ARTIFICIAL BANDWIDTH EXTENSION USING THE CONSTANT Q TRANSFORM

Pramod B. Bachhav[1]*, Massimiliano Todisco*[1]*, Moctar Mossi*[2]
Christophe Beaugeant[2] *and Nicholas Evans*[1]

[1]EURECOM, France
lastname@eurecom.fr
[2] INTEL, Sophia Antipolis, France
firstname.lastname@intel.com

ABSTRACT

Most artificial bandwidth extension (ABE) algorithms are based on the classical source-filter model of speech production. This approach generally requires the dual extension of each component through independent processing. Alternative approaches reported recently operate on the spectrum. With human perception thought to be largely insensitive to phase, most such approaches focus on the extension of the magnitude spectrum alone and rely on Fourier spectral analysis. This paper reports an approach to ABE based on the constant Q transform (CQT), a more perceptually motivated approach to spectral analysis. A Gaussian mixture model is used to estimate missing highband components from available narrowband components before resynthesis with phase estimates obtained from the upsampled narrowband signal. Objective assessment shows that energy normalisation is critical to performance. These findings and the appeal of CQT for ABE are confirmed through informal subjective tests based on the mean opinion score.

Index Terms— bandwidth extension, constant Q transform

1. INTRODUCTION

There is a fundamental link between the perceived quality of a speech signal and its bandwidth [1]. While wider bandwidths usually correspond to higher quality speech [2], this comes at the cost of higher bit rates [3]. As a result, speech signals are usually bandwidth-limited, often to either 4kHz or 8kHz.

Unvoiced phonemes typically exhibit significant energy across the wideband spectrum [1]. The higher frequency components are crucial to quality, with wideband speech signals yielding higher mean opinion scores approximately 1.4 times those of narrowband speech [1]. Even so, most legacy telephone networks operate within the 300-3400Hz range referred to as narrowband (NB). As a result, artificial bandwidth extension (ABE) algorithms have been developed to compensate for the consequential loss in intelligibility and quality by estimating the missing highband (HB) components above 3400Hz from the available NB components, thereby producing an estimated wideband (WB) signal.

Most ABE algorithms are based upon the classical source-filter model of speech production where a NB speech signal is represented by an excitation source and a vocal tract filter. The frequency content of these two components can be extended through independent processing before a WB signal is resynthesised. In practice, however, most approaches focus on the extension of the spectral envelope since it has the dominant impact on speech quality [1].

The use of many different feature representations has been reported, e.g. linear prediction coefficients (LPC) [4], line spectral frequencies (LSF) [5] and mel-frequency cepstral coefficients (MFCC) [6]. A mixed approach reported in [1] uses NB auto-correlation coefficients to estimate WB cepstral coefficients. All of these methods learn a mapping between NB and HB frequency components according to their correlation [7]. This relationship has been studied in terms of mutual information [8] and in combination with entropy [9] and with separability [10].

Other ABE algorithms which operate on the complex speech spectrum have also been reported. Notable examples include: a spectral folding approach and a shaping function learned from neural networks [11]; an adaptive spline neural network to estimate directly the missing HB spectral coefficients [12]; a deep neural network (DNN)-based approach using log-power spectrum features [13]; a sum-product network for the estimation of missing HB components [14], and a joint-dictionary approach which exploits sparsity [15].

Most of these approaches employ the short-time Fourier transform (STFT) for spectral analysis. The STFT has a fixed frequency resolution. This is equivalent to a bank of filters with variable Q factors. The latter is a measure of filter selectivity and is defined as the ratio of centre frequency and bandwidth. By contrast, the human auditory system exhibits constant Q characteristics between 500Hz to 20kHz [16].

This hypothesis is supported by [8] which shows that per-

P.B. Bachhav is supported by Intel.

ceptually inspired approaches to ABE may produce more natural speech than those which focus solely on the correlation between NB and HB components. Being a more perceptually motived approach to frequency analysis than Fourier counterparts, this paper thus reports our attempts to harness the CQT for bandwidth extension.

The rest of the paper is organised as follows: Section 2 presents the CQT; Section 3 describes its use for ABE; Section 4 describes objective and subjective assessments; conclusions are presented in Section 5.

