Acoustic and Language Model Adaptation in a Voice Interactive System for Elderly people

Saeideh Mirzaei (Aalto University)
Email: saeideh.mirzaei@aalto.fi
Jerome Boudy (Télécom SudParis)
Email: jerome.boudy@telecom-sudparis.eu
Pierrick Milhorat and Gerard Chollet (Télécom ParisTech)
Email: firstname.lastname@telecom-paristech.fr
Mikko Kurimo (Aalto University)
Email: mikko.kurimo@aalto.fi

***Abstract*—Automatic Speech Recognition (ASR) systems can perform better if trained for a specific application. Though, since we require a huge amount of information to train models it is not feasible to build such systems once ready for the user, but we could use adaptation to make the ASR system more appropriate for the final use. In this work we address adaptation for the vocal characteristics of the speaker, environmental noise and the language model. Acoustic model adaptation is done by Speaker Adaptive Training (SAT), linear Vocal Tract Length Normalization (lVTLN) and constrained Maximum Likelihood Linear Regression (cMLLR). Interpolation is applied for language model adaptation. The relative WER reduction by using cMLLR was 9.44%. The perplexity of the language model could be relatively improved by 14.47%.**

I. Introduction

The vAssist project (Voice Controlled Assistive Care and Communication Services for the Home) is focused on assisting the elderly and people with minor disabilities especially with daily medical care. A spoken dialogue system is used to interact with the user and has the role of recognizing the commands from the user. Although ASR systems have improved significantly, but they still need improvements in large vocabulary tasks.

To increase the performance of an ASR system, we could make it more user dependent by acoustic model adaptation methods and more task oriented by language model adaptation. Different techniques, such as vocal tract length normalization, maximum likelihood linear regression, maximum a posteriori, etc. have been proposed and in some cases the combination of these methods to compensate for acoustic mismatches. For language model adaptation as well, we could use linear or log-linear interpolation methods. In this work, we are focused on SAT, lVTLN, cMLLR in acoustic model adaptation and linear Interpolation for language model adaptation.

The data set used in this work is described in section 2. Section 3 gives an overview on speech recognition systems. In section 4 and 5 we go through the adaptation methods used in this work. Section 6 presents the experiments and results and section 7 concludes the work.

II. Data set

We have combined Ester - ISLRN: 110-079-844-983-7; ELRA-E0021, Catalogue ELRA (Evaluation des systèmes de transcription enrichie d'émissions radiophoniques) [4] and Etape - ANR ANR-09-CORD-009-05 (Evaluations en Traitement Automatique de la Parole) [5] to build the training and testing data sets. Both Ester and Etape are from French TV and Radio broadcasts. Etape compared to Ester contains more spontaneous speech and have more multiple speaker segments and so is more challenging for speech recognition. The audio files are recorded with a sampling rate of 16 kHz. The average length of each segment was 3.5 seconds.

We separated 18 speakers from data set for testing purpose. The number of speakers extracted from Ester and Etape are equal. Only single-speaker segments are reserved for testing and all segments in training set having any speaker from test set were removed from training. Details on the amount of data for training and evaluation is shown in Table I. Two speakers were common between the two data sets, the results for these two speakers in experiments are presented in a separate part as Ester and Etape.

Data set	*Ester*	*Etape*
Training	121 h	24 h
Test	9 h (8+2 speakers)	9 h (8+2 speakers)

TABLE I: Data set; 145 hours for training and 18 hours (18 speakers) for test

III. Speech Recognition System

The three main parts of each ASR system are feature extraction, acoustic and language models and decoding. In the following, we explain the general methods used in our speech recognition system. Feature vector is Mel-Frequency Cepstral Coefficients (MFCC). Speech signals are transformed by Fast-Fourier to frequency domain, an passed through mel-filters. After taking logarithm of the energy in each mel bin, a Discrete Cosine Transform is applied. The feature set could be built by appending the time derivatives, log-energy, etc. to the MFCCs or using Heteroscedastic Linear Discriminant

Analysis (HLDA) to reduce the feature dimensions after appending 3-9 consecutive feature vectors.

Hidden Markov Models (HMM) with continuous density represented by a mixture of Gaussians (GMM) are used to create the acoustic models. In a monophone model, each phoneme is modeled independently, while in triphone and quinphone models a context-dependency of ± 1 or ± 2 is taken into account. Since the number of triphones is large and for the probability of not observing all of them, states are tied as described in [12].

The language model defines the constraints for the sequences of the words that can appear in an utterance. This could be because of syntactic or semantic properties of the language. The grammatical constraint can be defined by n-gram language models. If the corpus is not large enough to cover all possible combination, we may use discounting methods such as Kneser-Ney. Otherwise, unseen combinations during training the language model will be given zero probability and leaves no flexibility in the recognition. Viterbi is widely used in speech recognition systems to perform the decoding. As Viterbi is computationally intensive, pruning methods and search networks are used for large vocabulary tasks, e.g. Weighted Finite States Transducer.

