Acoustic Modeling for Spoken Dialogue Systems Based on Unsupervised Utterance-based Selective Training

Tobias Cincarek, Tomoki Toda, Hiroshi Saruwatari, Kiyohiro Shikano

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

10.21437/Interspeech.2006-478

Abstract

The construction of high-performance acoustic models for certain speech recognition tasks is very costly and time-consuming, since it most often requires the collection and transcription of large amounts of task-specific speech data. In this paper acoustic modeling for spoken dialogue systems based on unsupervised selective training is examined. The main idea is to select those training utterances from an (untranscribed) speech data pool, so that the likelihood of a separate small (transcribed) development speech data set is maximized. If only the selected data are employed to retrain the initial acoustic models, a better performance is achieved than when retraining with all collected data. Using the proposed approach it is also possible to considerably reduce the costs for human-labeling of the speech data without compromising the performance. Furthermore, the method provides means for automatic task-adaptation of acoustic models, e.g. to adult or children speech. This is important, since detailed information about each automatically collected utterance is usually not available.
Index Terms: speech recognition, acoustic model, unsupervised training, utterance-based selection, spoken dialogue system

1. Introduction

Although there are many applications which make use of speech recognition technology, e.g. dictation systems, car navigation system, speech-based guidance system, train information system, call centers, etc., they are less widely used. One reason are the high costs for building the acoustic and the language model to realize high-performance speech recognition. For example, about half of the relative costs to develop an interactive dialogue system are due to speech database preparation [1]. The costs are mainly due to the tremendous effort necessary to collect and transcribe a large amount of task-specific speech data, which are required to build a robust acoustic model. Since speech recognition performance depends on various factors such as speaker characteristics (e.g. gender, age, accent), speaking style (e.g. read, spontaneous), domain (e.g. commands, dialogue) and acoustic conditions (e.g. background noise, reverberation, microphone), it is difficult to share one and the same acoustic model among different task domains.

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] or both [6] to reduce the effort necessary for speech data transcription, and training [4] or adaptation [7, 8] 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, 6].

By employing the above-mentioned approaches to active and unsupervised learning, costs for transcribing speech data can be reduced. However, their drawback is, that the selection of training utterances is restricted either to data with high recognition confidence (unsupervised learning) or data, which are difficult to recognize (active learning). For example, in case of a speech-based guidance system installed in a public place, the population of users originates from various speaker groups such as children, adults, natives, non-natives, a.s.o. In order to build a high-performance acoustic model for each desired target speaker group, appropriate training data subsets have to be selected. However, human-labeling of all collected data should not be necessary.

In this paper an approach for cost-effective construction of acoustic models for a spoken dialogue system is proposed and evaluated. It is based on utterance-based selective training [9], which enables automatic selection of training speech data from a large data pool so that the likelihood of the target model parameters given a small set of example speech data is maximized. In the beginning only a small amount of the collected speech data are labeled and the desired example data are extracted. All other speech data are employed to set up a large speech data pool, which is transcribed automatically by a speech recognizer. A data subset acoustically close to the example speech data are selected from this data pool. The selected data are finally employed to build the desired acoustic model.

The proposed method can also be employed to determine a subset of the collected data to be labeled by humans. Unnecessary to mention, that a higher recognition performance can be achieved if accurate transcriptions are employed for acoustic model training. However, this again is very costly and time-consuming. It is shown experimentally, that the number of data to be transcribed can be reduced considerably by employing the proposed unsupervised data selection method.

2. Proposed Approach

In this section an approach to automatic training data selection for acoustic model training is described. Here, the target is the construction of a children and adult acoustic model for a spoken dialogue system to be installed in a public place. Nevertheless, the proposed method can be applied easily to a different scenario. In the beginning, a certain amount of speech data is collected. The first few thousand inputs (noise and speech) are labeled by humans. The transcribed data are divided into a development data set of children and adult speech data. However, relatively noisy

speech utterances are discarded in advance.

All data collected beyond the initially transcribed data can be employed in two ways: (1) Transcribe the data automatically and use a subset of it for unsupervised learning. (2) Let humans transcribe a further subset of the data and employ them for supervised learning. In case of option (1), data selection is necessary, since automatic transcriptions are often erroneous and not all collected data will contain speech. In case of option (2), data selection is intended to reduce the transcription costs but without compromising the performance. Since it is necessary to divide the untranscribed data into a set of adult and children speech, selection based on a recognition confidence measure as in case of conventional unsupervised and active learning is not applicable.

To realize effective selection of the desired training data, an utterance-based selective training method is applied. It enables the selection of the subset of the untranscribed data, which maximizes the likelihood of the development data set. For example, if the development data set contains adult speech, most of the selected data will be adult utterances. The selection criterion (likelihood) and the selection algorithm are described in Section 2.1.

