BOOSTING LONG–TERM ADAPTATION OF HIDDEN–MARKOV–MODELS: INCREMENTAL SPLITTING OF PROBABILITY DENSITY FUNCTIONS

Udo Bub[1,2] *and Harald Höge*[1]

[1]Corporate Research and Development, Siemens AG, Munich, Germany
[2]Inst. for Human–Machine–Communication, Munich Univ. of Technology (TUM), Munich, Germany
Udo.Bub@mchp.siemens.de

ABSTRACT

The research described in this paper focuses on possibilities to avoid the tedious training of Hidden-Markov-Models when setting up a new recognition task. A major speaker independent cause for the decrease of recognition accuracy is a mismatch of the phonetic contexts between training and testing data. To overcome this problem, we introduced in previous work the idea of an update of task independent acoustic models by means of Bayesian learning. In this paper we introduce the new approach of adaptively splitting the probability density functions (pdfs) of a continuous density HMM. The goal is to model the appropriate state pdfs better so that they can more accurately match new contexts that are observed while the system is in service. Splitting AND Bayesian adaptation yields a remarkable reduction of word error rate compared to Bayesian adaptation only.

1. INTRODUCTION

Especially speech recognition systems with a small vocabulary have to be trained and tuned with often tedious efforts for a specific application. Word specific speech databases are necessary for sufficiently high recognition rates. On the other hand it is well–known that systems that have been designed for a certain assignment perform poorly when they are tested under conditions they have not been trained for (e.g. [1]). A major *speaker independent* cause for these deteriorations is a mismatch of the acoustic phonetic contexts between the phonemes of the training task and recognition task respectively. That is, the dictionaries which contain the words of a task and their phonetic transcriptions differ from training to testing. A problem occurs when training of a specialist model is not possible because at system development time there is not sufficient training data available or because the phonetic dictionary is being changed by the user during the application (*type–in*). In previous work [2, 3] we have developed a set–up to overcome these problems by means of a Bayesian re–estimation of the means of the probability density functions (pdfs) of a continuous density Hidden–Markov–Model (HMM). Doing so, the means are gradually adapted so they match better the new acoustic contexts that occur during the new application.

Using mixture Gaussians is a well–known means for adequate acoustic modeling of allophones. Various training techniques (e.g. [4]) start with a minimum number of states which are successively split during several training iterations. In the new approach shown here we start a recognition task with an HMM that has been thoroughly trained on phonetically balanced data. Now, while in service, the system not only adapts its means to the testing observations but also splits the pdfs to achieve superior modeling of newly occurring inter–phonetic contexts. The splitting criterion is based on entropy minimization and the goal is to open a new Gaussian whenever separate modeling of a context or an allophone is favorable. The proposed algorithm works online and is embedded in an incremental adaptation paradigm.

Our basic idea is to train first a phonetically balanced generalist HMM which can cope with every task equally well but which, by nature, offers a lower recognition performance on a specific task than a potential specialist HMM if this one could have been trained beforehand. Then, using the algorithms explained below, the acoustic models are gradually enhanced with the aim to match the accuracy of a specialist model.

2. ENTROPY BASED SPLITTING CRITERION

In previous work entropy has been applied as a good statistical estimate for expected increase or decrease of the recognition rate (e.g. [5]) of the acoustic models of a speech recognition system.

2.1. Entropy of approximated Gaussians

The entropy is defined as follows:

$$H_p = -\int_{-\infty}^{+\infty} p(\vec{x}) \log_2 p(\vec{x}) d\vec{x}. \qquad (1)$$

Assuming that $p(\vec{x})$ is a Gaussian distribution with a diagonal covariance matrix, i.e.

