Multi-Condition Training for Unknown Environment Adaptation in Robust ASR Under Real Conditions

J. Rajnoha

Automatic speech recognition (ASR) systems frequently work in a noisy environment. As they are often trained on clean speech data, noise reduction or adaptation techniques are applied to decrease the influence of background disturbance even in the case of unknown conditions. Speech data mixed with noise recordings from particular environment are often used for the purposes of model adaptation. This paper analyses the improvement of recognition performance within such adaptation when multi-condition training data from a real environment is used for training initial models. Although the quality of such models can decrease with the presence of noise in the training material, they are assumed to include initial information about noise and consequently support the adaptation procedure. Experimental results show significant improvement of the proposed training method in a robust ASR task under unknown noisy conditions. The decrease by 29 % and 14 % in word error rate in comparison with clean speech training data was achieved for the non-adapted and adapted system, respectively.

Keywords: speech recognition, environment adaptation, spectral subtraction, MLLR, noisy background.

1 Introduction

Automatic Speech Recognition (ASR) in a noisy environment has been a challenging issue in recent decades for many research centers, as the presence of noise significantly decreases the accuracy of ASR systems. There are several approaches to compensate the effect of unclean conditions, which can be combined together with more or less advantageous results.

The first class of these methods is applied before acoustic modelling in front-end signal preprocessing. The signal is standardly represented by auditory-based features PLPs [1] or MFCCs to minimize the effect of speaker variability. Then noise suppression methods, such as most widely used Spectral Subtraction (SS) [2], Wiener filtering, and Minimum Mean Square Error (MMSE) estimation [3], are applied within front-end signal processing to minimize the noise level background in the analyzed speech.

The second class involves approaches, that take effect in the modelling phase. The models of speech and pause are typically trained on clean speech data to ensure high quality of the final models of speech. Model adaptation transforms clean speech models to perform well in a noisy environment. Several adaptation techniques use background noise, which is combined with the speech signal e.g. in multi-environment models [4], or with acoustic models in parallel model combination (PMC) [5]. Other techniques use noisy speech data to adapt acoustic models for particular background conditions by simply retraining the clean speech models or by some transformation using maximum likelihood linear regression (MLLR) [6] or Maximum A Posteriori (MAP) adaptation [7]. The latter two schemes are also used also for speaker adaptation with only a small proportion of adaptation material.

Due to varying or unknown target background conditions, and due to the high costs of collecting speech data in a real environment, not enough data that matches the recognition conditions for the adaptation procedure is typically available. Therefore a set of data for "almost matched" conditions is often used for training or model adaptation [4, 8].

In [9], clean speech was mixed additively with real noise from a car to get adaptation data. The final models were then adapted on these recordings by MLLR and MAP, with a resulting improvement from 14.38 % to 5.73 %. The authors show the advantage of using noisy data for training speech models in a car environment using additive mixing of clean speech and noise. Similarly, an additive noise approach outperformed the recognition results of a baseline system in different environmental conditions trained and tested on the Aurora 2 database in [10].

Unlike using only additive noise, data recorded in real conditions is used in this paper. The aim of our work is to analyse the influence of using multi-environment training material for robust speech recognition in an unknown environment. The recordings from the real world are important from the point of view of the real influence of noisy conditions. Not only additive distortion but also convolutional distortion are taken into account.

As shown e.g. in [11], joint usage of spectral subtraction and MLLR adaptation seems to be a good framework for a recognition task under conditions with a high level of background noise. These techniques can be used for blind adaptation, and they are therefore also useful for unknown noise reduction. This paper describes the effect of multi-condition training in several phases of the noise reduction algorithm shown in Fig. 1.

