Sinusoidal-Based Lowband Synthesis for Artificial Speech Bandwidth Extension

Johannes Abel and Tim Fingscheidt, *Senior Member, IEEE*

***Abstract*—Conventional narrowband (NB) telephony suffers from limited acoustic bandwidth at the receiver side, leading to degraded speech quality and intelligibility. In this paper, artificial speech bandwidth extension (ABE) of NB speech toward missing frequencies below about 300 Hz (low-frequency band, LB) is proposed to enhance the speech quality. The LB-ABE in this paper is employed together with a preexisting ABE toward high-frequency components to obtain spectrally balanced speech signals. In an instrumental quality assessment, the spectral distance in the LB was improved by more than 5 dB compared to NB speech. In a subjective listening test, the gap of speech quality between wideband and NB speech was significantly reduced when employing the proposed ABE toward low frequencies. The LB extension was found to further improve the preexisting ABE toward higher frequencies by a significant 0.26 CMOS points.**

***Index Terms*—Artificial speech bandwidth extension, lowband, sinusoidal.**

I. Introduction

HUMAN speech production helps translating thoughts into speech. The message to be uttered is subject to phonological encoding, which subsequently triggers a chain of articulators on the physiological level for producing the sequence of sounds. Controlling the air flow arising from the lungs and thereby exciting the vocal cords, the glottis plays a key role in the speech production process. This muscle either periodically opens and closes with so-called fundamental frequency $F0$ or remains open while air passes through and thus creates a periodic or noisy excitation signal, respectively. In case of periodic excitation, additional resonant frequencies at integer multiples of the fundamental frequency emerge, which Zwicker and Fastl refer to as harmonic complex tones [1]. These harmonic complex tones or the aforementioned noisy excitation signal are then spectrally shaped by the vocal tract into vowels and consonants, where sounds based on harmonic excitation are called voiced sounds (e.g., /a/), and sounds with noisy excitation are called unvoiced sounds (e.g., /f/). The fundamental frequency (sometimes: pitch[1]) depends on the speaker's physiology (e.g., properties of the respective vocal tract), the speaker's psychological state (i.e., level of arousal), and the sound to be uttered. On average, men utter harmonic sounds with a fundamental frequency in the range of $50\ \text{Hz} \leq F0 \leq 250\ \text{Hz}$, while women produce harmonic excitation in the range of $120\ \text{Hz} \leq F0 \leq 500\ \text{Hz}$ [2]. Together with the fact that on average woman have a 2–3 cm shorter vocal tract and consequently formant frequencies are approx. 20% higher than those of men [3], [4], leads to the observation that (on average) men have deeper voices than women.

Traditional narrowband (NB) telephony limits the transmitted frequency components to a range of $300\ \text{Hz} \leq f \leq 3400\ \text{Hz}$ [5]. Obviously, for voiced sounds this leads to a degraded speech quality for most men and women with deep voice, since the first or even more of the harmonics are omitted. Remedy of this deficiency is achieved by using a wideband (WB) telephony call, where frequencies in the range of $100\ \text{Hz} \leq f \leq 7000\ \text{Hz}$ [5] are transmitted. In practice, however, successfully establishing a WB call depends on many circumstances: Both the end-user devices and the underlying infrastructure need to be WB-capable throughout. In addition, special arrangements need to be taken to enable WB calls from one operator to another. Whenever one of these factors is not given, only a NB call is possible, resulting in a call of minor speech quality. Even though more and more WB calls are established nowadays, customers are expected to cope again and again with bad telephone speech quality caused by lacking acoustic bandwidth still for many years. Artificial speech bandwidth extension (ABE) is a speech enhancement approach in the downlink path of a telephony call, improving the received NB speech signal by recovering missing frequency components. Therefore, ABE approaches can serve as a fallback solution to maintain a high and consistent speech quality whenever only a NB call is possible.

Most of the past ABE-related research efforts focused on estimating the upper band (UB), i.e., frequency components in the range $4\ \text{kHz} \leq f \leq 8\ \text{kHz}$ (or $\leq 7\ \text{kHz}$). These methods will be referred to as UB-ABE in this work. They mostly aim at improving speech quality, but some studies also reported an increased speech intelligibility [6]–[9]. The majority of

Manuscript received July 11, 2018; revised January 9, 2019; accepted January 22, 2019. Date of publication January 31, 2019; date of current version February 15, 2019. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Xiaodong Cui. *(Corresponding author: Tim Fingscheidt.)*

The authors are with the Institute for Communications Technology, Technische Universität Braunschweig, Braunschweig 38106, Germany (e-mail: j.abel@tu-bs.de; t.fingscheidt@tu-bs.de).

Digital Object Identifier 10.1109/TASLP.2019.2895969

[1]The term pitch refers to a metric describing the subjective impression how high or how low these harmonic complex tones are perceived by humans. In literature the term pitch is often used as synonym for fundamental frequency. At low frequencies, pitch and fundamental frequency of harmonic complex tones are almost identical [1], which might be one reason for the often imprecise usage of the term pitch.

UB-ABE approaches make use of the source-filter model for speech production by splitting the estimation task into two simpler estimation tasks, namely estimation of an UB excitation (a.k.a. residual) signal, and of an UB spectral envelope, independently. Most approaches obtain an UB residual signal by simple, yet effective, spectral folding [10], which is a modulation technique mirroring the existing NB residual into the UB. Due to the nature of this approach, the harmonic structure in the UB is destroyed. Investigations conducted in [11]–[13] revealed and confirmed that the harmonic structure in the UB is perceptually unimportant, thus justifying the use of spectral folding for UB-ABE approaches. The UB spectral envelope, on the other hand, is obtained by employing classifiers for estimating among quantized UB spectral envelopes, e.g., by nearest neighbor search in a pretrained NB codebook (codebook entries correspond to quantized UB spectral envelopes) [14], [15], Gaussian mixture models (GMMs) [16], [17], conventional neural networks [18], hidden Markov models (HMMs) with GMMs as acoustic model [19]–[23], or an HMM with feedforward deep neural networks (DNNs) as acoustic model [24]. Furthermore, DNNs have been employed for estimating the UB spectral envelope directly (regression) [25]–[28]. Opposed to the source-filter model, UB spectral magnitudes and UB phases can be estimated right away using sum-product networks (SPMs) [29], DNNs [30], [31], or recurrent neural networks (RNNs) [32], which can then be transformed back to the time domain by an overlap-add (OLA) structure. In several studies, an increased speech quality when using ABE solutions was shown [18], [33].

In addition to the extension towards high frequencies, lowband (LB) ABE (referred to as LB-ABE) can be performed, which extends the NB speech towards the low frequencies, i.e., $f \leq 300$ Hz. Note that missing LB frequencies are the usual case in NB calls, but may happen for tandeming or acoustical reasons also in WB calls. According to French and Steinberg [34], frequencies below 250 Hz do not contribute to the articulation index (AI), i.e., these frequencies play a negligible role w.r.t. speech intelligibility. Hence, the main objective of LB-ABE is speech quality enhancement. As seen for UB-ABE, many LB-ABE approaches also follow the source-filter model for speech production to obtain an LB speech component. In contrast to UB-ABE, the LB residual at voiced sounds needs to be as precise as possible in terms of a correct harmonic structure to prevent artifacts, such as shadow voices. Estimates of the LB residual signal are obtained, e.g., by non-linear processing of the NB residual. One approach is to convolve the NB residual with itself, and thus create harmonic components at the fundamental frequency and multiples of it [35]–[37]. In [38] a residual signal is taken from an internal NB representation of the code-excited linear prediction (CELP) decoder. Furthermore, a residual signal can be obtained by harmonic modeling [14]. Another approach leading to an estimated LB residual signal is to perform GMM-based classification among pretrained WB residual signals and subsequent lowpass filtering (known as modified data-driven voice source modeling) [39].

Spectral shaping of the LB residual signal can be achieved by simple amplification of the 100 Hz component by 10 dB [35] or via linear mapping [38]. More sophisticated approaches estimate the LB spectral envelope using nearest neighbor search (analog to UB spectral envelope estimation) [14], [36]. In [37] the LB spectral envelope is modeled by an autoregressive moving-average (ARMA)-filter. Employing statistical models, a GMM-based envelope estimation scheme was presented in [40] and furthermore an HMM estimating the temporal envelope and adapting the estimated residual signal by gain manipulation was presented in [39], implicitly also defining the resulting LB spectral envelope. Using a statistical model for regression, a conventional neural network for direct LB spectral envelope estimation was developed [36].

An LB residual can also be obtained in the frequency domain by constructing harmonic peaks and subsequent application of the inverse fast Fourier transform (IFFT) [40], [41]. Precise fundamental frequency estimators are of crucial importance for this approach, however, in [41] the fundamental frequency was taken from the clean WB signal. Apart from source-filter model approaches, the LB speech component can be completely estimated in the frequency domain using an HMM for magnitude estimation and a linear mapping function to obtain an LB phase estimate with subsequent IFFT [42].

