In **2004: A Speaker Odyssey – The Speaker and Language Recognition Workshop**, Toledo, 2004, pages 185–192.

Integrating Speaker Identification and Learning with Adaptive Speech Recognition

Gernot A. Fink, Thomas Plötz

Faculty of Technology
Bielefeld University, Germany
{gernot,tploetz}@techfak.uni-bielefeld.de

Abstract

Presently, speaker adaptive systems are the state-of-the-art in automatic speech recognition. A general baseline model is adapted to the current speaker during recognition in order to improve the quality of the results obtained. However, the adaptation procedure needs to be able to *distinguish* between data from different speakers. Therefore, in a general speaker adaptive recognizer *speaker recognition* has to be performed implicitly. The resulting information about the identity of the person speaking can be of great importance in many applications of speech recognition, e.g. in man-machine communication.

Therefore, we propose an integrated framework for *speech and speaker recognition*. Our system is able to detect new speakers and to identify already known ones. For a new speaker both an identification and an adapted recognition model are learned from limited data. The latter is then used for the recognition of utterances attributed to this speaker. We will present evaluation results with respect to speaker identification performance on two non-trivial speech recognition tasks that demonstrate the effectiveness of our integrated approach.

1. Introduction

For the purpose of speaker recognition, in principle, the speech content, i.e. the words and phrases actually spoken, conveys no relevant information[1]. The distinction between speakers is achieved by exploiting the different characteristics of speaker specific realizations of speech sounds. However, taking into account the speech content can substantially improve the quality of speaker recognition, as then much more detailed speaker specific models can be built. So-called *text dependent* speaker recognition can trivially be achieved if the prompting text to be uttered by a person is known beforehand as is the case in some speaker verification applications. Otherwise the speaker independent recognition of unconstrained speech would be necessary – a task which even with today's technology can't be solved in general. Therefore, most current speaker recognition systems operate in so-called *text independent* mode, i.e. they do not make use of any information about the speech content and are, consequently, very flexible with respect to their application.

For the purpose of speech recognition, i.e. the task of recovering a textual transcription of the speech content, in principle, the identity of the speaker is irrelevant. The spoken words are basically recognized from the speaker independent characteristics of speech sounds. Today, the majority of speech recognizers are so-called *speaker independent* systems, which means that they work rather well for a wide variety of speakers by modeling the expected variability in the realization of speech sounds. As this variability is large across different speakers and relatively small for a specific person, taking into account the identity of the speaker can substantially improve the recognition quality. However, in order to train such so-called *speaker dependent* models a large amount – i.e. several hours – of speech data needs to be available for the person in question, which is prohibitive for the majority of applications.

In order to be able to benefit from speaker specific modeling in applications as e.g. telephony-based services, speaker adaptation techniques were developed. Starting from a speaker independent system the model parameters are modified appropriately in order to better reflect the characteristics of a special speaker. Because a speaker independent system is used as a baseline the required amount of speaker specific speech data can be reduced substantially. By defining "clusters" of related model parameters that are to be modified similarly only a few seconds of speech can be sufficient for estimating a speaker specific model.

In contrast to telephony-based services, where a single speaker can be assumed per call, in more general applications as e.g. broadcast-news transcription or man-machine communication the identities of speakers can not be obtained that easily. Consequently, it is not clear on what data to adapt a specific model and on what speech to use it for recognition later. Therefore, general speaker

[1] Here we adopt a purely signal processing point of view ignoring the potential contribution of ideosyncrasies at the language level that can become relevant if larger speech corpora are analyzed.

adaptive systems also need to provide methods for detecting changes of speakers, which can be considered as a special case of un-supervised speaker identification.

If, however, adaptive speech recognition *requires* speaker identification to be performed implicitly, it will be beneficial to solve both tasks in an integrated framework where they can optimally complement each other. Such an integrated speech and speaker recognition system is then not only able to exploit speaker specific modeling in an optimal way for enhancing recognition performance, but it is also able to identify the person speaking.

This capability is extremely useful in the communication of several people with intelligent devices as e.g. a robotic assistant. In such a scenario it needs to be clear, which person requested some information, initiated some action, or is entitled to teach the robot some new behavior. Consider for example a robotic assistant in a hospital. It should probably provide general information about the location of facilities to everybody, but accept directions only from staff. Especially, only doctors should be allowed to order medications. Clearly, a robotic companion like this will always be equipped also with additional sensors that make person identification possible, e.g. based on visual data. Nevertheless, speech based identification can complement such identification processes or even substitute them in case of e.g. occlusions.