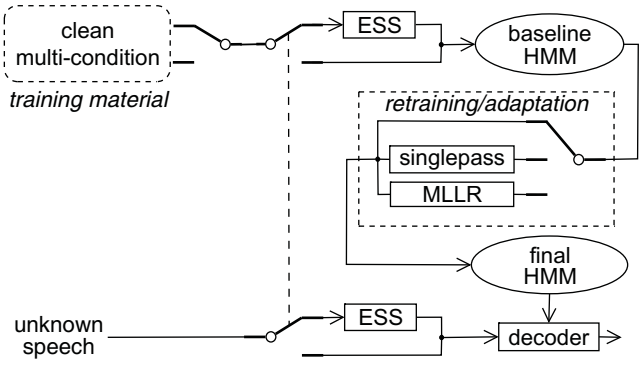

Fig. 1: Block scheme of the noise reduction algorithm

2 Noise compensation methods

The spectral subtraction technique is standardly used in front-end processing as a blind noise suppression method (see Fig. 1). Model parameters can be subsequently changed using single-pass retraining as a simple approach for offline adaptation or MLLR as a standard method which can be used for both offline and online adaptation.

2.1 Spectral subtraction

Spectral Subtraction (SS) is a technique frequently used for the suppressing the additive background noise component in the spectral domain to eliminate stationary or non-stationary noise with rather slow changes in characteristics. The characteristics of noise are estimated from speech pauses found e.g. by the Voice Activity Detector (VAD), which can often the limiting point of the algorithm. In our work, extended Spectral Subtraction (ESS) [12] uses modified adaptive Wiener filtering working without VAD.

2.2 Single-pass retraining

Single-pass retraining of the models is often used when a large amount of data with matching environmental conditions is available and offline retraining of acoustic models can be performed. The parameters of clean speech models are changed within one pass of the retraining procedure. This retraining is performed on the set of recordings with a matching environmental background. Such data will be called matching data in the following text. The disadvantage of this approach lies in the need for a sufficient amount of matching data for each model. This can be very difficult, mainly in the case of a specific environment. In addition a large number of speakers are needed for speaker independent recognition tasks. For these reasons, single-pass retraining was used as a low-bound result for unsupervised speaker adaptation experiments with an increasing amount of adaptation data in [6].

2.3 MLLR

As noted above, there is often not enough available data for the single-pass retraining procedure. MLLR uses a small amount of adaptation material to estimate an affine linear transform $\boldsymbol{A}$, $\boldsymbol{b}$ of the model parameters, which is found in terms of minimizing the likelihood of adaptation data. Based on our preliminary tests, we use only MLLR of mean vectors for our experiments, and other model parameters are unchanged. The new mean vector is then given by

$$\boldsymbol{\mu}_{\text{new}} = \boldsymbol{A}\boldsymbol{\mu}_{\text{old}} + \boldsymbol{b}\,. \tag{1}$$

The same transform $\boldsymbol{A}$, $\boldsymbol{b}$ can be applied to the mean vector of all models (global adaptation) or the models can be clustered on the basis of acoustic similarity into several classes. Separate transforms are then applied to particular classes. This clustering can represent the different effect of background distortion on particular speech phones. The regression class approach also enables us to cluster the models according to the amount of adaptation data to ensure sufficient quality of the transform. Binary regression class tree clustering [13] is used in this work.

3 Experiments

The experiments were performed on a small vocabulary speaker independent (SI) speech recognition task. The Czech digit sequence recogniser based on HMMs of monophones was used for this purpose.

3.1 Front-end setup

Front-end signal processing was carried out using the CtuCopy parametrization tool [14]. This enables similar functionality to the HTK HCopy tool [13] and provides additional noise reduction algorithms, e.g. VAD detection, spectral subtraction and LDA RASTA-like filtration.

Table 1 summarizes the overall setting of the recognition front-end.

Table 1: Front-end setup

segmentation window	25 ms Hamming window
segmentation step	10 ms
feature extraction	1 energy, 12 MFCCs +delta+acc. coeffs
models	HMMs of monophones 32 mixtures

3.2 Databases

The Czech Speecon database [15] was used for training and testing, i.e. 16 kHz data recorded in different environments using several types of microphones. The database involves utterances from almost 600 speakers with different content, e.g. phonetically rich sentences, digits, commands, etc.