During unvoiced sounds, the LB is perceptually unimportant [43], therefore an LB-ABE can be assumed not to be necessary for those sounds. The LB of voiced sounds, however, contains the perceptually important harmonics [43] and therefore the LB-ABE task can be simplified to recover only the missing harmonic peaks. A straightforward approach is to obtain an LB residual signal by first estimating the fundamental frequency from the NB signal and then synthesizing the missing harmonics with a sine generator in the time domain [43]–[47], which is referred to as sinusoidal synthesis. The main challenge of this approach is to estimate a precise fundamental frequency. Speech codec-specific approaches take an estimate of the fundamental frequency, e.g., from the CELP decoder [44], e.g., the adaptive multi-rate (AMR) decoder [45]. Being independent of the speech codec, some approaches employ separate fundamental frequency estimators [43], [46], [47]. The estimation of an envelope for the LB can be reduced to an estimation of the amplitude of the synthesized sinusoids, thus being a partial estimation of the LB spectral envelope. Estimating the harmonics' amplitudes may be done using a codebook mapping scheme [44], GMMs [45], neural networks [46], or can be as simple as taking the weighted amplitude of the first existing harmonic peak in the NB or taking the average of the existing harmonics in the NB [47]. Subjective listening tests for sinusoidal-based LB-ABE approaches were conducted in [43]–[45]. In the subjective listening test conducted in [43] it was found that speech signals processed by their LB-ABE approach contain more annoying distortions than the underlying NB signal with the LB missing. A preference test was conducted in [44], which revealed that in a direct comparison their LB-ABE approach mildly degrades the NB condition, while at the same time a slightly smaller speech quality gap between NB and WB was observed. The LB-ABE approach presented in [45] was able to slightly improve subjective speech quality of male speakers, however, slight additional artifacts were introduced into the resulting speech signal by LB-ABE, therefore eliminating the positive effect. The authors

in [46], [47] also reported artifacts in their respective LB-ABE approach.

Closely related to blind *estimation* of parameters for restoring lower or upper frequency components, sinusoidal-based analysis/synthesis methods exist, performing an explicit *calculation* of such parameters from a given speech signal (analysis) so that in a later step, speech synthesis can be performed to again obtain a speech signal based on the calculated parameters. Most prominently these approaches aim to compress speech signals for transmission (speech coding) in order to lower the bitrate. On the other hand, such analysis/synthesis methods provide the means to easily manipulate a speech signal at the parameter level, e.g., to change tonality of a given speech signal. Relevant research for the presented work was conducted McAulay and Quatieri [48] and Degottex and Erro [49], who investigated sinusoidal-based analysis/synthesis approaches.

In this work, we present a time-domain LB-ABE approach by sinusoidal synthesis of the missing harmonics for low-latency applications. Frame-wisely calculated amplitude and fundamental frequency estimates are processed by a sigmoid-based frame-to-sample interpolation function, which considers the human speech production w.r.t. the glottis' inertia and at the same time prevents discontinuities at the frame borders during synthesis. Prior work conducted in [43], [44], [47] employed linear techniques for conducting frame-to-sample interpolation, which bears the risk of producing discontinuities at the frame borders due to sudden parameter changes. While solutions presented in [44], [45] base their approaches on the fundamental frequency estimate from a speech decoder, we developed an LB-ABE approach which is independent of any speech coding employed in the transmission path. To obtain the fundamental frequency for synthesis, our scheme performs a low-latency estimation along with latency-free Viterbi-based tracking over time to ensure smooth and precise $F0$ estimates. In consequence, our approach is not restricted to be used with a specific hybrid speech codec such as the AMR codec [50], but can moreover be employed in a broader scope, e.g., on top of waveform speech codecs such as the wide-spread ITU-T G.711 [51] codec or even to enhance degraded (historical) speech recordings. To underline the capability of a broader use of the solution, we emulate practical employment by parameterizing and evaluating our LB-ABE approach on acoustically different speech databases. Our proposed approach introduces multiple means for error concealment of amplitude and fundamental frequency estimates. Regarding amplitude estimation, we refrain from employing non-linear statistical models such as GMMs [45] or NNs [46], in oder to keep full control of the system by using only a few hyperparameters. In prior work [43]–[45], the synthesized LB speech component suffered from artifacts, which unfortunately led to a degradation of the overall speech quality of the respective extended speech signals. Consequently, in this work, most efforts aim at minimizing the degrading effect of *each individual* estimation error. Therefore, we developed a supervision mechanism to guide each generated sinusoidal track from "birth" to "death". In detail, a novel state-based signal model for synthesizing the LB speech component has been developed, which prevents the propagation of estimation errors into the synthesized harmonic track(s) and parameterizes the employed frame-to-sample interpolation function according to estimated fundamental frequency, amplitude and the previous state. In consequence, the proposed system produces a robust and high-quality LB speech component.

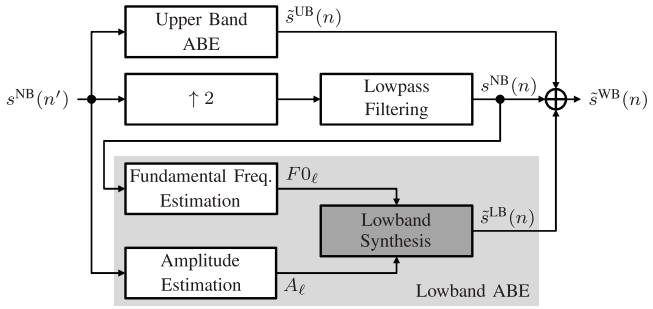

Fig. 1. Block diagram of the **ABE processing framework** including both an UB-ABE and an LB-ABE. Relevant processing steps for LB extension are found within the light gray-shaded box, here providing an LB signal estimate $\tilde{s}^{\mathrm{LB}}(n)$ at 16 kHz sampling rate with index n (8 kHz sampling rate has index n'). Sinusoids are generated during lowband synthesis, indicated as dark gray-shaded box.

The paper is structured as follows: Section II presents the algorithmic details of the LB-ABE approach, covering the main aspects, i.e., the employed fundamental frequency estimation, amplitude estimation, the signal model, and the sinusoidal synthesis. Following the experimental setup in Section III, results for instrumental and subjective assessment are presented and discussed in Section IV. Finally, conclusions are drawn in Section V.

II. Artificial Lowband Extension Framework

The ABE framework is depicted in Fig. 1. The system inputs the NB speech signal $s^{\mathrm{NB}}(n')$ and outputs the extended speech signal $\tilde{s}^{\mathrm{WB}}(n)$, with n' and n being the 8 and 16 kHz sample indices, respectively. In a framewise fashion, the LB-ABE approach estimates the fundamental frequency $F0_\ell$ and the amplitude A_ℓ of the missing harmonic components, with ℓ indexing frames of 5 ms speech. Subsequently, lowband synthesis is performed to generate sinusoids for each of the missing harmonic components below f_{max}, where the sum of all generated sinusoids constitutes the artificial LB speech signal $\tilde{s}^{\mathrm{LB}}(n)$. The LB-ABE outputs non-overlapping frames of 5 ms length, i.e., a frame length of $N = 80$ samples at a sample frequency of $f_s = 16$ kHz. Based on a parallel ABE structure, the output is calculated as superposition of the interpolated NB speech signal $s^{\mathrm{NB}}(n)$ (center), an estimated UB speech signal $\tilde{s}^{\mathrm{UB}}(n)$ (top), and the estimated LB speech signal $\tilde{s}^{\mathrm{LB}}(n)$ (bottom).

Subjective listening tests revealed that the existence of LB frequency components in speech contributes to an increased speech quality [52], however, it was found that the potential speech quality gain is even higher if the speech signal additionally contains higher frequency components, i.e., $f > 4$ kHz. Consequently, we assume the existence of an UB-ABE in the following. The interested reader may be referred to [28] for further information on our employed UB-ABE approach, but it could be any other as well.

A. Amplitude Estimation (A_ℓ)

Based on experiments conducted in [45], we simplify the amplitude estimation process by assuming that the harmonics below $f_h = 700$ Hz have an approximately identical amplitude. Justification of this somewhat ad-hoc assumption will be provided both by instrumental measures and subjective listening tests. The following efforts aim at finding the amplitude of the still existing harmonic frequencies below f_h.

The current NB speech frame $s_\ell^{\mathrm{NB}}(n'), n' \in \mathcal{N}'_\ell$ (with $\mathcal{N}'_\ell$ being the set of NB sample indices for frame ℓ) serves as input and contains 5 ms speech plus 5 ms look back. The resulting analysis window related to NB speech frame ℓ is of length $K' = 80$ samples. First, the discrete Fourier transform (DFT) is calculated following

$$S_\ell^{\mathrm{NB}}(k') = \sum_{n' \in \mathcal{N}'} s_\ell^{\mathrm{NB}}(n') \cdot e^{-j2\pi \frac{n'k'}{K'}}, \tag{1}$$

with $k' \in \mathcal{K}' = \{0, 1, \ldots, K'-1\}$. The frequency component having the highest amplitude below f_h is found at location

$$k'_{\max,\ell} = \underset{k' \in \mathcal{K}'_h}{\operatorname{argmax}} |S_\ell^{\mathrm{NB}}(k')|, \tag{2}$$

with $\mathcal{K}'_h \subset \mathcal{K}'$ being the set of DFT indices representing frequencies below f_h as being introduced before. The frequency component at bin $k'_{\max,\ell}$ is assumed to be among the lowest still received harmonics. The amplitude estimate for the missing LB harmonic(s) of the current frame ℓ is then calculated as

$$A_\ell = \alpha \left[\frac{1}{K'} \left| S_\ell^{\mathrm{NB}}(k'_{\max,\ell}) \right| \right] + (1-\alpha) A_{\ell-1}, \tag{3}$$

with $A_0 = 0$ and $\alpha = 0.35$ leading to a first-order IIR filter for smoothing the amplitude track. The fact that the amplitude of a harmonic track during voiced speech changes rather slowly over time has been exploited in the form of long-term prediction within speech coding schemes for decades [2]. Applied to LB-ABE, a slowly-changing amplitude estimate is the primarily expected outcome of the amplitude estimation. Besides simplicity, one of the main advantages of this approach is an increased robustness against outliers. At the same time, the estimate represents the amplitude characteristics of the still existing harmonic components and thereby automatically adapts the later generated sinusoidal to match the characteristics of the harmonic structure from the input signal which in consequence leads to linearity between input and output level of the proposed approach. The capability to adapt to the harmonics in the input signal while remaining robust against potential outliers is parameterized by α, which has been found by informal listening experiments.

