TASK ADAPTATION FOR DIALOGUES VIA TELEPHONE LINES

Udo Bub

Siemens AG, Munich, Germany
Munich University of Technology, Munich, Germany

E–mail: Udo.Bub@zfe.siemens.de

ABSTRACT

This paper describes our successful ongoing approaches toward better recognition accuracy for flexible interactive systems in automatic speech recognition. Degradation in performance of speech recognition systems is observed whenever any current application differs from the conditions during training time. Main speaker independent causes for these deteriorations are changes in transmission channels and changes in the task to be fulfilled. We present our results of research on changing tasks, i.e. more specifically on changing dictionaries. We propose an in–service adaptation technique that is speaker independent, works under unsupervised conditions, and has a long term memory. On 2000 adaptation words a reduction of error rate of more than 40% at negligible computational costs is achieved.

1. INTRODUCTION

It is well–known that automatic speech recognition (ASR) systems that have been designed for broad use perform during a special application poorly compared to systems that have been designed specifically for this very purpose. The reason for this degradation can be found in a mismatch between speaker characteristics, transmission channels, and task[1] (e.g. [1]) of the training data to those of the field assignment.

Especially applications of telephone based automated dialogue systems suffer from these limitations due to a very big population of possible speakers and huge differences in transmission channels that may vary from session to session. Also the task of real world applications is likely to change several times. For instance, an automatic telephone operator based on speech technology has to cope with a vast and ever changing variety of proper names of network users. This implies that the dictionary has to be kept flexibly and thus the specific task of the ASR system is not known during developing time.

[1] We claim that the recognition task is mainly characterized by the vocabulary and thus we are dealing particularly with the implications caused by changing dictionaries.

Systems that are determined to achieve broad user acceptance require an adaptation to specific customer needs without much reengineering effort. To overcome the described problems there is a recent trend to recognition systems that adapt their parameters to both changing speakers and channels (e.g. [1, 4, 7, 3]).

The problem of task adaptation has been recognized earlier (e.g. [5, 6]), but little is known about online adaptive acoustic modeling for this purpose. Modeling of such recognition set–ups faces following dilemma: Only a *generalist* Hidden–Markov–Model (HMM) that has been trained on phonetically balanced data can satisfactorily cope with all possible incoming recognition units of a task unkown during training time. However, a *specialist* model that has been trained on the same vocabulary as used during the application yields a considerably higher word recognition rate, mainly because it can make use of the coarticulations that it has already seen during the training phase.

We keep in mind that training of a giant vocabulary–independent HMM that can cope with all possible tasks like a specialist is prohibitive [5].
Real world applications demand strong constraints. Algorithms should be

- computationally inexpensive and easy to implement
- unsupervised
- speaker independent
- working online and should not require an advance adaptation set.

In the following we propose sequential in–service adaptation that meets above requirements.

2. BASELINE SYSTEM

Our baseline training and recognition tools use both continuous density Hidden–Markov–Models (CDHMMs) assuming multimodal Laplacian distributions. In this application the system allows for both context independent monophone modeling and context dependent diphone modeling [10].

As mentioned earlier we need to create a generalist seed HMM model as a starting–point for our adaptation. The telephone data base used for its training is SIETILL, an internal database. 6000 utterances of 1100 speakers are taken where they answered to various questions like when they were born, what is their phone number, from where they are calling, etc.

Given a different testing task this ensures that the seed model is trained vocabulary–independently and features no preferences for any specific recognitition unit. In our case the HMM is trained with monophone models, but training with tied diphones would be also straightforward. We refer to this model as the *generalist* from now on.
(We also train a different seed model on the German part of SpeechDat–I [9]. Training is performed on an excerpt of about 700 speakers with 10 utterances each. The speakers were asked to read 9 phonetically rich sentences from a newspaper and finally tell spontaneously what they had for breakfast on the day of the recording. All experimental results are very similar to those with the first generalist.)