2. THE CONSTANT Q TRANSFORM

Introduced by Youngberg and Boll [17] in 1978 and redefined by Brown [18] in 1991, the constant Q transform (CQT) is a perceptually motivated approach to time-frequency analysis. In contrast to Fourier-based analysis, CQT bin centres are geometrically distributed following the equal-tempered scale of Western music [19]. The CQT is popular in the field of music processing, e.g. [20, 21] and was recently applied to a number of speech processing problems [22, 23, 24].

The so-called Q factor reflects the selectivity of a filter used in time-frequency analysis and is defined as the ratio of its centre frequency and bandwidth:

$$Q = \frac{f_k}{f_{k+1} - f_k} \tag{1}$$

where $k = 1, 2, ..., K$ is the frequency bin index and where f_k is the centre frequency of bin k. When the bin frequencies are geometrically distributed as in the CQT transform, then Eq. 1 is simplied to $Q = (2^{1/B} - 1)^{-1}$ [18] where B is the number of bins per octave. The centre frequencies f_k are defined according to:

$$f_k = f_1 2^{(k-1)/B}$$

where f_1 is the centre frequency of the lowest bin. B thus determines the time-frequency resolution trade-off. Compared to the STFT, the CQT has a greater frequency resolution for lower frequencies but a greater temporal resolution for higher frequencies.

The CQT of a discrete signal $x(n)$ is defined by:

$$X^{CQ}(k,n) = \sum_{j=n-\lfloor N_k/2 \rfloor}^{n+\lfloor N_k/2 \rfloor} x(j) a_k^*(j - n + N_k/2)$$

where $a_k(n)$ are basis functions, $*$ is the complex conjugate and N_k is a variable window length. Full details of the CQT, including definitions of a_k and the inverse-CQT (ICQT) are presented in [25, 26] with efficient implementations.

3. ARTIFICAL BANDWIDTH EXTENSION

This section describes our approach to ABE using the CQT. The algorithm is illustrated in Fig. 1. ABE is performed using Gaussian mixture models (GMMs) learned using a database of parallel NB and WB speech utterances. Features are log-magnitude spectral estimates. Resynthesis is performed using artificially extended HB magnitude estimates and HB phase estimates obtained from the upsampled NB signal. Details of each step are given in the following.

3.1. Training

CQT feature extraction is applied exclusively to WB signal sampled at 16kHz. NB signals sampled at 16kHz are obtained from the treatment of WB signals with mobile station input (MSIN) highpass filtering [27] followed by lowpass filtering and level adjustment to active speech level of -26 dBov [28]. For NB signals (top pipeline of the training block in Fig. 1), features are extracted for NB components according to:

$$X^{NB} = \ln|X^{CQ}(k,n)|, \forall n, \forall k = \{k : f_k \epsilon [f_1, 3700] \text{ Hz}\} \tag{2}$$

For WB signals, features are extracted for HB components only according to:

$$Y^{HB} = \ln|Y^{CQ}(k,n)|, \forall n, \forall k = \{k : f_k \epsilon [3700, 8000] \text{ Hz}\} \tag{3}$$

The CQT is applied with values of $B = 48$ bins per octave, a lowest bin frequency of $f_1 = 250$Hz and a maximum bin frequency of $f_{max} = 8000$Hz. This gives features of dimensions 187 and 52 for NB and WB signals respectively which are mean and variance normalised (mvn) to give features X^{NB}_{mvn} and Y^{HB}_{mvn}. The two components are concatenated to give $Z = [X^{NB}_{mvn}, Y^{HB}_{mvn}]$, 239-variate feature vectors which are modeled as a mixture of 512 Gaussian components with full covariance matrices.

3.2. Extension

NB signals are first upsampled to 16kHz ($\hat{x}$) before feature extraction is applied as in Eq. 2. For every available normalised feature vector $\widehat{X}_{mvn}$, the missing normalised HB component $\widehat{Y}_{mvn}$ is estimated so as to minimise the mean square error (MSE):

$$MSE = E\left[\left\|\widehat{Y}_{mvn} - F(\widehat{X}_{mvn})\right\|^2\right]$$

where F is the mapping function obtained from the GMM of joint vectors Z. This standard mapping operation is described in [4]. Inverse mean and variance normalisation (mvn^{-1}) is then applied, using means and variances obtained from the training data, to obtain the HB log-magnitude estimate $\widehat{Y}$.

