ISCA Archive
http://www.isca-speech.org/archive

ITRW on Speech Recognition and
Intrinsic Variatioon (SRIV 2006)
Toulouse, France
May 20, 2006

UTTERANCE-BASED SELECTIVE TRAINING FOR COST-EFFECTIVE TASK-ADAPTATION OF ACOUSTIC MODELS

Tobias Cincarek, Tomoki Toda, Hiroshi Saruwatari and Kiyohiro Shikano

Graduate School of Information Science
Nara Institute of Science and Technology, Japan
cincar-t@is.naist.jp

ABSTRACT

The construction of acoustic models for speech recognition systems is a very costly and time-consuming process, since their robust training requires large amounts of transcribed speech data, which have to be collected and labeled by humans. This paper describes an approach for cost-effective construction of task-adapted acoustic models. Existing speech data(bases) are employed to set up a large training data pool. Apart from that, only a small amount of task-specific speech data is required. Based on an algorithm for utterance-based selective training of acoustic models, training utterances are selected from the training data pool so that the likelihood of the acoustic model given the task-specific speech data is maximized. The proposed method is evaluated for acoustic models with context-independent and context-dependent phonetic units. Results are reported for building an infant (preschool children) acoustic model with speech from elementary school children and an elderly acoustic model with adult speech. The proposed approach is already effective if there are only 20 task-specific utterances available. A relative improvement in word accuracy of up to 10% is achieved over conventional acoustic model construction and up to 2.8% over MAP and MLLR adaptation with the task-specific data. The gap in performance to an acoustic model trained on large amounts of task-specific data was reduced up to 76%.

1. INTRODUCTION

The development of applications making use of ASR technology requires the construction of a high-performance acoustic and language model. However, the costs for providing these models are very high. In this paper the focus is on the acoustic model. For example, about half of the relative costs to develop an interactive dialogue system are due to speech database preparation [1]. This is due to the fact, that at the beginning of the development cycle large amounts of task-specific speech data have to be collected and labeled by humans, which is very costly and time-consuming. It is impractical to provide enough human-labeled speech data for each possible combination of the various factors which have an influence on the recognition performance such as speaker characteristics (e.g. gender, age, accent), speaking style (e.g. read, spontaneous), domain (e.g. digits, commands, news, dialogue) and acoustic conditions (e.g. background noise, reverberation, microphone).

There are several proposals in literature to reduce the costs of acoustic modeling. Among them are attempts to build task-independent acoustic models, which are portable among different applications by combining speech data from multiple sources [2], employment of active learning [3, 4], unsupervised learning [5, 6] or both [7] to reduce the effort necessary for speech data transcription, and training [4] or adaptation [8, 9] methods, which make selective use of existing speech data resources. The application of active learning revealed, that the best model is not necessarily obtained when using all available training data but rather a subset and that a model with equal performance can be constructed with a smaller amount of carefully selected training data [4, 7].

By employing the above-mentioned methods for active and unsupervised learning, the costs for transcribing speech data can be reduced. However, both methods still require large amounts of task-specific speech data. Furthermore, the selection of training utterances is restricted either to data with high confidence (unsupervised training) or data, which are difficult to recognize (active learning).

In the following, an approach for cost-effective construction of task-adapted acoustic models is proposed, which can be effective even in case of only a few task-specific example data. Moreover, it provides means to select the desired training speech data based on a likelihood criterion. Instead of using all data available from multiple sources for training [2], a subset which is acoustically close to the task-specific data is selected. This is realized by the utterance-based selective training algorithm [10] which automatically chooses the training utterances from a large data pool so that the model likelihood given some task-specific example data is maximized. Unlike [8, 9] the proposed method is not limited to speaker-based selection and can already be effective with less (transcribed) task-specific development data than would be required for active or unsupervised learning.

