An Efficient Method for Making Un-Supervised Adaptation of HMM-based Speech Recognition Systems Robust Against Out-Of-Domain Data

Thomas Plötz and Gernot A. Fink

University of Dortmund, Germany,
Robotics Research Institute,
Intelligent Systems Group,
{Thomas.Ploetz,Gernot.Fink}@udo.edu

Abstract. Major aspects of cognitive science are based on natural language processing utilizing automatic speech recognition (ASR) systems in scenarios of human-computer interaction. In order to improve the accuracy of related HMM-based ASR systems efficient approaches for un-supervised adaptation represent the methodology of choice.
The recognition accuracy of speaker-specific recognition systems derived by online acoustic adaptation directly depends on the quality of the adaptation data actually used. It drops significantly if sample data out-of-scope (lexicon, acoustic conditions) of the original recognizer generating the necessary annotation is exploited without further analysis.
In this paper we present an approach for fast and robust MLLR adaptation based on a rejection model which rapidly evaluates an alternative to existing confidence measures, so-called log-odd scores. These measures are computed as ratio of scores obtained from acoustic model evaluation to those produced by some reasonable background model. By means of log-odd scores threshold based detection and rejection of improper adaptation samples, i.e. out-of-domain data, is realized.
By means of experimental evaluations on two challenging tasks we demonstrate the effectiveness of the proposed approach.

1 Introduction

Automatic speech recognition (ASR) represents one important pre-requisite for advanced natural language processing methods. In order to get more detailed insights into human cognition processes the application of sophisticated automatic speech recognition in real-world interaction scenarios has become a standard approach. Consequently, intelligent and most natural human-computer interaction substantially relies on robust ASR systems.

Humans are able to understand spoken language almost independently on the particular acoustic environment where it has been uttered. The key to this remarkable ability lies in adaptation to either different speakers, unfamiliar dialects

* The work described in this paper was partially supported by the German Research Foundation (DFG) within the Collaborative Research Centre "Situated Artificial Communicators" (SFB 360) at Thomas Plötz' former affiliation (University of Bielefeld, Germany, Faculty of Technology, Applied Computer Science Group).

or additional noise. Trying to imitate this behavior, in the last decade robust adaptation techniques of ASR systems based on stochastic models – most notably Hidden Markov Models (HMMs) [1] – have been developed. These techniques allow for the successful application of ASR systems to environments where the acoustic conditions are different from those met during training of the recognizer.

Since the acoustic environment of applications within the context of human-computer interaction, usually, changes almost continuously, rapid and automatic adaptation of ASR systems is of major interest. In these premises, in the last few years especially speaker adaptation based on Maximum Likelihood Linear Regression (MLLR) [2] has been applied successfully. Basically, this procedure offers two advantages:

1. Large amounts of general, i.e. speaker-independent speech data can be used for robustly estimating a base system.
2. Speaker-dependent training data can be exploited as effectively as possible for deriving high-quality speaker-specific recognizers from the original speaker-independent system.

For most interactive applications utilizing natural language processing unsupervised adaptation procedures represent the methodology of choice since labeled sample data uttered by specific speakers is, usually, rarely accessible. In these cases the annotation of adaptation data is obtained on-line from evaluating the base system appropriately. However, if the speaker-independent system provides poor recognition results the quality of the speaker-specific recognizer decreases, too.

Especially within "dynamic" speech-recognition domains, like e.g. in car environments or in the context of human-robot interaction, ASR adaptation is, according to our experiences, likely to fail for various reasons. Most prominently speaker-independent base systems are trained with respect to certain lexica but, naturally, naive ASR users often do not restrict themselves to such limited inventories. Either they do not know the particular lexicon or certain alternative conversation, e.g. to an instructor or another speaker, takes place which is (erroneously) used as adaptation samples. The resulting (false) hypotheses generated for this out-of-domain data by the base-system are then used for adaptation which is counterproductive for ASR improvement.

