Improving Children's Speech Recognition by HMM Interpolation with an Adults' Speech Recognizer

Stefan Steidl, Georg Stemmer, Christian Hacker, Elmar Nöth, and Heinrich Niemann *

Universität Erlangen-Nürnberg, Lehrstuhl für Mustererkennung, Martensstraße 3,
D-91058 Erlangen, Germany
stefan.steidl@informatik.uni-erlangen.de

Abstract. In this paper we address the problem of building a good speech recognizer if there is only a small amount of training data available. The acoustic models can be improved by interpolation with the well-trained models of a second recognizer from a different application scenario. In our case, we interpolate a children's speech recognizer with a recognizer for adults' speech. Each hidden Markov model has its own set of interpolation partners; experiments were conducted with up to 50 partners. The interpolation weights are estimated automatically on a validation set using the EM algorithm. The word accuracy of the children's speech recognizer could be improved from 74.6 % to 81.5 %. This is a relative improvement of almost 10 %.

1 Introduction

Traditionally, automatic speech recognition has been focusing on adults' speech while speech of children has been ignored almost completely. Nevertheless, the economic market for children's speech recognizers is growing. You just have to think of the huge number of children having already mobile phones which could be controlled via speech or of toys with speech recognition like SONY's entertainment robot AIBO. Unfortunately using a speech recognizer for adults to recognize children's speech yields only very poor results, because children's speech differs too much from adults' speech. One possible solution of this problem is the collection of large amounts of children's speech data what is expensive and time-consuming. Furthermore, finding test persons is much more difficult with children than with adults since the parents must agree, the children have to be picked up and brought home again and so on.

In literature, often MLLR (maximum likelihood linear regression) or MAP (maximum a posteriori) methods are applied to adapt a speech recognizer for adults' speech to children's speech. Another promising technique is vocal tract

* A part of this work was funded by the European Commission (IST programme) in the framework of the PF-STAR project under Grant IST-2001-37599. The responsibility for the content lies with the authors.

length normalization (VTLN). In this paper HMM interpolation is used to solve the problem: A children's speech recognizer is trained on a small amount of children's speech and afterwards the hidden Markov models (HMMs) are interpolated with the HMMs of an adult speech recognizer in order to increase the robustness of the models. Note that HMM interpolation is not in contrast to the techniques mentioned above. Especially a combination with VTLN makes sense and will be investigated in the near future.

The following issues are addressed in this paper: What are good interpolation partners? With how many partners should be interpolated? Which HMM parameters have to be interpolated? Which method is used for interpolation? In the following we describe a data-driven algorithm to choose an optimal set of interpolation partners for each hidden Markov model. The number of interpolation partners is optimized on a validation set and varies from one to 50. The parameters of the semi-continuous HMMs are interpolated linearly. The interpolation weights are estimated automatically on the basis of a validation set using the EM algorithm.

The idea to interpolate HMMs which have been trained on different datasets in order to achieve robust models is not new. For instance, K. Livescu [3] uses HMM interpolation to combine recognizers for non-native and native speech. Interpolation has also been employed for the same purpose by L. Mayfield Tomokiyo in [4]. Both approaches have in common that a single interpolation weight is shared by all HMMs and each hidden Markov model has only one *fixed* interpolation partner.

2 Interpolation of Hidden Markov Models

This paper focuses on the interpolation of semi-continuous hidden Markov models. In the following it is assumed that all HMMs share one common codebook consisting of K Gaussian densities. As each speech recognizer comes up with its own codebook, both codebooks have to be merged first. A greedy algorithm is used which selects sequentially the best pair $(\mathcal{N}_1, \mathcal{N}_2)$ of Gaussian densities and merges them into a new density $\mathcal{N}_3$ by taking the average of the density parameters. In our case only a simple mapping of the densities was performed. As a distance measure in order to choose the best pair of densities the increase of the entropy ΔH between the original densities $\mathcal{N}_1$ and $\mathcal{N}_2$ with their a priori probabilities p_1 and p_2 on the one hand and the resulting density $\mathcal{N}_3$ on the other hand is used:

$$\Delta H = (p_1 + p_2) \cdot H(\mathcal{N}_3) - \big(p_1 \cdot H(\mathcal{N}_1) + p_2 \cdot H(\mathcal{N}_2)\big) \ . \tag{1}$$

