AUTOMATIC RECOGNITION OF WIDEBAND TELEPHONE SPEECH WITH LIMITED AMOUNT OF MATCHED TRAINING DATA

Patrick Bauer, Johannes Abel*, Volker Fischer†, and Tim Fingscheidt**

*Institute for Communications Technology, Technische Universität Braunschweig, Germany
†European Media Laboratory GmbH, Heidelberg, Germany

{bauer, abel, fingscheidt}@ifn.ing.tu-bs.de, volker.fischer@eml-d.villa-bosch.de

ABSTRACT

Automatic speech recognition (ASR) for wideband (WB) *telephone* speech services must cope with a lack of matching speech databases for acoustic model training. This paper investigates the impact of mixing insufficient WB and additional narrowband (NB) speech training data. It turns out that decimation and interpolation techniques, reducing the bandwidth mismatch between the NB speech material in training and the WB speech data to be recognized, do not succeed in outperforming the pure NB ASR baseline. However, true WB ASR training supported by artificial bandwidth extension (ABE) reveals a performance gain. A new ABE approach that makes use of robust dynamic features and a Viterbi path decoder exploiting phonetic *a priori* knowledge proves to be superior. It yields a reduction of $1.9\ \%$ word error rate relative to the NB ASR baseline and $9.3\ \%$ relative to a WB ASR experiment trained on only a limited amount of WB speech data.

***Index Terms*—** bandwidth extension, speech recognition

1. INTRODUCTION

Due to the use of hidden Markov models (HMMs), the increase of computing power, and the availability of speech databases, more and more demanding tasks for automatic speech recognition (ASR) have been tackled within the past decades [1]. Particularly the rising quantity of training data was found to be essential [2]. However, ASR performance strongly depends on the acoustic speech bandwidth. Empirical results in [1] show a relative word error rate (WER) reduction of $20\ \%$ by doubling the sampling rate from 8 to 16 kHz. Accordingly, dictation and other large-vocabulary applications principally use a sampling rate of at least 16 kHz, especially in noisy environments. In the context of ETSI Aurora standardization for noise-robust distributed speech recognition [3], [4] reports a relative ASR performance gain of $14\ \%$ averaged over a weak, medium, and high recording mismatch between training and test conditions, when recognizing noisy speech sampled at 16 instead of 8 kHz. In a similar investigation using the Spanish SpeechDat-Car corpus, an average relative improvement of $16\ \%$ is obtained in [5].

As most telephone calls are still narrowband (NB) providing a speech bandwidth < 4 kHz, telephone-based interactive voice response (IVR) systems are used to operate at 8 kHz sampling rate. However, upcoming speech services now offer a wideband (WB) frequency range of $0.05 \ldots 7$ kHz [6, 7]. The influence of a conventional telephony network on phoneme recognition performance was evaluated in [8, 9] using the N-TIMIT corpus [10]. The phoneme error rate (PER) was found to be degraded by $24\ \%$ and $33\ \%$, respectively, in relation to direct WB speech. In [11], we obtained a similar result of $23\ \%$ relative PER degradation in a NB mobile network derivative of the TIMIT corpus. Additionally, we investigated the impact of a WB mobile network using the WTIMIT corpus [12]. It revealed an increase of $19\ \%$ PER rel. to direct WB speech and a reduction of $3\ \%$ PER rel. to the NB mobile network. However, such telephony network effects still remain to be validated on large-vocabulary instead of limited phoneme recognition tasks, particularly regarding telephone-based IVR systems in practice.

Since WB telephone speech databases are rarely available, WB telephony ASR systems must cope with a limited amount of matched training data. Some approaches make use of available NB speech material by compensating for the bandwidth mismatch via recognizer-specific solutions. On the one hand, [13] employs a WB-to-NB feature transform via maximum likelihood linear regression (MLLR) and [14] a modified Viterbi search driven by missing filterbank components. On the other hand, bandwidth extension techniques in the feature or acoustic model domain are given in [15, 16]. To be independent from the recognizer, [17] evaluates an artificial bandwidth extension (ABE) of *speech* on a SPEECON Car City subset, but without making use of matched data for WB acoustic model training. This paper investigates an ASR system based on [18] taking into account a limited amount of WB and more available NB speech training data from the German Verbmobil corpus [19]. Different strategies re-using the NB speech material are presented to improve the ASR performance. A new ABE approach further developed from [20] integrating robust dynamic features and a phonetically-driven optimal state sequence decoder turns out to be superior.

