ROBUST TIME-SYNCHRONOUS ENVIRONMENTAL ADAPTATION FOR CONTINUOUS SPEECH RECOGNITION SYSTEMS

Thomas Plötz, Gernot A. Fink

Faculty of Technology, Bielefeld University,
P.O. Box 100131, 33501 Bielefeld, Germany
{tploetz,gernot}@techfak.uni-bielefeld.de

ABSTRACT

In this paper we describe system architectures for robust MLLR based environmental adaptation of continuous speech recognition systems. Inspired by an existing broadcast news transcription system [1] we refined the identification of acoustic scenarios by using a combined GMM/HMM method. Thus environmental adaptation regarding arbitrary acoustic scenarios beyond speaker changes becomes possible. For deploying acoustic adaptation in interactive applications, such as human machine interaction, a time-synchronous adaptation approach is proposed. For different corpora the evaluation of our approaches shows significant improvements in recognition accuracy while satisfying the constraint of time-synchronous processing.

1. INTRODUCTION

The recognition performance of state of the art speech recognition systems is sensitive to possible mismatches between the acoustics of their training and test environments. Larger differences yield larger error rates. In order to overcome this problem acoustic adaptation techniques are the methodology of choice.

Depending on the processing strategy of an automatic speech recognition (ASR) system it is more or less suitable for applications within the field of human machine interaction. Systems following multi-pass approaches, such as the BBN BYBLOS recognizer [2] are well suited for batch processing tasks e.g. broadcast news transcription. Here the decoding process is split into several stages each increasing the restriction of the search space. In contrast to this, time-synchronous systems, such as the one developed at the technical university of Aachen [3] limit the decoding process to a single pass. Thus they are typically used for interactive applications within the domain of multi-modal human machine communication.

Requiring transcribed utterances of adaptation data the best transformation is estimated and applied to the particular ASR system. This procedure is called supervised adaptation. Besides the trivial case of batch processing where the transcription is explicitly given in most practical situations the annotation has to be generated automatically. One common approach for adaptation deployment is to manually decide whether (and when) the ASR system has to be modified or not. Once the modification is made it is valid for the further application progress. Following this procedure impressing improvements of recognition results are achievable as documented in e.g. [4, 5].

In general the approach outlined above cannot directly be transferred to interactive speech recognition applications. Online adaptation of an ASR system implicates the following essential constraints:

1. For preserving the interactive reactivity the computational effort of the adaptation procedure must not be substantial. Thus multi-pass processing approaches are not feasible. Besides this it is normally desirable to use only small amounts of adaptation data since gathering speech data on a larger scale is counterproductive for adapting an ASR system *fast*.
2. If adaptation is performed in an interactive environment the decision regarding the need for the (re-) adaptation of an ASR system should be made automatically. It is not desirable to monitor the deployment of the adaptation rule, e.g. due to explicitly signaling a speaker change.
3. For non-artificial applications (as benchmarks etc.) the annotation of the adaptation data needs to be created automatically using the best fitting acoustic model. If proceeding in such a manner in single-pass architectures the adaptational benefit can be utilized not until the adaptation data have been processed, which generally means too late. Within the field of human machine interaction, especially for dialog applications this is not satisfying since the adaptation fails if the acoustic scenarios change often. To overcome this in present system architectures multi-pass approaches are necessarily reintroduced determining in several stages the best fitting acoustic model for the current utterance. Due to the need of multiply processing every utterance this is beyond the intended time-synchronous approach required for interactive systems. The conflict of *immediately* benefiting from *acceptable* adaptation improvements needs to be resolved.

There is hardly any literature on approaches deploying adaptation methods in un-supervised time-synchronous speech recognition environments following the above described constraints. Zhang et al. proposed a transcription system for Japanese broadcast news including automatic speaker change detection and speaker adaptation [1]. Including all optimizations it is strictly speaking, however, a multi-pass system.