In this paper we will present an integrated speech and speaker recognition system intended for the use in a human-robot interaction scenario. In the following section we will briefly review some relevant related work. Then we will in detail describe the proposed integrated recognition framework. Evaluation results will be presented in section 4. A discussion of the capabilities and limitations of the proposed approach and a short summary conclude the paper.

2. Related Work

The general problem of speaker recognition can be further subdivided into the tasks of identification and verification. In speaker identification it has to be decided for a given speech sample which one of several known speakers produced it. In contrast, speaker verification only decides whether the speech sample originates from one designated speaker or not. Both tasks become much easier if the textual content of the speech sample investigated is known beforehand. However, due to their greater flexibility the majority of speaker recognition systems today operates in so-called *text independent* mode. This means that no side information about the spoken words or phrases is used during the decision process.

For modeling individual speakers' speech characteristics mainly so-called *generalized mixture models* (GMMs) are used [1]. These models operate on a suitable representation of speech as a sequence of feature vectors $\boldsymbol{X} = \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots \boldsymbol{x}_T$. The basis of a GMM Γ_i is formed by a Gaussian mixture model that describes a probability density in the space of feature vectors. The GMM Γ_i then defines a simple approximation for the cumulative density of utterance $\boldsymbol{X}$ originating from speaker i by computing the product of the baseline mixture model for all feature vectors $\boldsymbol{x}_t$

$$p(\boldsymbol{X}|\Gamma_i) = \prod_{t=1}^{T} \sum_{k=1}^{K_i} c_{ik} \mathcal{N}(\boldsymbol{x}_t|\boldsymbol{\mu}_{ik}, \boldsymbol{C}_{ik})$$

where the c_{ik} are the weights of the individual Gaussians with mean vectors $\boldsymbol{\mu}_{ik}$ and covariance matrices $\boldsymbol{C}_{ik}$.

Given a set of GMMs for each known speaker the task of speaker recognition can then be solved by deciding for speaker j whose associated model Γ_j maximizes the above probability density:

$$j = \underset{i}{\operatorname{argmax}}\, p(\boldsymbol{X}|\Gamma_i)$$

In order to be able to handle rejections an additional *background model* Γ_0 is used. This model can be very complicated in verification applications. A rather simple solution, which is often applied for speaker identification tasks, is to use a general GMM trained on data of a large number of speakers as the background model. Good reviews of current speaker recognition technology can be found in [2] or [3].

A classical theoretical framework for describing the basic modeling aspects of statistical speech recognition is the so-called *channel model* proposed by Jelinek and colleagues [4]. First, a speaker mentally formulates a word sequence $\boldsymbol{w}$ to be uttered with some probability $P(\boldsymbol{w})$. Then the word sequence is articulated, transmitted over a communication channel, and on the listeners side converted to some feature representation $\boldsymbol{X}$. This process can be described by a probability density $p(\boldsymbol{X}|\boldsymbol{w})$. Now the goal of the speech recognizer is to recover the original message $\boldsymbol{w}$ from the sequence of feature vectors $\boldsymbol{X}$. In a statistical sense the optimal solution to this problem is to compute the word sequence $\hat{\boldsymbol{w}}$ that maximizes the posterior probability $P(\boldsymbol{w}|\boldsymbol{X})$, which can be rewritten in terms of $P(\boldsymbol{w})$ and $p(\boldsymbol{X}|\boldsymbol{w})$:

$$\hat{\boldsymbol{w}} = \underset{\boldsymbol{w}}{\operatorname{argmax}} P(\boldsymbol{w}|\boldsymbol{X}) = \underset{\boldsymbol{w}}{\operatorname{argmax}} \frac{P(\boldsymbol{w})\, p(\boldsymbol{X}|\boldsymbol{w})}{p(\boldsymbol{X})}$$

By exploiting the fact that the probability of feature vector sequences *per se* $p(\boldsymbol{X})$ is a constant with respect to the maximization operation the above equation can be further simplified:

$$\hat{\boldsymbol{w}} = \underset{\boldsymbol{w}}{\operatorname{argmax}} P(\boldsymbol{w})\, p(\boldsymbol{X}|\boldsymbol{w})$$

From this formulation of the speech recognition task we can see that two modeling components are used. First,

there is a model describing the acoustic realization of a given word or word sequence $p(\boldsymbol{X}|\boldsymbol{w})$ which is usually realized by a *Hidden Markov Model* (HMM). An HMM is – roughly speaking – an extension of a GMM adding a memory component to the model which is realized by a set of discrete internal states. A GMM can be viewed as a simple HMM with only a single state. The second modeling component $P(\boldsymbol{w})$ is called the language model and is used to restrict the search for possible word hypothesis sequences to plausible solutions with respect to the application in question. Statistical language models are usually realized by so-called *n-gram* models. An excellent treatment of statistical speech recognition techniques can be found in [5]. A short introduction to the field with the focus on decoding an existing model is given in [6].