B. Fundamental Frequency Estimation ($F0_\ell$)

Fundamental frequency estimation is limited to a frequency range of $50\,\mathrm{Hz} \leq F0_\ell \leq 650\,\mathrm{Hz}$, with $F0_\ell$ being the fundamental frequency estimate for the current frame ℓ. To overcome low-resolution problems, the interpolated and lowpass-filtered speech signal $s^{\mathrm{NB}}(n)$ at 16 kHz sampling rate is input to the fundamental frequency estimation. Furthermore, to capture frequencies as low as $f = 50$ Hz, at least two periods $2T = 2/f = 40$ ms have to be considered, thus the analysis frame length is chosen to be $N_w = 8N = 640$ samples (40 ms). Due to delay restrictions, the analysis frame has no lookahead, instead, it contains the past seven frames as lookback to collect 40 ms of speech. Consequently, in the presented framework, fundamental frequency estimation is performed in frames $\ell \geq 8$.

The fundamental frequency estimation is mainly based on the robust algorithm for pitch tracking (RAPT) [53], but has been modified to determine fundamental frequency candidates without introducing any extra algorithmic delay. We achieve this by considering past $F0$ estimates. In the following we will briefly sketch this modified algorithm until we reach the $F0$ estimate in (10). The normalized cross-correlation function (NCCF) coefficients [53]

$$\mathrm{NCCF}_\ell(\nu) = \frac{\sum_{n \in \mathcal{N}_\ell^{(\nu)}} |s^{\mathrm{NB}}(n)| \cdot |s^{\mathrm{NB}}(n+\nu)|}{\sqrt{e_\ell(0) \cdot e_\ell(\nu)}} \tag{4}$$

are calculated, with $\mathcal{N}_\ell^{(\nu)}$ comprising $\{\mathcal{N}_{\ell-7}, \mathcal{N}_{\ell-6}, \ldots, \mathcal{N}_\ell\}$, however, with the last ν samples dropped (resulting in $8N-\nu$ samples). The energy function $e_\ell(\nu)$ is defined as

$$e_\ell(\nu) = \sum_{n \in \overline{\mathcal{N}}_\ell^{(\nu)}} |s^{\mathrm{NB}}(n)|^2, \ \nu \in \{0, \ldots, 8N-1\}, \tag{5}$$

with $\overline{\mathcal{N}}_\ell^{(\nu)}$ comprising $\{\mathcal{N}_{\ell-7}, \mathcal{N}_{\ell-6}, \ldots, \mathcal{N}_\ell\}$, however, with the first ν samples dropped (resulting in $8N-\nu$ samples).

Peaks found in the NCCF coefficients at index $\nu_\ell(j)$ stand for time periods $T_\ell(j) = \nu_\ell(j)/f_s$ of strong harmonic components present in the analysis frame, with $j \in \mathcal{P}_\ell$ being the j-th peak and $\mathcal{P}_\ell$ being the set of indices of all peaks found in the current frame ℓ. From these time periods, potential fundamental frequency candidates can easily be calculated as $f_\ell(j) = T_\ell(j)^{-1}$. The corresponding magnitudes are denoted as observations $b_\ell(j) := \mathrm{NCCF}(\nu_\ell(j))$. Please note that since RAPT was defined for batch processing of speech files, we do not use dynamic programming in the following to determine a fundamental frequency track over the entire speech file. Instead, we calculate a Viterbi score to also consider the fundamental frequency estimates from the past frames, leading to a smooth fundamental frequency track. Viterbi decoding without any algorithmic delay (remember that the analysis frame for fundamental frequency estimation does not contain any lookahead) is performed by calculating the score

$$\delta_\ell(j) = \left[\max_{i \in \mathcal{P}_{\ell-1}} \delta_{\ell-1}(i) \cdot a_\ell(i,j) \right] \cdot b_\ell(j). \tag{6}$$

Due to the recursive calculation of the score, information from all past frames is exploited to calculate the current score. The transition probability is modeled as

$$\begin{aligned} a_\ell(i,j) &= \mathrm{P}\left(f_\ell(j) | f_{\ell-1}(i)\right) \\ &= \max\left(1 - \left| \frac{f_\ell(j) - f_{\ell-1}(i)}{\Delta f} \right|^\beta, 0\right). \end{aligned} \tag{7}$$

If the distance between the consecutive frequencies is larger than Δf, the particular transition is forbidden. For the first frame,

the Viterbi score is obtained following

$$\delta_1(j) = \pi(i) \cdot b_\ell(j), \tag{8}$$

with $\pi(i)$ being a GMM with two modes, modeling the initial distribution of fundamental frequency values of women and men. The values $\pi(i)$ and the parameters $\beta = 0.3679$ and $\Delta f = 285$ Hz have been found beforehand in an optimization phase.

Finally, the index of the most plausible peak location

$$j_\ell^* = \underset{j \in \mathcal{P}_\ell}{\operatorname{argmax}}\ \delta_\ell(j) \tag{9}$$

is found, leading to the fundamental frequency estimate

$$F0_\ell = f_\ell(j_\ell^*). \tag{10}$$

C. Lowband Synthesis

In the continuous-time domain, a sine wave with amplitude A and frequency f is described by

$$A \cdot \sin\left(2\pi f t + \varphi\right),$$

with φ being the phase shift and t being the continuous-time index. Time discretization is achieved by substituting $t = n/f_s$, and together with the estimated harmonic frequency $h \cdot F0_\ell,\ h \in \{1, 2\}$, and amplitude A_ℓ, a sinusoid restoring the h-th harmonic component can be generated following

$$A_\ell \cdot \sin\left(2\pi \cdot (h \cdot F0_\ell) \cdot \left(\frac{n}{f_s}\right) + \varphi_\ell^{(h)}\right), \tag{11}$$

with $\varphi_\ell^{(h)}$ now being the phase shift of the h-th harmonic.[2] However, this approach suffers from the following three shortcomings:

1) Fundamental frequency and amplitude estimates are provided for each frame of 5 ms. Even small changes of these estimates over time lead to undesired switching effects at the frame borders during an active LB extension.
2) Abruptly starting or stopping of the LB extension causes switching effects in the generated sinusoidal track.
3) Assuming the phase shift $\varphi_\ell^{(h)}$ is provided by an estimator, each estimation error will inevitably disturb synthesis of a continuous waveform and thereby produces artifacts.

To tackle the first problem, a sigmoid-shaped interpolation function $\sigma(n; \Lambda_{\ell-1}, \Lambda_\ell)$, interpolating between some entities $\Lambda_{\ell-1}$ and Λ_ℓ (being fundamental frequency or amplitude estimates), is introduced, providing seamless transitions on *sample* level for the respective transition from frame $\ell-1$ to ℓ. By means of this interpolation function, the fundamental frequency and amplitude on sample level are obtained following

$$F0_\ell(n) = \sigma(n; F0_{\ell-1}, F0_\ell), \tag{12}$$

$$A_\ell(n) = \sigma(n; A_{\ell-1}, A_\ell), \tag{13}$$

respectively. Both values, $F0_\ell(n)$ and $A_\ell(n)$, replace $F0_\ell$ and A_ℓ in (11), respectively, in order to generate smooth sinusoidal tracks. For more details on the frame-to-sample interpolation $\sigma(\cdot)$ see Appendix A.

[2]Given the preprocessing of speech data and the resulting lower cut-off frequency in this work, the maximum number of harmonics to be synthesized is set to two. In other environments, the proposed LB-ABE can be parameterized to synthesize either more or even less than two harmonics.

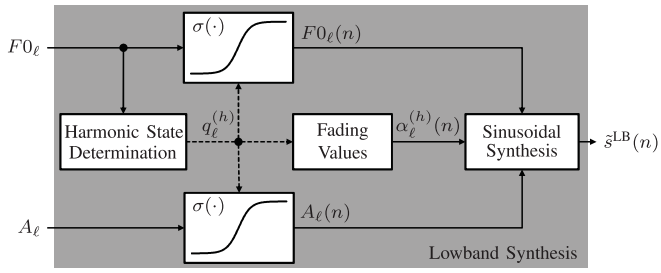

Fig. 2. Block diagram of **lowband synthesis** processing. First, the harmonic state $q_\ell^{(h)}$ is determined (see Appendix B) for each missing harmonic $h \in \{1, 2\}$ by means of the frame-wisely calculated fundamental frequency estimate $F0_\ell$. The harmonic state $q_\ell^{(h)}$ controls the sigmoid-shaped frame-to-sample interpolation function $\sigma(n; \cdot, \cdot)$ (see Appendix A) for amplitude estimate A_ℓ and fundamental frequency estimate $F0_\ell$. Fading values $\alpha_\ell^{(h)}(n)$ for carefully fading in and fading out the lowband component are also controlled by current state $q_\ell^{(h)}$ (see Appendix B). Finally, sinusoidal synthesis for generating $\tilde{s}_\ell^{\mathrm{LB}}(n)$ is performed, following (15).

The process of LB synthesis is controlled via a state machine,[3] which models the temporal process of fade-in and fade-out of the artificially generated sinusoidal tracks. For every frame and missing harmonic, indexed by h, the harmonic state $q_\ell^{(h)}$ is determined, as shown in Fig. 2. Based on the harmonic state $q_\ell^{(h)}$, the interpolation functions $\sigma(\cdot)$ in (12) and (13) are parameterized. A detailed description of the state machine can be found in Appendix B and Table V.

To solve the second issue, a fading strategy is implemented to overcome artifacts caused by switching on or off the LB-ABE by introducing fading values $\alpha_\ell^{(h)}(n)$ on sample level. The fading factors are also determined using the harmonic state machine as outlined in Appendix B, Table V. The fading factors are applied to (11), preventing sinusoids with high amplitudes when the sinusoidal synthesis starts, as well as preventing abruptly ending sinusoids with high amplitudes when the sinusoidal synthesis ends.

