Comparing Acoustic Model Adaption Methods for Non-Native Speech Recognition

H. Lang[1,2], M. Raab[1,3], R. Gruhn[1,2], W. Minker[2]
[1] *Harman Becker Automotive Systems, Speech Dialog Systems, Ulm, Germany*
[2] *Ulm University, Dept. of Information Technology, Ulm, Germany*
[3] *University of Erlangen Nuremberg, Chair of Pattern Recognition, Erlangen, Germany*
helmut.lang@uni-ulm.de

Introduction

In-car infotainment and navigation devices are typical examples where speech based interfaces are successfully applied. While classical applications are monolingual, such as voice commands or destination input, the trend goes towards multilingual applications such as music player control. However, the statistical framework that is currently employed for speech recognition has severe problems with non-native speech that has different characteristics than typical native training data. At the acoustic level those characteristics typical to non native speakers are insertions, deletions and substitutions of phonemes [1].

Previous work in non-native speech recognition has mostly only evaluated proposed techniques on limited test data due to non-availability of public databases with non-native speech from different accents. With the help of new non-native databases [2] that have recently become available more thorough evaluations become possible.

Experimental Setup

Our speech recognizer uses 11 Mel Frequency Cepstral Coefficients (MFCCs), where 9 neighboring vectors are stacked and the resulting multi-feature-vector is reduced in dimension by means of a linear discriminant analysis (LDA). The codebook has 1024 Gaussians utilizing full covariance matrices and is created with a variant of the LBG algorithm. To determine the most probable sequence of phonemes uttered according to the observed feature vectors semi-continuous Hidden Markov Models (HMMs) were employed, where each HMM either modeled a particular mono- or triphone. This means that the vectors contained in the codebook, in connection with a set of weight vectors assigned to each state of a given HMM, were utilized to calculate the likelihood that the corresponding state emitted an observed feature vector.

The baseline recognizer is trained on 200 hours of US Speecon data [3] and as our target was adaptation to non-nativeness the headset recordings were employed, in order to ensure similar channel conditions between the initial training data and the adaptation and test material.

To decide if monophone- or triphone-based recognizers are better suitable for non-native speech recognition all experiments were performed separately for mono- and triphones.

Non-Native Databases

For adaptation and testing we utilized the following two different non-native speech databases. To avoid channel adaptation we used the "clean" signal partitions of both corpora.

HIWIRE

The HIWIRE database [4] contains English from French, Italian, Greek and Spanish speakers. The exact distribution of speakers is listed in Table 1. The vocabulary of the HIWIRE corpus is quite task specific and covers only 133 words from a military aeronautic domain. Accordingly the number of different tri- and biphones found in the material is quite low.

Country	*# Speakers*	*# Utterances*
France	31	3100
Greece	20	2000
Italy	20	2000
Spain	10	999
Total	**81**	**8099**

Table 1: Distribution of speakers HIWIRE

There is a defined standard division into adaptation and test sets and 50 utterances of each speaker comprehend the test and adaptation set respectively. We adopted this partition for the experiments performed in order to keep reproducibility. Since speech of each test speaker is also used for adaptation, next to adapting to non-nativeness, speaker adaptation is also an issue that has to be considered interpreting the results in the following sections.

The HIWIRE database includes a constrained grammar in EBNF that was used in the test runs. Nevertheless we also performed each test with a word-loop based grammar to achieve comparability with the results attained on the tests with ISLE.

ISLE

The ISLE database [5] comprehends speech from 23 Italians and 23 Germans speaking English at different proficiency levels. The corpus is subdivided into different parts labeled from *A* to *G*. *A* to *C* comprise read speech from a non-fictional text describing the ascent of Mount Everest. Each speaker read 82 sentences of that text,

adding up to about 1300 words per speaker. The purpose of this partition was to cover a great range of vocabulary and 410 different words occur in this section. Blocks *D* to *G* focus on problem phones (as identified by language teachers), weak forms, tricky stress patterns and difficult consonant clusters.

Data was recorded in non-noisy environments using high-quality headset microphones and audio quality is thus comparable to the noise free part of the HIWIRE database.

The ISLE corpus lacks a predefined separation into training and test sets. Thus three different sets were defined:

D31-80 test set This set was used for testing after retraining on HIWIRE. It contains the utterances labeled blockd01_31 to blockd01_80 of every speaker separated by native language and consists of rather short utterances.

Training set To profit from the wide phonetic coverage of block *A* to *D* as well as from the special material from *E* to *G* the training set comprehends roughly 80 % of the material from all blocks.