Most current speech recognition systems are *speaker independent* which means that the acoustic models used – i.e. the HMMs – are trained on data from a large variety of speakers. Therefore, those systems usually perform reasonably well for a broad class of persons. A much more detailed modeling and, consequently, an improved recognition quality can be achieved by so-called *speaker dependent* systems. As those systems require a substantial amount of speech data from the person they are tailored to, their use is limited to only a few application areas as e.g. personal dictation systems.

The gap in recognition accuracy achieved by speaker independent or speaker dependent modeling is closed by model adaptation techniques. Given a speaker independent baseline system and data from a specific speaker the parameters of the acoustic model can be re-estimated to produce a recognition system specialized for the person in question. If enough data is available – i.e. several hours of speech – there is no principle difference between adaptation and training of HMMs. However, in practical applications adaptation data will always be limited and, therefore, classical training algorithms for HMMs will either fail completely or not produce sufficiently specialized parameter sets.

Therefore, the invention of *maximum likelihood linear regression* (MLLR) by Leggetter and Woodland [7] can be considered a major breakthrough in the field. The principle idea of the method is to define groups of model parameters that are modified by a common rule during the adaptation process. As a modification rule can have substantially fewer degrees of freedom than the total number of parameters it affects, the amount of samples required for parameter estimation is also greatly reduced. In MLLR the parameter groups are called *regression classes*. The parameters associated with each class are subject to a common affine transformation. For simplicity MLLR only transforms the mean vectors $\boldsymbol{\mu}_{ik}$ of the Gaussian mixtures within an HMM. The remaining parameters which are considered to be of lesser importance for the overall modeling quality are left unchanged. By condensing the re-estimation of model parameters into a small set of parameter transformations MLLR is able to produce adapted models on only a few seconds of speech which are comparable in recognition accuracy with purely speaker dependent models.

A problem which is mostly neglected in speaker adaptation is how different speakers are identified. In so-called batch adaptation a set of adaptation and testing data is given per speaker. Even in most online systems that perform adaptation directly during recognition speaker changes are usually available as an external side-information. The first approach to speaker adaptation that integrated the detection of speaker changes was proposed by Zhang and colleagues [8]. For every known speaker a GMM was used as a speaker model. However, the authors had to admit that the speaker recognition did not work reliably enough during online recognition. Therefore, the first online adaptive speech recognizer with online speaker change detection was proposed by ourselves [9] and forms the baseline for the proposed integrated framework.

In [10] also a speaker model was integrated into the recognizer, however, with the goal to improve the recognition quality for extremely short phrases by jointly evaluating both models. The aim of identifying speakers with speech recognition technology was pursued in [11]. The authors used a hybrid system combining HMMs and artificial neuronal networks. Due to peculiarities of the hybrid system they were only able to estimate verification models that could decide whether an utterance was from a given speaker or not. Additionally, for the training of these models large amounts of speaker specific data needed to be available as a complete hybrid recognition model was estimated.

3. Integrated Speech and Speaker Recognition

The most straight-forward way of exploiting speech recognition technology for speaker recognition would be to use speaker dependent acoustic models for each known speaker. Unfortunately, this approach is totally impractical as too much training data is required per speaker and as too much computational load is generated by evaluating all available acoustic models. Therefore, our approach aims at low computational complexity and at requiring as little speaker specific data as possible.

We use a speaker independent acoustic model as a baseline[2]. This model is adapted to new speakers us-

[2] Additionally, we make use of a statistical language model if such a model is available for the recognition task in question. This modeling component is, however, not adapted to a specific speaker but remains constant throughout all processing phases. Therefore, for reasons of simplicity, we will not mention the optional use of a statistical language model in the description of the integrated speech and speaker recognition algorithm.

ing MLLR. Furthermore, for every speaker detected a speaker model realized as a GMM is estimated which will be used for the identification. The rejection of speakers as new or unknown ones is robustly achieved by comparing the scores of the speaker independent acoustic model with the one delivered by the best matching adapted system. In the following this algorithm will be described in more detail.