This paper is structured as follows: Sec. 2 proposes several ABE versions to extend NB speech training data. Sec. 3 introduces the ASR system used for the experiments defined in Sec. 4. After having discussed the results in Sec. 5, Sec. 6 draws the conclusions.

2. ABE VERSIONS

The employed ABE algorithm is based on the statistical framework exploiting phonetic *a priori* knowledge introduced in [20, Sec. 3] and has been further modified to optimize performance particularly on critical fricatives /s,z/. As depicted in Fig. 1, it is divided into a main ABE processing path at the bottom, an ABE estimation part in the middle, and an access to pre-trained ABE parameters at the top.

The NB input speech samples $s_{\mathrm{NB}}(n')$ with index n' at 8 kHz sampling rate are subject to interpolation of factor two, linear prediction (LP) analysis filtering to remove the shape of the vocal tract, spectral folding to fill the upper frequency band, and LP synthesis filtering to form the vocal tract yielding the estimated WB output speech samples $\hat{s}_{\mathrm{WB}}(n)$ with index n at 16 kHz sampling rate.

This work was funded by the German Research Foundation (DFG) under grant no. FI 1494/4-1.

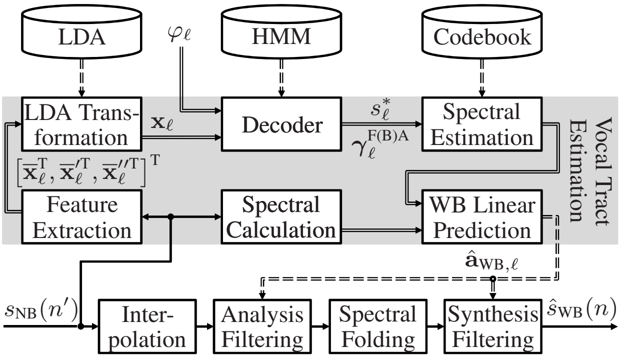

Fig. 1. ABE algorithm exploiting phonetic *a priori* knowledge[1].

To estimate the required WB LP filter coefficients $\hat{\mathbf{a}}_{\mathrm{WB},\ell}$ modeling the (inverse) vocal tract in frame $\ell = \lfloor n'/80 \rfloor = \lfloor n/160 \rfloor$, a 15-dimensional static feature vector $\overline{\mathbf{x}}_\ell$ is extracted from $s_{\mathrm{NB}}(n')$. In contrast to [20], it is augmented with first and second order dynamic feature vectors $\overline{\mathbf{x}}'_\ell$ and $\overline{\mathbf{x}}''_\ell$, respectively, that are derived via robust regression formulas spending five frames lookahead [21, Sec. 5.9]

$$\overline{\mathbf{x}}'_\ell = \frac{1}{28}\sum_{\lambda=1}^{3}\lambda(\overline{\mathbf{x}}_{\ell+\lambda} - \overline{\mathbf{x}}_{\ell-\lambda}), \quad \overline{\mathbf{x}}''_\ell = \frac{1}{10}\sum_{\lambda=1}^{2}\lambda(\overline{\mathbf{x}}'_{\ell+\lambda} - \overline{\mathbf{x}}'_{\ell-\lambda}). \quad (1)$$

The combined feature vector is transformed via a linear discriminant analysis (LDA) matrix into a 10-dimensional feature vector $\mathbf{x}_\ell$ serving as input of a first-order HMM. Following [20, Sec. 4.1.2], the HMM states $s_\ell = i$ are specified by $N = 24$ codebook entries $i = 1, ..., N$ representing upper-band spectral envelopes. They are trained on two phoneme classes: The first one ($i = 1, ..., 16$) represents all phonemes except for /s,z/, whereas the second one ($i = 17, ..., N$) is dedicated to /s,z/ only. This assignment is defined during ABE training by a mapping function $f(s_\ell = i) = \varphi_\ell$ that uniquely relates HMM states to phoneme class labels φ_ℓ. In offline ABE applications, such phonetic *a priori* knowledge is or can be made available, e.g., by manual phonetic transcription or forced Viterbi alignment. As further development from [20], a forward-backward algorithm (FBA) is employed for optimal state decoding. To access φ_ℓ, the forward and backward recursions (2)-(3) [22, Sec. 6.4.1] need to be reformulated taking into account the Gaussian mixture model (GMM)-based observation likelihood $b_i(\mathbf{x}_\ell, \varphi_\ell) = \mathrm{p}(\mathbf{x}_\ell, \varphi_\ell | s_\ell = i)$ as well as the state transition probability $a_{j,i} = \mathrm{P}(s_\ell = i | s_{\ell-1} = j)$ and the initial state probability $\pi_i = \mathrm{P}(s_1 = i)$:

$$\alpha_\ell(i) = b_i(\mathbf{x}_\ell, \varphi_\ell)\sum_{j=1}^{N} a_{j,i}\alpha_{\ell-1}(j), \quad \alpha_1(i) = \pi_i b_i(\mathbf{x}_1, \varphi_1), \quad (2)$$

$$\beta_\ell(i) = \sum_{j=1}^{N} b_j(\mathbf{x}_{\ell+1}, \varphi_{\ell+1}) a_{i,j}\beta_{\ell+1}(j), \quad \beta_L(i) = 1. \quad (3)$$

A posteriori probabilities $\gamma^{\mathrm{F(B)A}}_\ell$ are either computed by the complete FBA requiring all frames L (4) or just a forward algorithm (FA) (5):

$$\gamma^{\mathrm{FBA}}_\ell(i) = \mathrm{P}(s_\ell = i | \mathbf{x}^L_1, \varphi^L_1) = \frac{\alpha_\ell(i)\beta_\ell(i)}{\sum_{\nu=1}^{N}\alpha_\ell(\nu)\beta_\ell(\nu)}, \quad (4)$$

$$\gamma^{\mathrm{FA}}_\ell(i) = \mathrm{P}(s_\ell = i | \mathbf{x}^\ell_1, \varphi^\ell_1) = \frac{\alpha_\ell(i)}{\sum_{\nu=1}^{N}\alpha_\ell(\nu)}. \quad (5)$$

As an alternative to the FBA, a Viterbi path estimator is also employed [22, Sec. 6.4.2] by computing first a score $\delta_\ell(i)$ and a backtracking pointer $\psi_\ell(i)$ over the entire utterance for all states i:

$$\delta_\ell(i) = \max_j[\delta_{\ell-1}(j)a_{j,i}]b_i(\mathbf{x}_\ell, \varphi_\ell), \quad \delta_1(i) = \pi_i b_i(\mathbf{x}_1, \varphi_1), \quad (6)$$

$$\psi_\ell(i) = \arg\max_j[\delta_{\ell-1}(j)a_{j,i}]. \quad (7)$$

Backtracking is finally applied to decode the optimal state sequence

$$s^*_\ell = \psi_{\ell+1}(s^*_{\ell+1}), \ s^*_L = \arg\max_i[\delta_L(i)]. \quad (8)$$

The observation likelihood $b_i(\mathbf{x}_\ell, \varphi_\ell) = b_i(\mathbf{x}_\ell)\cdot\mathrm{P}(\varphi_\ell | s_\ell = i)$ can be split into a likelihood term $b_i(\mathbf{x}_\ell)$ that depends on the feature observation only and $\mathrm{P}(\varphi_\ell | s_\ell = i)$ denoting the elements of a $2 \times N$-dimensional phoneme class probability matrix defined as

$$\mathrm{P}(\varphi_\ell | s_\ell = i) = \begin{cases} 1-\epsilon, & \text{if } f(s_\ell = i) = \varphi_\ell \\ \epsilon, & \text{else} \end{cases} \quad \text{with} \ \ 0 < \epsilon \leq \frac{1}{2}, \quad (9)$$

which fulfills the normalization constraint $\sum_\varphi \mathrm{P}(\varphi_\ell | s_\ell = i) = 1$. After HMM state decoding, the upper-band spectral envelope is estimated using the codebook entries and assembled with the calculated NB spectrum to linearly predict the WB LP filter coefficients $\hat{\mathbf{a}}_{\mathrm{WB},\ell}$.

