SUMMARY Development of an ASR application such as a speech-oriented guidance system for a real environment is expensive. Most of the costs are due to human labeling of newly collected speech data to construct the acoustic model for speech recognition. Employment of existing models or sharing models across multiple applications is often difficult, because the characteristics of speech depend on various factors such as possible users, their speaking style and the acoustic environment. Therefore, this paper proposes a combination of unsupervised learning and selective training to reduce the development costs. The employment of unsupervised learning alone is problematic due to the task-dependency of speech recognition and because automatic transcription of speech is error-prone. A theoretically well-defined approach to automatic selection of high quality and task-specific speech data from an unlabeled data pool is presented. Only those unlabeled data which increase the model likelihood given the labeled data are employed for unsupervised training. The effectiveness of the proposed method is investigated with a simulation experiment to construct adult and child acoustic models for a speech-oriented guidance system. A completely human-labeled database which contains real-environment data collected over two years is available for the development simulation. It is shown experimentally that the employment of selective training alleviates the problems of unsupervised learning, i.e., it is possible to select speech utterances of a certain speaker group but discard noise inputs and utterances with lower recognition accuracy. The simulation experiment is carried out for several selected combinations of data collection and human transcription period. It is found empirically that the proposed method is especially effective if only relatively few of the collected data can be labeled and transcribed by humans.

**key words:** ASR application, real-environment, task-dependency, unsupervised training, selective training

1. Introduction

Although there are many applications for speech recognition technology, e.g., car navigation systems, speech-based guidance system, robots, etc., the breakthrough to consumer market has not been achieved yet. One reason is the high costs for building acoustic and language model of a speech recognition system. A completely human-labeled database which contains real-environment data collected over two years is available for the development simulation. It is shown experimentally that the employment of selective training alleviates the problems of unsupervised learning, i.e., it is possible to select speech utterances of a certain speaker group but discard noise inputs and utterances with lower recognition accuracy. The simulation experiment is carried out for several selected combinations of data collection and human transcription period. It is found empirically that the proposed method is especially effective if only relatively few of the collected data can be labeled and transcribed by humans.

For active learning, unsupervised learning and lightly supervised training, e.g. for a broadcast news transcription system [8] models are first bootstrapped with a small amount of human-transcribed data and then retrained with an additional subset of automatically transcribed data. In order to improve training data quality only speech segments with an automatic transcription matching closed-captions are employed. An improvement of lightly supervised training using consensus networks is proposed in [9].
cate, because utterances collected are from a large number of unknown speakers with speaker changes occurring frequently. Furthermore, the original idea of the speaker indexing methods is to identify segments belonging to the same speaker rather than the selection of speech data for arbitrary tasks.

Recently, we proposed a method for selective training of task-adapted acoustic models [13], [14]. The method is based on a greedy algorithm which adds or deletes training utterances from a large speech data pool. An utterance is only employed for training if the resulting model likelihood given a small set of arbitrary task-specific data increases. The algorithm was effective to build acoustic models for specific speaker groups by selecting an appropriate subset from a human-transcribed speech data pool.

Consider the case new speech data is collected for the development of a real-environment ASR application. To realize cost reduction only data collected during a restricted period after begin of system operation can be transcribed by humans. After that only unsupervised training with the unlabeled data can be carried out. However, unsupervised training requires a transcription of the unlabeled data. Automatic transcription by automatic speech recognition is always error-prone. The more transcription errors there are, the less effective will be unsupervised training. Furthermore, not all speech inputs collected should be employed for training, for example non-verbal inputs such as laughing or coughing, unintelligible inputs and inputs with strong background noise or multi-talk interference. Furthermore, it is often important to improve ASR performance by parallel decoding with multiple acoustic models. Therefore, it would be better to employ only a subset of the unlabeled data to construct each acoustic model.

In this paper, the selective training algorithm is applied to build task-adapted acoustic models if the data pool is completely untranscribed. In experiments it is investigated in how far the proposed method can select training utterances with a high transcription accuracy from the desired speaker group and whether it can discard too noisy or low quality inputs. The effect of the period of speech data collection and transcription on speech recognition performance is also analyzed.

