# IN–SERVICE ADAPTATION OF MULTILINGUAL HIDDEN–MARKOV–MODELS

*Udo Bub*[1,2], *Joachim Köhler*[1], *and Bojan Imperl*[3]

[1]Corporate Research and Development, Siemens AG, Munich, Germany
[2]Inst. for Human–Machine–Communication, Munich Univ. of Technology (TUM), Munich, Germany
[3]Faculty of Electrical Engineering and Computer Science, Univ. of Maribor, Maribor, Slovenia

## ABSTRACT

In this paper we report on advances regarding our approach to porting an automatic speech recognition system to a new target task. In case there is not enough acoustic data available to allow for thorough estimation of HMM parameters it is impossible to train an appropriate model. The basic idea to overcome this problem is to create a task independent seed model that can cope with all tasks equally well. However, the performance of such generalist model is of course lower than the performance of task dependent models (if these were available). So, the seed model is gradually enhanced by using its own recognition results for incremental online task adaptation. Here, we use a multilingual romanic/germanic seed model for a slavic target task. In tests on Slovene digits multilingual modeling yields the best recognition accuracy compared to other language dependent models. Applying unsupervised online task adaptation we observe a remarkable boost of recognition performance.

## 1. INTRODUCTION

The research described in this paper is part of our ongoing efforts towards flexible automatic speech recognition (ASR) systems that offer a maximum recognition performance for changing channels, speakers, and tasks (including languages).
It is well–known that ASR systems that have been designed for general use are being outperformed during a special application by a specialist system that has been designed for this one assignment only (e.g. [1]). A problem occurs when training of specialist models is impossible either

- because the final task is unknown during training time or
- because there is not enough task dependent data available for complete re–training.

This problem can be solved by creating a phonetically balanced Hidden–Markov–Model that can cope initially with all possible recognition tasks and then adapt it automatically online to a specific task given by the final user in order to improve the overall performance [2].

We introduce the new idea of adaptive online language transfer: First we generate a language independent, phonetically balanced multilinguist HMM from English, German, and Spanish [3]. Then this model is used as initial base for an examplary task of Slovene digits. Using task adaptation, the system learns gradually from the occurring utterances and updates its own HMM parameters. This offers a way to transfer a system to a new language if no data is available for the new target task/language at system development time.

The paper is structured as follows: In section 2 we describe the theoretic background of general task adaptation and discuss issues relevant to the final implementation. Then we focus on the procedures that are involved with building multilingual models that constitute the seed HMMs for adaptation to the final task. Consequently, we give next a brief introduction to the steps that are necessary for multilingual adaptation itself. In section 5 we first demonstrate the improvements that our task adaptation techniques yield for the case of changing dictionaries for a German recognizer. Using the knowledge derived from these experiments we finally carry out multilingual adaptation.

## 2. ONLINE TASK ADAPTATION

### 2.1. Adaptation Formulae

We assume that the relevant differences between tasks affect mainly the parameters of the HMM probability density functions (pdfs), or more specifically the location of their means in acoustic space. Due to simplicity we are going to discuss the formulae for 1–dimensional space. The practical relevant case of higher dimensions can be derived easily from the given formulations.
A Bayesian update of the mean $\mu$ of a normal density is achieved according to [4] by

$$\mu_n = \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}\mu_0 + \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2}m_n, \qquad (1)$$

where $m_n = \frac{1}{n}\sum_{k=1}^{n} x_k$ is the mean value of $n$ (with $n \geq 1$) new observations. $\mu_0$ and $\sigma_0^2$ are mean and variance of the previously trained task independent model, whereas $\sigma^2$ is the variance of the new task dependent adaptation data. Eliminating $\mu_0$ by counting in the explicit formulation of $\mu_{n-1}$ results in a recursive updating formula:

$$\mu_n = (1-\alpha_n)\mu_{n-1} + \alpha_n x_n, \tag{2}$$

with

$$\alpha_n = \frac{1}{n + \frac{\sigma^2}{\sigma_0^2}}. \tag{3}$$

Some algebra yields a new formulation of (1):

$$\mu_n = \mu_0 g_n(n) + \sum_{i=0}^{n-1} x_{n-i}\alpha_{n-i} g_n(i), \tag{4}$$

where we define $g_n(\cdot)$ as follows:

$$g_n(k) = \begin{cases} 1 & \text{if } k < 1 \\ \prod_{j=0}^{k-1}(1-\alpha_{n-j}) & \text{if } k \geq 1. \end{cases} \tag{5}$$