Regarding the third problem, previous research found that a high-fidelity temporal continuity of the phase is more important than trying to estimate the true phase of the missing harmonic(s) on a frame level (see, e.g., [45], [46], [54]). The latter approach would then produce phase estimation errors, which might have phase discontinuities at the frame borders (see also [31]). Therefore, we recursively calculate the phase shift in (11), serving as phase offset for continuous synthesis of the sinusoidal track. By means of the past fundamental frequency estimate $F0_{\ell-1}$ from (10), the sine's phase shift valid one instant *after* the previous frame $\ell - 1$ ended (last sample index from past frame is $n = N - 1$) for the next virtual sample (which could be indexed by $n = N$ for the moment), is obtained following

$$\varphi_{\ell-1}^{(h)} = 2\pi \cdot (h \cdot F0_{\ell-1}) \cdot \left(\frac{N}{f_s}\right) + \varphi_{\ell-2}^{(h)}. \tag{14}$$

[3]Please note that the states from this state machine are not related to the Viterbi states $j \in \mathcal{P}_\ell$, described in Sec. II-B.

Doing so results in a perfectly controlled continuous phase progression over time. For the first frame, this additional phase term is set to zero.

The artificial LB speech component to be added in Fig. 1 is finally obtained by means of sinusoidal synthesis (using fading values $\alpha_\ell^{(h)}(n)$, as well as (12), (13), and (14))

$$\tilde{s}_\ell^{\mathrm{LB}}(n) = \sum_{h\in\{1,2\}} \alpha_\ell^{(h)}(n)\cdot A_\ell(n) \cdot \sin\left(2\pi\cdot(h\cdot F0_\ell(n))\cdot\left(\frac{n}{f_s}\right)+\varphi_{\ell-1}^{(h)}\right), \quad (15)$$

as sum of the missing harmonics and thus concluding the LB-ABE presented in Fig. 2.

Estimation errors are unavoidable since RAPT (or *any* other fundamental frequency estimators) has limited capabilities, especially if the input signal suffers from any kind of distortion, such as speech coding or the application of filter masks to meet the requirements of speech transmission systems. To further minimize the influence of potential outliers, the moving average $\overline{F0}$ over the the past 3 seconds of fundamental frequency estimates is calculated. The sinusoidal synthesis is only active, if $\overline{F0} \leq f_{\max} = 300$ Hz. Furthermore, sinusoidal synthesis is only active, if the fundamental frequency estimate $F0_\ell$ is located in the range $\overline{F0} \pm \sigma$, with σ being the standard deviation of the averaged fundamental frequency estimates.

III. Experimental Setup

A. Speech Databases

In order to obtain parameter α for amplitude estimation (c.f. Sec. II-A), values $\pi(i)$ and parameters β and Δ for fundamental frequency estimation (c.f. Sec. II-B), speech data with annotated fundamental frequency values has been taken from the PTDB-TUG database [55], while the final instrumental evaluation uses speech data from the German and American English part of the NTT database [56]. Four female and four male speakers are included in each of the languages, providing 12 sentences per speaker. One sentence from each speaker has been used to obtain fundamental frequency estimates to calculate the initial moving average $\overline{F0}$ as described at the end of Section II-C and consequently were removed from the data set for instrumental and subjective assessment. Speech data in the PTDB-TUG database and the NTT database is sampled at 16 kHz.

B. Preprocessing

Speech data used in this work has been preprocessed to simulate transmission and device characteristics of a NB and WB phone call by following largely the preprocessing scheme presented in [28], [57]. Any delay introduced in the following preprocessing scheme was compensated for before applying instrumental measures. The speech data stemming from the speech databases is clean, has not been subject to speech coding, and thus is referred to as **WB direct**. Besides serving as reference signal for instrumental quality assessment, all further conditions are calculated from **WB direct** speech data.

For simulating a NB phone call, we follow Pulakka's preprocessing scheme [45]: First, the **WB direct** condition is highpass-filtered using a modified mobile station input (MSIN) filter, having the same magnitude response, however, is shifted along the frequency axis to have the -3dB point at $\approx$295 Hz (instead of $\approx$195 Hz). Subsequently, a second highpass filter is applied to the speech signal, which eliminates frequency components below 181 Hz. While Pulakka implemented the second filter in the spectral domain, we employ an FIR filter in the time domain for more convenient integration into our preprocessing framework. The FIR filter is characterized by 80 dB attenuation at 180 Hz and a passband starting at 300 Hz with a ripple of 0.01 dB. Please note that these preprocessing steps are in accordance with the sending sensitivity mask for handsets and handset equipment standardized by 3GPP [58]. Consequently, the LB-ABE generates speech components up to $f_{\max} = 300$ Hz, with a potential 5 Hz overlap between speech content from the NB signal and the synthesized LB signal to prevent gaps in the harmonic tracks. Subsequently, the signal is subject to decimation to 8 kHz sampling rate, 16-to-13-bit conversion, adaptive multirate (AMR) coding at 12.2 kbps and immediate decoding [59], and again 16-to-13-bit conversion. The resulting NB condition is referred to as **AMR** and provides the input $s^{\mathrm{NB}}(n')$ to the ABE framework, shown in Fig. 1. In case the UB-ABE has been applied to **AMR** speech data, we will refer to the resulting data as **UB-ABE**. If additionally the proposed LB-ABE is used, we refer to the resulting data as **LBUB-ABE**.

To simulate a WB phone call, **WB direct** data is P.341-filtered [60] and then subject to speech coding and decoding by the AMR-WB codec at 12.65 kbps. The resulting condition is referred to as **AMR-WB**.

IV. Evaluation and Discussion

A. Instrumental Evaluation

For instrumentally evaluating the synthesized LB speech components, the log-spectral distortion (LSD) metric is employed, following [61]:

$$LSD_\ell = \sqrt{\frac{1}{|\mathcal{K}|}\sum\nolimits_{k\in\mathcal{K}}\left[20\log_{10}\left(\frac{|S_\ell(k)|}{|\widehat{S}_\ell(k)|}\right)\right]^2},$$

with $\mathcal{K}$ being the set of DFT bin indices taken into consideration for LSD calculation. The spectra S and $\widehat{S}$ denote the DFT spectrum of the reference and the degraded speech signal under test, respectively. We will report a WB-LSD, i.e., considering frequencies $f < 7000$ Hz, an LB-LSD metric considering only the LB speech components ($0 < f < 200$ Hz), and additionally, an UB-LSD for a frequency range of $4000 < f < 7000$ Hz.

The resulting LSD values are presented in Tab. I. Obviously, for the **AMR** condition, the LSD values are the largest. Employment of the **UB-ABE** improves WB-LSD and UB-LSD values by a significant -7.20 dB and -12.43 dB, respectively. On average, the artificially estimated UB speech component leads to an improvement of $-12.43\,\mathrm{dB}/(7000-4000)\,\mathrm{Hz}\cdot 100 = -0.41$ dB per 100 Hz. The LB-LSD remains

TABLE I
INSTRUMENTAL ASSESSMENT USING WB-LSD (UP TO 7 KHZ), UB-LSD (4 . . . 7 KHZ), AND LB-LSD (0 . . . 200 HZ). THE BEST SCHEME BASED ON **AMR** INPUT DATA IS PRINTED IN **BOLD FACE**

Condition	WB-LSD	UB-LSD	LB-LSD
AMR	19.31	26.79	21.64
UB-ABE	12.11	14.36	21.60
LBUB-ABE	**11.74**	**14.34**	**16.45**
AMR-WB	7.21	8.26	5.93

TABLE II
INSTRUMENTAL ASSESSMENT USING WB-PESQ (**AMR** WAS UPSAMPLED) AND NB-PESQ (**UB-ABE**, **LBUB-ABE**, AND **AMR-WB** CONDITIONS WERE DOWNSAMPLED). THE BEST SCHEME BASED ON **AMR** INPUT DATA IS PRINTED IN **BOLD FACE**

Condition	WB-PESQ	NB-PESQ
AMR	2.68	**4.00**
UB-ABE	2.69	**4.00**
LBUB-ABE	**2.71**	3.95
AMR-WB	3.53	4.07

practically unchanged.[4] Adding the proposed LB-ABE (approach **LBUB-ABE**) improves the **AMR** condition by -5.15 dB in LB-LSD, which corresponds to an improvement of $-5.15\,\text{dB}/(300-0)\,\text{Hz}\cdot 100 = -1.72$ dB per 100 Hz, thus exhibiting more than 4 times the restoring capabilities as compared to the **UB-ABE** approach.

Regarding instrumental speech quality assessment, measures such as NB-PESQ [62], WB-PESQ [63], POLQA [64], or QABE [65] cannot be used for the presented LB-ABE approach, since these measures have not been developed for LB-ABE approaches. Still for information, Tab. II reports WB-PESQ [63] and NB-PESQ [62] results, the latter being measured on lowpass-filtered speech downsampled to 8 kHz sampling rate. Regarding WB-PESQ, the positive effect of **UB-ABE** is hardly visible. The benefit of employing an LB-ABE is also only slightly reflected by WB-PESQ predictions, while for WB-coded data a clear advantage over **AMR** data is attested. We also evaluated the conditions using WB-PESQ with corrigendum 2 applied, i.e., with a bug fix presented in early 2018 [66], however, the same behavior and thus the same conclusions compared to WB-PESQ resulted. With respect to NB-PESQ, obviously no difference between **AMR** and the **UB-ABE** condition can be observed, since the signals have been lowpass-filtered. Opposed to WB-PESQ, NB-PESQ attests the **LBUB-ABE** condition a speech degradation compared to **AMR**, while **AMR-WB** performs best. These small differences of speech quality among the tested NB and ABE conditions as predicted by WB- and NB-PESQ do not represent the results from informal listening tests conducted in our labs. Due to these inconclusive results, some spectral analysis (Section IV-B), and a subjective assessment (Section IV-C) become necessary.

[4]Please note that the application of any kind of speech enhancement approach such as the presented ABE schemes may result in very small deviations in LSD measurement at frequency ranges, which were not objective of the respective speech enhancement technique.

