

Speaker Adaptation using Regression Classes Generated by Phonetic Decision Tree-based Successive State Splitting

Se-Jin Oh[†], Kwang-Dong Kim[†], Duk-Gyoo Roh[†], Woo-Chang Sung[‡], Hyun-Yeol Chung[‡]

[†]Korean VLBI Network Group, Korea Astronomy Observatory, Korea

[‡]School of Electrical Eng. & Computer Science, Yeungnam University, Korea

{sjoh,kdkim,dgroh}@trao.re.kr, wcsung@hanmail.net, hychung@yu.ac.kr

Abstract

In this paper, we propose a new generation of regression classes for MLLR speaker adaptation method using the PDT-SSS algorithm so as to represent the characteristic of speaker effectively. This method extends the state splitting through clustering the context components of adaptation data into a tree structure. It enables to autonomously control a number of adaptation parameters (mean, variance) depending on the context information and the amount of adaptation utterances from a new speaker. Through the experiments, the phone and word recognition rates with adaptation have an average 34~37%, 9% higher accuracy than the speaker-independent acoustic models, respectively. The experimental results of Korean phone and word recognition confirmed the significant performance increase in small adaptation utterances compared with without any speaker adaptation.

1. Introduction

The most popular approach to model adaptation is through Bayesian formulation. For example, maximum a posteriori (MAP) estimation algorithm have been widely adopted recently and successfully applied to speaker adaptation [1].

Another category of adaptation techniques, which do not use the MAP framework, are often referred to as transformation-based approaches, such as vector field smoothing (VFS), maximum likelihood linear regression (MLLR) adaptation [1][2]. This techniques limits the number of free parameters by tying the HMM parameters or by applying some constraints on the parameters in order to improve recognition accuracies with a small amount of data. When the amount of adaptation data exceeds a certain value, however, the recognition accuracy often becomes inferior to that obtained with ML (Maximum Likelihood) estimation of the model parameters. This is because a model with a small number of free parameter could not fully utilize the potential information embedded in the large amount of data.

In the above adaptation techniques, we now focus on the study of MLLR adaptation. The MLLR has proven to be an effective speaker adaptation technique in the presence of limited adaptation data [2] [3]. A set of linear transformations for the mean, variance parameters of a mixture Gaussian HMM system is estimated such that the likelihood of the adaptation data is maximized. Since in general there is little adaptation data compared to the number of model parameters, it is necessary to cluster model parameters together into regression classes. It is assumed that all components in a given regression class transform in a similar fashion. Usually the number of regression classes is determined dynamically according to the amount of adaptation data available using a

regression class tree. If only few adaptation data are present a single transformation matrix at the tree root is calculated and applied to all models. The more adaptation data become available, the further the tree is descended and the more specific transformation matrices are computed.

An important question is which model parameters to cluster together into regression classes. Two approaches to design of regression class trees are common practice [3] [4].

- ① Phonetic knowledge: Here, expert knowledge is used to decide which components are to be transformed together. The components are split according to broad phonetic classes (e.g., nasals, glides) or, at a lower level, into phones. We denote this tree a "broad phonetic class" tree.
- ② Acoustic space: Components are clustered according to how close they are in acoustic space, irrespective of which phone they belong to. This has the advantage of being a "data-driven" approach with no need for expert knowledge. However, the resulting classes usually cannot be assigned a phonetic identity.

Neither approach, however, is necessarily optimal, in the sense of maximizing the likelihood of the adaptation data. In [3] it has been attempted to arrive at regression class trees that are closer to the maximum likelihood solution by employing an iterative procedure starting from an initial acoustic clustering. However, unfortunately no improvement in error rate has been observed.

The problem addressed here is as follows: first, we would like to get rid of the phonetic expertise required in the first approach but still obtain regression classes which represent broad phonetic classes and which deliver adaptation performance comparable to a hand-designed tree. If phonetic expertise is no longer required, the recognizer can be faster transferred to a new language, of which no in-depth phonetic expertise might be available. Second, adaptation data is closely related with the utterance style of speaker. That is, adaptation procedure is performed after getting the exact transcription and speaker utterance according to the text with selected various phones. It is highly influenced by the speaker's utterance style. Finally, the context information of speaker's utterance has an influence to adapt reference models.