**Bayesian Adaptation**

Now, we want to examine the contribution of past observation vectors to an update [5]. For this reason we derive a weighting function that calculates the weight of every past observation. Imagine we have just finished the $n$th adaptation pass for one specific mean. From (4) we can easily determine by inspection the weight of the $k$th (starting with 0) past observation:

$$w_n(k) = \begin{cases} \alpha_{n-k} g_n(k) & \text{if } 0 \leq k \leq n-1 \\ 0 & \text{else.} \end{cases} \tag{6}$$

In the same way the weight for the initial seed mean $\mu_0$ can be read as $g_n(n)$. As an example we determine now the weights after 30 adaptation passes. In our recognition system we do not model variances explicitely and replace them with a grand total variance instead. Consequently we set $\frac{\sigma^2}{\sigma_0^2} = 1$. This yields constant values of $\frac{1}{31}$ (including the weight for $\mu_0$), i.e. all incoming new observations and the initial seed mean are all weighted equally (figure 1). This means that for a constant ratio of variances no forgetting of past observations is possible.

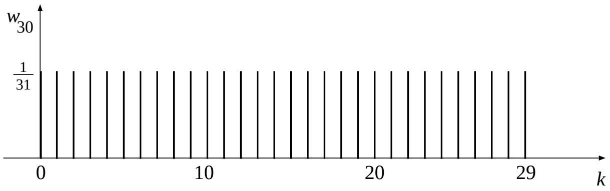

**Figure 1:** Weighting function for $n = 30$ **Bayesian** adaptation passes.

**Heuristic Adaptation**

However, for some applications such as varying adaptation conditions we would like to introduce the possibility of forgetting so that learning anew is feasible. This is important in the case of permanently changing tasks. In order to achieve this goal we simplify (3) to $\alpha = const.$ We have tested this case extensively in [2] where we determined heuristically the $\alpha$s that result in the best adaptation performance. The weighting function (6) reduces to $\alpha(1-\alpha)^k$ with $0 \leq k \leq n-1$. Then the weight of the initial value $\mu_0$ equals $(1-\alpha)^n$. It is important to note that a constant adaptation rate $\alpha$ results in an exponentially attenuated influence of past observation vectors (figure 2), i.e. learning anew is possible.

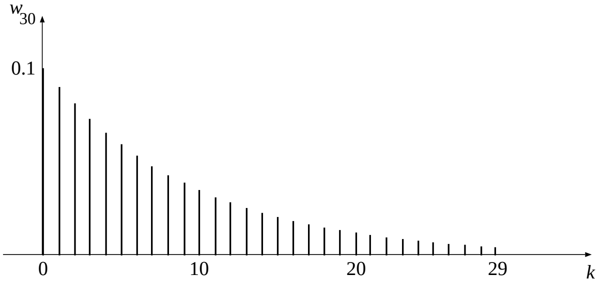

**Figure 2:** Weighting function for $n = 30$ **heuristic** adaptation passes and $\alpha = 0.1$.

## 2.2. Task Adaptation vs. Speaker Adaptation

We address the problem of task mismatch between training and testing processes. For a more thorough discussion please refer to [2]. Obviously above adaptation formulae merge previous knowledge (means of the pdfs) with new incoming observations (feature vectors) without taking the real adaptation purpose into account. The adaptation subject has to be given implicitely by making sure the new observations are generated fulfilling a specific criterion. Therefore it might be more appropriate to talk about a *hidden adaptation* or a *topic specific reestimation.* We have chosen the expression *Task Adaptation* in analogy to the well–known *Speaker Adaptation*, where very successful algorithms (e.g. [1]) are based on speaker specific reestimation. Other work on speaker adaption that is algorithmically related to our adaptation technique can be found e.g. in [6, 7].

The main differences between training data and the data of a new task is given by the new contexts that occur through new combinations of phonemes. So it might be straightforward to adapt only the context dependent states of our phoneme based 3–state HMMs. Although we can show in experiments that this yields already a considerable improvement, including the middle state in the updating process outperforms the first suggestion. The effect is mainly due to the fact that contexts (especially in small vocabulary systems) are spread over several phonemes and cannot be modeled appropriately by means of a diphone system.

In contrast to speaker adaptation we are not looking for a uniform shift of pdf–clusters in acoustic space. We would rather like to find a new arrangement of clusters that is given by the new task. Therefore we generally do not tie mixtures for the adaptation process.

## 2.3. Adaptation Strategy

The basic idea is to take an unbiased monophone seed model as a baseline and use its phonemic inventory for the working model when the recognition vocabulary changes. Whenever a new task that was previously unknown is being set up a diphone working model is automatically created corresponding to the new dictionary. The needed context dependent states are now copied from the corresponding context independent seed model. The working model is now subjected to task adaptation during the recognition process. Note that this adaptation has a long term memory, i.e. the acoustic models are always gradually adapted to the current application.