The entropy H of a Gaussian density $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is calculated as follows:

$$H(\mathcal{N}) = \int \mathcal{N}(\boldsymbol{x}) \cdot \ln\big(\mathcal{N}(\boldsymbol{x})\big)\,\mathrm{d}\boldsymbol{x} = \frac{1}{2} \ln\big((2\pi e)^D \cdot |\boldsymbol{\Sigma}|\big) \ . \tag{2}$$

D is the dimension of the feature vector $\boldsymbol{x}$. The algorithm is iterated until each density is merged. More details and extensions of the algorithm can be found in

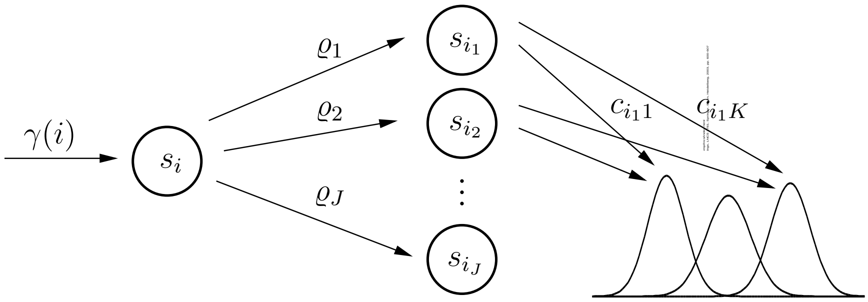

Fig. 1. The linear interpolation problem (3) can be interpreted as a hidden Markov model

[7]. In the next two sections, we describe the linear interpolation method and the estimation of the interpolation weights using the EM algorithm [6]. Afterwards the algorithm to choose the best interpolation partners is presented.

2.1 Linear Interpolation

We consider the general case of J interpolation partners. All J hidden Markov models are assumed to have the same number of states. The K mixture weights c_{ik} of the HMM state s_i are interpolated with the mixture weights $c_{i_j k}$ of the interpolation partners $s_{i_2}, \ldots, s_{i_J}$ as follows, where we set $s_{i_1} = s_i$ and $c_{i_1 k} = c_{ik}$:

$$\forall k : \quad \hat{c}_{i_k} = \varrho_1 \cdot c_{i_1 k} + \ldots + \varrho_J \cdot c_{i_J k} \quad \text{with} \sum_{j=1}^{J} \varrho_j = 1 \; . \tag{3}$$

In a second step the transition probabilities a_{ij} of state i are interpolated with the same interpolation weights ϱ_j.

2.2 Estimation of the Interpolation Weights

As each state of each HMM which has to be interpolated has its own set of interpolation weights ϱ_j, a tremendous number of parameters has to be estimated. This is done automatically on the basis of a validation set using the EM algorithm. The estimation formulas for the interpolation weights are based on [6, p. 305].

In order to use the EM algorithm to estimate the weights the problem (3) is interpreted as a discrete hidden Markov Model as shown in Fig. 1 [1]. As before, state $s_i = s_{i_1}$ is interpolated with the states s_{i_2} to s_{i_J}. The interpolation weights ϱ_j are interpreted as the transition probabilities from state s_i to the states s_{i_j}. The mixture weights $c_{i_j k}$ correspond to the output probabilities $b_{i_j}(k)$. The EM algorithm is an iterative parameter estimation technique which calculates new values of the parameters on the basis of existing estimates. The probability

$P(s_{i_j} \mid k, s_i, \boldsymbol{\varrho})$ is the probability of being in state s_{i_j} if the output is codeword k and an existing set of estimates $\boldsymbol{\varrho}$ is given. It's calculated as follows:

$$P(s_{i_j} \mid k, s_i, \boldsymbol{\varrho}) = \frac{P(s_{i_j}, k \mid s_i, \boldsymbol{\varrho})}{P(k \mid s_i, \boldsymbol{\varrho})} = \frac{\varrho_j \cdot c_{i_j k}}{\sum_{j=1}^{J} \varrho_j \cdot c_{i_j k}} \; . \tag{4}$$