Table 2 shows the division of the database in accordance with various environmental conditions. The whole database (ALL) was divided on the basis of type of recording environment (CLEAN and NOISY) or estimated SNR level (HISNR and LOSNR). Subsets with specific environment (OFFICE and CAR) were also created.

Each subset was divided into a training part and a testing part, taking into account a sufficient number of speakers for

Table 2: Description of SPEECON subsets and average estimated SNR

Name	Description	SNR [dB]	
		CS0	CS1
ALL	Whole SPEECON database	24.03	18.26
OFFICE	Very clean office recordings	26.91	19.88
CAR	Recordings in a car	13.33	8.43
CLEAN	Clean environment subset	27.15	20.80
NOISY	Noisy environment subset	21.25	15.44
HISNR	High SNR subset	27.51	20.36
LOSNR	Low SNR subset	13.76	12.07

SI recognition task. Training was performed on head-set microphone (CS0) data. Only the subsets ALL and OFFICE were used for training to simulate multi-condition training or clean data training, respectively.

Data from two different channels using a head-set microphone (CS0) and hands-free set (CS1) was used for testing. The CS1 channel is assumed to capture a higher level of background noise, which is illustrated in the estimated SNR values in Table 2. Each testing subset was divided for retraining or adaptation purposes according to the content, into a testing subset, which involves digits only, and a subset with the rest of the testing set, called the matched set.

As noted in sec. 2.3, the MLLR adaptation technique can work with a low amount of adaptation data. Subsets containing 20, 50, 100, 200, 500 and 1000 utterances were therefore created from each matched subset for comparison purposes. For the speaker-independent recognition task, each such subset involved as many speakers as possible. Not fewer than 18 speakers were present in the final subsets. This number can be considered as sufficient with regard to the number of speakers used in [9] (10–80) to get improvement in a speaker-independent task. Table 3 shows the average amount of adaptation data for different limits of utterances.

Table 3: Average amount of speech data for limited adaptation subsets

utter.	20	50	100	200	500	1000	all
time	65 s	2.8 min	5.7 min	10.9 min	26.9 min	57.2 min	7.8 h

3.3 Spectral subtraction in different conditions

Training the models on clean data, the presence of environmental distortion significantly decreases the recognition accuracy. As Table 4 shows, using ESS helps to suppress the influence of unclean conditions. Although the results are worse for matching conditions (Clean, CS0), the overall results give more than 8 % of WER enhancement.

A similar improvement was achieved for multi-condition training (Table 5). Although the unclean environment in the training phase decreases the quality of the resulting models, the overall contribution of using the multi-condition training database with ESS against the case of clean training data (Table 6) is almost 30 % of WER.

3.4 Single-pass retraining

All matching data for particular testing subsets was used for single-pass retraining in the case of clean or multi-condition training data. With regard to the results in section 3.3, ESS was used within front-end signal processing in the following experiments.

Table 7 shows that using multi-condition training data for training the initial models for single-pass retraining brings an improvement of over 22% against the clean speech models. All available matching data were used in this experiment, which led to the final set of 2400 (CAR) – 11600 (ALL) utterances for retraining.

3.5 MLLR

Single-pass retraining acts as a low-bound value for environment adaptation, as the amount of data for retraining is rather high. Only a limited amount of adaptation material for particular conditions is available in a real system, and decreasing the proportion of data for single-pass retraining procedure can lead to a significant decrease in recognition accuracy. MLLR-based adaptation removes this disadvantage. As shown in Fig. 2, even for a low amount of adaptation data the accuracy of a MLLR-adapted system outperforms the baseline and single-pass results.

Section 3.2 describes the adaptation subsets with a limited number of utterances, which reduces the computational load of the adaptation procedure. The recognition tests were performed on each subset and the results presented here show the value averaged over all these limited adaptation subsets.

Table 4: WER for different environmental conditions w/o and with ESS in front-end processing. The models are trained on clean data.