In order to facilitate the evaluation of the adaptation results we also generated a reference HMM that has been trained on the testing task. For this we use an internal database called VM. It consists of 850 speakers each uttering 61 isolated, German command words for ISDN applications including digits. Both monophone and diphone training for the reference model has been carried out on a subset containing 150 speakers. The diphone reference HMM is called the *specialist*.

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 occuring channel variations into account a short term online channel adaptation has been developed in our labs [4]: By maximum likelihood estimation an approximate distortion vector is determined and subsequently subtracted from the incoming feature vector. During training we carry out the improved computation of an LDA–Matrix [4] 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.

3. ONLINE TASK ADAPTION

The strategy is to take the generalist monophone model as baseline and use its phonemic inventory for the working diphone model whenever the dictionary changes and a new context dependent segment is needed for the changed task. The model to be adapted is updated online during the recognition process. Note that this adaptation has a long term memory and no reset is done. This means that the HMM is always gradually adapted to the current application. Following steps are carried out whenever the dictionary changes:

- Read the dictionary and find out the needed context dependent segments
- If an occurring segment is unknown so far, copy the corresponding context independent segment distributions from the generalist model into the new segment of the working model
- Recognition of incoming utterances
- Rejection of unsafe recognition results if desired
- Online retraining of working model with an appropriate adaptation formula using the incoming data

Many successful general adaptation techniques that can be found in literature make use of information that refers only implicitly to the subject of adaptation (speaker, channel, task). A common way is to take the data of an adaptation set, apply a certain algorithm, and then use the acquired data to balance the seed models according to a certain application. That is, the algorithm can be seen in many cases independently from the final application which means also that many known adaptation techniques may be used also for task adaptation if applied in an appropriate way.

3.1. Adaptation Formula

We assume that the relevant differences between tasks affect mainly the parameters of the HMM probability density functions, or more specifically the location of their means in acoustic space.

The feature extraction module transforms an incoming utterance into a series of observation vectors:

$$X = \{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_T\} \tag{1}$$

Using the viterbi algorithm each observation vector $\vec{x}_t$, $t = 1, 2, \ldots T$ can be mapped to a state θ_t^i of the best model i after recognition. Because we use multimodal Laplacian distributions to model the state emission probabilities, the corresponding probability density function of the sth state of an HMM is approximated by

$$b_s^i(\vec{x}) = \sum_{m=1}^{M_s^i} c_{s,m}^i e^{-\frac{\sqrt{2}}{\sigma}||\vec{x}-\vec{\mu}_{s,m,t}^i||}, \tag{2}$$

where M_s^i, $c_{s,m}^i$, and σ are constants determined during training. Given a mapping between observation vector and state we determine now the mean $\vec{\mu}_{s,m,t}^i$ that is nearest to $\vec{x}_t$ using the city block distance measure (n denotes the component of a vector)

$$||\vec{x} - \vec{\mu}|| = \sum_n |x_n - \mu_n|.$$

This nearest mean is now updated according to (3).

$$\vec{\mu}^{i}_{s,m,t+1} = (1-\alpha)\vec{\mu}^{i}_{s,m,t} + \alpha\vec{x}_t \quad (3)$$

A geometric interpretation of (3) can be given as follows: The update $\vec{\mu}^{i}_{s,m,t+1}$ lies on a straight line going through the old mean $\vec{\mu}^{i}_{s,m,t}$ and the current observation vector $\vec{x}_t$. The parameter α can be viewed as the adaptation rate. In the special case of $\alpha = 0$ no adaptation will be carried out at all and for $\alpha = 1$ the update equals $\vec{x}_t$.
It is noteworthy that a constant adaptation rate results in an exponentially attenuated influence of past observation vectors, i.e. learning anew is possible. The same adaptation formula has been used in [2] for speaker adaptation.

3.2. Rejection

In all applications for dialogue systems recognition errors may occur. The reason can be for instance an incorrect input by the user or simply a misclassification by the recognizer. In case such an error is detected a cooperative dialogue manager should ask the user for a better utterance.