Using this equation you can calculate the transition probabilities ϱ_j.

$$\varrho_j = P(s_{i_j} \mid s_i, \boldsymbol{\varrho}) = \sum_{k=1}^{K} P(k \mid s_i, \boldsymbol{\varrho}) \cdot P(s_{i_j} \mid k, s_i, \boldsymbol{\varrho}) \tag{5}$$

In order to get new estimates of the transition probabilities the term $P(k \mid s_i, \boldsymbol{\varrho})$ in (5) is replaced with the probability $\zeta(i,k) = P(s_i, k|\boldsymbol{X}, \boldsymbol{\lambda})$. This term is calculated on the validation set.

$$\tilde{\varrho}_j = \sum_{k=1}^{K} \zeta(i,k) \cdot \frac{\varrho_j \cdot c_{i_j k}}{\sum_{j=1}^{J} \varrho_j \cdot c_{i_j k}} \tag{6}$$

Due to this replacement the new estimates of the transition probabilities have to be normalized to meet the condition $\sum_{j=1}^{J} \varrho_j = 1$.

$$\hat{\varrho}_j = \frac{\tilde{\varrho}_j}{\sum_{j=1}^{J} \tilde{\varrho}_j} \tag{7}$$

The algorithm stops if the estimates of the transition probabilities don't change anymore. With the following measure of quality [6, p. 305] the success of the HMM interpolation can be evaluated quickly without having to re-compute the likelihood $P(\boldsymbol{X}|\boldsymbol{\lambda})$ of the validation set:

$$\ell(\varrho_1, \ldots, \varrho_J) = \log \prod_{k=1}^{K} \left(\sum_{j=1}^{J} \varrho_j \cdot c_{i_j k} \right)^{\zeta(i,k)} \tag{8}$$

$$= \sum_{k=1}^{K} \zeta(i,k) \log \left(\sum_{j=1}^{J} \varrho_j \cdot c_{i_j k} \right) \; . \tag{9}$$

2.3 Determination of the Interpolation Partners

We now can interpolate any hidden Markov Model with an arbitrary set of interpolation partners. The time required to calculate the interpolation weights and the amount of data available for a robust estimation of the interpolation weights is the only limiting factor to the number of interpolation partners. We found it reasonable to restrict the number of partners to at most 50 for our experiments. This raises the question which HMMs are good interpolation partners. In a first pass we therefor interpolate each HMM of the first speech recognizer with all

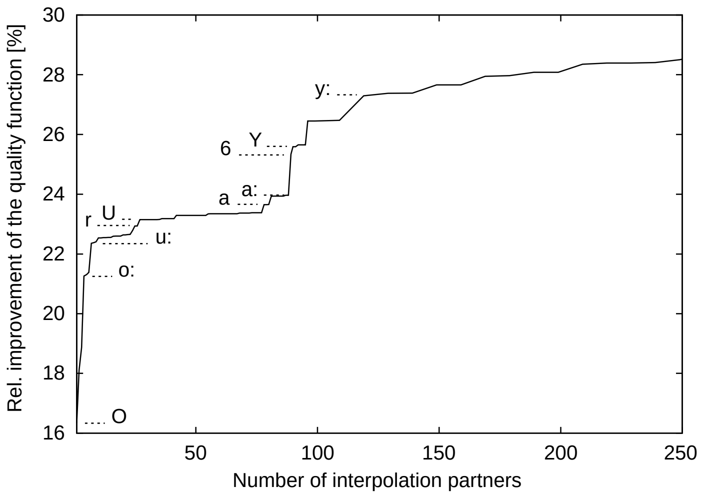

Fig. 2. Results of the interpolation of the monophone `o:` with the first n $(1 \leq n \leq 250)$ interpolation partners of the n-best list in terms of the relative improvement of the quality function. The marks indicate where new core phones appear for the first time, they are labeled with the name of the new core phone in SAMPA notation [5]