Test set The remaining 20 % of utterances were used as a second test set for ISLE after adaptation to ISLE. The exact division into training and test set is shown in Table 2.

Block	***German***			***Italian***		
	train	*test*	***tot.***	*train*	*test*	***tot.***
A	483	138	621	483	138	621
B	598	161	759	598	161	759
C	391	113	504	391	115	506
D	1466	368	1834	1469	368	1837
E	1143	298	1441	899	233	1132
F	158	46	204	160	46	206
G	207	46	253	207	46	253
tot.	**4446**	**1170**	**5616**	**4207**	**1107**	**5314**

Table 2: Utterance division, training and test set ISLE

The grammars used during testing with ISLE were a word-loop grammar containing each word appearing in the respective test set and a sentence based grammar that consisted of each sentence occurring.

Results

Retraining

Retraining involved a maximum of 8 additional expectation-maximization runs with non-native training data on the baseline system that was previously trained on native speech. During retraining the weight vectors and transition probabilities of the given models were re-estimated.

HIWIRE

The results attained after retraining and testing with the respective sets of HIWIRE are depicted in Figure 1.

There is a great increase of performance compared to

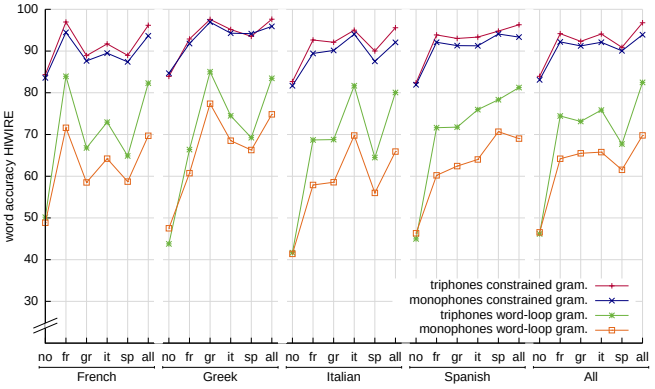

Figure 1: Results for retraining and testing on HIWIRE. The abbreviations below the x-axis identify the accent used for retraining and the languages beneath indicate the accent used for testing.

the baseline system ("no" for no retraining in Figure 1 indicates the baseline) after retraining regardless of accent retrained on. Every accent has its peak, when retraining was performed with the same accent. The only exception is Spanish in the case of triphones. As Spanish is the accent with the least amount of training data this is a hint that the amount of available training data is more crucial for triphone based adaptation than for monophones and in the case of monophones less data suffices to adequately re-estimate the baseline models, than is needed to re-estimate sophisticated triphone models. Contrary to [6], whose results for several accents show better performance in the case of monophones, no clear tendency towards mono- or triphones can be determined for the baseline system. An explanation for the difference between [6] and our results may be a different proficency level of speakers and the unlike task complexity. While HIWIRE covers rather short utterances with a vocabulary containing quite common words, the utterances of the test sets in [6] have a significantly higher complexity. According to this the number of pronunciation errors made in [6] may be expected to be much higher than for HIWIRE and those errors have greater influence on triphone based acoustic models than on monophone based ones.

After retraining triphones perform better than monophones. Hence we conclude that triphones are better suitable for non-native speech recognition but to circumvent the problem of data sparseness, which is always an issue in the context of non-native adaptation, monophones should be preferred.

Four possible reasons for the severe improvements obtained with this procedure come into mind:

- There are characteristics common to non-native speech, especially to the accents covered by HIWIRE. This could be a difference in fluency or speed that was not covered by the native training data. This common properties influence the model parameters in a way that improves non-native recognition regardless of the actual accent.

- In contrast to English spoken by natives, the HIWIRE accents share properties, like a different phone realizations or phoneme substitutions etc. that cause the recognition rates to improve for this particular and other similar accents.
- Due to different channel conditions between SPEECON and HIWIRE there is a great amount of channel adaptation and this channel adaptation is the reason for the observed improvements.
- Because of the limited vocabulary coverage of HIWIRE overfitting could be the reason for the great improvements.

To verify which of the explanations holds true we tested the models retrained on HIWIRE on the D31-80 test set of ISLE. The results of this experiment are shown in Figure 2.