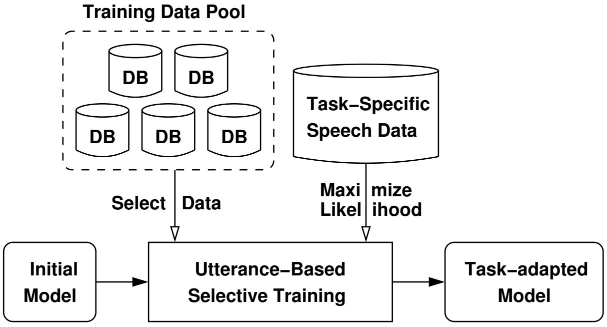

Fig. 1. Acoustic Model Construction Framework.

This paper is structured as follows: Section 2 explains the overall approach for cost-effective adaptation of acoustic models, the criterion for training data selection and the selection algorithm. Section 3 describes the setup of two evaluation experiments to verify the effectiveness of the proposed approach. Section 4 shows and discusses the experimental results. Section 5 summarizes results and mentions future work.

2. PROPOSED APPROACH

2.1. Acoustic Model Construction Framework

Figure 1 depicts the proposed framework for cost-effective task-adaptation of acoustic models based on selective training. It is assumed that one or more speech databases are available, which are combined to a large training data pool. Furthermore, a set of task-specific speech data is required. Although its size and configuration will depend on the desired target model complexity, it should be kept small in order to achieve cost reduction. The aim is to achieve good performance with less than 1,000 task-specific utterances. In order to be independent from the characteristics of a few speakers, the task-specific data set should ideally contain utterances from as many speakers as possible and cover the acoustic characteristics of the target environment for which the ASR system is to be built.

The procedure to build the task-adapted model is as follows: First, an initial acoustic model is build from the training data pool. Next, automatic selection of training utterances is carried out. The central idea is to select those utterances from the data pool, which maximize the likelihood of the task-specific data. Finally, the selected training data and the task-specific data are employed to build the task-adapted acoustic model, e.g. retraining the initial model with the selected data and optional adaptation with the task-specific data.

2.2. Utterance-based Selective Training Algorithm

In the following, the selection criterion and the selection algorithm as proposed in [10] are described briefly. A (sub-optimal) maximum likelihood estimate for the parameters of an HMM/GMM-based acoustic model can be obtained iteratively with the Expectation-Maximization (EM) algorithm [11]. The estimation is carried out so that the model likelihood given the training data is maximized. Here, the idea is to maximize the model likelihood given the task-specific data by selecting an appropriate subset for parameter estimation. Fast calculation of the likelihood is possible via the auxiliary Q-function using sufficient statistics. The Q-function is defined as follows:

$$Q(\mathbf{\Theta},\hat{\mathbf{\Theta}}) = \sum_{\vec{s}} P(\vec{s}|\mathcal{D},\mathbf{\Theta}) \log P(\vec{s},\mathcal{D}|\hat{\mathbf{\Theta}}) \qquad (1)$$

$\vec{s}$ denotes the (Gaussian) mixture and (HMM) state index sequence, $\mathcal{D}$ are the task-specific development data. In case of an HMM-based acoustic model with Gaussian mixture densities the output density part of Q can be written as

$$\propto \sum_q \sum_m y_{qm} \log \frac{\hat{w}_{qm}}{\sqrt{2\pi\hat{\sigma}^2_{qm}}} \qquad (2)$$

$$- \sum_q \sum_m \frac{z_{qm} - 2\hat{\mu}_{qm} o_{qm} + \hat{\mu}^2_{qm} y_{qm}}{2\hat{\sigma}^2_{qm}} \qquad (3)$$

where the sufficient statistics $\mathbf{S}_{\mathcal{D}}$ of the task-specific data $\mathcal{D}$ are given by the variables y_{qm} (occupancies), o_{qm} (means) and z_{qm} (second-order moments). The new model parameters $\hat{\mu}_{qm}$ (mean), $\hat{\sigma}^2_{qm}$ (variance) and $\hat{w}_{qm}$ (mixture weights) can be written as a function of the sufficient statistics $\mathbf{S}_{\mathcal{T}}$ of the training data $\mathcal{T}$, which are decomposable w.r.t. the training utterances $\mathbf{u}_i$. q and m denote the state and mixture index, respectively. An increase of the Q-function implies an increase of the likelihood $P(\mathcal{D}|\hat{\mathbf{\Theta}})$.