In this paper, to solve these problems, we propose the new generation of regression classes using the PDT-SSS (Phonetic Decision Tree-based Successive State Splitting) [7] algorithm on the supervised MLLR adaptation. Especially, contextual state splitting in the PDT-SSS algorithm is adopted to find the contextual information for the regression classes.

To test the effectiveness of proposed algorithm, we performed the phone and word recognition experiments using the ETRI (Electronics and Telecommunications Research Institute) and KLE (Korean Language Engineering) speech

databases. The speaker-independent reference models are trained by using the ETRI speech databases uttered by 400 speakers (200 male and 200 female) based on the hidden Markov network (HM-Net) [5]. This model structure is context-dependent model with optimal states. The speaker adaptation is carried out for the KLE each 35 male and 32 female speakers second utterance using the MLLR technique. The phone and word recognition experiments are performed using the one-pass Viterbi beam search algorithm [8] with phone-pair and word-pair grammar.

This paper is organized as follows: in section 2, we briefly introduce the MLLR and the use of regression class trees. In section 3, the proposed algorithm using the PDT-SSS is derived and section 4 presents experimental results. Finally in section 5, we summarize the conclusion.

2. MLLR adaptation

Maximum likelihood linear regression (MLLR) computes a set of transformations that will reduce the mismatch between an initial model set and the adaptation data. More specifically the MLLR is a model adaptation technique that estimates a set of linear transformations for the mean and variance parameters of a Gaussian mixture HMM system. The effect of these transformations is to shift the component means and alter the variances in the initial system so that each state in the HMM system is more likely to generate the adaptation data.

In the MLLR, the new estimate of the mean μ_i , is obtained by equation (1).

$$\mu_i = A_c \xi_i + b_c \quad (1),$$

where the matrix A_c and the bias term b_c are the MLLR parameters of regression class c . ξ_i is the original (speaker-independent) mean vector. The MLLR parameter A_c and b_c are estimated such that the likelihood of the adaptation data for class c is maximized.

A regression class tree consists of a hierarchy of regression classes and a set of base classes as leaves. All base classes below a tree node may share a common transformation matrix. The number of different transformation matrices is chosen according to the amount of adaptation data available. If only few adaptation data are present a single transformation matrix at the tree root is calculated and applied to all base classes. The more adaptation data become available, the further the tree is descended and the more specific transformation matrices are computed.

A clustering of base classes into regression classes is considered optimal if the resulting regression class tree maximizes the likelihood of the adaptation data. In [3] iterative schemes are proposed to approximate the computationally intractable optimal solution by optimizing the likelihood criterion locally at each split of the tree. The algorithm starts with an initial regression class tree obtained from acoustic clustering. In the next section we derive an algorithm to arrive at such an initial regression class tree.

3. Regression Class

The regression class tree [3] is constructed so as to cluster together components that are close in acoustic space, so that

similar components can be transformed in a similar way. Note that the tree is built using the original speaker-independent model set, and is thus independent of any new speaker. Fig. 1 is an example of a decision tree for the phone /k/, along with some actual questions. The root represents the collection of allophones of a given phone (the context-independent phone). Each split of an internal node is a disjoint partition of allophones residing in the parent node based on some question about the contextual information of allophones. Since the clustering is actually done by splitting from the root recursively, it is referred as top-down clustering, in contrast to bottom-up clustering. It is trivial to find the allophonic cluster for every allophone based on the decision tree, no matter whether the allophone is seen in the training data or not.

