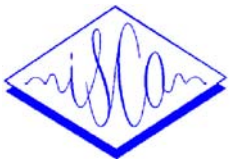
ISCA Archive
http://www.isca-speech.org/archive



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

Foreign-Accented Speaker-Independent Speech Recognition

Stefanie Aalburg, Harald Hoege

Siemens AG
Corporate Technology, 81730 Munich, Germany
firstname.lastname@siemens.com

Abstract

This research investigated whether acoustic-phonetic knowledge of the mother tongue of a non-native speaker can be used to adapt an existing target language phoneme HMM recognizer. For this purpose three sets of phoneme HMMs were generated, one representing the target language (German), one the mother tongue of the non-native speaker (Turkish), and the third the foreign-accented pronunciation of the target language (German spoken by Turkish speakers). The latter served as a benchmark for the tested adaptation methods. A derived Hidden Markov Model (HMM) clustering algorithm was applied on the target language phoneme HMM set using the mother tongue phoneme HMM set of the non-native speaker. Following the HMM adaptation a phoneme-level pronunciation technique was applied to generate phoneme mapping rules for the lexicon adaptation task. The results revealed a relative reduction of about 6% in WER for the adapted HMM. No further improvements were observed from the lexicon adaptation task.

1. Introduction

1.1. The Problem

Current research is focused on improving the speech recognition performance of recognition systems when confronted with foreign-accented speech. Typically speech recognition engines are monolingual, i.e. they are designed to recognize native speech of languages spoken by either a large group of people or representing economically strong countries. It is a well-known fact that the recognition performance degrads in case of non-native speech, since target phonemes are often assimilated or replaced by phonemes of the native language of the speaker [1] [2]. Furthermore the pronunciation space of the individual phonemes may differ due to different possible contexts due to the different phonology of the mother tongue of the non-native speaker. The different acoustic-phonetic properties of the foreign-accented speech are taken into account by Hidden Markov Model (HMM) adaptation.
To account for different word pronunciations of the non-native speaker, additional phonemic representations of the words in the lexicon may be required as well, thus lexicon adaptation is performed in addition to the HMM adaptation task.

1.2. Previous Work

To account for foreign-accented speech, recognition engines are often adapted in a speaker-dependent manner using the Maximum A Posteriori (MAP), or Maximum Likelihood Linear Regression (MLLR) matrixes for HMM adaptation [3].
An alternative to adaptation is the development of multi-lingual speech recognition engines [4]. Based on this approach a derived mean-vector clustering method using two sets of phoneme HMMs was developed [5]. This method is now applied on two sets of phoneme HMMs that are trained on native speech only, the one representing the target language and the other the mother tongue of the non-native speaker.
Based on the phoneme-level pronunciation scoring technique that has been used by [6] in interactive language learning, phoneme mapping rules are generated to extend the lexicon with additional word pronunciations. The adapted lexicon thus should be able to deal with different word pronunciations generated by the non-native speaker.

Section 2 describes the method followed in this approach. The following section 3 outlines the tested distance measures, the clustering methods used for phoneme-based HMM adaptation, and the lexicon adaptation technique. The experimental results and the respective discussion are given in section 4 and 5.

2. The Approach

The approach described here aims to improve the recognition performance of foreign-accented speech by incorporating acoustic-phonetic knowledge of the mother tongue of the non-native speaker into the target language phoneme HMMs, as well as varying word pronunciations into the lexicon.
The HMM adaptation is based on two sets of phoneme HMMs that were trained on native speech only, i.e. the one representing the target language and the other the mother tongue of the non-native speaker. A third set of phoneme HMMs is directly trained on foreign-accented speech and is used for benchmarking the recognition results of the adapted target language phoneme HMMs.
During adaptation the mean-vectors (*prototypes*) of the continuous probability density functions (pdfs) of the target language phoneme HMMs are merged according to some criteria with one or several pdfs from the mother tongue phoneme HMMs of the non-native speaker. As a fundamental criteria for merging prototypes served a distance value that is calculated between each prototype of the two sets of phoneme HMMs, thereby generating a full distance matrix. Several different distance measures have been tested in this approach.
Given the distances of the matrix, several methods of prototype clustering have been investigated:

