A SUBJECTIVE LISTENING TEST OF SIX DIFFERENT ARTIFICIAL BANDWIDTH EXTENSION APPROACHES IN ENGLISH, CHINESE, GERMAN, AND KOREAN

Johannes Abel[1], *Magdalena Kaniewska*[2], *Cyril Guillaumé*[2], *Wouter Tirry*[2], *Hannu Pulakka*[3], *Ville Myllylä*[3], *Jari Sjöberg*[3], *Paavo Alku*[4], *Itai Katsir*[5], *David Malah*[5], *Israel Cohen*[5], *M. A. Tugtekin Turan*[6], *Engin Erzin*[6], *Thomas Schlien*[7], *Peter Vary*[7], *Amr H. Nour-Eldin*[8,*], *Peter Kabal*[8], *and Tim Fingscheidt*[1]

[1] Institute for Communications Technology, Technische Universität Braunschweig, Germany
[2] NXP Software B.V., Leuven, Belgium; [3] Microsoft Phone Technology, Tampere, Finland
[4] Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland
[5] Department of Electrical Engineering, Technion, Israel Institute of Technology, Haifa, Israel
[6] Multimedia, Vision and Graphics Laboratory, Koç University, Istanbul, Turkey
[7] Institute of Communication Systems, RWTH Aachen University, Germany
[8] Department of Electrical & Computer Engineering, McGill University, Montreal, Canada

{j.abel,t.fingscheidt}@tu-bs.de, {magdalena.kaniewska,cyril.guillaume,wouter.tirry}@nxp.com;
{hannu.pulakka,ville.myllyla,jari.sjoberg}@microsoft.com, paavo.alku@aalto.fi;
Itai.Katsir@audiocodes.com, {malah,icohen}@ee.technion.ac.il;
{mturan,eerzin}@ku.edu.tr; {schlien,vary}@ind.rwth-aachen.de;
amr.nour-eldin@mail.mcgill.ca, peter.kabal@mcgill.ca

ABSTRACT

In studies on artificial bandwidth extension (ABE), there is a lack of international coordination in subjective tests between multiple methods and languages. Here we present the design of absolute category rating listening tests evaluating 12 ABE variants of six approaches in multiple languages, namely in American English, Chinese, German, and Korean. Since the number of ABE variants caused a higher-than-recommended length of the listening test, ABE variants were distributed into two separate listening tests per language. The paper focuses on the listening test design, which aimed at merging the subjective scores of both tests and thus allows for a joint analysis of all ABE variants under test at once. A language-dependent analysis, evaluating ABE variants in the context of the underlying coded narrowband speech condition showed statistical significant improvement in English, German, and Korean for some ABE solutions.

Index Terms— listening test, ACR, artificial bandwidth extension

1. INTRODUCTION

Artificial bandwidth extension (ABE) belongs to the class of speech enhancement algorithms and aims at improving speech quality as well as speech intelligibility by extending a speech signal in its acoustical bandwidth. Given an incoming narrowband (NB) speech signal, i.e., a signal sampled at $f_s' = 8$ kHz, ABE solutions estimate and subsequently synthesize frequency components in the upper band (UB), i.e., the frequency range $4\,\text{kHz} < f \leq 8\,\text{kHz}$, and thus close up to so-called HD voice calls. HD voice stands for coded wideband (WB) speech, i.e., speech signals sampled at $f_s = 16$ kHz with an acoustical bandwidth up to 7 kHz. In WB speech, syllable intelligibility rises from 90% to 98% [1], while at the same time, the perceived speech quality gains about 1.3 mean opinion score (MOS) points in German language [2]. However, HD voice calls require the participants of a call to be in WB-capable mobile cells, use WB-capable handsets, be client of a WB-capable operator (inter-operator HD voice calls are often a problem), and the complete transmission path between the mobile cells also has to be WB-capable [3]. Whenever at least one of these requirements is not met, ABE solutions can serve as fallback to maintain speech intelligibility and speech quality to a certain degree.

Most of today's ABE solutions divide the extension process by means of the source-filter model into two subproblems: estimation of a spectral envelope as well as generation of a suitable residual signal, both for higher frequency components. Besides solving these subproblems, some ABE solutions go further and also modify the NB input signal, e.g., via equalizing [4]. Known techniques for the estimation of spectral envelope are, for example, Gaussian mixture models (GMMs) [5], hidden Markov models (HMMs) [6–8], (deep) artificial neural networks (ANNs) [7, 9], and others. The generation of a residual signal might be based on noise and/or impulse generation [5], modulation of the NB residual signal [6, 7, 10], and others.

