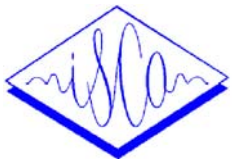

INTERSPEECH 2004 - ICSLP
8th International Conference on Spoken Language Processing
ICC Jeju, Jeju Island, Korea
October 4-8, 2004

Speech Recognition for Multiple Non-Native Accent Groups with Speaker-Group-Dependent Acoustic Models

Tobias Cincarek[1,2], *Rainer Gruhn*[1] *and Satoshi Nakamura*[1]

[1] ATR Spoken Language Translation Research Labs
2-2-2 "Keihanna Science City", Kyoto 619-0288, Japan
[2] University of Erlangen-Nuremberg, Germany
{tobias.cincarek,rainer.gruhn,satoshi.nakamura}@atr.jp

10.21437/Interspeech.2004-569

Abstract

In this paper, the recognition performance for non-native English speech with two different kinds of speaker-group-dependent acoustic models is investigated. The approaches for creating speaker groups include knowledge-based grouping of non-native speakers by their first language, and the automatic clustering of speakers. Clustering is based on speaker-dependent acoustic models in speaker Eigenspace. The acoustic model for each speaker group is obtained by bootstrapping with pre-segmented speech data or adaptation of a speaker-independent native baseline model. For the decoding of a non-native speaker's utterance not seen during the training or adaptation phase, the selection of a model suitable to cope with the accent characteristics of that speaker is necessary. Here, ideal selection via an oracle and parallel decoding are examined. Evaluation is conducted in a hotel reservation task for five major accent groups, including German, French, Indonesian, Chinese and Japanese speakers. Recognition results with speaker-dependent and an accent-independent non-native model will also be reported.

1. Introduction

Approaches for the recognition of non-native speech found in literature can be classified into three major classes: pronunciation modeling by altering word baseforms, acoustic modeling, and combinations of both. Pronunciation modeling can boost speech recognition performance for a certain foreign accent of the target language, e.g., by adding pronunciation variants to the dictionary, or better applying confusion rules to the phoneme or word lattice during decoding [2] [9]. Such an approach requires either in-depth knowledge of the target language and the first language of the non-native speaker to be able to design pronunciation variants manually, or large amounts of labeled speech data to extract them automatically. Furthermore, that approach covers only deletions, insertions and substitutions of target language phones. To account also for substitutions of target language phones with phones of the non-native speaker's first language, each word in the dictionary can be represented as a sequence of target language phones and a sequence of the non-native speaker's first language phones to form quasi-bilingual models [7] at the same time. An approach for several accents with multi-lingual acoustic models for recognizing digit strings by combining the phone sets and speech data of several native languages was shown to work better than MLLR adaptation of monolingual models with accented speech data [5]. The advantage of bi- and multi-lingual models is the availability of training data from native speech corpora without the need for collecting large amounts of accented data.

However, recent investigations of Flege et al. [6] suggest that non-native speakers may produce speech sounds which are either part of their first language or which were established by merging characteristics of a first language with a target language speech sound. These observations lead to the conclusion that adaptation of each acoustic-phonetic unit model is necessary and that pronunciation modeling with native phone models alone may not be the silver bullet to improve recognition performance for non-native speakers of any accent group.

Approaches which incorporate this phenomenon are the adaptation of a speaker-independent baseline model of the target language [11] or training from scratch with accented speech data [10], the merging of native models [12], and interpolation between native and non-native models [11] or only native models [12]. All these methods lead to a remarkable improvement in recognition performance of foreign accented speech.

Other than the method based on multi-lingual models, the methods summarized here were applied separately for each non-native accent group. To decode an unseen test utterance, the first language of the test speaker needs to be known in order to select an appropriate acoustic model. This step can in principle be carried out by an accent classification system, e.g., realized by an approach based on ergodic HMMs [8], which achieved an accuracy of 65% for six accent groups.

