Incremental word learning using large-margin discriminative training and variance floor estimation

Irene Ayllón Clemente[1,2], *Martin Heckmann*[2], *Alexander Denecke*[1,2], *Britta Wrede*[1], *Christian Goerick*[2]

[1]Research Institute for Cognition and Robotics, Bielefeld University, Germany
[2]Honda Research Institute Europe GmbH, Offenbach am Main, Germany
{iayllon,adenecke,bwrede}@cor-lab.uni-bielefeld.de, {firstname.lastname}@honda-ri.de

Abstract

We investigate incremental word learning in a Hidden Markov Model (HMM) framework suitable for human-robot interaction. In interactive learning, the tutoring time is a crucial factor. Hence our goal is to use as few training samples as possible while maintaining a good performance level. To adapt the states of the HMMs, different large-margin discriminative training strategies for increasing the separability of the classes are proposed. We also present a novel estimation of the variance floor when a very low number of training data is used. Finally our approach is successfully evaluated on isolated digits taken from the TIDIGITS database.

Index Terms: speech recognition, discriminative training, efficient learning

1. Introduction

We focus on an interactive learning scenario where a human tutor teaches a robot. This is inspired by the process of speech acquisition in children. For auditory learning in small children, a closed loop of speech perception and production plays an important role. While some authors concentrate on jointly solving both aspects [1], others [2] constrain their work on the generation of robust perception as it is in itself still a widely unsolved problem in automatic speech recognition (ASR) systems. Our current work also only considers the perceptual aspects.

Unfortunately, conventional ASR offline training techniques rely on a large amount of labelled training data which is not available in interactive learning. Because of this, and to maintain a short tutoring time, researchers aim either to train the system in an unsupervised manner (i.e. without a tutor) or to train it with a reduced number of samples.

One of the main drawbacks using a small amount of training data is that learning algorithms may fit the model parameters to some specific features of the dataset (overfitting). In our framework, we use Hidden Markov Models (HMMs), the most frequently used representations in ASR systems. Maximum likelihood (ML) estimation is often deployed to adapt the parameters of HMMs including the means and variances of its Gaussian Mixtures Models (GMM). A variance estimated from only a few training samples might not be representative for the underlying distribution. More precisely there is a tendency of underestimating the variances in such cases, if the value of the variance is too small. The Baum-Welch algorithm, the most commonly used ML training algorithm, can be modified by including a lower threshold on the variance parameters, a variance floor [3]. This translates the problem to the computation of this floor value. One simple way of computing the variance floor is estimating the average within state variances scaled by a predefined factor [4]. In [5], a method is used to adapt the variance floor to each dimension of the features. The method computes the average variance over all Gaussian components in each feature-dimension. Then this value is also scaled by some constant and used as variance floor.

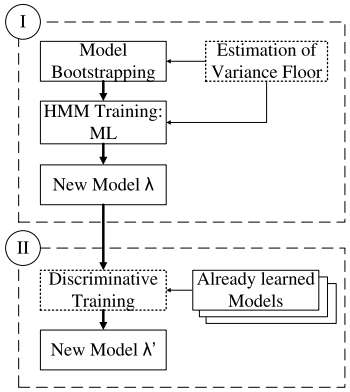

Figure 1: An overview of the incremental discriminative training system, which consists of two main stages. Stage I is explained in Sections 2.1 and 2.2 and the stage II in Section 2.3. ML stands for maximum likelihood estimation.

As mentioned before, when using a small number of training samples the resulting distributions generally differ from the true distribution of the speech segments. Additionally to ML estimation, discriminative training (DT) has been widely investigated for HMMs in ASR [6, 7]. The DT methods directly minimize the classification errors on the training data as the model estimation criterion. In our previous work [8] we investigated minimum classification error (MCE) estimation using the extended Baum-Welch (EBW) algorithm proposed in [9]. However, the application of incremental MCE training was not beneficial because of the very low number of training samples used. In this case, after ML all training data was already classified correctly. Recently, the generalization ability of HMMs has been further improved by taking the margin of the classifier into account, these techniques are called large-margin discriminative training [5, 10].