	ALL		OFFICE		CAR		CLEAN		NOISY		HISNR		LOSNR		AVG	avg	avg
	CS0	CS1	CS0	CS1	CS0	CS1	CS0	CS1	CS0	CS1	CS0	CS1	CS0	CS1		CS0	CS1
noSS	8.32	14.72	3.47	8.14	7.03	35.17	4.0	11.07	13.09	18.79	7.03	10.2	11.88	17.65	12.18	7.83	16.53
SS	8.46	13.53	4.17	8.01	5.2	29.97	4.45	9.82	11.31	16.56	6.83	9.41	11.76	16.31	11.13	7.45	14.8
Imp.	–1.68	8.08	–20.17	1.6	26.03	14.79	–11.25	11.29	13.6	11.87	2.84	7.75	1.01	7.59	8.66	4.82	10.48

Table 5: WER for different environmental conditions w/o and with ESS in front-end processing. The models are trained on multi-condition data

	ALL		OFFICE		CAR		CLEAN		NOISY		HISNR		LOSNR		AVG	avg	avg
	CS0	CS1	CS0	CS1	CS0	CS1	CS0	CS1	CS0	CS1	CS0	CS1	CS0	CS1		CS0	CS1
noSS	7.77	10.65	3.87	9.61	3.36	8.26	4.79	11.19	10.6	10.95	6.68	10.0	10.37	12.81	8.64	6.78	10.5
SS	7.04	10.65	3.87	7.21	1.53	8.87	4.68	9.7	8.55	10.77	5.74	9.55	9.78	12.58	7.89	5.88	9.9
Imp.	9.4	0.0	0.0	24.97	54.46	–7.38	2.3	13.32	19.34	1.64	14.07	4.5	5.69	1.8	8.59	13.17	5.63

Table 6: WER for clean (Clean) and multi-condition (M-C) training data and relative improvement for multi-condition training against clean training – no retraining/adaptation

	ALL		OFFICE		CAR		CLEAN		NOISY		HISNR		LOSNR		AVG	avg	avg
	CS0	CS1	CS0	CS1	CS0	CS1	CS0	CS1	CS0	CS1	CS0	CS1	CS0	CS1		CS0	CS1
Clean	8.46	13.53	4.17	8.01	5.2	29.97	4.45	9.82	11.31	16.56	6.83	9.41	11.76	16.31	11.13	7.45	14.8
M-C	7.04	10.65	3.87	7.21	1.53	8.87	4.68	9.7	8.55	10.77	5.74	9.55	9.78	12.58	7.89	5.88	9.9
Imp.	16.78	21.29	7.19	9.99	70.58	70.4	–5.17	1.22	24.4	34.96	15.96	–1.49	16.84	22.87	29.06	21.06	33.09

Table 7: WER for clean (Clean) and multi-condition (M-C) training data and relative improvement for multi-condition training against clean training – single-pass retraining

	ALL		OFFICE		CAR		CLEAN		NOISY		HISNR		LOSNR		AVG	avg	avg
	CS0	CS1	CS0	CS1	CS0	CS1	CS0	CS1	CS0	CS1	CS0	CS1	CS0	CS1		CS0	CS1
Clean	7.4	10.05	3.74	6.94	3.36	14.37	5.02	8.11	8.9	12.82	6.24	8.22	11.24	12.46	8.49	6.56	10.42
M-C	6.67	8.32	3.47	6.01	2.75	5.2	4.91	7.42	7.48	7.57	5.54	7.92	9.84	9.03	6.58	5.81	7.35
Imp.	9.86	17.21	7.22	13.4	18.15	63.81	2.19	8.51	15.96	40.95	11.22	3.65	12.46	27.53	22.5	11.42	29.46

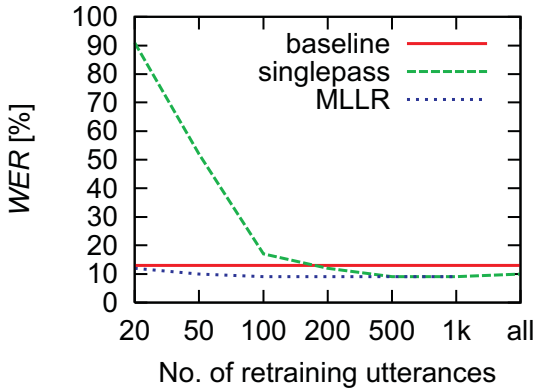

Fig. 2: WER for different amount of training data for single-pass retraining and MLLR adaptation

Various settings of model clustering for regression tree-based adaptation according to section 2.3 were used within the experiments. Global transformation and the division into 2, 4, 8, 16 and 32 regression classes were used, and the case with minimum achieved WER is reported in the following table.