We do not want to focus in this paper on the many possibilties how to carry out the error detection. Instead we simulate the option of selecting good utterances for adaptation by means of a rather simple statistical rejection strategy [8]. The score s_0 of the best and the score s_1 of the second best hyothesis after an n–best search are being considered:

$$rejectionflag = \begin{cases} 1 & \text{if } (s_1 - s_0) \leq r_{thresh} \\ 0 & else \end{cases} \quad (4)$$

If $rejectionflag$ equals 1 the corresponding utterance will be omitted by the adaptation algorithm. In order to find a value to specify r_{thresh} we first determine s_{mean} empirically as the mean score per word of incoming utterances. We found that $r_{thresh} = 0.005 s_{mean}$ yields a correct rejection of wrong utterances of 61.2% and thus adaptation is done on "cleaner" data than without rejection.

4. EXPERIMENTS

The scenario for dialogue applications is going to be in a way that each incoming utterance will be used after recognition immediately for reestimation of the current model. However, in order to get results that are comparable with each other we are using during the experiments canned data for both adaptation and testing. The adaptation set is always taken from the VM partition that has been used for the training of the specialist (see section 1).

It is important to mention that the adaptation utterances have a random order so that a hidden speaker adaptation is not possible. The test set consists during all experiments of 1500 utterances from a partition of the VM data base that has never been used in any other process involved with this paper. The perplexity is 61 and the task can be considered as difficult due to a high confusability of words.

Baseline Performance
In advance tests the generalist monophone model achieves on our test set a word accuracy of 85.5%. On the other hand the vocabulary–dependent monophone model boasts an accuracy of 95.4%. The specialist diphone model yields with 97.1% the highest performance.

Different Learning Rates
In this test we use 2000 utterances for adaptation. The goal is to evaluate the influence of changing αs on the recognition rate. Using adaptation formula (3) we achieve following online results (table 1).

α	Error Rate
0	14.2%
0.025	10.6%
0.05	**9.3%**
0.075	**9.3%**
0.1	**9.3%**
0.125	9.4%
0.15	9.5%
0.175	10.3%
0.2	9.8%

Table 1: Word error rates for different αs and a constant adaptation set.

Already small αs yield an improvement which means that online information even in a small dose can boost recognition performance. A broad optimum for α is found between 0.05 and 0.1. The performance deteriorates for bigger αs which means that a too big learning rate leads to overadaptation. The maximum reduction of error rate is 34.5%.

We also carry out one supervised iteration of HMM viterbi training on the identical 2000 word adaptation set using the same seed model. The resulting error rate on the test set is 7.8% which is only 1.5% better than the best adaptation result.

Different Learning Rates & Rejection
For the same adaptation set as before we want to examine now the influence of rejection. As can be seen from table 2 the overall error rate decreases compared to the no–rejection case.

A clearer minimum at $\alpha_{best} = 0.125$ can be observed and a reduction of error rate of 40.1% is achieved. The fact that α_{best} is bigger than without rejection means that bigger adaptation steps may be chosen when a cleaner adaptation set is available.
The overall word recognition rate lags now only 0.7% behind one iteration of supervised training on the identical data.

Constant Learning Rate, Growing Adaptation Set
In this experiment we want to determine the influence of the number of adaptation words on the performance of the de-

α	Error Rate
0	14.2%
0.025	9.7%
0.05	9.2%
0.075	9.0%
0.1	8.8%
0.125	**8.5%**
0.15	8.9%
0.175	8.7%
0.2	9.3%

Table 2: Word error rates **using rejection** for different αs and a constant adaptation set.

scribed process. No rejection is done and figure 1 shows the results.
A convergence towards the specialist's performance can be observed. After 6000 words the decrease of error rate is 52.9%. From real world applications we know that 2000 calls per day are a realistic calling rate, so the needed scope of adaptation utterances poses no problem.