3.3. Resynthesis

Time-domain speech signals are resynthesised using the ICQT. The magnitude component M is a concatenation

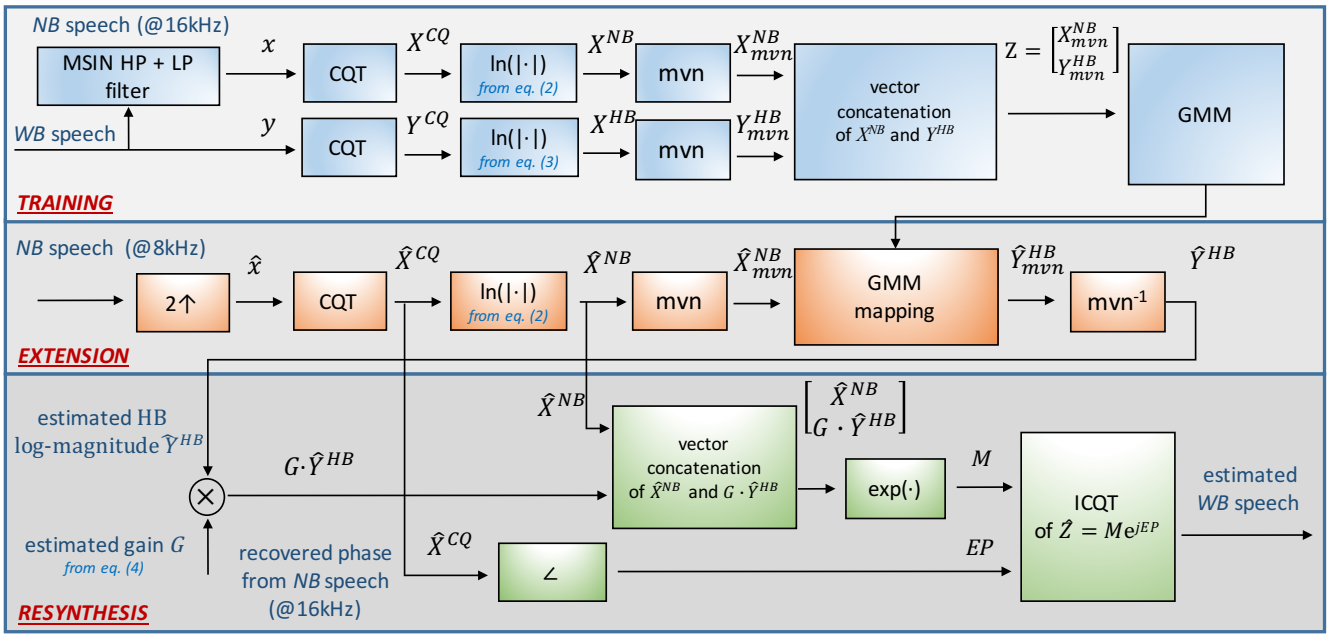

Fig. 1. *Block diagram of the CQT-based ABE system.*

of the original NB component $M_{NB} = \exp(\widehat{X}^{NB})$ and the artifically generated and gain-adjusted HB component $M_{HB} = G * \exp(\widehat{Y}^{HB})$. The phase component is extracted from an upsampled version of the NB input signal. This condition is referred to later as estimated phase (EP):

$$EP = \angle \widehat{X}^{CQ}(k,n), \forall n, \forall k = \{k : f_k \epsilon [f_1, 8000] \text{ Hz}\}$$

The extended signal is resynthesised by performing the ICQT on the vector $\widehat{Z} = M\exp(jEP)$. In order to gauge the degradation incurred as a result of using EP, contrastive ABE experiments were performed using oracle phase components extracted from the application of CQT to original WB signals. This condition is referred to as oracle phase (OP):

$$OP = \angle Y^{CQ}(k,n), \forall n, \forall k = \{k : f_k \epsilon [3700, 8000] \text{ Hz}\}$$

Resynthesis is then performed using the concatenated phase of the upsampled NB signal with OP defined above.

The gain adjustment G corrects for differences between the energy of estimated and original HB components and is estimated as follows. Estimates of the HB i.e. $\widehat{Y}^{HB}_{train}$ were obtained from X^{NB} for the entire training set. A polynomial regression [29] of order 4 was then performed between the root mean-square (RMS) values $\sqrt{E^{HB}_{oracle}}$ and $\sqrt{\widehat{E}^{HB}_{train}}$ where $E^{HB}_{oracle} = \sum |Y^{CQ}|^2$ is the true HB energy calculated from the original WB signal and $\widehat{E}^{HB}_{train} = \sum (\exp(\widehat{Y}^{HB}_{train}))^2$ is the energy of estimated HB for the training data.