In this paper we present an extension of the incremental word learning framework introduced in [8], where an unsupervised initialization of the parameters of a HMM is performed, followed by the retraining and construction of a new HMM using multiple sequence alignment (MSA). In Fig. 1 the main contribution of this paper compared to [8] is reflected in the blocks framed by dotted lines. In stage I, we initialize the parame-

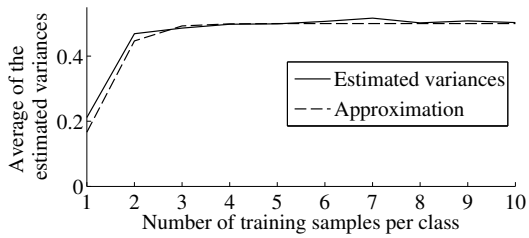

Figure 2: The average $\overline{X}_V(n)$ of the variances of the Gaussian Mixtures Models (GMM) over all the models for a different number of samples n and its approximation by the Gompertz function are depicted here.

ters of the HMMs by means of the model bootstrapping method proposed in [8] and a novel variance floor estimation, where we go a step further by not only estimating the variance floor depending on the feature-dimension [5] but also on the number of samples used. After retraining the HMMs using ML estimation, different large-margin discriminative training strategies are introduced and analyzed in stage II. These strategies also improve the recognition results when few training data are used.

The remainder of the paper is organized as follows. In Section 2 we give an overview of our incremental word learning framework. In Section 2.1, we briefly introduce the model bootstrapping presented in [8]. Next, the estimation of the variance floor and the different large-margin discriminative training strategies proposed are described in Sections 2.2 and 2.3, respectively. In Section 3 we report the results for our approach on an isolated digit recognition task and compare them to a standard approach. Finally, in Section 4 we discuss the results and give an outlook on future work.

2. Incremental word learning system

The incremental word learning system consists of two main stages (Fig. 1). First, the parameters of the Hidden Markov Models are initialized by means of a model bootstrapping method. This stage allows to estimate a good initial set of HMM parameters, which are trained by the Baum-Welch algorithm [11] (ML estimation) afterwards. In this stage, a suitable variance floor dependent on the number of samples used has to be chosen. Finally, in stage II different large-margin discriminative training strategies refine the estimates of the parameters computed in the stage before.

2.1. Model bootstrapping and ML estimation

As mentioned in Section 1, the Baum-Welch algorithm [11] is usually employed to train Hidden Markov Models. Unfortunately, this algorithm easily gets stuck in local minima. Thus, it is necessary to have a model bootstrapping which provides an adequate initialization of the HMM parameters to obtain good convergence.

The model bootstrapping system used here was presented in [8] and comprises three main steps: the unsupervised training of a generic HMM, in which a common HMM initialization model is built without using any labelled training data. Here an unrelated speech segment stored from an independent source (not implicitly including the words to learn) is used. Next, training of the previously obtained HMM using the Baum-Welch algorithm [11] on labelled training data is performed. This yields ergodic word-level HMMs. The main contribution in [8] was the proposal of an algorithm for transforming the ergodic word-level HMM into a left-to-right word-level HMM in the next step. The multiple sequence alignment (MSA) algorithm [8] iteratively merges the information contained in the Viterbi-decoding sequences of the training data into an optimal state sequence modelling the topology of a left-to-right HMM. The computational complexity of the MSA is $O(m^2T^2)$, where m is the number of training samples used and T is the length of the longest training sample of the model. These steps are the basis for the construction of a new word-level HMM, which is retrained by the Baum-Welch algorithm afterwards using the estimation of the variance floor described in the following section.

