4th International Workshop on Perceptual Quality of Systems (PQS 2013)
2-4 September 2013, Vienna, Austria

A Phonetic Reference Paradigm for Instrumental Speech Quality Assessment of Artificial Speech Bandwidth Extension

Tim Fingscheidt, Patrick Bauer

Institute for Communications Technology, Technische Universität Braunschweig
D–38106 Braunschweig, Germany
t.fingscheidt@tu-bs.de, patrick.bauer@tu-bs.de

Abstract

Today's instrumental speech quality measures are limited in their use as they "do not yet sufficiently include processing steps beyond the periphery of the auditory system" [1]. This becomes particularly obvious when using reference-based instrumental methods to assess the quality of artificial speech bandwidth extension (ABWE) approaches. While Blauert and Jekosch [1] have not proposed particular schemes, they advocate a model of sound quality representing layers of abstraction. In fact, once subjects are asked for opinion scores following any of ITU-T's definitions, they have already *understood* (or not) *what* was spoken. It is our firm conviction that in not-too-bad testing conditions this knowledge serves as internal reference for judging speech quality – which in consequence asks for a paradigm shift of reference-based instrumental speech quality measures. In consequence, not only some (direct wideband) reference speech data is useful, but also a phonetic transcription of the speech, serving as human-internal representation of what was spoken. The paper will give thoughts to support this thesis, along with a proof that not all sounds are equal, asking for a phoneme-specific processing of future reference-based instrumental speech quality assessment methods.
Index Terms: artificial speech bandwidth extension, instrumental listening speech quality assessment

1. Introduction

Instrumental speech quality assessment is an important means for cost-efficient and reproducible testing of either transmitted or enhanced speech signals. So-called *intrusive* or *reference-based* instrumental assessment methods use a reference signal (e.g., from before transmission), and a second signal whose quality is to be measured. Examples of such reference-based instrumental speech quality measures are PESQ (ITU-T P.862 [2]), Wideband-PESQ (ITU-T P.862.2 [3]), and the newer POLQA standard (ITU-T P.863 [4]). For an overview to current approaches the reader is referred to [5]. In quite some testing conditions such approaches show a reasonable correlation to absolute category rating (ACR) subjective listening test results. In other test conditions (background noise, artificially extended speech bandwidth) the use of these assessment methods is often limited.

In this paper we deal with instrumentally assessing the output speech of artificial bandwidth extension (ABWE) schemes, such as, e.g., [6–10]. The most important use case for ABWE is to convert narrowband speech (bandwidth $300...3400$ Hz, sampling rate 8 kHz) to wideband speech ($[50...]\,300...7000$ Hz, sampling rate 16 kHz). While a low-band extension to $50...300$ Hz solely aims at improving speech quality, a high-band extension to $3400...7000$ Hz also improves intelligibility, particularly for hearing-impaired persons [11]. Since this intelligibility improvement is important in communications, and since low-band extension is hard to achieve in a robust manner (i.e., in noisy conditions), high-band extension plays the more important role in literature and practice.

For ABWE in principle two reference signals could be identified: First the input narrowband (NB) speech signal, secondly the original wideband (WB) speech signal, which is, of course, unknown to the system. Using the NB signal as reference, the ABWE processed speech could be worse, similar, or better in quality as the input NB signal. Using the WB signal as reference, it is unlikely that ABWE processed speech could be judged better. Employing WB-PESQ [3] or POLQA [4] in their wideband mode for assessing ABWE speech, only the reference to the direct WB speech makes sense, since the reference speech is used to build up internal masking thresholds, etc. The test results, however, show insufficient correlation to respective subjective listening quality mean opinion scores (MOS) (see for example in [12]). Möller et al. may be the first to have investigated the applicability of these ITU-T speech quality assessment standards to a number of different ABWE schemes [12]. Beyond the insufficient correlation, the even more critical part is that P.862.2 as well as P.863 both do not correctly reflect the subjectively obtained rank order among the five investigated ABWE approaches.

In this paper we will present a novel paradigm of reference-based instrumental speech quality assessment,

10.21437/PQS.2013-6

using both a reference speech waveform as well as some reference phonetic labels. In Section 2 we start with some analysis of algorithmic approaches to ABWE, pointing out issues we can learn to assess the respective output speech. In Section 3 we present further arguments supporting the use of reference phonetics, along with the results of a small subjective listening experiment. The paper ends with some conclusions in Section 4.