Since there are extremely many possibilities to select an utterance subset from a large data pool, a heuristic selection strategy has to be employed. The delete scan (*ST_DelScan*) algorithm, a greedy search technique, examines each utterance in the training data pool only once. An utterance is discarded if its independent deletion results in a likelihood increase.

The main steps of the *ST_DelScan* algorithm are:

1. Calculate and store the sufficient statistics $\mathbf{S}_{\mathcal{T}}, \mathbf{S}_{\mathcal{D}}, \mathbf{S}_i$ for the whole training $\mathcal{T}$, the whole task-specific data $\mathcal{D}$ and each training utterance $\mathbf{u}_i$.

2. Calculate the initial model likelihood l given the task-specific data $\mathcal{D}$ via the Q-function based on the sufficient statistics $\mathbf{S}_{\mathcal{T}}$ and $\mathbf{S}_{\mathcal{D}}$.

3. For each utterance $\mathbf{u}_i$ in the training data pool $\mathcal{T}$ do:

   a. Exclude $\mathbf{u}_i$ from $\mathcal{T}$ temporarily.

   b. Calculate the likelihood l' of the model trained on the remaining utterances in $\mathcal{T}$ based on $\mathbf{S}_{\mathcal{T}}$, $\mathbf{S}_i$ and $\mathbf{S}_{\mathcal{D}}$.

   c. If the model likelihood increases, i.e. $l' > l$, $\mathbf{u}_i$ should not be used for training.

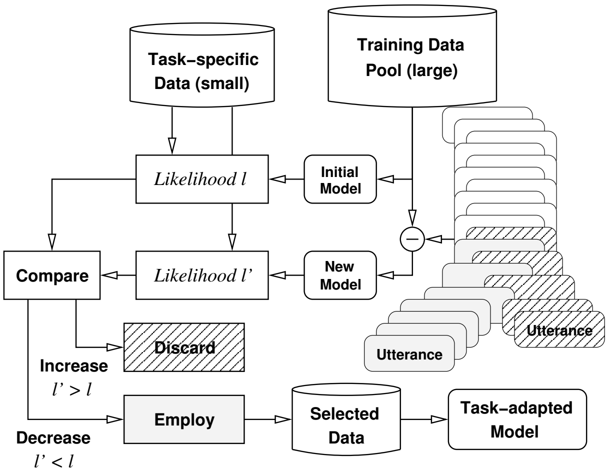

Fig. 2. Illustration of utterance-based selective training.

d. Otherwise $\mathbf{u}_i$ has to be used for training to prevent a likelihood decrease.

e. Put $\mathbf{u}_i$ back into the data pool $\mathcal{T}$.

4. Retrain the initial acoustic model for one or more iterations only with the utterances for which a likelihood decrease was observed.

See Figure 2 for a graphical illustration of the algorithm. In practice, a threshold for the minimum number of examples required per phone model and/or a threshold for the maximum relative change of the Q-function allowed must be set in order to prevent overfitting.

Other selection strategies are possible, e.g. floating search by deleting or adding one or more utterances while updating the set of selected training utterances immediately. However, drawbacks are a longer computation time, the impossibility of parallel computation and that the selection results depends on the order of processing utterances. Furthermore, previous experiments [10] could not show, that the floating search technique is superior to the greedy search technique.