In order to circumvent this kind of mis-adaptation the hypotheses provided by the speaker-independent base-system need to be judged in some way with respect to their usefulness for adaptation. Basically, this judgment can be obtained by evaluating confidence measures. Those hypotheses which, according to the confidence measure used, are classified as not suitable for proper adaptation should be rejected for speaker dependent specialization.

Within the domain of human-computer interaction computational facilities available for ASR are limited. Prominent examples are recognizers on embedded systems as e.g. in car environments or mobile robots. In addition to this the particular recognition system might need to fulfill certain more or less strict constraints with respect to processing-time. For multi-modal architectures where the speech recognizer is only one part of the overall system the problem further

increases. This implies that complex models can hardly be used for monitoring the actual speaker-adaptation process.

Surprisingly, there is only little literature on hypotheses judgment available focusing on rapid un-supervised adaptation approaches in dynamic environments including computational constraints as outlined above. In [3] confidence measures in terms of posterior word probabilities are exploited in order to "supervise" ASR adaptation. However, related systems, usually, consist of rather complex techniques which seems problematic for restricted computation facilities.

Contrary, in our work we concentrated on the development of a rejection scheme which can be used very efficiently for increasing the robustness of MLLR-based speaker adaptation of ASR-systems. Due to the limited resources available in a typical human-computer interaction scenario our approach proposes the application of a rather simple but effective technique principally known from detection applications utilizing discrete HMMs. The emission probabilities of semi-continuous HMMs are normalized prior to the model evaluation step. Given the transformed parameters the standard single-pass model evaluation results in so-called log-odd scores which can be compared directly to absolute thresholds. Consequently, poorly scoring hypotheses are rejected for adaptation which increases the overall quality of the speaker-dependent recognizer derived. In an experimental evaluation on two challenging recognition tasks simulating the typical dynamic environment as addressed by this paper we demonstrate the effectiveness of our new approach.

2 Related Work

In typical applications where only little sample data is available speaker-related specialization of ASR systems is, usually, limited to the adaptation of the acoustic model, i.e. the mixture parameters of the Hidden Markov Models used. For this purpose, certainly, one of the most promising approaches is the Maximum Likelihood Linear Regression (MLLR) technique.

Based on one or more 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 an affine transformation representing rotations and translations is applied to the appropriate means of the mixture densities. Using MLLR significant improvements of recognition accuracy are achievable with only small amounts of adaptation data. Since the computational effort required is not substantial MLLR has also been used successfully for online applications (cf. e.g. [4,5]).

In order to judge hypotheses obtained from ASR systems confidence measures, usually computed using some additional general-purpose or higher level recognizer, are widely applied. Such judgments can be used for out-of-vocabulary rejection, word spotting etc. [1].

For the first general kind of confidence measures Bayes' rule is directly exploited for the calculation of posterior probabilities $P(\mathbf{w}|\mathbf{x})$ for word hypotheses

$\mathbf{w}$ given the acoustic input $\mathbf{x}$, the acoustic model and the language model probability ($P(\mathbf{x}|\mathbf{w})$ and $P(\mathbf{w})$, respectively):

$$P(\mathbf{w}|\mathbf{x}) = \frac{P(\mathbf{w})P(\mathbf{x}|\mathbf{w})}{P(\mathbf{x})} = \frac{P(\mathbf{w})P(\mathbf{x}|\mathbf{w})}{\sum_{\mathbf{w}} P(\mathbf{w})P(\mathbf{x}|\mathbf{w})} \tag{1}$$

Based on this principle e.g. in [6] posterior probabilities are used as confidence measures for large vocabulary speech recognition. The general-purpose recognizer actually used for computing $P(\mathbf{x})$ is often referred to as filler model and posterior probabilities (cf. equation 1) give reasonable hints for hypotheses' confidences.

