EVALUATION OF DIFFERENT EXCITATION GENERATION ALGORITHMS FOR ARTIFICIAL BANDWIDTH EXTENSION

Jonas Sautter[1,2]*, Friedrich Faubel*[1]*, Markus Buck*[1]*, Gerhard Schmidt*[2]

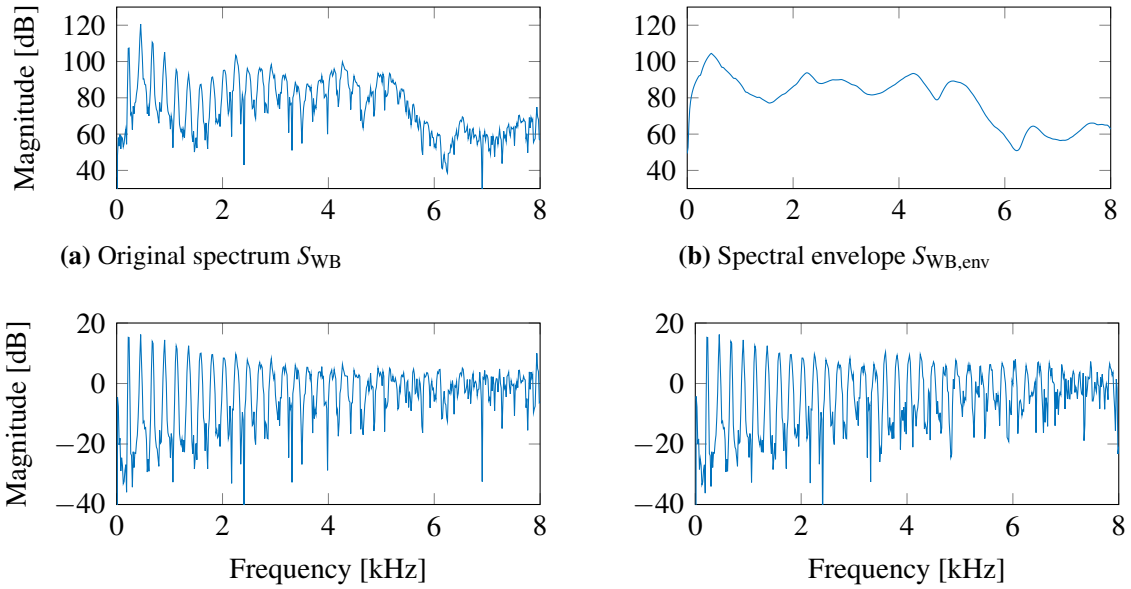
[1]*Acoustic Speech Enhancement Research, Nuance Communications Deutschland GmbH,*
[2]*Christian-Albrechts-Universität zu Kiel*
jonas.sautter@nuance.com

Abstract: Artificial bandwidth extension (ABWE) for speech signals is still an important topic in mobile telephony, especially when a 16 kHz wideband (WB) call suddenly falls back to an 8 kHz GSM connection. The aim of ABWE is to bridge the arising voice quality gap by reconstructing the WB signal. In order to achieve this, the speech signal is typically decomposed into a spectral envelope and an excitation signal, both of which are then extended separately. While the algorithms for envelope extension are getting increasingly more sophisticated, excitation generation is still often performed with rudimentary methods such as spectral folding (SF) or spectral shifting (SS). But this can introduce audible artifacts, especially for speech signals where the pitch frequency varies a lot. To reduce these artifacts, we introduce an algorithm that shifts parts of the spectrum multiple times by a smaller frequency shift. Additionally, we investigate if the speech quality can be further improved by interpolating the extended excitation signal with white noise. This is motivated by the fact that the SNR of the harmonic excitation decreases towards higher frequencies for real WB signals. The performance of the proposed algorithm is evaluated and compared to spectral folding and spectral shifting.