For the time being, subjective listening tests are the only reliable evaluation method for ABE solutions [11, 12], especially w.r.t. rank order prediction of different ABE schemes. Typical subjective evaluations in the context of ABE schemes follow testing methods, standardized in [13], namely absolute category rating (ACR), degradation category rating (DCR), and comparison category rating (CCR). In [14] several ABE solutions were also tested in an anonymous fashion and compared in terms of statistical reliability to instrumental measures for speech quality prediction. The underlying listening

*The author is now with Nuance Communications Canada, Inc.

test was conducted in German. In [7, 11] ACR and CCR tests were conducted in German to evaluate two ABE solutions. Subjective listening tests of variants of a single ABE solution were performed in three languages in [15].

This paper describes a unified ACR listening test setup, testing conditions in the same manner in each language and using data from the same database, thus ensuring comparability throughout languages and conditions under test. Technische Universität Braunschweig and NXP Software conducted the ACR listening tests in American English, Chinese (Mandarin), German, and Korean, evaluating 12 variants from six different institutions or consortia. In this study, we focused on WB ABE solutions that can be implemented at the receiving side of a (mobile) telephony call.

The remainder of this paper is organized as follows. First, in Sec. 2, the listening test design is described. Subsequently, the preprocessing chain to create the conditions under test is explained in Sec. 3. The ABE approaches under test are briefly described in Sec. 4. Afterwards, in Sec. 5 the results over the different conditions per language are shown and discussed. A conclusion is given in Sec. 6.

2. OVERVIEW OF THE TEST DESIGN

The test was prepared similarly as described in [11], following largely ITU-T Recommendation [13] for ACR listening tests. The test was conducted with 48 listeners in four languages: American English, Chinese (Mandarin), German, and Korean. For each language the speech material consisted of the recordings of two male and two female speakers, four utterances per speaker. One of these utterances per speaker was used in a preliminary familiarization phase. The remaining three utterances were used in the main test. Since 12 variants of ABE solutions were tested, two listening tests (LTs) were prepared, namely LT1 and LT2, each designed to evaluate 6 ABE variants. To create a point of reference for ABE algorithms, 16 (= 7 NB and 9 WB) anchor conditions were included into each test, giving in total 22 conditions per LT. The files of each LT were further divided into three listening panels, each of them representing a disjoint set of different files, presented in random order, while still all conditions were included. This enabled evaluating a larger set of samples (12 speech files per condition per LT) without excessively extending the test duration.

Age and gender distribution for all conducted listening tests are shown in Table 1. Participants were all native speaker of the respective test language and stated to not suffer from hearing impairment. Mono audio files at 79 dB SPL were played through a `Roland Octa-Capture` interface and listened to with a monaural closed-back `Sennheiser HD-25 II` headphone. A proper equalization was applied to compensate for the headphones' frequency response.

A preliminary listening test, containing 32 files selected from anchor conditions, was performed to provide a proper reference and familiarize the listeners with the test procedure. The speech codecs and ABE versions were simulating different telephone speech conditions, whereas the MNRU conditions mainly served as reference anchors to exploit the range of the ACR scale in MOS from 1 (bad) to 5 (excellent).

3. DATA PREPROCESSING

A speech corpus recorded by Speech Ocean [16] was used, employing the same recording environment for all of the tested languages. The corpus is sampled at 48 kHz. The preprocessing chain is depicted in Fig. 1. To create a point of reference for the ABE solutions under test, 16 anchor conditions, more precisely 7 NB and 9 WB anchor conditions were processed and became part of every listening test. The preprocessing is based on [17].

Language	LT	#Males	#Females	Average Age
English	1	20	4	49
	2	18	6	44
	1+2	38	10	46
Chinese	1	12	12	26
	2	14	10	35
	1+2	26	22	31
German	1	12	12	24
	2	14	10	26
	1+2	26	22	25
Korean	1	12	19	38
	2	15	9	20
	1+2	27	21	29

Table 1. Gender and age distribution of participants in each of the listening tests (LTs).