Following an extensive analysis of the Japanese system we present two enhanced approaches and system architectures for time-synchronous adaptation of ASR systems based on the ESMERALDA toolkit [6]. For both approaches MLLR-adaptation and *Gaussian Mixture Model (GMM)* based acoustic environment detection are used as proposed in [1]. Focusing on the domain of speech recognition within car environments we present system improvements yielding higher adaptation sensitivity to arbitrary acoustic mismatches beyond speaker changes.

In International Conference on Spoken Language Processing, volume 2, pages 1409-1412, Denver, 2002.

Based on our first system and the above given constraints for online adaptation systems we discuss the mandatory compromise between optimal adaptation improvements and its actual reachability in interactive ASR applications satisfying their essential constraints. We demonstrate the reliability of our second approach for online adapting such a speech recognition system using the wall street journal corpus (WSJ).

This paper is organized as follows. Section 2 gives a short review of adaptation technologies and the methods for acoustic environment discrimination used. Afterwards we introduce our system architectures enhancing the approach described in [1] and overcoming some of its drawbacks. In section 4 we present the evaluation of both systems using two different corpora.

2. BASIC TECHNOLOGY

For adapting the acoustic models of speech recognition systems presently the most promising approach is the technology of *Maximum Likelihood Linear Regression (MLLR)* proposed by Leggetter et al. [7]. Based on regression classes the maximum likelihood optimization for small amounts of adaptation data covering only parts of the acoustic model is generalized to the complete set of parameters. For each regression class a linear transformation including rotational and translational parts is applied to the appropriate means of the mixture densities. In the common case of diagonal and identical covariances of all codebook classes assigned to a regression class and a viterbi-like decoding process the transformation matrix can be established via the least squares regression. Using this procedure significant improvements of recognition accuracy are achievable with only small amounts of adaptation data and acceptable computational effort. The number of regression classes depends on the amount of available adaptation data. For online adaptation procedures generally only few amounts of data are available. Fischer and Stahl showed in [8] the excellent generalization ability of a single regression class using very small amounts of adaptation data.

Gaussian Mixture Models are well suited for the discrimination of acoustic environments as introduced by Reynolds and Rose in [9]. In principal GMMs are single state HMMs thus behaving like mixture classifiers. For classification prior class probabilities are used as emission probability weights and since there is only one state the transition probabilities are trivial.

3. ENHANCED SYSTEM ARCHITECTURES

The principle idea of online decision and adaptation architectures can be described as follows: Starting with a baseline model λ trained independently from special acoustic environments the adaptation is performed for each characteristic change in acoustics e.g. additional background noises or speaker changes. These changes have to be detected automatically. At run time or due to initially pre-loading several adapted models $\tilde{\lambda}$ are created each covering a distinct acoustic environment better than the original model. For each utterance it has to be decided which one of the adapted models or the baseline model is the best. Each time the original model performs better than all existing adapted models the current frame and optimal state are saved for adaptation. Once enough adaptation data are gathered this way the original model is modified using the standard MLLR-framework and a new specialized model $\tilde{\lambda}$ is added to the stock of adapted models. If the acoustic scenario is already known and, therefore, covered by an existing adapted model $\tilde{\lambda}'$, the final recognition process is performed using this special model.

3.1. Multi-pass approach

We developed our first system architecture inspired by the multi-pass approach of Zhang et al.[1]. Holding several adapted models in stock the automatic speaker change detection of their system is realized by the parallel decoding of a given utterance using all adapted models as well as the baseline model. Since the required computational effort is substantial they proposed the use of GMMs for automatic speaker change detection drastically reducing the decoding complexity. While speaker independently establishing the baseline HMM they also create a corresponding generic GMM in the preceding training stage. For every utterance alignments are produced in a first pass for all GMMs including the baseline model. The comparison of their scores yields the currently best fitting model which is used in a second pass for the actual HMM based recognition of the complete utterance.

