Approaches to Foreign-Accented Speaker-Independent Speech Recognition

Stefanie Aalburg, Harald Hoege

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

Abstract

Current research in the area of foreign-accented speech recognition focusses either on acoustic model adaptation or speaker-dependent pronunciation variation modeling.
In this paper both approaches are applied in parallel and in a speaker-independent fashion: the acoustic modeling part is based on a derived Hidden Markov Model (HMM) clustering algorithm and the lexicon adaptation is based on speaker-independent multiple pronunciation rules. The pronunciation rules are derived using phoneme-level pronunciation scores. Foreign-accented speech was simulated with Columbian Spanish and Spanish of Spain and the experiments showed an improved recognition performance for the acoustic modeling part and identical recognition results when adding pronunciation variants to the lexica. Both results are taken as indicators for an improved recognition performance when applied on real foreign-accented speech. The present limited availability of foreign-accented speech databases, however, clearly merits further investigations.

1. Introduction

1.1. The Problem

Currently the industrial development of speech recognition engines focuses on monolingual recognition engines for languages spoken by either a large group of people or languages representing countries with high economical power. Such systems are often confronted with foreign-accented speech from people with a different mother tongue. Generally, the recognition performance for foreign-accented speech is very poor [1], [2]. This paper focuses on recognition errors caused by:

- Assimilations or replacements of target phonemes by phonemes from the native language of the speaker.
- Differences of the acceptable pronunciation space of the phonemes for the accented and non-accented speech.

Besides the different acoustic quality of the target phonemes, the accented word pronunciations may also be based on different sequences of phonemes. Therefore, different phonemic representations of the words in the lexicon may be required as well. To account for both types of errors, a combination of acoustic modeling and lexicon adaptation is followed here.

1.2. Previous Work

The derived HMM clustering is similar to the approach of multilingual speech recognition, which uses similarities across languages to generate robust acoustic models, e.g. by developing a multilingual phoneme set [3]. In contrast to the work performed by [2] and [3], foreign phonemes introduced in the accented speech are not represented by own phoneme HMMs, but are merged with their best match of the target language phoneme HMMs. This is justified by the fact, that foreign phonemes are used in either replacement or assimilation of existing target phonemes, and thus inhibit acoustic-phonetic similarities to target-phonemes that are to be identified, rather than treated as complete new sounds.
Phoneme-level pronunciation scoring is a common technique in interactive language learning [4], [5]. Here, forced alignment is used to evaluate the quality of pronunciation, based on Viterbi-scores that are calculated on a frame- or phoneme-level. In this work, not only the scores are estimated but also the best alternative phonemes, which are then used to generate context-dependent phoneme-mappings between the accented and the target language. In contrast to the speaker-dependent adaptation techniques for foreign-accented speech, as presented by [6], the aim is to derive speaker-independent phoneme-mappings. Furthermore, no hand-written rules or decision trees, as used in previous work [6],[7] are needed here.
Section 2 describes the approach to acoustic model and lexicon adaptation, and section 3 the acoustic adaptation method, including an analysis of several distance measures. Section 4 illustrates the proposed lexicon adaptation method, based on phoneme-level pronunciation scoring. The experimental results are presented and analysed in section 5.

2. Approach

So far approaches in foreign accent recognition have always focussed on either the change of acoustic quality of the phonemes or a change in word pronunciation. In this approach both effects are taken into consideration in parallel.
First, two phoneme-based HMMs are trained: one for the target language and the other for the "accented" speech. Several distance measures are tested to identify acoustic-phonetic similarities between individual segments or probability density functions (pdfs) of the phoneme HMMs. Segments are sub-units of the phoneme HMMs and thus represent shorter speech fragments with assumed stationary acoustic properties. The two HMM sets are input to pdf- and segment-based clustering.
Second, the two HMMs are used to derive phoneme-level pronunciation scores during a cross-data forced-alignment. To generate the phoneme lattice, foreign phonemes contained in labeled speech are mapped to their best alternatives. During recognition the phoneme-based pronunciation scores are calculated and used to identify possible context-dependent alternative pronunciations. The extracted mapping rules are sorted according to their relative frequency of occurrence and used for lexicon adaptation.
Due to a lack of databases the effect of foreign-accented speech can only be simulated here with Columbian Spanish representing the target language (mother tongue) and Spanish spoken in Spain representing the foreign-accented speech. This approach

was taken because recognition tests revealed room for improvement for Spanish speech recognized by a Columbian Spanish recognition engine (see table (2)).