2. What We Can Learn from ABWE Technology

First of all, it appears to be instructive to shade some light on state-of-the-art approaches to ABWE. We start reporting an important finding: Different to the majority of speech enhancement algorithms, and different to virtually all speech codecs (where PESQ and POLQA mainly have been designed for), ABWE approaches nowadays contain a whole lot of technology from the pattern recognition field, or at least from the advanced decision and estimation theory field. To give some examples, in [7] and [8] an artifical neural network is employed. In the work by Jax and Vary [6] a hidden Markov model (HMM) has been first applied to ABWE. Bauer and Fingscheidt were able to show improvements of the HMM approach by introducing phonetic transcriptions into the training process of the HMM and of the upper band cepstral representations reflecting the states [9]. Yağli et al. [10] use a Viterbi algorithm to identify the best sequence of upper band envelope states, and in [13] even a language model is employed (i.e., the occurence of words and word sequences), both approaches being very close to automatic speech recognition (ASR) methodologies. As a consequence please note that – unlike in speech enhancement and much more as in speech coding – some language dependency has been found in ABWE approaches [14, 15].

When thinking of reliable instrumental speech quality assessment of ABWE speech we therefore consider it useful to take ASR quality measures into the picture. Typically word error rates are measured, or accuracy - both measures requiring reference data. Unlike PESQ and POLQA, the reference data for ASR quality assessment is the ground truth of spoken words in an orthographic or a phonetic representation. An important consequence of the fact, that modern ABWE approaches contain a whole lot of ASR-related components, is that we propose to perform research towards a *multi*-reference-based instrumental speech quality assessment method that relies in both, a (WB) reference speech signal, along with its phonetic transcription. Compared to PESQ and POLQA, we expect the additional information about what was spoken to be crucial for achieving high correlation to subjective MOS values, as well as for predicting the correct rank order among a number of ABWE approaches.

A further issue is to be learned from the ABWE field: In ABWE the necessity to extend the speech bandwidth differs for particular sounds of speech. In Fig. 1(a) a spectrogram of the American English sentence "Those answers will be straight forward if you think them through carefully first" spoken by a male speaker is displayed. It is well known and obvious to see that the upper band (i.e., from $4...7$ kHz) spectral content is predominant for fricative sounds such as /s/, while vocals are typically well reprensented already by the lower band. Many state-of-the-art ABWE approaches, however, are known to suffer either from a lisping problem (the sharp /s/ is not represented well enough in the upper band), or from artifacts (there are too many sharp extensions in the upper band; an /f/ may sound then like an /s/...). Typically, ABWE approaches have to be parameterized to trade-off these two effects. Good ABWE approaches are often characterized by a better trade-off of this kind.

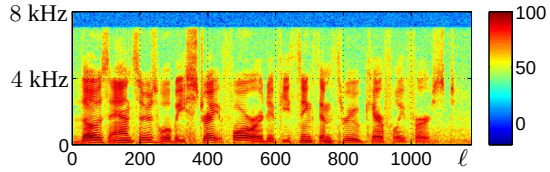
The fact that lisping or /s/-like artifacts play such a predominant role in ABWE research and development should lead us to the understanding, that *not all sounds are equal*(ly important). If not all sounds are equal, however, we should take profit from the knowledge of which sound was spoken, when designing a reference-based instrumental quality assessment method.

3. Perspectives for a Multi-Reference-Based Instrumental Speech Quality Assessment

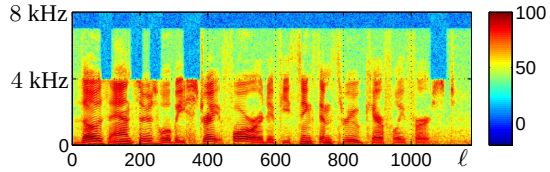
Apart from the experience we can gain from ABWE research, there are other hints towards a multi-reference-based quality assessment. One aspect is very simple: The way subjective (listening or conversational) speech quality tests are performed according to ITU-T Recommendations is that subjects give their votes *after* having listened to a sentence, or *after* having had a small conversation. Even if it is asked for *quality*, not for *intelligibility*, we can assume that knowing *what* was spoken influences the votes of subjects[1].

Given the fact that an ACR vote (in contrast to a CCR or DCR vote) is not explicitly comparing one speech sample with another, one could ask why most of the avaliable reference-based instrumental speech quality assessment methods aim at predicting an ACR MOS score. Humans do not have the reference speech signal available. In fact, however, "humans, through their experience, have acquired knowledge of normal and abnormal phenomena in speech sounds" [5]. What we would like to emphasize in this context is that these human-internal models of speech sound phenomena are sound–dependent, or one could say phone– or even phoneme–dependent. If these models are phoneme–dependent, then they play a very similar role as the acoustic models in automatic speech

[1] If the quality was so bad, that speech was (partly) unintelligible, this will of course influence the votes as well, but in that case without an internal *phonetic* reference. An internal reference in that case may be one of a lower layer.