Our first system in principle follows this approach. Figure 1 shows the resulting adaptation algorithm. As Zhang et al. stated the usage of a baseline GMM is somehow crucial. Indeed we have been unable to produce acceptable decision results based on pure GMM discrimination for noisy environments. The newly created GMMs in most of the applications outperform the generic decision model. This becomes clear since a structurally equal classifier always performs better if trained with the 'test data'. Strictly speaking this is what happens, when establishing new GMMs presenting the current data. Thus a robust scenario identification based on pure GMM evaluations seems only achievable for special cases like speaker change detection but not for arbitrary acoustic adaptation. The discrimination performance of the 'adapted' GMMs among themselves is much better. Because of their common basis the selection of the best *adapted* model works very well. For the final decision we use the scores of both the best adapted HMM ($P(\vec{o}, \vec{s}^*|\lambda')$) and the base model ($P(\vec{o}, \vec{s}^*|\lambda)$) (upper comparison in figure 1). The comparison shown at the bottom of figure 1 is necessary when readapting the models in stock. If this fails the baseline model is used as fall-back for the final recognition process. We have refined the configuration for detecting arbitrary changes in acoustic environments within the ESMERALDA speech recognizers [6] as follows:

- a single global regression class
- 2 000 – 6 000 frames (10 ms each) of adaptation data,
- 10 - 15 GMM-classes created by an algorithm for vector quantizer design.

Using the multi-pass system yields satisfying improvements of online speech recognizers with medium size vocabularies (cf. section 4).

When applying the above described multi-pass architecture to large vocabulary speech recognizers several drawbacks concerning the computational effort become clear. For large lexica it is not feasible to keep the whole sets of active states for each frame in memory in order to retrieve the globally optimal state s_i^* for each frame $\vec{o}_i$ while transcribing the utterance. Even the usage of huge amounts of memory and sophisticated pruning methods does not solve the problem at all: there are too many active states that have to be explored. Following the reasoning in section 1 for interactive applications it is impracticable to pass each utterance at least

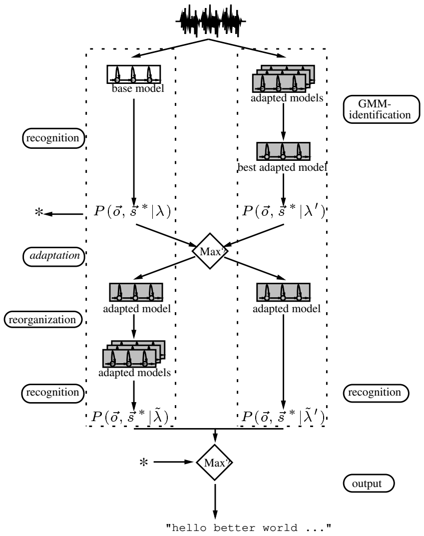

Fig. 1. Multi-pass system architecture

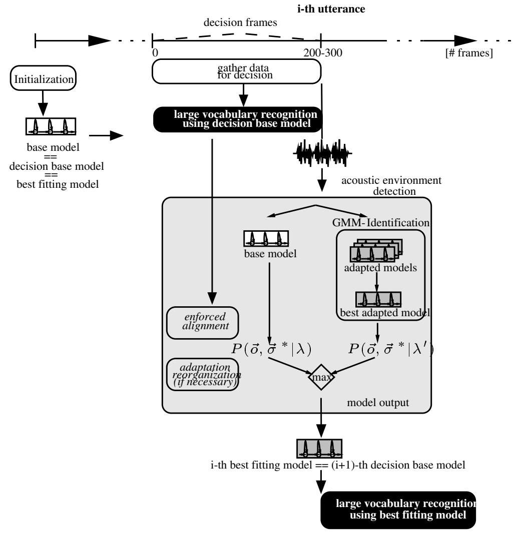

Fig. 2. Single-pass system architecture

twice using a large vocabulary recognition system, first for model decision and second for actual recognition (and transcription for further adaptation).

3.2. Single-pass approach

Keeping in mind the above described drawbacks of the multi-pass approach we developed our second system satisfying the constraints discussed in section 1. The refinements of our multi-pass approach described in the previous section (especially the combined GMM/HMM scenario decision process) have been kept since they provide the ability of arbitrary environment adaptation.