2.1. Utterance-based Selective Training Algorithm

In the following, a brief description of utterance-based selective training as proposed in [9] is given. Figure 1 shows a graphical illustration of the proposed selective training method.

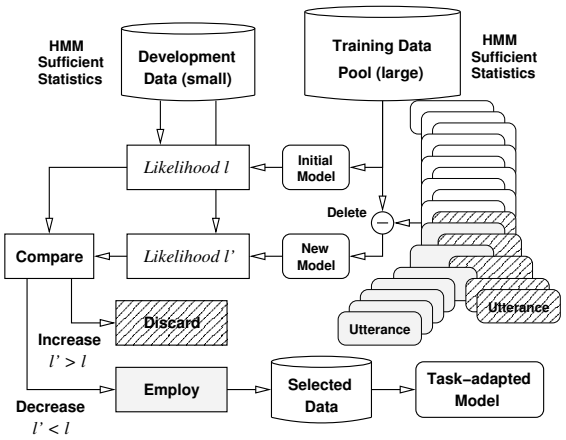

Figure 1: Utterance-based selective training by using a greedy maximum likelihood (ML) training data selection strategy.

The starting point is a large training data pool $\mathcal{T}$ and a small development data set $\mathcal{D}$. This data pool consists of existing and/or newly collected speech data. The development data set is set up with examples of speech data for which an acoustic model is to be built. The central idea of the proposed selective training method is to select a subset $\mathcal{S} \subseteq \mathcal{T}$ from the training data pool, so that the likelihood $P(\mathcal{D}|\hat{\mathbf{\Theta}})$ of the parameters $\hat{\mathbf{\Theta}}$ estimated on the selected data $\mathcal{S}$ is maximized. This can also be interpreted as to select those speech data from the pool which are acoustically close to the example data.

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. A training utterance is discarded if its independent deletion results in a likelihood increase. Other selection strategies are possible, e.g. floating search by deleting or adding one (*ST_DelAdd*) 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 result depends on the order of processing utterances.

A reader might argue, that every time an utterance is added to or deleted from the data pool, calculation of the model parameters and their likelihood is computationally intensive. Actually, calculation of the likelihood given the development data $P(\mathcal{D}|\hat{\mathbf{\Theta}})$ is possible implicitly in a very short amount (about one second) of time via the auxiliary Q-function using HMM sufficient statistics in the framework ML estimation of HMM parameters using the EM algorithm. Technical aspects of HMM parameter estimation can be found in [10], the selective training algorithm in [9].

3. Experimental Setup

The speech data employed for experiments and the speech recognition system are described in Sections 3.1 and 3.2, respectively. The experimental setup is shown in Figure 2. Data shorter than 0.7 seconds are discarded since they most often do not contain any speech data. All data collected during the first three months (approx. 9k data) are assumed to be labeled manually. From these data a development set of relatively clean 2,000 children utterances and 1,000 adult utterances is extracted to be used for likelihood calculation during selective training. The evaluation data are selected randomly from data collected during the first two years, so that there is one utterance each from a male and female speaker, or one utterance each from an elementary and junior-high school child for the most frequent 1,000 utterance transcriptions. The data pool consists of all collected data with a duration of 0.7 seconds or more except the data from the first three weeks and the evaluation data. The data pool is transcribed automatically using the speech recognition system as described in Section 3.2. After greedy ML utterance selection based on the *ST_DelScan* algorithm is carried out, the selected data are employed for retraining of the initial acoustic model.

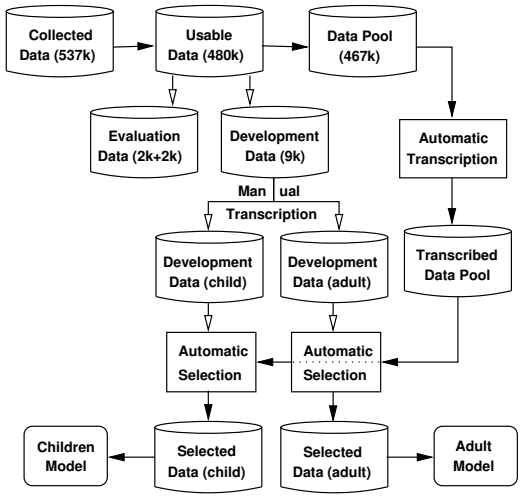

Figure 2: Experimental setup for unsupervised data selection and unsupervised acoustic model construction.

