AN HMM-BASED ARTIFICIAL BANDWIDTH EXTENSION EVALUATED BY CROSS-LANGUAGE TRAINING AND TEST

Patrick Bauer, Tim Fingscheidt

TU Braunschweig, Institute for Communications Technology,
Schleinitzstr. 22, D – 38106 Braunschweig, Germany
{p.bauer,t.fingscheidt}@tu-bs.de

ABSTRACT

Artificial bandwidth extension techniques can be employed in mobile terminals to improve the quality of the far-end speaker's signal at the receiver. To accomplish this, usually statistical models are trained requiring wideband speech material from a language that is expected to be used in the conversation. In practice however, the language of a certain phone conversation is not known to the user equipment. Therefore we investigated the performance of an HMM-based multilingually trained artificial bandwidth extension on speech signals of which the language was unseen in training. The cross-language training and test turned out to cause only minor degradations compared to the use of monolingually trained acoustic models of the language used in test. Our findings indicate that artificial bandwidth extension can be efficiently trained with multilingual speech data without significant losses in speech quality.

Index Terms— speech enhancement, artificial bandwidth extension

1. INTRODUCTION

Artificial bandwidth extension (ABWE) in general performs speech enhancement by upsampling of narrowband (telephony) speech and estimating further frequency components of interest. Speech enhancement systems in mobile phones usually improve the quality of the near-end speech signal that is perceived by the far-end conversation partner.

There are, however, obstacles to face before artificial bandwidth extension techniques can be widely employed in phone terminals. One is the often observed high-frequency whistling and lisping effect of artificially bandwidth extended speech as tackled, e.g., in [1, 2]. Especially fricatives such as /s/, /z/, /f/, and partly /S/, /Z/ are difficult to be estimated based upon only a narrowband speech signal, because a considerable portion of their energy is located in higher frequency components. This effect is particularly observed if the narrowband speech signal was bandlimited to an upper cutoff frequency of 3.4 kHz as it is the case in landline telephony. Speaker-independent training of acoustic models used in the ABWE also tends to increase the lisping effect.

A further obstacle is the language. The ABWE acoustic models and classification schemes are usually trained in a particular language one expects the system to work with. Even most of the recently proposed systems such as [3–5] do not explicitly address an operation in more than one language. For a phone application, the language however cannot be deducted simply from the user interface language the user has selected. A (reliable) language identification on the speech signal of a phone conversation appears to be a somewhat too massive solution in terms of computational power to be implemented in a phone terminal. In [6] a system has been proposed along with test results in 3 languages, however, no test results have been included for the language the classification scheme was optimized for.

In this paper we present an HMM-based artificial bandwidth extension technique whose acoustic models can be trained with wideband speech data of any language. For a total of 4 European languages we perform monolingual training and test (speaker-dependent and speaker-independent), and we also test the use of multilingual training speech data excluding the test language (i.e., a *cross-language* test).

In section 2 we present the basics of our ABWE approach. Section 3 discusses the experimental setup and our simulation results for the crosslingual training case as compared to the monolingual training case. Also examples of spectral distortion over time are given for the simulated cases, which allow a deeper analysis of the effects observed. Finally in section 4 conclusions are drawn.

2. THE HMM-BASED ABWE SYSTEM

The artificial bandwidth extension scheme we propose in Fig. 1 employs an HMM-based statistical model similar to [3]. A brief overview and some specifics of the system will be given in the following. The narrowband ($f_\mathrm{s} = 8$ 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.

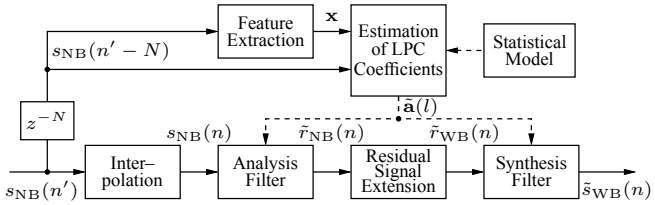

Fig. 1. Block diagram of the artificial bandwidth extension.

2.1. Analysis, Extension, and Synthesis

The actual processing of the interpolated speech signal consists of a mere (wideband) LP analysis filter, some residual signal extension technique, and final LP synthesis filtering with the very same coefficients as they were used in the analysis filter. Due to the time-varying sets of wideband LPC coefficients it turned out that it is highly recommended to employ a direct form I filter implementation as opposed to the commonly used transposed direct form II filter structure. The reason is that with the direct form I filter structure either the input (for the analysis filter) or the output signals (for the synthesis filter) are fed into a delay line without any modification. Multiplication with coefficients is then performed on the pure input/output signals in the delay lines. A switch of coefficients of course has impact on the next output sample, but is based on all but one old input/output samples in the delay line. This smoothes the output waveform of the algorithm in case of coefficient transitions.