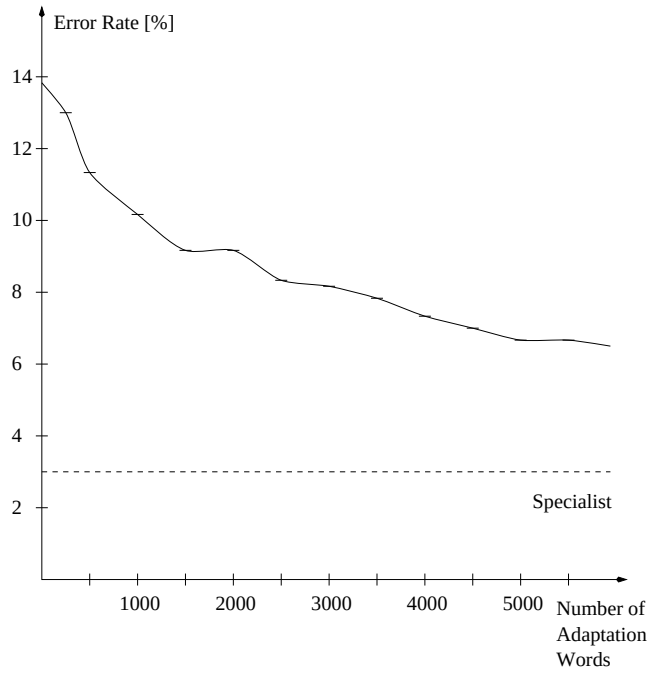

Figure 1: Word error rate corresponding to number of adaptation utterances for $\alpha = 0.1$

5. CONCLUSION

We have pointed out the need for task adaptive systems and have proposed a technique that works online without supervision at negligible computational costs. The scenario results in speaker independent improvement of the acoustic models of an ASR system given a task that was unknown during training phase. The algorithm works better the cleaner the data are.
Future work will focus on optimization of the proposed algorithm regarding faster convergence and better recognition.

6. ACKNOWLEDGEMENTS

The author would like to thank all members of the Siemens Speech Group for their help that was essential to carry out this research. Especially Harald Höge – who supervised this work – contributed significantly with many fruitful ideas.

This research was sponsored by a Siemens grant and was made possible by a cooperation between Siemens AG and the Institute for Human–Machine–Communication of the Munich University of Technology.

7. REFERENCES

1. Digalakis V.V., Rtischev R., Neumeyer L.G.; "Speaker Adaptation Using Constrained Estimation of Gaussian Mixtures"; *IEEE Trans. Speech and Audio Processing*, vol. 3, no. 5, pp. 357–365; 1995
2. Dobler S., Ruehl H.W.; "Speaker Adaptation for Telephone Based Speech Dialogue Systems"; *Proc. Eurospeech*, pp. 1139–1142; Madrid, 1995
3. Gauvain J.L., Lee C.H.; "Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains"; *IEEE Trans. on Speech and Audio Processing, Vol. 2, No. 2*, pp. 291–298; 1994
4. Hauenstein A., Marschall E.; "Methods for Improved Speech Recognition Over Telephone Lines"; *Proc. IEEE Intern. Conf. on Acoustics, Speech and Signal Processing*, pp. 425–428; Detroit MI, 1995
5. Hon H.W., Lee K.F.; "On Vocabulary–Independent Speech Modeling"; *Proc. IEEE Intern. Conf. on Acoustics, Speech and Signal Processing*, pp. 725–728; Albuquerque NM, 1990
6. 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
7. Leggetter C.J., Woodland P.C.; "Speaker Adaptation Using Linear Regression"; *Technical Report* CUED/F–INFENG/TR. 181, Cambridge Engineering Department; Cambridge, 1994
8. Littel B.; *personal communications*; Siemens AG 1995
9. Winski R. et al.; "Specification of Telephone Speech Data Collection"; *Technical Report* LRE–63314–D1.4–1, Commission of the European Communities, DG13/E–4; Luxembourg, 1996
10. Zünkler D.; "An ISDN Speech Server Based On Speaker Independent Continuous Hidden Markov Models"; *Proc. NATO–ASI*; Cetraro, 1990