The recognition results in Table 8 again show the improvement for using multi-condition training material for initial models. Only the case for very clean conditions (Clean, CS0) brings a slight decrease in WER. The contribution is evidentavg mainly for channel mismatch (CS1).

3.6 Overall improvement

Fig. 3 summarizes the contribution of using multi-condition training data for initial training in particular phases of the noise reduction procedure. The use of multi-condition training data leads to a significant improvement in all phases of the system.

The proposed noise reduction method led to the enhancement of WER by 48 %. The improvement achieved by

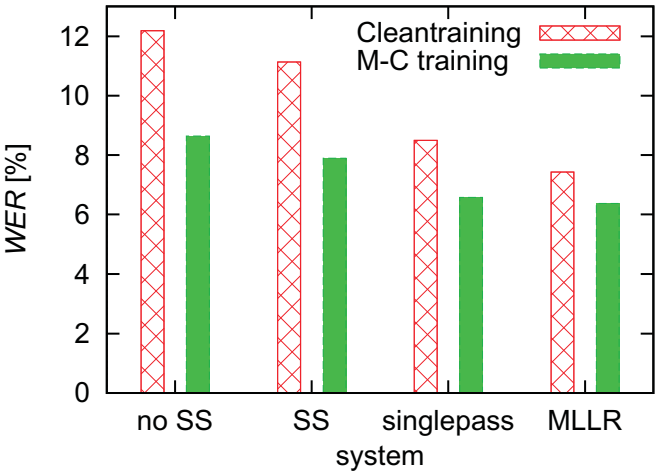

Fig. 3: Average *WER* in particular phases of noise reduction for clean and multi-condition training

Table 8: WER for clean (Clean) and multi-condition (M-C) training data and relative improvement for multi-condition training against clean training – MLLR adaptation

	ALL		OFFICE		CAR		CLEAN		NOISY		HISNR		LOSNR		AVG	avg	avg
	CS0	CS1	CS0	CS1	CS0	CS1	CS0	CS1	CS0	CS1	CS0	CS1	CS0	CS1		CS0	CS1
Clean	7.82	8.78	3.81	5.7	2.86	5.05	4.45	7.1	10.18	10.75	6.84	7.24	10.66	10.52	7.43	6.81	8.01
M-C	7.18	7.44	3.54	5.36	2.4	3.01	4.53	6.68	8.46	7.7	5.99	7.2	9.53	8.83	6.38	6.1	6.66
Imp.	8.29	15.21	7.05	5.88	16.11	40.37	–1.68	5.89	16.88	28.31	12.42	0.58	10.56	16.03	14.06	10.4	16.84

multi-condition training brought a more than 14 % decrease in recognition error.

4 Conclusion

The paper shows the advantages of using a multi-condition training data for robust ASR in unknown background conditions. The main contribution of the work is in using recordings from a real environment, which reflects the real influence of noise in a robust recognition task.

The results can be summarized in the following points.

- Multi-condition (M-C) training brings significant improvement to recognition accuracy, even in the case without any other noise reduction method. In the results presented here, multi-condition training outperforms the system that uses spectral subtraction and clean training data by 22 %.

  A combination of M-C training and spectral subtraction algorithm resulted in more than 29 % enhancement of WER against the baseline system. An increase in recognition accuracy by more than 70 % can be observed for data recorded in a car.
- Single-pass retraining gives a robust offline procedure for correcting acoustic models when enough matching data is available for a high variety of speakers and a rich phonetic content. The main contribution was observed for channel mismatch, and the use of M-C trained initial models brought an additional improvement to these results.
- Advantageous clustering of models based on available adaptation data within MLLR adaptation is shown to bring an improvement over single-pass retraining. The final improvement using M-C trained models only slightly outperforms the single-pass results.

