Multilingual Speech Interfaces for Resource-constrained Dialog Systems

Martin Raab[1,3], Rainer Gruhn[2], Elmar Nöth[3]

[1] Nuance Communications, Speech Technologies, Ulm, Germany
martin.raab@informatik.uni-erlangen.de,
WWW home page: http://www.nuance.com
[2] University of Ulm, Dept. of Information Technology, Ulm, Germany
[3] University of Erlangen, Chair of Pattern Recognition, Erlangen, Germany

Abstract. This paper presents results of a combination of two algorithms for the fast and effective support of multilingual speech in a dialog system. Previously only results of the individual algorithms were published. The Multilingual Weighted Codebook algorithm generates sets of Gaussians (codebooks) that cover multiple languages well, especially it was designed to provide optimal performance for one designated main language in the system. It makes sense to operate with such a main language, as human users also prefer the use of their native language when it is possible. The second algorithm projects a Gaussian mixture distribution to a new set of Gaussians. With this algorithm, time and cost intensive iterative reestimations of a distribution for one speech sound are avoided. Together both algorithms allow a much faster provision of acoustic models for multilingual speech recognition with semi-continuous Hidden Markov Models that are based on one single global codebook.

1 Introduction

It is a frequent problem that spoken dialog systems have to handle speech from multiple languages. One example are speech operated navigation systems that should allow destination input in many countries. In the near future companies are also preparing for the rollout of music players that are operated via speech. A major problem in these examples is that the dialog systems operate with constrained resources, and current speech recognition technology typically requires more resources for each language that has to be covered.

An additional problem in the above examples is that the users are not perfect in the pronunciation of the foreign city names or the foreign song titles. They utter the utterances with non-native accent. This accent is amongst others influenced by the native language of the user, and referred to as main language in the rest of this work.

A common approach to reduce the resource need for each additional language in the speech recognition system is to model similar sounds in different languages with the same models. In this approach, phonemes from different languages can share one acoustic model when they have the same IPA (International

Phonetic Alphabet, [Ladefoged, 1990]) symbol. Examples are [Weng et al., 1997] [Koehler, 2001] [Schultz and Waibel, 2001] [Niesler, 2006]. The sharing can also be based on acoustic model similarity determined by a distance measure. For example, [Koehler, 2001] [Dalsgaard et al., 1998] measure the log-likelihood difference on development data to determine the similarity of phonemes, as motivated by [Juang and Rabiner, 1985].

The advantages of all these approaches are that they cover many languages with much less parameters than a combination of all the monolingual recognizers. Thus they are very appropriate in all cases where one really needs all languages equally. However, in the examples mentioned at the beginning, the languages are not of equal importance. There is one device, which is typically owned by one user with one native language and that language is more important for the system than the other languages as the user usually utters commands, spellings and digit sequences in that language. Hence it is vital for a commercial system to recognize this main language with maximum performance.

Motivated by [Koehler, 2001] who showed that parameter sharing achieves better performance than sharing of full Hidden Markov Models (HMM), we developed the idea of Multilingual Weighted Codebooks (MWC) for a semi-continuous HMM [Raab et al., 2008a] [Raab et al., 2008b] . The advantage of semi-continuous HMMs is that all HMMs use one codebook with a small number of Gaussians compared to the number of Gaussians that are used in continuous HMMs. However, the performance of a semi-continuous HMM with a suboptimal codebook from a foreign language is significantly reduced [Raab et al., 2008b]. The MWCs solve this problem by adding the most different Gaussians from the foreign language codebooks to the main language codebook. The MWCs were shown to help for native and non-native speech [Raab et al., 2008b].

The problem is that MWCs depend on the actual language combination. This leads to an unacceptably high training effort for more than a couple of languages, as the acoustic models have to be retrained for each MWC, thus for every language combination. A projection of a GMM that is defined on one set of Gaussians to another codebook solves this problem, as it can project the once trained Gaussian mixture distribution to the Gaussians that are currently available. Thus, once a monolingual acoustic model exists for each language, the Gaussians can be exchanged without retraining with this projection.