1 Introduction

Wideband calls with a bandwidth of about 50 Hz to 7 kHz are today available in most urban areas. However, there are still remote areas, e.g., in the countryside, where the mobile telephony network only supports GSM narrowband (NB) calls (about 300 Hz to 3.5 kHz). While moving to these areas, the bandwidth of the call is reduced and the speech quality gets worse. The aim of artificial bandwidth extension (ABWE) is to reduce the speech quality degradation by extending the NB signal to wideband (WB). Most approaches are based on separating the speech signal into its excitation and its spectral envelope, following the source-filter model of speech generation [1–6]. The decomposition of a WB spectrum is shown as an example in figs. 1a to 1c. Subsequently, the excitation and the spectral envelope can be extended separately, which reduces the complexity of ABWE. Most excitation extension methods can be classified into the following groups: spectral folding (SF) and spectral shifting (SS) [7], non-linear characteristics, and function generators [8]. In most of the current ABWE approaches, SF or SS are used [6].

In this paper, we introduce a new algorithm for the extension of the NB excitation signal. This is motivated by the fact that SF and SS can lead to audible artifacts [7], which can mostly be ascribed to the wrong continuation of the harmonic excitation above 4 kHz. In a voiced speech spectrogram, the change of the pitch frequency, i.e., the first harmonic, between two adjacent time frames $f_{\Delta,1}$ must be multiplied by n to get the delta frequency of the n-th harmonic $f_{\Delta,n}$. Consequently, with SS, the curve that is described by the upper harmonics directly above 4 kHz is too flat. This can sometimes be heard as a ringing tone. This effect can especially be observed



(a) Original spectrum S_{WB}

(b) Spectral envelope $S_{\mathrm{WB,env}}$

(c) Excitation spectrum $S_{\mathrm{WB,exc}}$

(d) Extended excitation spectrum with MSSCN $S_{\mathrm{MSSCN,exc}}$

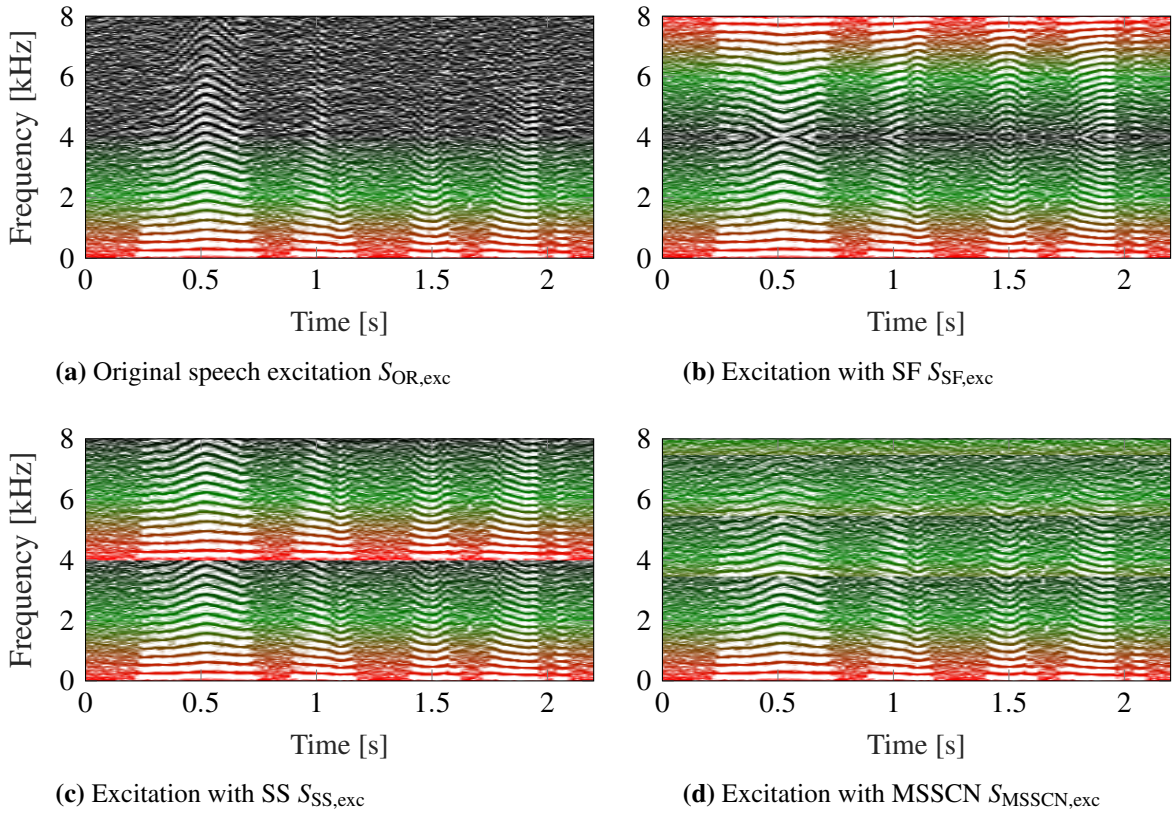
Figure 1 – Spectrum of the vowel 'a' in 'favor'. The original spectrum S_{WB} (a) is decomposed into the spectral envelope $S_{\mathrm{WB,env}}$ (b) and the excitation spectrum $S_{\mathrm{WB,exc}} = S_{\mathrm{WB}}/S_{\mathrm{WB,env}}$ (c). The excitation spectrum with MSSCN is depicted in (d).