Generally, multi-condition training material for initial training of speech models seems to bring an improvement to the recognition task in unknown environmental conditions. As the training and testing data in our experiments come from the same source, future work will be oriented to higher mismatches in adaptation and recognition conditions.

Acknowledgements

This research has been supported by grants GAČR 102/08/0707 “Speech Recognition under Real-World Conditions”, GAČR 102/08/H008 “Analysis and modelling biomedical and speech signals”, and by research activity MSM 6840770014 “Perspective Informative and Communications Technicalities Research”.

References

[1] Hermansky, H.: Perceptual Linear Predictive (PLP) Analysis of Speech. *J. Acoust. Soc. Am.*, vol. **87** (1990), no. 4, p. 1738–1752.

[2] Kang, G. S., Fransen, L. J.: Quality Improvement of LPC-Processed Noisy Speech by Using Spectral Subtraction. *IEEE Trans. on ASSP*, Vol. **37** (1989), No. 6, p. 939–942, June 1989.

[3] Ephraim, Y., Malah, D.: Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator. *IEEE Trans. on Acoustics, Speech and Signal Processing*, Vol. ASSP-32 (1984), no. 6, December 1984.

[4] Ming, J., Jancovic, P., Hanna, P., Stewart, D.: Modeling the Mixtures of Known Noise and Unknown Unexpected Noise for Robust Speech Recognition. *European Conference on Speech Communication and Technology* (Eurospeech’2001), Aalborg, Denmark, September 2001, p. 579–582.

[5] Gales, M. J. F., Young, S. J.: Parallel Model Combination for Speech Recognition in Noise. *Technical report CUED/F-INFENG/TR 135*, Cambridge, England, 1993.

[6] Leggetter, C. J., Woodland, P. C.: Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models. *Computer Speech & Language*, Vol. **9** (1995), No. 2, (April 1995), p. 171–185.

[7] Gauvain, J. L., Lee, C. H.: Maximum a posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains. *IEEE Transactions on Speech and Audio Processing*, Vol. **2** (1994), No. 2, p. 291–298.

[8] Liao, Y. F., Fang, H. H., Hsu, C. H.: Eigen-MLLR Environment/Speaker Compensation for Robust Speech Recognition. *Proceeding Interspeech’08*, Brisbane, Australia, September 2008, pp. 1249–1252

[9] Bippus, R., Fischer, A., Stahl, V.: Domain Adaptation for Robust Automatic Speech Recognition in Car Environments. *Proc. Eurospeech’99*, Budapest, Hungary, 1999, p. 1943–1946.

[10] Ming, J. Hou B.: Speech Recognition in Unknown Noisy Conditions. Chapter 11, in book *Robust Speech Recognition and Understanding*, M. Grimm and K. Kroschel (eds.), I-TECH Education and Publishing, 2007, p. 175–186.

[11] Matassoni, M., Omologo, M., Santarelli, A., Svaizer P.: On the Joint Use Of Noise Reduction and MLLR Adaptation for In-Car Hands-Free Speech Recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing* (ICASSP’02), 2002, p. 289–292.

[12] Sovka, P., Pollak, P., Kybic, J.: Extended Spectral Subtraction. *Proc. EUSIPCO’96*, Trieste, Italy, September 1996.

[13] Young, S. et al.: *The HTK Book* (for HTK Version 3.2.1), Cambridge University Engineering Department, 2002.

[14] Fousek, P., Pollak, P.: Additive Noise and Channel Distortion-Robust Parametrization Tool – Performance Evaluation on Aurora 2 & 3. *Proc. Eurospeech’03*, p. 1785–1788.

[15] SPEECON project page, URL: http://www.speechdat.org/speecon.

Josef Rajnoha
e-mail: rajnoj1@fel.cvut.cz

Department of Circuit Theory

Czech Technical University in Prague
Faculty of Electrical Engineering
Technická 2
166 27 Praha, Czech Republic