Alternatively recognition hypotheses can be judged by confidence measures which are computed separately by (more or less heuristically) combining the results of various so-called predictor models. As an example in [7] several numerical values, obtained e.g. from parallel N-best evaluation or alternative, less complex recognizers (e.g. "phone-only" decoding), have been exploited.

The majority of applications utilizing confidence measures is directed to the general improvement of the recognition accuracy of ASR systems aiming at e.g. more robust dialogue control. As one example in [8] confidence scores are applied in order to improve hands-free speech based navigation in continuous dictation systems. However, there is hardly any literature addressing the confidence measure based improvement of un-supervised speaker-adaptation. In [3] a two-pass adaptation strategy based on word posterior probabilities used as confidence measures is described. Given confidence scores for the hypotheses obtained from the first recognition pass a word graph is generated which is the base for calculating word posterior probabilities. All frames for a word with low confidence are rejected for adaptation.

The authors report improvements for the recognition accuracy of speaker-dependent ASR systems derived by adaptation automatically supervised by evaluating confidence measures. However, the rather complex computation of confidences seems problematic for applications as addressed by this paper which aims at a simpler rejection model.

3 Log-odd scores based rejection

When addressing automatic speech recognition in interactive applications where computational facilities are rather limited certain constraints need to be respected. This includes the best possible avoidance of multi-pass recognition procedures as well as the restriction to approaches which are computationally simple but still provide reasonable recognition results. The situation gets even more problematic when ASR represents only a single component of a multi-modal processing framework.

Nevertheless, especially for dynamic application domains with changing speakers and additional conversation out of the recognizer's scope, i.e. lexicon and acoustic conditions, additional robust un-supervised ASR adaptation is a very important but challenging task.

In these promises we apply MLLR-based adaptation to speaker-independent recognizers within an online adaptation framework (cf. [4] for details). Recognition results obtained by evaluating the base system are used for acoustic adaptation towards speaker specific models. Comparable to the procedure presented in [3] adaptation is monitored by confidence measures, i.e. based on threshold comparison non-confident hypotheses are rejected for MLLR adaptation. In our approach, however, confidence measures are computed using a normalization technique applied to the scores obtained from evaluating the acoustic model. Since this normalization can be applied in advance (see below) its computational effort is almost neglectable.

Basically, raw scores $P(\mathbf{x}, \mathbf{s}|\lambda_{\mathbf{w}})$ obtained from Viterbi alignments of speech data $\mathbf{x}$ to a particular acoustic model $\lambda_{\mathbf{w}}$ along the most probable state path $\mathbf{s}$ could already be used as confidence measures. However, these scores are dependent on the lengths of the particular utterances which prevents direct exploitation.

Inspired by their successful use in alternative applications, e.g. within the bioinformatics domain (cf. e.g. [9]), aiming at the detection of certain patterns modeled by HMMs we use so-called log-odd scores as confidence measures. Originally developed for discrete HMMs the basic idea is to normalize the particular emission probabilities $b_j(o_k)$ of an HMM state j for generated symbols o_k to some reasonable background distribution $P(o_k)$:

$$b'_j(o_k) = \frac{b_j(o_k)}{P(o_k)} \tag{2}$$

When using semi-continuous (SC)HMMs [10] mixtures $\mathcal{N}(\mathbf{x}|\mu_k, \mathbf{C}_k)$ are shared between all HMM states j and individually weighted by c_{jk} resulting in modified emission probabilities $b_j(\mathbf{x})$:

$$b_j(\mathbf{x}) = \sum_{k=1}^{K} c_{jk} \mathcal{N}(\mathbf{x}|\mu_k, \mathbf{C}_k) \tag{3}$$