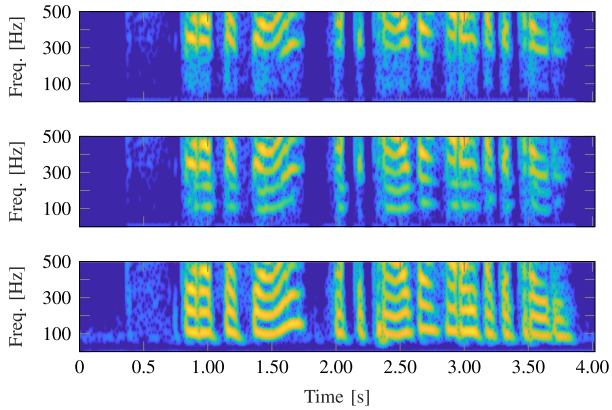

Fig. 3. Spectral analysis of a male speaker's speech signal (ge01m002) in the conditions **AMR** (top), **LBUB-ABE** (center), and **WB direct** (bottom), limited to approx. 4s of speech and a maximum frequency of 500 Hz.

B. Spectral Analysis

To visualize the effect of our proposed LB-ABE, Fig. 3 shows a spectral comparison of the conditions **AMR** (top), **LBUB-ABE** (center), and **WB direct** (bottom) for frequencies below 500 Hz of a speech signal from a male speaker. The top plot illustrates the effect of NB preprocessing: All harmonic components below 295 Hz, however, most certainly below 181 Hz (c.f. Section III-B), are removed during calculation of the **AMR** condition, which serves as input to the proposed LB-ABE approach. In this particular example, the first two harmonics were completely lost, while the third harmonic is partly removed at times where the third harmonic track fell below 300 Hz. The center plot shows the output of the proposed LB-ABE approach, i.e., the artificially generated sinusoids, which, compared to the spectrum in the bottom plot denoting the **WB direct** condition, restore most of the missing harmonics in the LB frequency range.

In another experiment, we deactivated the Viterbi algorithm presented in Section II-B and simply chose as $F0_\ell$ estimate the fundamental frequency candidate with the highest NCCF_ℓ. Doing so leads to a 0.07 dB increase of LB-LSD, indicating a slightly worse performance. In informal listening tests, however, we notice severe artifacts. For a more meaningful analysis, Fig. 4 shows the spectra of the **LBUB-ABE** approach, once without Viterbi (top) and once with Viterbi (center), and for comparison, also the **WB direct** spectrum (bottom). It can be observed, that without Viterbi more LB content is generated compared to the LB-ABE including the Viterbi. Between 0 and 0.25s, the non-Viterbi-based approach correctly detects a missing harmonic component and accordingly synthesizes a sinusoid (top plot, $\boxed{1}$). However, at 0.6s an estimation error occurs over several frames (top plot, $\boxed{2}$), destroying the harmonic structure and therefore causing a severe artifact. Opposed to the non-Viterbi approach, the proposed Viterbi-based approach prevents this misplaced sinusoid (center plot, $\boxed{2}$) and thereby an audible artifact, due to consideration of the past frames by means of the Viterbi score, described in (6). On the other hand, the Viterbi-based approach lacks the missing harmonic components in the

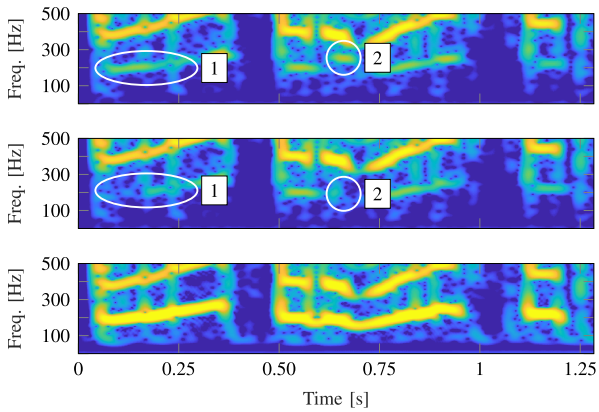

Fig. 4. Spectral analysis of a female speaker's speech signal (ge03f075) in the conditions **LBUB-ABE** w/o Viterbi (top), **LBUB-ABE** w/ Viterbi (center), and **WB direct** (bottom), limited to approx. 1.25s of speech and a maximum frequency of 500 Hz.

TABLE III
OVERVIEW OF MEAN, MINIMUM (MIN.), AND MAXIMUM (MAX.) FUNDAMENTAL FREQUENCY $F0$ OF EACH SPEAKER PROVIDING THE SIGNALS UNDER TEST FOR SUBJECTIVE ASSESSMENT OF SPEECH QUALITY. FUNDAMENTAL FREQUENCY STATISTICS MEASURED ON SPEECH DATA IN **WB DIRECT** CONDITION USING PRAAT [68]

Speaker	Mean $F0$	Min. $F0$	Max. $F0$
Female A (ge01f)	221 Hz	156 Hz	361 Hz
Female B (ge03f)	202 Hz	134 Hz	382 Hz
Male A (ge01m)	120 Hz	80 Hz	192 Hz
Male B (ge03m)	134 Hz	83 Hz	196 Hz

time between 0 and 0.25s (center plot, 1), revealing a trade-off between possible artifacts and correctly synthesized LB content. Reducing the impact of fundamental frequency estimation errors and thereby taking the risk of missing a harmonic component was found to be more important than synthesis of as much sinusoids as possible.

While the LSD difference is small, we observed a clear perceptual difference in the above described spectral analysis experiment. These inconsistent results together with the implausible WB- and NB-PESQ results found in Sec. IV-A, inevitably lead to the necessity of a subjective listening test, which is conducted in the next subsection.

C. Subjective Assessment

Subjective speech quality assessment is done via a semi-formal comparison category rating (CCR)[67, Annex E] test, in which two conditions A and B are compared to each other on the CMOS scale, i.e., -3 (B is much worse than A) up to 3 (B is much better than A) with integer steps in between. The signals under test stem from two female and two male speakers of the German part of the NTT database. Regarding the potential of the proposed LB-ABE in a subsequent subjective listening test, a rough idea can be obtained by analyzing the fundamental frequency of the speakers under test using Praat [68]. This is done on the basis of the **WB direct** data to obtain fundamental frequency tracks. Statistics calculated from these fundamental frequency contours are presented in Tab. III. For all speakers, the mean fundamental frequency is lower than $f_{\text{max}} = 300$ Hz, thus the speech signals provided by these speakers will mostly lack at least the first harmonic component and are suitable for processing with the proposed LB extension approach. A maximum $F0$ larger than f_{max} for female A and B indicates fade-outs of the LB-ABE approach during active periods. Both male speakers have a maximum fundamental frequency smaller than f_{max}, and thus synthesis of at least the first harmonic is potentially possible without fade-out. Regarding the LB-ABE approach, this set of speakers exhibits high potential for speech quality improvement but also for degradation due to estimation errors.

Four sentences are taken from each speaker and the conditions **AMR**, **UB-ABE**, **LBUB-ABE**, and **AMR-WB** are derived according to Sec. III-B, normalized to -26 dBov following ITU-T P.56 [69], and finally converted to 48 kHz sampling rate. Comparing all four conditions to each other leads to six CCR comparisons, denoted by ①–⑥. During a preliminary familiarization phase, 12 comparisons were scored by the test subjects to get acquainted to the CCR scale and to find a suitable playback level. In this familiarization phase, for which one of the four sentences from each speaker was used, all six CCR comparisons were presented. The three remaining sentences per speaker were then used in the main test, which is the basis for the reported results in the following. All possible comparisons (4 speakers x 3 sentences x 6 CCR comparisons $= 72$) were split into two listening test panels, while remaining balanced over conditions and speakers. Each comparison is tested in both directions (i.e., A vs. B *and* B vs. A), leading to 72 randomized comparisons in each of the two listening test panels. A total of 16 German native speakers without any known hearing impairment served as test subjects, which were given a service charge for their participation. The test subjects were equally assigned to one of the two listening test panels. The speech signals under test were presented to the test subjects via a conventional PC with external `RME Fireface 400` over `AKG K-271 MK II` headphones in diotic fashion. Subjects were allowed to repeatedly listen to the signals under test.

In Tab. IV, the speaker-dependent CMOS values with the respective 95% confidence intervals (CI95) for the CCR comparisons ①–⑥ are presented, including the overall means and confidence intervals.[5] Please note that on average, the LB extension is active in about 27% of the frames in all speech files under test, with speaker-dependent percentages being in the range from 24% to 29%.

Obviously, switching from **AMR** to **AMR-WB** is rewarded by the test subjects with the highest CMOS value of 2.42 over the entire subjective listening test (①). The two following CCR comparisons ② and ③ answer the question, how the gap as observed in ① shrinks by employing **UB-ABE** and further **LBUB-ABE**. On average, **UB-ABE** leads to a smaller gap of 2.03 CMOS points, while additionally enabling the proposed LB extension leads to an even smaller gap of 1.81 CMOS points, proving the effectiveness of the proposed approach. If, during a WB call, the speech codec has to be switched to an NB

[5]Please note that Tab. IV shows only half of the confidence interval width to simplify the presentation. The actual confidence interval can be calculated as follows: [CMOS − CI95, CMOS + CI95], by means of the CMOS and related CI95 values as presented in Tab. IV.