- The usage of distance thresholds to define similarity between components
- The setting of a maximal number of prototypes that are used for clustering
- The introduction of individual component weights to emphasize or attenuate the influence of a component during clustering

Following the adaptation HMM adaptation, the phoneme-level pronunciation scoring technique is applied to generate phoneme

mapping rules for the lexicon adaptation, as explained in [5]. The mappings are derived during a forced-alignment recognition process using the adapted phoneme HMMs with the target language label files. During recognition, for each phoneme a score is calculated that expresses the pronunciation "goodness". In case of a low score the best alternative phoneme of the adapated phoneme set is marked as a possible candidate for a phoneme mapping rule. Later the extracted phoneme mapping rules are sorted according to their relative frequency of occurrence and only the most frequent rules are selected for lexicon adaptation.

3. Acoustic Model and Lexikon Adaptation

3.1. Distance Measures for Pdf-Based Clustering

Acoustic-phonetic similarities between the prototypes of the target language HMM set and mother tongue HMM set can be identified using distance measures. An extensive survey about the most frequently used distance measures is given in [4], out of which the following four distance measures were tested for the prototype clustering:

1. Euclidean Distance Metric:

$$D_{Euk}(i,j) = \frac{1}{D}\sum_{d=1}^{D}(\mu_{d,i} - \mu_{d,j})^2 \quad (1)$$

   with i and j representing the prototypes of the target language HMM set and the mother tongue HMM set respectively. D denotes the dimension of the feature vectors.

2. L1 Distance Metric:

$$D_{L1}(i,j) = \frac{N_i N j}{N_i + N_j}\sum_{d=1}^{D}|\mu_{d,i} - \mu_{d,j}| \quad (2)$$

   where N corresponds to a weighting factor for each pdf, i.e. the weight corresponds to the number of times a pdf was seen during training.

3. Approximated Divergence:

$$D_{div}(i,j) = \frac{1}{D}\sum_{d=1}^{D}\frac{(\mu_{d,i} - \mu_{d,j})^2}{\sigma_{d,i}\sigma_{d,j}} \quad (3)$$

   where σ stands for the variance of the pdfs.

4. Log-Likelihood Distance:

$$D_{LL}(\lambda_i, \lambda_j) = \log p(X_i|\lambda_i) - \log p(X_i|\lambda_j) \quad (4)$$

   with X_i representing the observation data of HMM_i (λ_i). The cross distance $D_{LL}(\lambda_j, \lambda_i)$ is defined accordingly. Since the distances are not symmetric the Log-Likelihood distance between two HMMs (λ) is defined as:

$$D_{LL}(\lambda_i; \lambda_j) = \frac{1}{2}(D_{LL}(\lambda_i, \lambda_j) + D_{LL}(\lambda_j, \lambda_i)) \quad (5)$$

3.2. Pdf-Based Adaptation Techniques

Given the distance values between the individual prototypes there are several possibilities to perform pdf-based adaptation. At first a full distance matrix between all pdf mean-vectors (prototypes) is calculated using one of the above distance measures. In a second step the components that are to be merged are identified.