3. Adaptation of Acoustic Models

3.1. Distance Measures for Data-Driven Clustering

In order to identify acoustic-phonetic similar segments or pdfs of two sets of HMMs, a suitable distance measure is required. [3] gives a survey and analysis of the commonly used distance measures used for merging HMMs. Since the *Bhattacharrya distance*, and *Mahalanobis distance* measure require full covariance matrixes, they are not considered in the following analysis. Instead, five different distance measures are tested in respect of their quality for phoneme-based HMM clustering:

1. The Log-Likelihood distance compares the HMM fit to its own and another HMMs' observation data, i.e.:

$$D_{LL}(\lambda_i, \lambda_j) = logp(X_i|\lambda_i) - logp(X_i|\lambda_j), \quad (1)$$

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

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

2. The Delta-Log-Likelihood distance measure is used to express the loss of information when two HMMs are merged. The smaller the distance value the smaller the loss of information (see [3]).

$$D_{\Delta LL}(i,j) = logp(X|\lambda_i) + logp(X|\lambda_j) - logp(X|\lambda_{i\cup j}), \quad (3)$$

where the first two terms denote the Log-Likelihood of the HMMs λ_i and λ_j, and $LL_{i\cup j}$ represents the Log-Likelihood of the merged HMM.

3. The Approximated Divergence is a common Gaussian distance measure for pdfs with diagonal covariances.

$$D_{div}(i,j) = \frac{1}{D}\sum_{d=1}^{D}\frac{(\mu_{d,i} - \mu_{d,j})^2}{\sigma_{d,i}\sigma_{d,j}}, \quad (4)$$

where D denotes the dimension of the feature vectors, and μ represents the mean-vector of the pdfs.

4. Assuming identical variances for all pdfs, equation (4) is reduced to the Euclidean distance.

$$D_{euk}(i,j) = \frac{1}{D}\sum_{d=1}^{D}(\mu_{d,i} - \mu_{d,j})^2 \quad (5)$$

5. A common alternative to the Euclidean distance is the L1-norm distance, which introduces a weighting factor N for each pdf, i.e. the weight corresponds to the number of times a pdf was seen during training.

$$D_{L1}(i,j) = \frac{N_i N_j}{N_i + N_j}\sum_{d=1}^{D}|\mu_{d,i} - \mu_{d,j}| \quad (6)$$

3.2. Adaptation

Clustering can either be based on the distance between individual pdfs or entire segments. Pdfs can be merged across segments and phonemes, in contrast to segments. At first, the algorithm searches for foreign phonemes and merges the respective segments and pdfs with target segments or pdfs. Second, all other segments or pdfs are merged, when their distance falls below a pre-set threshold.

The following describes the procedure for pdf-based clustering, which remains identical for all tested distance measures:

1. Calculate full distance matrix for all pdfs
2. Search for pdfs belonging to foreign phonemes and merge them according to the nearest neighbor criterium
3. While ($D_{min} < D_{thres}$) merge pdfs according to nearest neighbor criterium

D_{min} represents the minimum distance found for each of the 5 distance measures, and the upper limit D_{thres} is found experimentally. Using pdf-based clustering the number of initial and final pdfs remains unchanged.

For segment-based clustering the minimum distance between segments is either found with the Log-Likelihood distance, or as a sum of the distances of the pdfs of the two segments, as described by (7):

$$D = min(\sum_{i=1}^{N}\sum_{j=1}^{M} dist_{ij}), \quad (7)$$

where N is the number of pdfs belonging to segment i, and M is the number of pdfs belonging to segment j. $dist_{ij}$ represents the Euclidean distance measure, since it has a lowest computational complexity.