The paper is organized as follows. The target application, a speech-oriented guidance system operated in a real environment is described in Sect. 2. The proposed approach for selective unsupervised training of acoustic models using unlabeled data is explained in Sect. 3. The setup of experiments for evaluating the effect of the proposed method is described in Sect. 4. The evaluation result of applying each training method independently as well as a powerful combination is shown in Sect. 5. A promising conclusion is given in Sect. 6.

2. Target Application

As outlined in the introduction, employing one acoustic model for different applications and environments is difficult due to the task-dependency of speech recognition. In the following, acoustic model construction for a real-environment speech-oriented guidance system is considered as a realistic scenario for application development.

The purpose of a speech-oriented guidance system is to offer a certain group of users convenient access to proper information in a certain environment. While the information society is at the verge to an ubiquitous society, there is growing demand for this kind of services in any place. Although entering search queries via keyboard is still the prevailing method for accessing information, formulating one’s question freely in natural language and using speech is a far more natural way to human-machine communication.

The Takemaru system [12] installed inside the entrance hall of the North community center in Ikoma city, Nara Prefecture, Japan (Fig. 1). The indoor environment is relatively calm with a background noise level of approx. 50 dB (A). The place is frequently visited by adults and children, because it is a public facility with a library, a branch office for residential services and there are weekly events. The system uses the mascot character of Ikoma city, Takemaru, as agent. The system can handle queries related to the agent, general information such as time, date, weather and news, the facility itself, surrounding area and sightseeing.

3. Proposed Approach

3.1 Problem Description

Various kinds of inputs are collected when deploying an ASR application in a real environment. There are speech and non-verbal inputs from several user groups as well as noise inputs from the surrounding environment. The data collected by Takemaru during the first two years is shown in Fig. 2.

For system development, a large amount of the collected data is most often transcribed and labeled with tags
Fig. 2 Inputs collected with Takemaru during the first two years of operation.

(e.g. noise, validity, speaker group classification) by humans. This is a very costly and time-consuming process. If accurate transcriptions and speaker group labels are available for each utterance, it is straightforward to build high-quality speaker-group-dependent acoustic models. The purpose of building extra models for different speaker groups is to improve overall recognition performance by parallel decoding, and to optimize the performance separately for each speaker group. For example, the Takemaru system uses different models for recognizing and responding to inputs of adult and children users.

In order to reduce the costs of acoustic modeling, it is imperative to reduce the amount of data to be transcribed by humans. However, it is desirable that also the unlabeled data can be used effectively for model training. To achieve this, the employment of unsupervised training is inevitable.

3.2 Proposed Training Procedure

The proposed training procedure is a combination of unsupervised and selective training. Selective training is employed to alleviate the problems which arise when using unsupervised training.

3.2.1 Employment of Unsupervised Training

Unsupervised learning requires to transcribe the unlabeled data automatically. The initial model for automatic transcription can be built from scratch or by adapting an existing model with the labeled data. The automatically transcribed data can then be employed together with the labeled data to retrain the initial model. However, with introduction of unsupervised learning for real-environment data the following problems arise:

1. Automatic transcriptions are always error-prone
2. Speaker (group) of unlabeled inputs is unknown
3. Data kind is unknown (speech, noise, non-verbal)

Consequently, it is necessary to select an appropriate subset of the unlabeled data for acoustic modeling. Ideal training utterances are speech-only inputs, which have a high transcription label accuracy, do not contain strong noise interferences and belong to a certain speaker group.

3.2.2 Employment of Selective Training

It is possible to automatically select training utterances with these characteristics by applying our previously proposed method for selective training [14]. A graphical illustration of the selective training algorithm is given in Fig. 3. The starting point is a large training data pool T and a small development data set D. Here, the data pool consists of all unlabeled, collected inputs. Human-transcribed utterances from a certain speaker group form the development set. Let \( \Theta \) denote the model parameters estimated on the selected data S. The main idea of the proposed selection algorithm is to select a subset of utterances S \( \subseteq T \) from the data pool so that the model likelihood \( P(D|\Theta) \) given the development data is maximized. Since D consists of human-prepared example data, it can be expected that the algorithm selects high-quality speech utterances from the data pool matching the desired target task, e.g. speaker group.

