Experiments on Adaptation Methods to Improve Acoustic Modeling for French Speech Recognition

Saeideh Mirzaei[3], Pierrick Milhorat[4], Jérôme Boudy[1], Gérard Chollet[2] and Mikko Kurimo[3]
[1]*Department of Electronics and Physics, Telecom SudParis, Evry, France*
[2]*CNRS - Laboratoire Traitement et Communication de l'Information, Telecom ParisTech, Paris, France*
[3]*Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland*
[4]*Media Archiving Research Laboratory, Kyoto University, Kyoto, Japan*

Keywords: Speech Recognition, Speaker Adaptation, Linear Regression, Vocal Tract.

Abstract: To improve the performance of Automatic Speech Recognition (ASR) systems, the models must be retrained in order to better adjust to the speaker's voice characteristics, the environmental and channel conditions or the context of the task. In this project we focus on the mismatch between the acoustic features used to train the model and the vocal characteristics of the front-end user of the system. To overcome this mismatch, speaker adaptation techniques have been used. A significant performance improvement has been shown using using constrained Maximum Likelihood Linear Regression (cMLLR) model adaptation methods, while a fast adaptation is guaranteed by using linear Vocal Tract Length Normalization (lVTLN).We have achieved a relative gain of approximately 9.44% in the word error rate with unsupervised cMLLR adaptation. We also compare our ASR system with the Google ASR and show that, using adaptation methods, we exceed its performance.

1 INTRODUCTION

Automatic Speech Recognition (ASR) systems can play a great role in today's Human-Machine Interactive (HMI) systems. As ASR systems are introduced to a wide range of applications, the accuracy of these systems becomes significantly important. It depends on a number of factors; whether the ASR system is continuous or not, the extent of the task domain, the type of speech, planned or not and so on.

In our large vocabulary task, a continuous speech recognition system and data from different broadcast programs, containing both planned and spontaneous speech, we focus on the speaker dependency of the acoustic models. A Speaker Dependent (SD) system is trained using data from only one speaker, whereas a Speaker Independent (SI) system contains features from a large number of speakers. SD systems have shown to perform better than SI systems but as training models demands a large amount of data, SD systems are not feasible in practice. Hence, adaptation methods are deployed to improve SI models using a small amount of data from a new user. Without labels provided for the adaptation data, estimating parameters here is done in an unsupervised manner.

Previous work on the French broadcast news data reported a word error rate between 12 and 26 percent in the campaign held between 2007 and 2009 (Galliano et al., 2009). The data here is different from the original setup in this work and hence, results are not comparable with those reported in the aforementioned paper. The objective of our work is to investigate the improvement achieved by speaker adaptation and so the data has been rearranged based on the speakers.

This paper is organized as follows; Section 2 gives an overview of the adaptation methods with details on those implemented in this work, which includes vocal tract length normalization and linear regression methods to estimate the transformation parameters. In section 3, we present the tools used to carry out the experiments. Section 4 describes the data set and how it is arranged for training and test. In section 5, we introduce the evaluation metrics in our experiments. Section 6 presents the results giving precise information on the adjustments used to implement the experiments. We conclude the paper in section 7 and provide a perspective for the possible future work.

Mirzaei, S., Milhorat, P., Boudy, J., Chollet, G. and Kurimo, M.
Experiments on Adaptation Methods to Improve Acoustic Modeling for French Speech Recognition.
DOI: 10.5220/0005703702780282
In *Proceedings of the 5th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2016)*, pages 278-282
ISBN: 978-989-758-173-1

2 ACOUSTIC MODEL ADAPTATION

Acoustic model adaptation techniques, in general, are used to reduce the mismatch between the trained model parameters and the test data conditions, the channel and environmental effects or characteristics of a new speaker voice. This is possible by transforming the feature set or adjusting the model parameters. The model parameters here are Gaussian Mixture Model specifications (GMM means and convariances) as part of Hidden Markov Models (HMM). The desired qualities of adaptation techniques are to be fast, to require a small amount of adaptation data and to asymptotically converge to the maximum likelihood estimate of the parameter.

