Spectral Restoration of Narrowband Speech Recordings Supported by Phonetic Transcriptions

P. Bauer, T. Fingscheidt

TU Braunschweig, Institute for Communications Technology, Germany, Email: {bauer, fingscheidt}@ifn.ing.tu-bs.de

Abstract

Due to limitations of the acoustic bandwidth, historic and telephone recordings suffer from poor speech quality and intelligibility. Nevertheless they are widely found in radio, television, internet, as well as in storage media in the context of newscast reports, documentations, or large databases with archived speech material. Spectral restoration of such narrowband speech recordings improves the auditory impression, as well as intelligibility. As the recordings are available offline, phonetic transcriptions already exist, or can be extracted from the speech data manually or automatically, and support the spectral restoration process. Hence, subjective quality and intelligibility much closer to wideband speech can be expected. In this paper, an artificial bandwidth extension (ABWE) is presented to spectrally restore 8 kHz sampled speech signals by upsampling to 16 kHz and estimating further frequency regions of interest. It makes use of phonetic transcriptions in order to improve the partially insufficient ABWE performance on critical phonemes, i.e., mainly fricatives /s/, /z/, and /f/. We found the spectrally restored speech significantly enhanced, particularly the typical lisping effect disappeared in many instances.

Introduction

Artificial bandwidth extension (ABWE) usually performs speech enhancement by upsampling of narrowband speech, e.g., telephone speech at $f_s = 8$ kHz sampling rate, and estimating further frequency components of interest, i.e., up to 7 kHz at $f_s = 16$ kHz. Examples of typical ABWE systems are given in [1–3]. Often high-frequency whistling and lisping effects are observed, which are tackled, e.g., in [4, 5]. Fricatives, such as /s/, /z/, and partly /f/, /S/, /Z/, are difficult to be estimated based upon only a narrowband speech signal [6]. A considerable portion of their energy is located at higher frequencies, while the low-frequency portion is highly confusable among these sounds.

Some authors have pushed ABWE quality further by transmitting low rate side information [7]. This is also done in speech codecs, such as the adaptive multirate wideband (AMR-WB) codec [8]. It turns out that a few hundred extra bits per second allow for a high-quality wideband speech reconstruction, while only a narrowband speech signal is transmitted.

In this paper we show how side information of a different kind can be exploited in the ABWE estimation process: We assume the availability of time-aligned phonetic transcriptions along with the narrowband speech waveforms in training *and* test. Note that transcriptions can always be generated offline, either by human transcriptors or automatically by forced Viterbi alignment. So in contrast to [7], our approach does not require any (far-end) availability of wideband speech at all.

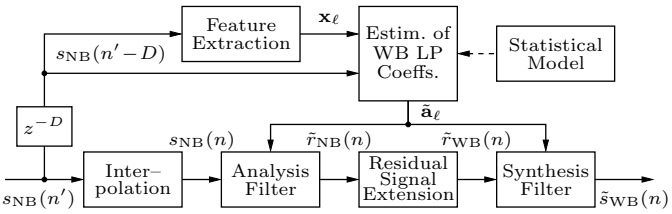

Figure 1: State-of-the-art artificial bandwdith extension

A possible field of application using phonetic transcriptions in ABWE is spectral restoration of historic and telephone speech recordings. Due to limitations of their acoustic bandwidth, they suffer from poor speech quality and intelligibility. However, they are widely found in radio, television, internet, as well as in storage media in the context of newscast reports, documentations, or large databases with archived speech material. Note that Hansen et al. have already worked on speech enhancement applications that exploit phonetic a–priori knowledge for noise reduction [9].

Our paper is organized as follows: First, the state-of-the-art ABWE scheme will be recapitulated that works only on speech waveforms. Then we will show how time-aligned phonetic transcriptions can be included into the ABWE training and estimation process. Finally, some experimental results will be discussed in terms of achieved speech quality.

