ROBUST AUTOMATIC SPEECH RECOGNITION USING ACOUSTIC MODEL ADAPTATION PRIOR TO MISSING FEATURE RECONSTRUCTION

Ulpu Remes, Kalle J. Palomäki, and Mikko Kurimo

Adaptive Informatics Research Centre, Helsinki University of Technology
P.O. Box 5400, FIN-02015 HUT, email: firstname.lastname@hut.fi

ABSTRACT

When speech recognition is used in real-world environments, simultaneous speaker and environmental adaptation and compensation for time-varying noise effects is needed. Noise compensation methods like missing feature reconstruction should be combined with adaptation methods like constrained maximum likelihood linear regression (CMLLR). This is only straightforward if reconstruction is used prior to CMLLR. In this work, reconstruction is modified so that we can estimate CMLLR transformations prior to reconstruction. The new approach is evaluated on large vocabulary speech data recorded in noisy public and car environments and compared to using reconstruction prior to CMLLR estimation. The results suggest the noise environment determines which approach is better. Using adaptation prior to reconstruction has the better performance when evaluated on data from public environments. The relative reductions in letter error rate were 47–50 % compared to the baseline and 13–19 % compared to using either adaptation or reconstruction alone.

1. INTRODUCTION

Automatic speech recognition works reasonably well when the system is tested and trained under the same conditions, but speaker and environmental variation and environmental noise make the recognition task difficult. In this work, we discuss an important but insufficiently solved problem which arises when automatic speech recognition is applied in real-world environments: how to achieve simultaneous speaker and environmental adaptation and compensation for time-varying noise effects? Adaptation methods such as the constrained maximum likelihood linear regression (CMLLR) [1] are designed to improve statistical robustness in general rather than compensate for noise, so adaptation should be combined with a separate noise compensation method.

Methods intended specifically for noise compensation include the missing feature methods proposed in [2][3]. The missing feature methods are applied in compressed spectral domain, where noise corruption is typically such that some spectrotemporal regions are reliable and represent mostly speech, while some represent mostly noise. The missing feature methods 1) find the noise corrupted regions, where the speech informations is missing, and 2) handle speech recognition with missing valued. In missing feature reconstruction, the missing values are replaced with estimates calculated based on the reliable, speech dominated features and clean speech statistics [3]. Missing feature reconstruction methods have performed well under various noise conditions [3][4], but since noise robustness alone is often not enough, reconstruction needs to be combined with, for example, speaker compensation.

In previous work [5], we combined missing feature reconstruction and CMLLR for large vocabulary continuous speech recognition in noisy environments. When reconstruction was applied prior to CMLLR, using CMLLR improved the speech recognition performance modestly. In this work, we examine if the speech recognition performance further improves when CMLLR transformations are estimated prior to reconstruction (from the noisy features) rather than after reconstruction (from the reconstructed features). We propose to modify cluster-based missing feature reconstruction [3] so that CMLLR transformations can be estimated prior to reconstruction and their effect accounted for in missing feature reconstruction. This is referred to as using adaptation prior to reconstruction in the rest of this work. The proposed method is evaluated and compared with previous approaches on Finnish large vocabulary continuous speech data recorded in noisy public environments, such as parks and cafeterias, and inside cars.

2. METHODS

2.1 Baseline system

Our large vocabulary continuous speech recognition system uses a morph-based growing n-gram language model [6] which is trained on 145 million words of book and newspaper data. Since all words and word forms can be represented with the unsupervised morphs, the decoding vocabulary is in practise unlimited [7]. The decoder is a time-synchronous beam-pruned Viterbi token-pass system [8] and the acoustic models are state-clustered hidden Markov triphone models constructed with a decision-tree method [9]. Each state is modelled with 16 Gaussians, and the states are also associated with gamma probability functions to model the state durations [10].

The speech signal is represented with 12 MFCC and a log energy feature. Features are used with their first and second order differentials, and treated with cepstral mean subtraction (CMS) and maximum likelihood linear transformation (MLLT) [11] optimised in training. In addition, features are adapted with constrained maximum likelihood linear regression (CMLLR) [1] in order to reduce the mismatch between training and testing conditions (i.e. compensate for speaker and environmental variation). CMLLR is essentially a model adaptation method, but it can be formulated as a linear feature transformation. CMLLR transformations are estimated from the test data in an unsupervised manner.