Maximum A Posteriori (MAP) (Gauvain and Lee, 1994) estimates each parameter after a sample containing that parameter is observed. Although converging to maximum likelihood estimates, adaptation using MAP is very slow considering the large number of existing parameters in a model. In fact, a large amount of adaptation data is required for MAP to be effective. This led to other methods to bind the parameters together and make each estimation valid for each class. Structural MAP (Shinoda and Lee, 1997) and Regression-based Model Prediction(Ahadi and Woodland, 1997) are techniques based on MAP. Some other proposed methods use regression analysis or pooling techniques to accelerate adaptation. In this section the three methods used in this work are explained.

2.1 Vocal Tract Length Normalization

Differences in the vocal tract shape and length among speakers result in the fundamental frequencies to vary from one speaker to another. This can be noticed between a male and a female speaker, the fundamental frequency of a typical female speaker is higher than that of a typical male speaker. VTLN normalizes the perceived voice during feature extraction based on the position of the formants to reduce this variation. A piece-wise linear approach has been used, i.e. the transformation function is linear (Eide and Gish, 1996), and only one warping factor has to be estimated.

$$\hat{f} = Af + b \tag{1}$$

where, $\hat{f}$ and f are the warped and unwarped frequencies respectively. There exists only one parameter to be estimated, the warping factor, with a defined range, e.g. between 0.8 and 1.2.

2.2 Maximum Likelihood Linear Regression

The Maximum Likelihood Linear Regression (MLLR) (Leggetter and Woodland, 1995) method uses a linear transformation matrix to re-estimate the model parameters. The transformation is applied on either the Gaussian parameters or the features, which defines the unconstrained or constrained MLLR respectively. Different variations of MLLR estimate transformation matrices for means only, for means and variances, etc. With constrained MLLR (cMLLR) the same transformation matrix is used to transform both means and covariances of the Gaussians.

$$\hat{\mu} = A\mu + b \tag{2}$$

$$\hat{\Sigma} = A\Sigma A^T \tag{3}$$

The transformed mean vector $\hat{\mu}$ and covariance matrix $\hat{\Sigma}$ are obtained by applying the transformation matrix A on the original values, μ and Σ.

To make adaptation robust, Gaussians close together in the acoustic space or Gaussians in the same state can be grouped and the same transformation matrix can be applied to that same class. This is essential when the adaptation data is small and the probability of not observing the effect of some parameters is high. In the case where a large amount of adaptation data is available, finer transformations can be applied to a smaller group of Gaussians. Special attention must be paid to the amount of adaptation data and the size of the transformation matrix, to prevent overfitting the model parameters.

2.3 Speaker Adaptive Training

Speaker Adaptive Training (SAT) is used to train the speaker independent acoustic model on the average voice (Anastasakos et al., 1996). Either VTLN or MLLR can be used to eliminate the inter-speaker variability during estimation of HMM parameters. The specifications of the transformation matrix must be estimated jointly with the HMM parameters.

$$(\hat{\lambda}, \hat{G}) = \arg\max_{\lambda, G} \prod_r \mathcal{L}(O^r; G^r(\lambda)) \tag{4}$$

where, λ is the HMM parameters vector and G is a block matrix including the speaker specific transformation matrices. The training is performed on observations from each speaker, by maximizing the likelihood of the observed data, O^r, from speaker r with the given transformation matrix specified for the same speaker, G^r, and the model parameters. SAT gives

better results when used with adaptation methods. Its main drawback is the large memory requirement to store all the transformation matrices.