The segment-based clustering is performed as follows:

1. Find nearest segments for all foreign segments:
   - pool all pdfs of the two segments in one cluster
   - while ($N_{pdfs} > N_{thres}$)
     { reduce number of pdfs with pruning
2. While ($D_{min} < D_{thres}$) find similar segments
   - pool all pdfs of the two segments in one cluster
   - while ($N_{pdfs} > N_{thres}$)
     { reduce number of pdfs with pruning.

D_{thres} is experimentally found and N_{thres} is set to $\frac{1}{2}(N+M)$.

4. Modeling of Pronunciation Variants

4.1. Phoneme-Based Pronunciation Scores

Modeling pronunciation variations means to investigate the closeness of pronunciation between two languages. This is done with forced-alignment to create a phoneme-lattice that serves as input for the phoneme-level pronunciation scoring. For each phoneme of the lattice the corresponding Viterbi-score is calculated and set into relation with the best Viterbi-score found during an open loop recognition. In case the best matching phoneme for the given time-period is identical to the one given by the lattice, the score will equal "1". The algorithm calculates the mean score for each phoneme after having processed several thousands of utterances.

The phoneme-based pronunciation scores are calculated as follows:

$$S_{NLog} = \frac{-\sigma^2 logP(X|\lambda_{p\_Lattice})/No.ofFrames}{min_{p\in P} - \sigma^2 logP(X|\lambda_p)/No.ofFrames}, \quad (8)$$

where the numerator represents the Viterbi-score for the given feature vector sequence X and the phoneme ($p\_Lattice$) taken from the lattice. The denominator represents the best phoneme p found during an open loop Viterbi-search, where p is a member of the phoneme set P. Since the Viterbi probabilities are applied to a negative logarithmic transform, high probabilities are represented by low scores, hence the search for the minimal score ($min_{p\in P}$). The applied transform is a common method to convert the small Viterbi probabilities to higher numbers, and to replace the computationally rich multiplications by less complex additions during the Viterbi-score calculation.

4.2. Context-Dependent Phoneme-Mappings

Besides the pronunciation score, possible phoneme mappings are extracted, e.g. in case $\lambda_{p\_Lattice}$ and λ_p of equation (8) are not identical, phoneme p represents a possible mapping for the phoneme $p\_Lattice$. For each possible mapping the left and right neighbor of $p\_Lattice$ are taken from the lattice and categorized into one of the following groups:

1. Front Vowel (**FV**) = a, e, i
2. Back Vowel (**BV**) = o, u
3. Plosives (**P**) = p, t, k, b, d, g
4. Fricatives (**F**) = f, T, s, x
5. Nasal (**N**) = m, n, J
6. Liquids (**L**) = l, L, r, rr
7. Semivowels (**S**) = j,w

Since the aim was to derive general context-dependent phoneme-mapping rules, the number of different categories was reduced to a minimum. The resulting context-dependent mappings per phoneme are sorted according to their relative frequency.

5. Experimental Results

5.1. Experimental Setup

The recognizer consists of speaker-independent continuous density phoneme HMMs in Bakis topology. Each phoneme HMM is divided into three segments, where each segment has two tied states that share the same Gaussian mixture density. The segments serve as fundamental recognition unit and their pdf mean vectors consist of 24-dimensional Mel-Filter Cepstrum coefficients (MFCC) vectors, that were derived from a LDA of 39-dimensional super vectors. After the LDA transform the feature vectors are de-correlated and the covariance matrix of the Gaussian densities is reduced to a diagonal format.

The Columbian phoneme HMMs were trained on 3846 utterances of the SALA (SpeechDat Across Latin-America) database. The selected utterances contained a total of 35 isolated application words. The Spanish phoneme HMMs were trained on a total of 8423 utterances of the Spanish SpeechDat(II) database. The Spanish training and test set contained a selection of 32 isolated application words.