in speech signals with a highly varying pitch frequency over time, e.g., in fig. 2c between 0.2 and 0.7 seconds. With SF, the same effect occurs for the harmonics directly below 8 kHz (see fig. 2b between 0.2 and 0.7 seconds). To reduce these effects, this paper proposes to just use the upper part of the NB spectrum for spectral shifting. In this part of the spectrum, the curve described by the harmonics is not as flat as in the lower frequencies. However, the shifting needs to be done multiple times in order to fill the WB excitation spectrum with a repetition of the selected frequency range. Consequently, we call this method multiple spectral shifting (MSS).

Another effect that can lead to artifacts is a mismatch in the harmonics to noise ratio (HNR). For all voiced phonemes in natural human speech, the harmonic structure of the excitation is getting increasingly weaker in higher frequencies, as it can, e.g., be observed in the excitation spectrum that is shown in fig. 1c. This stands in contrast to the excitation extension methods that have been presented so far, where the upper band spectrum is always filled with shifted or mirrored parts of the NB spectrum. As a consequence, the HNR is higher than for natural speech in the frequency region where the excitation spectrum is extended. The mismatch with respect to the HNR is reduced by MSS compared to SF and SS because the HNR in the frequency range between 1.5 and 3.5 kHz is already lower than it is for frequencies below 1.5 kHz. To further reduce this mismatch, comfort noise can be inserted above the frequency where excitation extension begins. To achieve a smooth transition, the MSS excitation spectrum is interpolated with comfort noise (MSSCN) along frequency (see fig. 1d).

To validate the superior performance of MSS, we conducted subjective listening tests that compare MSS to the often used excitation extension methods SF and SS as well as the original WB excitation (OR). Additionally, MSSCN was compared to MSS and OR in order to evaluate if an improvement can be achieved by comfort noise insertion. The spectral envelope was always taken from the WB signal to avoid an influence on the comparisons.

Section 2 deals with the different excitation extension methods. In section 3, the setup and the results of the subjective listening tests are described. A conclusion is given in section 4.



(a) Original speech excitation $S_{\mathrm{OR,exc}}$

(b) Excitation with SF $S_{\mathrm{SF,exc}}$

(c) Excitation with SS $S_{\mathrm{SS,exc}}$

(d) Excitation with MSSCN $S_{\mathrm{MSSCN,exc}}$

Figure 2 – Spectrogram of a speech excitation signal. The upper band excitation was generated using (a) the original WB excitation (OR), (b) spectral folding (SF), (c) spectral shifting (SS), and (d) multiple spectral shifting with comfort noise (MSSCN). The upper band starts at 4 kHz in (b) and (c) and at 3.5 kHz in (d). The colors indicate where the spectral parts in the upper frequency band originate from.

2 Extension of the Excitation Signal

The excitation short-term spectrum $S_{\mathrm{NB,exc}}(k,l)$ of the NB speech signal is the quotient of the short-term spectrum $S_{\mathrm{NB}}(k,l)$ of the NB speech signal and the corresponding spectral envelope $S_{\mathrm{NB,env}}(k,l)$:

$$S_{\mathrm{NB,exc}}(k,l) = \frac{S_{\mathrm{NB}}(k,l)}{S_{\mathrm{NB,env}}(k,l)} = \frac{S_{\mathrm{NB}}(k,l)}{\bar{S}_{\mathrm{NB}}(k,l)}. \tag{1}$$

Here, $\bar{S}_{\mathrm{NB}}$ denotes the short-term spectrum S_{NB} after spectral smoothing along frequency, k is the frequency index and l the time frame index. The time index l will be omitted in the following for reasons of readability. The envelope can, e.g., be extracted by linear prediction analysis (LPC) in the time domain (TD) or by spectral smoothing along frequency in the frequency domain (FD).

