EUROSPEECH '97
5th European Conference on Speech
Communication and Technology
Rhodes, Greece, September 22-25, 1997

SPEAKER NORMALIZATION TRAINING FOR MIXTURE STOCHASTIC TRAJECTORY MODEL

Irina Illina

CRIN/CNRS, INRIA-Lorraine
B.P. 239, 54506 Vandœuvre-lès-Nancy, France
illina@loria.fr

Yifan Gong

Speech Research
Media Technologies Laboratory
Texas Instruments
Dallas TX 75265, U.S.A.
Yifan.Gong@ti.com

ABSTRACT

In this paper we are interested in speaker and environment adaptation techniques for speaker independent (SI) continuous speech recognition. These techniques are used to reduce mismatch between training and the testing conditions, using a small amount of adaptation data. In addition to reducing this mismatch during the adaptation, we propose to reduce the variation due to speakers or environments during the training itself in the context of Speaker Normalisation (SN) approach, using MLLR transformation. SN also includes a combination of the context-dependent, phone dependent and broad phonetic class dependent information. The use of linear regression to model broad phonetic class dependent information assures our model to be used in the case that the adaptation data or training data is not given for some phonetic symbols. SN is developed for Mixture Stochastic Trajectory Model, a segment based model. The approach can be used for speaker, gender or environment normalization. We show the performance of SN compared to SI recognition and to MLLR speaker adaptation, through experiments on continuous speech recognition.

1. INTRODUCTION

The accuracy of speech recognition systems may degrade significantly when they are operated in test conditions that mismatch the training conditions. The mismatch may be due to changing speaker characteristics, varying speaker environments, different task constraints or a combination of these factors. Compensation of this mismatch is very important to practical application of recognition systems in real-world situations. Generally, distorsion appears as a combination of various acoustic differences and hits exact form is unknown [6]. In this paper, we are interested in modelling space representation of mismatch and structure-based compensation techniques. We assume that the differences between training and test conditions can be modeled by a linear transformation in the context of Maximum Likelihood Linear Regression (MLLR) [7].

Adaptation techniques use a small amount of adaptation data to reduce this mismatch. For example, in the context of speaker adaptation, they decrease the difference in performance between speaker dependent (SD) and SI system. However, such adaptation techniques only reduce the variation between training and testing conditions and has no effect on the variation due to speakers or environments during the training itself. Such variations may give high overlap among distributions of different speech units.

Recently, some research to reduce this overlap has been proposed. In [2, 3, 8], the model of decoupling speaker variation and phonetic variation is proposed for continuous density HMM. The speaker specific variation is modeled by linear transformation in the same way as in the adaptation techniques. with the difference that the transformation is integrated in the training process. This speaker normalization (SN) makes it possible to reduce the variance and hence the overlap of the acoustic models. In [1], the SN and the combination of context dependent and context independent information provide a new method of the SI training. In [5] a normalized SI model is generated by removing speaker characteristics during the training using a shift vector obtained by the MLLR technique. The experimental study of [9] demonstrates that speaker normalisation continues to be important even after significant amounts of speaker adaptation.

In this paper, we propose a SN approach for Mixture Stochastic Trajectory Model (MSTM) [4]. Compared to the above approaches for SN, our technique has several originalities. Firstly, we use a combination of context-dependent, phone dependent and broad phonetic class dependent information. It is possible to use the model when no adaptation data or training data is given for some phonetic symbols. Secondly, we integrate SN approach in a segment based model, a Mixture Stochastic Trajectory Model (MSTM). Finally, for training estimation problem we use EM algorithm and Bayesian estimation for adaptation. Through linear transformation used for normalization, the approach can be used for speaker, gender or environment normalization within the context of SI continuous speech recognition. In our implementation we use supervised batch adaptation, however, the technique can be extended to unsupervised and online adaptation modes.

The paper is organized as follows. We begin by brief presentation of MSTM and by giving the speaker normalisation model for MSTM. Next, the training and adaptation parameter estimation are presented in section 2. Section 3 give the experimental result for French continuous speech recognition. Finally, section 4 concludes this paper.