IV. Acoustic Model Adaptation

The mismatch between the vocal features of a specific speaker and the speaker independent model can be compensated by adjusting the model parameters or transforming the features to better match the model parameters. Maximum a Posteriori (MAP) adjusts each parameter given the observed data and considering a prior distribution for that parameter. In MAP a weight is given to the prior distribution and the observed data. As the amount of observed data is increased the effect of prior information becomes less. MAP needs a large amount of data to be effective, therefore other methods have been proposed. In this work, we have implemented vocal tract length normalization and maximum likelihood linear regression for adaptation.

A. Vocal Tract Length Normalization

The vocal tract shapes and length of each speaker is different from the other. This causes variation in the formants of the voice between speakers. The fundamental frequency of the voice of a typical female speaker is higher than that of a typical male speaker. Vocal Tract Length Normalization (VTLN) compensates for this change by warping the frequency [3]. To do so, a piece-wise linear function can be implemented and the only parameter to be estimated is the warping factor. As there is only one parameter to be estimated, the method is fast in adaptation.

B. Speaker Adaptive Training

SAT was proposed in [1] to cancel the inter-speaker variability while training the model. The idea is to estimate the HMM parameters during training while taking into account the speaker specific vocal features. This can be done by normalizing the features or applying vocal tract normalization. In [1], HMM parameters are jointly estimated with features transformation for each speaker.

$$(\bar{\lambda}, \bar{G}) = \arg\max_{\lambda, G} \prod_{r} \mathcal{L}(O^r; G^r(\lambda)) \tag{1}$$

where λ is the HMM parameters and G is the speaker specific transformation.

To estimate the parameters, using expectation-maximization, first the means and covariances of the Gaussians are kept fixed and the parameters of HMMs (the Q function) are estimated. In the second step, considering HMM parameters and the Gaussian covariances fixed, the means are estimated and in the last step the covariances are estimated. These parameters are iteratively estimated until convergence.

C. Maximum Likelihood Linear Regression

Maximum Likelihood Linear Regression method uses a linear transformation to estimate the model parameters given the adaptation data. This transformation is applied on the Gaussians means and variances. Sometimes, we could assume the main differences between speakers are the parameters means, in which case we only use one transformation matrix to estimate the means. Otherwise, not only the means but also the covariances are needed to be estimated. In the unconstrained case different transformation matrices for the means and covariances should be obtained, first the means are estimated assuming the covariances fixed and then covariances are estimated.

$$\bar{\mu} = A\mu + b, \quad \bar{\Sigma} = L\Sigma L' \tag{2}$$

in which μ and σ are mean and covariance matrices respectively.

In the constrained MLLR (cMLLR) the same transformation matrix is used for both means and covariance estimation. 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 them. This is vital where the adaptation data is small and the probability of not observing some parameters is high. In the case a large amount of adaptation data is available, more precise transformations can be applied to a smaller group of Gaussians. Special attention must be paid to the amount of the adaptation data and the size of the transformation matrix, to prevent overfitting the model parameters. We have used cMLLR in the experiments in this work.

V. Language Model Adaptation

Language models play a great role in speech recognition performance. Depending on the domain of the application, some words or their combinations are more probable to occur. Therefore, we could improve the language model by giving more weight to those more probable words sequences. Here we have implemented Interpolation for language model adaptation.

A. Linear Interpolation

This is one of the most popular and simplest methods that combines the language models giving a weight between zero and one to each model conditioned that the sum of the weights (interpolation coefficients) be equal to one.

$$P(W|h) = \sum_i \lambda_i P_i(W|h), \quad \sum_i \lambda_i = 1 \tag{3}$$

The interpolation coefficients (λ_i) are estimated on held-out data empirically or by maximum likelihood[2].

VI. Experimental Results

A. Set up

We have used Kaldi [9] for the experiments in this work. SRILM [11] is also used to build the language models. The MFCCs are extracted over every 25ms-length frame and by frame shifting of 10 ms. To compensate for the channel effects, we have made use of Cepstral Mean and Variance Normalization (CMVN); The means and variances of the features in each utterance are normalized (to zero and one respectively) during both training and testing phase. The feature set to train monophone and basic triphone models has 39 dimensions; 13 MFCCs, with their first and second derivatives appended. For models trained by SAT, we used HLDA and the feature set has 40 dimension; 7 consecutive feature vectors are appended and then the 91 dimension feature set is mapped to a 40 dimension feature set.