Here we will present a practical system for the recognition of continuous non-native speech of multiple accent groups. For acoustic model selection, parallel decoding with speaker-group-dependent models is employed. Investigation is conducted to determine whether models built with speech data from speakers of knowledge-based speaker groups or models trained with data from data-drivenly created speaker groups are more suitable for recognition. To take account of Flege et al.'s findings, acoustic models are constructed by bootstrapping with pre-segmented non-native speech data. With this approach non-native phones and pronunciation can be learned automatically and coded statistically as HMM mixture distributions.

2. Overall approach

The proposed approach consists of an off-line step for the construction of speaker-group-dependent acoustic models and an on-line step for model selection and test utterance decoding. Figures 1 and 2 illustrate the processing of both steps.

Speaker group formation is necessary for the off-line step. Two methods are investigated: a straightforward knowledge-based approach by grouping the non-native speakers by their first language to build accent-dependent models, and an

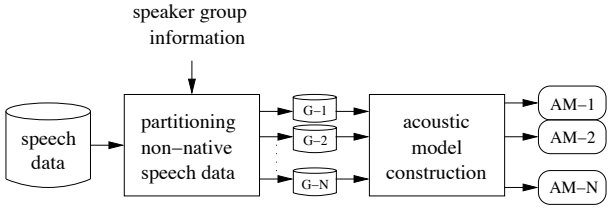

Figure 1: *Off-line step: Construction of speaker-group-dependent acoustic models. The non-native speech data is partitioned into several training data sets by taking together all the training data from the speakers of each group. From each data set one acoustic model is constructed.*

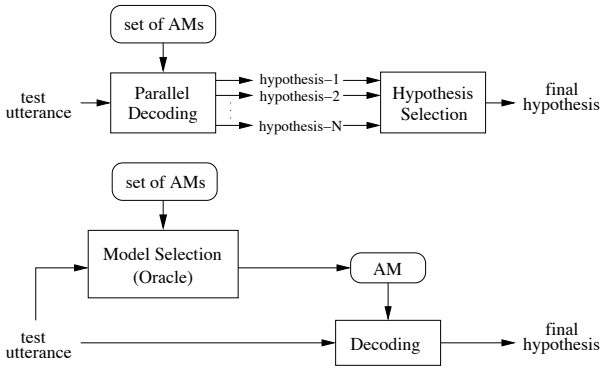

Figure 2: *On-line step: Decoding of a test utterance. Top: A test utterance is decoded in parallel with each acoustic model available. The hypothesis with maximum likelihood is selected as the final hypothesis. Bottom: Assuming that a perfect speaker group classification or speaker identification system is available, oracle-based acoustic model selection is simulated to get a reference value for maximum possible performance.*

automatic approach by clustering the speakers in speaker Eigenspace to build cluster-dependent models.

2.1. Speaker clustering

A clustering method in speaker Eigenspace-based on Eigenvoices was introduced in [4] and applied to cluster native speakers. Here the clustering scheme is applied to non-native speakers. Eigenspace-based methods have the advantage, that complex representations can be transformed into simpler ones with few parameters while retaining most of the original information.

Speaker-dependent (SD) models were constructed by MAP adaptation of mean vectors of a single mixture speaker-independent monophone model. This adaptation procedure has the advantage, compared with training from scratch, that all remaining model parameters remain the same. Only the mean vectors need to be considered in further processing.

The mean vectors of each SD model were extracted and concatenated to a high-dimensional (39 features $*$ 44 models $*$ 3 states) supervector. Constructing these vectors from GMMs or multiple mixture HMMs would pose some difficulty of alignment between the mixture components of each speaker's model. The correspondence is already given implicitly for single mixture HMMs by model names and state topology.

Principal Component Analysis (PCA) based on the correlation matrix was applied to the supervectors to obtain a basis for an Eigenspace covering most of the sample variance. Finally, the supervectors were projected into this Eigenspace in order to obtain a low-dimensional representative for each speaker, which is suitable for clustering.

As clustering algorithms, k-means and agglomerative hierarchical clustering with four different kinds of inter-cluster distance measures as described in [3] were examined. Inter-vector distances were measured by the Euclidean distance. For the k-means algorithm, each cluster was initialized with the speakers of each first language group.