2. SPEAKER NORMALIZATION MODEL FOR MSTM

2.1. Acoustic Model of MSTM

MSTM is a segment-based model using phonemes as speech units and uses *a posteriori* distribution as acoustic model [4]. In order to model durational constraints, an observed segment Y of duration d for phoneme $s \in \mathbb{P}$ is rescaled linearly to a fixed-length sequence X:

$$Y = (y_1, \dots, y_d) \mapsto X = (x_1, \dots, x_Q), \quad x_i, y_i \in \mathbb{R}^D \qquad (1)$$

The probability density function (pdf) of the fixed-length sequence X^s is modeled using a mixture of trajectories T^s and is written as:

$$\begin{aligned} p(X|d,s,\lambda) &= \sum_{t_k \in T^s} Pr(t_k|d,s,\lambda) p(X|t_k,d,s,\lambda) \\ &\triangleq \sum_{t_k \in T^s} Pr(t_k|s,\lambda) p(X|t_k,d,s,\lambda) \qquad (2) \end{aligned}$$

where $Pr(t_k|s,\lambda)$ is the probability of trajectory cluster t_k, given the phoneme s and the model λ. T^s is the set of all trajectory clusters of phoneme s. The probability $p(X|t_k,d,s,\lambda)$ of the sequence X given trajectory cluster t_k and phoneme s is modelled by a multivariate Gaussian distribution defined on the whole observation sequence of s and is given by:

$$p(X|t_k,d,s,\lambda) \triangleq \mathcal{N}(X, \mu_k^s, \Sigma_k^s) \qquad (3)$$

where μ_k^s is a mean vector of dimension $M \times 1$, $M \triangleq D \times Q$, D is the dimension of the parameter space and Σ_k^s is a $M \times M$ covariance matrix. In this paper, we use diagonal covariance matrix. Thus the model parameter set λ is given by:

$$\lambda = \{\alpha_k^s, \mu_k^s, \Sigma_k^s\} \quad \forall t_k \in T^s, s \in \mathbb{P} \qquad (4)$$

where $\alpha_k^s \triangleq Pr(t_k|s)$. *A posteriori* distribution is used during the recognition and is given as follows:

$$Pr(s|X,d,\lambda) = \frac{p(X|d,s,\lambda)Pr(d|s,\lambda)Pr(s|\lambda)}{\sum_{v \in \mathbb{P}} p(X|d,v,\lambda)Pr(d|v,\lambda)Pr(v|\lambda)} \qquad (5)$$

where $Pr(d|s,\lambda)$ is the phone duration probability of symbol s, modeled by Gamma distribution, and $Pr(s|\lambda)$ is *a priori* phoneme probability of s. Sentence searching is accomplished by using $Pr(s|X^s,d,\lambda)$ in a dynamic programming algorithm. More details on this search as well as on parameter estimation of acoustic and duration model can be found in [4].

2.2. Speaker Normalization Model

In this section, we propose to separate the modeling of speaker variability by integrating the speaker normalization as part of the MSTM parameter estimation (training) problem. This allows reducing the inter-speaker variability of the training data and generating a more accurate acoustic model for speaker or environment adaptation. Such models have lower overlap between the distributions of different speech units due to their reduced inter-speaker variability and their potentially smaller variance. We also propose to combine broad phonetic class information, phone-dependent and context-dependent information. This allows the reduction of the overlap between different distributions of speech units.

Assume that the training data consists of speech from different speaker clusters. The pdf of speech trajectory $X^{l,s}$ from speaker cluster l for the phoneme s is defined as:

$$\begin{aligned} p(X^{l,s}|d,s,\lambda) &= \sum_{k \in T^s} Pr(t_k|s,\lambda) p(X^{l,s}|t_k,d,s,\lambda) \\ &= \sum_{k \in T^s} Pr(t_k|s,\lambda) \mathcal{N}(X^{l,s}, \mu_k^{l,s}, \Sigma_k^s) \qquad (6) \end{aligned}$$