During resynthesis, estimated gain is then given by

$$G = \sqrt{E^{HB}_{reg} / \widehat{E}^{HB}} \tag{4}$$

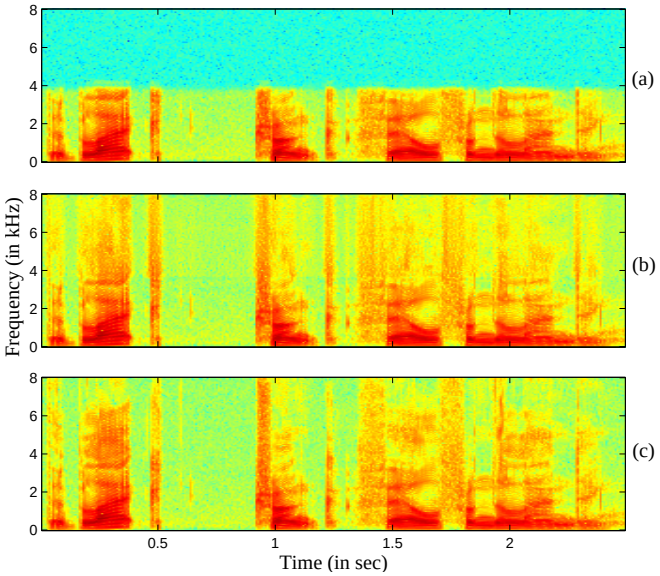

Fig. 2. *Spectrograms of an upsampled NB speech (top), artificially extended WB speech (middle) and original WB speech (bottom).*

where $\sqrt{E^{HB}_{reg}}$ is obtained through regression from $\sqrt{\widehat{E}^{HB}}$ and $\widehat{E}^{HB} = \sum (\exp(\widehat{Y}^{HB}))^2$ is the energy of estimated HB. Two further experimental conditions aim to assess the benefit of gain adjustment. The first is performed with the gain estimated as described above. This condition is referred to as estimated gain (EG). A contrastive condition uses the oracle gain (OG) given by $\sqrt{E^{HB}_{oracle} / \widehat{E}^{HB}}$. Spectrograms of NB, artifically extended and original WB utterances are illustrated

in Fig. 2.

4. EXPERIMENTAL SETUP AND RESULTS

This section describes the experimental setup and results for both objective and subjective assessments.

4.1. Database

ABE experiments are performed using the TSP speech database [30] which consists of 1378 phonetically balanced Harvard sentences spoken by 12 male and 12 female speakers and recorded with a sampling frequency of 48kHz. WB versions were created by downsampling the original files to 16kHz. For training, parallel NB speech signals were created as described in Section 3.1.

4.2. Objective metrics

Objective assessments were first performed using the RMS log-spectral distortion (RMS-LSD) [31] which is known to correlate well with subjective assessment results [32]. It is given by:

$$\text{RMS-LSD} = \sqrt{\frac{1}{\triangle F}\int_{\triangle F}\left[20log_{10}\left|\frac{g}{H(f)} - \frac{\hat{g}}{\hat{H}(f)}\right|\right]^2 df}$$

where $\triangle F = [3700, 8000]$Hz, $H(f)$ and $\hat{H}(f)$ are the original and estimated envelopes calculated in the frequency range $\triangle F$ using linear prediction analysis and where g and $\hat{g}$ are the gains of their respective excitation components. In practice, a summation is used to calculate the integral.

Table 1 shows RMS-LSD assessment results for ABE using oracle and estimated gain and phase. Without gain normlisation, the RMS-LSD is high. As expected, the lowest RMS-LSD is achieved when using OG and OP. The increase in RMS-LSD is greater when EG is used in place of OG. There is little difference between RMS-LSD obtained using OP and EP, thereby indicating greater sensitivity to gain than phase. This result is not suprising given the relative insensitivity of human perception to phase. The practical ABE system with EG and EP gives RMS-LSD figures of 2.46 and 4.64dB for training and testing sets respectively. These figures are marginally higher than those of 1.89 and 3.13dB for OG and OP.