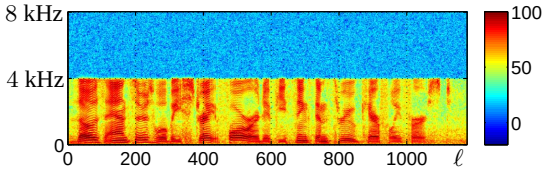


(a) Wideband speech

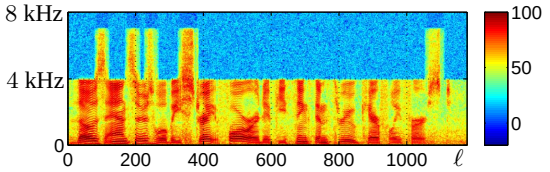


(b) Wideband speech with implanted narrowband /s/

Figure 1: Time-frequency plot of wideband speech signals; frame index ℓ, frame (DFT) size $K = 512$.



(a) Narrowband speech



(b) Narrowband speech with implanted wideband /s/

Figure 2: Time-frequency plot of upsampled narrowband speech signals; frame index ℓ, frame (DFT) size $K = 512$.

recognition. Any deviation from such an acoustic model leads to a lower Viterbi score in ASR, so does any deviation from "normal phenomena" in speech sounds in human speech perception.

But how does the human listener know to which of his phoneme-dependent acoustic models to compare? This brings us back to the beginning of this section: As long as the human listener is in the process of speech recognition, he indeed does not know to which models to compare. Still it may be the case that he uses these models already for speech recognition, in conjunction with a language model, maybe quite similar to how automatic speech recognition works. But all this is unproven. What is of interest here, is that the human listener at some point has recognized what was spoken, and afterwards he is being asked to judge quality. In that case we expect that he concatenates his phoneme-dependent models *according to what he recognized*, and performs a *rescoring of the representation in his acoustic memory*. Accordingly, a reference-based speech quality assessment method should also have access to what was said in order, e.g., to perform a similar concatenation of phoneme–dependent acoustic models.

This thesis is in-line with Blauert's and Jekosch's layer model of sound quality [1]: They even state that listeners only rarely react to what they *hear*, instead, they respond to what the "auditory events actually *mean*". Understanding meaning requires plain recognition of what was said as a prerequisite. For our proposed multi-reference paradigm we only ask for the knowledge of what was spoken.

A Small Phonetic Experiment

In order to provide further support for our paradigm of multi-reference instrumental speech quality assessment, we performed a simple phonetic experiment. Starting with some original wideband speech sample whose spectrogram is shown in Fig. 1(a), we at first removed the upper band of all five /s/ sounds ("Those answers will be straight forward if you think them through carefully first"), see Fig. 1(b). In Fig. 2(a) the respective narrowband signal is shown, while in Fig. 2(b) the complimentary operation has taken place: The original wideband /s/ sound has been implanted into the narrowband signal. Note that all four signals are sampled with 16 kHz. The signal in Fig. 1(b) can be seen as an ideal ABWE, however, with lisping artifacts (all five /s/ sounds sound like an /f/), while the signal in Fig. 2(b) is actually a narrowband signal enriched by a wideband /s/.

Figure	1 (a) WB	1 (b) WB with NB /s/	2 (a) NB	2 (b) NB with WB /s/
WB-PESQ	4.64	3.34	4.27	3.41
MOS	3.81	3.00	3.19	3.69

Table 1: Instrumental (WB-PESQ) and subjective (ACR test MOS) results for the four speech samples with spectrograms sketched in Figs. 1-2.

These four files have now been evaluated in a simple ACR-like informal listening test. The four files can be put into 24 orders, of which three were presented to one of eight subjects (diotic presentation using AKG headphones K-271 MkII). While each first set of four files has been used as warming up, the following two sets of four files have been scored, yielding 8 scores per person, 64 scores in total, and 16 scores per file. We used the 5-point MOS score definition for ACR listening tests [16, Annex B]. In addition, we used P.862.2 WB-PESQ [3] with the WB signal from Fig. 1(a) as a reference.