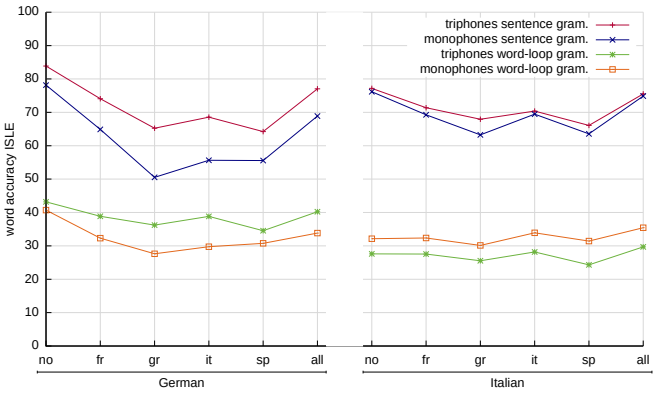

Figure 2: Results for retraining on HIWIRE and testing on ISLE. The labeling follows the same scheme as that of Figure 1.

Performance decreases for both accents of ISLE. So from the previously stated explanations only channel adaptation and overfitting still come into consideration. The only exception, where there is a slight increase in performance after retraining is the monophone based recognizer with word-loop grammar in case of Italian after retraining on Italian accent. This is a first hint that overfitting causes the observed decrease in performance. In order to further proof this assumption we decided to perform retraining on ISLE and test the thus re-estimated models on HIWIRE. If adverse channel conditions between HIWIRE and ISLE caused the decrease in performance after retraining on HIWIRE, performance on HIWIRE should, compared to the baseline system, decrease for a similar amount after retraining on ISLE.

ISLE

The results of retraining on ISLE and testing on HIWIRE are shown in Figure 3. For monophones there is a consistent improvement, but this holds not true for triphones. What shows again that monophones should be preferred for non-native adaptation. The results verify that the reduction of performance on ISLE after retraining on HIWIRE was due to overfitting and hence in order to adapt to non-native speech in general phonetically rich databases have to be employed.

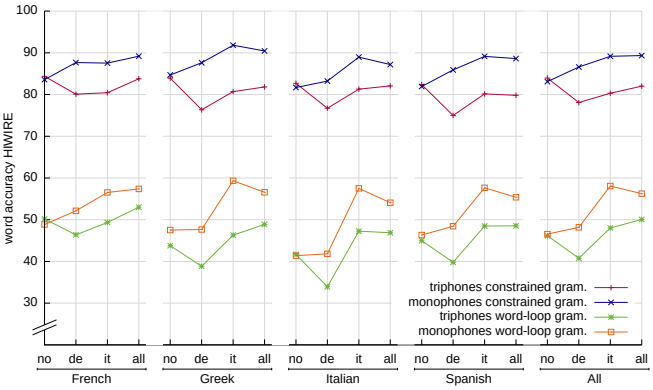

Figure 3: Results for retraining on ISLE and testing on HIWIRE. The labeling follows the same scheme as that of Figure 1.

In contrast to HIWIRE, where training on all accents did not cause a significant reduction of performance compared to the models trained on single accents, German seems to reduce the positive influence of Italian and thus represses the performance gain achieved with Italian. The only exclusion is French which benefits most from training on both accents. An interpretation of this results could be that there is a closer relation between Spanish, Greek and Italian accented English, than between Italian and French and there is a stronger relation between French and German than between German and the other accents. Conforming to this hypothesis and to the observed negative effects of German on Spanish and Italian accents are the results of [7] where performance at German accented speech decreased significantly when training was performed on all accents (Danish, German, British, **Spanish**, **Italian** and Portuguese) compared to solely training on German.

MAP Adaptation

In a separate experiment we adapted the means of the Gaussians comprehending the codebook using a maximum a posteriori (MAP) estimation approach, corresponding to the training data of the respective training set from the ISLE corpus.

The transcription of the utterances was known to the system beforehand, thus the adaptation was implemented in a supervised manner.

As the target was adaptation to non-nativeness and not to particular speakers, MAP was performed on all speakers of a particular accent, as well as altogether on all speakers of both accents.

The results attained are shown in Figure 4. Similar to retraining, there are consistent improvements in the case of monophones, but triphones sometimes perform worse than the baseline system.