TABLE IV
SUBJECTIVE SPEECH QUALITY ASSESSMENT: RESULTS FROM A CCR TEST, EVALUATING THE **UB-ABE** BASELINE AND THE **LBUB-ABE** APPROACH VS. NB AND WB-CODED SPEECH SIGNALS. ON AVERAGE, CONDITION B (MARKED BY A *) ALWAYS TURNS OUT TO HAVE BETTER QUALITY THAN CONDITION A

CCR Comparison	①		②		③		④		⑤		⑥	
Condition A	**AMR**		**UB-ABE**		**LBUB-ABE**		**UB-ABE**		**AMR**		**AMR**	
Condition B	**AMR-WB***		**AMR-WB***		**AMR-WB***		**LBUB-ABE***		**UB-ABE***		**LBUB-ABE***	
Speaker	CMOS	CI95	CMOS	CI95	CMOS	CI95	CMOS	CI95	CMOS	CI95	CMOS	CI95
Female A	2.40	0.18	1.96	0.16	1.75	0.18	0.08	0.30	1.17	0.24	1.25	0.26
Female B	2.50	0.23	2.42	0.23	2.17	0.29	0.25	0.26	1.08	0.30	1.21	0.29
Male A	2.46	0.18	1.81	0.32	1.77	0.26	0.48	0.33	1.33	0.26	1.35	0.27
Male B	2.31	0.14	1.92	0.31	1.56	0.26	0.21	0.30	1.31	0.20	1.35	0.21
Average	2.42	0.09	2.03	0.14	1.81	0.13	0.26	0.15	1.22	0.13	1.29	0.13

speech codec (e.g., during a mobile cell handover procedure or if cell reception falls below a certain threshold), *the employment of* **LBUB-ABE***reduces the resulting speech quality drop from 2.42 CMOS points (observed in ①) to only 1.81 CMOS points (observed in ③), which accounts to about 25% less drop in speech quality*.

On the other hand, the speech quality increase of **UB-ABE** and **LBUB-ABE** compared to the **AMR** condition is investigated in ⑤ and ⑥. On average, both ABE conditions significantly improve the underlying **AMR** condition, with **LBUB-ABE** having a slightly higher CMOS mean. In the following, we will analyze the impact of LB extension on female and male speakers individually.

1) Female Speakers: The superiority of **LBUB-ABE** over **UB-ABE** for both female speakers becomes apparent when comparing CCR conditions ⑤ and ⑥. Test subjects noticed a difference whether or not the LB extension is applied, with the result that CMOS scores in ⑥ are less gender-dependent than in ⑤. This leads to the conclusion that *a combination* of both LB-ABE and UB-ABE is useful especially for female speakers. Both female speakers A and B take profit from LB extension as can be seen when comparing ② and ③, i.e., the additional employment of LB extension accounts for more than 0.2 CMOS points improvement. This observation may be explained by the fact, that the LB extension contributes to restoring the spectral balance of female speakers, who might suffer from an imbalanced spectral distribution after only applying UB-ABE. To understand this, we have to consider that female speakers in general have about 20% higher formant frequencies [3], [4], i.e., a spectral centroid on average at a higher frequency than men. Most of the formant frequencies are still present in the **AMR** condition, however, with a missing LB, a spectral imbalance towards high frequencies results, which consequently affects female speakers stronger than male speakers. Of course, UB-ABE improves speech quality [28], [33], however, it does not sufficiently restore spectral balance over sounds, especially for female speakers. In [12] it was already stated, that only the simultaneous extension towards high *and* low frequencies leads to the maximum improvement possible, rather than the exclusive use of only one of the techniques.

2) Male Speakers: Due to their respective low mean fundamental frequencies, for both male speakers most of the time two harmonics were synthesized (c.f. Tab. III). Therefore, it is no surprise that for these speakers the CCR comparison ③ attests an even higher-than-average capability to reduce the speech quality gap towards **AMR-WB** as observed in ①. The direct comparison of **UB-ABE** and **LBUB-ABE** in ④ reveals a speech quality gain of a very strong and even on individual basis significant 0.48 and a good 0.21 CMOS points for male A and male B, respectively. This is not surprising, as male A has the lowest mean fundamental frequency (c.f. Tab. III) over all of the four tested speakers and thus the highest potential for improvement by LB extension, followed by male B and female B. Obtaining the highest CMOS value of all speakers is therefore plausible and proves at the same time that the LB extension successfully synthesizes multiple tracks of missing harmonics, leading to an increased speech quality.

In summary, test condition ④ shows that each speaker, male and female, takes profit from an LB extension. *On average over speakers, LB extension improves the speech quality by a significant 0.26 CMOS points.*

V. CONCLUSIONS

In this work, a lowband (LB) artificial speech bandwidth extension (ABE) approach is presented, restoring missing harmonic components at frequencies below 300 Hz, which were omitted in telephone calls. The proposed time-domain approach is based on sinusoidal synthesis, while employing a sophisticated harmonic state machine, controlling the signal generation process, and therefore preventing annoying artifacts and enabling the synthesis of natural LB speech components. The proposed LB extension is shown to be useful both for female and male speakers, particularly when an ABE towards some upper band (UB) is already being used, since the perceptually important spectral balance of speech is then restored.

In a subjective listening test, the speech quality drop observed between wideband speech and narrowband (NB) speech could be reduced by 25%, if the NB signal is processed by the combination of LB and UB extension. Employment of the LB extension additionally improves the UB extension by a significant 0.26 CMOS points.

APPENDIX A
INTERPOLATION FUNCTION

As can be seen in Fig. 5, the interpolation function is based on a sigmoid function, having small slope at the frame borders and thus being able to compensate for large frequency or amplitude

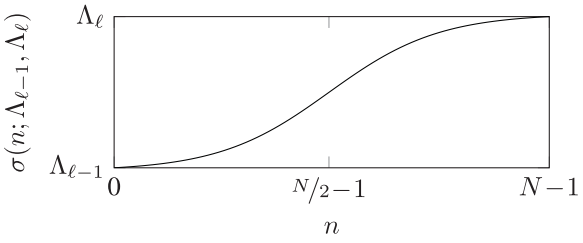

Fig. 5. Sigmoid-based interpolation function $\sigma(n;\Lambda_{\ell-1},\Lambda_\ell)$ to obtain smooth transitions on sample basis using frame-wise calculated entities $\Lambda_{\ell-1}$ (valid for $n=0$) and Λ_ℓ (valid for $n=N-1$).

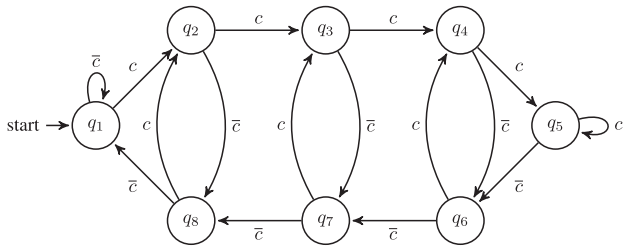

Fig. 6. State machine diagram for controlling sinusoidal synthesis. Depending on the current state $q_\ell^{(h)}=q_z$, with $z\in\{1,\ldots,8\}$, a different parameter set for sinusoidal synthesis is chosen (c.f. Tab. V for more details). State transitions are evaluated by means of a control variable c, which is either c (true) or $\bar{c}$ (false). After carefully fading in the sinusoids, the state machine will converge into q_5 ("on"), otherwise the state machine will remain in state q_1 ("off").

shifts over time. The general sigmoid function that we deploy is defined as

$$\sigma'(n)=\frac{1}{1+e^{-\alpha\left(n-\frac{N}{2}\right)}},\quad \text{with } 0\le n\le N-1,$$

with N being the frame length (i.e., 5 ms) and parameter $\alpha=0.1$. To adapt the sigmoid function to any desired start and end value, $\Lambda_{\ell-1}$ and Λ_ℓ, respectively, min-max scaling of $\sigma'(n)$ is performed following

$$\sigma(n;\Lambda_{\ell-1},\Lambda_\ell)=\frac{\sigma'(n)-\Lambda'_{\text{min}}}{\Lambda'_{\text{max}}-\Lambda'_{\text{min}}}\cdot(\Lambda_\ell-\Lambda_{\ell-1})+\Lambda_{\ell-1},\quad (16)$$

with Λ'_{max} and Λ'_{min} being the respective maximum and minimum value of $\sigma'(n)$ in $0\le n\le N-1$. The resulting interpolation function for general values $\Lambda_{\ell-1}$ and Λ_ℓ is shown in Fig. 5.

Please note that for descending transitions, i.e., $\Lambda_{\ell-1}>\Lambda_\ell$, $\sigma'(n)$ is multiplied by -1 before min-max scaling, which, however, for simplicity reasons is not explicitly indicated in this work.

APPENDIX B
HARMONIC STATE MACHINE

Considering once again the physiology of human speech production, the movement of the glottis is also influenced by its inertia, so that particularly a periodic excitation needs a short amount of time to start and also to come to an end. To model the glottis' inertia on the signal level, a state machine is defined, introducing intermediate steps on the path to switching on and off the LB extension process in (15) in a step-wise soft manner. The state flow diagram is shown in Fig. 6. The state machine executes transitions as shown in Fig. 6 by evaluating the control variable c, which is either $c=c_\ell^{(h)}=1$ (true) or $\bar{c}=c_\ell^{(h)}=0$ (false). State q_1 is the initial state and stands for a deactivated sinusoidal synthesis. Opposed to the initial state, state q_5 stands for a fully active sinusoidal synthesis. The activation states $q_2,\ldots,q_5$ use the amplitude and fundamental frequency estimates directly for sinusoidal synthesis, with amplitude and fundamental frequency transitions modeled via (16) on sample level. For each LB harmonic $h\in\{1,2\}$ a separate state machine is active, meaning that effectively two, one, or no LB harmonic may be generated. The control variable is updated, after fundamental frequency $F0_\ell$ from (10) has been estimated,

TABLE V
TABLE SHOWING DEFINITION OF AMPLITUDE $A_\ell(n)$, FREQUENCY $F0_\ell(n)$, AND FADING VALUES $\alpha_\ell^{(h)}(n)$ USED FOR SINUSOIDAL SYNTHESIS FOR THE DIFFERENT STATES $q_\ell^{(h)}=q_z$

State q_z	Amplitude $A_\ell(n)$	Frequency $F0_\ell(n)$	Fading $\alpha_\ell^{(h)}(n)$
q_1 (off)	$0\ \forall n$	$0\ \forall n$	$0\ \forall n$
q_2			$\sigma(n;0.00,0.25)$
q_3	$\sigma(n;A_{\ell-1},A_\ell)$	$\sigma(n;F0_{\ell-1},F0_\ell)$	$\sigma(n;0.25,0.80)$
q_4			$\sigma(n;0.80,1.00)$
q_5 (on)	$\sigma(n;A_{\ell-1},A_\ell)$	$\sigma(n;F0_{\ell-1},F0_\ell)$	$1.00\ \forall n$
q_6	$A^*\ \forall n$	$F0^*\ \forall n$	$\sigma(n;1.00,0.80)$
q_7	$A^*\ \forall n$	$F0^*\ \forall n$	$\sigma(n;0.80,0.25)$
q_8	$A^*\ \forall n$	$F0^*\ \forall n$	$\sigma(n;0.25,0.00)$

following

$$c_\ell^{(h)}=\begin{cases}1 & \text{if } 0\text{ Hz}<h\cdot F0_\ell<f_{\text{max}},\\ 0 & \text{else.}\end{cases}\quad (17)$$

Subsequently a state transition step is performed and the resulting state controls the computation of fundamental frequency $F0_\ell(n)$ and amplitude $A_\ell(n)$ according to (12) and (13), respectively, and of fading values $\alpha_\ell^{(h)}(n)$ for sinusoidal synthesis as defined in Tab. V.