2.2. Estimation of the variance floor

In Section 1, the relevance of setting a variance floor was motivated when a very small number of training samples is used. To decide which floor constant is optimal for our task (see Section 3.1), first it is necessary to evaluate how the variances decrease when the number of samples n is reduced. In Fig. 2 the average $\overline{X}_V(n)$ over the variances of all Gaussian Mixture Models for all feature-dimensions and models after one iteration of the Baum-Welch algorithm is displayed. One can observe that the variances abruptly grow after one sample and then slightly increase until they saturate at $\overline{X}_V(\infty) = 0.5$. This behaviour can be easily modelled using a sigmoid function, called the Gompertz function:

$$G(n; a, b, c) = a \cdot e^{b \cdot e^{c \cdot n}} \tag{1}$$

with parameters a, b and c adapted to the data. In Fig. 2 the dashed line represents the following Gompertz function: $v(n) = -(0.5 + G(n; -1, -3, -2))$, where n is the number of samples. Once the behaviour of the variances is modelled, $v(n)$ is normalized such that $v_{f1}(n)$ converges to 1 for $n \to \infty$, resulting in $v_{f1}(n) = \overline{X}_V(\infty)/v(n) = 0.5/v(n)$. However, as the number of training samples decreases a larger variance floor is required [3]. The reinforcement term $r_f(n) = 1 + e^{-n}$ compensates this effect, by increasing the value of the function $v_{f_1}(n)$ for a very small number of training samples. Finally, by multiplying the reinforcement term $r_f(n)$ and the normalized variance function $v_{f1}(n)$ estimated above, we obtain a function of the variance floor depending on the number of samples used:

$$v_f(n) = \frac{0.5 \cdot (1 + e^{-n})}{-(0.5 + G(n; -1, -3, -2))} \tag{2}$$

The variance floor estimation used in [5] which is based on a larger number of training samples is shown in Eq. 3. Here K is the scaling constant and $\overline{X}_{V,d}$ is the average variance over all Gaussian components in each dimension d.

$$V_F(d) = K \cdot \overline{X}_{V,d} \tag{3}$$

Our variance floor estimation is represented in Eq. 4. The variance floor function $v_f(n)$ and the scaling factor K are multiplied by the average variance $\overline{X}_{V,d}$ over all Gaussian components in each feature-dimension. This yields a variance floor value depending on the dimension of the feature d and the number of samples n used.

$$V_F^*(d, n) = K \cdot v_f(n) \cdot \overline{X}_{V,d} \tag{4}$$

This variance floor V_F^* is used in each iteration of the Baum-Welch algorithm.

2.3. Large-margin discriminative training

The main idea of the large-margin principle is to estimate the HMMs parameters in such a way that the decision boundary determined by the estimated HMMs achieves the maximum classification margin.

For a word sample s_i, assuming that it belongs to class W_i, the multi-class separation margin for s_i is defined as [10]:

$$\begin{aligned} d(s_i) &= F(s_i|\lambda_{W_i}) - \max_{W_j \in \Omega; j \neq i} F(s_i|\lambda_{W_j}) \\ &= \min_{W_j \in \Omega; j \neq i} [F(s_i|\lambda_{W_i}) - F(s_i|\lambda_{W_j})] \end{aligned} \quad (5)$$

where Ω denotes the set of all possible words, F is a discriminant function and λ_{W_j} is the word-level HMM of the class W_j.

Different optimization algorithms have been proposed for large-margin computation. The most prominent ones are gradient descent [10] and semidefinite programming [12]. Semidefinite programming algorithms provide the best results, however they have a high computational cost. In on-line learning it is fundamental to reduce the computation time, hence we use the limited memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) algorithm [13] to ensure a faster convergence.

As mentioned in Section 1, when a small number of samples is used, the HMM models may be overtrained. In this case, the models already seem well separated. This can be characterized by large distances between them. Therefore, optimizing the model in each iteration with the minimum distance criterion according to Eq. 5 does not improve recognition results significantly.

To optimize the models, one heuristic is to slightly increase the variances in order to produce overlapping models. In this way we multiply the variances with a scaling factor S to intentionally worsen the estimations and to force the algorithms to recalculate the means of the models. In addition to the scaling factor S, we propose two alternative strategies to the distance criterion of Eq. 5 used in combination with [10]. These strategies select a set of classes to optimize their models in each iteration, thereby reducing the computational cost, which means, saving the retraining of some models, and improving the recognition results for a small number of training samples.