State-of-the-Art ABWE System

The state-of-the-art artificial bandwidth extension shown in Fig. 1 on high level exploits only the speech waveform [10]. In Fig. 2 the estimation of the wideband (WB) linear prediction (LP) coefficients is further detailed. The ABWE system employs a statistical model based on a hidden Markov model (HMM) quite similar to [1]. A brief overview and some specifics will be given in the following: The narrowband ($f_s = 8$ kHz) speech signal $s_{\mathrm{NB}}(n')$ with sample index n' is subject to interpolation yielding the upsampled speech signal $s_{\mathrm{NB}}(n)$ with sample index n referring to 16 kHz sampling rate. The actual processing of the interpolated speech signal consists of a mere WB LP analysis filter, extension of the resulting narrowband residual $\tilde{r}_{\mathrm{NB}}(n)$ by spectral folding (zeroing every other sample), and final LP synthesis filtering of the extended excitation signal $\tilde{r}_{\mathrm{WB}}(n)$ with the very same coefficients

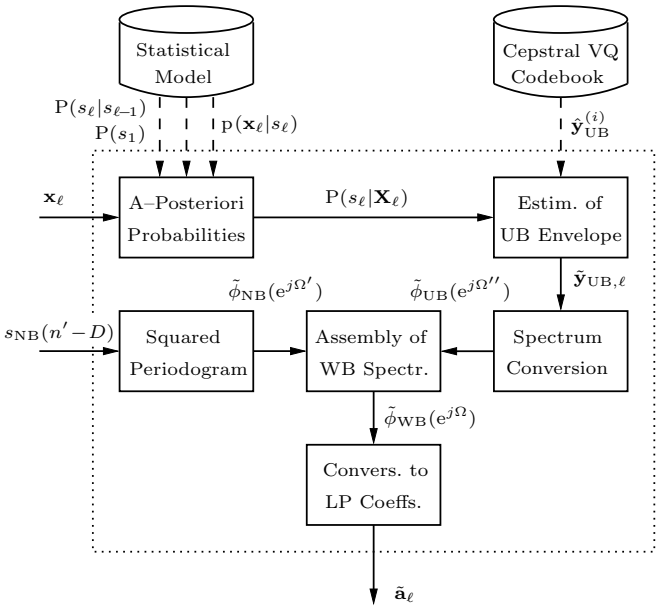

Figure 2: State-of-the-art estimation of wideband linear prediction coefficients $\tilde{\mathbf{a}}_\ell$

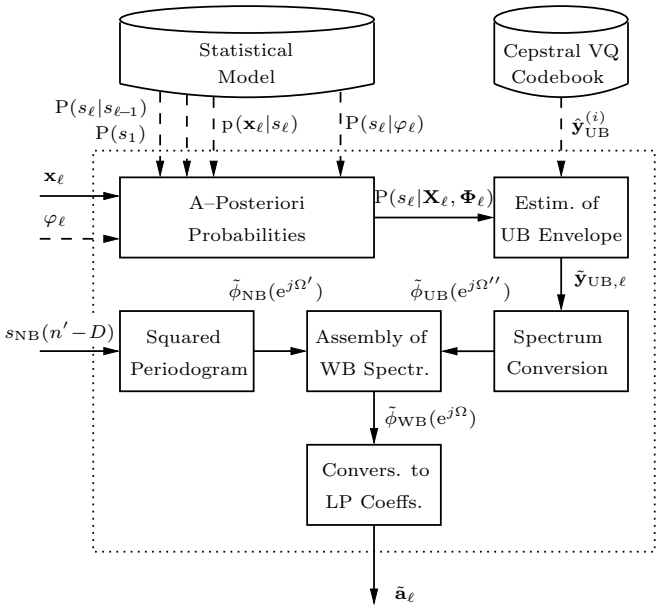

Figure 3: Estimation of wideband linear prediction coefficients $\tilde{\mathbf{a}}_\ell$ using phonetic transcriptions φ_ℓ

as they were used in the analysis filter. Since there are only modifications of the upper frequency band, and the LP analysis and synthesis filters are totally inverse, this scheme is transparent towards the narrow band of the resulting estimated WB speech signal $\tilde{s}_{\mathrm{WB}}(n)$. In the following, we describe how the estimated WB LP coefficient vector $\tilde{\mathbf{a}}_\ell \in \mathbb{R}^{16}$ in frame ℓ is computed from the narrowband speech signal.

Estimation of Wideband LP Coefficients

After delay compensation yielding a narrowband signal $s_{\mathrm{NB}}(n'-D)$ that is time-aligned to its interpolated version at 16 kHz, feature extraction is performed. It operates with a frame length of 20 ms and a frame shift of 10 ms, accordingly the WB LP coefficients are updated every 10 ms. The primary features are 10 autocorrelation coefficients, the zero crossing rate, gradient index, normed relative frame energy, local kurtosis, and spectral centroid, as proposed in [11]. A linear discriminant analysis (LDA) is employed to reduce the dimension of the primary feature vector from $d_0 = 15$ to $d = 5$. The resulting feature vector $\mathbf{x}_\ell \in \mathbb{R}^d$ is subject to a statistical model.