0. Initialization: The only thing which needs to be available beforehand is a speaker independent acoustic model λ_0. If some speakers should be known to the system *a priori* then for each of these an identification model – a GMM Γ_i – and an adapted acoustic model λ_i need to be available too.

For every new utterance $\boldsymbol{X} = \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots \boldsymbol{x}_T$ the following processing steps are executed:

1. Identification I: First, an initial segment $\boldsymbol{Y} = \boldsymbol{x}_1, \ldots \boldsymbol{x}_n$ of the complete utterance $\boldsymbol{X}$ is selected. If there are already some speakers known to the system the best matching speaker k is identified among these based on $\boldsymbol{Y}$ using the existing speaker models Γ_i. Note that in this phase of processing no rejections are allowed. The detection of new speakers will be handled separately and will be described below. The length of the utterance prefix $\boldsymbol{Y}$ needs to be long enough to allow a robust identification procedure and at the same time as short as possible in order not to introduce too much processing delay. Currently, we use a prefix length between 2 and 3 seconds of speech.

2. Recognition I: In the second processing step recognition results – i.e. the optimal word hypothesis chains – are computed on the utterance prefix $\boldsymbol{Y}$. The decoding of the speaker independent acoustic baseline model λ_0 delivers an associated score y_0 for the recognition result on $\boldsymbol{Y}$. If some speaker k was selected in the previous processing step the associated speaker adapted model λ_k is also decoded yielding a recognition score y_k.

3. Identification II: If the recognition score y_k achieved with the speaker adapted acoustic model is better than the score y_0 delivered by the baseline model the current utterance is attributed to the known speaker k. Otherwise the utterance is assumed to be produced by some unknown new speaker m. Note that this is also the case if no speakers are known to the system and, therefore, no identification models or adapted acoustic models exist.

4. Learning & Adaptation: This processing step is only necessary, if in the previous phase it was decided, that an utterance originating from a yet unknown speaker m was detected. In this case a new identification model – i.e. a new GMM Γ_m – and a new speaker adapted acoustic model – an HMM λ_m – are created. These models then need to be estimated on appropriate data starting with the current utterance $\boldsymbol{X}$. In order to achieve a reasonable compromise between fast availability of these models – the new speaker can only be detected implicitly via the rejection criterion until a dedicated identification model is available – and robust estimation currently 60 seconds of speech are required for training both the identification and the adapted acoustic model. As this amount of data will in general span several shorter utterances the estimation of speaker specific models is carried out in parallel to the general course of the main algorithm. For the estimation of the speaker identification GMM Γ_m we apply the k-means algorithm [12] directly to the sequence of feature vectors. The number of clusters to be estimated typically ranges between 8 and 32 and needs to be determined empirically. The adapted acoustic model λ_m is created from the baseline model λ_0 by first estimating the transformation parameters on the feature data via MLLR [7] and then applying the resulting model transformation to λ_0.

5. Recognition II: Up to this point recognition results are only available for the initial utterance prefix $\boldsymbol{Y}$. The task of this final processing step is now to deliver a recognition result for the complete utterance using the best available acoustic model. There are two major configurations that have to be considered: First, the current utterance was possibly rejected in the second identification step and is, therefore, attributed to a new speaker m. As training a speaker adapted model requires more data than the utterance prefix the availability of a speaker adapted model may be delayed at least until the second utterance of the new speaker. In general, however, an adapted acoustic model will be available only after some few utterances that are used for model estimation. Until then the best available model for recognition is the baseline speaker independent model λ_0. In the second case the current utterance prefix was successfully identified to be produced by a known speaker k. As for this speaker an adapted acoustic model λ_k already exists it can immediately be used to produce the best possible recognition results. Note that for both models – the baseline one and the one adapted to speaker k – recognition hypotheses are available for the first part of the utterance. These results can now simply be extended to cover the complete utterance $\boldsymbol{X}$ thus saving a repeated processing of the initial part.

4. Results

We experimentally evaluated our integrated speaker identification approach using two different corpora. They cover two different acoustic environments that are typical for speaker and speech recognition tasks: quiet office environment and noisy surroundings in a car. For both scenarios we performed two types of evaluation: In the

first type of experiments the speaker identification models Γ_i as well as the speaker adapted acoustic models λ_i were estimated beforehand on rather large sets of speaker specific enrollment data. In contrast, these speaker specific models were automatically learned online using significantly smaller amounts of adaptation data in the second type of experiments.