2.2 Noise mask estimation

Noise that originates from sources uncorrelated with speech corrupts the speech signal additively in power spectral domain. Thus, in logarithmic mel-spectral domain, when speech dominates over noise, the time-frequency components $Y(\tau, i)$ may be considered as reliable estimates of the clean speech values $X(\tau, i)$ which would have been observed if the speech signal had not been corrupted with noise. In other words, for the reliable components, $X_r(\tau, i) \approx Y_r(\tau, i)$. The components in the noise dominated regions are, on the other hand, unreliable and provide only an upper bound for the corresponding clean speech values, $X_u(\tau, i) \leq Y_u(\tau, i)$. Labels dividing the noisy speech mel-spectrogram to reliable and unreliable parts are referred to as a spectrographic mask (Figure 1).

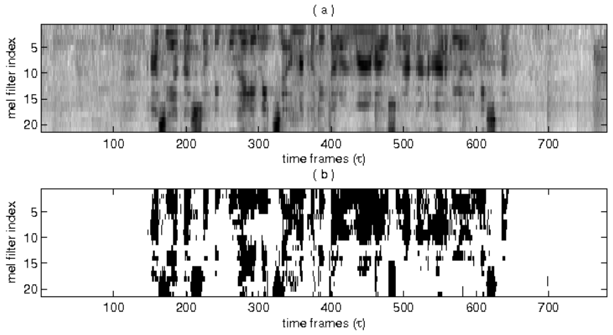

Figure 1: Logarithmic mel-spectrogram for noisy speech and an example spectrographic mask that divides the features to reliable (black) and unreliable (white) regions.

In this work, the spectrographic masks are constructed based on local signal-to-noise ratio (SNR) estimates. The estimates are derived from noise estimates that we calculate during speech pauses, which are detected using a Gaussian mixture based speech/non-speech classifier. The non-speech frames i.e. the frames $Y(\tau)$ that have been classified as non-speech are collected and temporally smoothed to produce the noise estimate $N(\tau)$. Now, time-frequency components are taken to be unreliable if their observed value $Y(\tau, i)$ does not exceed the estimated noise power $N(\tau, i)$ with minimum of γ dB. The speech/non-speech classifier and noise power estimation are described in our previous work [5]. The threshold $\gamma = 3$ dB which has been decided based on experiments with the far recorded parameter optimisation data (see Section 3 for dataset description).

2.3 Missing feature reconstruction

Recognising partially observed speech is possible if either (i) the classifier (decoder) is modified to marginalise over the missing values or (ii) data imputation is used and the features reconstructed. Classifier modification methods such as bounded marginalisation [2] have been efficient when tested with a limited vocabulary (e.g. in a connected digit recognition task) but using these methods limits the speech recogniser to operate on the log-spectral features. When data imputation is used instead, the unreliable components are replaced with estimates that correspond to clean speech. In this case, the reconstructed log-spectral features may be further processed as usual and the decoder needs not be modified (Figure 2). This is especially important in large vocabulary continuous speech recognition where the state-of-the-art systems typically use various normalisation and feature transformation methods.

In this work, we use the cluster-based feature reconstruction method proposed in [3] for data imputation. Here, the log-spectral clean speech features $X(\tau)$ are assumed independent and identically distributed according to a Gaussian mixture model (GMM)

$$P(X) = \sum_{\nu} c(\nu) N(X; \mu(\nu), \Sigma(\nu)) \tag{1}$$

where $c(\nu)$ are the component weights (prior probabilities), $\mu(\nu)$ the components means and $\Sigma(\nu)$ the covariance matrices. Reconstructed values for the unreliable feature components are chosen so that the reconstructed feature maximises the likelihood of the clean speech model but does not exceed the observed values Y_u. The bounded MAP estimate for the unreliable values is given as

$$\hat{X}_u = \arg\max_{X_u} P(X_u | X_r, X_u \le Y_u, \theta\,) \tag{2}$$

where θ are the GMM parameters from Equation (1). If the $\Sigma(\nu)$ here are diagonal, Equation (2) can be solved analytically. In this work, however, we use full covariance matrices, and to calculate the bounded MAP estimate, we iterate over the frequency channels i as proposed in [3]:

1. Initialise $\hat{X}(i) = Y(i) \;\forall\; i$
2. Repeat until $\hat{X}$ is converged: for each $X(k) \in X_u$

$$\hat{X}(k) = \arg\max_{X(k)} P(X(k) | X(i) = \hat{X}(i) \,\forall\, i \neq k, \theta\,) \tag{3}$$

$$\hat{X}(k) = \min\{\hat{X}(k), Y(k)\} \tag{4}$$

The clean speech model used in this work is a 5-component GMM trained with 96-minute extract from the SPEECON training set described in Section 3. The clusters and distribution parameters are jointly estimated using the expectation-maximisation (EM) algorithm in the GMMBAYES Matlab Toolbox [12].

2.4 Motivation

The missing feature methods could be a solution for speech recognition especially in changing and unpredictable noise conditions since they make minimal assumptions about the noise and do not utilise noise estimates. In this work, a simple noise estimate is used in spectrographic mask estimation, but if the mask was estimated based on e.g. perceptual criteria, as suggested in [13], noise estimation would be unnecessary. Common approaches to handling the missing features in speech recognition include missing feature reconstruction and bounded marginalisation. Bounded marginalisation as defined in [2] limits the recogniser to use log-spectral features and operate under the assumption that frequency channels are uncorrelated. As these limitations are not well-suited for HMM-based large vocabulary speech recognition, we choose to use reconstruction instead.

There is one problem particular to the reconstruction approach: while reconstructing the features reduces noise interference, it also produces artefacts in the observed features. Now, although the net effect from the reduced interference and increased artefacts remains positive on speech recognition performance, the effects on speaker and environmental adaptation may be privative. In addition to the increased artefacts, since reconstruction is carried out based on a low-complexity speaker-independent GMM, the process is likely to remove or smooth speaker-dependent characteristics. This affects the feature statistics used for CMLLR estimation and can degrade adaptation performance, as suggested in [5]. Therefore, it appears we should look for reversing the order between reconstruction and adaptation: adaptation prior to reconstruction.

2.5 Using adaptation prior to reconstruction

Adaptation was, in previous work [5], applied after reconstruction because missing feature methods operate in the log-spectral domain while adaptation is applied after the differantial features have been calculated and the features treated with DCT, CMS, and MLLT transformations as described in Section 2.1. This will be referred to as the acoustic model domain. If we wish to estimate the CMLLR transformations prior to reconstruction, the reconstruction process needs to be modified to take later adaptation into account. While the feature likelihoods are normally evaluated in the log-spectral domain as illustrated in Equation (2), the clean speech model in Equation (1) is now trained in the acoustic model domain. The noisy log-spectral features Y are transformed into the acoustic model domain and also treated with adaptation. Now, the bounded MAP estimates are solved iteratively from

$$\hat{X}_u = \arg\max_{X_u} P(\, A \cdot T(X) + b \mid X_r, X_u \le Y_u, \theta\,'\,) \tag{5}$$

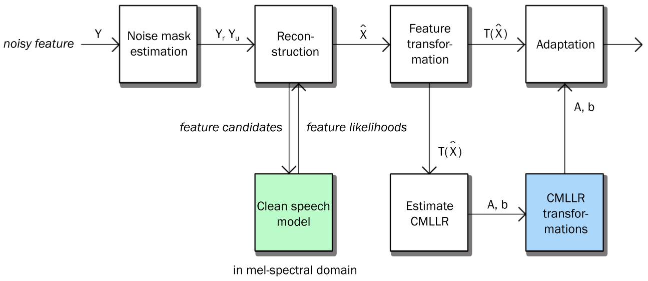

Figure 2: Feature extraction and missing feature reconstruction. After reconstruction in mel-spectral domain, the reconstructed features are processed like normal features.