Using the Viterbi algorithm each observation vector $\vec{x}_t$, $t = 1, 2, ...T$ can be mapped to a state $\theta_t^i$ of the best model $i$ after recognition. To model the state emission probabilities we are using multivariate Laplacian distributions. Given a mapping between observation vector and state we determine now the mean $\vec{\mu}^i$ that is nearest to $\vec{x}_t$ using the city block distance measure. This nearest mean is now updated after each utterance according to (2). Depending on how fast the new task is going to change we use either Bayesian or heuristic adaptation.

# 3. MULTILINGUAL PHONEME MODELING

Cross language transfer of speech technology requires huge amounts of speech data to train a recognizer in the new language. To avoid this problem we exploit in our approach the acoustic–phonetic similarities of sounds across languages. In previous work [3] we have developed multilingual phoneme models based on Hidden-Markov-Modeling for a variety of languages. The rationale behind this is to create a balanced non–specialist HMM seed model that unifies the properties of several languages. This model is then used as an improved starting model for adaptation/transfer into any target language, so that convergence is being sped up and overall performance is getting better.

**Creating Multilingual Phoneme Models (MLPMs)**
We model acoustic–phonetic similar phonemes across languages as multilingual phonemes. The MLPMs are trained on the OGI ML-TS corpus for the three languages German, Amercian English and American Spanish [8]. Each phoneme model consists of a 3–state continuous density HMM. The states are modeled by Laplacian mixture densities where the number of pdfs depends on the number of occurences in the training material. The 125 language dependent phoneme models are mapped to a universal phoneme set using a log likelihood based distance measure. For the languages which are included in the training corpus the mixture weights are estimated for each language separately whereas the pdfs are tied across the languages. Hence, there are pdfs which are used in all of the three languages and pdfs used only in a single language. For the MLPMs which are used in a new target language the mixture weights are averaged across the languages. Finally, the multilingual inventory constists of 72 context independent phoneme models. The MLPMs provide broad acoustic–phonetic models for a variety of languages [3].

# 4. ADAPTIVE ONLINE LANGUAGE TRANSFER

The goal of using multilingual phoneme models is to use them as a base for improved transfer to a new language. Here we use them as seed models for adaptation to a slavic language (Slovene) that has not been included in the multilingual inventory. Although all of the IPA phonemes for Slovene are included in the MLPMs one has to adapt and optimize the parameters for Slovene.

We will show that our MLPMs speed up the cross language transfer of speech technology. Other related strategies (e.g. [9, 10]) work offline, require an advance training set, and do not make use of unsupervised acoustic adaptation.

**Mapping from Multilingual to Slovene Phonemes**
In this work the Slovene phoneme set in the acoustic dictionary used for the digit task is mapped manually to the MLPMs by phonetic knowledge rather than by a statistical method because there is currently no labeled speech data available. After the mapping we obtain 18 Slovene phoneme models with 2002 pdfs.

# 5. EXPERIMENTS AND RESULTS

First we evaluate the influence of the described possibilities to carry out task adaptation towards a known task. From the knowledge derived from these experiments we adopt the best parameters for optimal multilingual adaptation. All experiments are carried out with continuous density HMMs and 8 kHz telephone data is being used exclusively.

## 5.1. Experiments on Task Adaptation

For these experiments we first trained a monophone seed model that covers general German task–independently. To carry out this training we used the phonetically balanced sentences from the German part of the SpeechDat–1 database [11]. As target we use a German isolated word task with a vocabulary perplexity of 62. The task mismatch is given by the new recognition vocabulary. It is always ensured that the utterances of the adaptation set have random order so that no implicit speaker adaptation is possible. Adaptation and test sets are both recorded on identical channel characteristics. Regarding the heuristic adaptation we found that $\alpha = 0.1$ yields a good adaptation performance [2]. Results are shown in figure 3. The hypothesis that the heuristic adaptation yields better results than the Bayesian adaptation during the first adaptation steps is confirmed experimentally. However, Bayesian adpatation outperforms the heuristic one as more adaptation data become available.

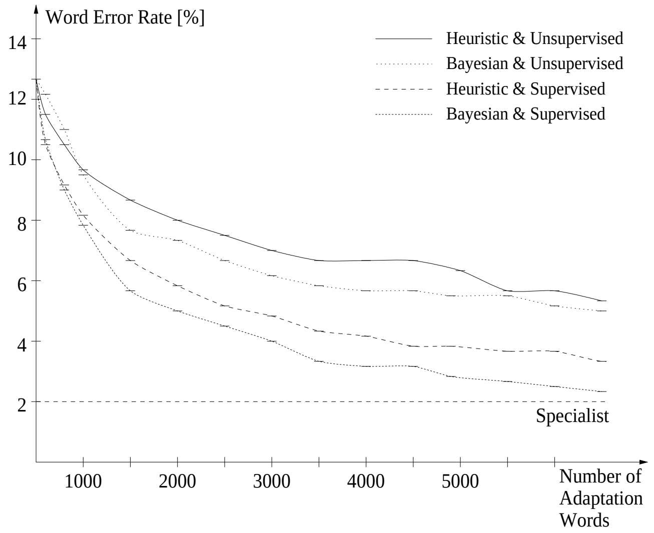

**Figure 3:** Word error rates corresponding to number of adaptation utterances; $\alpha = 0.1$ is used for both heuristic adaptation set–ups.