There are 2^n possible subsets for a pool with n data. Since investigating all data subsets is computationally infeasible, a heuristic selection strategy has to be employed. With a greedy search technique each utterance in the training data pool is examined only once for deletion. Data are discarded if their independent deletion increases the model likelihood.

If T and S are large, estimation of model parameters and calculation of the likelihood will be computationally intensive. Nevertheless, in the framework of maximum likelihood estimation of HMM parameters with the expectation-maximization algorithm [15], it is possible to calculate the likelihood \( P(D|\Theta) \) with low computational costs via the auxiliary Q-function

\[
Q(\Theta, \tilde{\Theta}) = \sum_{\tilde{\xi}} P(\tilde{\xi}|D, \Theta) \log P(\tilde{\xi}, D|\Theta)
\]
where $\Theta$ and $\hat{\Theta}$ denote initial and updated model parameters, respectively. $s$ is the state and mixture index sequence. An increase of the $Q$-function implicates an increase of the model likelihood $P(D|\hat{\Theta})$. The $Q$-function can be expressed as a function of only the HMM sufficient statistics of the development data $D = \{x_t\}$ and the HMM sufficient statistics of either the training data $T$ or the subset $S$.

For simplicity of notation, data $x_t$ are assumed to be one-dimensional and the HMM transition probabilities are neglected. Nevertheless, it is easy to define the equations for multivariate data and to include the transition probabilities.

The output density part of $Q$ can be rewritten for Gaussian parameters $\mu_{qm}$, $\sigma_{qm}$ and mixture weights $\hat{\omega}_{qm}$ to be proportional to the expression

$$\sum_q \sum_m \gamma_{qm}(t) \log \frac{\hat{\omega}_{qm}}{\sqrt{2\pi\sigma_{qm}}}$$  \hspace{1cm} (2)

$$- \sum_q \sum_m \gamma_{qm}(t) \frac{1}{2}(x_t - \mu_{qm})^2 \frac{1}{\sigma_{qm}}$$  \hspace{1cm} (3)

where $\gamma_{qm}(t)$ is the state occupation probability of data frame $x_t$ for state $q$ and mixture component $m$. This expression can be transformed to

$$\sum_q \sum_m y_{qm} \log \frac{\hat{\omega}_{qm}}{\sqrt{2\pi\sigma_{qm}}}$$  \hspace{1cm} (4)

$$- \sum_q \sum_m z_{qm} - 2\hat{\mu}_{qm}y_{qm} + \hat{\mu}_{qm}^2 \frac{1}{2\sigma_{qm}}$$  \hspace{1cm} (5)

where variables $\gamma_{qm}$, $\sigma_{qm}$ and $z_{qm}$ denote the sufficient statistics (SS) of the development data $D$. The SS are obtained by calculating the sum over $t$ for each corresponding term in advance. It is clear that the SS of the development data need only to be calculated once and that they are independent from the training data. Therefore, the computational complexity of the $Q$-function is lower than that of likelihood function $P(D|\hat{\Theta})$, which requires complete calculation for every speech frame of the development data whenever the model parameters change.

HMM sufficient statistics have the property to be additive, i.e. the statistics for the whole data pool are given as the sum of the statistics of each training utterance. Deleting an utterance from the pool means subtracting the corresponding sufficient statistics.

Further technical details on the selective training algorithm and HMM parameter estimation can be found in [14] and [16], respectively.

4. Simulation Experiments

In the following, details and conditions of a simulation experiment using the proposed method for acoustic model construction are described. Furthermore, in order to show the effectiveness of the proposed approach for the selection of arbitrary task data it is compared with a conventional approach using GMM-based selection.

4.1 Proposed Method

In the beginning, only a small amount of human-labeled data are employed for active system development. After that the system will retrain itself automatically with newly collected data. This is interesting for practical system development, since human efforts are only necessary at the beginning. The performance will improve over time without additional costs and human intervention.

In a simulation experiment, the influence of data collection and transcription period on speech recognition performance is analyzed. The less data are labeled by humans the lower the development costs but also the performance. Therefore, it is investigated how performance improves as more collected data for selective unsupervised training become available (cf. Figure 4).