Technically, SCHMMs can be interpreted as discrete HMMs containing an integrated "soft" vector quantizer where c_{jk} represent the emission probabilities of the discrete model weighted by means of the density values $\mathcal{N}$. In order to obtain confidence measures these coefficients c_{jk}, and thus implicitly the actual emission probabilities $b_j(\mathbf{x})$, are now (cf. equation 2) normalized with respect to certain background distribution. Furthermore, the scores are converted into the negative log-domain resulting in log-odd scores for SCHMMs:

$$b'_j(\mathbf{x}) = -\sum_{k=1}^{K} \ln \frac{c_{jk}}{P(c_{jk})} + \ln \mathcal{N}(\mathbf{x}|\mu_k, \mathbf{C}_k) \tag{4}$$

Normalizing the emission probabilities implies a length normalization of the acoustic scores $P(\mathbf{x}, \mathbf{s}|\lambda_{\mathbf{w}})$. Thus, the resulting scores can be compared directly to an absolute threshold. Those hypotheses scoring worse than this threshold are rejected for adaptation.

Basically, the background distribution $P(c_{jk})$ used for computing log-odd scores corresponds to some sort of "random" model. We evaluated two different random models with respect to their usefulness for computing confidence measures. First, all emission probabilities are normalized with respect to a uniform state-specific background distribution of the mixture weights c_{jk} – referred to as *Flat* background model. The second type of background model consists of the prior probabilities of the mixtures the semi-continuous HMMs are based on. This corresponds to reasonably biasing the random model to the actual statistics of the underlying mixture model and is referred to as *Prior* model.

Technically the computation of confidence measures based on log-odd scores can be performed very efficiently. The usual calculations are just to be enhanced by the additional but non-complex normalization step. Furthermore, this normalization can be performed implicitly prior to any actual calculations, i.e. without increasing the computational complexity for model evaluation. Therefore, the mixture weights c_{jk} of all states of the acoustic models are converted using the normalization term of equation 4 and standard HMM evaluation is performed using the converted weights.

4 Experimental Evaluation

In order to demonstrate the effectiveness of our new log-odd scores based rejection model aiming at improved MLLR adaptation of speaker-independent ASR systems within dynamic environments we performed various experiments. For better reproducibility and generalization we report here the results of systematic evaluations using combinations of known datasets to realistically cover the dynamic environments as addressed by this paper. This includes sets of speaker dependent adaptation samples mixed with additional utterances out of the original recognizers' scope (language, lexicon etc.).

In fact this corresponds to a typical scenario for human-computer interaction with naive users in the loop. Often certain utterances containing e.g. out of vocabulary words are mixed with utterances which can actually be used for acoustic adaptation. The first kind of utterances might originate from conversation with an instructor or from the "learning" phase of the human user not being aware of the actual lexicon of the ASR system. According to our practical experiences with the interaction of humans and a mobile robot [11] the quality of speaker adaptation drops substantially without detection and proper treatment of utterances actually not suitable for adaptation.

4.1 Datasets

The first set of experiments is directed to speaker adaptation within car environments. Therefore, the *SLACC (Spoken LAnguage Car Control)* corpus [12] has been exploited. It 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 different cars and environments (highway or city traffic) by several speakers (lexicon size: 856 words). MLLR adaptation is performed separately for three speakers where the particular adaptation sets consist of SLACC utterances as well as of speech data originating from a completely different corpus, namely from the *Wall-Street-Journal (WSJ0)* task [13]. Non-SLACC utterances have been selected randomly, and the ratio of WSJ0- to SLACC-adaptation data is approximately 1:3 per speaker. Adapted recognizers are evaluated on speaker-specific test data from SLACC (about 300 utterances per speaker).

Additionally, we tested our new approach directly on the WSJ0-corpus. The 5k closed vocabulary speaker independent recognition base-system used was trained on about 15 hours of speech and (in summary) tested on 330 utterances with approximately 40 minutes of speech. For each of the 8 speakers included the official adaptation sets have been mixed with utterances randomly chosen from the SLACC-corpus (see above) as well as from the VERBMOBIL corpus [14]. Again, the ratio of out-of-domain and in-domain sample data is approximately 1:3.