$$\mathcal{N}(\vec{\mu}, \sigma_n) = \frac{1}{\sqrt{(2\pi)^N}} \frac{1}{\prod_n \sigma_n} e^{-\frac{1}{2}\sum_n \frac{(x_n-\mu_n)^2}{\sigma_n{}^2}} \quad (2)$$

yields

$$H_p = \sum_{n=1}^{N} \log_2 \sqrt{2\pi e}\sigma_n. \quad (3)$$

In our acoustic models we approximate $p(\vec{x})$ with $\hat{p}(\vec{x}) = \mathcal{N}(\vec{\hat{\mu}}, \sigma_n)$, where $\vec{\hat{\mu}} = \frac{1}{L}\sum_{l=1}^{L} \vec{x}_l$ is the mean of L observations. The corresponding entropy as a function of $\hat{\mu}$ is given by

$$H_{\hat{p}}(\hat{\mu}) = -\int_{-\infty}^{+\infty} p(\vec{x}) \log_2 \hat{p}(\vec{x}) d\vec{x}, \quad (4)$$

which results in

$$H_{\hat{p}}(\hat{\mu}) = H_p + \sum_{n=1}^{N} \frac{(\mu_n - \hat{\mu}_n)^2}{\sigma_n^2} \log_2 \sqrt{e}. \quad (5)$$

The expectation $E\{(\mu_n - \hat{\mu}_n)^2\}$ equals $\frac{1}{L}\sigma_n^2$, so the expectation of $H_{\hat{p}}(\hat{\mu})$ is given as

$$H_{\hat{p}} = E\{H_{\hat{p}}(\hat{\mu})\} = H_p + \frac{N}{L} \log_2 \sqrt{e}. \quad (6)$$

2.2. Splitting of Gaussians

Let $\hat{p}(\vec{x}) = \mathcal{N}(\vec{\hat{\mu}}, \sigma_n)$ be the pdf to be split. We assume that the two Gaussians that result from a splitting process have the same deviation σ^s and are both weighted equally. This yields new pdfs:

$$\hat{p}^s(\vec{x}) = \frac{1}{2}\mathcal{N}(\vec{\hat{\mu}}_1^s, \sigma^s) + \frac{1}{2}\mathcal{N}(\vec{\hat{\mu}}_2^s, \sigma^s) \quad (7)$$

Using (4) it is possible to compute the entropy $H_{\hat{p}}^s$ of the approximated split Gaussian. Assuming that $\mu_1 \approx \hat{\mu}_1$, $\mu_2 \approx \hat{\mu}_2$, and μ_1 is sufficiently far away from μ_2 yields an easier solution of the integral because then the overlap of some Gaussians can be neglected. This finally results into

$$H_{\hat{p}}^s = 1 - \sum_{n=1}^{N} \log_2 \sqrt{2\pi e}\sigma_n^s + \frac{1}{2}(\log_2 \sqrt{e}\frac{N}{L_1} + \log_2 \sqrt{e}\frac{N}{L_2}) \quad (8)$$

As splitting criterion we demand a decrease of the entropy, i.e.

$$H_{\hat{p}} - H_{\hat{p}}^s > C, \quad (9)$$

where C (with $C > 0$) is a constant that indicates the desired magnitude of the entropy decrease. For simplicity we choose $L_1 = L_2 = \frac{L}{2}$ and obtain following splitting criterion:

$$\sum_{n=1}^{N} \log_2 \frac{\sigma_n}{\sigma_n^s} > \log_2 \sqrt{e}\frac{N}{L} + 1 + C. \quad (10)$$

For practical purposes we simplify (10) to

$$\frac{L}{N}\sum_{n=1}^{N} \log_2 \frac{\sigma_n}{\sigma_n^s} > C', \quad (11)$$

where C' has to be determined empirically.

3. COMBINING SPLITTING WITH INCREMENTAL ADAPTATION

Due to simplicity we are going to discuss the the formulae in the 1–dimensional space. A Bayesian update of the mean $\hat{\mu}$ of a normal density after adaptation step l is achieved by

$$\hat{\mu}_l = (1 - \alpha_l)\hat{\mu}_{l-1} + \alpha_l x_l, \quad (12)$$

with

$$\alpha_l = \frac{1}{l + \frac{\sigma^2}{\sigma_0^2}}. \quad (13)$$

x_l is the lth observation mapped to a specific state during the Viterbi decoding and σ^2 is the variance of the adaptation material. This recursive adaptation formula is initialized with $\hat{\mu}_0$ and σ_0^2 which are mean and variance of the previously trained task independent model. In our experiments we use

$$\alpha_t = const. \quad (14)$$

Please refer to [3] for an explanation of the implications that come along with this simplification.