For the 7 NB anchor conditions, first, a decimation from 48 kHz to 16 kHz using a high-quality (HQ) low-pass filter HQ3 is performed. The resulting signal is then subject to the mobile station input (MSIN) high-pass filter [18], simulating handset microphone characteristics. The signal is then decimated using another high-quality low-pass filter HQ2 to 8 kHz and then adjusted to an active speech level [19] of -26 dBov. Simulating a mobile NB phone call, the intermediate result **NB′** is subject to 13 bit conversion [18], encoding and subsequently decoding (ENC/DEC) using the adaptive multirate narrowband (AMR-NB) speech codec [20] at 12.2 kbps and again the 13 bit conversion [18]. Following an interpolation to 16 kHz, the result is referred to as the **AMR-NB** anchor condition. In addition, **NB′** is processed via the modulated noise reference unit (MNRU) [21] with speech-to-modulated-noise power ratios of 6 dB, 12 dB, 18 dB, 24 dB, 30 dB, and ∞ dB (direct). After interpolation to 16 kHz, this processing path leads to 6 **NB-MNRU** anchor conditions.

For the 9 WB anchor conditions, first, a 50 Hz high-pass filter (HP50) is applied, followed by a decimation by a factor of three using HQ3 and active speech level adjustment to -26 dBov [19]. The intermediate result **WB′** is then the basis for further processing. For simulation of mobile HD-Voice calls, **WB′** is converted to 14 bit [18], encoded and subsequently decoded by the adaptive multirate wideband (AMR-WB) speech codec [22] at bitrates 8.85 kbps, 23.05 kbps, and 23.85 kbps and again converted to 14 bit representation [18]. The three resulting WB anchor conditions are referred to as **AMR-WB**. In addition, **WB′** is subject to MNRU processing [21] with speech to modulated noise power ratios of 5 dB, 15 dB, 25 dB, 35 dB, 45 dB, and ∞ dB (direct), leading to 6 **WB-MNRU** anchor conditions.

The ABE solutions under test are applied to the **AMR-NB** condition sampled at 8 kHz. Finally, *all* files are postprocessed by P.341 filtering [18], i.e., limited to an acoustical bandwidth of 0.05-7 kHz and interpolated to 48 kHz sampling rate.

4. ABE APPROACHES

In this section, the 6 ABE approaches under test are briefly described. All ABE approaches are based on the source-filter model for speech production. Some of the institutions or consortia participated with several ABE schemes or parameter settings, resulting in a total of 12 **ABE** variants / test conditions. The contributing partners were asked to use blind ABE schemes with a maximum of 30 ms algorithmic delay. Please note that the approaches are ordered alphabetically after the contributing institutions, with the ordering unrelated to that of the results presented later in Section 5.

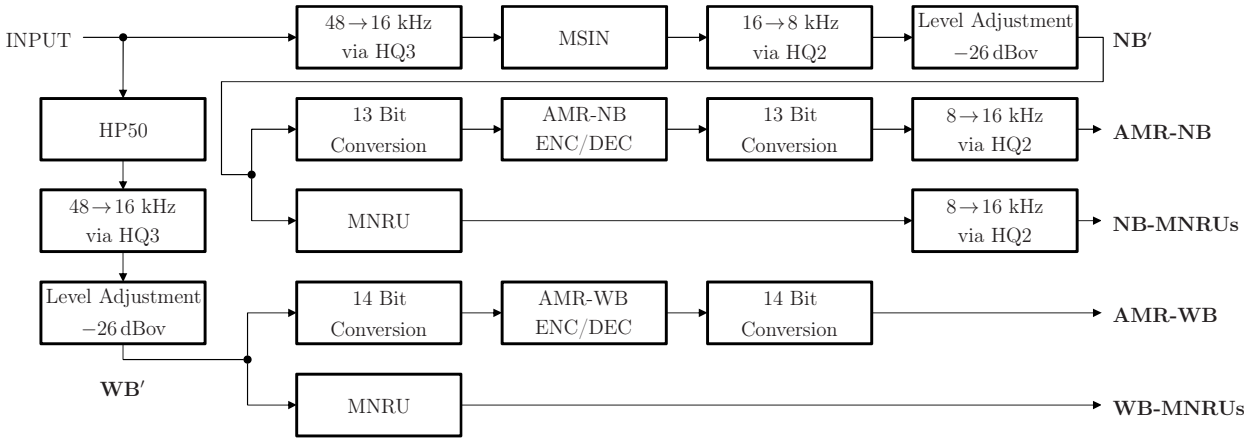

Fig. 1. Block diagram of the data preprocessing steps; **ABE** conditions are processed subsequently to the **AMR-NB** condition.