The procedure of acoustic model construction is illustrated in Fig. 5. The main steps are as follows.

1. Train or adapt JNAS [17] model with the labeled data using Baum-Welch or MLLR-MAP adaptation depending on the amount of available training data
2. Obtain initial model for automatic transcription and later retraining
3. Recognize all data in the unlabeled data pool

![Fig. 4 Initial human development. After that automatic development.](image)

![Fig. 5 Procedure for acoustic model construction.](image)
4. Obtain transcriptions for the unlabeled data
5. Selective training to maximize model likelihood given the labeled data
6. Obtain subset of the data pool
7. Retrain initial model with the labeled and selected utterances
8. Obtain the final acoustic model

Experimental conditions for acoustic model training, adaptation and evaluation, and the language model for automatic transcription are given in Table 1. When much training data is available, the acoustic model is trained using the Baum-Welch algorithm. Otherwise it is adapted in a two-step approach using MLLR-MAP: Firstly, mean vectors and diagonal covariance matrices of Gaussians are adapted using MLLR transforms [3]. After that MAP estimation [4] of mean vectors is carried out for the MLLR-transformed model.

Real-environment data collected with the Takemaru system are employed for the development simulation. The complete two-year Takemaru database is shown in Table 2. Training and evaluation data sets are given in Table 3. Evaluation data were selected randomly so that there is one utterance each for the most frequent 1,000 utterance transcriptions. The adult test set is gender-balanced with 1,000 utterances each from male and female speakers. The children test set consists of 1,000 utterances each for preschool, lower grade and higher grade school children. For supervised training or supervised adaptation only valid user inputs are employed, i.e. utterances with strong interfering noise and unintelligible inputs are discarded. Moreover, evaluation, labeled and unlabeled data are always selected to be mutually disjoint.

4.2 GMM-Based Selection

A conventional method to select training data for a certain task is a pattern classification approach using statistical models such as Gaussian Mixture Models (GMM). Their application has already been successful, e.g. for speaker identification/verification [21] or the rejection of unusable inputs to a spoken dialogue system [22]. The disadvantage of the GMM-based approach over the proposed method for selective training is that it lacks an optimization criterion which is directly linked to the model quality. The proposed selective training method adopts the model likelihood given the task-specific data as optimization criterion.

Classification models are built for three classes: One GMM each is constructed from the labeled adult (and elderly), (preschool and school) child and noise (and non-verbal) data of each transcription period (cf. grouping in Table 2). The same 25-dimensional MFCC-based feature vector as for speech recognition is employed for constructing the GMM for each class from scratch. The number of mixture components is increased incrementally until 64 Gaussians are reached.

For automatic data selection, the unlabeled data of each corresponding data collection period are partitioned automatically into three classes. Data classified as noise are discarded completely. Data classified as adult (or child) are sorted by the GMM likelihood. The top-ranked data are employed together with the labeled data for building the adult and child acoustic model, respectively. For a fair comparison w.r.t. the number of training data the same selection rate as determined by the proposed selective training algorithm is used.

5. Experimental Results

The proposed method, which employs labeled and unlabeled selected data (A+C) for training, is compared with supervised training when using only labeled data (A) and semi-supervised training when using labeled and unlabeled data (A+B). Furthermore, it is analyzed whether the selective training algorithm actually selects speech data of the expected characteristics from the unlabeled data pool.

5.1 Comparison of Training Methods

The case of a fixed and short transcription period of only one week is considered first. The performance is shown in Figs. 6 and 7 for adults and children, respectively. It can be observed that the performance with semi-supervised training increases for children but decreases for adult speakers. This is due to the fact that most of the collected data...
are from children users. Consequently, performance improves remarkably for adult speakers when using the proposed method. For children there are slight improvements for short as well as long data collection periods. It is clear that the proposed method outperforms both supervised and semi-supervised training.

The influence of more human transcriptions on performance after 18 months (fixed) of data collection is shown in Figs. 8 and 9. The more human-transcribed data is available for training, the higher is the overall performance. The difference between the proposed method and supervised training becomes small for both adult and children after three months. This shows that the proposed method is effective especially when only very few data can be transcribed by humans. Moreover, it is clear that the conventional GMM-based method does not outperform the proposed method in selecting arbitrary task-specific training data. While the GMM-based method is effective in selecting adult utterances with a performance close to selective training, its overall performance is lower than both semi-supervised training and the proposed method for building the child model.