The splitting of densities is carried out in contextual domain and can be also interpreted as parallel state splitting [4]. Now, a key question is how to choose the location of the means of the two new densities resulting from a split. In this case we choose $\hat{\mu}_1^s$ to stay as $\hat{\mu}$ and $\hat{\mu}_2^s$ to be the mean value of the adaptation material mapped to $\hat{\mu}$ during the Viterbi path. Splitting is done for fine tuning when necessary. After every utterance the criterion (11) is evaluated and splitting is carried out accordingly. The adaptation and splitting processes are initialized with a seed HMM model with already approximately 6000 densities.

4. BASELINE SYSTEM

For both training and recognition telephone speech data is sampled at 8 kHz. Every 10 ms a feature vector is computed based on the data of an overlapping 25 ms Hamming window. A feature vector consists of 51 elements of which are 24 cepstrally smoothed spectral coefficients, 12 Δcepstral, and 12 $\Delta\Delta$cepstral components as well as 1 energy, 1 Δenergy, and 1 $\Delta\Delta$energy component.

In order to take occurring channel variations into account we apply a short–term maximum likelihood channel compensation to the feature extraction. During training we

carry out the computation of a Linear Discriminant Analysis (LDA) resulting in a superior class separation capability. Before LDA we build a 2–frame super vector from which we retain after transformation only the 24 most significant components as input for the Viterbi search. We use 6–state continuous density HMMs.

5. APPLICATION AND RESULTS

The splitting is implemented for long–term, incremental adaptation processes. The adaptation is supposed to work in an unsupervised mode in real world applications. However, here we are using supervised adaptation for the evaluation of the algorithms. An application in the field of speaker adaptation might be thinkable but is not straightforward because many adaptation data is needed and the contextual or allophonic differences are rather task dependent than speaker dependent.

For these experiments we first train a monophone seed model that covers general German task–independently. This model is called the *generalist* in the further text. It is used as the model with which the process is initialized. To carry out the training for the generalist we use the phonetically balanced sentences from the German part of the SpeechDat–1 database [6] and from the database SieTill. Please refer to [7] for further information. Altogether we have 16500 utterances from 1500 speakers for training. We use standard Viterbi training and the generalist model has 6000 mixture densities overall. Because the pdfs have been trained iteratively on phonetically varying utterances the model can cope with every phonetic contexts equally well, but offers a lower recognition performance on a specific task than a specialist.

As target of the adaptation we use a German isolated word task with a vocabulary perplexity of 62 (our internal database VM). We choose this task because we have plenty of speech material available for this database. So we can train a specialist model whose performance we can compare to the adapted model's performance in order to get an idea of the viability of the algorithms. The pdfs are iteratively trained and thus highly specialized in the acoustic contexts given by the training material.

5.1. Baseline

The baseline performance of the system can be seen in table 1. The test set consists of 1500 utterances from VM database that have not been involved with any other training/adaptation process described in this paper. During training of the monophone specialist we investigated also the influence on the error rate of the amount of the training material and the number of parameters that constitute the HMM. As can be seen in figure 1 the recognition accuracy improves

generalist (monophone)	specialist (monophone)	specialist (diphone)
93.1%	97.1%	97.9%

Table 1: Baseline performance.

the more data are available for training. Furthermore, the minimum error rate is found at a higher pdf number the more data has been used, i.e. the more data is available the more pdfs should be used for the acoustic modeling. The ef-

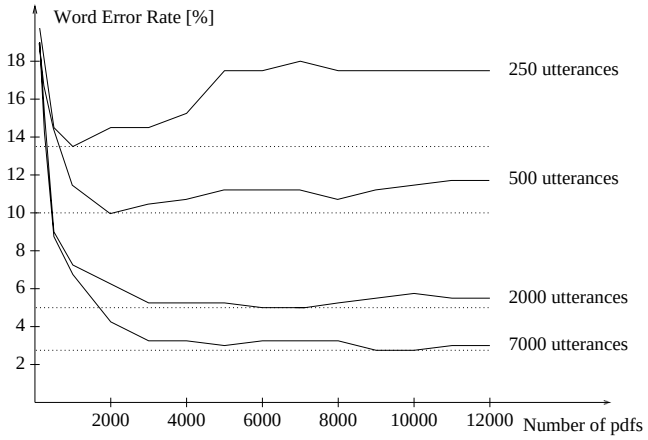

Figure 1: Word error rate corresponding to the number of training utterances and pdfs.

fect can be explained as follows: the more pdfs are available (provided there is enough training data) the more accurate the modeling for the context can be realized. For example, different allophones can be modeled more precisely now. In the next step we would like to make use of this fact for boosting the re-estimation process.