2.3.1. Retraining the last introduced word-model (RLM)

As we are dealing with an incremental word learning system, we consider that the word-models are improved incrementally. That means, if word-model 1 and 2 have already been optimized, word-model 3 can only affect the relation between 1 and 3, and 2 and 3, but it may not influence the relation between word-models 1 and 2. This approach assumes that the only word-model to readapt in each step is the last model added. This strategy reduces the computational cost significantly, because only one model is retrained in each iteration. This method is referred to as retraining last model (RLM).

2.3.2. Selecting word-models via confidence intervals (CBS)

As mentioned before, it is not possible to get a reliable estimation of the word-models when using a very small number of training data. Furthermore, not all models are uniformly estimated. Hence, instead of using the distance of each sample to each already learned model (see Eq. 5), we propose to select the word-models by using the probability distribution of their discriminant functions F, i.e., confidence based selection (CBS).

To construct each distribution, the discriminant function $F(s_j|\lambda_{W_j})$ is calculated for each sample s_j belonging to each class W_j, which is modelled by the word-level HMM λ_{W_j}. Once the distributions of the discriminant functions are computed, we analyze the number of samples s_i which may be wrongly generated by the class W_j using the discriminant function $F(s_i|\lambda_{W_j})$ of the sample s_i and confidence intervals. Finally, the models of the classes W_j wrongly showing a high confidence that sample s_i belongs to these classes are retrained.

3. Experiments

3.1. Experimental procedure

In the unsupervised phase described in Section 2.1, the database used is a subset of TIMIT [14] with alternating utterances from men and women. TIMIT contains recordings of 630 speakers, each reading ten phonetically rich sentences. To evaluate the following phases of our incremental word learning system, a subset of the TIDIGITS corpus [15] containing only isolated digits was used. This subset contains utterances from 112 men and women collected from 21 regions of the United States. There are a total of eleven words (digits) in the corpus vocabulary (digits of "1" to "9", plus "oh" and "zero"). From this subset, several datasets are generated. First, the subset is split up into a test set and several training sets. The test set contains 224 samples for each digit from men and women and each training set up to 10 labelled samples for each digit. Each training segment that we have selected was uttered by a different speaker, where exclusively men were used. Further, for the estimation of the behaviour of the variances depicted in Fig. 2, a random set of training samples was also selected.

In our experiments, the 45-dimensional acoustic feature vectors consist of 15 RASTA-PLP coefficients [16] and their first and second order time derivatives. The RASTA-PLP features were first decorrelated by means of Principal Component Analysis (PCA) and afterwards normalized. The PCA coefficients and the normalization parameters are computed from a subset of TIMIT. All data is sampled at a rate of 16 KHz. The models used in our experiments are continuous density HMMs (CDHMMs) with 16 hidden states, where each state is described by a Gaussian Mixture Model (GMM) with 3 components. As baseline system, a Hidden Markov Models framework implemented as in [11] using a statistical Matlab Toolbox called NETLAB [17] is used.

First, we start evaluating the stage I of our incremental word learning system (Fig. 1). We compare our variance floor estimation method V_F^* presented in Section 2.2 with the method V_F from [5] using as model bootstrapping the conventional initialization of the baseline system BL proposed in [11]. The value of K was set to 1. After that, the model bootstrapping MSA explained in Section 2.1 is also evaluated when using our proposed variance floor estimation method V_F^*. This is the final configuration used as basis for stage II.

After performing the experiments for stage I, we evaluate the advantage of adding stage II by analyzing the two proposed large-margin discriminative training strategies (RLM and CBS) described in Section 2.3. Thereby only the means of the GMMs are updated. The scaling factor S is set to 1.1 for RLM. In CBS, S is set to 1 if the model of the class W_j overlaps more than 2 models of different classes W_i, to 1.1 if it overlaps 1 or 2 models and to 1.2 if it does not overlap any model. Setting the length of the confidence interval to $\pm\sigma$, with σ being the standard deviation of the distribution of the discriminant functions F, saves computation time and does not impair the recognition results.