3. EXPERIMENTAL SETUP

3.1. Speech Data

For experiments, spontaneous Japanese speech from the Takemaru database is employed. Takemaru-kun [12] is a speech-oriented dialogue system intended to provide the user information on the weather, news, the surrounding environment, public transportation system, Internet pages, a.s.o. The system is very popular among children, because it is based on an animated character. It is a working system installed in a public place in Nara, Japan. Speech data is collected automatically since November 2002 from users who speak to the system. All data recorded from the first two years (about 120 hours) is completely transcribed, labeled with tags (e.g. noise) and classified subjectively into one of five speaker groups: infants (preschool children), elementary school children, junior-high school children, adults and elderly persons. The transcribed part of the database contains more than 250,000 utterances (including speech and non-speech inputs) in total.

Table 1. The objective of experiment (1) is to build an infant (preschool children) model using speech from elementary school children. In experiment (2), adult speech is employed to build an elderly acoustic model. (# sentences / time)

Speech Data Sets	(1) School Children → Infant Model	(2) Adults → Elderly Model
Data Pool	29,776 / 17 hrs	17,874 / 9 hrs
Task Data	500 / 17 min	53 / 2 min
Evaluation	1,554 / 53 min	400 / 12 min

Table 1 gives details about the part of the speech data employed in experiments. The aim is to build one acoustic model each for infants (preschool children) and elderly people. Since it is comparably difficult to collect speech data from these two speaker groups, model construction by selecting training data from elementary school children and adult speakers is examined, respectively. Although speech data has been collected for more than two years, the Takemaru database contains only very few utterances from elderly people.

3.2. Acoustic and Language Model

The acoustic feature vector is 25-dimensional including ΔE, 12 MFCC and 12 Δ MFCC. A monophone and a PTM acoustic model is built from scratch with all utterances in the corresponding data pool using HTK [13]. The monophone model consists of 3-state HMMs with up to 16 Gaussians densities (diagonal covariance matrix) per state. There is one HMM for each of the 40 phonemes in the standard Japanese phoneme set plus three silence HMMs (utterance begin, utterance end and short pause). Evaluation is also carried out for phonetic tied-mixture (PTM) models [14], which share one codebook of 32 Gaussians per state among state-clustered triphones with the center phone in common, but with mixture weights untied. Information about the complexity of each PTM acoustic model employed in experiments (1) and (2) is given in Table 2. PTM acoustic models enable fast decoding with the open-source LVCSR engine Julius [15] while maintaining a high recognition performance.

For decoding the infant test set (1,554 sentences / 5,742 words), a task-specific 4k language model trained on transcriptions of infant utterances, and for decoding the elderly test set an (400 sentences / 1,609 words) an open 40k word language model trained on utterance transcriptions from the Takemaru database as well as texts from e-mails and Internet

Table 2. The total number of physical HMMs, the number of distinct HMM states and the total number of parameters (means, covariances, weights and transition probabilities).

Nr.	AM Type	# phys. models	# states	# params
(1)	PTM	765	785	210k
(2)	PTM	572	628	200k

pages is employed.

4. EXPERIMENTAL RESULTS

4.1. Infant-adapted Acoustic Model

The word accuracy of the initial monophone and PTM model built with all utterances in the data pool (containing only speech from elementary school children) is 46.9% and 53.0%, respectively. When applying selective training using 200 infant utterances for likelihood computation, the accuracy increases up to 10% relative for the monophone (51.7%) and 5.5% relative for the PTM model (55.9%). At the same time the gap in performance to a high-cost PTM model trained on 10,000 infant utterances is reduced by 76%. 35% of the utterances in the data pool were selected.

4.2. Elderly-adapted Acoustic Model

The word accuracy of the initial monophone and PTM model built with all utterances in the data pool (containing only adult speech) is 73.6% and 76.7%, respectively. 53 utterances from elderly people are employed for likelihood-based utterance selection. There is a relative improvement of recognition accuracy of up to 3.1% for the monophone (75.9%) and up to 2.0% for the PTM model (78.2%). The selection rate of training utterances in the data pool was 44%.

4.3. Retraining with the Selected Data

Table 3 shows the relationship between the number of EM training iterations to train the initial acoustic model with the selected speech data and the recognition performance. Except for the context-independent monophone model for elderly people (peak after the second iteration), the recognition accuracy has the tendency to increase with a growing number of training iterations. Retraining of the initial acoustic model with the whole data pool did not improve the performance of the initial model.