4.1. Koç University, Istanbul

The spectral envelope is estimated along the Viterbi path of the NB spectral envelope with minimum mean square error estimators [8]. The excitation extension is done using synchronous overlap and add on the NB residual spectrum [23].

4.2. McGill University, Montreal

The ABE scheme is based on [5]. A GMM models the spectral envelope using MFCCs while the UB excitation signal is obtained by generation of white Gaussian noise modulated by the NB-based equalized 3-4 kHz midband signal. In one variant based on [24], speech temporal information is accounted for by incorporating delta features into feature vectors, with the optimal static and delta feature dimensionalities determined via empirical optimization. In a second variant based on the idea of tree-based GMM extension [25], temporal information is accounted for by modeling the high-dimensional distributions of long-term feature vectors via temporally-extended GMMs.

4.3. Microsoft / Aalto University, Finland

The ABE scheme is based on [9]. An excitation signal is generated from the linear prediction residual of the NB input signal by spectral folding. An ANN is used to estimate the spectral envelope of the UB from input features, and the spectrum is shaped with a time-domain filter bank. An additional variant with a more conservative extension was also provided.

4.4. RWTH Aachen University

The ABE algorithm is based on [10, 26]. The excitation is extracted from the NB signal, spectrally flattened and copied to the UB with some additive white noise. The spectral envelope applied afterwards is estimated with the help of an HMM model with 128 states and 16 Gaussian mixture components per state based on zero crossing rate and 13 MFCC features. A quadrature mirror filter synthesis filter bank is used to combine NB and artificial UB to the final WB signal.

4.5. Technion, Haifa

The ABE approach is based on [4]. An excitation signal is generated using a simple spectral copying technique. The spectral envelope is estimated by means of a phonetic and speaker dependent statistical approach. Speech phoneme information is extracted using an HMM. Speaker vocal tract shape information is extracted by a codebook search. Further processing of the estimated vocal tract shape includes iterative tuning. Low frequencies in the NB signal are emphasized using an equalizer filter.

4.6. Technische Universität Braunschweig / NXP Software

The ABE approach is based on [27]. Following the source-filter model, the spectral envelope is estimated using an HMM, while the NB residual is extended applying spectral folding. HMM states are defined in favor of critical phonemes to reduce misrepresentation [28]. Additional classifiers are employed to adaptively correct overestimations [7, 11].

5. RESULTS

It is not recommended to perform a language-independent analysis of the obtained results, since merging of listening tests would change statistical properties, e.g., rank order of the different conditions. However, the overall test is designed to allow mapping of listening tests LT1 and LT2 *within* a language and thus enabling language-dependent comparisons of all conditions at once. This assumption was verified via hypothesis testing [31] of anchor conditions checking for equality throughout both tests. Therefore, the subjective votes of anchor conditions are put together and an anchor-condition-based and language-dependent mean is calculated. Afterwards, linear regression coefficients for a mapping of the anchor conditions towards the former mentioned language-dependent means are calculated and applied to the scores of the **ABE** conditions. The following language-dependent analysis is based on the results of this mapping process, merging LT1 and LT2 into one single listening test per language, LT1+2.

Table 2 presents condition-based means for every language, after linear mapping was applied. Clearly, the scores for both **NB-** and **WB-MNRU** conditions are proportional to the speech-to-modulated-noise-power ratio of the respective condition, thus showing to which extent the MOS scale was used. The **WB-MNRU** at ∞ dB condition was scored the highest over all languages. Interestingly, the score gap between **NB-MNRU** at ∞ dB and **WB-MNRU** at ∞ dB shows a high dependency on language. On one hand, German participants differentiate these two conditions by about 1.38 MOS points, thus substantiating the results obtained in [2] also on coded speech. On the other hand, Chinese participants scored the **WB-MNRU** at ∞ dB condition only 0.3 MOS points higher than the respective NB condition. The gaps for English and Korean are between 0.53 and 0.95 MOS points, respectively. Compared to, for example, American English, Chinese contains fewer fricative sounds [29, 30], the energy of which lies mostly at the higher frequencies, and hence, could be one of the explanations for the rather small noticeable difference between NB and WB in this language.