3. Non-native speech database

Read speech data of about 100 non-native English speakers was collected at ATR. It is clean speech recorded at 16-kHz sampling frequency and 16-bit precision. The data consists of 48 phonetically balanced sentences of the TIMIT set and six hotel reservation dialogs. To be able to abstract from variability introduced by gender, only the data of 75 non-native male speakers was actually used. The speaker set utilized for experiments consists of 15 Japanese, 15 Chinese, 15 French, 15 German and 15 Indonesian natives. All speakers utter the same sentences. The training and adaptation data set comprises 88 utterances ($\approx$ 10 minutes), the validation data set ten utterances ($\approx$ 1 minute) and the test data set 23 utterances ($\approx$ 3 minutes) per speaker.

For comparison of recognition performance of natives vs. non-natives, speech data of six native English speakers uttering the same test sentences as the non-native speakers was used.

4. Baseline system

HTK was employed for training and adaptation of all acoustic models, building of the language model and decoding in all evaluation experiments. The configuration of the baseline system is as follows:

4.1. Acoustic model

More than 60 hours (37,413 utterances) of speech data from the LDC Wall Street Journal corpus (WSJ) were used to build three speaker-independent native English acoustics with different complexity:

1. 44 Monophone 3-state HMMs with 16 mixtures
2. State-clustered biphone model with about 3,000 states and 10 mixtures
3. State-clustered crossword triphone model with about 9,600 states and 12 mixtures

39 acoustic features, 12 MFCC coefficients and energy with first and second derivation, were extracted every 10 ms. The word accuracy of these three acoustic models on the Hub2 5K evaluation task was 80.8% for the monophone, 86.8% for the biphone and 93.6% for the triphone model.

Since only speech data from male non-native speakers are considered in this research a gender-dependent monophone model was built by MAP adaptation of the SI baseline AM with the speech data of all male speakers from WSJ.

4.2. Language model

The n-gram probabilities were estimated from a database with 235 dialogs in the hotel reservation domain comprising 6,460 utterances with 65,893 words in total. The lexicon contained about 8,800 entries for about 7,300 words including compounds. The perplexity for the 344-word evaluation task, two dialogs with 23 utterances in total, was 32.

5. Results

For evaluation, 75-fold leave-one-speaker-out cross validation was carried out for all experiments with speaker-group-dependent models in order to obtain a realistic estimate of performance. Speaker-group-dependent models consist of 42 HMMs with ten mixtures. In each table the average word accuracy for each speaker group is shown. Initially three approaches for construction of the accent-dependent models were examined. Performance was best for bootstrapping the models with pre-segmented non-native speech data obtained by forced-alignment with the monophone SI native baseline model. Results were slightly lower for training the models from scratch, followed by MAP adaptation of the native monophone SI baseline.

5.1. Speaker-independent models

Recognition accuracy with the speaker-independent baseline system as described in section 4 varies remarkably for each acoustic model and speaker group. For native speakers there was an increase, or at least no decrease in accuracy on average when decoding with the biphone and the triphone model in comparison to the monophone model. However, for non-native speech severe degradations can be observed (see Table 1). While the relative drop in accuracy was rather low for German speakers with about 8%, error rates almost doubled for Japanese speakers. This may be due to phonetic errors and different coarticulation of speech sounds especially for speakers whose first language is Japanese and Chinese, who have fewer speech sounds in common with the English language than German, French or Indonesian. A comparison of IPA-based [1] phone sets revealed that German has at least 28, Indonesian 26, French 25, Mandarin Chinese 21 and Japanese 19 phones in common with American English. In further experiments monophone models are employed, because they are more robust to accent variability and require less data for training and adaptation than context-dependent models.

Table 1: *Recognition performance with the speaker-independent (SI) native English baseline acoustic models.*

Model	Eng	Ger	Fre	Ind	Jap	Chi
SI mono	82.4	75.7	71.9	70.7	55.4	63.3
SI biph	82.5	72.5	66.9	63.1	42.9	54.9
SI triph	85.5	69.6	62.1	51.7	31.4	39.0

5.2. Speaker clustering

Several clustering methods with different distance measures were examined. Hierarchical clustering produced rather balanced clusters for the furthest neighbour distance but rather sparse clusters for the centroid distance and average inter-vector distance, and very sparse clusters for the nearest neighbour metric. The tendency of producing sparse clusters also increased with the dimension of the Eigenspace. Since clusters generated by the k-means algorithm were more balanced, even when setting the Eigenspace dimension to 20, being equivalent to capturing nearly 95% of sample variance, their speaker configurations were used for building the cluster-dependent models. The distribution of speakers' first languages in each cluster is shown in table 2.