4.4. Variation of the Task-Specific Data

The performance in case of larger and smaller task-specific speech data sets for experiment (1) is depicted in Figure 3. It is clear that selective training is already effective with only 20 task-specific utterances. Maximum performance seems to be reached with about 100-200 utterances. Furthermore, it is apparent that selective training can provide a better model than standard adaptation methods such as MAP adaptation of means or MLLR adaptation of means and variances if there are only few task-specific data available. The combination of selective training and MLLR adaptation was not effective for the monophone model, but there were improvements for the PTM model.

Table 3. Relationship between the number of EM training iterations with the selected training data and the recognition performance (word accuracy in %).

	Training Iteration					
Monophone	init	1	2	3	5	8
(1) Infant	46.9	50.2	49.7	50.2	**51.7**	**51.7**
(2) Elderly	73.6	75.1	**75.9**	75.1	75.0	74.7
PTM AM	init	1	2	3	5	8
(1) Infant	53.0	55.5	**55.9**	**55.9**	55.7	55.3
(2) Elderly	76.7	77.9	77.5	77.7	77.7	**78.2**

Table 4 shows that the number of utterances selected from the data pool increases with the size of the task-specific data set, although not at the same rate. Even in case of only 20 utterances the selected training data suffice to train the initial acoustic model robustly.

Table 4. Relationship between the number of task-specific data and the number of utterances selected from the data pool.

# Task-specific	Experiment / Acoustic Model	
Utterances	Infant/Mono	Infant/PTM
10	8,715 (29%)	9,108 (31%)
20	9,426 (32%)	9,544 (32%)
50	9,609 (32%)	9,825 (33%)
100	10,300 (35%)	10,434 (35%)
200	10,252 (34%)	10,311 (35%)
500	10,852 (36%)	10,793 (36%)

4.5. Comparison to High-Cost Models

This section compares the performance of the low-cost acoustic models build with selective training to high-costs acoustic models. The costs are due to the collection and labeling of the task-specific speech data. While speech data collection can in principle be carried out automatically, accurate transcription of the speech data has to be carried out by humans.

The experimental results so far showed the effectiveness and practical applicability of the proposed method to the problem of building a task-adapted acoustic model with only a few task-specific speech data. However, it is not clear yet,

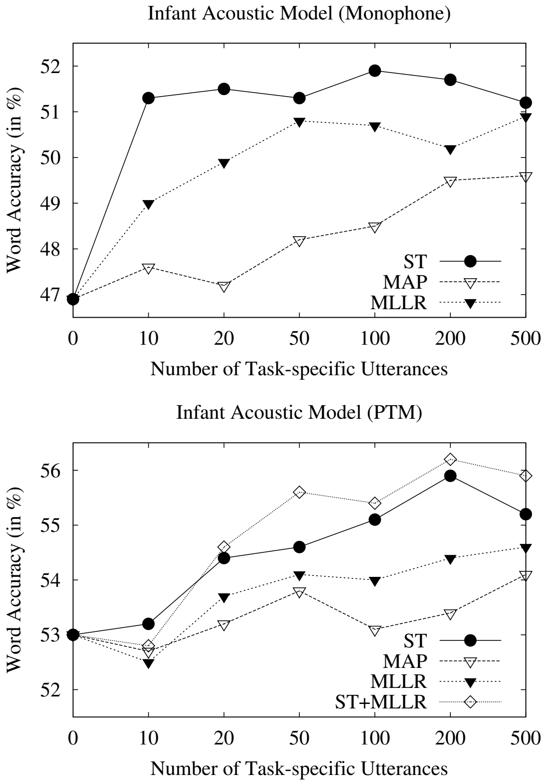

Fig. 3. Influence of the amount of task-specific speech data on the performance of selective training (ST) and standard adaptation methods. Retraining with the selected data was conducted for three (monophone model) or eight (PTM model) iterations, respectively. MLLR adaptation is carried out for two, MAP adaptation for one iteration (Experiment 1).