Instead of the most probable state-sequence $\vec{s}^* = \{\vec{s}_i^*\}$ globally determined via the backtracing in the viterbi algorithm we use the sequence $\vec{\sigma}^* = \{\vec{\sigma}_i^*\}$ of *local* best states. These states are retrieved frame-wise and time-synchronously within the forward exploration. Experiments showed that the average loss of adaptation improvement is about two percent (cf. section 4). Compared to the newly attained ability of applying adaptation generally to large vocabulary recognition systems this seems negligible.

Our first method has been modified for becoming a single-pass system architecture as illustrated in figure 2. Its processing principles are shown by means of an exemplary utterance whereas the processing timeline is given at the top of the figure. The shaded boxes represent the computational effort: the darker the background the higher the computational effort is.

As required for interactive applications our second system actually fulfills the single-pass approach. Every utterance is effectively processed only *once* for the final recognition using the full-blown recognizer (inverted boxes). Investigating the optimal amount of data being necessary for a working scenario discrimination we found out that small amounts of data (200 to 300 'decision frames') are already sufficient when using the combined GMM/HMM decision process. For the decision frames the model, determined for the previous utterance as best fitting, represents the model currently used. As long as there are no adapted models this for obvious reasons is the baseline model.

Using the current model the conventional recognition is performed for the decision frames by the large vocabulary recognizer. Its results simultaneously represent the final recognition output for the whole recognition system as well as the decision transcription for the first frames. This transcription is then used as the basis of an enforced alignment of both the best adapted and the baseline HMM for the final HMM decision as described in section 3.1. Performing an enforced alignment gives us the opportunity of using an elementary recognition system as used for medium or small size vocabularies. Once the decision has been established its result is valid for the rest of the utterance and the decision frames of the following utterance. This heuristic model look-ahead is motivated by the following observation: After processing the decision frames in every case the best fitting model for the current utterance is selected. Since in typical domains the ratio of decision frames is small compared to the length of the complete utterance and even suboptimal models are not completely 'out of range' possible failures of our heuristics are negligible. The main part of the utterance is treated using the optimal model.

4. EVALUATION

We tested our systems using two different corpora. First we demonstrate the general capabilities for detecting arbitrary acoustic environments beyond speaker changes using our refined adaptation methods. Therefore the domain of automatic speech recognition in car environments has been chosen. In order to demonstrate the general applicability of our second approach experiments with larger vocabularies have been performed.

The *SLACC (Spoken LAnguage Car Control)* corpus consists of read speech containing instructions (ca. 9 hours for training and about 100 minutes for test) for the control of non safety-relevant functions in car environments, e.g. mobile phone or air-condition. They were recorded in various cars and in different environments e.g. highway or city traffic uttered by several speakers (lexicon size: 658 different words) [10].

Additionally, we tested the systems on the *Wall-Street-Journal* task (WSJ0) [11]. The 5k closed vocabulary speaker independent recognition system used was trained on about 15 hours of speech (the phonotypical transcription of the vocabulary was supplied by “Carnegie Mellon Pronouncing Dictionary” Version 0.6) and tested on 330 utterances with approx. 40 minutes of speech.

Figure 3 summarizes the performance of our combined GMM /HMM approach for detecting the actual acoustic realities. By means of a SLACC experiment using the multi-pass system beside speaker changes there is evidence for the strong sensitivity of the scenario detection approach. The experiment subsumes read utterances of several speakers who have been on two trips in the same car. The first part of the experiment contains the utterances of all speakers on their first trip and the second part includes the utterances of the speakers second trip. Whereas the x-axis represents the processed utterances vertically the models used are plotted. Additionally, at the upper x-axis the particular acoustical environments, namely city and highway traffic are mentioned.

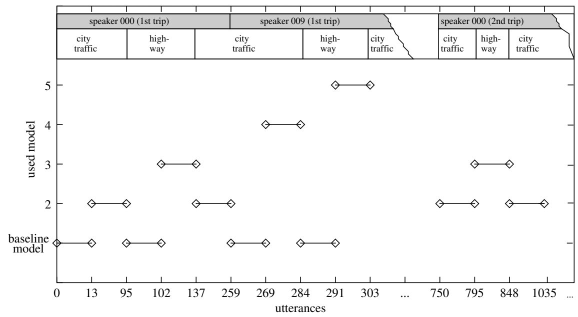

Fig. 3. GMM/HMM decision performance beyond speaker change detection (SLACC-corpus)

It can be seen that beyond the detection of speaker changes for every speaker separate models for city traffic as well as for driving on the highway have been established. Beside this the figure shows the advantage of having several models in stock: Once an adapted model has been created for a special acoustic scenario e.g. speaker 000 at city traffic (first trip) it can be instantaneously reused when the scenario reappears (second trip). Thus without any delays the best fitting model can be used for optimal recognition.

