Speaker- and Language Dependency of Artificial Bandwidth Extension

Patrick Bauer, Tim Fingscheidt

TU Braunschweig, Institute for Communications Technology, Schleinitzstr. 22, 38106 Braunschweig, Germany

Email: {p.bauer,t.fingscheidt}@tu-bs.de

Abstract

Artificial bandwidth extension techniques can be employed in mobile terminals to improve the intelligibility and quality of the far-end speaker's speech signal at the receiver. To accomplish this, usually statistical models are trained requiring wideband speech material from the conversational partner, or at least from the language that is expected to be used in the conversation. In practice however, both, the speaker and language of a certain phone conversation are 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 speaker and language were 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. The experimental results further showed that both of these speaker-independent methods could even keep up with the speaker-dependent technique to a large extent. Our findings indicate that artificial bandwidth extension can be effciently trained with speaker- and language-independent speech data without significant losses in speech intelligibility and quality.

Introduction

Artificial bandwidth extension (ABWE) in general performs speech enhancement by upsampling of narrowband (telephony) speech and estimating further frequency components of interest.

There are, however, obstacles to face before ABWE techniques can be widely employed in phone terminals. One is the often observed high-frequency whistling and lisping effect as tackled, e.g., in [1]. 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.

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 [2, 3, 4] 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 [5] 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 the sequel we describe our experimental setup. Simulation results are then discussed for different ABWE scenarios concerning speaker- and language dependency. Measurement results of spectral distortion are given for the simulated cases, which allow a deeper analysis of the effects observed. Finally the conclusions are drawn.

Experimental Setup

In this paper we evaluate the speaker- and language dependency of an artificial bandwidth extension technique that is proposed in [6]. It employs an HMM-based statistical model similar to [2].The ABWE acoustic models can be trained with wideband speech data of any language.

For a total of 4 European languages we perform 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. In all cases the required set of test signals is excluded during training (leave-one-out method), so that SD excludes the current test signal, SI the current test speaker, and CL the current test 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.

For the experiments 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) between bandwidth-extended signal and its original wideband speech reference serves as our performance evaluation measure. 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: Total results of log-spectral distortion (LSD) over all 4 languages: speaker-dependent (SD), speaker-independent (SI), and cross-lingual (CL) training and test.

	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

Simulation Results

Table 1 provides the total LSD results of the SD, SI and CL experiments. It turns out that the speaker-dependent ABWE performance is 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 really 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.

In general, these results are confirmed by Fig. 1 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. 1 shows some further interesting details: While the SI and the CL performance are similar (German, Spanish), French is worse when trained in a crosslingual fashion, whereas 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.

Conclusions

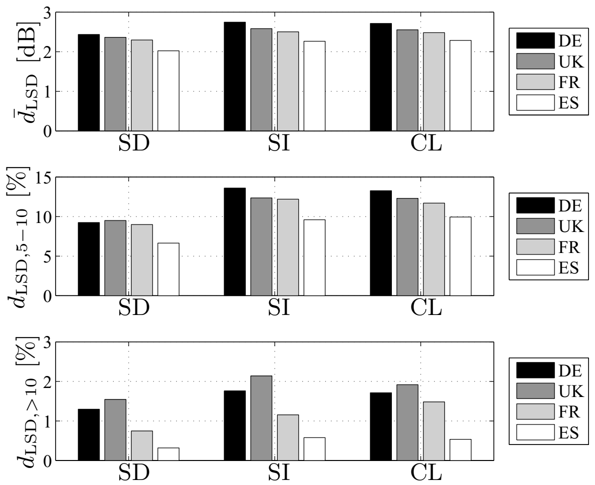

Figure 1: Figures from top to bottom: Results for (a) mean log-spectral distortion $\overline{d}_{\mathrm{LSD}}$ [dB], (b) LSD outliers in the 5...10 dB range [%], (c) LSD outliers beyond 10 dB [%].

In this paper we have evaluated speaker- and language dependency of artificial bandwidth extension. An HMM-based technique has been implemented that was shown to provide comparable performance in the speaker-independent monolingual and cross-lingual case. 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 conversational partner or language is available.

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] 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.

[3] 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.

[4] 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.

[5] 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.

[6] P.Bauer and T. Fingscheidt, "An HMM-Based Artificial Bandwidth Extension Evaluated by Cross-Language Training and Test," in *Proc. of ICASSP'08*, Las Vegas, Nevada, USA, Apr. 2008.