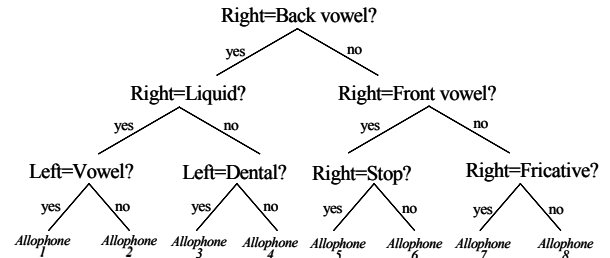


Figure 1. An example of a decision tree that clusters the allophones of the phone /k/.

3.1. Generation algorithm for regression class

After constructing the question set for decision trees, we must decide which question can be used for a node to achieve optimal splitting. The objective of a decision tree is used to reduce the uncertainty of the event being decided upon.

The PDT-SSS algorithm, which is used to generate the regression class, is the method to grow the phonetic decision tree by iterative clustering the node from the root of tree according to the phone question. The growing of tree is carried out according to the state position of context-independent HMM. Basic idea is that all splitting states are located in the root of tree and the state splitting is iteratively carried out by best question set according to the splitting reference. And then the state is shared in each leaf which is terminated. The most notable feature of the PDT-SSS is that acceptable context classes are split by phonetic questions. And furthermore, when a state is split, there two mixtures aren't associated with the new states respectively, but new single Gaussian distributions are made from the appropriate training samples. Hence, highly accurate HM-Nets can be generated with the PDT-SSS since they can represent any context, and their context classes are split appropriately. By considering this point of view, we propose the new generation of regression class using the contextual domain splitting of the PDT-SSS algorithm, described as follows:

- ① Train the context-dependent acoustic models (triphone) and its statistic file.
- ② Locate a set of appropriate linguistic questions manually and prepare the prototype one-state HM-Net without time domain splitting.
- ③ Select the state with the most variability using equation (2)(see below).
- ④ Split the state on a contextual domain for each question.

- a) Split the acceptable context class and derive two single Gaussian distribution according to the question, where each Gaussian is corresponding to either *yes* or *no*.
- b) Assign each context class and Gaussian to the new states.

- ⑤ Choose the best regression class on the contextual domain based on the context class of questions.

In step 3, a variability of a state $S(i)$ is calculated by following equation.

$$d_i = n_i \sum_{p=1}^P \frac{\sigma_{ip}^2}{\sigma_{Tp}^2} \quad (2),$$

where σ_{ip}^2 and σ_{Tp}^2 are the p-th variances of training samples associated with the states $S(i)$ and all other states, respectively, n_i is the number of training samples of the states $S(i)$, and P is the dimension of the feature vectors.

The details of step 4 are described as followed. In this step, the state $S(m)$ chosen in the previous step is split into two states $S'(m)$ and $S(M)$ where m, M are the states number of models ($0 \leq m < M$).

On the contextual domain, the objective is to determine the most suitable question to split $S(m)$, and concatenates $S'(m)$ and $S(M)$ in parallel. Since the frame set can be divided into two subsets by each question, the new single Gaussian distribution can be divided by calculating only the mean and the variance of each subset. In state $S(m)$, $C(m)$ defines acceptable context classes, and c_k is k-th factor. In the contextual state splitting, for the purpose of calculating the whole frame f_k , the mean μ_k , and the variance σ_k^2 are assigned to each c_k , the distribution of *yes* and *no* can be calculated by following equations with divided $C(m)$ for the question q :

$$\mu_{q,yes} = \frac{\sum_{c_k \in Q_q} f_k \mu_k}{\sum_{c_k \in Q} f_k} \quad (3)$$

$$\sigma_{q,yes}^2 = \frac{\sum_{c_k \in Q_q} f_k \{ \sigma_k^2 + (\mu_k - \mu_{q,yes})^2 \}}{\sum_{c_k \in Q_q} f_k} \quad (4)$$

$$\mu_{q,no} = \frac{\sum_{c_k \in Q_q} f_k \mu_k}{\sum_{c_k \in Q} f_k} \quad (5)$$

$$\sigma_{q,no}^2 = \frac{\sum_{c_k \in Q_q} f_k \{ \sigma_k^2 + (\mu_k - \mu_{q,no})^2 \}}{\sum_{c_k \in Q_q} f_k} \quad (6),$$

where $\mu_{q,yes}, \sigma_{q,yes}^2, \mu_{q,no}, \sigma_{q,no}^2$ are the mean and the variance of *yes/no* side with divided $C(m)$ for the question q , respectively. Therefore Q_q is the context class to be a 'yes' by the question q .