Assuming a certain state $s_\ell = i$, $i \in \mathbb{S}' = \{1, \ldots, N'\}$ of the HMM model, the observation probability density function (PDF) $\mathrm{p}(\mathbf{x}_\ell|s_\ell = i)$ for the known observation $\mathbf{x}_\ell$ is computed from a Gaussian mixture model (GMM) obtained in training. By combining the observation PDF with the pre-trained state transition probabilities $\mathrm{P}(s_\ell = i|s_{\ell-1} = j)$, the state a–posteriori probabilities of frame ℓ can be calculated in a recursive fashion

$$\mathrm{P}(s_\ell = i|\mathbf{X}_\ell) = C \cdot \mathrm{p}(\mathbf{x}_\ell|s_\ell = i) \cdot \sum_{j=1}^{N'} \mathrm{P}(s_\ell = i|s_{\ell-1} = j) \cdot \mathrm{P}(s_{\ell-1} = j|\mathbf{X}_{\ell-1}), \quad (1)$$

where $\mathbf{X}_\ell = \{\mathbf{x}_\ell, \mathbf{x}_{\ell-1}, \ldots, \mathbf{x}_1\}$ denotes the sequence of observations. Factor C just normalizes the sum of the a–posteriori probabilities over all states s_ℓ to one. Using the vector quantizer (VQ) codebook of upper band (UB) cepstral coefficient vectors $\hat{\mathbf{y}}_{\mathrm{UB}}^{(i)}$ with index $i = 1, \ldots, N'$, as obtained during training, the a–posteriori probabilities are utilized to perform a minimum mean square error (MMSE) estimation of the upper frequency band in the cepstral domain:

$$\tilde{\mathbf{y}}_{\mathrm{UB},\ell} = \sum_{i=1}^{N'} \hat{\mathbf{y}}_{\mathrm{UB}}^{(i)} \cdot \mathrm{P}(s_\ell = i|\mathbf{X}_\ell).$$

After having assembled both spectra, of the narrow band (NB) $\tilde{\phi}_{\mathrm{NB}}(\mathrm{e}^{j\Omega'})$ and of the upper band $\tilde{\phi}_{\mathrm{UB}}(\mathrm{e}^{j\Omega''})$, the resulting WB power spectrum $\tilde{\phi}_{\mathrm{WB}}(\mathrm{e}^{j\Omega})$ is finally converted via the Levinson-Durbin recursion into the required WB LP coefficient vector $\tilde{\mathbf{a}}_\ell$. Note that the normalized frequencies Ω', Ω'' cover only the NB and UB of the WB spectrum, respectively.

Training Process

Using WB speech training material, a statistical model is trained [1, 11]. As a first step, a VQ codebook of UB cepstral coefficients $\hat{\mathbf{y}}_{\mathrm{UB}}^{(i)} = \mathrm{E}\{\mathbf{y}_{\mathrm{UB}}|s_\ell = i\} \in \mathbb{R}^9$ is obtained by selective linear prediction, computation of cepstral coefficients, and LBG training with $N' = 16$ entries. This VQ is then framewise applied to the WB training database yielding a certain state $s_\ell = i$ as classification result for frame ℓ. In a second step, these classification results are used to train an LDA matrix as part of the feature extraction. In a third step, initial state probabilities $\mathrm{P}(s_1)$ and state transition probabilities $\mathrm{P}(s_\ell = i|s_{\ell-1} = j)$ are trained and stored for ABWE test. Finally, using the LDA-transformed feature vectors $\mathbf{x}_\ell$, the parameters of a GMM-based observation PDF $\mathrm{p}(\mathbf{x}_\ell|s_\ell = i)$ are derived from expectation maximization (EM) training: a scalar weighting factor, a mean vector, and a diagonal covariance matrix of every d-dimensional normal distribution. For each HMM state, a separately trained GMM of $G = 8$ mixtures is used.