The **AMR-NB** and **AMR-WB** coded conditions were also

Condition	Chinese	English	German	Korean
NB-MNRU				
6 dB	1.15	1.08	1.06	1.09
12 dB	1.63	1.55	1.40	1.32
18 dB	2.21	2.16	1.89	1.84
24 dB	2.82	2.64	2.29	2.42
30 dB	3.36	3.24	2.93	2.96
∞ dB	4.12	3.76	3.31	3.54
WB-MNRU				
5 dB	1.13	1.10	1.03	1.03
15 dB	1.64	1.67	1.56	1.54
25 dB	2.58	2.66	2.44	2.44
35 dB	3.65	3.64	3.58	3.57
45 dB	4.23	4.20	4.57	4.29
∞ dB	4.43	4.29	4.70	4.49
AMR-NB	3.98 (07)	3.48 (09)	3.07 (07)	3.37 (08)
ABE$_{01}$	4.11 (02)	3.62 (04)	3.30 (04)	3.52 (06)
ABE$_{02}$	4.16 (01)	3.78 (03)	3.42 (01)	3.75 (01)
ABE$_{03}$	4.04 (04)	3.78 (02)	3.34 (03)	3.64 (02)
ABE$_{04}$	3.99 (05)	3.61 (05)	3.21 (06)	3.58 (03)
ABE$_{05}$	2.96 (13)	3.04 (11)	2.53 (11)	3.09 (11)
ABE$_{06}$	2.96 (12)	2.96 (12)	2.53 (12)	2.84 (12)
ABE$_{07}$	4.08 (03)	3.57 (06)	3.41 (02)	3.53 (04)
ABE$_{08}$	3.98 (06)	3.82 (01)	3.27 (05)	3.48 (07)
ABE$_{09}$	3.71 (09)	3.56 (07)	3.05 (08)	3.53 (04)
ABE$_{10}$	3.47 (10)	3.16 (10)	2.69 (10)	3.32 (09)
ABE$_{11}$	3.72 (08)	3.50 (08)	2.75 (09)	3.32 (09)
ABE$_{12}$	3.05 (11)	2.87 (13)	2.48 (13)	2.63 (13)
AMR-WB				
8.85 kbps	3.97	3.90	3.98	3.91
23.05 kbps	4.37	4.22	4.44	4.41
23.85 kbps	4.27	4.12	4.47	4.27

Table 2. Language-dependent results of listening tests after linear regression of LT1 and LT2 towards anchor conditions mean resulting in LT1+2.; (..) shows the rank order of **ABE** and **AMR-NB** conditions.

scored plausibly, with the higher bit rate and acoustical bandwidth of **AMR-WB** being rewarded by the participants. Interestingly, **AMR-NB** and **AMR-WB** at the lowest bit rate were scored similarly in Chinese.

For **ABE** and **AMR-NB** conditions, Table 2 also shows the rank order of the condition-based mean opinion scores. In general, ABE solutions are not perceived similarly across languages. While the rank of **AMR-NB** is roughly in the center of the ABE ranks, the rank of certain ABE solutions varies quite a lot. **ABE**$_{08}$ as an example is the best ABE variant in American English, however, in Korean the same ABE approach is ranked at position 7.

To analyze the question, whether an ABE solution improves the underlying **AMR-NB** condition, a simple comparison of condition-based mean values is not sufficient. Instead, a pair-wise comparison of the condition-based means of all **ABE** conditions vs. **AMR-NB** condition using two-sample t-test [31] was performed. In detail, for each of the pair-wise comparisons, the null hypothesis $H_0 : \mu_1 = \mu_2$ that both condition means μ_1, μ_2 are equal to each other is tested. The resulting p-values are shown in Table 3. The higher the p-values, the more likely is the null hypothesis. The following conclusions assume that if $p < 0.05$ then the difference between the means of the two conditions is statistically significant.