3.1. Data and Labels

Spontaneous Japanese speech from the Takemaru database is employed. Takemaru-kun [11] 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. The system collected more than 500,000 noise and speech inputs since November 2002. All data recorded during the first two years are completely transcribed, labeled with tags (e.g. noise) and classified subjectively into speaker groups: infants (preschool children), elementary school children, junior-high school children, adults and elderly persons (cf. Table 2)

Table 1: Speech data collected with the Takemaru-kun dialogue system: relative share, total number and total duration noise and speech inputs for each speaker group.

Category	rel.share	# data	Time
Preschool Children	10.1%	27,535	14.3 hrs
Elementary School	39.0%	106,797	57.7 hrs
Junior-high School	11.5%	31,402	15.8 hrs
Adults, Elderly	11.3%	31,100	14.1 hrs
Noise, Non-Verbals	28.1%	76,864	19.3 hrs
Labeled		273,698	121.2 hrs
Unlabeled		263,973	120.1 hrs
Total		537,671	241.3 hrs

Table 2: Speech data actually employed in experiments. The (unlabeled) data pool consists of all collected data except inputs shorter than 0.7 seconds.

Experiment:	Adult Model		Children Model	
Data Set	# Data	Time	# Data	Time
Data Pool	466,511	227 hrs	466,511	227 hrs
Development	1,000	31 min	2,000	66 min
Test Data	2,000	65 min	2,000	61 min

3.2. Speech Recognizer

The initial acoustic model of the speech recognizer is built with the Japanese Newspaper Article Sentences (JNAS) database [12]. The standard set of 40 Japanese phonemes plus three silence HMMs are employed. The acoustic feature vector is 25-dimensional including ΔE, 12 MFCC and 12 Δ MFCC. There is one 3-state HMM per phoneme model. A phonetic-tied mixture (PTM) acoustic model [13] is synthesized from a state-clustered triphone (about 2000 different states) and a monophone model with 64 Gaussian densities (diagonal covariance matrix) per state. PTM models enable fast decoding with the open-source LVCSR engine Julius [14] while maintaining a high recognition performance. The same language model is employed for evaluation and the automatic transcription of the data pool. There is one open 40k word (Japanese morpheme) trigram language model each for adults and children which are trained on texts collected from local Internet pages, e-mails, hypothesized questions and a few utterance transcriptions.

4. Experimental Results

4.1. Unsupervised Data Selection and Unsupervised Training

In Figure 3 unsupervised EM (Uns.EM [pool]) is compared to unsupervised selective EM training (Uns.EM [select]). Evaluation of both methods is carried out for data pools of sizes 10k up to 400k (the data are not drawn randomly but employed in the same chronological order as they have been collected). Training speech data selection is carried out separately for building the adult- and the child-dependent model. Each point of the graph's lower line shows the performance of the model trained on all pool data. The upper line shows the performance of the models trained on the data selected from the corresponding data pool of given size.

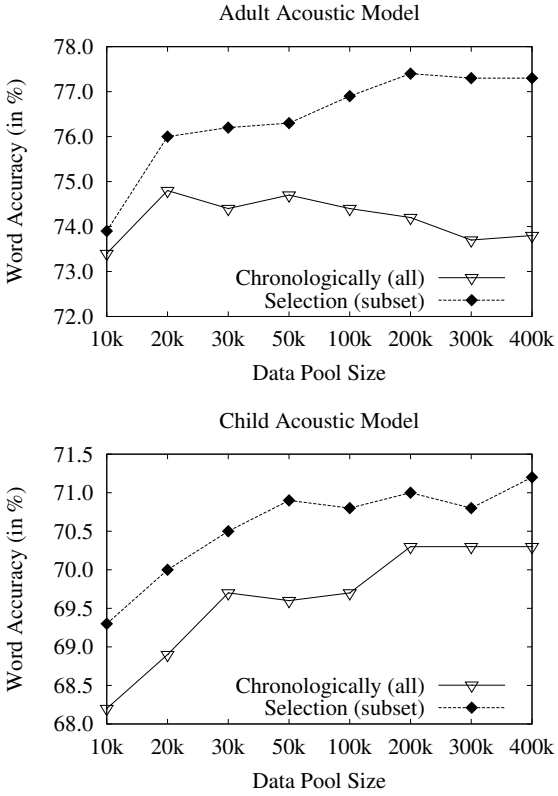

Figure 3: Comparison of one (unsupervised) EM iteration with all collected data to one (unsupervised) EM iteration with the data subset selected using two iterations of the *ST_DelScan* algorithm.