1. Minimum Distance:
   - while $D_{min} < D_{threshold}$ merge mother-tongue prototype with the target language prototype, thereby *iteratively* adapting it.
   - adaptation occurs across phonemes
   - no maximum number of prototypes used for merging is defined, $D_{threshold}$ is the sole constraint during adaptation
2. Equally Weighted Distances:
   - define a maximum number of non-native prototypes that are used for adapting the target language prototype
   - estimate the adaptated target language mean-vector as follows:

$$\hat{\mu} = N\mu_0 + N\mu_1 + ... + N\mu_{max} \quad (6)$$

     with μ_0 representing the original target language pdf mean-vector, and μ_1 ... μ_{max} representing the maximum number of mother-tongue prototypes. N is equal $\frac{1}{max+1}$.
3. Distance Dependent Weighting:
   - define a maximum number of mother-tongue prototypes that are used for adapting the target language prototype
   - estimate the adapted target language mean-vector as follows:

$$\hat{\mu} = N_0\mu_0 + N_1\mu_1 + ... + N_{max}\mu_{max} \quad (7)$$

     with

$$N_0 = 1 - \alpha$$

     and α represents a *learning rate* that is used to determine the influence of the mother tongue prototypes during adaptation, i.e. how fast are the original values "forgotten" [7].

$$N_i = \alpha \times \left(\frac{1 - (D_i/D_{threshold})}{\sum_{i=1}^{max} 1 - (D_i/D_{threshold})}\right)$$

     with $i = 1...max$ and D_i representing the distance of the prototype μ_i to μ_0.
4. Learning-Rate Dependent Weighting:
   - use maximum number of non-native prototypes for adapting the target language prototype
   - estimate the adapted prototype as follows:

$$\hat{\mu} = \mu_0(1-\alpha)^{max+1} + \sum_{i=0}^{max-1}\mu_{max-i}\alpha(1-\alpha)^i \quad (8)$$

     with α being the *learning rate* as defined in 7.

3.3. Lexicon Adaptation

The lexicon adaptation is applied after the pdf-based phoneme HMM adaptation, i.e. the adapted phoneme HMMs are forced-aligned with the target language label files, thus creating a phoneme lattice that serves as input for a phoneme-level pronunciation scoring. For each phomeme of the lattice the Viterbi-score is calculated and set into relation with the best Viterbi-score that is found during an open-loop phoneme recognition. The score will be equal to "1" in case both phonemes, the one given by the lattice and the one found during the open-loop recognition, are identical. See [5] for more details.

4. Experimental Results

4.1. Experimental Setup

The experiments are conducted with speaker-independent continuous density phoneme HMMs, where each phoneme consists of three segments. Each segment has two tied states that share the same Gaussian mixture density. The pdf mean-vectors (prototypes) consist of 24 Mel-Filter Cepstrum coefficients and are derived from a Linear Discriminant Analysis (LDA) of a 39 dimensional feature vector.
Three sets of phoneme HMMs were trained for testing and evaluating the described adaptation methods.

1. Target Language Phoneme HMMs (German), trained on SpeechDat(II) German database
2. Non-Native Phoneme HMMs (Turkish), trained on Orientel Turkish database
3. Forreign-Accented Phoneme HMMs (accented German), trained on Orientel Database of German spoken by Turkish speakers

Since all three sets of phoneme HMMs should be as similar as possible in respect to the represented acoustic properties and recognition abilities, the choice of training and testing material is crucial for successful adaptation. For each phoneme set the training material consisted of a selection of phonetically balanced utterances and isolated spoken words. All utterances were recorded through the fixed telephone network and spoken in a quite environment, e.g. home or office.
All databases contain mobile phone recordings and noisy recordings but due to the constraint to generate similar training conditions for all three sets of phoneme HMMs the chosen utterances belonged to the conditions above.
The test set consisted of isolated spoken application words again recorded through the fixed telephone network and spoken in home or office environment.
The following describes the training and test material in detail:

1. The German phoneme HMMs were trained on a total of 6140 utterances and tested on 3786 utterances representing 37 isolated spoken application words.
2. The Turkish training and testing set consisted 7150 and 3144 utterances respectively, where the vocabulary comprised 26 application words.
3. The foreign-accented German phoneme set was trained on a total of 1543 utterances and tested on 630 utterances, representing 34 applicaton words.