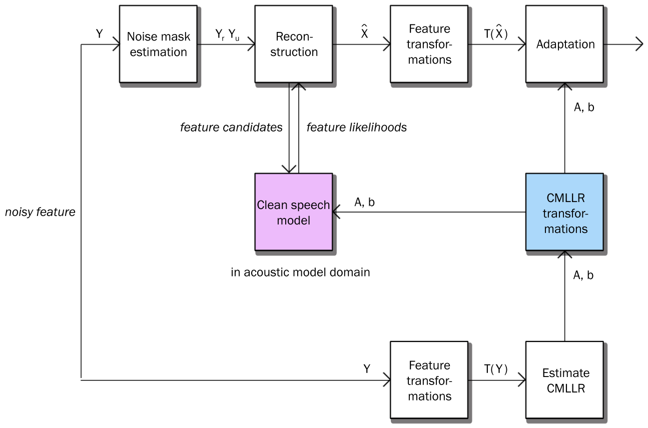

Figure 3: Feature extraction when CMLLR transformations are estimated prior to missing feature reconstruction and their effect on transformed features compensated for in the reconstruction process (adaptation prior to reconstruction).

where $X = X_r \cup X_u$ and $T(X)$ the transformation from log-spectral to acoustic model domain, matrix A and vector b are the speaker-dependent CMLLR transformation parameters, and θ' the GMM parameters estimated in the acoustic model domain. Note that the missing values are still infilled in the log-spectral domain although the optimal values are chosen based on likelihood scores calculated after adaptation in the acoustic model domain. Thus, reconstruction has changed from finding the optimal clean speech estimates to finding the optimal clean speech estimates when it is known that a specific linear transformation will be used for adaptation later in the process. With adaptation thus compensated for, the speaker-dependent transformations estimated prior to reconstruction are applicable to the reconstructed features as such, and the feature extraction process becomes as illustrated in Figure 3.

It should be noted that when the candidate reconstructed features are propagated to the acoustic model domain for likelihood evaluation, the CMS coefficients and first and second differentials are calculated from the noisy rather than reconstructed features. This is because both CMS and differential features at time τ are calculated based on features $X(\tau - T) \ldots X(\tau + T)$ i.e. over several frames, while reconstruction processes one frame at a time. After the features $X(\tau - T) \ldots X(\tau + T)$ have all been reconstructed in the mel-spectral domain, CMS and differential features for $X(\tau)$ are recalculated from the reconstructed features. Thus, the features used for speech recognition are completely recalculated based on the reconstructed mel-spectral features.

3. DATA

The acoustic model is trained with data selected from the Finnish SPEECON database [14]. The 26-hour training set contains clean speech recorded with a close-talk microphone (a headset) in quiet conditions. 208 speakers, including both male and female subjects, are represented in the data. Among the utterances used for training are words, sentences and free speech. The speech/non-speech classifier required for noise estimation, on the other hand, is trained with television news data from the Finnish Broadcasting Company (YLE) as described in [5] except that the data has been artificially corrupted with babble noise to improve the classifier performance.

Also the test data used in this work is from SPEECON. The utterances used for parameter optimisation and evaluation are all read sentences recorded either a) in public places both indoors and outdoors where speech, footsteps, unspecified clatter etc. appears in the background or b) in car environments. The sentences are excerpts from Internet texts and occupy a large (unlimited) vocabulary. The 60-minute and 29-minute parameter optimisation sets for public and car environments have speech from 20 and 10 speakers, respectively, and the 94-minute and 57-minute evaluation sets have been collected from 30 and 20 speakers, respectively. The evaluation sets do not share speakers with each other, the parameter optimisation sets, or the acoustic model training data. The data is recorded in 3-minute sessions with one speaker in one environment. CMLLR transformations are estimated per session, so the system is adapted to both the speaker and the environment. The environmental differences that CMLLR can compensate for include e.g. the distance between the speaker and the microphone and static noise components.

The proposed method is tested under three conditions: we use data recorded with the headset, data recorded with a lavalier microphone, and data recorded from 0.5 m–1 m distance (microphone mounted on the rear-view mirror in car environments). The three recordings have been made simultaneously, so they have exactly the same speech contents. In the speech data recorded in public environments, SNR values estimated with the recording platform are on average 24 dB in the headset data, 14 dB in the lavalier data, and 9 dB in the far recorded data, and in the speech data recorded in car environments, 13 dB in the headset data, 5 dB in the lavalier data, and 8 dB in the far recorded data. Speech recognition performance on the headset data recorded in public environments corresponds well to performance on speech data recorded in quiet environments, whereas in the lavalier and far recorded data, both environmental noise and reverberation affect the speech signal and decrease performance.