The SAMPA (Speech Assessment Methods Phonetic Alphabet) notation was used to represent the phonemes in the lexicon. There was a large overlap of phonemes for the Columbian and Spanish phonemes, except for the two Spanish phonemes /T/ and /J/ that did not occur in the Columbian phoneme set. These two phonemes serve as examples of foreign phonemes that may be introduced by non-native speakers. Thus for the recognition of Spanish speech by the Columbian or clustered HMM sets, the phonemes /T/ and /J/ were mapped as follows:
a) /J/ → /j/ b) /T/ → /s/

5.2. Results on Acoustic Modeling

Table (1) shows the matches found for the foreign phonemes /J/ and /T/ for those distance measures that lead to the best recognition results after merging.
After merging, the pdfs or segments of the new phoneme-HMMs are evaluated in a "cross-database" recognition test. The following table shows the recognition results of the base HMMs tested on their respective language and in "cross- database" test, as well as the recognition results for the best clustered HMMs tested on the SALA and SpeechDat(II) database.
The pdf-based clustering leads to slight improvements of the recognition results for both test sets. The small improvement is due to the fact, that the phonemes do not contain any quality degradation caused by assimilation of a foreign phoneme towards a target phoneme. In the experiments foreign phonemes occurred as replacements only but never as assimilations. Nevertheless a slight improvement was observed and serves as indication for a possible improvement for recognition of real foreign accented speech.

Table 1: *Best mapping of foreign phonemes to target phonemes for pdf-based and segment-based clustering*

Foreign Phonemes	Pdf-based Clustering Distance Measure: *Aprox. Divergence*	Seg.-based Clustering Distance Measure: *Log-Likelihood*
/J/	/n/, /j/, /b/, /d/,/L/, /tS/ /tS/ , /e/, /i/	/tS/
/T/	/s/, /f/, /r/, /g/, /x/ /j/, /r/, /e/, /i/	/x/

Table 2: *Recognition Results*

HMMs	tested on SpecchDat(II)	tested on SALA
Spanish	99.1 %	94.3 %
Columbian	78.2 %	96.2 %
Pdf-based Clustered HMM	79.7 %	96.4 %
Seg.-based Clustered HMM	78.2 %	96.2 %

5.3. Results of Lexicon Adaptation

For the calculation of phoneme-based pronunciation scores, the Columbian phoneme HMMs are aligned to a total of 8423 Spanish speech files. In the cross-database alignment the Spanish HMMs are aligned to a total of 3846 Columbian speech files. In the first case, the phonemes /J/ and /T/ of the Spanish label files are foreign to the Columbian phoneme set and were mapped according to section 5.1.
During alignment phoneme-based pronunciation scores are calculated according to equation (8). The table (3) gives some example pronunciation scores and context-dependent phoneme-mappings with the highest frequency.

Table 3: *Pronunciation scores and phoneme mappings for the Columbian phoneme HMMs aligned to Spanish labeled speech*

Phonemes of Lattice	Scores and Mapping Rules
/j/	1.01139
	7 % /i/ context: F - FV
/x/	1.03667
	42 % /s/ context: FV - FV
/l/	1.04748
	22 % /n/ context: FV - WE
/s/	1.00695
	6 % /f/ context: WB - FV

The contexts **WB** and **WE** represent **w**ord **b**eginning and **w**ord **e**nding respectively.
The scores indicate the closeness of pronunciation between Spanish and Columbian Spanish. The percentage indicates the relative frequency of a phoneme mapping compared to all other mappings. In table (3) the score of the phoneme /s/, representing the Spanish pronunciation, indicates that only very few times a different Columbian phoneme would have been preferred during

recognition. Therefore the mapping of /T/ → /s/ during forced-alignment did not introduce high confusion for /s/. Furthermore, the low percentage of the mapping rule indicates a large variety of possible mappings rather than a clear context dependent match to any other phoneme, in contrast to the Spanish phoneme /x/.
Next, forced-alignment of the Spanish phoneme HMMs on the Columbian data as a cross test is performed to evaluate the mapping rules. The cross-alignment also reveals possible context-dependent mappings of the Spanish phonemes /T/ and /J/ to alternative Columbian phonemes. Table (4) shows such alternative mappings for the phoneme /T/. The cross-test did not reveal other alternatives for phoneme /J/.

