Towards unsupervised online word clustering

Holger Brandl, Frank Joublin, Christian Goerick

2008

Preprint:

This is an accepted article published in Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP). The final authenticated version is available online at: https://doi.org/[DOI not available]

TOWARDS UNSUPERVISED ONLINE SPEECH ACQUISITION

Holger Brandl, Britta Wrede

Applied Computer Science
Bielefeld University
hbrandl@techkfak.uni-bielefeld.de

Frank Joublin, Christian Goerick

Honda Research Institute Europe GmbH
Offenbach am Main
Frank.Joublin@honda-ri.de

ABSTRACT

Almost all current speech recognition systems fail to integrate learning into the recognition process. Here we propose a system which is able to recognize and learn the structure of speech online in a unified framework. To do so we've extended HMM-based filler-free keyword spotting with acoustic model acquisition (AMA). To evaluate the inherent dynamics of the combined acquisition-recognition process we propose measures of model activity, model correlation and speech coverage. Based on these criteria the emerging acoustic model is embedded into a regulating framework which was designed to maximize model activation sparseness and speech coverage. First experiments on a speech corpus containing isolated words lead to a coverage of 92% for a training confusion ratio of 0.69 and an averaged model correlation of 6.4%.

Index Terms— One, two, three, four, five

1. INTRODUCTION

The holy grail of speech recognition research is to build systems which automatically acquire the structure and meaning of spoken language. But to this day common automatic speech recognition (ASR) frameworks are designed to detect predefined words using a predefined grammar. To make it even worse, no online learning at all is possible with such systems: the underlying models are trained offline using an annotated speech database and remain fixed during recognition. But although it is clear that human-like speech processing involves learning also during recognition, not too much effort were spent to develop online-learning systems.

Here a new approach to learn the acoustical structure of speech based on incrementally trained Hidden Markov word models is proposed. The idea of this work is to combine simple unsupervised and supervised speech segmentation methods to bootstrap a model-based language representation. Essential to this approach is the regulative feedback loop which controls the acquisition behavior.

Recently some authors claimed to work in the direction of unsupervised AMA (ie. [1], [2], [3]). But most of these works describe only methods for acoustic model (AM) bootstrapping using a small set of annotated speech data: An initial AM is trained supervised with this annotated training sample and is employed to label a larger set of untranscribed speech. These automatically labeled utterances are used to reestimate the model parameters. Sometimes this process is used iteratively to further increase AM goodness. As stated in [4] *lightly* supervised AMA seems to be a more appropriate name for such approaches.

Related to our work are the CELL framework proposed in [5] and the incremental HMM training method for syllable-like units described in [6]. The former defines a framework for multi-modal learning where object labels and semantic categories are learned simultaneously. It lacks of an implementation and evaluation of a top-down feedback loop necessary to ensure a meaningful lexicon. Besides that, its speech processing back end is an ANN-based phoneme recognizer, which was shown to be less powerful for speech recognition than context-depending HMMs (cf. [7]). The approach of [6], which groups similar segments to define syllable models, lacks of the possibility to train models in a time-incremental manner.

The remainder of this work is organized as follows. In section 2 we describe the implemented speech acquisition architecture. We introduce the special kind of MAP-training used to estimate the word models. Subsequently section 3 presents measures which are suitable to reflect the current state of an acoustic model and defines how to integrate these into a unified regulation framework for speech acquisition. Results are presented in section 4 and discussed subsequently in section 5.

2. SYSTEM ARCHITECTURE

As depicted in figure 1 incoming speech is analyzed in a twofold way to detect segments using an energy based voice activity tracker and a keyword spotting system which setups on the word models contained in the acoustic model. Word model acquisition is triggered by voice activity segments of word length and is regulated based on measures of AM completeness, orthogonality and stability.

At first no word models are contained in the acoustic model. Incoming speech is analyzed solely by an unsupervised voice-activity-based speech segmentation module. Inspired by the properties of child directed speech uttered by adults to ease the word model bootstrapping of their children, we assume the input speech to occasionally contain isolated words. Segments of high voice activity are tested whether they obey length constraints to ensure that only segments of word-length are used for model training. These segments are used within the acquisition module to update existing word models or to create new ones.