In all recognition and identification experiments speech samples were represented as sequences of 39-dimensional feature vectors consisting of one energy coefficient and 12 cepstral coefficients together with their first and second smoothed derivative. In order to remove effects of different recording channels and environments we apply cepstral mean subtraction. For acoustic modeling we used semi-continuous HMMs with tri-phone subword units. A robust estimation of the model parameters is achieved by automatically creating state clusters using an entropy-based similarity criterion.

The graphs in figure 1 show the results of the evaluation for both types of experiments. The x-axis represents the sequence of utterances forming the test-set. The IDs of the test speakers are plotted along the y-axis. Therefore, a consecutive part of the test data uttered by the same speaker is shown as a grey horizontal bar. The speaker IDs detected automatically are inserted into the graphs as small filled triangles at the appropriate position. If more than one speaker model was created in the online learning mode detections based on such additional models are shown with a small vertical offset. Rejected utterances are marked similarly in the lowest row of the graph which is labeled "rejection".

4.1. Experiments on Clean Speech

Initially, we tested our approach on the *Wall-Street-Journal* task (WSJ0) [13]. A 5k closed vocabulary speaker independent recognition system 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 approximately 40 minutes of speech. The acoustic models use a shared codebook of 1400 Gaussians with diagonal covariances and tri-phone sub-word units with approximately 5500 state clusters. From this baseline speaker independent model the predefined speaker adapted acoustic models were created by carrying out an MLLR adaptation on the official adaptation sets. The same data was used for establishing the GMMs (22 Gaussians each) required for speaker identification via the k-means algorithm.

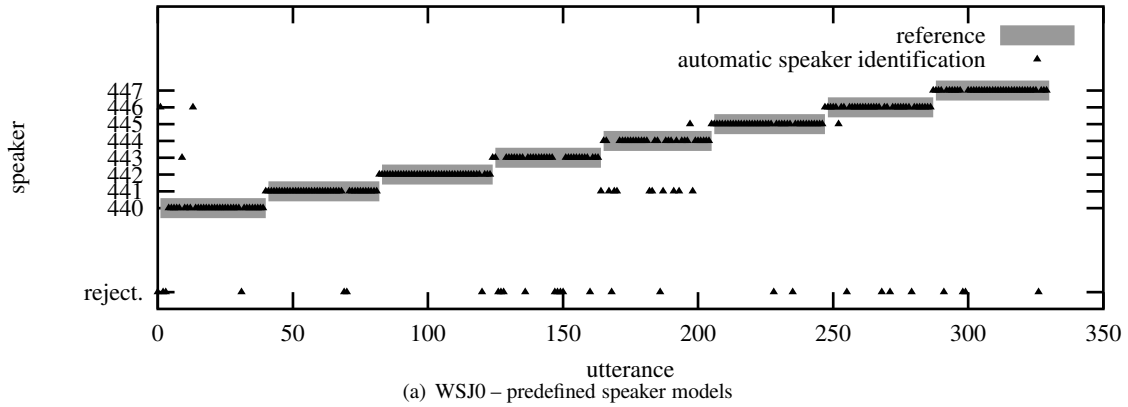
Figures 1(a) and 1(b) illustrate the performance of the speaker identification for the WSJ0 task. In the upper graph 1(a) the identification results are shown for the first type of experiment. Using sufficient amounts of adaptation data specialized HMMs as well as GMMs were created for every speaker before performing the evaluation. In this setup the speaker recognition performance quite satisfactory. For the 330 test utterances of 8 speakers the classification error is below 5% and less than 9% of all utterances are falsely rejected. Even for the more difficult task of learning the speaker dependent models online, illustrated in the lower graph 1(b), the classification error remains at the same low level of approximately 5%, though only rather small amounts of adaptation data – about 60 seconds per speaker[3] – were available for establishing the speaker specific models. Compared to the first experiment where speaker dependent models were created offline on significantly larger adaptation sets, the rejection rate is increased to approximately 34%. Due to changes in the acoustic environment, for the speakers 441, 446 and 447 more than one adapted model was created each, covering these specialties.

