TOWARDS UNSUPERVISED ONLINE WORD CLUSTERING

Holger Brandl

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

Frank Joublin, Christian Goerick

Honda Research Institute Europe GmbH
Offenbach am Main, Germany
{firstname.lastname}@honda-ri.de

ABSTRACT

Understanding the bootstrapping process of speech representation in infants is one key issue towards systems which may provide human-like speech recognition abilities some day. Until now, almost all current speech recognition systems have failed to integrate learning into the recognition process. Here we propose a system for unsupervised word-clustering, 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. To evaluate and control the dynamics of the combined acquisition-recognition process we propose measures for model activity, model correlation and speech coverage.

Index Terms— Speech recognition, Unsupervised learning, Clustering methods, Hidden Markov models

1. INTRODUCTION

The holy grail of speech recognition research is to build systems which automatically acquire the structure and meaning of spoken language. However state of the art ASR frameworks are only designed to detect predefined words using a predefined grammar. No learning is possible with such systems, although it is clear that human-like speech processing involves learning also during recognition.

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

Recently some authors have claimed to work in the direction of unsupervised acoustic model acquisition (AMA) (ie. [1], [2], [3]). But most of these works only describe 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 [2] *lightly*, supervised AMA seems to be a more appropriate name for such approaches.

Related to our work are the CELL framework proposed in [4] and the incremental HMM training method for syllable-like units described in [5]. The former defines a framework for multi-modal learning where object labels and semantic categories are learned simultaneously. It lacks the 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 has been shown to be less powerful for speech recognition than Hidden Markov Models (cf. [6]). The approach of [5] which groups similar segments to learn syllable models, lacks 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. Section 3 presents suitable measures 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 twice to detect speech segments: using an energy based voice activity tracker and a keyword spotting system which sets up on the word models contained in the initially empty acoustic model. 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. These segments trigger the word model acquisition process, which regulation is based on measures of AM completeness, orthogonality and stability.

To avoid the usually difficult choice of a filler model for the keyword spotter, the approach proposed by [7] was integrated: the different keyword models analyze the speech input in parallel in order to create segment hypotheses. All word models were chosen to be HMMs with Bakis topology containing 8 states. Each state modeled the feature space with a Gaussian mixture model comprising 4 component densities. Mel-frequency cepstral coefficients, normalized energy, and their first and second-order derivatives were used to define the 39-dimensional feature-space of the system.

Keyword spotting was preferred against a continuous speech recognition approach in order to compute regulative measures for acquisition control as depicted in section 3. Spotted segments might be further employed within a multi-modal semantic learning framework, and - as discussed in section 5 - could be used to derive training segments also within continuous speech utterances.

1-4244-1484-9/08/$25.00 ©2008 IEEE

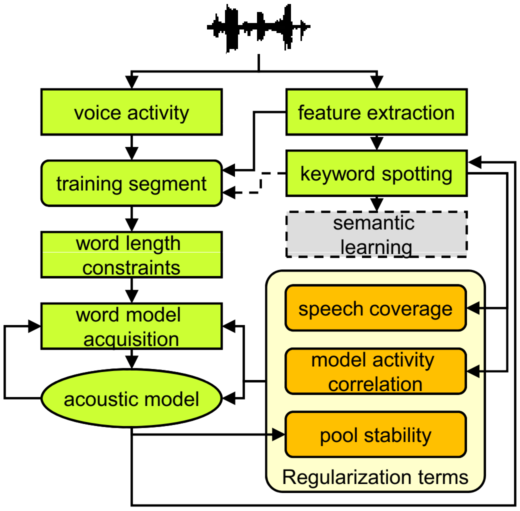


Fig. 1. The combined acquisiton-recognition architecture.

2.1. The Bootstrapping Process

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

The unsupervised clustering method to bootstrap the AM is as follows: let the acoustic model $\mathcal{M}$ contain at least one word model. A new training segment X will be processed in two steps. 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 segment likelihoods of λ^* to be approximated by a probability distribution with the density $f_{\lambda^*}(p)$. The corresponding cumulative distribution function F_{λ^*} is subsequently used to map $P(X|\lambda^*)$:

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

Given a threshold θ, two cases need to be considered (cf. figure 2):

1. $\nu(\lambda^*, X) \leq \theta$: In this case the model λ^* seems not to be an appropriate model for X. But because λ^* was found to be the best matching model for X, a new model λ_{new} is created using the model parameters of λ^* for initialization. The new model is shifted towards X by performing a first parameter update.
2. $\nu(\lambda^*, X) > \theta$: The model λ^* seems to be appropriate to model the current segment X, which will then be used to improve/reestimate λ^*. If a defined amount of segments was accumulated to estimate the model, it is tagged as *stable*.