To avoid the usually difficult choice of a filler model in the implementation of the keyword spotter the approach proposed by [8] was integrated. The different keyword models analyze the speech

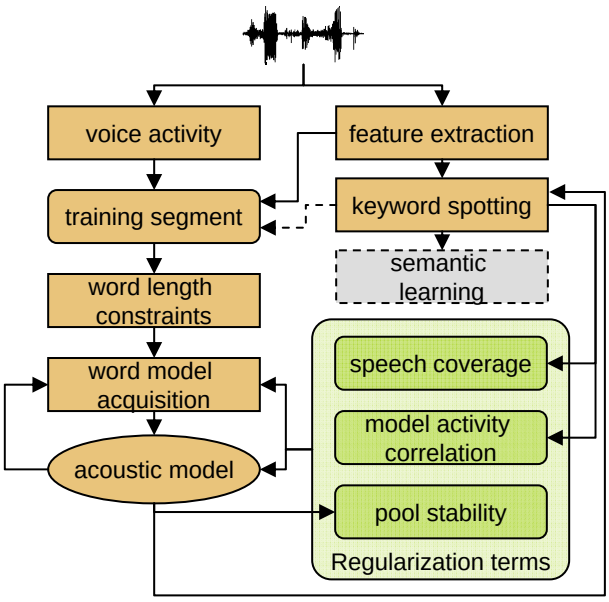


Fig. 1. The speech acquisition loop.

input independently in order to create segment hypotheses. Mel-frequency cepstral coefficients along with normalized energy extended with their first and second-order derivatives were used to give a 39-dimensional feature vector as input for the keyword spotter. All word models were chosen to be Hidden Markov models with Bakis topology containing 8 states. Each state modeled the feature space with a Gaussian mixture model comprising 4 component densities.

Spotted segments are used to update regulative measures and might be further employed within a multi-modal semantic learning framework. As discussed in section 5 keyword spotting could also be used to create training segments within continuous speech utterances.

2.1. The Bootstrapping Process

The AM is empty at the beginning and becomes populated with word models over time. Given an empty AM, incoming training segments can be used to train a first word model. Because it can not be assumed that all initial training segments contain the same word this model should be thought as a general word model and not as a model of a specific word.

The unsupervised clustering method to bootstrap the AM proceeds as follows: Let the acoustic model $\mathcal{M}$ contain at least one word model. A new training segment X will be processed in a twofold way. First the model λ^* which is most likely to explain the given segment is determined by

$$\lambda^* = \arg\max_{\lambda:\mathcal{M}} P(X|\lambda) \tag{1}$$

Thereby $P(X|\lambda)$ denotes the data likelihood. For the second step we assume the histogram of former training to be approximated by a probability distribution with the density $f_{\lambda^*}(p)$. The corresponding cumulative distribution function F_{λ^*} is than used to map $P(X|\lambda^*)$:

$$\nu(\lambda^*, X) = F_{\lambda^*}(P(X|\lambda^*)) = \int_{-\infty}^{P(X|\lambda^*)} f_{\lambda^*}(p)dp \tag{2}$$

Two cases have to be considered (cf. figure 2):

1. $\nu(\lambda^*, X) \geq \theta$: In this case the model λ^* seems not to be an appropriate model for X. But because λ^* was found to be the best model for X in the pool, a new model λ_{new} is created using the model parameters of λ^* for initialization. To make the new model to be different of λ^* the segment X is utilized to perform a first parameter update.
2. $\nu(\lambda^*, X) < \theta$: The model λ^* seems to be appropriate to model the current segment X, which therefore will be used to improve/reestimate λ^*.

The selectable threshold θ used to test whether the current segment is likely to has been generated by the best word model. In this case the model becomes updated, and otherwise a new model is derived. After each processed segment $f_{\lambda_{update}}(p)$ is incrementally updated with $P(X|\lambda_{update})$.

Given that a specific amount of training segments was used to estimate the parameters of a word HMM, it is tagged as *stable*. Subsequently it can be employed to derive new models within the depicted loop. Compared with supervised AMA the resulting acoustic model will contain only word models which are actually required to model the already processed speech utterances. Because the proposed framework combines training and recognition into one integrated framework, new words are modeled based on their appearance in time.