3 TOOLS

The tools used to perform the experiments were all open source. The language models were built using SRILM (Stolcke et al., 2002). The same tool was used to assess these language models. The selected language model in ARPA format was then transformed to FST format by OpenFST (Allauzen et al., 2007) to be readable by Kaldi (Povey et al., 2011). Kaldi, a powerful ASR tool, was used to build the acoustic models, perform adaptation methods and produce the outputs for the final evaluation of the ASR system.

4 DATA

The data from Ester - ISLRN: 110-079-844-983-7; ELRA-E0021, Catalogue ELRA (Evaluation des systèmes de transcription enrichie d'émissions radiophoniques) (Galliano et al., 2006) and Etape - ANR ANR-09-CORD-009-05 (Evaluations en Traitement Automatique de la Parole) (Gravier et al., 2012) were combined to form our data set for training and testing. The data is from French TV and Radio broadcasts. Etape, compared to Ester, contains more spontaneous speech and has more multiple-speaker segments, and so it is more challenging for speech recognition tasks. The sampling rate of the audio files is 16 kHz. After a manual segmentation of audio files to extract speech parts only, the average length of the resulting files was 3.5 seconds.

Given the nature of the work, we needed to rearrange the data based on the speakers. 18 speakers with the highest amount of speech data from both sets were used as the evaluation set. An equal number of speakers were extracted from Ester and Etape and only single-speaker segments were preserved for testing. The rest of the data was used for training after excluding those segments containing any speaker from the test set. In total 145 hours of speech were used for training and 18 hours for test. The test part contains data from both Ester and Etape: 8 hours from Ester only, 8 hours from Etape only and 2 hours common to the two sets. The two hours of shared data comes from the two speakers appearing in the two data sets, the results for these two speakers in experiments are presented in a separate part as Ester-Etape. Test sets from Ester, Etape and Ester-Etape include 46155, 60372 and 16329 words respectively.

5 EVALUATION METRICS

The criteria to select the language model is the perplexity which by definition is the language model confusion to predict the next word. It is formulated as follows:

$$Perplexity = \sqrt[N]{\frac{1}{\prod_{i=1}^{N} Pr(W_i|H)}} \quad (5)$$

To evaluate the performance of the ASR system, the word error rate (WER) is used. It is defined based on the Levenshtein distance and is calculated as follows:

$$WER = \frac{I+D+S}{T} \quad (6)$$

with I, the number of Inserted, D, Deleted, S, Substituted words and T, the Total number of original words.

The confidence interval is 1% maximum with the confidence of 95%, having 45000-word (Ester), 60000-word (Etape) and 16000-word (Ester-Etape) test sets. The error margin is obtained by using the following formula:

$$I_c = 1.96 \times \sqrt{\frac{\bar{x}(1-\bar{x})}{n}} \quad (7)$$

with $\bar{x}$, the error rate and n, the sample size. The value of 1.96 is obtained from a standard distribution with the coverage of 95%.

6 EXPERIMENTAL RESULTS

We present the experimental setup in this section. It describes the system, and then all the results of the evaluation are presented in the following parts.

6.1 Set Up

The feature set is constructed using Mel-Frequency Cepstral Coefficients (MFCCs) (Davis and Mermelstein, 1980). The MFCCs are extracted over 25 ms-length frames with a frame shifts of 10 ms. The first 13 coefficients form the basic feature vector. In the unadapted model system, these coefficients and their first and second derivatives are adjoined to build the 39-element feature vector. The feature set for the adapted model system is obtained by using Heteroscedastic Linear Discriminant Analysis (HLDA) (Kumar and Andreou, 1998). HLDA is implemented on Gaussians using only means and the new classes

are assumed to have fixed variances except for the general model which has a unit variance. With a context dependency of length three (appending 7 consecutive feature vectors), a 91-dimensional vector is formed and then reduced to 40-D, out of which 1 is the general model for all the rejected dimensions. Cepstral Mean and Variance Normalization (CMVN) (Prasad and Umesh, 2013) is applied segment-wise both during training and testing sessions to cancel the channels effects.