Table 2: *Distribution of non-native speakers in clusters created by k-means clustering in a 20-dimensional Eigenspace.*

Cluster	1	2	3	4	5
Chinese	-	4	2	5	4
French	7	5	-	3	-
German	3	8	-	1	3
Indonesian	-	2	-	3	10
Japanese	-	-	13	1	1

5.3. Oracle experiments

Knowing the first language, cluster membership or identity of the test speaker, several oracle-based experiments for obtaining reference values for maximum recognition performance can be carried out. Results are summarized in Table 3.

An estimate of maximum possible performance for each non-native speaker can be obtained by decoding with speaker-dependent (SD) models. There is a remarkable increase in accuracy for all non-native speakers in comparison to the gender-dependent baseline model. The performance with SD models for six native English speakers was 92.6%, indicating that non-native speech is indeed more variable than native speech.

The performance with accent-dependent (AD) models is also high, suggesting that the accent characteristics of speakers having the first language in common are similar. The difference in accuracy between SD and AD models is largest for the Chinese speaker group, which can be explained by the fact that this group consists of speakers from several areas of China, also including some speakers whose first language is Cantonese.

Recognition with cluster-dependent (CLD) models still leads to good performance, but is slightly lower than that with accent-dependent models.

Table 3: *Recognition performance with the gender-dependent (GD) native baseline monophone model, the speaker-dependent, the accent-dependent and the cluster-dependent models.*

Model	Ger	Fre	Ind	Jap	Chi
GD baseline	75.7	73.8	71.9	55.6	63.7
Cluster-dep.	80.1	82.8	82.9	82.1	75.5
Accent-dep.	82.7	84.4	85.4	82.2	77.3
Speaker-dep.	87.2	87.5	87.7	84.6	82.8

5.4. Parallel decoding

In order to build a practical ASR system for non-native speech recognition, parallel decoding with the accent-dependent or the cluster-dependent models was employed. This procedure yields one recognition hypothesis from each acoustic model for an unseen utterance. The hypothesis with maximum acoustic likelihood was selected as the final recognition result.

The results for parallel decoding are summarized in Table 4. There is a small drop in recognition accuracy in comparison to the oracle experiment of section 5.3 for the accent-dependent models, but no performance decrease for the cluster-dependent models. Furthermore, both model types yield better results than the GD baseline. The difference in accuracy between AD and CLD models is significant for the Japanese speaker group only.

While the cluster classification accuracy (64.6%) was higher than the accent classification accuracy (52.5%), paral-

Table 4: *Non-native speech recognition performance by parallel decoding with accent-dependent (AD) or cluster-dependent (CLD) models.*

Model	Ger	Fre	Ind	Jap	Chi
AD parallel	80.6	83.6	83.0	80.4	75.5
CLD parallel	80.1	82.8	82.9	82.0	75.5

lel decoding with CLD models may in practice perform better than decoding with AD models if data of more speakers become available. The results for each speaker group are summarized in Figure 3.