Transcription-Supported ABWE

Phonetic investigations about ABWE recently demonstrated that the representatives of the UB cepstral coefficient codebook obtained from the state-of-the-art training process are insufficient for the reconstruction of sharp /s/- and /z/-sounds [6]. Lisping effects are the consequence forming *the* major obstacle for the acceptance of ABWE.

Training Process Using Transcriptions

In order to solve the typical lisping problem, the LBG-based codebook training has been modified by using phonetic transcriptions: The redesigned codebook consists of $N = 24$ entries, those of the state-of-the-art system that have been LBG-trained, augmented by another $N - N' = 8$ entries that have been trained solely on /s/-phoneme instances. The training of these 8 entries was found to be advantageously conducted by performing a 64-entry LBG training on /s/ frames only, and keeping only the 8 *sharpest* representatives. They are determined as those entries with the largest positive cepstral distance to the mean of the 64 found representatives. The $N - N'$ additional codebook entries are used to represent /s/ and /z/, since both fricatives can be treated in the upper frequency band as a group, i.e., their characteristics (voicing properties) are mostly contained in the NB [6]. Based on this redesigned codebook yielding N HMM states $s_\ell = i$, $i \in \mathbb{S} = \{1, \ldots, N', \ldots, N\}$, the LDA matrix and the statistical model are trained in analogy to the state-of-the-art ABWE training.

Estimation of WB LP Coefficients Using Transcriptions

In contrast to (1), the modified ABWE system illustrated in Fig. 3 requires framewise phonetic transcriptions φ_ℓ for the recursive computation of the state a–posteriori probabilities. To support those sub-codebook states that are dedicated to represent sharp /s/- and /z/-sounds, *state probabilities given the phonetic transcription*

$$\mathrm{P}(s_\ell = i|\varphi_\ell) = \begin{cases} \mathrm{P}(i)\,, & \text{if } \varphi_\ell \in \{/\mathrm{s}/,\ /^*\mathrm{s}/,\ /\mathrm{z}/\}, \\ \frac{1}{N}\,, & \text{else} \end{cases} \quad (2)$$

are included. They are driven by the phonetic labels φ_ℓ using the probability mask

$$\mathrm{P}(i) = \begin{cases} \epsilon\,, & i = 1, \ldots, N' \\ \frac{1-N'\epsilon}{N-N'}\,, & i = N'+1, \ldots, N, \end{cases}$$

with $\epsilon = 10^{-4}$ being just a small value and $\sum_{i=1}^{N} \mathrm{P}(i) = 1$. Note that $N - N'$ classes in $\mathbb{S}$ refer to transcriptions $\varphi_\ell \in \{/\mathrm{s}/,\ /^*\mathrm{s}/,\ /\mathrm{z}/\}$, while N' classes cover the rest of the phoneme labels. The transcription-supported a–posteriori probabilities can be recursively computed by

$$\mathrm{P}(s_\ell = i|\mathbf{X}_\ell, \mathbf{\Phi}_\ell) = C \cdot \mathrm{p}(\mathbf{x}_\ell|s_\ell = i) \cdot \frac{\mathrm{P}(s_\ell = i|\varphi_\ell)}{\mathrm{P}(s_\ell = i)} \cdot \sum_{j=1}^{N} \mathrm{P}(s_\ell = i|s_{\ell-1} = j) \cdot \mathrm{P}(s_{\ell-1} = j|\mathbf{X}_{\ell-1}, \mathbf{\Phi}_{\ell-1}), \quad (3)$$

	ABWE	ABWE-PHON
$\overline{d}_{\mathrm{LSD}}$ [dB]	3.32	3.43
$d_{\mathrm{LSD},5-10}$ [%]	13.72	16.25
$d_{\mathrm{LSD},>10}$ [%]	0.37	0.48

Table 1: Total results of log-spectral distortion (LSD) for the state-of-the-art ABWE system (ABWE) and for the modified system using phonetic transcriptions (ABWE-PHON)

where $\mathbf{\Phi}_\ell = \{\varphi_\ell, \varphi_{\ell-1}, \ldots, \varphi_1\}$ denotes the phoneme sequence. Note that in (2) offsets /s*/ and /z*/ are implicitly covered by the state-of-the-art ABWE technique, since $\mathrm{P}(s_\ell = i|/\mathrm{s}^*/) = \mathrm{P}(s_\ell = i|/\mathrm{z}^*/) = \frac{1}{N}$. This takes into account the fact that due to the recursive nature of (3), the influence of the use of transcriptions /s/ or /z/ in (2) is propagated anyway to some extent into the following frames. Note further that in contrast to onsets /*s/, onsets /*z/ were found not to take advantage of the transcriptions and therefore were also computed by means of $\mathrm{P}(s_\ell = i|/^*\mathrm{z}/) = \frac{1}{N}$. As concerns $\mathrm{P}(s_\ell = i)$ in (3), we simply assumed $\mathrm{P}(s_\ell = i) = \frac{1}{N}$ in any case to avoid a certain overemphasis of classes $i = N' + 1, ..., N$.