The results are displayed in Table 1. The pure wideband and narrowband signals receive WB-PESQ values of 4.64 and 4.27, respectively. In the subjective listening

test the MOS values are 3.81 and 3.19, respectively. We note – but do not overemphasize – the big absolute difference, which may be due to the fact that in this simple informal listening test we have not used any other signals serving as quality anchors (such as modulated noise reference unit (MNRU) signals). However, we observe that the quality difference between WB and NB is somewhat lower in the instrumental measurement. Much more interesting is the performance of the ideal (however, lisping) ABWE in Fig. 1(b). While human subjects rated it only about 0.2 MOS points worse than NB, WB-PESQ rated it more than 0.9 MOS points worse than NB. The most surprising result, however, is the narrowband signal with implanted WB /s/ sounds in Fig. 2(b): While the human subjects acknowledged the clearly improved quality and intelligibility by an *increase of 0.5 MOS points* vs. NB (just 0.12 MOS points below WB!), WB-PESQ judged this sample as being *more than 0.8 MOS points worse* than NB!

Recent ABWE approaches already follow the strategy to *extend the bandwidth only where necessary* – which is clearly the concept as shown in Fig. 2(b); while such behaviour seems to be clearly favourable vs. NB, instrumental measures such as WB-PESQ do not value such a design – in contrary: they punish it. Concerning WB-PESQ and measures alike, we assume that a major problem is the segmental kind of time integration of framewise disturbances in WB-PESQ and POLQA[2]: With our experience from ABWE development, and from this simple phonetic experiment we can state that a single lisping error is already a major quality problem, while producing wideband /s/ sounds is the most important goal of ABWE design.

The location of /s/ sounds in the speech material could be easily given to some multi-reference instrumental speech quality assessment method by a separate phonetic reference input. In order not to introduce language–specific solutions one could, e.g., adopt X-SAMPA as the phonetic format. Framewise disturbances could be easily weighted according to some phoneme-dependent costs, once they are integrated over time to yield the final quality estimate. We believe that with this new multi-reference paradigm a new family of instrumental assessment methods could be designed, which could also show adequate quality prediction capability in artificial bandwidth extension conditions – and beyond.

4. Conclusions

In this paper we have outlined a couple of reasons why a multi-reference instrumental speech qaulity assessment is a promising approach to be able to predict also the speech quality in artificial speech bandwidth extension conditions. Besides a reference speech waveform we advocate a phonetic transcription of the speech as further reference, since this follows the fact that human subjects also rate speech quality after having recognized what was spoken. The proposed paradigm is in full alignment of Blauert's multi-layer model of sound quality [1]; it even provides concrete perspectives towards a powerful new instrumental measure. Our thesis has been supported by a surprising phonetic experiment where state-of-the-art approaches totally fail.

[2]This time integration in PESQ and POLQA actually mainly follows an ℓ^2-norm, which is still too much a segmental measure.

5. References

[1] J. Blauert and U. Jekosch, "A Layer Model of Sound Quality," *J. Audio Eng. Soc*, vol. 60, no. 1/2, pp. 4–12, 2012. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=16160

[2] "ITU-T Recommendation P.862, Perceptual Evaluation of Speech Quality (PESQ)," ITU, Feb. 2001.

[3] "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.

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

[5] S. Möller, W.-Y. Chan, N. Côté, T. Falk, A. Raake, and M. Wältermann, "Speech Quality Estimation: Models and Trends," *IEEE Signal Processing Magazine*, vol. 28, no. 6, pp. 18–28, Nov. 2011.

[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] J. Kuntio, L. Laaksonen, and P. Alku, "Neural Network-Based Artificial Bandwidth Expansion of Speech," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 3, pp. 873–881, Mar. 2007.

[8] 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.

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

[10] C. Yağli, M. Turan, and E. Erzin, "Artificial Bandwidth Extension of Spectral Envelope Along a Viterbi Path," *Speech Communication*, vol. 55, no. 1, pp. 111–118, Jan. 2013.

[11] 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 ICASSP 2013*, Vancouver, BC, Canada, May 2013, pp. 7039–7043.

[12] 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 INTERSPEECH 2013*, Lyon, France, Aug. 2013.

[13] J. Han, G. Mysore, and B. Pardo, "Language Informed Bandwidth Expansion," in *Proc. of 2012 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, Santander, Spain, Sept. 2012, pp. 1–6.

[14] P. Bauer and T. Fingscheidt, "An HMM-Based Artificial Bandwidth Extension Evaluated by Cross-Language Training and Test," in *Proc. of ICASSP'08*, Las Vegas, Nevada, USA, Apr. 2008.

[15] H. Pulakka, L. Laaksonen, M. Vainio, J. Pohjalainen, and P. Alku, "Evaluation of 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] "ITU-T Recommendation P.800, Methods for Subjective Determination of Transmission Quality," ITU, Aug. 1996.