Please note that among the following ABE versions (a) - (e) used for the ASR experiments in Sec. 4.3, phonetic *a priori* knowledge is exclusively exploited in version (d) by choosing $\epsilon \neq \frac{1}{2}$:

- **(a)** ABE version using robust dynamic features (1) and the FA (2), (5) without exploiting phonetic information (i.e., $\epsilon = \frac{1}{2}$),
- **(b)** version (a), but using the complete FBA (2)-(4),
- **(c)** version (a), but using the Viterbi path decoder (6)-(8),
- **(d)** version (c), but exploiting phonetic information with $\epsilon = \frac{1}{6}$,
- **(e)** cheat version with the original upper frequency band >4 kHz.

3. ASR SYSTEM

Our framework for acoustic model training and recognition relies on the RWTH Aachen University open source speech recognition toolkit [18]. The ASR components relevant for the experiments described in Sec. 4.3 are briefly reviewed in the following.

3.1. Signal Processing Front End

The acoustic front end uses a Hamming window of 25 ms length and 10 ms frame shift to extract Mel-frequency cepstral coefficient (MFCC) features from the speech signal [23]. The MFCCs are subject to utterance-based cepstral mean normalization. For 16 kHz-sampled speech signals 16 MFCCs are computed, whereas the number of MFCCs is reduced to 12 for speech signals sampled at 8 kHz. High-frequency components of 8 kHz-sampled speech signals are slightly accentuated by applying a first order finite impulse response pre-emphasis filter $y(n) = x(n) - 0.97 \cdot x(n-1)$ prior to the filter bank, with $x(n)$ and $y(n)$ being the input and output samples of index n, respectively. Temporal dynamics are captured by concatenating nine consecutive feature vectors centered around the current frame. An LDA transformation is finally applied to reduce the dimension to 45.

3.2. Acoustic Model Training

Acoustic model training largely follows the recipes outlined in [18]. Linear segmentation of the signal (also known as flat start) is applied for the initialization of context independent, single-state HMMs. Two cycles of decision tree training and LDA are used for the creation of a triphone HMM inventory in the plain MFCC feature space. The HMM state emission probabilities are modeled by GMMs with

[1] Note that single/double lines denote a sample-/frame-wise processing.

a globally pooled, diagonal covariance matrix. The well-known expectation maximization algorithm is used to iteratively train the GMM parameters. The acoustic resolution of the model is increased by a splitting step that creates small perturbations of the mean vectors and is typically performed after 3 to 6 training iterations. The HMM state transition probabilities are fixed depending on the type of the transition (i.e., loop, forward, or skip).

3.3. Recognition System

Acoustic models are used with a variant of the speech recognizer, which employs stochastic n-gram language models and a pronunciation lexicon organized as a prefix tree in a time-synchronous beam search algorithm [24]. Both acoustic and language models have been trained on a subset of the German Verbmobil corpus (see Sec. 4.2 for details), whereas the recognition lexicon has been created on the entire corpus. After some preliminary experiments for parametrization purposes not reported in detail here, the weighting exponent of the language model term has been decided and was kept fixed. Other key characteristics of the recognition system are as follows:

- 6000 classification and regression tree (CART)-tied, context-dependent triphone HMMs,
- roughly 700,000 densities of dimension 45 with a globally pooled, diagonal covariance matrix,
- a trigram language model with a total of 100,000 n-grams trained via modified Kneser-Ney smoothing,
- a lexicon with about 11,400 words and 12,500 pronunciations.

4. EXPERIMENTAL SETUP

After having explained the preprocessing of simulated mobile telephone speech conditions in Sec. 4.1, Sec. 4.2 defines several data sets being required for the ASR experiments designed in Sec. 4.3.

4.1. Preprocessing of Speech Data

The single steps of preprocessing are carefully selected in order to simulate realistic speech conditions for mobile telephony IVR systems. Please note that *real* telephone speech data being transmitted over both NB and WB (mobile) telephony networks is hardly available until now[1]. Based on originally 16 kHz-sampled speech data, the preprocessing creates NB, ABE and WB speech conditions.