Additionally, moderate smoothing of the cepstral estimates is performed in a way that the log-spectral distortion (LSD) computed from the UB cepstral coefficients $\tilde{\mathbf{y}}_{\mathrm{UB},\ell-1}$ and $\tilde{\mathbf{y}}_{\mathrm{UB},\ell}$ for consecutive frames satisfies

$$\sqrt{2\left(\tfrac{10}{\ln 10}\right)^2 \cdot ||\tilde{\mathbf{y}}_{\mathrm{UB},\ell-1} - \tilde{\mathbf{y}}_{\mathrm{UB},\ell}||^2} \leq 30\,\mathrm{dB}.$$

Experimental Results

We performed experiments for both ABWE systems based on the *SpeechDat-Car* database in US English [12]. WB speech training material was obtained from 6 male and 6 female speakers. Each of the speakers provided 2 speech sessions. The resulting 24 speech sessions were excluded from test. For test purposes, 404 speech sessions obtained from 202 speakers were used. The WB LSD measure

$$d_{\mathrm{LSD}} = \sqrt{2\left(\tfrac{10}{\ln 10}\right)^2 \cdot ||\mathbf{y}_{\mathrm{WB}} - \tilde{\mathbf{y}}_{\mathrm{WB}}||^2}$$

served as a performance evaluation measure with $\mathbf{y}_{\mathrm{WB}}$ and $\tilde{\mathbf{y}}_{\mathrm{WB}}$ being the 64-dim. cepstral coefficient vectors of the original WB speech signal and of the bandwidth-extended signal, respectively. Besides the mean LSD $\overline{d}_{\mathrm{LSD}}$, we computed the percentage of LSD outliers in the range of $5 \ldots 10$ dB ($d_{\mathrm{LSD},5-10}$) and beyond 10 dB ($d_{\mathrm{LSD},>10}$).

Table 1 shows total LSD results of both ABWE experiments. It turns out that the mean LSD is slightly increased in case of the modified ABWE system. Also the numbers of LSD outliers in both ranges are somewhat higher. The state-of-the-art ABWE system benefits from the fact that it is solely based on true LBG states, which provide on average the smallest cepstral distance. However, Fig. 4 exemplarily demonstrates for the utterance *"less poisonous"* how sharp /s/- and /z/-sounds are significantly improved by the transcription-supported ABWE approach. Due to a much better spectral representation of high-frequency energy components,

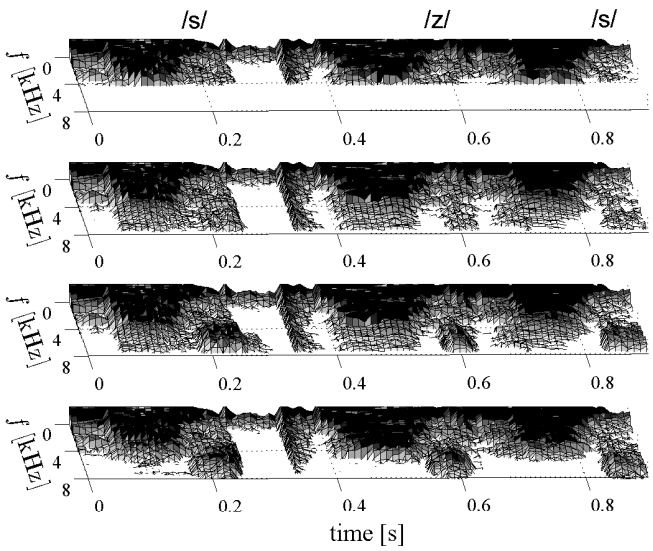

Figure 4: Figures from top to bottom: Spectrograms of the utterance *"less poisonous"* for (a) 8 kHz original speech, (b) state-of-the-art ABWE speech, (c) transcription-supported ABWE speech, (d) 16 kHz original speech

the typical lisping effect disappears in many instances. It turns out that a cepstral distance measure (which by nature computes averages) does not sufficiently reflect the remaining challenge of fricative bandwidth extension, at least for /s/- and /z/-sounds.