4.2 Results

For all experiments performed the particular training data is used in order to set up a speaker-independent recognition system consisting of semi-continuous HMMs and mixture densities with diagonal covariances. Furthermore, for the WSJ0-system a Bi-gram model is incorporated. The annotation necessary for MLLR adaptation is obtained automatically using the particular base system. Given the confidence measures computed as log-odd scores based on the acoustic model evaluation, and the comparison to an absolute threshold the resulting hypotheses are either respected for adaptation or explicitly rejected. Note that, actually, hypotheses are rejected which not necessarily corresponds to the rejection of complete utterances. All experiments have been performed using our own HMM toolkit ESMERALDA [15].

In tables 1 and 2, respectively, the results for the experimental evaluation of the speaker-dependent systems obtained using MLLR adaptation and our new rejection model based on log-odd scores are summarized.[1] The figures reported have been averaged over speaker-wise evaluations.

The application of our rejection model based on the simple log-odd scores based confidence measures substantially improves the recognition accuracy of speaker-dependent recognizers derived on poor adaptation sets. When rejecting all hypotheses for adaptation with log-odd scores larger than -100 the word-error rate (WER) can be reduced by more then 18% relative for the SLACC-task (compared to standard adaptation without rejection). When using this absolute threshold for the WSJ0-task the WER decreases by more than 16% relative.

[1] Note that baseline experiments without any adaptation showed no significant differences between results obtained when using log-odds scores or their un-normalized counterparts.

Threshold for Rejection $[-\ln(P(O\|\lambda))]$	WER [%]	Δ WER [%]	Rejection [%]
No Rejection (Base)	27.7	–	–
Flat Background Model			
-100	**22.6**	**-18.4**	**87.0**
-50	24.7	-10.8	63.3
0	27.7	0.0	7.00
Prior Background Model			
-100	**22.6**	**-18.4**	**87.7**
-50	25.1	-9.4	62.7
0	27.8	+0.4	5.6

Table 1. Results for SLACC-based evaluation (WER for "clean" adaptation sets without rejection: 20.9%).

Background Model – threshold: -100 –	WER [%]	Δ WER [%]	Rejection [%]
None (Base without Rejection)	15.1	–	–
Flat	12.7	-15.9	70.4
Prior	**12.6**	**-16.6**	**71.4**

Table 2. Results for WSJ0-based evaluation using the rejection threshold determined in SLACC-based experiments.

The differences between the particular background models are almost negligible. For both Prior- and Flat-background model the abovementioned substantial reductions in WER can be reached.

Comparing the results obtained when processing poor adaptation sets using our rejection model to those related to model specialization exploiting "clean" adaptation samples, the effectiveness of our newly developed approach becomes manifest. As demonstrated for SLACC-experiments when applying our rejection model on poor adaptation data the results for optimal conditions, i.e. using "clean" adaptation data, can almost be reached. Thus, even in dynamic environments as addressed by this paper, speaker-dependent recognition systems can be obtained very robustly and efficiently by applying MLLR adaptation and the new rejection model based on log-odd scores to a speaker-independent base system.

5 Summary

The existence of robust automatic speech recognition systems is of major importance for probably all kinds of natural language processing research within

cognitive science. As great ideal human listeners are able to understand speech even in noisy environments or from unknown speakers.

MLLR adaptation has been established as state-of-the-art for the improvement of the recognition accuracy of automatic speech recognition systems based on HMMs. However, according to our experiences within the domain of human-computer interaction, especially in dynamic application domains with naive speakers in the loop adaptation may fail due to improper adaptation data. This data often originates from conversation out of the particular recognizer's scope, i.e. containing words beyond the lexicon of the original recognition system or utterances of poor acoustic quality. Unfortunately, when ASR represents only one part of a complex (e.g. multi-modal) interaction system computational facilities for adaptation are rather limited.