In the residual signal extension block energy is allocated in the upper band at frequencies beyond 4 kHz. As was shown already in [7] and recently confirmed in [8], preserving pitch harmonics is of no perceptual importance beyond 4 kHz, therefore we simply use spectral folding (setting every other signal sample to zero). 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 lower band of the resulting estimated wideband speech signal $\tilde{s}_{\mathrm{WB}}(n)$. In the following we describe how the wideband LPC coefficients $a_1, ..., a_{16}$ are estimated from the narrowband speech signal.

2.2. Feature Extraction

Unlike the approach in [3], the feature extraction operates directly on the narrowband speech signal sampled at 8 kHz. It has a frame length of 15 ms and a frame shift of 10 ms, accordingly the wideband LPC 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 [9]. An LDA is employed to reduce the dimensionality of the primary feature vector from 15 to 5. The resulting feature vector $\mathbf{x}$ is subject to a statistical model.

2.3. Statistical Model Training

Using wideband speech training material, a statistical model is trained. As a first step a vector quantizer (VQ) codebook of upper band cepstral coefficients is obtained by selective linear prediction, computation of cepstral coefficients, and LBG training. This VQ is then framewise applied to the wideband training database and yields a certain state $S_i(l)$ 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, state probabilities $\mathrm{P}(S_i)$ and state transition probabilities $\mathrm{P}(S_i(l)|S_j(l-1))$ are trained and stored for the ABWE system. Finally, using the LDA transformed feature vectors $\mathbf{x}$, the parameters of a GMM-based observation probability density function (pdf) $\mathrm{p}(\mathbf{x}|S_i)$ are derived from EM training: a scalar weighting factor, a mean vector and a covariance matrix of every 5-dimensional multivariate normal distribution. For each of the 16 HMM states S_i a separately trained GMM of 8 mixtures is used.

2.4. Estimation of LPC Coefficients

Assuming a certain state S_i of the HMM model, the observation pdf $\mathrm{p}(\mathbf{x}|S_i)$ for the known feature vector $\mathbf{x}$ can be computed from the GMMs. In order to compute state a posteriori probabilities for frame l in a recursive fashion, the trained state (transition) probabilities are combined with the observation pdf following

$$\mathrm{P}(S_i(l)|\mathbf{X}(l)) = C \cdot \mathrm{p}(\mathbf{x}(l)|S_i(l)) \cdot \sum_{j=1}^{16} \mathrm{P}(S_i(l)|S_j(l-1)) \cdot \mathrm{P}(S_j(l-1)|\mathbf{X}(l-1)),$$

with $\mathbf{X}(l) = \{\mathbf{x}(l), \mathbf{x}(l-1), \ldots\}$. The factor C just normalizes the sum of the a posteriori probabilities over all states to one. Using the vector quantizer (VQ) codebook of upper band cepstral coefficient vectors $\mathbf{c}_i$ as obtained during training, the a posteriori probabilities are utilized to perform an MMSE estimation of the upper frequency band cepstral coefficients. By assembling the respective power spectra of the lower and higher frequency band, the estimated wideband spectrum is finally converted via the Levinson-Durbin recursion into the required LPC coefficient set $\tilde{\mathbf{a}}(l)$.

3. EXPERIMENTAL SETUP AND RESULTS

We performed experiments in three ABWE scenarios investigating to which extent the performance appears to be data-dependent concerning the speaker or language. The first scenario comprises speaker-dependent (SD) training and test data, while the remaining ones include speaker-independent monolingual (SI) and crosslingual (CL) data, respectively.

	SD	SI	CL
mean LSD [dB]	2.3	2.54	2.52
5...10 dB outliers [%]	8.65	12.03	11.88
>10 dB outliers [%]	1.0	1.43	1.44

Table 1. Total results of log-spectral distortion (LSD) over all 4 languages: SD (speaker-dependent), SI (speaker-independent), and CL (cross-lingual) training and test.

In all cases the required set of test signals is left out for training purposes (leave-one-out method), so that SD excludes the current test signal, SI the current speaker, and CL the current language. The speaker-independent experiments shall thereby represent quite demanding but realistic ABWE applications with unknown speakers and/or with speakers of a language unseen in training.

We used the NTT wideband speech database with the languages German (DE), British English (UK), French (FR), and Spanish (ES). The training data available amounts to approximately 70s for SD, 9min for SI and $\frac{1}{2}$h for CL experiments. The number of 384 test signals is equal in all cases. All four languages are covered, each with four male and female speakers, respectively, and with 12 utterances of 8s duration per speaker. Appropriate narrowband signals are achieved by high-quality sample rate conversion with cutoff frequency 3.8 kHz.