During the transition from q_1 (LB-ABE off) to q_5 (LB-ABE on), the activation states q_2, q_3, and q_4 are passed, which control a fade-in function to carefully introduce the LB speech component into the resulting WB speech signal. Deactivation states q_6, q_7, and q_8 are responsible for carefully fading out the LB speech component. While fading out, the currently estimated amplitude and fundamental frequency are not trustworthy. Therefore, A^* and $F0^*$ denote the latest estimated amplitude and fundamental frequency value, respectively, for which $c_\ell^{(h)}=1$ held. Both values are updated in states q_2, q_3, q_4, q_5 and taken from memory for states q_6, q_7, q_8. Considering a state transition from q_3 to q_7, then A^* and $F0^*$ stand for the estimated amplitude and fundamental frequency values which were valid in state q_3 (i.e., values estimated for frame $\ell-1$). On the other hand, considering a state transition from q_6 to q_7, then A^* and $F0^*$ stand for the estimated amplitude and fundamental frequency value valid in state q_5 (i.e., values estimated for frame $\ell-2$ in this case).

If, during voiced speech, a fundamental frequency is estimated leading to $c_\ell^{(h)} = 0$, two things might have happened: The speech in the current frame is in fact not voiced anymore and the sinusoidal synthesis needs to fade-out, or the estimated fundamental frequency is likely to be wrong. Analog to this scenario, during unvoiced speech, a fundamental frequency estimate could lead to $c_\ell^{(h)} = 1$, meaning either that the current speech frame actually contains voiced speech and consequently that sinusoidal synthesis has to fade-in, or again that the estimated fundamental frequency is likely to be wrong. For all of these cases, artifacts due to falsely synthesized sinusoids have to be prevented, while fading in or out the generated sinusoids has to be performed as quickly as possible. To solve this, according to Tab. V, the first activating state q_2 (coming from q_1) only fades in towards 25%, while the first deactivating state q_6 (coming from q_5) only fades out towards 80%. If the control variable c in the next frame is then again calculated based on a correct fundamental frequency estimate, LB extension can easily recover from this error without introducing artifacts in the LB speech signal.

The fading values $\alpha_\ell^{(h)}(n)$ are chosen to allow seamless fading from state to state, as shown in Tab. V. Preliminary informal subjective listening tests revealed that fast fading in is of higher importance than fast fading out. Consequently, in Tab. V an asymmetrical fading strategy is implemented, i.e., fading in towards 0.25 vs. fading out only by $1.00 - 0.80 = 0.20$.

REFERENCES

[1] E. Zwicker and H. Fastl, *Psychoacoustics—Facts and Models*, 2nd ed. New York, NY, USA: Springer, Jan. 1999.

[2] P. Vary and R. Martin, *Digital Speech Transmission: Enhancement, Coding and Error Concealment*, 2nd ed. Hoboken, NJ, USA: Wiley, Mar. 2006.

[3] E. Eide and H. Gish, "A parametric approach to vocal tract length normalization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Atlanta, GA, USA, May 1996, pp. 346–348.

[4] M. Giurgiu and A. Kabir, "Comparison of vocal tract length normalization technique applied for clean and noisy speech," in *Proc. Int. Conf. Telecommun. Signal Process.*, Aug. 2011, pp. 351–354.

[5] *ITU-T Recommendation P.10, Vocabulary for Performance, Quality of Service and Quality of Experience*, ITU, Geneva, Switzerland, Nov. 2017.

[6] L. Laaksonen, H. Pulakka, V. Myllyla, and P. Alku, "Development, evaluation and implementation of an artificial bandwidth extension method of telephone speech in mobile terminal," *IEEE Trans. Consum. Electron.*, vol. 55, no. 2, pp. 780–787, May 2009.

[7] L. Laaksonen, J. Kontio, and P. Alku, "Artificial bandwidth expansion method to improve intelligibility and quality of AMR-coded narrowband speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Philadelphia, PA, USA, Mar. 2005, pp. 809–812.

[8] P. Bauer, J. Jones, and T. Fingscheidt, "Impact of hearing impairment on fricative intelligibility for artificially bandwidth-extended telephone speech in noise," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Vancouver, BC, Canada, May 2013, pp. 7039–7043.

[9] P. Bauer, M.-A. Jung, J. Qi, and T. Fingscheidt, "On improving speech intelligibility in automotive hands-free systems," in *Proc. IEEE Int. Symp. Consum. Electron.*, Braunschweig, Germany, Jun. 2010, pp. 1–5.

[10] J. Makhoul and M. Berouti, "High-frequency regeneration in speech coding systems," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Washington, DC, USA, Apr. 1979, pp. 428–431.

[11] H. Carl, "Untersuchung verschiedener Methoden der Sprachkodierung und eine Anwendung zur Bandbreitenvergrößerung von Schmalband-Sprachsignalen," Ph.D. dissertation, Ruhr-Universität Bochum, Bochum, Germany, 1994.

[12] P. Jax, "Enhancement of Bandlimited Speech Signals: Algorithms and Theoretical Bounds," Ph.D. dissertation, RWTH Aachen, Aachen, Germany, 2002.

[13] H. Pulakka, P. Alku, L. Laaksonen, and P. Valve, "The effect of highband harmonic structure in the artificial bandwidth expansion of telephone speech," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, Antwerp, Belgium, Aug. 2007, pp. 2497–2500.

[14] H. Carl and U. Heute, "Bandwidth enhancement of narrow-band speech signals," in *Proc. Euro. Signal Process. Conf.*, Edinburgh, U.K., Sep. 1994, pp. 1178–1181.

[15] T. Unno and A. McCree, "A robust narrowband to wideband extension system featuring enhanced codebook mapping," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2005, pp. 805–808.

[16] A. H. Nour-Eldin and P. Kabal, "Memory-based approximation of the Gaussian mixture model framework for bandwidth extension of narrow-band speech," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, Florence, Italy, Aug. 2011, pp. 1185–1188.

[17] Y. Wang, S. Zhao, Y. Yu, and J. Kuang, "Speech bandwidth extension based on GMM and clustering method," in *Proc. Commun. Syst. Netw. Technol.*, Gwalior, India, Apr. 2015, pp. 437–441.

[18] H. Pulakka and P. Alku, "Bandwidth extension of telephone speech using a neural network and a filter bank implementation for highband mel spectrum," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2170–2183, Sep. 2011.

[19] P. Jax and P. Vary, "Wideband extension of telephone speech using a hidden Markov model," in *Proc. IEEE Workshop on Speech Coding*, Delavan, WI, USA, Sep. 2000, pp. 133–135.

[20] P. Bauer, J. Abel, and T. Fingscheidt, "HMM-based artificial bandwidth extension supported by neural networks," in *Proc. Int. Workshop Acoust. Signal Enhancement*, Juan les Pins, France, Sep. 2014, pp. 1–5.

[21] C. Yagli, M. A. T. Turan, and E. Erzin, "Artificial bandwidth extension of spectral envelope along a viterbi path," *Speech Commun.*, vol. 55, pp. 111–118, Jan. 2013.

[22] 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. Annu. Conf. Int. Speech Commun. Assoc.*, Dresden, Germany, Sep. 2015, pp. 2588–2592.

[23] I. Katsir, D. Malah, and I. Cohen, "Evaluation of a speech bandwidth extension algorithm based on vocal tract shape estimation," in *Proc. Int. Workshop Acoust. Signal Enhancement*, Aachen, Germany, Sep. 2012, pp. 1–4.

[24] J. Abel, M. Strake, and T. Fingscheidt, "Artificial bandwidth extension using deep neural networks for spectral envelope estimation," in *Proc. Int. Workshop Acoust. Signal Enhancement*, Xi'an, China, Sep. 2016, pp. 1–5.

[25] Y. Li and S. Kang, "Artificial bandwidth extension using deep neural network-based spectral envelope estimation and enhanced excitation estimation," *IET Signal Process.*, vol. 10, no. 4, pp. 422–427, 2016.

[26] Y. Wang, S. Zhao, W. Liu, M. Li, and J. Kuang, "Speech bandwidth expansion based on deep neural networks," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, Dresden, Germany, Sep. 2015, pp. 2593–2597.

[27] J. Abel and T. Fingscheidt, "A DNN regression approach to speech enhancement by artificial bandwidth extension," in *Proc. IEEE Workshop Appl. Signal Process. to Audio Acoust.*, New Paltz, NY, USA, Oct. 2017, pp. 219–223.

[28] J. Abel and T. Fingscheidt, "Artificial speech bandwidth extension using deep neural networks for wideband spectral envelope estimation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 1, pp. 71–83, Jan. 2018.