The NB condition is obtained by bandpass filtering to a range of about $0.2 \ldots 3.6$ kHz via the MSIN highpass filter, as specified by the ITU-T in [25], and a flat lowpass filter slightly adapting the FLAT1 filter in [25]. After decimation to a sampling rate of 8 kHz, which involves a lowpass filter for anti-aliasing and subsequent downsampling, the adaptive multirate narrowband (AMR-NB) speech codec [26] is applied at bitrate 12.2 kbps. The resulting NB condition serves as input for the ABE versions (a) - (e) in Sec. 2 to create five ABE conditions. In contrast, the WB condition is derived from the original, 16 kHz-sampled speech data by transmitter-sided P.341 filtering to a range of about $0.05 \ldots 7.0$ kHz according to [25, 27] and transcoding via the adaptive multirate wideband (AMR-WB) speech codec [28] at bitrate 12.65 kbps. The preprocessed speech conditions are denoted as ***NB***, ***ABEa-e*** and ***WB*** in the following. Please note that some of the speech conditions used for the ASR experiments in Sec. 4.3 furthermore require another decimation ($\downarrow 2$) or interpolation ($\uparrow 2$). Corresponding to decimation, interpolation involves upsampling and subsequent lowpass filtering.

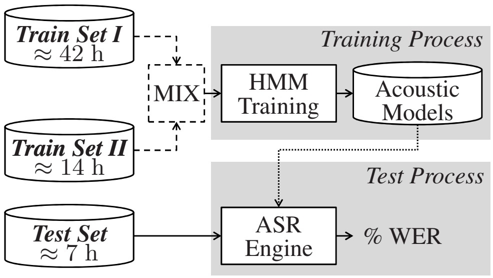

Fig. 2. Processing scheme and data sets of the ASR experiments.

ID	***Train Set I***	***Train Set II***	***Test Set***	f_s [kHz]
①	***WB***	***WB***	***WB***	16
②	***NB***	***NB***	***NB***	8
③	–	***WB***	***WB***	16
④	***NB***	***WB***$\downarrow 2$	***WB***$\downarrow 2$	8
⑤	***NB***$\uparrow 2$	***WB***	***WB***	16
⑥a - ⑥e	***ABEa-e***	***WB***	***WB***	16

Table 1. *Assignment of preprocessed data sets to ASR experiments.*

4.2. Definition of Database Subsets

The speech data for our experiments originates from the German part of the Verbmobil spontaneous speech corpus [19] that we consider to be large enough for the purpose of this study. The speech recording was done over close-talk and room microphones, as well as a telephone. Since the preprocessing in Sec. 4.1 requires WB speech, we only selected the close-talk and room microphone recordings sampled at 16 kHz. In total we used about 70 hours of German WB speech data divided in four speaker-disjoint, gender- and age-balanced subsets of different size.

The first two subsets ***Train Set I*** and ***Train Set II*** have portions of about 42 hours (i.e., 60 %) and 14 hours (i.e., 20 %), respectively. They contain speech data to train the HMMs in our experiments. We chose different data sizes to take into account the fact that for the acoustic model training of telephone-based IVR systems there is usually available more NB than WB speech material. In this use case, the smaller subset ***Train Set II*** provides WB speech, while the larger ***Train Set I*** only includes NB speech. Those ASR experiments in Sec. 4.3 taking into account ABE therefore extend ***Train Set I*** to create additional speech data for WB acoustic model training. Based on the given data volume of 70 hours, we assume that a ratio of 3:1 between NB and WB HMM training speech data is a good show case to demonstrate realistic effects of imbalanced data sets in practice. The two remaining subsets both have portions of 7 hours (i.e., 10 % each). One of these subsets, denoted by ***Test Set***, is used for ASR evaluation, while the other one is dedicated to ABE training.

4.3. ASR Experiments

Based on the preprocessed speech conditions in Sec. 4.1 and the database subsets defined in Sec. 4.2, we designed ASR experiments of practical relevance. Fig. 2 depicts a generic block diagram of the ASR processing scheme. The subsets ***Train Set I*** and ***Train Set II***

[1] The WTIMIT corpus [12] seems to be the first published WB mobile telephone speech corpus, however, its data volume of about 5.5 hours is not sufficient for the large-vocabulary ASR investigations presented here.