Addressing such interactive applications with computational restrictions we presented a rejection model based on the evaluation of confidence scores applying log-odd scores for semi-continuous Hidden Markov Models. For this simple normalization technique, which can be applied in advance thus maximally limiting the additional computational effort, the ratio of actual acoustic scores as obtained by HMM evaluation to a reasonable background model is computed. Based on these values hypotheses' scores can directly be compared to an absolute threshold and rejected for adaptation if necessary. Two variants of the background model based either on a uniform distribution of mixture coefficients involved or their prior probabilities have been investigated.

The effectiveness of our approach has been demonstrated by means of experimental evaluations on two challenging tasks. Therefore, MLLR adaptation has been applied to speaker-independent base systems processing data sets containing both in-domain and out-of domain utterances. This corresponds to a very common scenario in e.g. human-robot interaction where often adaptation data out-of-the-scope of the recognizer (lexicon etc.) need to be processed. When applying the rejection model the adaptation process for interactive speech recognition applications can be improved substantially.

References

1. Huang, X., Acero, A., Hon, H.: Spoken Language Processing – A Guide to Theory, Algorithm, and System Development. Prentice Hall PTR (2001)
2. Leggetter, C.J., Woodland, P.C.: Maximum likelihood linear regression for speaker adaptation of continuous density Hidden Markov Models. Computer Speech & Language (1995) 171–185
3. Pitz, M., et al.: Improved MLLR speaker adaptation using confidence measures for conversational speech recognition. In: Int. Conf. Spoken Lang. Proc. (2000)
4. Plötz, T., Fink, G.A.: Robust time-synchronous environmental adaptation for continuous speech recognition systems. In: Int. Conf. Spoken Lang. Proc. Volume 2. (2002) 1409–1412
5. Zhang, Z., Furui, S., Ohtsuki, K.: On-line incremental speaker adaptation with automatic speaker change detection. In: Proc. Int. Conf. on Acoustics, Speech, and Signal Processing. (2000)

6. Wessel, F., Schlüter, R., Macherey, K., Ney, H.: Confidence measures for large vocabulary continuous speech recognition. IEEE Trans. on Speech and Audio Processing **91** (2001)
7. Chase, L.: Word and acoustic confidence annotation for large vocabulary speech recognition. In: Proc. European Conf. on Speech Communication and Technology. (1997)
8. Feng, J., Sears, A.: Using confidence scores to improve hands-free speech-based navigation in continuous dictation systems. ACM Transactions on Computer-Human Interaction **11** (2004) 329–356
9. Durbin, R., Eddy, S., Krogh, A., Mitchison, G.: Biological sequence analysis: Probabilistic models of proteins and nucleic acids. Cambridge University Press (1998)
10. Huang, X.D., Jack, M.A.: Semi-continuous Hidden Markov Models for speech signals. Computer Speech & Language **3** (1989) 239–251
11. Haasch, A., et al.: BIRON – The Bielefeld Robot Companion. In: Proc. Int. Workshop on Advances in Service Robotics, Fraunhofer IRB Verlag (2004) 27–32
12. Schillo, C.: Der SLACC Korpus. Technical report, Faculty of Technology, Bielefeld University (2001)
13. Paul, D.B., Baker, J.M.: The design for the Wall Street Journal-based CSR corpus. In: Speech and Natural Language Workshop. (1992)
14. Kohler, K., et al.: Handbuch zur Datenaufnahme und Transliteration in TP 14 von VERBMOBIL – 3.0. Technical Report 11, Institut für Phonetik und digitale Sprachverarbeitung, Universität Kiel (1994)
15. Fink, G.A.: Developing HMM-based recognizers with ESMERALDA. In: Text, Speech and Dialogue. Volume 1692 of Lecture Notes in Artificial Intelligence. Springer, Berlin Heidelberg (1999) 229–234