To build the language models a lexicon of 54k words was used with 39 phonemes. Two 3-gram language models were built. One was build on the training part of Ester and Etape data sets with Kneser-Ney smoothing (Chen and Goodman, 1999) applied on the model. The other language model was produced by using Google n-gram counts made available in 2009 for French data. In this latter, combinations with probability of less than 10^{-7} were pruned out. The size and perplexity (tested on the test corpus) of these two language models are presented in Table 1. We used the language model trained on Ester and Etape data sets to perform the experiments because of its smaller size and lower perplexity.

Table 1: Language models and their perplexities.

Data set	Perplexity	Size
Ester-Etape training set	150	4.5M
Google n-gram counts	289	104M

The monophone model was built with 132 states and 1000 Gaussians in total. The triphone model was built with approximately 3000 states and a total number of 56000 Gaussians (18 Gaussians per state). All the settings during training and decoding were left to their default values; 35 iterations for estimation, the scaling factor of 0.083333 dedicated to the acoustic likelihood. During decoding the same acoustic likelihood scaling factor was used. The maximum number of states at each frame was 7000 and the beam factor of 13 was used as pruning beam during graph search, and determining the lattices after decoding. SAT was then implemented to build the final speaker independent model with normalized GMMs by applying cMLLR.

6.2 Results

Unsupervised adaptation was implemented. During adaptation and decoding, the data was fed to the system in a batch mode. With the first pass of decoding, the first lattices were produced by using which the transformation matrices were estimated. Lattices from the second pass were used to readjust the parameters and produce the final lattices.

Table 2 shows the results for the basic model and unsupervised adapted models using lVTLN and cMLLR. Compared to Etape, tests on Ester data set revealed better results in general since this set includes mostly planned speech.

Both adaptation methods improved the performance of the basic system (Triphone Model in Table 2) but cMLLR proved to be more effective than lVTLN in all test sets. With Ester test set, cMLLR improved the performance by 11.3% while the gain obtained by lVTLN with the same set was 7.4%. cMLLR provided a relative gain of 8.2% for the Etape test data. The improvement for the same data by lVTLN was 5.6%. All relative gains are calculated in respect to the results of the basic model.

Table 2: WER%; a gain between 6-12 percent is obtained by adaptation.

	Ester	Etape	Ester-Etape
Triphone Model	28.2%	53.7%	52.3%
SAT+lVTLN	26.1%	50.7%	47.7%
SAT+cMLLR	25.0%	49.3%	46.2%

In the second experiment, in which only cMLLR was implemented for adaptation, we increased the number of adaptation utterances to investigate the improvement achieved corresponding to each amount. The number of utterances was increased from 1 to 10. The results in this part are compared with the performance of Google ASR and the basic model in Figure 1, with test on Ester data set only. The confidence interval with this set is 0.4%.

The horizontal axis in Figure 1 shows the number of adaptation utterances (for the adapted model) and the vertical axis shows the WER. The two horizontal lines in the figure display the WER for the basic model (dashed line) and the Google ASR outputs (solid line). The WER by Google ASR and the basic model were 26.1% and 28.2% respectively.

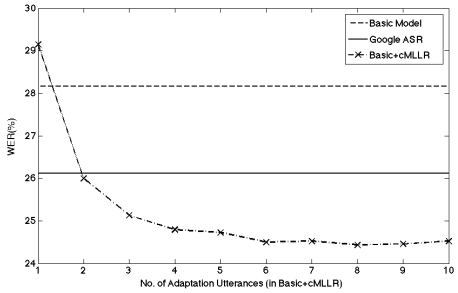

Figure 1: A comparison between the Basic Model, the Google ASR and the incrementally Adapted Model

We observe how the performance changes by increasing the number of adaptation utterances; using only one utterance as the adaptation data decreases the performance resulting in a higher WER, with two and more utterances up to 6, the performance gets improved gradually. Afterwards, with more than 6 utterances, no more gain is obtained. The line is expected to reach the value of 25% (the WER in Table 2) if the adaptation data was increased. We also observe that the adapted model reaches the Google ASR performance with two utterances and outperforms it with more adaptation utterances.