ID	% WER	% WER relative to			
		①	②	③	④, ⑤
①	36.83	±0.0	**−6.9**	−14.0	−7.8
②	39.58	+7.5	±0.0	−7.6	−0.9
③	42.83	**+16.3**	**+8.2**	±0.0	+7.2
④	39.95	+8.5	**+0.9**	**−6.7**	±0.0
⑤	39.95	+8.5	**+0.9**	**−6.7**	±0.0
(6a)	39.30	+6.7	**−0.7**	−8.2	−1.6
(6b)	39.25	+6.6	**−0.8**	−8.4	−1.8
(6c)	39.07	+6.1	**−1.3**	−8.8	−2.2
(6d)	38.83	+5.4	**−1.9**	**−9.3**	**−2.8**
(6e)	38.05	**+3.3**	**−3.9**	**−11.2**	**−4.8**

Table 2. *Results of the ASR experiments in Table 1.*

are both devoted to the HMM training. They can either be mixed, which means that they are used in combination, or one of them is discarded. After having trained the acoustic models, the ASR evaluation takes place on the subset ***Test Set***. The evaluation results are reported in terms of % WER.

Tab. 1 assigns the preprocessed database subsets to consecutively numbered ASR experiments. The sampling rate f_s indicates that the acoustic front end either operates on 16 kHz- or 8 kHz-sampled speech data. Experiments ① and ② serve as a reference of the upper-bound ASR performance in the pure WB and NB case, respectively. Due to the lack of WB speech data for HMM training in practice, experiment ③ must do without the large ***Train Set I***. The remaining experiments are designed under three constraints:

- WB telephone speech is received to be recognized somehow,
- large ***Train Set I*** is only available as NB telephone speech,
- smaller ***Train Set II*** is available as WB telephone speech.

We consider these constraints as describing a situation of high practical relevance. Experiment ④ represents the simple case that incoming WB telephone speech is recognized by a NB ASR system, therefore, all employed WB speech data is decimated prior to training and recognition, respectively. In contrast, experiment ⑤ leaves the WB speech training and test data untouched, but interpolates the available NB speech material for training. In experiments (6a) - (6e) this interpolation is replaced by the ABE versions (a) - (e) defined in Sec. 2.

5. RESULTS AND DISCUSSION

The results of the performed ASR experiments are depicted in Tab. 2. The absolute WERs vary roughly between 37 % and 43 % indicating that the given ASR task is rather challenging[2]. Anyway, the goal of this contribution is not to achieve the lowest absolute WER, but to point out WER differences influenced by limitations of training data and speech bandwidth. Hence, the following discussion mainly focuses on relative WERs.

As expected, the WB baseline ① outperforms the NB baseline ② by 6.9 % WER relative, given full access to matched training data. A limitation of the WB baseline to 14 out of 56 hours (i.e., 25 %) of WB speech training data in ③, however, degrades the WER by 16.3 % relative to ① and 8.2 % relative to ②, respectively. It turns out that the lack of training data causes a huge drop in WB ASR performance. In order to solve this problem, the remaining experiments make additionally use of available NB speech data for training, in spite of the mismatch in speech bandwidth.

On the one hand, simple decimation prior to the recognition helps to reduce the bandwidth mismatch in ④, so that the recognizer consistently operates at 8 kHz sampling rate. Indeed, the WER decreases by 6.7 % relative to ③, however, it is still 0.9 % higher relative to the NB baseline ②, also due to codec mismatch. On the other hand, ⑤ does not require any decimation prior to the recognition, since the NB speech training data has been interpolated to 16 kHz matching the sampling rate of the recognizer. Surprisingly, ⑤ achieves the same results as ④, despite the bandwidth mismatch between the interpolated training data and the WB speech data to be recognized.

Instead of just interpolating training data as in ⑤, the ABE versions (a) - (e) are applied in experiments (6a) - (6e) to reduce the bandwidth mismatch. Obviously, all ABE versions are found to be beneficial, leading to a WER lower than in experiments ② to ⑤. According to Sec. 2, ABE estimation based on the complete FBA in (6b) turns out to perform somewhat better than just using the FA in (6a). For the first time, the NB baseline ② is slightly improved by 0.7 − 0.8 % WER relative. A further improvement is achieved by means of the Viterbi path decoder in (6c). It reveals a WER reduction of 1.3 % relative to the NB baseline ②. By additionally exploiting phonetic information for Viterbi path estimation in (6d), the WER is even reduced in relation to ② by 1.9 %. Please note that the cheat experiment (6e) demonstrates the upper-bound ASR performance for ABE by reconstructing the original upper frequency band > 4 kHz. Due to the degradation of the lower frequency band by means of a telephone bandpass filter and an AMR-NB speech codec according to Sec. 4.1, (6e) still exceeds the WER of the WB baseline ① by 3.3 % relative. However, the NB baseline ② is outperformed by 3.9 % WER relative. Assuming that the ASR performance for ABE ranges from a simple interpolation in ⑤ to a perfect reconstruction of the original upper frequency band in (6e), the potentially achievable relative WER gap is 4.8 %. The best ABE version (d) already bridges more than half of this gap by reducing the WER relative to ⑤ by 2.8 %. It furthermore reveals a significant WER improvement of 9.3 % in relation to the WB baseline with limited amount of matched training data in ③, which is close to the relative WER gain of the ABE upper-bound performance (6e) resulting in 11.2 %.