where $\mathcal{N}(X^{l,s}, \mu_k^{l,s}, \Sigma_k^s)$ is a multivariate Gaussian distribution with mean $\mu_k^{l,s}$ and covariance matrix Σ_k^s. The basic assumption is to model $\mu_k^{l,s}$ as follows:

$$\mu_k^{l,s} = A^{c(s),l} \overline{\mu}^s + \Delta_k^s \qquad (7)$$

where $A^{c(s),l}$ is a $M \times (M+1)$ linear regression matrix including an additive bias vector and $\overline{\mu}^s$ is a extended mean vector of dimension $M+1$. $c(s)$ is the broad phonetic class of phoneme s, where $c(s) \in \{1, 2, \dots, R\}$ and R is the number of regression classes. Index l can represent the speaker cluster for speaker normalization, gender cluster for gender normalization or environment cluster for environment normalization. If the number of speakers is low, we can use one speaker per speaker cluster. If the number of speakers is high, the number of classes is chosen to assure trainability. Δ_k^s is a M dimension vector. $\mu_k^{s,l}$ represents the combination of speaker specific variation (transformation matrix $A^{c(s),l}$) and speaker independent information that models the phonetic variation ($\overline{\mu}^s$, Δ_k^s, Σ_k^s).

Our approach is a generalisation of the approach of [1]. Compared to [1], instead of using the combination of context-dependent ($\delta_{n,m}$) and context-independent (μ_r^l) information, we use the combination of context-dependent (Δ_k^s, Σ_k^s), phone-dependent ($\overline{\mu}^s$) and broad-phonetic class dependent ($A^{c(s),l}$) information. This allows to reduce the overlap between differents context-dependent units, the overlap between differents phone units and the overlap between broad phonetic classes. This model is more flexible than the [1]'s, because we use the regression matrix $A^{c(s),l}$ that allow training and adaptation when the adaptation data or training data are not given for some phonetic symbols for a speaker l.

2.3. Training Parameter Estimation

The whole parameter set λ to be estimated during training is:

$$\begin{aligned} \lambda = \{\overline{\mu}^s, A^{r,l}, \{\Delta_k^s, \Sigma_k^s, \alpha_k^s\}_{k=1}^{T^s} | s \in \mathbb{P}, \\ l = 1, \dots, L, \quad r = 1, \dots, R\} \end{aligned}$$

where $\alpha_k^s \triangleq P(t_k|s)$, $\mathbb{P}$ is the set of phonetic symbols and L is the number of speaker clusters. The parameters λ are derived according to the maximum likelihood estimation (MLE) criterion. The Expectation-Maximization (EM) algorithm is used to perform MLE estimation. The observed data is $\mathcal{X} = \{\mathcal{X}^{l,s} | s \in \mathbb{P}, \ \ l = 1, \dots, L\}$, $\ \mathcal{X}^{l,s} = \{X^{l,s} \in \mathbb{R}^M, \ \ s \in \mathbb{P}, \ \ l = 1, \dots, L\}$ and missing data is $\mathcal{Y} = \{\{t_k\}_{k=1}^{T^s} \ | \ s \in \mathbb{P}\}$.

The joint estimation of parameters given above is difficult because of their inter-dependency. We use iterative estimation by optimize over one parameter set while keeping the other fixed. This approach has the advantage of lower computation complexity.

In the following, we give the reestimation formulas for each parameter for the case of diagonal covariance matrix Σ_k^s:

- tied linear transformation $A^{r,l}$ is obtained by resolving the M systems of linear equations. Each system has $M+1$ linear equations with $M+1$ unknowns:

$$\begin{aligned} \sum_{s:c(s)=r} \sum_{k \in \mathcal{T}^s} \sum_{X^{l,s} \in \mathcal{X}^{l,s}} \beta_k^{s,l} (\Sigma_k^s)^{-1} (X^{l,s} - \Delta_k^s) \overline{\mu}^{s\#} \\ = \sum_{k \in \mathcal{T}^s} \sum_{X^{l,s} \in \mathcal{X}^{l,s}} \beta_k^{s,l} (\Sigma_k^s)^{-1} A^{r,l} \overline{\mu}^s \overline{\mu}^{s\#} \end{aligned}$$

where $\beta_k^{s,l} \triangleq Pr(t_k|X^{l,s}, s, \lambda')$, $\#$ stands for transposition operation and λ' is parameter set of previous EM-iteration;