2.2. Model training

To reduce computational costs for training, Viterbi-alignment was applied to split training segments into state-dependent training samples. Doing so the estimation problem reduced to the adaption of the state dependent output probability functions (OPDF). These OPDFs were updated by using *maximum a-posteriori*-training (MAP) procedure to overcome the issue of few training data, to allow an incremental training procedure and to integrate prior knowledge into the speech modeling process (cf. [9]).

Additionally, to further increase the model quality MAP-trained models are updated using the ML algorithm as soon as a defined amount of training data becomes available (cf. [10]). Because of the dominant effect of the state data likelihoods transition probabilities were chosen to be fixed.

3. REGULATION

Regulation may take place at different processing stages. In contrast to supervised AMA we can not rely on aligned labels here. Therefore several measures are introduced which are intended to reflect the current state of the acoustic model. Based on these properties, methods for regulative feedback to control creation, updating and pruning of models are introduced.

Model spotting coverage $\Gamma(t)$ describes how well a speech signal can be modeled at time t given the current acoustic model. It is

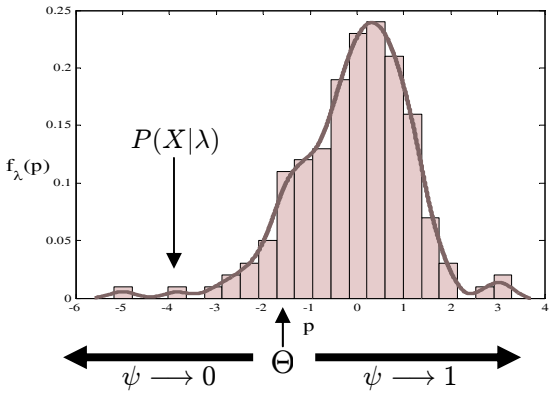


Fig. 2. Adaptive threshold selection for model splitting. Given low coverage the splitting threshold θ is increased to ease the creation of new models.

defined as the ratio of speech covered by at least one of the detected keyword-segments to the overall amount of speech.

Model coactivity describes how sparse the overall spotting activity is, i.e. how many of the models are generating segment hypotheses for in a given time. The more of them are active the more redundant is AM. Ideally one model is active at a time. It is measured pairwise in terms of correlated keyword spotting activity. For two models i and j the model coactivity is denoted with $\eta(\lambda_i, \lambda_j, t)$.

Pool stability $\psi(t)$ is defined as the ratio of stable models to the non stable models.

This triplet defines a concrete implementation of the regularization terms commonly used for unsupervised learning tasks: completeness Γ, orthogonality η and stability ψ. To compute Γ and η a history interval needs to be defined.

Based on these terms the acquisition problem can be reformulated as an optimization problem to provide a unified framework for speech acquisition:

$$\Gamma + \psi - |\eta\| \rightarrow \max! \tag{3}$$

Thereby $|\bullet|$ denotes a common matrix norm.

3.1. Regulation heuristics

Given the method of section 2.1 it is clear that the number of word models in the unsupervised trained AM will grow monotonously over time. It is therefore crucial to limit the size of the AM either by pruning models or by regulating the splitting process.

(I) A first method to limit the pool growth is chosen to be based on pool stability. New models are created only if

$$\psi(t) > \Gamma(t) \tag{4}$$

Otherwise the best pool model is updated. Using this heuristic the creation of new models is eased if speech coverage is low. Vice versa the rule prevents to create new models if the current AM is already able to model the speech input sufficiently.

(II) Whereas the default acquisition loop assumes $\nu(\lambda^*, X)$ to be greater than a fixed threshold it might be more appropriate to use an adaptive threshold. Such a threshold can be chosen by:

$$\theta = \theta_0 \cdot (1 + \beta \cdot \psi) \tag{5}$$

This regulation (cf. figure 2) is inspired by the idea to ease the creation of new models if the AM is in sufficiently stable. Low stability prevents the creation of new models, to allow existing models to reach a stable state by acquiring additional training data.

β and θ_0 are constants to be defined. If $\psi \approx 1$ the θ is chosen to be the default splitting threshold θ_0. Otherwise the stability weight β defines the increasing effect of ψ.