Conclusion

Comparing the results attained with retraining and those after MAP adaptation of Gaussian means (see Figure 5) the superiority of retraining becomes evident. Similar

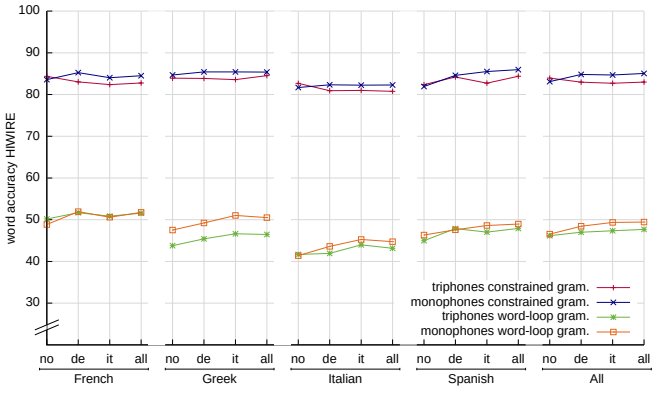

Figure 4: Results for MAP adaptation of Gaussian means according to ISLE and testing on HIWIRE. The labeling follows the same scheme as that of Figure 1.

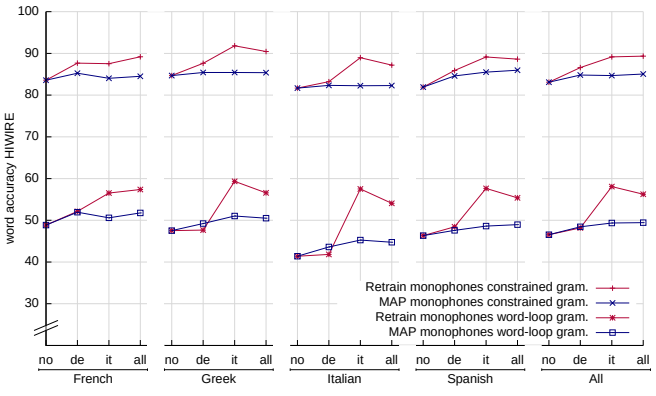

Figure 5: Results for MAP adaptation of Gaussian means compared to the results of retraining on ISLE and testing on HIWIRE. The labeling follows the same scheme as that of Figure 1.

results are reported in [8]

The only case, where MAP performed better than retraining, is for the word-loop grammar with adaptation to German accent and testing on Greek and Italian. So it may be stated that for interrelated accents retraining performs significantly better than MAP. This allows to draw the conclusion that MAP adaptation of Gaussian means is less accent specific than retraining and thus retraining is the better choice for non-native adaptation.

When the triphones appearing in the test were re-estimated on a sufficient amount of training data triphones showed better performance than monophones. This was the case, when retraining as well as testing was performed on HIWIRE and with this setup we were able to achieve an average WA of 96.6 % after retraining, compared to an average WA of 83.3 % for the baseline system solely trained on native speech.

But even with less adequate training data, great improvements could be achieved. After retraining on the Italian part of ISLE there was a relative average increase in WA of 21.7 % on the accents of HIWIRE (excluding Italian).

Hence our results clearly verify the effectiveness of retraining for non-native speech recognition.

References

[1] D. van Compernolle, "Recognizing Speech of Goats, Wolves, Sheep and ... Non-Natives", in *Speech Communication*, **35**, pp. 71–79, Elsevier Science Publishers B. V., Amsterdam, 2001.

[2] M. Raab, R. Gruhn, and E. Nöth, "Non-Native Speech Databases", in *Proceedings ASRU*, pp. 413–418, 2007.

[3] D. Iskra, B. Grosskopf, K. Marasek, H. van den Heuvel, F. Diehl, and A. Kiessling, "Speecon - Speech Databases for Consumer Devices: Database Specification and Validation", in *Proceedings LREC*, pp. 329–333, 2002.

[4] J. C. Segura et al., "The HIWIRE Database, a Noisy and Non-Native English Speech Corpus for Cockpit Communication", 2007. URL: `http://www.hiwire.org/`.

[5] W. Menzel et al., "The ISLE Corpus of Non-Native Spoken English", in *Proceedings LREC*, pp. 957–963, 2000.

[6] T. Cincarek, R. Gruhn, and S. Nakamura, "Speech recognition for multiple non-native accent groups with speaker-group-dependent acoustic models", in *Proceedings ICSLP*, pp. 1509–1512, 2004.

[7] C. Teixeira, I. Trancoso, and A. Serralheiro, "Recognition of Non-Native Accents", in *Proceedings Eurospeech*, pp. 2375–2378, 1997.

[8] L. Mayfield Tomokiyo and A. Waibel, "Adaptation Methods for Non-Native Speech", in *Proceedings of Multilinguality in Spoken Language Processing*, 2001.