4.2. Experiments on Noisy Speech

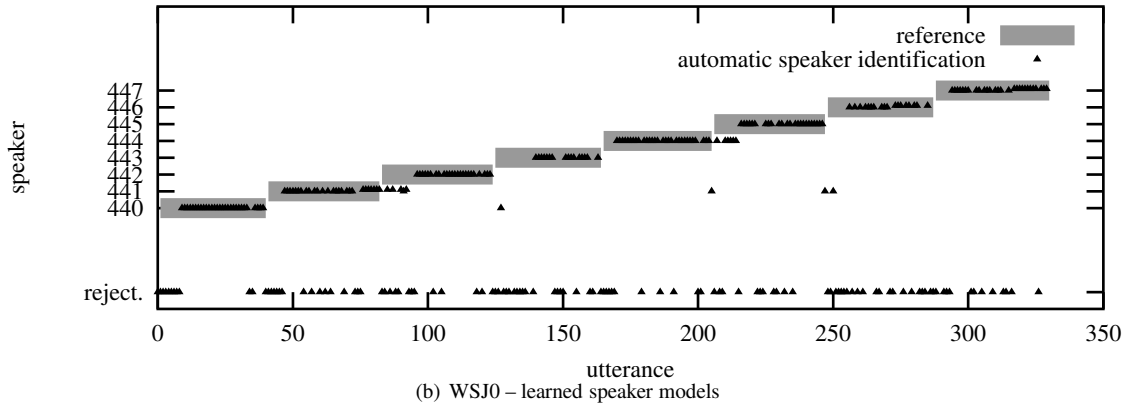
The second set of experiments was performed regarding the task of speaker and speech recognition in cars. The *SLACC (Spoken LAnguage Car Control)* corpus consists of read speech containing instructions (more than 9 hours for training and about 100 minutes for testing) for the control of non safety-relevant 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) [14]. The speaker independent acoustic baseline model uses a shared codebook of 512 Gaussians with diagonal covariances and tri-phone sub-word units with approximately 1500 state clusters. For the first type of experiment we used on average 270 utterances to create speaker adapted acoustic models by MLLR adaptation for each of the three test speakers. The same data was used for estimating GMMs (16 Gaussians each) for speaker identification. The evaluation itself was performed on a set of more than 800 utterances overall. In figure 1(c) the excellent performance of the proposed approach is illustrated for noisy environments with frequently changing acoustic characteristics (city traffic, highway, rain etc.). There is hardly any classification error (only 0.1%) and also the rejection rate is very small (approximately 8%).

In the second type of SLACC related experiments the speaker specific models were established online. The performance of the proposed integrated approach is illustrated in figure 1(d). Similar to the experiments on the WSJ0 task the classification error remains very low at approximately 1.7%. The rejection rate, however, is increased to 34%. For the different acoustic scenarios multiple specialized models were created for all three speakers (especially for speaker 014).

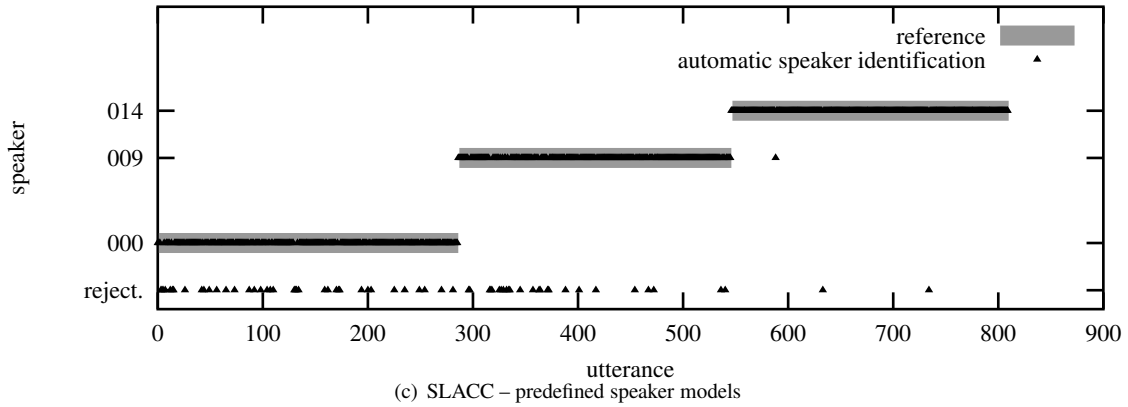
[3]The amount of adaptation data used was determined empirically in informal experiments. Its size mainly affects the speaker models and can be substantially smaller for MLLR adaptation.