(III) Independent of the control of model acquisition, models which represent the same acoustical entity will occasionally emerge. Therefore a pruning criterion is necessary to remove such redundant models from the AM. Given an pruning sensitivity $\alpha \in [0, 1]$ a pruning rule can be defined by

$$\eta(\lambda_i, \lambda_j) \; > \; (1 - \alpha \cdot \Gamma) \quad \Rightarrow \quad \text{Delete model } \lambda_i \tag{6}$$

Thereby a model is pruned if the model coactivity exceeds a coverage-adapted threshold. Given low coverage values, the adaption rule avoids pruning in order to allow a continuing model adaption.

4. RESULTS

The speech acquisition system was evaluated on subsets of a single-speaker speech database containing subsets including 10 (20min), 20 (40min) and 30 (60min) mainly mono-syllabic uniformly distributed isolated words (0.7 words/second). These speech subsets of different complexity were chosen in order to evaluate the properties of the regulating framework. By Using only one speech set it would not have come clear, whether regulation takes places as expected or the system parameterization only accounts for the emerging model pool.

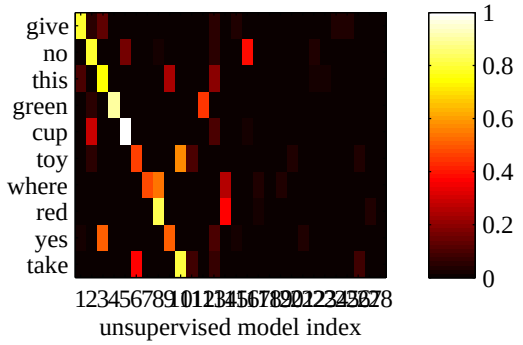
Additionally continuous speech utterances embedding between 0 and 5 of these words into out-of-vocabulary speech were used for evaluation purposes. Because we focus on acquisition and regulation, and not on environment- and speaker- robustness, all speech data was generated by a single speaker in a noise free environment.

4.1. Performance Measures

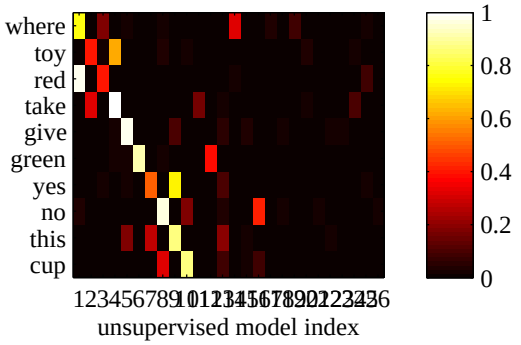
To ensure the training of meaningful models it is necessary to evaluate the system behavior when assigning training segments to models. This is only possible using additional supervised information. Given supervised labels *training confusion matrices* $T_{conf}(t)$ were being computed (cf. [11]) by combining the training histograms of all models. Subsequently the matrix trace was maximized over all column permutations.

As opposed to supervised machine learning tasks the number of models M does not necessarily equal the number of classes C in hierarchical clustering methods like the one proposed in section 2.1. Therefore, to ensure a meaningful trace maximization the classification matrix was extended with dummy columns if there were less AM models than labels.

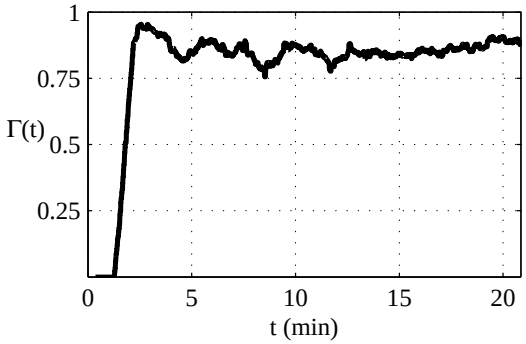
To ease comparative evaluations between different system parameterizations as well as to provide a mean for training process visualization *training kappa* κ_t and *overall training accuracy* p_t were computed (cf. [11]). Additionally, to evaluate the detection performance of the emerging AM, we computed D_{conf}, κ_d and p_d based on the keyword detection activities.



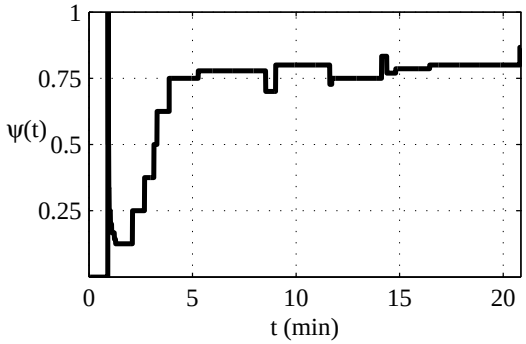
(a) Training confusion T_{conf} of the final AM. Because of the applied trace maximization the relation between models and labels is evident. Because of missing regulation an overhead of 5 additional models is contained in the model pool.