models of the second recognizer individually and evaluate the improvement of the quality function (9). In doing so you obtain a list of n possible interpolation partners. Figure 2 shows the results of the interpolation of the monophone `o:` with the first n $(1 \leq n \leq 250)$ interpolation partners of the n-best list in terms of the relative improvement of the quality function. Two aspects become evident: Firstly, choosing only the first 50 interpolation partners yields only a suboptimal result. Secondly, the graph shows noticeable steps. These steps are caused by HMMs which represent polyphones[1] with identical core phone and similar right and left context. In Fig. 2 marks indicate where new core phones appear for the first time, they are labeled with the name of the new core phone in SAMPA notation [5]. Similar HMMs yield nearly the same result if they are interpolated separately. But in combination the results can't be improved any further. Hence it makes sense not to choose the first n entries of the n-best list, but to choose only those polyphones whose distance to the interpolation partners which are already chosen is larger than a given threshold. As a distance measure the Kullback-Leibler divergence between corresponding HMM states (10) is used.

$$d(s_i, s_j) = \sum_{k=1}^{K} c_{ik} \cdot \log \frac{c_{ik}}{c_{jk}} \tag{10}$$

[1] Polyphones are the generalization of the well-known concepts of bi- or triphones and allow a variable-sized context.

Table 1. Partitioning of the children's speech corpus for the training of the speech recognizer and for the interpolation of the hidden Markov models

Task	Speakers	Texts
Training Speech Recognizer	40	*Zürcher Lesetest*
	6 of 40	*Nordwind und Sonne*
HMM Interpolation	6 - 40	*Nordwind und Sonne*
Evaluation	20	*Nordwind und Sonne*

3 Speech Database

In this paper we describe the interpolation of a speech recognizer for children with a recognizer for adults. The children's speech corpus consists of read speech of 62 children (29 male and 33 female) at the age of 10 to 12 years. The pupils read four different German texts: *Nordwind und Sonne (The North Wind and the Sun)* and three texts of the reading test *Zürcher Lesetest* [2]. Each text is about 90 words long. The vocabulary consists of 227 entries. Altogether 3.5 hours of read children's speech are available.

To train the adults' speech recognizer a subset of the recordings of the Verbmobil Project [8] was used. It consists of 28 hours of spontaneous dialogues between humans in German (11,762 turns of 610 dialogues). The vocabulary contains 6825 entries.

4 Experimental Results

4.1 Baseline Recognizer

As an adults' speech recognizer we use the Verbmobil recognizer for spontaneous speech. Its codebook consists of 500 densities. On the Verbmobil test set a word accuracy of 76.1 % is achieved using a 4-gram language model. If this recognizer is used to recognize children's speech (vocabulary reduced to *Nordwind und Sonne*, no language model) a word accuracy of 61.9 % is achieved.

In order to obtain a baseline system for children's speech we retrained this recognizer using the texts of the *Zürcher Lesetest* of 40 children. The testing of the recognizer and the evaluation of the HMM interpolation is performed with the text *Nordwind und Sonne* of the remaining 20 children. The data of the speakers of the training set reading *Nordwind und Sonne* is used for HMM interpolation. In order to include polyphones of the *Nordwind und Sonne* text in the model inventory of the children's speech recognizer the corresponding recordings of 6 of the 40 training speakers are added to the speech recognizer's training data. Table 1 shows the partitioning of the children corpus. The speakers of the training and test sets are disjoint. To evaluate the effects of the HMM interpolation we don't use any language model. Our baseline speech recognizer yields a word accuracy of 74.6 %.

Table 2. Results of the HMM interpolation with a varying number of interpolation partners

Experiment	Word Accuracy
Baseline	74.6 %
1 partner	79.2 %
5 partners	79.8 %
10 partners	80.1 %
20 partners	**80.9 %**
30 partners	80.7 %
40 partners	**80.9 %**
50 partners	80.8 %

4.2 HMM Interpolation

In the experiments described in this paper we interpolate our baseline recognizer for children's speech with the VERBMOBIL recognizer for adults' speech. The first group of experiments evaluates the optimal number of interpolation partners. The method to choose the interpolation partners is described in Sect. 2.3. The full validation set consisting of 40 speakers reading *Nordwind und Sonne* is used. Table 2 shows the results of these experiments. With only one interpolation partner, the word accuracy of our speech recognizer can be improved from 74.6 % to 79.2 %. As expected, you get even better results with more interpolation partners. The maximum of 80.9 % is reached with 20 resp. 40 partners. This is equivalent to a relative improvement of 8.4 %.