6. CONCLUSIONS

Telephone-based interactive voice response systems supporting wideband (WB) speech services severely suffer from the lack of WB telephone speech databases for acoustic model training. This paper investigates potential automatic speech recognition (ASR) designs re-using *unmatched* training data, i.e., available NB telephone speech material. Decimation of incoming WB telephone speech prior to the recognition – reducing the bandwidth mismatch to the additional narrowband (NB) speech training data – even leads to a lower ASR performance than the pure NB baseline. Interpolation of the NB speech training data for recognizing WB speech performs comparably, in spite of the bandwidth mismatch. An improved ASR performance is achieved by extending the available NB speech training data via artificial bandwidth extension (ABE). The use of a Viterbi path decoder exploiting phonetic information turns out to be beneficial for ABE. It reveals a substantial word error rate improvement of 1.9 % relative to the NB baseline (more than halfway toward an ABE performance upper bound), and even of 9.3 % relative to the WB baseline with limited amount of matched training data.

[2] Obviously, the absolute WER is much higher than, e.g., in [29, 30], however, the employed Verbmobil 1996 evaluation set with a total duration of 43 min is not directly comparable to our test set of 7 h length.

REFERENCES

[1] X. Huang, A. Acero, and H. W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*, Prentice Hall, 2001.

[2] K. W. Church and R. L. Mercer, "Introduction to the Special Issue on Computational Linguistics Using Large Corpora," *Computational Linguistics*, vol. 19, no. 1, pp. 1–24, Mar. 1993.

[3] "Speech Processing, Transmission and Quality Aspects (STQ); Distributed Speech Recognition; Advanced Front-End Feature Extraction Algorithm; Compression Algorithms," ETSI, Jan. 2007.

[4] D. Macho and Y. M. Cheng, "On the Use of Wideband Signal for Noise Robust ASR," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Hong Kong, Apr. 2003, vol. 2, pp. 109–112.

[5] C. Nadeu and M. Tolos, "Recognition Experiments with the SpeechDat-Car Aurora Spanish Database Using 8 kHz- and 16 kHz-Sampled Signals," in *Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding*, Madonna di Campiglio, Italy, Dec. 2001, pp. 135–138.

[6] S. Ferraz de Campos Neto and K. Järvinen, "Wideband Speech Coding Standards and Wireless Services [Guest Editorial]," *IEEE Communications Magazine*, vol. 44, no. 5, pp. 56–57, May 2006.

[7] T. Fingscheidt, "The Silent Speech Bandwidth Revolution in Mobile Telephony," IEEE Speech and Language Processing Technical Committee's Newsletter, Aug. 2012.

[8] P. J. Moreno and R. M. Stem, "Sources of Degradation of Speech Recognition in the Telephone Network," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Adelaide, Australia, Apr. 1994, vol. 1, pp. 109–112.

[9] B. Chigier, "Phonetic Classification on Wide-Band and Telephone Quality Speech," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Toronto, ON, Canada, May 1991, pp. 291–295.

[10] C. Jankowski, A. Kalyanwamy, S. Basson, and J. Spitz, "NTIMIT: A Phonetically Balanced, Continuous Speech, Telephone Bandwidth Speech Database," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Albuquerque, NM, U.S.A., Apr. 1990, vol. 1, pp. 109–112.

[11] P. Bauer, D. Scheler, and T. Fingscheidt, "WTIMIT: The TIMIT Speech Corpus Transmitted Over the 3G AMR Wideband Mobile Network," in *Proc. of ITG Conference on Speech Communication*, Bochum, Germany, Oct. 2010.