(b) Detection confusion D_{conf} of the final AM. Because D_{conf} lacks of the orthogonality amount found for T_{conf} it seems reasonable to conclude that the acquired word models are sufficient to classify input speech but not sufficient to be used as keyword spotting models.



(c) Speech Coverage Γ and mean detection c. Already after a short training period 95% of all input segments become detected acquired models, which are embedded in the keyword spotter. Besides a slight decay the acquired speech models are continuously able to detect most of the input words



(d) Pool stability ψ. Because only stable model are used to derive new models the model pools is completely stable in the moment the first moment is tagged as stable.

Fig. 3. Results for isolated word acquisition

4.2. Isolated words

Firstly, the basic acquisition loop of section 2.1 combined with regulation type (I) was evaluated. Previous experiments showed that without any pool stability ψ regulation too many models are created leading to a corrupted AM because of the lack of sufficiently large training samples for each model. Figure 4.2 visualizes the acquisition process and the properties of final acoustic model. The system was parameterized with $\theta = 0.05$ and processed the complete 10 words evaluation set. The acquisition system performed in 0.4x realtime using a single-core CPU with a frame-latency of $5.4ms \pm 2ms$.

The final AM contains 15 models which is an overrepresentation of the 10 classes to learn. Compared with the training confusion matrix in 3(a) the detection confusion matrix in figure 3(b) lacks of the low orthogonality. To overcome the latter problem all final models were embedded into a search graph using a flat grammar. Doing so, effects due to a possibly erroneous implementation of the keyword spotter are canceled out and competition between different word models becomes inherently integrated by using Viterbi decoding for recognition.

By assigning labels to appropriate models based on the orthogonality information gained from T_{conf} common word error rates (WER) are applicable to reflect the quality of the acoustic model. Detections of non-assigned supernumerary models were treated as detection errors hereby.

WERs based on an additional training set comprising 5 minutes of 10 isolated uniformly distributed words were computed every 60 seconds using the current model pool. The results are depicted in figure 4. Starting from 100% WER (because of an initial empty AM) the system approaches ???% when the acquisition process was interrupted. Compared with high detection confusion observed in figure 3(b) the final WER of ???% indicates that the AM quality is quite high and therefore rather the used keyword spotter still needs improvements.

Additionally model coactivity is summarized by a Gaussian with mean $\mu_\eta(t)$ of all $\eta(i, j, t)$ its and variance $\sigma^2_\eta(t)$ and visualized as density shape within figure 4. Every time a new model is created σ^2_η increases because the new model and the model it was derived of show the same detection activity as long as the former has not processed enough training samples.

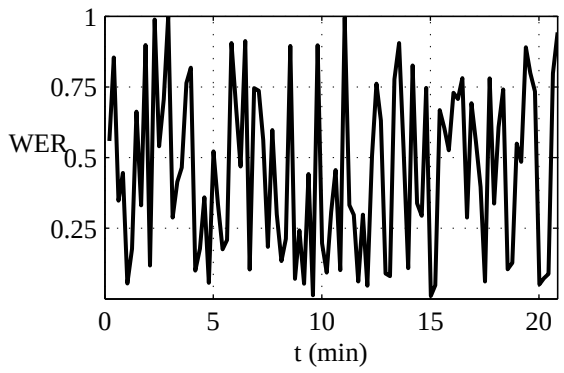


Fig. 4. WER and summerized model coactivity during the acquisition process. σ_η is visualized as tube around the model coactivity mean μ_η

4.3. Regulation

To further convergence speed as well as model quality the combined regulation by the evaluation setup of the previous subsection was extended with adaptive splitting threshold regulation (cf. III). Figure 4.3 depicts the results in terms of unsupervised measures, confusionality and WER.

... . A less sensitive choice of α (cf. eq.6) will lead to more specific word models to be contained in the acoustic model, which entails o higher degree of over-segmentation.