4. Experimental Results

4.1. Speech databases and analysis

The Korean speech databases of ETRI for training and KLE for adapting and testing were used. The ETRI speech databases consist of 10,000 words, 10,000 digits, and 100,000 sentences uttered by 1000 speakers (500 male and 500 female) with words, digits, single syllable, and sentences per speaker. In this paper, 500 speakers' utterances (250 male and 250 female) were used for assessing on this system in the ETRI databases. Among the selected 400 speakers' utterances (200 male and 200 female) were used for building gender independent context-dependent acoustic models (HM-Net). In addition, the remaining 100 speakers (50 male and 50 female) were used for testing the reference models. The KLE speech databases consist of 452 words with phoneme-balanced words (PBWs) uttered twice by 35 male and 32 female speakers. The first utterances of speakers were used for the speaker-independent and speaker-dependent test. The second utterances of speakers were used for each speaker adaptation for reference models.

All speech data were sampled at 16 kHz with 16 bits quantization, pre-emphasized with a transfer function of $1 - 0.97z^{-1}$ and windowed using a 25 ms Hamming window with a 10 ms shift. The data parameterization consists of 12 LPC MEL-cepstrum coefficients and is normalized log-power plus 1st and 2nd order differences.

The 152 questions were prepared for the PDT-SSS training, where 76 phone classes based on Korean phonological rules [9] were paired with the contextual factors (base, preceding, and succeeding phones). Initial topology of reference HM-Net model was a 126-state in which 43 context-independent phone models were connected in parallel. An HM-Net model was constructed to 2000 states with 4 mixtures.

4.2. Adaptation Experiments

For the following experiments we used 75 question set as base classes to generate the regression class tree. The regression class tree generated by the PDT-SSS is shown in Fig. 2. To evaluate the quality of the regression class, recognition experiments with offline the supervised MLLR adaptation have been carried out on the KLE speech databases. We carried out the three experiments, such as speaker and task-independent word recognition, preliminary speaker-dependent word recognition to find the best threshold, and speaker-dependent phone/word recognition. Speech recognition was performed using the one-pass Viterbi beam search algorithm with word-pair grammar for word recognition and phone-pair grammar for phone recognition.

The reference speaker-independent phone and word recognition rate without any adaptation is an average 22.8%, 88.7% for the 35 male speakers test database and an average 18.1%, 88.7% for the 32 female speakers, respectively.

To find the best threshold of the MLLR appropriate for the amount of adaptation data available, the sample speaker-dependent word recognition experiments have been carried out on the various thresholds. In case of the KLE adaptation data, 600 selected the best threshold. Through this threshold and regression classes, the speaker-dependent phone and word recognition rate with the MLLR adaptation is an average 57.4%, 97.6% for the 35 male speakers and an

average 56.0%, 98.0% for the 32 female speakers, respectively. It can be seen from the results in Table 1 and Fig. 3.

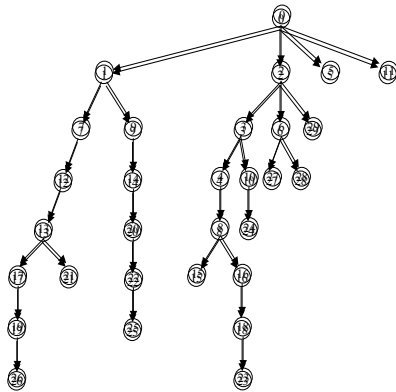


Figure 2. The regression class tree obtained from clustering of phones with contextual state splitting of the PDT-SSS.

Table 1: The recognition accuracy for 452 phone and word.