[29] R. Peharz, G. Kapeller, P. Mowlaee, and F. Pernkopf, "Modeling speech with sum-product networks: Application to bandwidth extension," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Florence, Italy, May 2014, pp. 3699–3703.

[30] B. Liu, J. Tao, Z. Wen, Y. Li, and D. Bukhari, "A novel method of artificial bandwidth extension using deep architectures," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, Dresden, Germany, Sep. 2015, pp. 2598–2602.

[31] J. Abel, M. Strake, and T. Fingscheidt, "A simple cepstral domain DNN approach to artificial speech bandwidth extension," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Calgary, AB, Canada, Apr. 2018, pp. 5469–5473.

[32] Y. Gu, Z.-H. Ling, and L.-R. Dai, "Speech bandwidth extension using bottleneck features and deep recurrent neural networks," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, San Francisco, CA, USA, Sep. 2016, pp. 297–301.

[33] J. Abel *et al.*, "A subjective listenining test of six different artificial bandwidth extension approaches in English, Chinese, German, and Korean," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Shanghai, China, Mar. 2016, pp. 5915–5919.

[34] N. R. French and J. C. Steinberg, "Factors governing the intelligibility of speech sounds," *J. Acoust. Soc. Amer.*, vol. 19, no. 1, pp. 90–119, Jan. 1947.
[35] Y. Qian and P. Kabal, "Combining equalization and estimation for bandwidth extension of narrowband speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Montreal, QC, Canada, May 2004, pp. 713–716.
[36] B. Iser and G. Schmidt, "Neural networks versus codebooks in an application for bandwidth extension of speech signals," in *Proc. Euro. Conf. Speech Commun. Technol.*, Geneva, Switzerland, Sep. 2003, pp. 565–568.
[37] U. Kornagel, "Techniques for artificial bandwidth extension of telephone speech," *Signal Process.*, vol. 86, no. 6, pp. 1296–1306, Jun. 2006.
[38] C. Avendano, H. Hermansky, and E. A. Wan, "Beyond nyquist: Towards recovery of broad-bandwidth speech from narrow-bandwidth speech," in *Proc. Euro. Conf. Speech Commun. Technol.*, Madrid, Spain, Sep. 1995, pp. 165–168.
[39] M. Thomas, J. Gudnason, P. Naylor, B. Geiser, and P. Vary, "Voice source estimation for artificial bandwidth extension of telephone speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Dallas, TX, USA, Mar. 2010, pp. 4794–4797.
[40] I. Uysal, H. Sathyendra, and J. G. Harris, "Bandwidth extension of telephone speech using frame-based excitation and robust features," in *Proc. Euro. Signal Process. Conf.*, Antalya, Turkey, Sep. 2005, pp. 1–4.
[41] Y. Hu and P. C. Loizou, "Effects of introducing low-frequency harmonics in the perception of vocoded speech," *J. Acoust. Soc. Amer.*, vol. 128, no. 3, pp. 1280–1289, Sep. 2010.
[42] K. Kalgaonkar and M. Clements, "Sparse probabilistic state mapping and its application to speech bandwidth expansion," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Taipei, Taiwan, Apr. 2009, pp. 4005–4008.
[43] H. Gustafsson, U. Lindgren, and I. Claesson, "Low-complexity feature-mapped speech bandwidth extension," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 2, pp. 577–588, Feb. 2006.
[44] J. S. Park, M. Y. Choi, and H. S. Kim, "Low-band extension of CELP speech coder by harmonics recovery," in *Proc. Intell. Signal Process. Commun. Syst.*, Seoul, South Korea, Nov. 2004, pp. 147–150.
[45] 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 Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 8, pp. 2219–2231, Oct. 2012.
[46] J.-M. Valin and R. Lefebvre, "Bandwidth extension of narrowband speech for low bit-rate wideband coding," in *Proc. IEEE Workshop Speech Coding*, Delavan, WI, USA, Sep. 2000, pp. 130–132.
[47] G. Miet, A. Gerrits, and J. C. Valire, "Low-band extension of telephone-band speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Istanbul, Turkey, Jun. 2000, pp. 1851–1854.
[48] R. McAulay and T. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 34, no. 4, pp. 744–754, Aug. 1986.
[49] G. Degottex and D. Erro, "A uniform phase representation for the harmonic model in speech synthesis applications," *EURASIP J. Audio, Speech, Music Process.*, vol. 38, pp. 1–16, Oct. 2014.
[50] *Speech Codec Speech Processing Functions: AMR Wideband Speech Codec; Transcoding Functions*, 3GPP TS 26.190, Rel. 6, Dec. 2004.
[51] *Pulse Code Modulation (PCM) of Voice Frequencies*, ITU-T Rec. G.711, 1972.
[52] W. Krebber, "Sprachübertragungsqualität von Fernsprech-Handapparaten," Ph.D. dissertation, RWTH Aachen, Aachen, Germany, 1995.
[53] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," in *Speech Coding and Synthesis*. Amsterdam, The Netherlands: Elsevier, 1995, pp. 497–518.
[54] P. Mowlaee, J. Kulmer, J. Stahl, and F. Mayer, *Single Channel Phase-Aware Signal Processing in Speech Communication: Theory and Practice*. Hoboken, NJ, USA: Wiley, 2016.
[55] G. Pirker, M. Wohlmayr, S. Petrik, and F. Pernkopf, "A pitch tracking corpus with evaluation on multipitch tracking scenario," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, Florence, Italy, Aug. 2011, pp. 1509–1512.
[56] "Multi-Lingual Speech Database for Telephonometry," NTT Advanced Technology Corporation (NTT-AT), Tokyo, Japan, 1994.
[57] *EVS Permanent Document EVS-7c: Processing Functions for Characterization Phase*, 3GPP S4 141126, V. 1.0.0, Aug. 2014.
[58] *Universal Mobile Telecommunications System (UMTS); LTE; Terminal Acoustic Characteristics for Telephony; Requirements*, 3GPP TS 26.190, Rel. 14, Jul. 2017.
[59] *Mandatory Speech Codec Speech Processing Functions: AMR Speech Codec; Transcoding Functions*, 3GPP TS 26.090, Rel. 6, Dec. 2004.
[60] *ITU-T Recommendation G.191, Software Tool Library 2009 User's Manual*, ITU, Geneva, Switzerland, Nov. 2009.
[61] I. Katsir, I. Cohen, and D. Malah, "Speech bandwidth extension based on speech phonetic content and speaker vocal tract shape estimation," in *Proc. Euro. Signal Process. Conf.*, Barcelona, Spain, Aug. 2011, pp. 461–465.
[62] *ITU-T Recommendation P.862, Perceptual Evaluation of Speech Quality (PESQ): An Objective Methodfor End-to-End Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs*, ITU, Geneva, Switzerland, Feb. 2001.
[63] *ITU-T Recommendation P.862.2, Wideband Extension to Recommendation P.862 for the Assessment of Wideband Telephone Networks and Speech Codecs*, ITU, Geneva, Switzerland, Nov. 2007.
[64] *ITU-T Recommendation P.863, Perceptual Objective Listening Quality Assessment*, ITU, Geneva, Switzerland, Jan. 2011.
[65] J. Abel, M. Kaniewska, C. Guillaumé, W. Tirry, and T. Fingscheidt, "An instrumental quality measure for artificially bandwidth-extended speech signals," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 2, pp. 384–396, Feb. 2017.
[66] *ITU-T Recommendation P.862.2, Wideband Extension to Recommendation P.862 for the Assessment of Wideband Telephone Networks and Speech Codecs - Corrigendum 2*," ITU, Geneva, Switzerland, Mar. 2018.
[67] *ITU-T Recommendation P.800, Methods for Subjective Determination of Transmission Quality*, ITU-T, Geneva, Switzerland, Aug. 1996.
[68] P. Boersma and D. Weenink, Praat: Doing Phonetics by Computer. Mar. 2018. [Online]. Available: http://www.praat.org/
[69] *ITU-T Recommendation P.56, Objective Measurement of Active Speech Level*, ITU, Geneva, Switzerland, Dec. 2011.

Johannes Abel received the M.Sc. degree in computer and communications systems engineering from Technische Universitt Braunschweig, Braunschweig, Germany. During his studies, he worked as a Student Assistant in the field of speech enhancement and wrote his master thesis at the Institute for Communications Technology on artificial bandwidth extension for automatic speech recognition. In 2013, he started working toward the Ph.D. degree in the field of artificial bandwidth extension for telephony applications. His research interests include speech enhancement, machine learning, and automatic speech recognition.

Tim Fingscheidt (S'93–M'98–SM'04) received the Dipl.-Ing. degree in electrical engineering in 1993 and the Ph.D. degree in 1998 from RWTH Aachen University, Aachen, Germany. He further pursued his work on joint speech and channel coding as a consultant with the Speech Processing Software and Technology Research Department, AT&T Labs, Florham Park, NJ, USA. In 1999, he entered the Signal Processing Department, Siemens AG (COM Mobile Devices) in Munich, Germany, and contributed to speech codec standardization in ETSI, 3GPP, and ITU-T. In 2005, he joined Siemens Corporate Technology, Munich, Germany, leading the speech technology development activities in recognition, synthesis, and speaker verification. Since 2006, he has been a Full Professor with the Institute for Communications Technology, Technische Universität Braunschweig, Braunschweig, Germany. His research interests include speech and audio signal processing, enhancement, transmission, recognition, and instrumental quality measures. From 2008 to 2010, he was an Associate Editor for the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, and since 2011 he serves as a member of the IEEE Speech and Language Processing Technical Committee. He received several awards, among them are a prize of the Vodafone Mobile Communications Foundation in 1999 and the 2002 prize of the Information Technology branch of the Association of German Electrical Engineers (VDE ITG). In 2017, he co-authored the ITG award-winning publication, but the ITG prize is only awarded once in a life time. He has been a speaker of the Speech Acoustics Committee ITG AT3 since 2015.