Finally table 1 shows the improvements for large vocabulary speech recognition systems when using our enhanced single-pass approach. The minor loss when changing from optimal state adaptation as originally required by MLLR to local states is shown for an experiment on the SLACC corpus.

	SLACC	WSJ0
WER baseline	23.8%	14.5%
WER $\vec{\sigma}^*$-adaptation	21.8%	12.9%
Δ WER	8.4%	12.4%
WER $\vec{s}^*$-adaptation	21.5%	–
Δ WER	≈2%	–

Table 1. Online adaptation results for both corpora using the single-pass system architecture (additionally for SLACC: adaptation results if using viterbi-states $\vec{s}^*$ instead of local states $\vec{\sigma}^*$)

5. CONCLUSION

We presented system architectures for robust environmental adaptation of continuous speech recognition systems.

Due to a combined GMM/HMM scenario detection the acoustic adaptation could be extended towards arbitrary environmental changes. Experiments within the domain of speech recognition in car environments showed the improved sensitivity of our approach. Satisfying the constraints for online adaptation as needed in e.g. human machine interaction we proposed a second system architecture following the time-synchronous approach. Due to a strict single-pass approach the conflict of immediately benefiting from adaptation improvements could be resolved which was evaluated using two different corpora.

The combination of a powerful scenario detection and the time-synchronous approach yields robust acoustic adaptation for interactive speech recognition applications.

6. REFERENCES

[1] Zhi-Peng Zhang, Sadaoki Furui, and Katsutoshi Ohtsuki, “On-line incremental speaker adaptation with automatic speaker change detection,” in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing*, 2000.

[2] Long Nguyen, Spyros Matsoukas, Jay Billa, Richard Schwartz, and John Makhoul, “The 1999 BBN BYBLOS 10xRT broadcast news transcription system,” in *2000 Speech Transcription Workshop*, Maryland, 2000.

[3] H. Ney, L. Welling, K. Beulen, and F. Wessel, “The RWTH speech recognition system and spoken dokument retrieval,” in *IECON*, Aachen, 1998, vol. 4, pp. 2022–2027.

[4] Vasilios V. Digalakis, “Online adaptation of hidden markov models using incremental estimation algorithms,” in *IEEE Trans. on Speech and Audio Processing*, May 1999.

[5] C. J. Leggetter and P. C. Woodland, “Flexible speaker adaptation using maximum likelihood linear regression,” in *Workshop on Spoken Language Systems Technology*. ARPA, 1995, pp. 110–115.

[6] Gernot A. Fink, “Developing HMM-based recognizers with ESMERALDA,” in *Lecture Notes in Artificial Intelligence*, Václav Matoušek, Pavel Mautner, Jana Ocelíková, and Petr Sojka, Eds., Berlin Heidelberg, 1999, pp. 229–234, Springer.

[7] C. J. Leggetter and P. C. Woodland, “Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models,” *Computer Speech & Language*, pp. 171–185, 1995.

[8] Alexander Fischer and Volker Stahl, “Database and online adaptation for improved speech recognition in car environments,” in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing*, 1999.

[9] D.A. Reynolds and R.C. Rose, “Robust text-independent speaker identification using gaussian mixture speaker models,” in *IEEE Trans. on Speech and Audio Processing*, May 1999, pp. 253–261.

[10] Christoph Schillo, “Der SLACC Korpus,” Tech. Rep., Faculty of Technology, Bielefeld University, 2001.

[11] D. B. Paul and J. M. Baker, “The design for the Wall Street Journal-based CSR corpus,” in *Speech and Natural Language Workshop*. 1992, Morgan Kaufmann.