The SAMPA (Speech Assessment Methods Phonetic Alphabet) [8] notation was used for the lexica generation. The German set of phonemes consists of 39 phonemes, whereas the Turkish of 36 phonemes. The following Turkish SAMPA phonemes are not included in the German phoneme set: /e/, /i/, /o/, /2/, /y/, /w/, and /Z/.
Out of this set the phonemes /w/ and /Z/ do not have any acoustic-phonetic similar equivalent in the German phoneme set, thus are candidates for assimilations or replacements in the accented German speech.

4.2. Results of the Pdf-based phoneme HMM Adaptation

The following table gives the word recognition rates (WRR) for each of the three sets of phoneme HMMs, i.e. the German phoneme HMMs tested on SpeechDat(II), the Turkish phoneme HMMs tested on the Orientel Turkish database, and the foreign-accented phoneme HMMs tested on the Orientel database German spoken by Turkish.

Furthermore a cross-test recognition test was performed to obtain the baseline recognition performance of the German and the foreign-accented set of phoneme HMMs. The results are summarized in Table 1.

Table 1: *WRR in % of the Three Baseline Phoneme HMM Sets*

Database / Phoneme HMMs	SpeechDat(II) Database	Orientel Turkish Database	Orientel German Spoken by Turkish Speakers Database
German HMMs	94.0 %		86.7 %
Turkish HMMs		91.3 %	
Foreign-Accented HMMs	90.6 %		87.9 %

The WRR of the foreign-accented phoneme HMMs is quite low in comparison with the other two phoneme sets. This is due to the limited amount of training material, consisting of fixed network and home/office recordings. These recording conditions were the constraint for successful training of the Turkish phoneme HMM set, which suffered a great loss in recognition accuracy if noisy or mobile phone recordings were included in the training set.

The pdf-based clustering was applied to the German set of phoneme HMMs, using the Turkish (mother tongue) phoneme HMMs for adaptation. The adaptation performance was tested for the different distance measures and adaptation methods, where the maximum number of mother-tongue prototypes used for clustering, the "learning rate" α, and the distance threshold were found experimentally.

Table 2 illustrates the best WRR obtained for each distance measure and adaptation method. The adaptation methods are represented by numbers according to their presentation in section 3.2.

Table 2: *WRR in % of the Adapted HMM Sets*

Distance Measure	Adaptation Method	Speechdat(II) Database	Orientel German Spoken by Turkish Speakers Database
Euclidean Distance	1	93.9 %	85.6 %
	2	93.9 %	85.6 %
	3	94.0 %	85.6 %
	4	93.9 %	85.8 %
L1 Distance	1	93.9 %	85.9 %
	2	93.9 %	85.8 %
	3	93.9 %	85.9 %
	4	93.8 %	85.9 %
Approx. Divergence Distance	**1**	**94.2 %**	**87.5 %**
	2	94.0 %	86.9 %
	3	94.0 %	87.2 %
	4	93.9 %	87.2 %
Log-Likelihood Distance	1	93.8 %	87.2 %
	2	93.7 %	86.9 %
	3	93.8 %	86.7 %
	4	93.5 %	87.0 %

4.3. Results of the Lexicon Adaptation

The lexicon adaptation was performed on the best adapted target language phoneme HMM. As Table 2 indicates the best results were obtained with the approximate divergence distance measure and the minimum distance adaptation method.
Table 3 gives the result of the phoneme-level pronunciation scoring technique, which revealed some alternative phoneme pronunciations of the non-native speaker.

Table 3: *Phoneme Mappings of the Non-Native Speaker*

Phoneme of / Lattice	Best Phoneme of Open-Loop Recognition
/o:/	/O/
/e:/	/E/
/N/	/n/
/a:/	/a/
/I/	/i:/
/O/	U
/@/	/E/

The left column of Table 3 gives the phonemes specified by the lattice and the right column gives the phonemes that received higher Viterbi-scores, thus reflecting the effect of incorporating Turkish phoneme pronunciations.
After lexicon adaptation a new recognition test was performed. Table 4 presents the WRR of the adapted HMM with and without lexicon adaptation.