Informal listening tests revealed a significant speech quality improvement, when the modified ABWE system is compared to the state-of-the-art ABWE system. Comparing ABWE speech of the modified system directly to the original WB speech, informal listening tests revealed a smaller quality margin, as if ABWE speech of the state-of-the-art system was compared to the WB speech reference. This proves the potential of our approach to significantly narrow down the remaining speech quality's gap of ABWE by using phonetic transcriptions. For demonstration purposes, a spectrally restored speech sample from a historical recording of Franklin D. Roosevelt's first inaugural address delivered in 1933 will be presented at the conference.

Conclusions

In this paper we have proposed a modified artificial bandwidth extension (ABWE) basing on speech waveform *and* phonetic transcription to spectrally restore narrowband speech recordings. It aims at high-quality ABWE by reducing the lisping effects that typically appear in the state-of-the-art ABWE system regarding critical fricatives /s/, /z/. The core of our present work is the modification of the recursive a–posteriori probability computation in (3), which allows many ways now for the phonetic information to be involved into the restoration process. Although objective log-spectral distortion measures were found not to document significant improvements, informal subjective listening tests revealed an improved performance. In particular, the typical lisping effect disappeared in many instances, which was exemplarily confirmed by spectrograms, too. Speech samples of a historical recording will be given at the conference.

Acknowledgements

The authors have been greatly supported by D. Scheler who helped a lot with the transcription. We are thankful to Harald Höge from Siemens AG, Munich, Germany, for providing the SpeechDat-Car database in US English. The work was supported by the German Research Foundation (DFG) under grant no. FI 1494/2-1.

References

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

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

[3] M.L. Seltzer, A. Acero, and J. Droppo, "Robust Bandwidth Extension of Noise-Corrupted Narrowband Speech," in *Proc. of INTERSPEECH*, Lisbon, Portugal, Sept. 2005, pp. 1509–1512.

[4] M. Nilsson and W.B. Kleijn, "Avoiding Over-Estimation in Bandwidth Extension of Telephony Speech," in *Proc. of ICASSP*, Salt Lake City, Utah, USA, May 2001, pp. 869–872.

[5] S. Yao and C.-F. Chan, "Block-Based Speech Bandwidth Extension System with Separated Envelope Energy Ratio Estimation," in *Proc. of EUSIPCO*, Antalya, Turkey, Sept. 2005.

[6] P. Bauer, T. Fingscheidt, and M. Lieb, "Phonetic Analysis and Redesign Perspectives of Artificial Speech Bandwidth Extension," in *Proc. of ESSV*, Frankfurt a.M., Germany, Sept. 2008.

[7] B. Geiser, P. Jax, and P. Vary, "Robust Wideband Enhancement of Speech by Combined Coding and Artificial Bandwidth Extension," in *Proc. of IWAENC*, Eindhoven, The Netherlands, Sept. 2005.

[8] "Speech Codec Speech Processing Functions: AMR Wideband Speech Codec; Transcoding Functions (3GPP TS 26.190, Release 5)," 3GPP; TSG SA, Dec. 2001.

[9] J. H.L. Hansen and B. L. Pellom, "Text-Directed Speech Enhancement Employing Phone Class Parsing and Feature Map Constrained Vector Quantization," *Speech Communication*, pp. 169–190, Apr. 1997.

[10] P. Bauer and T. Fingscheidt, "An HMM-Based Artificial Bandwidth Extension Evaluated by Cross-Language Training and Test," in *Proc. of ICASSP*, Las Vegas, NV, USA, Apr. 2008.

[11] P. Jax, *Enhancement of Bandlimited Speech Signals: Algorithms and Theoretical Bounds*, Ph.D. thesis, vol. 15 of P. Vary (ed.), Aachener Beiträge zu digitalen Nachrichtensystemen, Nov. 2002.

[12] A. Moreno, B. Lindberg, C. Draxler, G. Richard, K. Choukri, S. Euler, and J. Allen, "SpeechDat-Car: A Large Database for Automotive Environments," in *Proc. of LREC*, Athens, Greece, May 2000.