The wideband log-spectral distortion (LSD)

$$d_{\mathrm{LSD}} = \sqrt{2\left(\tfrac{10}{\ln 10}\right)^2 \cdot \sum_{d=1}^{64} (c_{\mathrm{WB},d} - \tilde{c}_{\mathrm{WB},d})^2}$$

serves as our performance evaluation measure with $c_{\mathrm{WB},d}$ and $\tilde{c}_{\mathrm{WB},d}$ being the cepstral coefficients of the original wideband 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 to 10 dB ($d_{\mathrm{LSD},5-10}$) and beyond 10 dB ($d_{\mathrm{LSD},>10}$). These ranges were found to reasonably document the speech quality in the context of artificial bandwidth extension from 8 to 16 kHz sampled speech.

Table 1 provides the total LSD results of the SD, SI and CL experiments. It turns out that the speaker-dependent ABWE performance is significantly better than the speaker-independent one. However, it should be noted that this advantage of SD training versus SI training is not consistent over all speakers: There are speakers gaining a lot from SD training, while others perform just as good as in the SI case. As somewhat surprising we found that the crosslingual (CL) performance of the ABWE scheme is not worse than the monolingual one (SI). Taking these results it can be concluded for speaker-independent ABWE scenarios that — at least for languages related to each other — crosslingual training and test does not cause a loss in speech quality.

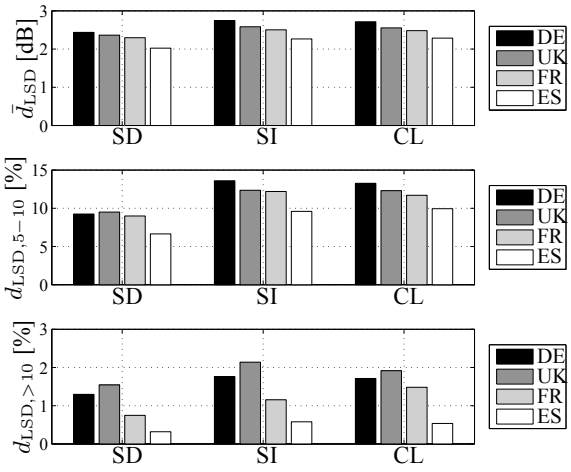

Fig. 2. Figures from top to bottom: Results for (a) mean log-spectral distortion $\overline{d}_{\mathrm{LSD}}$ [dB], (b) LSD outliers in the 5...10 dB range [%], (c) LSD outliers beyond 10 dB [%].

In general, these results are confirmed by Fig. 2 which displays the LSD results for each single language. As can be seen however, there are quite significant differences between the languages. In the SI test condition the LSD performance of German and English bandwidth-extended signals is somewhat worse than that of French or Spanish ones. German ABWE shows the worst mean LSD, while UK English ABWE obviously generates more LSD outliers beyond 10 dB, which are usually the really perceivable ones. French figures are better than both English and German, and Spanish ABWE performance is best. This result is interesting in so far, as it resembles automatic speech recognition performance reported in these languages.

The bottom diagram of Fig. 2 shows some further interesting details: While the SI and the CL performance of German and Spanish are quite the same, French is worse when trained in a crosslingual fashion, while English is even better. An explanation could be that English takes profit from being related to the other three European languages and in that sense from the larger amount of fitting training data in the CL case. In contrary, a French ABWE trained by English, German, and Spanish produces in CL simulations a bit more sounds that may not really be part of the French language than in the SI case. This impression is further supported from informative auditive listening tests.

In its upper four plots Fig. 3 depicts the spectrograms of an original wideband speech signal and all corresponding bandwidth-extended versions (SD, SI and CL). The signal displayed is the German utterance *"Einen Apfel"*. Obviously most times there is quite a high similarity between the original and the speaker-dependent bandwidth-extended signal. Some differences become visible concerning the speaker-independent spectrograms. However, among the