In this section, the different methods of excitation extension are described, starting with SF and SS. The newly proposed methods MSS and MSSCN are described based on SS. Example excitation spectrograms for OR, SF, SS, and MSSCN are given in fig. 2. In the following, the NB excitation signal $s_{\mathrm{NB,exc}}$ is sampled with a sampling rate of 16 kHz in accordance with the WB excitation signal $s_{\mathrm{WB,exc}}$ that shall be estimated.

2.1 Spectral Folding

The excitation extension method of spectral folding (SF) can be realized both in the TD or in the FD. In the TD, SF can be applied by sub-sampling with a factor of 2 and an amplification. This is equivalent to setting every second sample to zero and multiplying the remaining signal

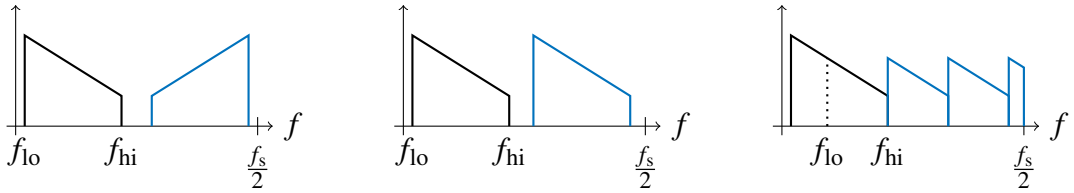


Figure 3 – Schematic spectra for the excitation extension methods (a) SF, (b) SS, and (c) MSS. The blue parts stand for the generated extension spectrum, the black parts for the original NB spectrum.

with a factor of 2. The aliasing components, which occur due to the sub-sampling, form the upper part of the spectrum and therewith create the extended excitation (see fig. 3a). This gives the following approximation of the WB excitation:

$$s_{\mathrm{SF,exc}}(m) = \begin{cases} 2 \cdot s_{\mathrm{NB,exc}}(m) & \text{for } m \text{ even} \\ 0 & \text{for } m \text{ odd} \end{cases} . \tag{2}$$

The sub-sampling is equivalent to a modulation with the modulation frequency $\omega_{0,\mathrm{SF}} = \pi$. SF can also be performed in the FD by mirroring the NB spectrum upwards at the horizontal axis with a frequency of 4 kHz. Additionally, the complex conjugate of the values has to be taken in order to preserve phase consistency. Subsequently, the WB excitation for the current frame can be written:

$$S_{\mathrm{SF,exc}}(k) = \begin{cases} S_{\mathrm{NB,exc}}(k) & \text{for } k \leq \frac{K}{2} \\ {S_{\mathrm{NB,exc}}}^*(K-k) & \text{for } k > \frac{K}{2} \end{cases} \qquad \forall k \in \{0,1,\ldots,K\} \tag{3}$$

where S^* denotes the complex conjugate of S, k is the frequency bin, $K = N_{\mathrm{FFT}}/2$ is the number of frequency bins and N_{FFT} denotes the length of the fast Fourier transform (FFT).

2.2 Spectral Shifting

Spectral shifting (SS) shifts the lower half of the spectrum to the upper half with a shifting frequency of $f_{0,\mathrm{SS}} = 4\,\mathrm{kHz}$ (see fig. 3b). It can also be achieved in the TD through modulation. The corresponding modulation frequency is $\omega_{0,\mathrm{SS}} = 2\pi f_0/f_s = \pi/2$. The aliasing terms have to be removed to avoid interference with the NB signal. This is achieved by applying a convolution with a high-pass filter h_{HP} with a cutoff frequency at $f_c = 4\,\mathrm{kHz}$. The estimated WB excitation can then be written as

$$s_{\mathrm{SS,exc}}(m) = s_{\mathrm{NB,exc}}(m) + s_{\mathrm{NB,exc}}(m) \cos\left(m\frac{\pi}{2}\right) * h_{\mathrm{HP}} . \tag{4}$$