how much more speech data would have to be collected in order to achieve the same performance as with selective training. Table 5 lists up the performance of acoustic models trained on either many thousand infant utterances collected with the Takemaru dialogue system or a corpus containing more than 50,000 utterances from about 300 different elderly persons (a database description can be found in [16]). The decision to use this speech corpus for comparison is due to the fact that only very few utterances from elderly people were collected by the dialogue system. In case of experiment (1), if there are 10,000 transcribed infant utterances available for retraining the initial model, a higher performance than with selective training can be achieved. Nevertheless, the difference in performance between this well-trained (56.8%) and the initial model (46.9%) is reduced by 76% (relative). Furthermore, using only 5,000 infant utterances for EM training would not be enough to outperform selective training. In case of experiment (2), an acoustic model trained on a large database of elderly speech could not beat the performance of the initial acoustic model trained on adult speech collected with the Takemaru system. From this comparison it is clear that many times more task-specific speech data would have to be collected in order to reach the performance of the model obtained with selective training.

Table 5. Performance (word accuracy) of high-cost models as more transcribed data collected with the Takemaru system (infant speech) or from a separate database (elderly speech) becomes available.

Model	# Training Data	Word Accuracy
Infant Monophone	2,000 infant	49.9%
	3,000 infant	50.4%
	5,000 infant	51.5%
Infant PTM	3,000 infant	54.9%
	5,000 infant	55.5%
	10,000 infant	56.8%
Elderly PTM	56,604 elderly	73.9%

4.6. Summary of Experimental Results

A summary of experimental results is given in Table 6. There are significant improvements in word accuracy over the initial model by employing selective training. The column "Adapt" shows the maximum performance obtained using MAP or MLLR adaptation with the task-specific data. The difference in performance to selective training is large enough to be able to consider the proposed algorithm as a reasonable alternative for task-adaptation of acoustic models.

Table 6. Summary of experimental results (word accuracy in %). Selective training (SelTrain) is effective in each experimental setup and is able to provide a better task-adapted model than conventional acoustic model construction (Initial) and MLLR or MAP adaptation (Adapt) with the task-specific data.

Data → Model	Type	Initial	Adapt	SelTrain
Takemaru DB	Mono	46.9	50.7	**51.7**
Element → Infant	PTM	53.0	54.4	**55.9**
Takemaru DB	Mono	73.6	75.0	**75.9**
Adult → Elderly	PTM	76.7	77.5	**78.2**

4.7. Computation Time and Disk Space

Table 7 shows that the proposed approach for automatic selection of training utterances is computationally practical. For example, when using a PTM model with about 210k parameters and a data pool containing about 30k utterances in case of experiment (1), the required disk space is about 4.7GB and the processing time to select relevant utterances is only three hours (one CPU, no network access). Most of the disk space is required to store the sufficient statistics of the training utterances. There is much room for speedup by partitioning the data pool into subsets of equal size and distribute the computation among multiple CPUs, since the *ST_DelScan* algorithm is parallelizable.

Table 7. Run time and disk space required when using a standard PC with one 3.2 GHz CPU (no network access).

# Utterances	# Parameters	Time	Space
29,776	Monophone / 100k	20 min	2.6 GB
29,776	PTM / 210k	3 hrs	4.7 GB

5. CONCLUSION

A framework for cost-effective task-adaptation of acoustic models using utterance-based selective training was proposed and evaluated. Training utterances are selected from a large data pool so that the model likelihood given a small amount of task-specific speech data is maximized. The selected training utterances are employed to retrain the initial model in order to obtain a task-adapted model.

Results of two evaluation experiments for building an infant and elderly acoustic model show, that the approach is already effective in case if there are only a few task-specific utterances available. Furthermore, the performance was better than with conventional acoustic model construction as well as MLLR and MAP adaptation using the task-specific data. The gap in performance to a high-cost acoustic model is reduced up to 76% relative in case of the infant acoustic model.