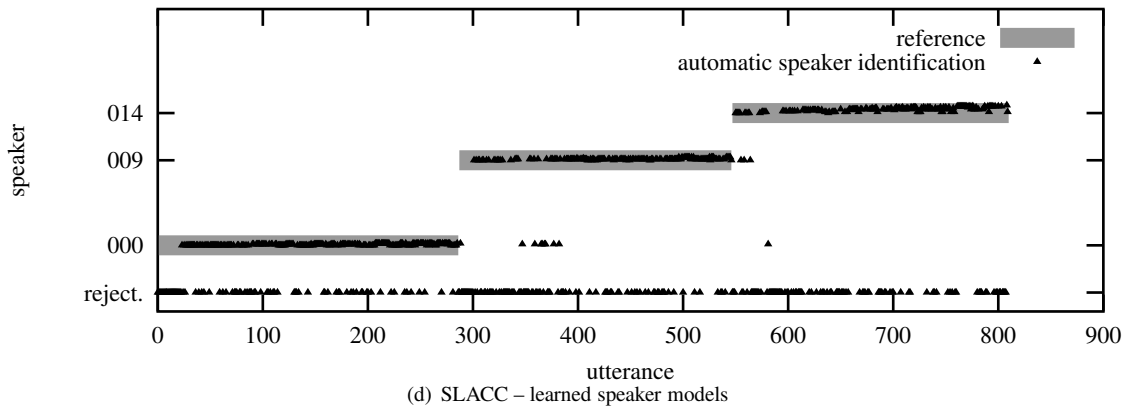
(a) WSJ0 – predefined speaker models



(b) WSJ0 – learned speaker models



(c) SLACC – predefined speaker models



(d) SLACC – learned speaker models

Figure 1: Performance of the integrated speaker identification approach for the WSJ0 and the SLACC corpus

	WSJ0		SLACC	
	CE	RR	CE	RR
predefined	4.54	8.5	0.1	8.0
learned	5.15	34.0	1.72	34.0

Table 1: Speaker identification results using predefined or automatically learned models: Classification error (CE) and rejection rate (RR) in percent.

	WSJ0		SLACC	
	WER	ΔWER	WER	ΔWER
baseline	14.5	–	25.3	–
predefined	12.5	13.8	20.1	21.9
learned	13.6	6.2	19.8	20.6

Table 2: Improvements in speech recognition achieved by speaker adaptation with respect to the speaker independent baseline: Word error rate (WER) and relative improvement (ΔWER) in percent.

4.3. Identification and Recognition

In table 1 the speaker identification results are summarized for both corpora. The classification error (CE) as well as the rejection rate (RR) are given in percent. For each corpus in the first row the results using predefined speaker specific models are shown. In the second row the corresponding figures are given for the online learning scenario.

Though in this paper the focus is clearly on identification performance the proposed integrated framework would be questionable if no improvements in speech recognition quality could be achieved. Therefore, we also measured the relative reduction in word error rate (ΔWER) resulting from speaker adaptive recognition with respect to the speaker independent baseline system without any adaptation procedure. As expected, the WER could be reduced significantly in all experimental configurations (cf. table 2). The best results were obtained with the predefined speaker specific models which is reasonable, as these models were trained on larger amounts of adaptation data. Additionally, in this type of experiment substantially fewer rejections occurred which means that fewer utterances were decoded using the speaker independent acoustic model.

5. Discussion

The evaluation results presented in the previous section demonstrate that speaker identification and speech recognition methods can be combined successfully. But especially in biometric applications, where the prime concern is to achieve the highest possible classification accuracy while at the same time rejecting a minimum number of persons, speaker recognition models learned online on severely limited data should surely be replaced by well trained models. Also speaker recognition applications that need to be truly text independent can not make use of our integrated approach as today's speech recognition systems are always limited to some domain and will, therefore, not perform satisfactorily on out-of-domain data.

The proposed technique is, however, intended to be used in application areas, where speech recognition is required anyway, as is the case for the majority of man-machine interaction scenarios. As state-of-the-art speech recognition systems will be speaker adaptive, the possibility to automatically distinguish between different speakers' speech will already substantially improve the flexibility of the recognizer itself. Additionally, high identification accuracies can be achieved even in noisy environments. Only when learning the models online a relatively high rejection rate has to be accepted. This is, however, not an issue in an interactive scenario, because there the identification of a person does not need to be performed on *every single* utterance that it produces, but the individual cues can be accumulated over a longer period of time. As the detection of unknown speakers is achieved by implicitly performing text dependent speaker verification, the rejection criterion is extremely robust making low identification error rates possible.