The number of labeled, unlabeled pool and unlabeled selected data used to train each corresponding model is given in Table 4. The sum of the number of labeled and unlabeled pool data is not equal for each column, because noise data, non-verbal data, unintelligible data and utterances except from the target speaker group are excluded from the labeled data. Selective training can augment the existing labeled data with an appropriate subset of the unlabeled data. This avoids data insufficiency for model training and improves performance especially when the data transcription period is short.

### 5.2 Validity of Data Selection

The selected data of the one week transcription and 18 months data collection experiment are analyzed w.r.t. to the human labels and speech recognition accuracy on utterance basis.
A breakdown w.r.t. human-assigned age group and noise labels is shown in Table 5. It is clear that the proposed selective training algorithm is able to reduce the relative share of noise data and also increases the relative share of utterances from the desired speaker group. This means that the second and third initially mentioned problems arising from the employment of unsupervised learning are already alleviated by the proposed method.

Selected utterances had more often a relatively high, discarded utterances a relatively low recognition accuracy. When building the adult acoustic model, the average word accuracy (correct rate) of selected and discarded speech inputs is 78.6% (81.7%) and 49.5% (48.5%), respectively. For the children experiments the rates are 57.2% (63.4%) and 53.0% (59.5%), respectively. This shows that utterances with erroneous transcriptions are more likely to be discarded so that the first initially mentioned problem of unsupervised learning is also addressed.

6. Conclusion

In this paper, a powerful combination of unsupervised and selective training is proposed to reduce the costs of acoustic modeling for real-environment speech-based applications. The employment of unsupervised learning is inevitable to avoid the very costly and time-consuming process of transcribing large amounts of speech data by humans as much as possible. The purpose of applying selective training is to alleviate the problems which arise when using unsupervised training. The idea to select additional training utterances from an unlabeled data pool, so that the model likelihood given the labeled data increases, is promising. Experimental results show that the proposed selective training algorithm can select automatically utterances of the desired speaker group, e.g. adult or child, and discard most of the noise-only inputs as well as data with a lower recognition accuracy. From analyzing the influence of the data collection and transcription period on the recognition performance it was clear that the proposed method is especially effective when the amount of data labeled by humans is restricted.

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 (MEXT), Japan.

References

[18] HTK Speech Recognition Toolkit.
http://htk.eng.cam.ac.uk/
http://julius.sourceforge.jp/

Table 5 Breakdown w.r.t. human-assigned labels of the pool data before selection and the subset selected with the selective training algorithm.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Adult Model</th>
<th>Child Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Pool → Selected</td>
<td>140k → 8k</td>
<td>139k → 34k</td>
</tr>
<tr>
<td>Adult Inputs</td>
<td>12.0% → 73.9%</td>
<td>12.3% → 8.4%</td>
</tr>
<tr>
<td>Child Inputs</td>
<td>64.4% → 14.7%</td>
<td>63.0% → 78.4%</td>
</tr>
<tr>
<td>Noise Inputs</td>
<td>23.6% → 11.4%</td>
<td>23.7% → 13.1%</td>
</tr>
</tbody>
</table>
Appendix: Complete Simulation Result

The complete view on the development simulation is given in Figs. A.1–A.4. The performance of combinations of selected transcription (vertical axis) and collection (horizontal axis) periods are shown. The overall improvement through the proposed method is clear by comparing Fig. A.1 with Figs. A.2 and A.3 with Fig. A.4.

When comparing the leftmost boxes with the right boxes in Figs. A.4 or A.2, pairs of collection and transcription periods with equal performance can be identified. For example, by transcribing only two weeks of data collected during 12 months the same performance as when transcribing the data from two months can be obtained.