Note that although SNR estimates are higher for the data recorded in public environments, the dynamic noise scene is rather challenging for the conventional noise compensation methods. The car engine noise is more static and concentrated on low-frequencies. Note how in car environments, the average SNR in far recorded data is higher than in lavalier data even if the lavalier microphone is positioned closer to the speaker. This is likely because the lavalier microphone picks the low-pass component from the car engine noise more strongly than the far microphone (which is in the rear-view mirror). Low-pass noise decreases SNR but does not mask important speech frequency regions. Therefore it has a small effect on the actual speech recognition performance.

4. RESULTS

We examine how speech recognition performance changes when cluster-based missing feature reconstruction and adaptation with constrained maximum likelihood linear regression (CMLLR) are used in a noisy speech recognition task

Table 1: *Public environments.* Speech recognition results over the evaluation data recorded in noisy public environments such as parks and cafeterias. The headset data (H) is practically clean speech with average SNR 24 dB, while the lavalier (L) and far recorded (F) data are noisy speech with average SNR 14 dB and 9 dB, respectively. The best results obtained in each condition are underlined.

		H	L	F
WER	(a) BASELINE	13.8	43.8	67.6
	(b) ADA	$\underline{12.3}$	28.8	44.9
	(c) REC	14.1	30.5	46.8
	(d) REC→ADA	12.9	27.2	41.2
	(e) ADA→REC	12.5	$\underline{25.7}$	$\underline{39.1}$

		H	L	F
LER	(a) BASELINE	3.4	22.1	34.6
	(b) ADA	$\underline{2.7}$	12.7	22.0
	(c) REC	3.4	13.6	22.2
	(d) REC→ADA	3.0	11.7	20.4
	(e) ADA→REC	2.8	$\underline{11.0}$	$\underline{18.3}$

Table 2: *Car environments.* Speech recognition results over the evaluation data recorded in cars. The headset data (H) has average SNR 13 dB, the lavalier data (L) average SNR 5 dB and the far recorded data (F) average SNR 8 dB. The far microphone has been in the rear-view mirror, where the engine noise is not as loud as in the lavalier position. The best results obtained in each condition are underlined.

		H	L	F
WER	(a) BASELINE	14.7	51.5	95.7
	(b) ADA	$\underline{10.7}$	24.7	72.1
	(c) REC	13.9	40.9	68.4
	(d) REC→ADA	10.8	$\underline{23.7}$	$\underline{51.9}$
	(e) ADA→REC	11.3	28.2	63.2

		H	L	F
LER	(a) BASELINE	4.2	33.7	60.1
	(b) ADA	$\underline{2.5}$	11.5	42.7
	(c) REC	3.7	23.3	38.6
	(d) REC→ADA	2.5	$\underline{10.6}$	$\underline{33.5}$
	(e) ADA→REC	2.7	13.5	37.2

with large vocabulary continuous speech data. Speech recognition performance is evaluated with the following systems: (a) baseline system without adaptation, (b) baseline system with adaptation, (c) missing feature reconstruction without adaptation, (d) missing feature reconstruction prior to adaptation, (e) adaptation prior to missing feature reconstruction as proposed in Section 2.5. The results for speech recorded in noisy public environments are given in Table 1 and the results for speech recorded in car environments in Table 2. Since the words in Finnish are often long and consist of several morphemes, speech recognition performance is measured primarily in letter error rate (LER). In this work, also the word error rates (WER) are reported, but system comparisons are based on the letter error rates. Statistical significance is tested using the Wilcoxon signed rank test with significance level $p = 0.05$. This test is used for all pairwise system comparisons in this work.

The results in Tables 1 and 2 indicate that adaptation (b) improves speech recognition performance in all test conditions. Reconstruction (c) improves the results when applied on noisy data i.e. in all test conditions excluding the headset data from public environments. The improvements from baseline (a) are statistically significant ($p < 0.05$). The best results with the headset data in both public and car environments were obtained with the baseline system with adaptation (b). With the speech data from public environments, the best results with lavalier and far recorded data were obtained with the system using adaptation prior to reconstruction (e), and with the speech data from car environments, the best results with lavalier and far recorded data were obtained with the system using reconstruction prior to adaptation (d).