## 5.2. Experiments on Adaptive Language Transfer

We test our multilingual approach on a Slovene digit task. The acoustic lexicon contains 12 Slovene digits. Overall we have 646 isolated words, uttered by 54 speakers, for adaptation.

In order to measure the advances achieved by multilingual modeling we also create purely language dependent German, American English, and American Spanish models that we use also as seed for adaptation to Slovene modeling.

The recognition results are presented in table 1. It can be seen that the mapped HMMs yield a modest recognition performance with the multilingual HMM outperforming the others. It is important to keep in mind that these are bootstrapping recognition rates. Language and task dependent training – if possible – would yield better results, of course. One has to mention that from a phonetical point of view German already offers the phoneme inventory that is closest to Slovene when compared to English and Spanish. Therefore the gap of performance between English or Spanish modeling and multilingual modeling is bigger.

Since we are dealing with a long–term process we are using Bayesian adaptation here. In all cases in–service task adaptation causes a clearly noticeable boost in performance, with the best result achieved through multilingual HMM modeling.

## 6. CONCLUSION

We reported on recent advances in the fields of task adaptation and multilingual phoneme modeling. In a novel approach of combining these two technologies we could show that in–service bootstrapping of an automatic speech recognition system to a new language is possible without prior training data. The proposed strategy outperforms conventional, non–adaptive HMM modeling.

| *TASK ADAPTATION* | *OFF* | *ON* |
|---|---|---|
| German HMM | 73.4% | 83.4% |
| American English HMM | 62.7% | 69.0% |
| American Spanish HMM | 65.9% | 76.6% |
| **Multilingual HMM** | **76.5%** | **85.0%** |

**Table 1:** Recognition Performance on the Slovene digits task.

## 7. ACKNOWLEDGEMENTS

The authors are indebted to Harald Höge due to his contributions and encouragements regarding this research.

## 8. REFERENCES

1. Lee C.H., Gauvain J.L.; "Speaker Adaptation Based on MAP Estimation of HMM Parameters", *Proc. IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing*, pp. II-558–II-561, Minneapolis MN, 1993
2. Bub U.; "Task Adaptation for Dialogues via Telephone Lines", *Proc. Intern. Conf. on Spoken Language Processing*, pp. 825–828, Philadelphia PA, 1996
3. Köhler J.; "Multi–Lingual Phoneme Recognition Exploiting Acoustic–Phonetic Similarities of Sounds", *Proc. Intern. Conf. on Spoken Language Processing*, pp. 2195-2198, Philadelphia PA, 1996
4. Duda R.O., Hart P.E.; "Pattern Classification and Scene Analysis", John Wiley & Sons, New York, 1973
5. Ruske G.; "Automatische Spracherkennung", Oldenbourg, Munich, 1994
6. Dobler S., Rühl H.W.; "Speaker Adaptation for Telephone Based Speech Dialogue Systems", *Proc. Eurospeech*, pp. 1139–1142, Madrid, 1995
7. Thelen E.; "Long Term On–Line Speaker Adaptation for Large Vocabulary Dictation", *Proc. Intern. Conf. on Spoken Language Processing*, pp. 2139–2142, Philadelphia PA, 1996
8. Cole A., Muthusamy Y., Oshika B.; "The OGI Multi–Language Telephone Speech Corpus", *Proc. Intern. Conf. on Spoken Language Processing*, pp. 895–898, Banff, 1992
9. Dalsgaard P., Andersen O., Barry W.; "Data-Driven Identification of Poly– and Monophonemes for four European Languages", *Proc. Eurospeech*, pp. 759–762, Berlin, 1993
10. Wheatley B., Kondo K., Anderson W., Muthusamy Y.; "An Evaluation of Cross Language Adaptation for Rapid HMM Development in a New Language", *Proc. IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing*, pp. 237–240, Adelaide, 1994
11. Höge H., Tropf H., Winski R., van den Heuvel H., Haeb–Umbach R., Choukri K.; "European Speech Databases for Telephone Applications", *Proc. IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing*, to appear, Munich, 1997