The monophone model has 132 states with 1000 Gaussians in total. The triphone model has approximately 3000 states with a total number of 56000 Gaussians. The language model is a 3-gram language model trained with a 54k words lexicon. The training transcription of our experiments (the combination of Ester and Etape) was used to build the model. It is smoothed with Kneser-Ney discounting method. Word Error Rate (WER) is used to evaluate the recognition performance whereas Perplexity is used as a criteria for language model evaluation.

B. Results

SAT was implemented to build the final speaker independent model with normalized GMMs by applying cMLLR. Decoding was first done with the speaker independent system. These hypotheses were used to estimate the lVTLN or cMLLR transformation matrices. The probability of each hypothesis is used as a weight in the estimation of the transformation matrix. The transformed features were used to re-score the lattices to produce the final transcriptions [8]. Table II shows the error rates for the speakers from each set.

	Ester	*Etape*	*Ester and Etape*
Triphone Model	28.2%	53.7%	52.3%
$HLDI + SAT + lVTLN$	26.1%	50.7%	47.7%
$HLDI + SAT + cMLLR$	25.0%	49.3%	46.2%

TABLE II: WER%; a gain between 8-11 percent is obtained by adaptation

The performance of the system was improved by adaptation using either of the two methods. However, cMLLR demonstrates better results than lVTLN. A gain of about 6.53% was obtained by lVTLN while the gain was 9.44% in adaptation with cMLLR.

Two language models were constructed to build the final language model; one on the Google n-gram counts and the other on Ester and Etape training corpus. The statistics of these language models are given in TableIII, both are trigram. The third column shows the interpolated language model. As it can be seen the perplexity of the final language model is lower than that of each previous language models.

TABLE III: The Perplexity for the three language models

Language Model	*Pruning*	*Smoothing*	*Perplexity*	*oov*	*size*
Google	10^{-7}	-	289	5.7%	104M
Ester-Etape	10^{-7}	Kneser-Ney	152	0	4.6M
Interpolated LMs	-	-	130	0	110M

VII. Conclusion

Here we presented a speech recognition system with unsupervised adaptation. The performance of these systems were compared with the speaker independent system testing on the Ester and Etape data. The results were presented for three test sets, Etape (with more spontaneous speech), Ester (with mostly single-speaker segments) and a combination of these two sets. Two methods lVTLN and cMLLR were used in adaptation. lVTLN is faster in adaptation and needs less adaptation data while cMLLR performs better and the gain obtained by this adaptation approach is higher than lVTLN. The basic model which was a triphone model was significantly improved by applying HLDA, SAT and cMLLR, it was shown that the performance was improved by a relative 9.44 percent reduction in WER. Further improvements could be obtained by improving the language model. We showed that the perplexity of the language model could be reduced by adapting to the task. We presented the perplexity for language models from Google n-gram counts and the language model based on training corpus and compared them with the language model obtained by interpolating these two language models.

VIII. Acknowledgments

This work was supported by the vAssist project from Ambient Assisted Living (AAL) program and funded by the French Research Agency (ANR).

References

[1] Anastasakos, Tasos, et al. "A compact model for speaker-adaptive training." Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on. Vol. 2. IEEE, 1996.

[2] Bellegarda, Jerome R. "Statistical language model adaptation: review and perspectives." Speech communication 42.1 (2004): 93-108.

[3] Eide, Ellen, and Herbert Gish. "A parametric approach to vocal tract length normalization." Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on. Vol. 1. IEEE, 1996.

[4] Galliano, Sylvain, et al. "Corpus description of the ester evaluation campaign for the rich transcription of french broadcast news." Proceedings of LREC. Vol. 6. 2006.

[5] Gravier, Guillaume, et al. ”The ETAPE corpus for the evaluation of speech-based TV content processing in the French language.” LREC-Eighth international conference on Language Resources and Evaluation. 2012.
[6] Kumar, Nagendra, and Andreas G. Andreou. ”Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition.” Speech communication 26.4 (1998): 283-297.
[7] Leggetter, Christopher J., and Philip C. Woodland. ”Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models.” Computer Speech & Language 9.2 (1995): 171-185.
[8] Mangu, Lidia, Eric Brill, and Andreas Stolcke. ”Finding consensus among words: lattice-based word error minimization.” Eurospeech. 1999.
[9] Povey, Daniel, et al. ”The Kaldi speech recognition toolkit.” (2011).
[10] Prasad, N. Vishnu, and Srinivasan Umesh. ”Improved cepstral mean and variance normalization using Bayesian framework.” Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on. IEEE, 2013.
[11] Stolcke, Andreas. ”SRILM-an extensible language modeling toolkit.” INTERSPEECH. 2002.
[12] Young, Steve J., Julian J. Odell, and Philip C. Woodland. ”Tree-based state tying for high accuracy acoustic modelling.” Proceedings of the workshop on Human Language Technology. Association for Computational Linguistics, 1994.