![Fig. A.1](image1.png)
**Fig. A.1** Training with labeled and unlabeled pool data. (Children)

![Fig. A.2](image2.png)
**Fig. A.2** Training with labeled and selected unlabeled data. (Children)

![Fig. A.3](image3.png)
**Fig. A.3** Training with labeled and unlabeled pool data. (Adults)

![Fig. A.4](image4.png)
**Fig. A.4** Training with labeled and selected unlabeled data. (Adults)

Tobias Cincarek was born in Mannheim, Germany. He received his diploma degree in computer science from Friedrich-Alexander University, Erlangen-Nuremberg, Germany in 2004. During 2003-2004 he was an intern at the Department of Acoustics and Speech Research, ATR Spoken Language Translation Research Laboratories, Japan, where he had been engaged in research on acoustic modeling and pronunciation scoring for non-native speech. Currently, he is enrolled in the doctoral course at the Graduate School of Information Science, Nara Institute of Science and Technology (NAIST). His research interests include speech recognition and spoken dialogue systems. He is a member of the Institute of Electrical and Electronics Engineers (IEEE) and the Acoustical Society of Japan (ASJ).
Tomoki Toda received the B.E. degree in electrical engineering from Nagoya University, Nagoya, Japan, in 1999 and the M.E. and Ph.D. degrees in engineering from the Graduate School of Information Science, Nara Institute of Science and Technology (NAIST), Nara, Japan, in 2001 and 2003, respectively. From 2001 to 2003, he was an Intern Researcher at the ATR Spoken Language Translation Research Laboratories, Kyoto, Japan. He was a Research Fellow of the Japan Society for the Promotion of Science in Graduate School of Engineering, Nagoya Institute of Technology from 2003 to 2005. He was a Visiting Researcher at the Language Technologies Institute, Carnegie Mellon University, Pittsburgh, USA, from October 2003 to September 2004. He is currently an Assistant Professor in the Graduate School of Information Science, NAIST, and a Visiting Researcher at the ATR Spoken Language Communication Research Laboratories. He received the TELECOM System Technology Award for Student from the Telecommunications Advancement Foundation in 2003. He has been a member of the Speech and Language Technical Committee of the IEEE Signal Processing Society since January 2007. He is a member of IEEE, ISCA, and ASJ. His research interests include voice transformation, speech synthesis, speech analysis, and speech recognition.

Hiroshi Saruwatari was born in Nagoya, Japan, on July 27, 1967. He received the B.E., M.E. and Ph.D. degrees in electrical engineering from Nagoya University, Nagoya, Japan, in 1991, 1993 and 2000, respectively. He joined Intelligent Systems Laboratory, SECOM CO., LTD., Mitaka, Tokyo, Japan, in 1993, where he engaged in the research and development on the ultrasonic array system for the acoustic imaging. He is currently an associate professor of Graduate School of Information Science, Nara Institute of Science and Technology. His research interests include array signal processing, blind source separation, and sound field reproduction. He received the Paper Awards from IEICE in 2001 and 2006. He is a member of the IEEE and the Acoustical Society of Japan.

Kiyohiro Shikano received the B.S., M.S., and Ph.D. degrees in electrical engineering from Nagoya University in 1970, 1972, and 1980, respectively. He is currently a professor of Nara Institute of Science and Technology (NAIST), where he is directing speech and acoustics laboratory. From 1972, he had been working at NTT Laboratories, where he had been engaged in speech recognition research. During 1990-1993, he was the executive research scientist at NTT Human Interface Laboratories, where he supervised the research of speech recognition and speech coding. During 1986-1990, he was the Head of Speech Processing Department at ATR Interpreting Telephony Research Laboratories, where he was directing speech recognition and speech synthesis research. During 1984-1986, he was a visiting scientist in Carnegie Mellon University. He received the IEICE (Institute of Electronics, Information and Communication Engineers of Japan) Yonezawa Prize in 1975, IEEE Signal Processing Society 1990 Senior Award in 1991, the Technical Development Award from ASJ (Acoustical Society of Japan) in 1994, IPSJ (Information Processing Society of Japan) Yamashita SIG Research Award in 2000, Paper Award from the Virtual Reality Society of Japan in 2001, IEICE Paper Award in 2005 and 2006, and IEICE Inose Best Paper Award in 2005. He is a fellow member of IEEE and IPSJ. He is a member of ASJ, Japan VR Society, and International Speech Communication Association.