- phone model extended mean vector $\overline{\mu}^s$ of dimension $M+1$ is obtained by resolving a system of $M+1$ linear equations with $M+1$ unknowns:

$$\sum_{l=1}^{L} \sum_{k \in \mathcal{T}^s} \sum_{X^{l,s} \in \mathcal{X}^{l,s}} \beta_k^{s,l} A^{c(s),l\#} (\Sigma_k^s)^{-1} (X^{l,s} - \Delta_k^s)$$

$$= \sum_{l=1}^{L} \sum_{k \in \mathcal{T}^s} \sum_{X^{l,s} \in \mathcal{X}^{l,s}} \beta_k^{s,l} A^{c(s),l\#} (\Sigma_k^s)^{-1} A^{c(s),l} \overline{\mu}^s$$

  where $A^{c(s),l}$ are updated;

- speaker independent vector Δ_k^s:

$$\Delta_k^s = \frac{\sum\limits_{l=1}^{L} \sum\limits_{X^{l,s} \in \mathcal{X}^{l,s}} \beta_k^{s,l} (X^{l,s} - A^{c(s),l} \overline{\mu}^s)}{\sum\limits_{l=1}^{L} \sum\limits_{X^{l,s} \in \mathcal{X}^{l,s}} \beta_k^{s,l}}$$

  where $A^{c(s),l}$ and $\overline{\mu}^s$ are updated;

- trajectory covariance matrix Σ_k^s:

$$\Sigma_k^s = \frac{\sum\limits_{l=1}^{L} \sum\limits_{X^{l,s} \in \mathcal{X}^{l,s}} \beta_k^{s,l} G^{\#} G}{\sum\limits_{l=1}^{L} \sum\limits_{X^{l,s} \in \mathcal{X}^{l,s}} \beta_k^{s,l}}$$

  where $A^{c(s),l}$, $\overline{\mu}^s$, Δ_k^s are updated and $G \triangleq X^{l,s} - A^{c(s),l} \overline{\mu}^s - \Delta_k^s$;

- a priori distribution of trajectory component α_k^s:

$$\alpha_k^s = \frac{\sum\limits_{l=1}^{L} \sum\limits_{X^{l,s} \in \mathcal{X}^{l,s}} \beta_k^{s,l}}{\sum\limits_{l=1}^{L} card(\mathcal{X}^{l,s})}$$

  where $card(\mathcal{X}^{l,s})$ stands for cardinality of set $\mathcal{X}^{l,s}$.

2.4. Adaptation Parameter Estimation

For the adaptation, the parameter set to be estimated is $\lambda = \{A^{r,l}\}$, where index l represent the adaptation speaker and can be omitted. We assume that $A^{r,l}$ is a diagonal matrix of dimension $M \times (M+1)$. Since for adaptation the available data may be sparse, we use Tied MAP Estimation of General Linear Transformation [12]. Bayesian approach provides a convenient method for combining sample observations and prior information. It may take full advantage of large amounts of adaptation data, as asymptotically the estimates converge to speaker-dependent values.

We give the estimation formulas for the case of diagonal transformation with the additive bias. For notational simplicity, we separate the M dimensional problem of λ-estimation in M independent problems of one dimensional problems of estimates $\lambda_i = \{a_i^r, b_i^r\}$, $i = 1, \ldots, M$, where b_i^r is a transformation additive bias. We assume that the conjugate prior distribution for the univariate Gaussian vector has a normal-gamma prior density. For each i-th estimation problem, the MAP estimate of a_i^r and b_i^r is a solution of the system:

$$\begin{bmatrix} r_0 + A & r_0 w + B \\ r_0 + B & r_0 w^2 + C \end{bmatrix} \begin{bmatrix} b_i^r \\ a_i^r \end{bmatrix} = \begin{bmatrix} r_0 w + D - (r_0 + A)\delta_k^s \\ r_0 w^2 + E - (r_0 w + B)\delta_k^s \end{bmatrix}$$

where r_0 is a conjugate prior distribution parameter, that can be adjusted experimentally. We denote:

$$A \triangleq \sum_{s:c(s)=r} \sum_{k \in \mathcal{T}^s} \sum_{X \in \mathcal{X}^s} \beta_k^s \sigma_k^s, \; B \triangleq \sum_{s:c(s)=r} \sum_{k \in \mathcal{T}^s} \sum_{X \in \mathcal{X}^s} \beta_k^s \sigma_k^s m^s,$$

$$C \triangleq \sum_{s:c(s)=r} \sum_{k \in \mathcal{T}^s} \sum_{X \in \mathcal{X}^s} \beta_k^s \sigma_k^s (m^s)^2,$$

$$D \triangleq \sum_{s:c(s)=r} \sum_{k \in \mathcal{T}^s} \sum_{X \in \mathcal{X}^s} \beta_k^s \sigma_k^s x^s,$$

$$E \triangleq \sum_{s:c(s)=r} \sum_{k \in \mathcal{T}^s} \sum_{X \in \mathcal{X}^s} \beta_k^s \sigma_k^s m_s x^s$$

where $\beta_k^s \triangleq Pr(t_k|X^s, s, \lambda')$, σ_k^s is the i-th element of $(\Sigma_k^s)^{-1}$, m^s is the i-th element of $\overline{\mu}^s$, x^s is the i-th element of adaptation data X^s, w is the i-th element of $\overline{\mu}^s$ and δ_k^r is the i-th element of a vector $\overline{\Delta}_k^r$ defined as follow: $\overline{\Delta}_k^r = \frac{\sum_{s:c(s)=r} \sum_{k \in \mathcal{T}^s} \beta_k^s \Delta_k^s}{\sum_{s:c(s)=r} \sum_{k \in \mathcal{T}^s} \beta_k^s}$

3. EXPERIMETS AND RESULTS

3.1. Experimental conditions

Experiments deal with a French continuous speech corpus recorded by the CRIN/INRIA laboratory. For training, 79 phonetically rich sentences were read by 7 French speakers (1 female). On the average, there are about 70 observations per phoneme for each speaker. For testing, 241 sentences were recorded for each speaker. There is only a small overlap between training and test vocabularies. The observation vectors are 13 MFCC including a normalized energy. For this corpus, 32 context-independent phone models, including one silence model, are built. The language model has a word-pair equivalent perplexity of 49 and a 2010 words vocabulary. In all experiments, the covariance matrix is assumed to be diagonal. The training is performed with the speech of 6 speakers and testing with the speech of the 7-th speaker. For speaker classification we used ascendent speaker classification with the distance measure TR2 of [10]. The number of speaker clusters is 2 ($L = 2$). We use 3 regression classes: one for vowels, one for silence symbol and one for all other phonemes. For adaptation, 10 sentences (about 20 seconds of speech) is used. We use supervised batch adaptation.

3.2. Results

Figure 1 give the results of 4 approaches in term of word error rate (WER): SI speech recognition, Maximum Likelihood Linear Regression adaptation (MLLR) [11] (MLLR is used only during the adaptation), SN training and adaptation (SNT) with MLE ($r_0 = 0$) and SNT with MAP ($r_0 = 3.$). The parameter r_0 is adjusted experimentally. The best result is obtained with SNT with MAP approach, using about 1200 pdfs. This represents 24% reduction in error rate compared to SI recognition, 8% error reduction compared to MMLR approach and 2% reduction compared to SNT with MLE approach with the same number of pdf for all approach.