At first, we experimented with mathematically optimal projections with respect to the L2 distance [Raab et al., 2009b]. However, we later showed that similar performance can be achieved with an approximated projection that considers only the expected values of Gaussians and HMM states [Raab et al., 2009a]. This projection has the additional advantage that it runs in almost no time (fractions of a second) for the projection of one language to a new codebook.

In this paper, we now present the combination of the MWC and the projection algorithm. Together they form a scalable mechanism that allows to rapidly provide speech interfaces for new language combinations, with similar performance than our baseline system with one monolingual codebook. At the same

time, the immense additional effort for the provision of systems has significantly been reduced.

The remainder of this paper is organized as follows. In the next section we present our multilingual baseline system. Section 3 describes the two algorithms and their combination. In Section 4 the experimental setup is described. Section 5 presents the experimental results. Finally, a conclusion is drawn in Section 6.

2 Benchmark and Baseline System

The monolingual systems are benchmark systems, as they are built to obtain maximum performance on one language. The problem with running monolingual recognizers in parallel is that the number of resources needed increases linearly with the number of languages, as there is no parameter tying at all between the languages.

Our baseline system reduces the resource need by keeping only the codebook of the main language. Then all additional language HMMs are added to the acoustic model. However, they have to be retrained, as they were initially trained on another set of Gaussians. Through the application of only one codebook the number of Gaussians remains the same, thus not requiring any more resources. Only the evaluation of the additional HMMs needs some more resources, but not too much as an HMM evaluation reduces to a multiplication of two vectors in a semi-continuous HMM system.

3 Proposed System

3.1 Multilingual Weighted Codebook

To improve the performance on the additional languages, the monolingual codebook in our baseline system is replaced by a MWC. The MWC is basically the main language codebook plus some additional Gaussians. Figure 1 depicts an example for the extension of a codebook to cover an additional language. From left to right one iteration of the generation of MWCs is represented.

The picture to the left shows the initial situation. The Xs are mean vectors from the main language codebook, and the area that is roughly covered by them is indicated by the dotted line. Additionally, the numbered Os are mean vectors from the second language codebook. Supposing that both Xs and Os are optimal for the language they were created for, it is clear that the second language contains sound patterns that are not typical for the first language (Os 1,2 and 3).

The middle picture shows the distance calculation. For each of the second language codebook vectors, the nearest neighbor among the main language Gaussians is determined. These nearest neighbor connections are indicated by the dotted lines. Our previous experiments showed that using the Mahalanobis distance produces the best results [Raab et al., 2008a].

The right picture presents the outcome of one iteration. From each of the nearest neighbor connections, the largest one (O number 2) was chosen as this is obviously the mean vector which causes the largest vector quantization error. Thus, the Gaussian O number 2 was added to the main language codebook.

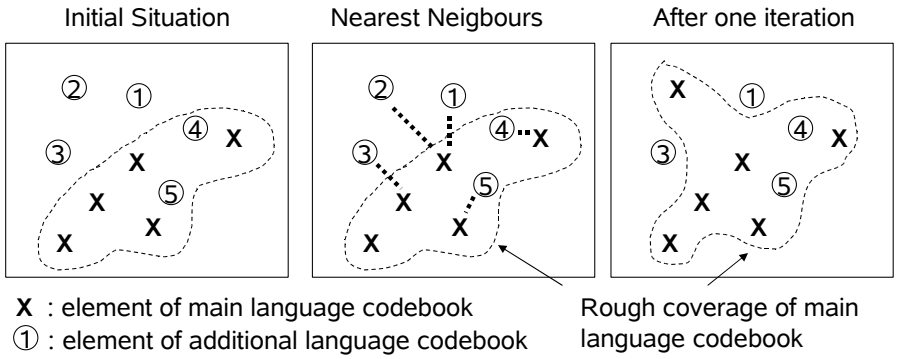

Fig. 1. Basic idea of Multilingual Weighted Codebooks

3.2 Projection of a GMM

The goal of this algorithm is to project a Gaussian Mixture Model (GMM) distribution that is modeled with a set of Gaussians to another set of Gaussians. In [Raab et al., 2009b] we presented mathematically optimal projections that minimize the L2 distance between the reference distribution and the estimated distribution. The reference distribution is a Baum-Welch trained state emission probability defined on one codebook. However, in [Raab et al., 2009a] we could show that similar performance can be achieved with approximated projections. In this paper, we only introduce the best of the approximated projections.