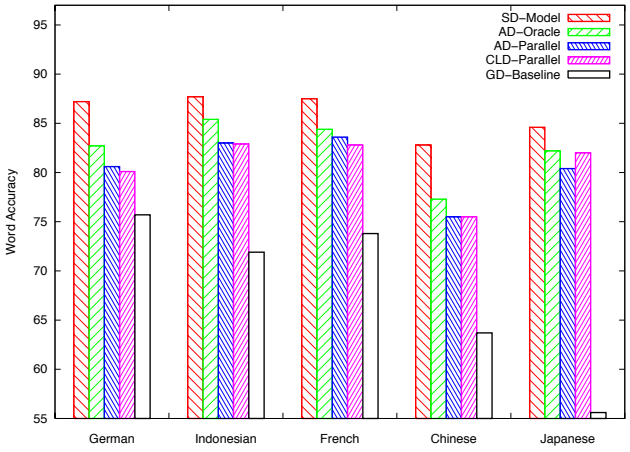

Figure 3: *Comparison of experimental results.*

5.5. Non-native model

To investigate whether the pronunciation variations of all speakers from the considered five accent groups can be captured by only one monophone acoustic model, the data of 50 non-native speakers, 10 from each accent group, were used to train one 16 mixture non-native monophone model (NN). Evaluation was carried out with the remaining 25 speakers, doing 3-fold cross-validation. The speakers for each training and test set were randomly selected, taking care that the first languages of speakers in training and test sets were distributed uniformly.

Table 5: *Performance with a non-native monophone model.*

Model	Ger	Fre	Ind	Jap	Chi
SI NN	80.5	83.1	83.2	79.7	79.8

As Table 5 illustrates, the performance is almost equal to parallel decoding with AD or CLD models, except for the rather accent-inhomogeneous Chinese speaker group, which may be due to the higher robustness of the non-native model, which was trained with 50 speakers, than the AD and CLD models, for which data of only 15 speakers were available. Since accuracy with AD models in the oracle experiments of section 5.3 is still significantly higher than that with the NN model, accent-dependent models may in principle perform better than accent-independent models, whenever an accent classification system with high accuracy is available.

6. Conclusion

A practical approach for non-native ASR was introduced. It is based on parallel decoding with several speaker-group-dependent monophone acoustic models and maximum likelihood hypothesis selection. Good accuracy with accent-dependent models was achieved for five non-native accents groups with a relative improvement of 6% up to 44% on average to a GD native baseline depending on the speaker group.

The maximum recognition performance with monophone models is limited. However, as long as large corpora of non-native speech are not available, training of robust context-dependent acoustic models is infeasible. Assuming rather consistent pronunciation variations of non-native speakers within each accent group, higher accuracy may be possible with accent- and context-dependent models.

7. Acknowledgment

This research was supported in part by the National Institute of Information and Communications Technology (NICT) of Japan.

8. References

[1] *Handbook of the International Phonetic Association*. Cambridge University Press, 1999.

[2] Nobert Binder, Rainer Gruhn, and Satoshi Nakamura. Recognition of non-native speech using dynamic phoneme lattice processing. In *Proceedings of Acoustical Society of Japan*, pages 203–204, March 2002.

[3] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification*. John Wiley & Sons, Inc., 2001.

[4] R. Faltlhauser and G. Ruske. Robust speaker clustering in eigenspace. In *IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 57–60, 2001.

[5] V. Fischer, E. Janke, and S. Kunzmann. Recent progress in the decoding of non-native speech with multilingual acoustic models. In *European Conference on Speech Communication and Technology*, pages 3105–3108, 2003.

[6] James E. Flege, Carlo Schirru, and Ian R. A. MacKay. Interaction between the native and second language phonetic subsystems. *Speech Communication*, 40:467–491, 2003.

[7] S. Matsunaga, A. Ogawa, Y. Yamaguchi, and A. Imamura. Non-native english speech recognition using bilingual english lexicon and acoustic models. In *Proc. of ICASSP*, volume 1, pages 340–343, 2003.

[8] Carlos Teixeira, Isabel Trancoso, and Antonio Serralheiro. Accent identification. In *Proc. of ICASSP*, volume 3, pages 1784–1787, 1996.

[9] Laura Mayfield Tomokiyo. Lexical and acoustic modeling of non-native speech in LVCSR. In *Proc. of the ICSLP*, pages 1619–1622, 2000.

[10] Laura Mayfield Tomokiyo and Alex Waibel. Adaptation methods for non-native speech. In *Proceedings of Multilinguality in Spoken Language Processing*, 2001.

[11] Zhirong Wang, Tanja Schultz, and Alex Waibel. Comparison of acoustic model adaptation techniques on non-native speech. In *Proc. of ICASSP*, pages 540–543, 2003.

[12] Silke Witt and Steve Young. Off-line acoustic modelling of non-native accents. In *European Conference on Speech Communication and Technology*, volume 3, pages 1367–1370, 1999.