The corresponding FD implementation is obtained by copying the NB spectrum to the upper frequency band ($f > 4\,\mathrm{kHz}$ or $k > K/2$):

$$S_{\mathrm{SS,exc}}(k) = \begin{cases} S_{\mathrm{NB,exc}}(k) & \text{for } k \leq \frac{K}{2} \\ S_{\mathrm{NB,exc}}(k - \frac{K}{2}) & \text{for } k > \frac{K}{2} \end{cases} \qquad \forall k \in \{0,1,\ldots,K\} \tag{5}$$

2.3 Multiple Spectral Shifting

As stated in section 1, just the upper part of the NB spectrum is shifted in MSS. After informal listening tests, we set this frequency range $[f_{\mathrm{lo}}, f_{\mathrm{hi}}]$ to $f_{\mathrm{lo}} = 1.5\,\mathrm{kHz}$ and $f_{\mathrm{hi}} = 3.5\,\mathrm{kHz}$. This frequency range is modulated multiple times to fill up the estimated WB spectrum (see fig. 3c).

The modulation frequency corresponds to the frequency range $f_{0,\mathrm{MSS}} = f_{\mathrm{hi}} - f_{\mathrm{lo}} = 2\,\mathrm{kHz}$ and has to be set to $\omega_{0,\mathrm{MSS}} = {}^{2\pi f_{0,\mathrm{MSS}}}/_{f_s} = {}^{\pi}/_{4}$. In the TD, only integral multiples of the bandpass-filtered part of the spectrum are modulated:

$$s_{\mathrm{MSS,exc}}(m) = s_{\mathrm{NB,exc}}(m) + (s_{\mathrm{NB,exc}}(m) * h_{\mathrm{BP}}) \cdot \left(\cos(m\frac{\pi}{4}) + \cos(2m\frac{\pi}{4})\right) * h_{\mathrm{HP}} \qquad (6)$$

where h_{BP} denotes a bandpass filter for the frequency range $[f_{\mathrm{lo}}, f_{\mathrm{hi}}]$. The corresponding FD implementation with the frequency range $[k_{\mathrm{lo}}, k_{\mathrm{hi}}]$ is

$$S_{\mathrm{MSS,exc}}(k) = \begin{cases} S_{\mathrm{NB,exc}}(k) & \text{for } k < k_{\mathrm{hi}} \\ S_{\mathrm{NB,exc}}(k_{\mathrm{lo}} + ((k - k_{\mathrm{hi}}) \bmod k_{\Delta})) & \text{for } k \geq k_{\mathrm{hi}} \end{cases} \qquad \forall k \in \{0, 1, \ldots, K\}, \quad (7)$$

with $k_{\Delta} = k_{\mathrm{hi}} - k_{\mathrm{lo}}$.

2.4 Multiple Spectral Shifting with Comfort Noise

For MSSCN, the extended spectrum $S_{\mathrm{MSS,exc}}$ is mixed with comfort noise in order to reduce the HNR in higher frequencies. This is achieved by interpolating with a generated complex white noise spectrum $W(k)$ using a frequency dependent weighting factor $\beta(k)$. The real and the imaginary part of $W(k)$ are normally distributed with $\mu = 0$ and $\sigma^2 = 1$. Note that the mean energy of the excitation is already normalized to about 0 dB (see fig. 1c). As excitation extension starts at k_{hi} in MSS, we insert comfort noise for $k > k_{\mathrm{hi}}$ only:

$$S_{\mathrm{MSSCN,exc}}(k) = S_{\mathrm{MSS,exc}}(k) \cdot \sqrt{1 - \beta(k)^2} + W(k) \cdot \beta(k) \qquad \forall k \in \{0, 1, \ldots, K\} \qquad (8)$$

where the weight β is calculated according to:

$$\beta(k) = \begin{cases} 0 & \text{for } k < k_{\mathrm{hi}} \\ \left(\frac{k - k_{\mathrm{hi}}}{K - k_{\mathrm{hi}}}\right)^{\alpha} & \text{for } k \geq k_{\mathrm{hi}} \end{cases}. \qquad (9)$$

The exponent α was set to 0.7 after informal listening tests.