First of all, none of the evaluated ABE approaches was able

Condition μ_1	μ_2	p-value for H_0 true and H_1 false: Chinese	English	German	Korean
ABE$_{01}$	**AMR-NB**	0.49	0.26	<0.05 (+)	0.37
ABE$_{02}$		0.19	<0.05 (+)	<0.01 (+)	<0.01 (+)
ABE$_{03}$		0.92	<0.05 (+)	<0.05 (+)	0.08
ABE$_{04}$		0.71	0.30	0.34	0.13
ABE$_{05}$		<0.01 (−)	<0.01 (−)	<0.01 (−)	<0.05 (−)
ABE$_{06}$		<0.01 (−)	<0.01 (−)	<0.01 (−)	<0.01 (−)
ABE$_{07}$		0.17	0.49	<0.01 (+)	0.10
ABE$_{08}$		0.75	<0.01 (+)	<0.05 (+)	0.22
ABE$_{09}$		<0.01 (−)	0.51	0.84	0.09
ABE$_{10}$		<0.01 (−)	<0.05 (−)	<0.01 (−)	0.86
ABE$_{11}$		<0.05 (−)	0.80	<0.01 (−)	0.84
ABE$_{12}$		<0.01 (−)	<0.01 (−)	<0.01 (−)	<0.01 (−)

Table 3. Results of the two-sample t-test for null hypothesis test: **ABE** solutions vs. **AMR-NB**; (+): **ABE** condition mean is higher than **AMR-NB**, (−): vice versa.

to give a consistent improvement over **AMR-NB** in all languages. Particularly in Chinese, all ABE variants failed to show significant improvement over **AMR-NB**. Furthermore, half of the tested ABE solutions even degraded compared to the **AMR-NB** condition. This might also be explained by the former mentioned smaller difference between NB and WB speech w.r.t. the language and the resulting lack of perceived higher acoustical bandwidth. At the same time, artifacts introduced by ABE solutions remain and cause a degradation of subjective speech quality. For Korean, only **ABE**$_{02}$ showed improvement over the **AMR-NB** condition, 8 out of 12 ABE variants did not show significant difference compared to **AMR-NB** and the last three degraded the quality with statistical significance. It is worth noting that most of the ABE methods did not use Chinese and Korean speech data in training hence potentially explaining the poor ABE result for these languages.

In English, ABE solutions **ABE**$_{02}$, **ABE**$_{03}$, and **ABE**$_{08}$ show significant improvement over the **AMR-NB** condition. The same is valid for the first three ABE conditions as well as **ABE**$_{07}$ and **ABE**$_{08}$ in German.

If an ABE solution is significantly better than **AMR-NB**, how much of the gap between **AMR-NB** and **AMR-WB** at 23.05 kbps could be closed? To calculate this measure of performance, the condition-based mean of the ABE solution is subtracted by the corresponding mean of the **AMR-NB** condition and then divided by the MOS distance between the **AMR-NB** and **AMR-WB** conditions. In English, **ABE**$_{02}$ and **ABE**$_{03}$ filled the gap by 40% while **ABE**$_{08}$ closed the gap by 46%. For German, **ABE**$_{08}$, **ABE**$_{01}$, **ABE**$_{03}$, **ABE**$_{07}$, and **ABE**$_{02}$ could fill the gap by 15%, 17%, 20%, 24%, and 25%, respectively. In Korean, **ABE**$_{02}$ closed the gap by 36%.

6. CONCLUSIONS

In this work, listening tests in American English, Chinese, German, and Korean were conducted, evaluating 12 variants of ABE algorithms processed by six institutions and consortia. Due to the large number of conditions under test, listening tests were split into two separate tests per language. A carefully chosen listening test design enabled merging the scores from both listening tests via linear mapping, thereby making it possible to perform a language-dependent analysis of all ABE variants in a joint manner.

It was shown that some ABE solutions were able to improve the underlying coded narrowband speech signal with statistical significance in English, German, and Korean. In these languages it was possible to close the gap between coded narrowband and wideband speech by up to 46%, 25%, and 36%, respectively.

7. REFERENCES

[1] 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.

[2] W. Krebber, *Sprachübertragungsqualität von Fernsprech-Handapparaten*, Ph.D. thesis, vol. 10, no. 357 of VDI Fortschrittsberichte, 1995.

[3] Global mobile Suppliers Association, "Mobile HD voice: Global Update report," Information Papers, Apr. 2015.

[4] I. Katsir, D. Malah, and I. Cohen, "Evaluation of a Speech Bandwidth Extension Algorithm Based on Vocal Tract Shape Estimation," in *Proc. of International Workshop on Acoustic Signal Enhancement (IWAENC)*, Aachen, Germany, Sept. 2012, pp. 1–4.