The second group of experiments evaluates the influence of the size of the validation set used to calculate estimates of the interpolation weights. The experiments are conducted with 20 and with 50 interpolation partners. Table 3 shows the results. It could be expected that you will need a large validation set to get robust estimates of the interpolation weights. Fortunately, this is not the case. The size of the validation set has only little influence on the HMM interpolation. The best results are achieved with even a small validation set consisting of only 6 resp. 12 speakers. Using 50 interpolation partners, a maximal word accuracy of 81.5 % is reached. Compared to the baseline system, this is a relative improvement of 9.2 %. The fact that a relatively small validation set is sufficient is an important result because if a large validation set was required it could have been better to use this data for training of the (baseline) speech recognizer instead for interpolating the hidden Markov models. Further experiments on other speech data will show whether this is a fortunate coincidence or not.

5 Conclusion and Outlook

Our experiments show two things: Firstly, if a speech recognizer has poorly trained models because of a lack of training data it can be improved by interpolating its models with the models of a second speech recognizer although the speech databases of both recognizers are different. Our new approach to choose

Table 3. Results of the HMM interpolation with 20 and with 50 partners and a varying number of speakers in the validation set

Partners	Validation Set	Word Accuracy
20 partners	6 speakers	**81.4 %**
	12 speakers	81.1 %
	18 speakers	80.7 %
	24 speakers	81.0 %
	30 speakers	80.8 %
	40 speakers	80.9 %
50 partners	6 speakers	81.3 %
	12 speakers	**81.5 %**
	18 speakers	81.3 %
	24 speakers	81.1 %
	30 speakers	81.4 %
	40 speakers	80.8 %

a different set of interpolation partners with up to 50 partners for each model is successful. This method is more promising than the HMM interpolation with only one fixed partner. A direct comparison between both methods is still missing. Secondly, adults' speech can help to recognize children's speech although both kinds of speech differ quite much. In our concrete case, the word accuracy of our children's speech recognizer could be improved by almost 10 %. Further experiments combining our approach with VTLN will be carried out. Due to the fact that all children read the same four texts, it would be better to add the validation set to the training of the baseline recognizer. We therefore plan to redo the experiments using a new children corpus with a much bigger vocabulary.

References

1. Jelinek, F. and Mercer, R. L.: Interpolated Estimation of Markov Source Parameters from Sparse Data. In: Gelsema, E. S., Kanal, L. N. (eds.): Pattern Recognition in Practice. North Holland Publishing Co., Amsterdam (1980) 381–397
2. Linder, M. and Grissemann, H.: Zürcher Lesetest. 6th edn. Testzentrale Göttingen, Robert-Bosch-Breite 25, 37079 Göttingen (2000), http://www.testzentrale.de
3. Livescu, K.: Analysis and Modeling of Non–Native Speech for Automatic Speech Recognition. Master Thesis, Massachusetts Institute of Technology (1999)
4. Mayfield Tomokiyo, L.: Recognizing Non–Native Speech: Characterizing and Adapting to Non–Native Usage in LVCSR. PhD Thesis, Carnegie Mellon University (2001)
5. SAMPA – Computer Readable Phonetic Alphabet. http://www.phon.ucl.ac.uk/home/sampa/home.htm
6. Schukat-Talamazzini, E. G.: Automatische Spracherkennung – Grundlagen, statistische Modelle und effiziente Algorithmen. Vieweg (1995)
7. Steidl, S.: Interpolation von Hidden Markov Modellen. Diploma Thesis (in German), Chair for Pattern Recognition, University of Erlangen-Nuremberg (2002)
8. Wahlster, W.: Verbmobil: Foundations of Speech-to-Speech Translation. Springer (2000)

Erschienen in: Michaelis, Krell (Eds.), Pattern Recognition, 25th DAGM Symposium, Proceedings, LNCS 2781, Springer, Berlin, Heidelberg, 2003, pp. 600-607