3 Subjective Listening Tests

For the quality evaluation of ABWE algorithms, objective instrumental measures cannot fully replace subjective listening tests [9–11]. We expect that this also holds for excitation extension algorithms. Hence, the quality of the algorithms is here assessed in subjective listening tests rather than through objective measures. As a rating scale, we used the comparison category rating (CCR) CMOS scale, which is defined in ITU–T Rec. P.800 [12], because CCR methods offers a higher sensitivity than absolute category rating (ACR) methods if approximately equivalent conditions are rated [13]. Therein, relative, discrete rating choices from *much worse* to *much better* are mapped to 7 integer values from -3 to 3. Section 3.1 is about the data generation process. In section 3.2, the test setup is described, and section 3.3 deals with the test results.

3.1 Data Generation

For the evaluation, we randomly selected 100 utterances from the well-known TIMIT corpus [14], which comprises 630 American English speakers, each reading 10 phonetically rich sentences. 50 of these utterances were spoken by female and the other 50 by male speakers. All sound files were normalized to -5 dBFS. The spectral envelope and the original excitation were extracted from the true WB signal as described in section 2. For excitation extension methods, the corresponding NB signals were created by applying a bandpass filter to the true WB signals. The frequency response of this filter is shown in fig. 4.

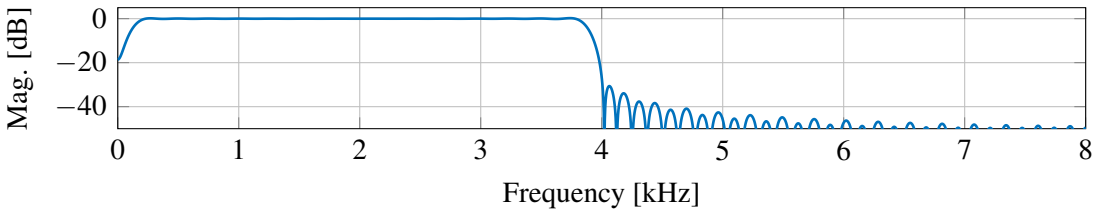


Figure 4 – Bandpass filter to generate the NB signals.

3.2 Test Setup

In the subjective evaluation, 36 listeners, who were between 23 and 51 years old, participated, 28 of which were male and 8 female. Each one of them rated the relative overall speech quality in 24 A-B comparisons. Every method in OR, SF, SS, and MSS was compared to all other methods twice. Additionally, MSSCN was compared to OR and MSS twice. This led to $4 \cdot 3 + 3 \cdot 2 = 18$ comparisons. Additionally, OR, SF, and SS were compared with themselves twice for reliability checks, which resulted in a total of $18 + 3 \cdot 2 = 24$ comparisons per listener. The utterances were chosen randomly from the generated dataset. The order of the A-B tests was randomized, as was the order of speech files. The listeners could play the sound files as often as they wanted. The volume of the playback was also set by the listeners. For reference, a short introduction was given before the test, where the listeners could listen to 10 A-B comparisons.

3.3 Results

This subsection presents the results of the listening tests. These are summarized in figs. 5a to 5f. The bars show the mean rating values of the direct A-B comparison. The length of the red error bars stands for the standard error, which is the standard deviation of the mean

$$\sigma_{\bar{x}} = \sqrt{\frac{{\sigma_x}^2}{N}} \tag{10}$$

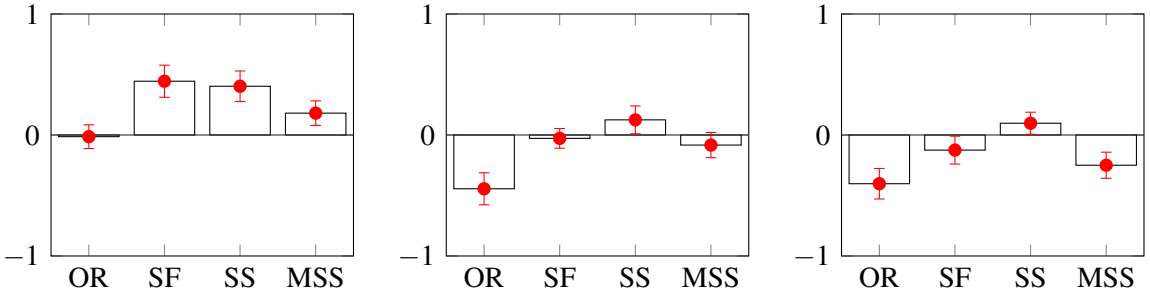
where ${\sigma_x}^2$ denotes the variance:

$${\sigma_x}^2 = \frac{1}{N-1} \sum_{i=0}^{N-1} |x_i - \mu|^2 . \tag{11}$$

In these equations, x is a vector of ratings, μ is the mean of x and N is the number of ratings or the length of the vector x. All absolute mean results are smaller than 1. This indicates that all differences were rated less than *slightly better* or *slightly worse* in the mean. In informal listening tests, we observed that differences between the excitation extension algorithms can only be heard on some files. This could be the reason for the low absolute ratings as the files are selected randomly. Figure 5a summarizes all direct comparisons between OR, SF, SS, or MSS and the original excitation. While SF and SS are rated worse than OR, MSS is much closer to OR. Regarding the error bars, just the differences between SF or SS and OR are statistically significant. In the direct comparisons between SF and SS methods, SF performs slightly better than SS. MSS is rated better than SS with statistical significance but just marginally better than SF. The comparison of MSS and MSSCN did not result in a reliable preference of the listeners.

As stated in section 3.2, we also conducted comparisons of OR, SF, and SS with themselves for reliability measures. To check the reliability of each listener, we created a measure called *mean difference voting ratio* (*MDVR*) that divides the mean absolute voting of all A-A comparisons $\bar{x}_{\mathrm{AA}}$ through the mean absolute voting of all A-B comparisons $\bar{x}_{\mathrm{AB}}$:

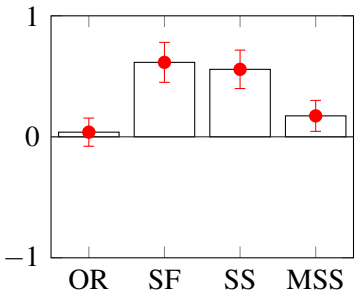
$$MDVR(x) = \frac{\bar{x}_{\mathrm{AA}}}{\bar{x}_{\mathrm{AB}}} = \frac{\frac{1}{N_{\mathrm{AA}}} \sum_{i=0}^{N_{\mathrm{AA}}-1} \left| x_{\mathrm{AA},i} \right|}{\frac{1}{N_{\mathrm{AB}}} \sum_{i=0}^{N_{\mathrm{AB}}-1} \left| x_{\mathrm{AB},i} \right|}. \tag{12}$$



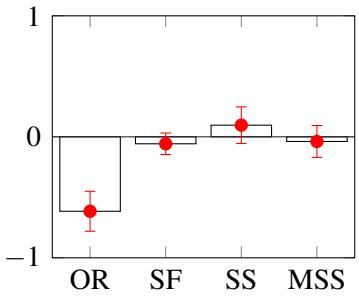
(a) Ratings for OR by all listeners

(b) Ratings for SF by all listeners

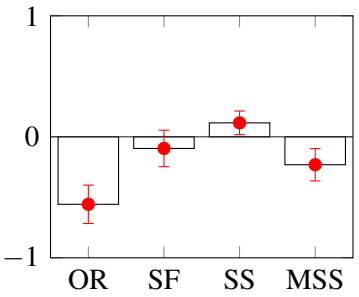
(c) Ratings for SS by all listeners



(d) Ratings for OR by reliable listeners only



(e) Ratings for SF by reliable listeners only



(f) Ratings for SS by reliable listeners only

Figure 5 – Mean ratings $\bar{x}$ (bars) and standard error $\sigma_{\bar{x}}$ (red error plots) of the subjective listening tests. (a) to (c) show the results for all listeners, (d) to (f) for the reliable listeners only. Positive values indicate that the listeners prefer OR in (a) and (d), SF in (b) and (e), or SS in (c) and (f) to the respective method on the x-axis. Note that the absolute value of 1 on the y-axis stands for *slightly better/worse* and that the rating range was $[-3,3]$ in the subjective test.