5.2. Adaptation without Splitting

The new task contexts are 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. We use $\alpha = 0.1$ for the adaptation. Looking at the regular monophone adaptation (table 2) we observe an improvement of the recognition accuracy, especially during the first adaptation steps. Adding more adaptation material in this case does not seem to improve the accuracy. In order to get an idea how better adaptive context modeling can improve the re–estimation process we try also a diphone adaptation: First we determine the occurring diphone contexts from the dictionary of the new task. Then we fill the state distributions with

the distributions from the monophone generalist model, so that the middle states represent the same phoneme. This diphone model has 17700 densities now. That is, from the monophone seed model we create an equivalent diphone seed model which, besides the diphone modeling, has more parameters ready for adaptation. As can be seen, diphone

no. of adaptation words	0	2000	7000
acc. monophone adapt.	93.1%	95.4%	95.4%
acc. diphone adapt.	93.1%	96.4%	97.3%

Table 2: Word accuracy using regular adaptation.

adaptation outperforms the monophone adaptation remarkably and does not get saturated as early as the monophone model. After 7000 utterances the performance is close to the one of the specialist.

5.3. Influence of Splitting

Now, we evaluate the influence of splitting on the adaptation. First we carry out splitting without any underlying adaptation process of the kind of formula (12). Table 3 shows that splitting alone already yields an improvement. A limitation of the total number of pdfs might be useful if the real time performance of the system cannot be guaranteed for large numbers. The results of a combination of the splitting technique together with regular adaptation are shown in table 4. Applying splitting *and* regular monophone adaptation during 7000 adaptation words yields an improvement from 95.4% to 96.3% as compared to regular adaptation only.

no. of adaptation words	0	2000	7000
word accuracy	93.1%	93.4%	93.5%
number of densities	6000	6048	6354

Table 3: Splitting only (monophone modeling).

no. of adaptation words	0	2000	7000
word accuracy	93.1%	95.7%	96.3%
number of densities	6000	6028	6148

Table 4: Splitting with Bayesian adaptation (monophone modeling).

6. CONCLUSION AND FUTURE WORK

We have introduced the new idea of adaptive online splitting and have derived an entropy based splitting criterion. The ultimate goal is to adapt the acoustic modeling of phoneme contexts and allophones to a new task while the system is in service. Provided enough training data, modeling by means of a higher number of pdfs is preferable. The proposed system uses new incoming data to gradually learn new contexts and increase the number of used pdfs accordingly. Splitting AND Bayesian adaptation on 7000 adaptation words yields a 19.6% reduction of word error rate compared to Bayesian adaptation only. As a next step, we are looking into the possibility to merge two pdfs whenever another one is split. The advantage would be that there would always be a constant number of parameters of the HMM. The above algorithms are currently being tested for unsupervised adaptation. The first results confirm the quality of the results shown before, but with a lower overall recognition accuracy.

7. 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] Bub U., Köhler J., Imperl B.: “In–Service Adaptation of Multilingual Hidden–Markov–Models”, *Proc. IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing*, pp. 1451–1454, Munich, 1997.

[4] Takami J., Sagayama S.: “A Successive State Splitting Algorithm for Efficient Allophone Modeling”, *Proc. IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing*, pp. 573–576, Toronto, 1992.

[5] Hon H.W.: “Vocabulary–Independent Speech Recognition: The VOCIND System”, *PhD Thesis*, CMU–CS–92–108, School of Computer Science, Carnegie Mellon University, Pittsburgh PA, 1992.

[6] 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*, pp. 1771–1774, Munich, 1997.

[7] URL: “http://www.icp.grenet.fr/ELRA”