Two practical aspect of the integrated speech and speaker recognition method that should be stressed are its rather low computational complexity and its capability to run *online* i.e. processing input speech with only a minor delay defined mostly by the length of the utterance prefix used for speaker identification. The latter feature makes it ideally suited for scenarios where an appropriate low reaction time of the artificial system is expected by a user interacting with it. Due to its moderate computational requirements the method can even be applied on mobile systems – as e.g. robotic companions – that only have limited computational power available. In such an interactive scenario one "peculiarity" of the approach – i.e. the fact that often multiple models per speaker are learned online – can be compensated by additional knowledge sources about the identity of the current communication partner. With such external feedback multiple speaker specific models for a single speaker could be merged appropriately.

6. Conclusion

In this paper we presented an approach for integrating speaker identification and learning with adaptive speech recognition. Presently, adaptive systems are the methodology of choice for speech recognition applications. Based on a speaker independent recognizer specialized models are created for different speakers using moderate amounts of speaker specific adaptation data.

As the adaptation procedure needs to distinguish between data from different speakers in order to modify the baseline system appropriately, speaker recognition has to be performed implicitly.

Therefore, our our system combines both speaker identification as well as speech recognition in a combined decoding process, In a two stage identification procedure GMMs are used for the pre-selection of speaker specific models. Subsequently, the speech recognition results of speaker adapted as well as speaker independent acoustic systems are used for the final speaker recognition. Our system is able to detect new speakers and to identify already known ones. For new speakers specific models are learned in an un-supervised mode.

Experimental evaluations on two non-trivial speech recognition tasks showed that our integrated speaker recognition method is able to achieve robust speaker identification for both predefined speaker models and automatically learned ones.

7. References

[1] Douglas A. Reynolds and Richard C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. on Speech and Audio Processing*, vol. 3, no. 1, pp. 72–83, 1995.

[2] George R. Doddington, Mark A. Przybocki, Alvin F. Martin, and Douglas A. Reynolds, "The NIST speaker recognition evaluation – Overview, methodology, systems, results, perspective," *Speech Communication*, vol. 31, no. 2–3, pp. 225–254, 2000.

[3] Douglas A. Reynolds, "An overview of automatic speaker recognition technology," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing*, Orlando, Florida, 2002, vol. 4, pp. 4072–4075.

[4] Frederick Jelinek, Lalit R. Bahl, and Robert L. Mercer, "Design of a linguistic statistical decoder for the recognition of continuous speech," *IEEE Trans. on Information Theory*, vol. 21, no. 3, pp. 250–256, 1975.

[5] Xuedong Huang, Alex Acero, and Hsiao-Wuen Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*, Prentice Hall, Englewood Cliffs, New Jersey, 2001.

[6] Hermann Ney and Stefan Ortmanns, "Dynamic programming search for continuous speech recognition," *IEEE Signal Processing Magazine*, vol. 16, no. 5, pp. 64–83, 1999.

[7] C. J. Leggetter and P. C. Woodland, "Maximum Likelihood Linear Regression for speaker adaptation of continuous density Hidden Markov Models," *Computer Speech & Language*, vol. 9, pp. 171–185, 1995.

[8] 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*, Istanbul, 2000.

[9] Thomas Plötz and Gernot A. Fink, "Robust time-synchronous environmental adaptation for continuous speech recognition systems," in *Proc. Int. Conf. on Spoken Language Processing*, Denver, 2002, vol. 2, pp. 1409–1412.

[10] Larry P. Heck and Dominique Genoud, "Integrating speaker and speech recognizers: Automatic identity claim capture for speaker verification," in *2001: A Speaker Odyssey – The Speaker Recognition Workshop*, Crete, 2001, pp. 249–254.

[11] Dominic Genoud, Dan Ellis, and Nelson Morgan, "Combined speech and speaker recognition with speaker-adapted connectionist models," in *Proc. Workshop on Automatic Speech Recognition and Understanding*, Keystone, Colorado, 1999.

[12] J. McQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1967, pp. 281–296.

[13] Douglas B. Paul and Janet M. Baker, "The design for the Wall Street Journal-based CSR corpus," in *Speech and Natural Language Workshop*. 1992, Morgan Kaufmann.

[14] Christoph Schillo, "Der SLACC Korpus," Tech. Rep., Faculty of Technology, Bielefeld University, 2001.