Artificial Bandwidth Extension Algorithms," in *Proc. of Annual Conference of the International Speech Communication Association*, Lyon, France, Aug. 2013.

[18] "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," ITU, Feb. 2001.

[19] "ITU-T Recommendation P.862.2, Wideband Extension to Recommendation P.862 for the Assessment of Wideband Telephone Networks and Speech Codecs," ITU, Nov. 2007.

[20] "ITU-T Recommendation P.800, Methods for Subjective Determination of Transmission Quality," ITU, Aug. 1996.

[21] "ITU-T Recommendation P.863, Perceptual Objective Listening Quality Assessment," ITU, Jan. 2011.

[22] T. Ramabadran and M. Jasiuk, "Artificial Bandwidth Extension of Narrow-Band Speech Signals via High-Band Energy Estimation," in *Proc. of European Signal Processing Conference*, Lausanne, Switzerland, Aug. 2008.

[23] B. Fodor and T. Fingscheidt, "Reference-free SNR Measurement for Narrowband and Wideband Speech Signals in Car Noise," in *Proc. of 10th ITG Conference on Speech Communication*, Braunschweig, Germany, Sept. 2012, pp. 199–202.

[24] "Multi-Lingual Speech Database for Telephonometry," NTT Advanced Technology Corporation (NTT-AT), 1994.

[25] "ITU-T Recommendation G.191, Software Tool Library 2009 User's Manual," ITU, Nov. 2009.

[26] "ITU-T Recommendation P.56, Objective Measurement of Active Speech Level," ITU, Dec. 2011.

[27] "ITU-T Recommendation P.810, Modulated Noise Reference Unit (MNRU)," ITU, Feb. 1996.

[28] "Mandatory Speech Codec Speech Processing Functions: AMR Speech Codec; Transcoding Functions (3GPP TS 26.090, Rel. 6)," 3GPP; TSG SA, Dec. 2004.

[29] "ITU-T Recommendation P.341, Transmission Characteristics for Wideband Digital Loudspeaking and Hands-Free Telephony Terminals," ITU, Mar. 2011.

[30] "Speech Codec Speech Processing Functions: AMR Wideband Speech Codec; Transcoding Functions (3GPP TS 26.190, Rel. 6)," 3GPP; TSG SA, Dec. 2004.

[31] "Quality Assessment Characterisation/Optimisation step1 Test Plan for the ITU-T G.729 Based Embedded Variable Bit-rate (G.729EV) Extension to the ITU-T G.729 Speech Codec, Version 1.1," ITU, Nov. 2005.

[32] Sebastian Möller, *Assessment and Prediction of Speech Quality in Telecommunications*, Kluwer Academic Publishers, 2000.