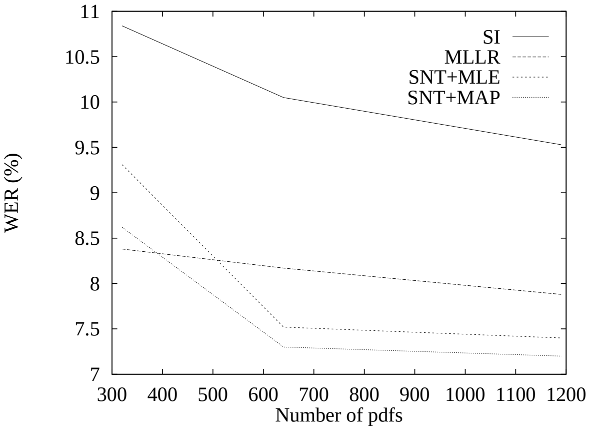

Figure 1: Word Error Rate (%) as function of number of pdfs and the approach

4. CONCLUSION

In this paper, we have reported an extension of the work in speaker or environment adaptative training for SI continuous speech recognition. We have proposed a speaker normalization training for the Mixture Stochastic Trajectory Model. Using speaker-dependent transformation and speaker-independent mean and variance (Eq-7), the approach models separately the speaker (or environment) variation and phonetic variation, and therefore reduces the distribution overlap between SI models of differents speech units. The model can be used in the case where, for some phonetic units, the adaptation or training data are not given. During the training, the parameters of the model are estimated according to the MLE criterion, and during the supervised batch adaptation according to Bayesian estimation. The results show that the new acoustic models give more efficient mismatch reduction between training and the testing conditions than SI recognition or MLLR adaptation approach.

Out future work includes evaluating the technique to environment and gender normalisation, and integrating SN approach in the context of unsupervised adaptation and incremental adaptation.

5. ACKNOWLEDGEMENT

The authors would like to express their appreciation to Dr. Mohamed Afify for fruitful discussions about this work and to Prof. Jean-Paul Haton for encouragement throughout this work.

6. REFERENCES

[1] A. Acero and X. Huang. Speaker and Gender Normalization for Continuous-Density Hidden Markov Models. *In Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP'96*, 1:342–345, 1996.

[2] T. Anastasakos, J. McDonough, and J. Makhoul. Speaker Adaptive Training: a Maximum Likelihood Approach to Speaker Normalization. *In Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP'97*, 2:1043–1046, 1997.

[3] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul. A Compact Model for Speaker-Adaptive Training. *In Proc. of Int. Conf. on Spoken Language Processing, ICSLP'96*, 2, 1996.

[4] Y. Gong. Stochastic Trajectory Modeling and Sentence Searching for Continuous Speech Recognition. *IEEE Trans. on Speech and Audio Processing*, 5(1):33–44, Jan. 1997.

[5] J. Ishii and M. Tonomura. Speaker Normalization and Adaptation Based on Linear Transformation. *In Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP'97*, 2:1055–1558, 1997.

[6] C.-H. Lee. On Feature and Model Compensation Approach to Robust Speech Recognition. *Workshop of ESCA-NATO*, pages 45–54, 1997.

[7] C. J. Leggetter and P. C. Woodland. Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density HMMs. *Computer, Speech and Language*, 9(2):171–185, 1995.

[8] J. McDonough, T. Anastasakos, G. Zavaliagkos, and H. Gish. Speaker-Adapted Training on the Switchboard Corpus. *In Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP'97*, 2:1059–1062, 1997.

[9] V. Nagesha and L. Gillick. Studies in Transformation-Based Adaptation. *In Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP'97*, 2:1031–1033, 1997.

[10] K. T. Ng, H. Li, and J. P. Haton. Some Nonparametric Distance Measures in Speaker Verification. *In Proc. of European Conf. on Speech Communication and Technology*, 1:317–320, September 1995. Madrid, Spain.

[11] O. Siohan, Y. Gong, and J.-P. Haton. Comparative Experiments of Several Adaptation Approaches to Noisy Speech Recognition using Stochastic Trajectory Model. *Speech Communication*, 18(4):335–352, 1996.

[12] G. Zavaliagkos. Maximum a Posteriori Adaptation Techniques for Speech Recognition. *Ph.D. Thesis, Northeastern University*, May 1995.