4.3. Subjective listening test

Subjective assessments were performed using comparison-based mean-opinion score (MOS) tests [31]. Tests were performed for artificially extended WB signals through comparison to NB and original WB signals, using EG-EP and OG-EP configurations. Tests were performed with 10 listeners who were asked to compare the quality of 10 pairs of speech signals A and B. They were asked to rate the quality of signal B with respect to A according to the following scale: -3 (much worse), -2 (slightly worse), -1 (worse), 0 (about the same), 1 (slightly better), 2 (better), 3 (much better). All speech files used for subjective tests are available online[1].

Table 1. *RMS-LSD results (in dB) with and without gain normalization and different phase extensions. OG - oracle gain, EG - estimated gain, OP - oracle phase, EP - estimated phase. EG-EP is the proposed method.*

Gain	Phase	Train Mean (σ)	Test Mean (σ)
-	OP	3.01 (0.72)	5.28 (1.51)
-	EP	3.21 (0.71)	5.39 (1.49)
OG	OP	1.89 (0.37)	3.13 (0.67)
OG	EP	2.16 (0.38)	3.30 (0.67)
EG	OP	2.46 (0.40)	4.64 (1.06)
EG	**EP**	2.66 (0.42)	4.77 (1.05)

Table 2. *Comparison based MOS for EP with EG and OG. EG-EP is the proposed method.*

Comparison B $\rightarrow$ A	MOS
EG-EP $\rightarrow$ NB	1.12
OG-EP $\rightarrow$ NB	1.14
EG-EP $\rightarrow$ WB	-1.42
OG-EP $\rightarrow$ WB	-1.03

MOS scores are illustrated in Table 2. While both EG-EP and OG-EP systems were rated poorer than original WB signals, they were rated better than NB speech signals. As expected, the OG-EP system gave higher MOS scores than the EG-EP system, althought the difference is modest. Lastly, the improvement in MOS scores correlates with the improvement in objective RMS-LD scores.

5. CONCLUSIONS

The constant Q transform (CQT) is a perceptually motivated approach to time-frequency analysis. This paper reports its first application to artificial bandwidth extention (ABE). Both objective and subjective experimental results show that the proposed approach using the CQT produces higher-quality, higher-bandwidth speech signals. While phase is shown to be relatively unimportant, the accurate estimation of spectral magnitude and gain is critical. Future work should target better approaches to gain estimation and the optimisation of CQT analysis for ABE. Alternative narrowband to wideband mapping or regression techniques such as those rooted in deep learning may also deliver further improvements.

[1] http://audio.eurecom.fr/content/media

6. REFERENCES

[1] P. Jax and P. Vary, "On artificial bandwidth extension of telephone speech," *Signal Processing*, vol. 83, no. 8, pp. 1707–1719, 2003.

[2] A. Shahina and B. Yegnanarayana, "Mapping neural networks for bandwidth extension of narrowband speech." in *INTERSPEECH*, 2006.

[3] S. Voran, "Listener ratings of speech passbands," in *Proc. of IEEE Workshop on Speech Coding For Telecommunications*, 1997, pp. 81–82.

[4] K.-Y. Park and H. S. Kim, "Narrowband to wideband conversion of speech using GMM based transformation," in *Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2000, pp. 1843–1846.

[5] Y. Qian and P. Kabal, "Dual-mode wideband speech recovery from narrowband speech." in *INTERSPEECH*, 2003.

[6] A. H. Nour-Eldin and P. Kabal, "Mel-frequency cepstral coefficient-based bandwidth extension of narrowband speech." in *INTERSPEECH*, 2008, pp. 53–56.

[7] Y. M. Cheng, D. O'Shaughnessy, and P. Mermelstein, "Statistical recovery of wideband speech from narrowband speech," *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 4, pp. 544–548, 1994.

[8] M. Nilsson, H. Gustaftson, S. V. Andersen, and W. B. Kleijn, "Gaussian mixture model based mutual information estimation between frequency bands in speech," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2002, pp. I–525.

[9] A. H. Nour-Eldin and P. Kabal, "Objective analysis of the effect of memory inclusion on bandwidth extension of narrowband speech." in *INTERSPEECH*, 2007, pp. 2489–2492.