[5] A. H. Nour-Eldin, *Quantifying and Exploiting Speech Memory for the Improvement of Narrowband Speech Bandwidth Extension*, Ph.D. thesis, Dept. of Electrical and Computer Engineering, McGill University, 2013.

[6] P. Jax and P. Vary, "Wideband Extension of Telephone Speech Using a Hidden Markov Model," in *Proc. of IEEE Workshop on Speech Coding*, Delavan, WI, USA, Sept. 2000, pp. 133–135.

[7] P. Bauer, J. Abel, and T. Fingscheidt, "HMM-Based Artificial Bandwidth Extension Supported by Neural Networks," in *Proc. of International Workshop on Acoustic Signal Enhancement (IWAENC)*, Juan les Pins, France, Sept. 2014, pp. 1–5.

[8] C. Yagli, M. A. T. Turan, and E. Erzin, "Artificial Bandwidth Extension of Spectral Envelope Along a Viterbi Path," *Speech Communication*, vol. 55, pp. 111–118, Jan. 2013.

[9] 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.

[10] T. Schlien, F. Heese, M. Schäfer, C. Antweiler, and P. Vary, "Audiosignalverarbeitung für Videokonferenzsysteme," in *Proc. of Workshop Audiosignal- und Sprachverarbeitung; INFORMATIK 2013*, Koblenz, Germany, Sept. 2013, pp. 2987–3001.

[11] P. Bauer, C. Guillaumé, W. Tirry, and T. Fingscheidt, "On Speech Quality Assessment of Artificial Bandwidth Extension," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Florence, Italy, May 2014, pp. 6082–6086.

[12] H. Pulakka, V. Myllylä, A Rämö, and P. Alku, "Speech Quality Evaluation of Artificial Bandwidth Extension: Comparing Subjective Judgments and Instrumental Predictions," in *Proc. of Annual Conference of the International Speech Communication Association (Interspeech)*, Dresden, Germany, Sept. 2015, pp. 2583–2587.

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

[14] S. Möller, E. Kelaidi, F. Köster, N. Côté, P. Bauerand 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 (Interspeech)*, Lyon, France, Aug. 2013.

[15] H. Pulakka, L. Laaksonen, M. Vainio, J. Pohjalainen, and P. Alku, "Evaluation of an Artificial Speech Bandwidth Extension Method in Three Languages," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 6, pp. 1124–1137, Aug. 2008.

[16] Speech Ocean, http://www.speechocean.com.

[17] "EVS: Processing Functions for Characterization Phase, v1.0.0 (3GPP S4 141126, V. 1.0.0)," 3GPP; TSG SA, Aug. 2014.

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

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

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

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

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

[23] M. A. T. Turan and E. Erzin, "Synchronous Overlap and Add of Spectra for Enhancement of Excitation in Artificial Bandwidth Extension of Speech," in *Proc. of Annual Conference of the International Speech Communication Association (Interspeech)*, Dresden, Germany, Sept. 2015, pp. 2588–2592.

[24] A. H. Nour-Eldin and Peter Kabal, "Combining Frontend-Based Memory with MFCC Features for Bandwidth Extension of Narrowband Speech," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Taipei, Taiwan, Apr. 2014, pp. 4001–4004.

[25] A. H. Nour-Eldin and Peter Kabal, "Memory-based approximation of the Gaussian mixture model framework for bandwidth extension of narrowband speech," in *Proc. of Annual Conference of the International Speech Communication Association (Interspeech)*, Florence, Italy, Aug. 2011, pp. 1185–1188.

[26] P. Jax and P. Vary, "On Artificial Bandwidth Extension of Telephone Speech," *Signal Processing*, vol. 83, no. 8, pp. 1707–1719, Aug. 2003.

[27] 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 (EUSIPCO)*, Glasgow, Scotland, Aug. 2009, pp. 1839–1843.

[28] 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, pp. 36–39.

[29] Ching Y. Suen, *Computational Analysis of Mandarin*, Birkhäuser Basel, 1979.

[30] M.A. Mines, B.F. Hanson, and J.E. Shoup, "Frequency of Occurrence of Phonemes in Conversational English," *Language and Speech*, vol. 21, no. 3, pp. 221–241, July 1978.

[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.