The word accuracy of the initial model is 75.2% for the adult and 58.3% for the children test set. It is clear from Figure 3, that performance for adult speech degrades if all pool data are employed for training. This is due to the fact, that most of the collected speech data are from children. In case of the 200k data pool, employing only the automatically selected data for retraining, recognition accuracy rises to 77.4%. Performance for the children increases to 70.3% even if all collected data are employed for training. A further relative improvement of 1-2% is achieved in case of unsupervised training data selection. The algorithm selected 8.1% of the data pool for building the adult, 20.0% for the children model. While the relative share of child and adult data among the selected training speech data was as high as 70.9% and 61.5%, respectively,

the relative share of noise-only data was only about 10%.

This shows, that the proposed method effectively selects the desired training data in unsupervised manner, and is able discard data which are likely to have negative effects on the recognition performance at the same time.

4.2. Human Transcription of the Selected Data

The data selected could also be transcribed and employed for supervised EM training. Given the untranscribed 200k data pool, 55k and 79k data are selected for training the adult and children model, respectively, using one *ST_DelScan* selection iteration each. Retraining the initial acoustic model with all relatively clean 15k adult and all relatively clean 51k children data contained in this first automatically selected and then human-transcribed data set (Uns.Sel.+Trans.), a performance equal to when transcribing all collected data is achieved. In the latter case, the adult and children models are trained separately on the speech data of the corresponding age group (Sup.EM [group]).

Taking all selected data (for adult and children) together, the number of (unique) utterances was 106k. Consequently, only about half of the collected data need to be transcribed without a compromise in performance.

A final comparison is made with supervised MLLR+MAP adaptation using the human-labeled development data. Employing the whole development data set, a word accuracy of 77.2% and 72.0% are achieved for the adult and the child-adapted model, respectively. Although supervised MLLR+MAP adaptation cannot outperform the proposed approach, further improvements may be possible from their combination. A summary of experimental results is given in Table 3.

Table 3: Experimental results for the 200k data pool and two training iterations with each method. The column # data shows the number of training utterances actually employed for each method.

Experiment →	Adult Model		Children Model	
↓ Method	# data	W.Acc.	# data	W.Acc.
Initial Model	0	75.2	0	58.3
Adaptation	1,000	77.2	2,000	72.0
Uns.EM [pool]	200,000	74.2	200,000	71.1
Uns.EM [select]	20,418	**77.4**	38,381	**72.5**
Sup.EM [group]	14,899	78.7	67,458	75.9
Uns.Sel.+Trans.	14,518	**78.8**	50,585	**75.7**

5. Conclusion

The proposed method for acoustic model construction based on unsupervised greedy ML data selection is a promising alternative to conventional unsupervised and active learning. Speech data of the desired speaker group are selected effectively while most of the noisy data are discarded successfully. Unsupervised EM training with the selected data gives a higher performance than when using all collected data. If the proposed method is employed to select those of the collected data to be transcribed by humans, almost half of the transcription costs can be cut.

6. Acknowledgments

A part of this work is supported by the MEXT COE and e-Society project, Japan.

7. References

[1] Y. Gao, L. Gu, and H.-K. J. Kuo, "Portability Challenges in Developing Interactive Dialogue Systems," in *International Conference on Acoustics, Speech, and Signal Processing*, 2005, pp. 1017–1020.

[2] F. Lefevre, 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, G. Riccardi, and A. Gorin, "Active Learning for Automatic Speech Recognition," in *International Conference on Acoustics, Speech, and Signal Processing*, 2002, vol. 4, 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*, 2004, pp. 1095–1098.

[5] 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.

[6] G. Riccardi and D. Hakkani-Tür, "Active and Unsupervised Learning for Automatic Speech Recognition," in *European Conference on Speech Communication and Technology*, 2003, pp. 1825–1828.

[7] C. Huang, 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*, 2004, pp. 1001–1004.

[8] S. Yoshizawa, A. Baba, K. Matsunami, Y. Mera, M. Yamada, 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*, 2001, pp. 1219–1222.

[9] T. Cincarek, T. Toda, 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*, 2005, pp. 168–173.

[10] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," in *Proceedings of the IEEE*, 1989, vol. 77, pp. 257–286.

[11] R. Nishimura, Y. Nishihara, R. Tsurumi, A. Lee, 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*, 2003, pp. 70–78.

[12] K. Ito, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi, "JNAS: Japanese Speech Corpus for Large Vocabulary Continuous Speech Recognition Research," *The Journal of the Acoustical Society of Japan*, vol. 20, pp. 199–206, 1999.

[13] A. Lee, T. Kawahara, K. Takeda, and K. Shikano, "A New Phonetic Tied-Mixture Model for Efficient Decoding," in *International Conference on Acoustics, Speech, and Signal Processing*, 2000, pp. 1269–1272.

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