Task	Speaker	Adaptation	Recog. Rate(%)
Phone 452	35 male	without	22.8
		with	57.4
	32 female	without	18.1
		with	56.0
Word 452	35 male	without	88.7
		with	97.6
	32 female	without	88.7
		with	98.0

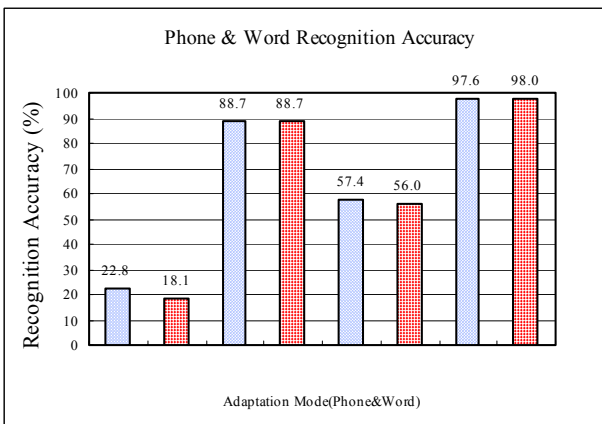


Figure 3. Recognition accuracy.

As can be seen in Table 1 and Fig. 3, we obtained about 34~37% relative improvement for phone recognition experiments, and also obtained about 9% relative improvement for word recognition experiments using the regression class of the MLLR adaptation. Through the experimental results, we have verified the effectiveness of state splitting algorithm to generate the regression classes.

5. Conclusions

In this paper, we have proposed a new generation of regression classes for the MLLR adaptation using the PDT-SSS algorithm in order to represent the characteristic of

adaptation speaker effectively. The proposed algorithm is as follows: we first made the question set with center of phone environments for phonetic decision tree manually. And then, the regression classes are generated by the contextual domain state splitting of the PDT-SSS and by the question set to find out the effect of context information. Finally, the MLLR adaptation is carried out using the regression class for each speaker. In this paper, the contextual state splitting is only adopted to find the contextual information for the regression class in the PDT-SSS. To test the effectiveness of proposed algorithm, we have performed the phone and word recognition experiments using the ETRI and KLE speech databases. Through the experiments, the phone and word recognition rates with the adapted model have an average 34~37%, 9% higher accuracy than the speaker-independent acoustic models. These results show that regression classes generated by the PDT-SSS are well adopted to the MLLR adaptation.

6. Acknowledgements

The authors would like to thank Takaaki HORI of NTT for his contribution at Japan and would also like to thank the ETRI and KLE for providing us with Korean large vocabulary speech databases.

7. References

- [1] J.L. Gauvain and C.H. Lee, "Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains," *IEEE Trans. Speech Audio Processing*, 2:291-298, 1994.
- [2] M.J.E. Gales and P.C. Woodland, "Variance Compensation within the MLLR Framework," *Technical Report CUED/F-INFENG/TR242*, Cambridge University, 1996.
- [3] M.J.E. Gales, "The Generation and Use of Regression Class Trees for MLLR Adaptation," *Technical Report CUED/F-INFENG/TR263*, Cambridge University, 1996.
- [4] Reihold Haeb-Umbach, "Automatic Generation of Phonetic Regression Class Trees for MLLR Adaptation," *IEEE Trans. on Speech and Audio Processing*, 9(3):299-302, 2001.
- [5] J. Takami and S. Sagayama, "A Successive State Splitting Algorithm for Efficient Allophone Modeling," *Proc. Of ICASSP'92*, 1:573-576, 1992.
- [6] M. Ostendorf and H. Singer, "HMM Topology design Using Maximum Likelihood Successive State Splitting," *Computer Speech and Language* 11:17-41, 1997.
- [7] T. Hori, M. Katoh, A. Ito, and M. Kohda, "A Study on HM-Nets using Decision Tree-based Successive State Splitting," *Proc. of ICSP'97*, 2:383-387, 1997.
- [8] L.R. Rabiner and B.H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, 1993.
- [9] H. Y. Lee, *Korean Phonetics*, Taehak Press, Korea, 1996.