3.2. Experimental results

Compared to the variance floor estimation method (Eq. 3) in [5], our method (Eq. 4) improves the results obtained in the baseline system [11] when a very small number of training samples is used (see Table 1). However, when $v(n)$ approaches $\overline{X}_V(\infty)$, both methods are very similar and the recognition results are not improved. Furthermore, the MSA model bootstrapping described in [8] clearly reduces the word error rates (WER) com-

Nr.	BL(V_F)		BL(V_F^*)		MSA		RLM		CBS	
Data	M	W	M	W	M	W	M	W	M	W
10	1.0	33.0	1.0	33.0	0.4	28.8	0.3	27.4	0.1	25.5
9	1.1	33.1	1.1	33.1	0.5	29.6	0.4	28.4	0.2	26.9
8	1.4	34.7	1.4	34.7	0.6	31.3	0.5	29.9	0.3	28.4
7	1.6	36.9	1.6	36.9	0.7	33.1	0.5	30.8	0.4	29.4
6	1.7	41.1	1.7	39.4	1.0	35.2	0.8	33.4	0.6	32.0
5	2.3	42.0	2.0	40.2	1.2	36.7	1.1	35.3	0.8	33.3
4	5.6	48.2	3.1	46.4	2.0	42.7	1.6	39.8	1.6	39.1
3	10.5	56.1	6.3	54.3	3.7	50.9	3.1	46.8	3.0	45.5
2	35.3	79.1	22.8	71.7	13.4	67.6	12.1	64.9	12.0	64.0
1	75.8	88.4	44.9	80.9	32.6	77.3	31.6	75.4	32.3	75.0

Table 1: Word Error Rates (WER %) of the model bootstrapping method used, the baseline system and the different large-margin discriminative training strategies. For each method, the WER values represent the mean of a 20-fold cross-validation on the male training data set, evaluated on separated male (M) and female (W) test data sets. MSA stands for the multiple sequence alignment bootstrapping method proposed in [8], BL for the baseline system, V_F for the variance floor estimation in [5] and V_F^* for the variance floor estimation presented in Section 2.2. In the case of discriminative training, RLM and CBS stand for the strategy 1 and 2 respectively.

pared to the baseline system as shown in Table 1. In Fig. 3, the achieved improvement of MSA against the baseline system BL(V_F) is 41% in men and almost 15% in women when 6 training samples are used.

Additionally, Table 1 and Fig. 3 show that large-margin discriminative training strategies in combination with MSA model bootstrapping outperform the baseline system for male and female voices, independent of the strategy used. Nevertheless, CBS provides the best results with a 65% relative improvement in men and 22% in women with respect to the baseline system BL(V_F) when 6 training samples are used.

4. Discussion and Summary

We have proposed an incremental word learning system [8] extended by different large-margin discriminative training strategies and a variance floor estimation dependent on the feature-dimension [5] and the number of training samples.

First, we have demonstrated that our variance floor estimation method improves the recognition results when a very small number of training samples is used. Further it is not necessary to tune any other parameter to obtain convergence in comparison with [8]. Second, we have shown that using the semi-supervised model bootstrapping method in [8] outperforms a conventional initialization baseline system [11] in all test cases, e.g. for 6 training samples an improvement of 41% in male voices was obtained. Finally, two large-margin discriminative training strategies used in combination with the semi-supervised model bootstrapping mentioned before were presented. Each of the strategies outperforms the preceding algorithms as shown in Section 3.2. The RLM strategy is faster than CBS, however the CBS method provides superior results. Furthermore, the generalization power of our system is shown via the female test sets, where recognition results are also improved although no training samples containing female voices were used. Additionally, our system is optimized to operate online.