This projection is based on a combination of the best match of expected values of Gaussians and of HMM states. This projection maps all HMMs of all L languages to one fixed set of N Gaussians (=Recognition Codebook, RC). Each Gaussian $\mathcal{N}$ is represented by its mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

The Gaussian map (map_G) is based on the smallest Mahalanobis distance (Gaussian Distance D_G).

$$map_G(\mathcal{N}^i_{MC^l}) = \mathcal{N}^j_{RC},\ 0 \leq i < M^l, 0 \leq j < N, 0 \leq l < L$$
$$j = \arg\min_k D_G(\boldsymbol{\mu}^i_{MC^l}, \boldsymbol{\mu}^k_{RC}, \boldsymbol{\Sigma}^i_{MC^l}) \tag{1}$$

The state map $\mathbf{map_S}$ is based on the minimum Mahalanobis distance (D_S) between the expected values of their probability distributions. The covariance which is needed for the Mahalanobis distance is a global diagonal covariance $\boldsymbol{\Sigma}_{All}$

estimated on all training samples. With D_S we define our state based mapping for each state $\mathbf{s}$

$$\begin{aligned} \mathbf{map_S}(\mathbf{s}_l^i) &= \mathbf{s}_{RS}^j, 0 \leq i < S_l, 0 \leq j < RS, 0 \leq l < L \\ j &= \arg\min_k D_S(E(\mathbf{s}_l^i), E(\mathbf{s}_{RS}^k), \boldsymbol{\Sigma}_{All}) \end{aligned} \tag{2}$$

with the main language states RS (=Recognition States).

The combined map $\mathbf{map_{G+S}}$ of the previous two mappings is defined as

$$\begin{aligned} &\mathbf{map_{G+S}}(\mathbf{s}_l^i) = \\ &\gamma_{G+S}\, \mathbf{map_S}(\mathbf{s}_l^i) + (1-\gamma_{G+S}) \begin{pmatrix} w_{\mathbf{s}_l^i}^1 map_G(\mathcal{N}_{MC^l}^1) \\ w_{\mathbf{s}_l^i}^2 map_G(\mathcal{N}_{MC^l}^2) \\ \vdots \\ w_{\mathbf{s}_l^i}^{M^l} map_G(\mathcal{N}_{MC^l}^{M^l}) \end{pmatrix} \\ &0 \leq l < L, 0 \leq i < S_l \end{aligned} \tag{3}$$

with the combination weight γ_{G+S} and the weight w for each individual Gaussian. In this work γ_{G+S} was set to 0.5, and slowly decreased when more Gaussians were added. This was done to account for the fact that the more Gaussians are in the MWC, the better the Gaussian distance, as the MWC algorithm exactly adds the Gaussian that have no match in the other codebook. This projection projects a language with 1800 three state HMM models to a new codebook in 15 seconds, or in 0.3 seconds if some elements in the distance calculation like the distances of all states to each other are precomputed [Raab et al., 2009a].

3.3 Combined System

All the systems without projections require significantly more resources and/or effort. The resource need of the benchmark system increases linearly with the number of languages, the training effort for the baseline system increases with the number of languages squared. The MWC systems keep the low resource need of the baseline system with increased performance, but increase the training effort exponentially. However, the projections alone are maybe also not desirable, as they perform worse than every other system

The question of interest is how well a combination of the MWC algorithm and the projections works. This means that first an appropriate MWC is built for the current task, and then the required HMMs are projected to the task. This removes the additional training effort almost completely, and should still give good performance due to the well matching codebook.