5. DISCUSSION

CMLLR adaptation (b) and missing feature reconstruction (c) significantly improve the speech recognition results over the noisy, lavalier and far recorded data from public environments (Table 1). When adaptation and reconstruction are used together (results (d) and results (e) in Table 1), the system performance further improves. The best results are obtained when adaptation is used prior to reconstruction (e). This confirms the hypothesis suggested in our previous work [5], where missing feature reconstruction and CMLLR were also tested in noisy public environments, but reconstruction was only used prior to adaptation. However, the situation changes in car environments (Table 2). For the lavalier and far recorded data, the best results are obtained when reconstruction is used prior to adaptation. To understand why the results are different in the two environments, we shall discuss in detail the results from using adaptation prior to reconstruction (e) and reconstruction prior to adaptation (d).

5.1 Adaptation prior to reconstruction

Using adaptation prior to reconstruction means that the CMLLR transformations are estimated from the noisy rather than reconstructed features. In public environments, the noise interference is typically not static and contains noise events such as footsteps or car passing the scene. CMLLR transformations, on the other hand, are estimated over several sentences and cannot compensate for dynamic effects. Thus, even if the transformations are estimated from noisy features, they are likely to compensate mostly for speaker variation and other mismatched static elements such as the microphone position. Reconstruction and the CMLLR transformations estimated prior to reconstruction compensate for different elements in the acoustic scene, so using reconstruction after adaptation (e) improves the results from adaptation (b) as indicated in Table 1. The difference is statistically significant ($p < 0.05$).

The results are different when we use adaptation prior to reconstruction in car environments (Table 2). Evaluated on the lavalier data, adaptation (b) gives better results than adaptation prior to reconstruction (e) ($p < 0.05$). Since the car engine noise is quite static, adaptation can compensate for the noise, and reconstruction is not necessary. The relative error reduction from adaptation (b) alone is

66 %. Evaluated on the far recorded data, adaptation prior to reconstruction (e) gives better results than adaptation (b) ($p < 0.05$) unlike in the previous case. The state sequence hypothesis (corresponding to the baseline result (a)) used for unsupervised adaptation contains so many errors that the CMLLR estimates become less than optimal; only in this case are better results obtained with reconstruction (c) than adaptation (b) ($p < 0.05$). Since adaptation does not do well in compensating for noise (or other variation), using reconstruction improves the results. However, also with the far recorded data, the results from using adaptation prior to reconstruction (e) in car environments are not as good as the results from using reconstruction prior to adaptation (d).

5.2 Reconstruction prior to adaptation

In previous work [5], we tested using reconstruction prior to adaptation on speech data recorded in noisy public environments same as here. This did not significantly improve the speech recognition results compared to using reconstruction without adaptation. In this work, the difference between the results from reconstruction (c) and reconstruction prior to adaptation (d) in Table 1 is statistically significant ($p < 0.05$), and the results from adaptation prior to reconstruction (e) are even better ($p < 0.05$). We believe this is because reasons discussed in Section 2.4.

Evaluated on the data from car environments, reconstruction prior to adaptation (d) gives better results than adaptation prior to reconstruction (e) ($p < 0.05$). When the CMLLR transformations are estimated from noisy features, adaptation and reconstruction both compensate for noise, but when reconstruction is used prior to adaptation and the CMLLR transformations are estimated from reconstructed features, adaptation learns mismatched elements other than noise. Thus, even if reconstruction probably degrades adaptation performance, using reconstruction prior to adaptation (d) is better than having both methods compensate for noise, which is what happens when we use adaptation prior to reconstruction (e) in car environments. Evaluated on the lavalier data, the difference between using reconstruction prior to adaptation (d) and adaptation (b) in Table 2 is not statistically significant ($p = n.s.$). Evaluated on the far recorded data, reconstruction prior to adaptation (d) is the better system ($p < 0.05$) in all pairwise comparisons, but there is another reason for this: the error-filled baseline result (a) is not well-suited for unsupervised adaptation, so in this case, adaptation really benefits from the better state sequence hypothesis that is constructed after reconstruction.