7 CONCLUSIONS

Here we presented a large vocabulary continuous speech recognition system based on a GMM-HMM system. We implemented adaptation methods to improve the system. Two methods, lVTLN and cMLLR, were used for unsupervised acoustic model adaptation. The performance of these systems were compared with the speaker independent system by testing on the Ester and Etape data sets. The basic model, which was a triphone model, was improved by applying SAT and lVTLN/cMLLR. It was shown that the performance was improved by a relative 9.44 percent reduction in WER by using cMLLR. In the end the basic model and the adapted model using cMLLR were compared with the Google ASR. It was shown that the adaptation with cMLLR could improve the basic system to overpass Google ASR.

We also observed in general the system worked better for the data set including more planned speech. This shows the importance of a good language model. Therefore, we believe further gain could be obtained by improving the language model, e.g. combining the Google n-gram counts with n-gram language model from training set using interpolation methods.

REFERENCES

Ahadi, S. and Woodland, P. C. (1997). Combined bayesian and predictive techniques for rapid speaker adaptation of continuous density hidden markov models. *Computer speech & language*, 11(3):187–206.

Allauzen, C., Riley, M., Schalkwyk, J., Skut, W., and Mohri, M. (2007). Openfst: A general and efficient weighted finite-state transducer library. In *Implementation and Application of Automata*, pages 11–23. Springer.

Anastasakos, T., McDonough, J., Schwartz, R., and Makhoul, J. (1996). A compact model for speaker-adaptive training. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, volume 2, pages 1137–1140. IEEE.

Chen, S. F. and Goodman, J. (1999). An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4):359–393.

Davis, S. B. and Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 28(4):357–366.

Eide, E. and Gish, H. (1996). A parametric approach to vocal tract length normalization. In *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, volume 1, pages 346–348. IEEE.

Galliano, S., Geoffrois, E., Gravier, G., Bonastre, J.-F., Mostefa, D., and Choukri, K. (2006). Corpus description of the ester evaluation campaign for the rich transcription of french broadcast news. In *Proceedings of LREC*, volume 6, pages 315–320.

Galliano, S., Gravier, G., and Chaubard, L. (2009). The ester 2 evaluation campaign for the rich transcription of french radio broadcasts. In *Interspeech*, volume 9, pages 2583–2586.

Gauvain, J.-L. and Lee, C.-H. (1994). Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. *Speech and audio processing, ieee transactions on*, 2(2):291–298.

Gravier, G., Adda, G., Paulson, N., Carré, M., Giraudel, A., and Galibert, O. (2012). The etape corpus for the evaluation of speech-based tv content processing in the french language. In *LREC-Eighth international conference on Language Resources and Evaluation*, page na.

Kumar, N. and Andreou, A. G. (1998). Heteroscedastic discriminant analysis and reduced rank hmms for improved speech recognition. *Speech communication*, 26(4):283–297.

Leggetter, C. J. and Woodland, P. C. (1995). Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. *Computer Speech & Language*, 9(2):171–185.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlíček, P., Qian, Y., Schwarz, P., et al. (2011). The kaldi speech recognition toolkit.

Prasad, N. V. and Umesh, S. (2013). Improved cepstral mean and variance normalization using bayesian framework. In *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*, pages 156–161. IEEE.

Shinoda, K. and Lee, C.-H. (1997). Structural map speaker adaptation using hierarchical priors. In *Automatic Speech Recognition and Understanding, 1997. Proceedings., 1997 IEEE Workshop on*, pages 381–388. IEEE.

Stolcke, A. et al. (2002). Srilm-an extensible language modeling toolkit. In *INTERSPEECH*.