[10] P. Jax and P. Vary, "Feature selection for improved bandwidth extension of speech signals," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2004, pp. I–697.

[11] J. Kontio, L. Laaksonen, and P. Alku, "Neural network-based artificial bandwidth expansion of speech," *IEEE trans. on audio, speech, and language processing*, vol. 15, no. 3, pp. 873–881, 2007.

[12] A. Uncini, F. Gobbi, and F. Piazza, "Frequency recovery of narrow-band speech using adaptive spline neural networks," in *Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 1999, pp. 997–1000.

[13] K. Li and C.-H. Lee, "A deep neural network approach to speech bandwidth expansion," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 4395–4399.

[14] R. Peharz, G. Kapeller, P. Mowlaee, and F. Pernkopf, "Modeling speech with sum-product networks: Application to bandwidth extension," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 3699–3703.

[15] J. Sadasivan, S. Mukherjee, and C. S. Seelamantula, "Joint dictionary training for bandwidth extension of speech signals," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 5925–5929.

[16] C. Schörkhuber and A. Klapuri, "Constant-Q transform toolbox for music processing," in *7th Sound and Music Computing Conference, Barcelona, Spain*, 2010, pp. 3–64.

[17] J. Youngberg and S. Boll, "Constant-Q signal analysis and synthesis," in *Proc. of IEEE Int Conf on Acoustics, Speech, and Signal Processing (ICASSP)*, 1978, pp. 375–378.

[18] J. Brown, "Calculation of a constant Q spectral transform," *Journal of the Acoustical Society of America*, vol. 89, no. 1, pp. 425–434, January 1991.

[19] R. E. Radocy and J. D. Boyle, *Psychological foundations of musical behavior*. C. C. Thomas, 1979.

[20] E. Dorken and S. H. Nawab, "Improved musical pitch tracking using principal decomposition analysis," in *Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 1994, pp. II/217–II/220.

[21] G. Costantini, R. Perfetti, and M. Todisco, "Event based transcription system for polyphonic piano music," *Signal Process.*, vol. 89, no. 9, pp. 1798–1811, Sep. 2009.

[22] M. Todisco, H. Delgado, and N. Evans, "A new feature for automatic speaker verification anti-spoofing: Constant Q cepstral coefficients," in *Speaker Odyssey Workshop*, Bilbao, Spain, 2016.

[23] T. Kinnunen, M. Sahidullah, I. Kukanov, H. Delgado, M. Todisco, A. Sarkar, N. B. Thomsen, V. Hautamki, N. Evans, and Z.-H. Tan, "Utterance verification for text-dependent speaker recognition: a comparative assessment using the RedDots corpus," in *Proc. of INTERSPEECH (to appear)*, 2016.

[24] H. Delgado, M. Todisco, M. Sahidullah, A. Sarkar, N. Evans, Z.-H. Tan, and T. Kinnunen, "Further optimisations of constant Q cepstral processing for integrated utterance and text-dependent speaker verification," in *Submitted to SLT IEEE Workshop on Spoken Language Technology*, 2016.

[25] G. A. Velasco, N. Holighaus, M. Dorfler, and T. Frill, "Constructing an invertible constant-Q transform with nonstationary Gabor frames," in *Proc. Digital Audio Effects (DAFx-11)*, 2011.

[26] C. Schörkhuber, A. Klapuri, N. Holighaus, and M. Dörfler, "A Matlab toolbox for efficient perfect reconstruction time-frequency transforms with log-frequency resolution," in *Audio Engineering Society (53rd Conference on Semantic Audio)*, G. Fazekas, Ed., AES (Vereinigte Staaten (USA)), 6 2014.

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

[28] "ITU-T Recommendation P. 56, Objective measurement of active speech level," *ITU*, 2011.

[29] S. Rogers and M. Girolami, *A first course in machine learning*. CRC Press, 2015.

[30] P. Kabal, "TSP Speech Database," *McGill University, Database Version : 1.0*, pp. 02–10, 2002.

[31] D. Zaykovskiy and B. Iser, "Comparison of neural networks and linear mapping in an application for bandwidth extension," in *Proc. of Int. Conf. on Speech and Computer (SPECOM)*, 2005, pp. 1–4.

[32] P. Jax and P. Vary, "An upper bound on the quality of artificial bandwidth extension of narrowband speech signals," in *Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2002, pp. I–237.