Table 4: *WRR in % of Best Adapted HMM with and without Lexicon Adaptation*

	Best Adapted / HMM	Orientel German spoken by Turkish Speakers Database
without Lexicon Adaptation	Aprox. Divergence and min. Dist adapted HMM	87.5 %
with Lexicon Adaptation	Aprox. Divergence and min. Dist adapted HMM	87.5 %

The lexicon adaptation did not lead to any further improvement, which might be due to the fact that the size of test vocabulary was too small and the word confusability comparatively low.

5. Conclusions and Discussion

An approach to improve the recognition performance of foreign-accented speech by a mono-lingual speech recognition system was presented. The basis for pdf-based HMM adaptation were two native phoneme HMM sets, one representing the target language of the non-native speaker, and the other the mother tongue of the non-native speaker. The idea was to investigate whether the collection of foreign-accented speech databases can be avoided by using the acoustic-phonetic knowledge of the mother tongue of the non-native speaker for a rather simple adaptation procedure.
For evaluation purposes a phoneme HMM set of foreign-accented speech was trained, representing the unrealistic case of having foreign-accented speech available for training or adaptation. Thus, the WRR of the foreign-accented set of phoneme HMMs given in table 1 represent the **best case** of foreign-accented speech recognition, whereas the WRR of the German phoneme HMM set tested directly on the foreign-accented speech represents the **worst case** of foreign-accented speech recognition. As shown in Table 1 the margin between these two cases is only 1.2 % of absolut improvement when using a foreign-accented speech recognition engine.
Considering the small margin of 1.2 % absolut WRR improvement, the results presented in bold face in Table 2 and in Table 4 correspond to a relative improvement of 66 % of the WRR of the adapted phoneme HMM set and 0.8 % absolut increase of WRR. As an interesting outcome the recognition results for the target language also show a slight improvement.
Therefore it appears useful to incorporate the different acoustic-phonetic properties of the mother tongue of the non-native speaker into the target language recognition engine.
The lexicon adaptation technique did not improve the recognition results but reveals interesting acoustic-phonetic changes of the adapted phoneme HMM set.
Further tests will be carried out to analyse pronunciation habits of the non-native speaker, i.e. forced-alignment of the foreign-accented phoneme HMM set with the German label files will be performed. Another interesting aspect is the analysis of the influence of the German orthography on the foreign-accented pronunciation.

6. References

[1] Gornonzy, S., Sahakyan, M., and Wokurek, W. “Is Non-Native Pronunciation Modelling Necessary”, Eurospeech, Aalborg, 2001.

[2] Stemmer, G., Noeth, E., and Niemann, H., “Acoustic Modeling of Foreign Words in a German Speech Recognition System”, Eurospeech, Aalborg, 2001.

[3] Mayfield Tomokiyo, L., “Recognizing Non-Native Speech: Characterizing and Adapting to Non-Native Usage in LVCSR”, Pittburgh, Carnegie Mellon University, Phd. Thesis, 2001.

[4] Koehler, J., “Erstellung einer statistisch modellierten multi-lingualen Lautbibliothek fuer die Spracherkennung”, Muenchen, Technische Universitaet, Dissertation, Shaker Verlag, Aachen, 1999.

[5] Aalburg, S. and Hoege, H., “Approaches to Foreign-Accented Speaker-Independent Speech Recognition”, Eurospeech, Geneva, 2003.

[6] Witt, S., M., and Young, S., J., “Phone-Level Pronunciation Scoring and Assessment for Interactive Language Learning”, Speech Communication, Vol. 30, 2000, pp. 95-108.

[7] Bub, U., “Anwendungsspezifische Online-Anpassung von Hidden-Markov-Modellen in automatischen Spracherkennungssystemen”, Muenchen, Technische Universitaet, Dissertation, Herbert Utz Verlag, Muenchen, 1999.

[8] http://www.phon.ucl.ac.uk/home/sampa/turkish.htm