[12] P. Bauer and T. Fingscheidt, "WTIMIT 1.0," Linguistic Data Consortium, Philadelphia, 2010.

[13] M. Karafiat, L. Burget, J. Cernocky, and T. Hain, "Application of CMLLR in Narrow Band Wide Band Adapted Systems," in *Proc. of Annual Conference of the International Speech Communication Association*, Antwerp, Belgium, Aug. 2007, pp. 282–285.

[14] Y.-F. Liao, J.-S. Lin, and W.-H. Tsai, "Bandwidth Mismatch Compensation for Robust Speech Recognition," in *Proc. of European Conference on Speech Communication and Technology*, Geneva, Switzerland, Sept. 2003, pp. 3093–3096.

[15] M. L. Seltzer and A. Acero, "Training Wideband Acoustic Models using Mixed-Bandwidth Training Data via Feature Bandwidth Extension," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Philadelphia, PA, U.S.A., Mar. 2005, vol. 1, pp. 921–924.

[16] M. L. Seltzer and A. Acero, "Training Wideband Acoustic Models Using Mixed-Bandwidth Training Data for Speech Recognition," *IEEE Transactions on Audio Speech and Language Processing*, vol. 15, no. 1, pp. 235–245, Jan. 2007.

[17] P. Bauer, M.-A. Jung, and T. Fingscheidt, "Investigations on Offline Artificial Bandwidth Extension of Telephone Speech Databases," in *Proc. of ITG Conference on Speech Communication*, Bochum, Germany, Oct. 2010.

[18] D. Rybach, C. Gollan, G. Heigold, B. Hoffmeister, J. Lööf, R. Schlüter, and H. Ney, "The RWTH Aachen University Open Source Speech Recognition System," in *Proc. of Annual Conference of the International Speech Communication Association*, Brighton, U.K., Sept. 2009, pp. 2111–2114.

[19] W. Wahlster, Ed., *Verbmobil: Foundations of Speech-to-Speech Translation*, Springer, 2000.

[20] P. Bauer and T. Fingscheidt, "A Statistical Framework for Artificial Bandwidth Extension Exploiting Speech Waveform and Phonetic Transcription," in *Proc. of European Signal Processing Conference*, Glasgow, Scotland, Aug. 2009, pp. 1839–1843.

[21] University of Cambridge, "The HTK Book (for HTK Version 3.4)," http://htk.eng.cam.ac.uk/, Dec. 2006.

[22] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, 1993.

[23] L. Welling, N. Haberland, and H. Ney, "Acoustic Front-End Optimization for Large Vocabulary Speech Recognition," in *Proc. of European Conference on Speech Communication and Technology*, Rhodes, Greece, Sept. 1997, pp. 2099–2102.

[24] D. Nolden, H. Ney, and R. Schlüter, "Time Conditioned Search in Automatic Speech Recognition Reconsidered," in *Proc. of Annual Conference of the International Speech Communication Association*, Makuhari, Japan, Sept. 2010, pp. 234–237.

[25] "ITU-T Recommendation G.191, Software Tool Library 2009 User's manual," ITU, Nov. 2009.

[26] "Mandatory Speech Codec Speech Processing Functions: Adaptive Multi-Rate (AMR) Speech Codec; Transcoding Functions (3GPP TS 26.090)," 3GPP; TSG-SA, Aug. 1999.

[27] "ITU-T Recommendation P.341, Transmission Characteristics for Wideband Digital Loudspeaking and Hands-Free Telephony Terminals," ITU, Mar. 2011.

[28] "Speech Codec Speech Processing Functions: Adaptive Multi-Rate-Wideband (AMR-WB) Speech Codec; Transcoding Functions (3GPP TS 26.190)," 3GPP; TSG-SA, Mar. 2001.

[29] L. Welling, S. Kanthak, and H. Ney, "Improved Methods for Vocal Tract Normalization," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Phoenix, AZ, U.S.A., Mar. 1999, vol. 2, pp. 761–764.

[30] M. Finke, P. Geutner, H. Hild, T. Kemp, K. Ries, and M. Westphal, "The Karlsruhe-Verbmobil Speech Recognition Engine," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Munich, Germany, Apr. 1997, vol. 1, pp. 83–86.