4 Experimental Setup

Our semi-continuous speech recognizer uses 11 MFCCs with their first and second derivatives per frame. Monolingual recognizers for English, French, German,

Spanish and Italian are trained on 200 hours of Speecon data [Iskra et al., 2002] with 1024 Gaussians in the codebook. The HMMs are context dependent and the codebook for each language is different. Table 1 describes the native test sets

Table 1. Descriptions of the native test set for each language

Testset	Language	Words	Vocabulary
GE_City	German	2005	2498
US_City	English	852	500
IT_City	Italian	2000	2000
FR_City	French	3308	2000
SP_City	Spanish	5143	3672

Table 2. Description of the non-native test sets

Testset	Accent	Words	Vocabulary
Hiwire_FR	French	5192	140
Hiwire_SP	Spanish	1759	140
Hiwire_IT	Italian	3482	140
IFS_MP3	German	831	63

and Table 2 the non-native test sets. The native tests are city names from an in-house database. The Word Accuracy (WA) differences that the results show between the languages are due to different noise conditions in the different tests.

The first three non-native test sets contain command and control utterances in accented English from the Hiwire database [Segura et al., 2007]. The Hiwire database was identified as the most appropriate existing and available database after a review of existing non-native databases [Raab et al., 2007]. To indicate that the language information in the test name only specifies the accent, the language information is given after the test name. The last non-native accent test was collected specifically for this work and contains Italian, French and Spanish song titles names spoken by Germans.

5 Experiments

5.1 Native Speech

In this section native speech of five languages is evaluated, first with German as main language, then in a second set of experiments with English as main language. The codebook of the baseline system only contains Gaussians from the main language, and the MWC systems always contain all Gaussians from the main language. However, the MWC systems also add some additional Gaussians for the language that is tested.

Figure 2 shows the Word Accuracies for city name tests. The x-axis indicates both which test is performed as well as the size of the MWC that is applied. The curves show that in the German test all three system perform equal, the MWC based on the German codebook generated through the projections, the baseline and the benchmark system. This is actually a strength of the MWC approach, as commonly parameter sharing methods for multilingual speech recognition have slightly negative impacts for all languages. This is not the case for the main language in the MWC system.

The picture is different for the additional languages. Here it becomes evident that benchmark systems are better than other systems with parameter tying across languages. In the case that no additional Gaussians were added to the codebook, the baseline system also outperforms the projection system. However, when the two algorithms are really combined, and Gaussians are added before the projection is executed, the combined system outperforms the baseline system. This out performance requires some additional resources at runtime for the additional Gaussians, but a provision of the combined system requires less effort than a provision of the baseline system. The set of results that is missing in this figure are the results of the MWC with a standard retraining. It is quite certain that this would give better performance than the MWC performance shown, but only for the experiments in this figure the same effort as building 20 monolingual recognizers would need to be done. In general it is unrealistic to provide Baum-Welch trained MWC systems for many languages and all their combinations today.

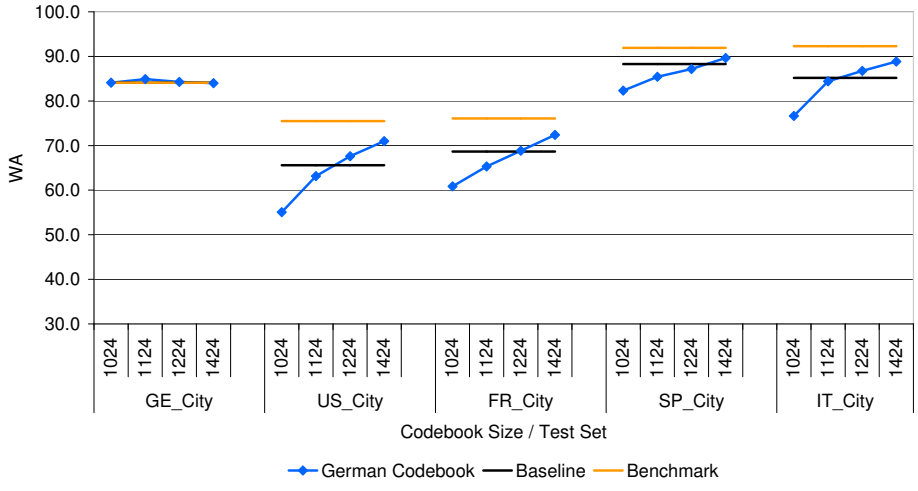

Fig. 2. Performance of the scalable architecture on native speech of different languages with German as main language

To verify that these results are not just due to some peculiarities of the German codebook, the same experiments are redone with English as main language in Figure 3. The main difference is that now nothing happens for the English system, and it always achieves the benchmark performance. For the other tests, the trends remain the same, the baseline system is better than projections alone and the combined system can outperform the baseline when more Gaussians are added. However, in general the performance is a little worse for the additional languages as compared to the results with German as main language. We attribute this to the fact that German has more phonemes than English, and that there are thus more sounds that not well covered with an English codebook.