Here, N_{AA} denotes the length of the vector x_{AA} and N_{AB} denotes the length of the vector x_{AB}. For all listeners who heard more differences between equal sound files than between different sound files, $MDVR(x)$ becomes larger than 1. This was the case for 10 out of 36 listeners. These listeners were excluded in fig. 5d to fig. 5f as their answers cannot be taken to be reliable if they hear more differences between identical files than between different files. The high percentages of about 23% of the experts and 40% of the non-experts with $MDVR > 1$ indicate that most differences were at the border of audibility.

The results in fig. 5d to fig. 5f show comparable results to fig. 5a to fig. 5c. The biggest difference is that SF and SS were now rated significantly worse than the proposed method MSS, while the difference between OR and MSS was not statistically significant (see fig. 5d).

4 Conclusion

In this paper, we introduced a new method for excitation extension from NB (0.3-3.5 kHz) to WB (0.3-8 kHz), which aims at reducing the artifacts introduced by well-known methods such as SF and SS. We performed a CCR listening test to assess the quality of the different excitation extension methods. The spectral envelope was taken from the original WB signal. The mean ratings were all relatively close to 0, which stands for *about the same* speech quality. SF and SS were rated worse than the proposed method MSS, while the difference between MSS and OR was not statistically significant. MSS should be preferred to SF and SS because it can be implemented efficiently in FD and it can lead to better speech quality. The insertion of comfort noise did not result in statistically significant improvements of the speech quality.

References

[1] CARL, H. and U. HEUTE: *Bandwidth enhancement of narrow-band speech signals*. In *Proceedings EUSIPCO*, vol. 2. 1994.

[2] AVENDANO, C., H. HERMANSKY, and E. A. WAN: *Beyond Nyquist: towards the recovery of broad-bandwidth speech from narrow-bandwidth speech.* In *EUROSPEECH*. 1995.

[3] NAKATOH, Y., M. TSUSHIMA, and T. NORIMATSU: *Generation of broadband speech from narrowband speech using piecewise linear mapping.* In *EUROSPEECH*. 1997.

[4] ENBOM, N. and W. B. KLEIJN: *Bandwidth expansion of speech based on vector quantization of the mel frequency cepstral coefficients*. In *IEEE Workshop on Speech Coding Proceedings. Model, Coders, and Error Criteria*. 1999.

[5] JAX, P. and P. VARY: *Wideband extension of telephone speech using a hidden Markov model*. In *IEEE Workshop on Speech Coding. Proceedings. Meeting the Challenges of the New Millennium*. 2000.

[6] ABEL, J., M. KANIEWSKA, C. GUILLAUME, W. TIRRY, H. PULAKKA, V. MYLLYLÄ, J. SJÖBERG, P. ALKU, I. KATSIR, D. MALAH, I. COHEN, M. A. T. TURAN, E. ERZIN, T. SCHLIEN, P. VARY, A. NOUR-ELDIN, P. KABAL, and T. FINGSCHEIDT: *A subjective listening test of six different artificial bandwidth extension approaches in English, Chinese, German, and Korean*. In *ICASSP*. 2016.

[7] MAKHOUL, J. and M. BEROUTI: *High-frequency regeneration in speech coding systems*. In *ICASSP*, vol. 4. 1979.

[8] ISER, B., W. MINKER, and G. SCHMIDT: *Bandwidth extension of speech signals*. Springer Publishing Company, Incorporated, 1 edn., 2008.

[9] BAUER, P., C. GUILLAUMÉ, W. TIRRY, and T. FINGSCHEIDT: *On speech quality assessment of artificial bandwidth extension*. In *ICASSP*. 2014.

[10] PULAKKA, H., V. MYLLYLÄ, A. RÄMÖ, and P. ALKU: *Speech quality evaluation of artificial bandwidth extension: Comparing subjective judgments and instrumental predictions*. In *INTERSPEECH*. 2015.

[11] MÖLLER, S., 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 *INTERSPEECH*. 2013.

[12] *Methods for subjective determination of transmission quality*. ITU-T Recommendation P. 800, 1996.

[13] MÖLLER, S.: *Assessment and prediction of speech quality in telecommunications*. Springer Science & Business Media, 2012.

[14] GAROFOLO, J. S., L. F. LAMEL, W. M. FISHER, J. G. FISCUS, D. S. PALLETT, and N. L. DAHLGREN: *DARPA TIMIT acoustic phonetic continuous speech corpus*. Web Download. Philadelphia: Linguistic Data Consortium, 1993.