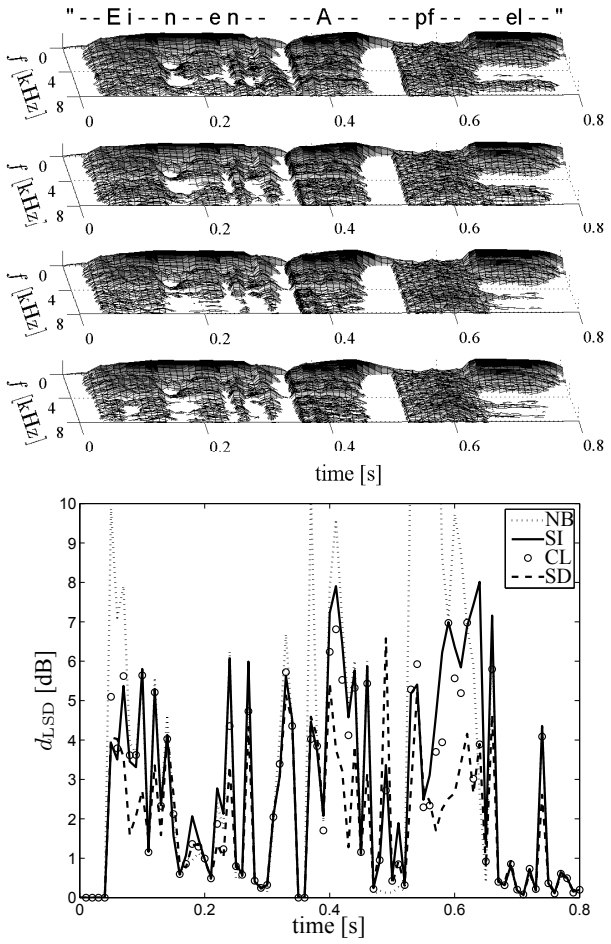

Fig. 3. Figures from top to bottom: Spectrograms for (a) 16 kHz original speech, (b) bandwidth-extended speech (SD) (c) bandwidth-extended speech (SI), (d) bandwidth-extended speech (CL). The utterance spoken was *"Einen Apfel"*. Bottom plot: (e) Log-spectral distortion [dB] over time for the narrowband signal (dotted curve), and the artificially bandwidth-extended signals with SI technique (solid), CL (circle markers), and SD (dashed).

speaker-independent schemes, the crosslingual results again keep up with the monolingual ones. This is confirmed by the corresponding LSD curves over time as drawn in the bottom plot of Fig. 3. While the SD technique in most cases represents a lower bound, the SI scheme several times yields a higher log-spectral distortion than the CL case. As expected the SD and SI techniques differ considerably in quality around time 0.6 s during the rather critical phonems /p/ (plosive) and /f/ (fricative) in the word *"Apfel"*. Even here the CL results perform slightly better than the SI ones. The additional dotted curve of the narrowband LSD (NB) demonstrates that the ABWE significantly improves the band-limited speech quality in terms of log-spectral distortion.

4. CONCLUSIONS

In this paper we have addressed the problem of enhancing speech quality by artificial bandwidth extension with speaker-independent monolingual and crosslingual training and test data. An HMM-based technique has been presented that was shown to provide comparable performance in both cases. A significant improvement versus the narrowband speech quality is reported by evaluation of log-spectral distances. Language-dependent characteristics of the ABWE performance were found providing a performance ranking as known from ASR techniques. Our findings indicate that artificial bandwidth extension can be efficiently trained and employed in a crosslanguage scenario which makes it useful for real-world telephony applications, where no knowledge about the phone conversation's language is available.

5. REFERENCES

[1] M. Nilsson and W.B. Kleijn, "Avoiding Over-Estimation in Bandwidth Extension of Telephony Speech," in *Proc. of ICASSP'01*, Salt Lake City, Utah, USA, May 2001, pp. 869–872.

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

[3] 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.

[4] 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.

[5] M.L. Seltzer, A. Acero, and J. Droppo, "Robust Bandwidth Extension of Noise-Corrupted Narrowband Speech," in *Proc. of INTERSPEECH'05*, Lisbon, Portugal, Sept. 2005, pp. 1509–1512.

[6] H. Pulakka, L. Laaksonen, and P. Alku, "Quality Improvement of Telephone Speech by Artificial Bandwidth Expansion – Listening Tests in Three Languages," in *Proc. of ICSLP'06*, Pittsburgh, Pennsylvania, Sept. 2006, pp. 1419–1422.

[7] H. Carl, *Untersuchung verschiedener Methoden der Sprachkodierung und eine Anwendung zur Bandbreitenvergrößerung von Schmalband-Sprachsignalen*, Ph.D. thesis, vol. 4 of U. Heute (ed.), Arbeiten über Digitale Signalverarbeitung, 1994.

[8] H. Pulakka, P. Alku, L. Laaksonen, and P. Valve, "The Effect of Highband Harmonic Structure in Artificial Expansion of Telephone Speech," in *Proc. of INTERSPEECH'07*, Antwerpen, Belgium, Aug. 2007.

[9] 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.