As future work, the behavior of the proposed method has to be evaluated if the data pool consists of multiple speech databases and whether it is effective in case the source of the training data and the task-specific data is different. Moreover, since further cost reduction of acoustic modeling could be achieved in general if the transcription of the training utterances needs not to be provided, it is worth considering acoustic model construction based on unsupervised selective training.

6. ACKNOWLEDGMENTS

A part of this work is supported by the COE and e-Society project of the Ministry of Education, Culture, Sports, Science and Technology, Japan.

7. REFERENCES

[1] Y. Gao and L. Gu and H.-K. J. Kuo, "Portability Challenges in Developing Interactive Dialogue Systems," in *International Conference on Acoustics, Speech, and Signal Processing*, pp., 1017–1020.

[2] F. Lefevre and J.-L. Gauvain and L. Lamel, "Genericity and Portability for Task-dependent Speech Recognition," *Computer Speech and Language*, vol. 19, pp. 345–363, 2005.

[3] D. Hakkani-Tür and G. Riccardi and A. Gorin, "Active Learning for Automatic Speech Recognition," in *International Conference on Acoustics, Speech, and Signal Processing*, vol., pp., 3904–3907.

[4] T. M. Kamm and G. G. L. Meyer, "Robustness Aspects of Active Learning for Acoustic Modeling," in *Proceedings of the International Conference on Spoken Language Processing*, pp., 1095–1098.

[5] T. Kemp and A. Waibel, "Unsupervised Training of a Speech Recognizer: Recent Experiments," in *European Conference on Speech Communication and Technology*, pp., 2725–2728.

[6] F. Wessel and H. Ney, "Unsupervised Training of Acoustic Models for Large Vocabulary Continuous Speech Recognition," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2001.

[7] G. Riccardi and D. Hakkani-Tür, "Active and Unsupervised Learning for Automatic Speech Recognition," in *European Conference on Speech Communication and Technology*, pp., 1825–1828.

[8] C. Huang and T. Chen and E. Chang, "Transformation and Combination of Hidden Markov Models for Speaker Selection Training," in *Proceedings of the International Conference on Spoken Language Processing*, pp., 1001–1004.

[9] S. Yoshizawa and A. Baba and K. Matsunami and Y. Mera and M. Yamada and A. Lee and K. Shikano, "Evaluation on Unsupervised Speaker Adaptation based on Sufficient HMM Statistics of Selected Speakers," in *European Conference on Speech Communication and Technology*, pp., 1219–1222.

[10] T. Cincarek and T. Toda and H. Saruwatari and K. Shikano, "Selective EM Training of Acoustic Models based on Sufficient Statistics of Single Utterances," in *Automatic Speech Recognition and Understanding Workshop*, pp., 168–173.

[11] A. P. Dempster and N. M. Laird and D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *J.R. Statistical Society*, vol. 1, no. 39, pp. 1–38, 1977.

[12] R. Nishimura and Y. Nishihara and R. Tsurumi and A. Lee and H. Saruwatari and K. Shikano, "Takemaru-kun: Speech-oriented Information System for Real World Research Platform," in *International Workshop on Language Understanding and Agents for Real World Interaction*, pp., 70–78.

[13] "HTK Speech Recognition Toolkit, `http://htk.eng.cam.ac.uk/`,".

[14] A. Lee and T. Kawahara and K. Takeda and K. Shikano, "A New Phonetic Tied-Mixture Model for Efficient Decoding," in *International Conference on Acoustics, Speech, and Signal Processing*, pp., 1269–1272.

[15] "Julius, an Open-Source Large Vocabulary CSR Engine - `http://julius.sourceforge.jp/`,".

[16] A. Baba and S. Yoshizawa and M. Yamada and A. Lee and K. Shikano, "Elderly acoustic model for large vocabulary continuous speech recognition," in *European Conference on Speech Communication and Technology*, pp., 1657–1660.