In human-robot interactive learning, a reduction from 1.7% WER when using only 6 training samples in a speaker-independent task obtained by the baseline system to 0.6% WER with our approach is a significant and highly relevant improvement. In other words, the performance obtained by the baseline system 1% WER for 10 training samples can be maintained while reducing the tutoring time to half (0.8% WER for 5 training samples). In future work, we will investigate if our results also hold for more complex tasks and expand our system towards a multimodal learning framework.

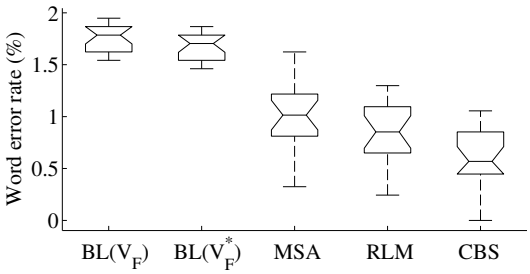

Figure 3: The box plots of the WER of all methods presented are displayed for 6 training samples only considering the tests on male voices. For abbreviations, see Table 1.

5. References

[1] Minematsu, N., "A modulation-demodulation of speech communication," in *Proc. 5th Int. Conf. on Speech Prosody*, 2010.

[2] ten Bosch, L., Boves, L., Van Hamme, H. and Moore, R. K., "A computational model of language acquisition: the emergence of words," *Fundam. Inf.*, vol. 90, no. 3, pp. 229–249, 2009.

[3] Melin, H., Koolwaaij, J.W., Lindberg, J. and Bimbot, F., "A comparative evaluation of variance flooring techniques in hmm-based speaker verification," in *Proc. 5th Int. Conf. on Spoken Language Processing*, 1998.

[4] Young, S. J., Evermann, G., Gales, M. J. F., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V. and Woodland, P. C., *The HTK Book, version 3.4*, Cambridge University Engineering Department, Cambridge, UK, 2006.

[5] Dong, Y., Deng, L., He, X. and Acero, A., "Large-margin minimum classification error training: A theoretical risk minimization perspective," in *Comput. Speech Lang.*, 2008, vol. 22, pp. 415–429.

[6] Juang, B. H., Hou, W. and Lee, C. H., "Minimum classification error rate methods for speech recognition," in *IEEE Trans. on Speech and Audio Process.*, 1997, pp. 257–265.

[7] McDermott, E., *Discriminative training for speech recognition*, Ph.D. thesis, Waseda University, 1997.

[8] Ayllón Clemente, I., Heckmann, M., Sagerer, G. and Joublin, F., "Multiple sequence alignment based bootstrapping for improved incremental word learning," in *Proc. 35th IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2010.

[9] He, X., Deng, L. and Chou, W., "Discriminative learning in sequential pattern recognition - a unifying review for optimization-oriented speech recognition," *IEEE Signal Processing Magazine*, pp. 14–36, 2008.

[10] Li, X., Jiang, H. and Liu, C., "Large margin hmms for speech recognition," in *Proc. 30th IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2005.

[11] Rabiner, L. R., "A tutorial on hidden markov models and selected applications in speech recognition," *Proc. of the IEEE*, vol. 77, pp. 257–286, 1989.

[12] Li, X. and Jiang, H., "Solving large margin estimation of hmms via semidefinite programming," in *Proc. INTERSPEECH*, 2006.

[13] Nocedal, J. and Wright, S. J., *Numerical optimization*, Springer Verlag, NY, 1999.

[14] Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S. and Dahlgren, N. L., "Darpa timit acoustic-phonetic continuous speech corpus cd-rom," Linguistic Data Consortium, 1993.

[15] Leonard, R., "A database for speaker-independent digit recognition," in *Proc. 9th IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 1984.

[16] Hermansky, H. and Morgan, N., "Rasta processing of speech," *IEEE Trans. Speech and Acoustics*, vol. 2, pp. 587–589, 1994.

[17] Nabney, I. T., *NETLAB: algorithms for pattern recognition*, Springer Advances in Pattern Recognition Series, 2002.