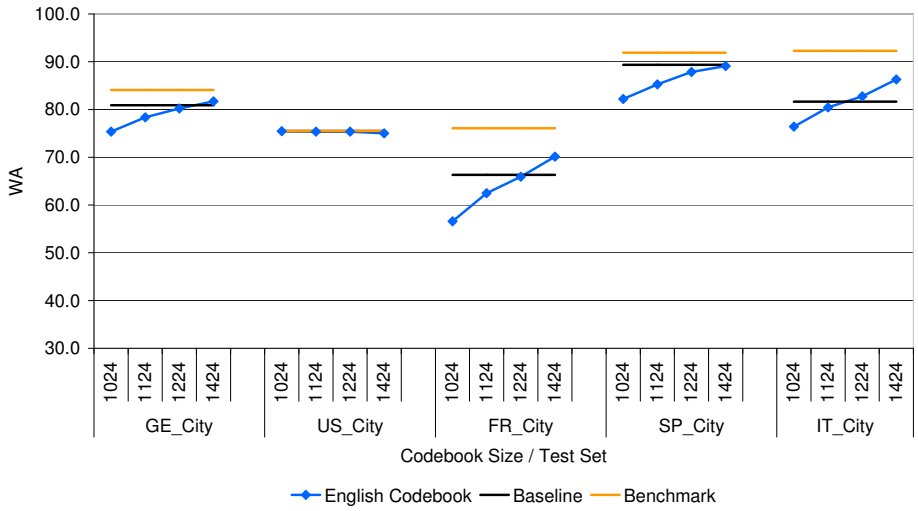

Fig. 3. Performance of the scalable architecture on native speech of different languages with English as main language

5.2 Non-native Speech

This section evaluate the performance on non-native accents. The first three tests are non-native English by Spanish, French and Italian speakers. The last test contains song titles by Germans. The last test is also strongly accented speech, as not all speakers were familiar with the language from which the name originated. Figure 4 shows the performance on these four tests. A major difference to the previous charts is that this time the main language was different for each test set, as it was always set to the native language of the speakers. It is also the case that no benchmark performance for the last test is given, as no monolingual system can recognize speech from three languages.

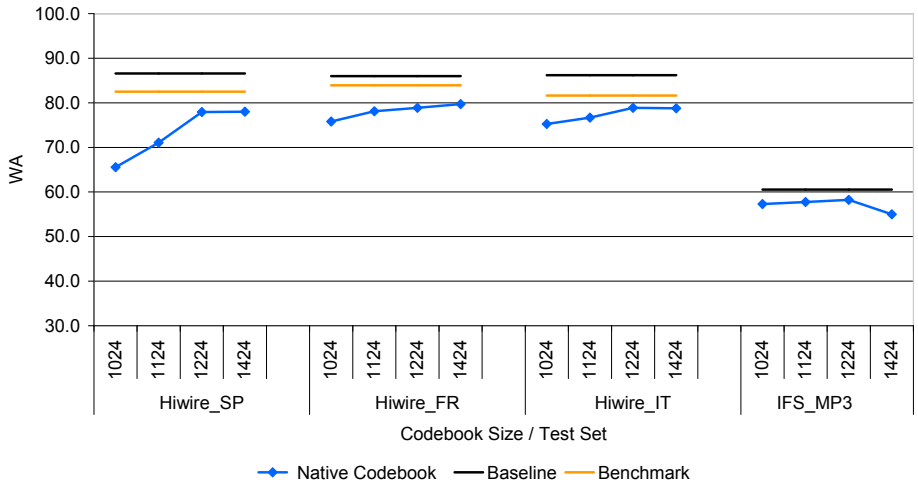

Fig. 4. Performance of the scalable architecture on non-native English with, the main language is always the mother tongue of the speakers

That the application of the codebook from the native language of the users is a good choice is supported by the fact that this time the baseline systems outperform the benchmark systems. This means that our systems are better in the recognition of accented speech if they apply a codebook of the native language of the speakers.