Table 4: *Pronunciation scores and phoneme mappings for the Spanish phoneme HMMs aligned to Columbian labeled speech*

Phonemes of Lattice	Scores and Mapping rules
/f/	1.03018
	26 % /T/ context: F - FV
/x/	1.1425
	25 % /T/ context: L - FV
/s/	1.02162
	23 % /T/ context: N - FV

Having identified context-dependent phoneme-mappings, the lexicon is updated using several different thresholds for the pronunciation scores and the percentages of the mapping rules. Adapted words in the lexicon thus received another possible pronunciation transcript, e.g.:

- anterior / a n t e r j o r /
- anterior / a n t e r i o r /
- agenda / a x e n d a /
- agenda / a s e n d a /
- español / e s p a J o l /
- español / e s p a n o l /
- mensajería / m e n s a x e r i a /
- mensajería / m e n s a s e r i a /

Table (5) below summarizes the recognition results for all tested lexica.

Table 5: *Recognition Results for Clustered HMMs and Adapted Lexicon*

HMMs	tested on Spanish	tested on Columbian
Pdf-based Clustered HMMs	79.7 %	96.4 %
Seg.-based Clustered HMMs	78.2 %	96.2 %

As table (5) indicates, the recognition result remained unchanged for all tested lexica. Although the word confusions changed, the overall recognition performance remained constant for all lexica variations. This proves that the obtained recognition results are due to other effects than pronunciation variants of the words. Nevertheless, no decrease of recognition performance was observed either. The observed recognition performance of the clustered HMMs thus depends on the acoustic property only.

6. Conclusions and Discussion

A combined approach to foreign-accented speech recognition was presented. Several distance measures were tested in regard to their quality for HMM clustering. It was found that pdf-based clustering based on the approximated divergence Gaussian distance measure obtained the best results. Although pdf-based clustering occurred across segments and phonemes, it appeared to lead to an improved robustness for both the target language and the simulated foreign-accented speech.
No further improvement of the recognition performance was achieved with pronunciation modeling. However, the speech recognition performance did not decrease either. This might be due to the fact, that the clustered HMMs as well as the introduced mappings for /J/ and /T/ according to section 5.1 already accounted well for the few pronunciation variants.

Further experiments will be run on real non-native accented speech, that is currently collected within the European funded project OrienTel, which aims to develop language resources for speech-based telephony applications across the area between Morocco and the Gulf States. The HMMs will then be trained on phonetically balanced sentences to avoid vocabulary dependent phoneme-mapping rules. The overall aim is to test the algorithms on native speech databases only, where the accented speech database serves as a mean of evaluation. The acoustic modeling part will thus receive a special focus in future.

7. References

[1] Goronzy S., Sahakyan M., and Wokurek W., "Is Non-Native Pronunciation Modeling Necessary?", Eurospeech, Aalborg, 2001.

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

[3] Koehler J., "Erstellung einer statistisch modellierten multilingualen Lautbibliothek fuer die Spracherkennung", Muenchen, Technische Universitaet, Dissertation, Shaker Verlag, Aachen, 1999.

[4] Eskenazi M., "Using a Computer in Foreign Language Pronunciation Training: What Advantages?", CALICO Journal, Vol. 16, No. 3, 1999, pp. 447-469.

[5] Witt S.M., Young S.J., "Phone-Level pronunciation scoring and assessment for interactive language learning", Speech Communication, Vol. 30, 2000, pp.95-108.

[6] Goronzy S., Kompe R., Rapp S., "Generating Non-Native Pronunciation Variants for Lexicon Adaptation", Eurospeech Satelite Workshop, Aalborg, 2001.

[7] Ward W., Krech H., Yu X., Herold K., Figgs G., Ikeno A., Jurafsky D., Byrne W., "Lexicon Adaptation for LVCSR: Speaker Idiosyncracies, Non-Native Speakers, and Pronunciation Choice", Pronuncation Modeling and Lexicon Adaptation for Spoken Language (PMLA) workshop, 2002.