6. CONCLUSIONS AND FUTURE WORK

We proposed a new method to combine constrained maximum likelihood linear regression (CMLLR) [1] and cluster-based missing feature reconstruction [3] and evaluated the methods in noisy speech recognition task with data recorded in public and car environments. In our experiments, using both adaptation and reconstruction consistently improved the speech recognition results on the noisy lavalier and far recorded data. Using adaptation prior to reconstruction gave the best results on data recorded in public environments, while on the other hand, using reconstruction prior to adaptation gave the best results on data recorded in car environments. We suggested that the result is due to the fundamental differences in the noise conditions between public and car environments.

In our previous work [5], adaptation did not significantly improve the speech recognition results when applied on the reconstructed features. While this is not the case here, and the proposed method significantly improved the results on noisy speech recognition, we believe adaptation and reconstruction could do better still. In future, we aim to investigate other methods for using adaptation prior to reconstruction. For example, with minor changes to the feature extraction process, it would be possible to approximately propagate the adapted features to the log-spectral domain.

7. ACKNOWLEDGMENTS

This work was supported by the Academy of Finland in the projects *Auditory approaches to automatic speech recognition*, *New adaptive and learning methods in speech recognition*, and *Adaptive Informatics* and by the Helsinki Graduate School in Computer Science and Emil Aaltosen Säätiö.

REFERENCES

[1] M. J. F. Gales. Maximum likelihood linear transformations for HMM-based speech recognition. *Computer Speech and Language*, 12:75–98, April 1998.

[2] M. Cooke, P. Green, L. Josifovski, and A. Vizinho. Robust automatic speech recognition with missing and unreliable acoustic data. *Speech Communication*, 34:267–285, June 2001.

[3] B. Raj, M. L. Seltzer, and R. M. Stern. Reconstruction of missing features for robust speech recognition. *Speech Communication*, 43:275–296, September 2004.

[4] M. Van Segbroeck and H. Van hamme. Vector-quantization based mask estimation for missing data automatic speech recognition. In *Proc. INTERSPEECH*, pages 910–913, Antwerp, Belgium, August 27–31 2007.

[5] U. Remes, K. J. Palomäki, and M. Kurimo. Missing feature reconstruction and acoustic model adaptation combined for large vocabulary continuous speech recognition. In *Proc. EUSIPCO*, Lausanne, Switzerland, August 25–29 2008.

[6] V. Siivola and B. Pellom. Growing an n-gram language model. In *Proc. INTERSPEECH*, pages 1309–1312, Lisbon, Portugal, September 4–8 2005.

[7] T. Hirsimäki, M. Creutz, V. Siivola, M. Kurimo, S. Virpioja, and J. Pylkkönen. Unlimited vocabulary speech recognition with morph language models applied to Finnish. *Computer Speech and Language*, 20:515–541, October 2006.

[8] J. Pylkkönen. An efficient one-pass decoder for Finnish large vocabulary continuous speech recognition. In *Proc. 2nd Baltic Conference on Human Language Technologies*, pages 167–172, Tallinn, Estonia, April 4–5 2005.

[9] J. J. Odell. *The use of context in large vocabulary speech recognition.* PhD thesis, University of Cambridge, 1995.

[10] J. Pylkkönen and M. Kurimo. Duration modeling techniques for continuous speech recognition. In *Proc. INTERSPEECH*, pages 385–388, Jeju Island, Korea, October 4–8 2004.

[11] M. J. F. Gales. Semi-tied covariance matrices for hidden Markov models. *IEEE transactions on speech and audio processing*, 7:272–281, May 1999.

[12] GMMBAYES. http://www.it.lut.fi/project/gmmbayes/.

[13] J. Barker. Robust automatic speech recognition. In D. L. Wang and G. J. Brown, editors, *Computational Auditory Scene Analysis.* Wiley-Interscience, 2006.

[14] D. Iskra, B. Grosskopf, K. Marasek, H. van den Heuvel, F. Diehl, and A. Kiessling. SPEECON - speech databases for consumer devices: Database specification and validation. In *Proc. LREC*, pages 329–333, Las Palmas, Canary Islands, Spain, May 29–31 2002.