In the case of the speakers which were familiar with the spoken language, the proposed system benefits from the addition of the additional Gaussians, and comes close to the performance of the benchmark system when Gaussians are added. In the case of the less familiar speakers in the IFS_MP3 test, however, the addition of Gaussians did not help. The performance on this test is also worse than for the other tests, which might indicate that their speech is just too different from the speech of native speakers of Spanish, French and Italian.

6 Conclusion

This paper has combined two of our previous algorithms for the support of multilingual speech interfaces in resource-constrained dialog system. The first algorithm builds Multilingual Weighted Codebooks that allow the improved recognition of many languages with semi-continuous HMMs. The second algorithm allows the faster generation of a system with parameter sharing between languages. The results in this paper show the fruitful combination of these two algorithms that allows to rapidly generate new acoustic models with high performance for the required languages.

References

[Dalsgaard et al., 1998] Dalsgaard, P., Andersen, O., and Barry, W. (1998). Cross-language merged speech units and their descriptive phonetic correlates. In *Proc. ICSLP*, page no pagination, Sydney, Australia.

[Iskra et al., 2002] Iskra, D., Grosskopf, B., Marasek, K., van den Huevel, H., Diehl, F., and Kiessling, A. (2002). Speecon - speech databases for consumer devices: database specification and validation. In *Proc. LREC*, pages 329–333, Las Palmas de Gran Canaria, Spain.

[Juang and Rabiner, 1985] Juang, B. H. and Rabiner, L. R. (1985). A probabilistic distance measure for Hidden Markov Models. *AT&T Technical Journal*, 64(2):391–408.

[Koehler, 2001] Koehler, J. (2001). Multilingual phone models for vocabulary-independent speech recognition tasks. *Speech Communication Journal*, 35(1-2):21–30.

[Ladefoged, 1990] Ladefoged, P. (1990). The revised international phonetic alphabet. *Language*, 66(3):550–552.

[Niesler, 2006] Niesler, T. (2006). Language-dependent state clustering for multilingual speech recognition in Afrikaans, South African English, Xhosa and Zulu. In *Proc. ITRW*, Stellenbosch, South Africa.

[Raab et al., 2009a] Raab, M., Aradilla, G., Gruhn, R., and Nöth, E. (2009a). Online generation of acoustic models for multilingual speech recognition. In *Proc. Interspeech*, pages 2999–3002, Brighton, UK.

[Raab et al., 2007] Raab, M., Gruhn, R., and Nöth, E. (2007). Non-native speech databases. In *Proc. ASRU*, pages 413–418, Kyoto, Japan.

[Raab et al., 2008a] Raab, M., Gruhn, R., and Nöth, E. (2008a). Multilingual weighted codebooks. In *Proc. ICASSP*, pages 4257–4260, Las Vegas, USA.

[Raab et al., 2008b] Raab, M., Gruhn, R., and Nöth, E. (2008b). Multilingual weighted codebooks for non-native speech recognition. In *Proc. TSD*, pages 485–492, Brno, Czech Republic.

[Raab et al., 2009b] Raab, M., Schreiner, O., Herbig, T., Gruhn, R., and Nöth, E. (2009b). Optimal projections between Gaussian mixture feature spaces for multilingual speech recognition. In *Proc. DAGA*, pages 411–414, Rotterdam, Netherlands.

[Schultz and Waibel, 2001] Schultz, T. and Waibel, A. (2001). Language-independent and language-adaptive acoustic modeling for speech recognition. *Speech Communication*, 35:31–51.

[Segura et al., 2007] Segura, J., Ehrette, T., Potamianos, A., Fohr, D., Illina, I., Breton, P.-A., Clot, V., Gemello, R., Matassoni, M., and Maragos, P. (2007). The HIWIRE database, a noisy and non-native English speech corpus for cockpit communication. http://www.hiwire.org/.

[Weng et al., 1997] Weng, F., Bratt, H., Neumeyer, L., and Stolcke, A. (1997). A study of multilingual speech recognition. In *Proc. Eurospeech*, pages 359–362, Rhodes, Greece.