5. DISCUSSION

We proposed a method for word recognition merged with unsupervised AMA by combining ideas of unsupervised and supervised speech processing. So far the approach relies on speech which contains isolated words for acquisition. The key concepts of the approach include a regulation scheme which ensures high model activity sparseness as well as low model correlation. Additionally the number of models was bounded by using model pruning based on model correlation. We could show that our current system is able to learn a stable set of word models independently of the number of words to model. Because the approach is based on time-continuous keyword spotting and time-incremental training the method is suited to be used for online speech acquisition systems.

Here we restricted model acquisition to use voice activity segments only. Although this step was necessary to get a deeper insight into ongoing processes during unsupervised word model acquisition, the use of segments generated by word models itself for training is going to be the next step towards unsupervised speech acquisition. Given that the system won't rely on isolated words as input for training anymore. Assuming the input to possess properties of child directed speech the approach might be able to model some aspects of the early speech acquisition process of children.

Speech acquisition makes sense only in the context of other sensorial modalities. For example visual precepts might provide labels in order to ground the acquired acoustical word models. Given the current performance of the proposed approach it seams reasonably to assume that integration of contextual information could greatly improve the system performance.

6. REFERENCES

[1] Thomas Kemp and Alex Waibel, "Unsupervised training of a speech recognizer: Recent experiments," in *Proc. Eurospeech*, 1999.

[2] Dilek Hakkani-Tr and Giuseppe Riccardi, "Active and unsupervised learning for automatic speech recognition," in *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH)*, 2003.

[3] Frank Wessel and Hermann Ney, "Unsupervised training of acoustic models for large vocabulary continuous speech recognition," in *Automatic Speech Recognition and Understanding Workshop*, 2001.

[4] Lori Lamel, Jean luc Gauvain, and Gilles Adda, "Unsupervised acoustic model training," *ICASSP*, 2002.

[5] D. Roy, *Learning Words from Sights and Sounds: A Computational Model*, Ph.D. thesis, MIT, 1999.

[6] Hema A. Murthy, T. Nagarajan, and N. Hemalatha, "Automatic segmentation and labeling of continuous speech without bootstrapping," in *EUSIPCO*, 2004, Poster-presentation.

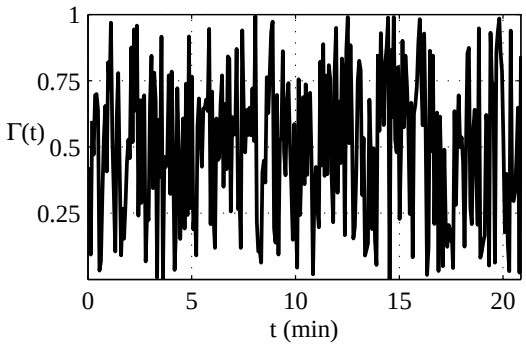
[7] Xuedong Huang, Alex Aceero, and Hsiao-Wuen Hon, *Spoken Language Processing*, Prentice Hall PTR, 2001.

[8] Jochen Junkawitsch, *Detektion von Schlsselwortern in fliessender Sprache*, Ph.D. thesis, Technical University of Munich, 2000.

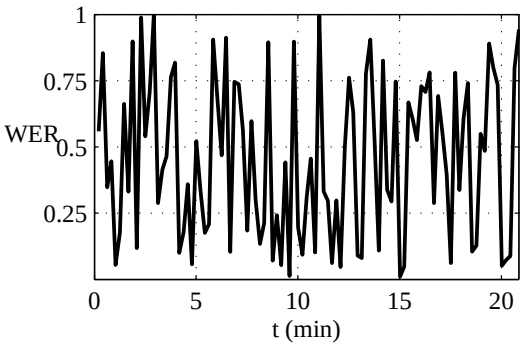
[9] Jean-Luc Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," 1994.

[10] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society. Series B (Methodological*, vol. 39, no. 1, pp. 1–38, 1977.

[11] Gebhard Banko, "A review of assessing the accuracy of classifications of remotely sensed data and of methods including remote sensing data in forest inventory," Tech. Rep., International Institute for Applied Systems Analysis, 1998.



(a) Model coverage Γ and Pool stability ψ for regulation by (I) and (II) using $\theta_0 = 0.05$ and $\beta = 0.2$



(b) WER progression for different threshold adaption modes

Fig. 5. Results for coverage adapted splitting thresholds