This approach is related to Leader-Follower Clustering (cf. [8, Chap. 10.11]). Word classes are represented by points in HMM-space. New data points are not HMMs itself, but speech feature segments which are sampled from an existing or still unknown word/class-HMM. Existing classes are matched against segments by computing the data-likelihood for each class. This induces an ordering relation within the current AM which defines the class generating process.

2.2. Model training

To reduce computational costs for training, Viterbi-alignment was applied to split training segments into state-dependent training samples. This way, the estimation problem reduced to the adaption of the state dependent output probability density functions. These were updated by using *maximum a-posteriori* (MAP) training to overcome the issue of few training data, to allow an incremental online learning procedure and to integrate prior knowledge into the speech modeling process by deriving new models from already existing ones (cf. [9]).

Additionally, to further increase the model quality, MAP-trained models were reestimated once using the *maximum likelihood* training as soon as they become tagged as stable . Transition probabilities were chosen to be fixed because of the dominant effect of the state density values.

3. REGULATION

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

Model spotting coverage $\Gamma(t)$ describes how well a speech signal can be modeled at time t given the current acoustic model. It is defined as the ratio of speech covered by at least one of the detected keyword-segments to the overall amount of speech.

Model coactivity $\eta(t)$ describes how sparse the overall spotting activity is, i.e. how many of the models are generating segment hypotheses at a given time. The more of them are active the more redundant is the AM. Ideally, only one model is active at a given 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 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:

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

where $| \bullet |$ denotes a matrix norm.

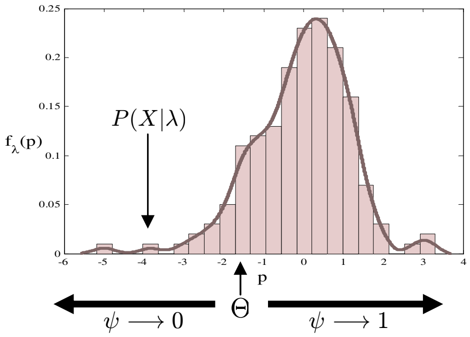


Fig. 2. Adaptive threshold selection for model splitting. Given low coverage the splitting threshold θ is increased to ease the creation of new models.

3.1. Regulation heuristics

Due to the nature of speech, problem 3 cannot be solved analytically. Therefore, in order to bootstrap a meaningful speech representation we use the proposed AM properties to regulate the bootstrapping process presented in section 2.1. To do so, we define some heuristics which are intended to bound the AM size and to to regulate the splitting process.

(I) A first criterion to limit the number of models is based on pool stability and speech coverage: 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 the creation of new models if the current AM is already able to model the speech input well enough.

(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 as:

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

where β and θ_0 are constants to be defined. If $\psi \approx 0$ the θ is chosen to be the default splitting threshold θ_0.

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

(III) Independently 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 a 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.

Number of words	**10**	**20**	**30**	**Ref20**
Processed speech	18min	36min	54min	36min
Speech coverage Γ	94.9%	95.5%	98%	95%
Pool stability ψ	0.85	0.91	0.92	1
# Models / # words	1.4	1.3	1.43	1
WER	19%	45%	42%	4%

Table 1. Final AM properties for corpora of different size. The last column contains the baseline results for the 20 words scenario: In method 2.1 the assignment/splitting decision is done always optimal with respect to the corpus annotation

4. RESULTS

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* $T_{\text{conf}}(t)$ were computed by combining the training histograms of all models. Subsequently, the matrix trace was maximized over all column permutations to make relation between models and labels more evident (cf. figure 3(a)).

Additionally, to evaluate the detection performance of the emerging AM, we computed $D_{\text{conf}}(t)$ based on the keyword detection activities.

Assuming that human speech perception and the metric used within the bootstrapping process match, we assigned labels to models based on the orthogonality information gained from T_{conf}. Doing so, word error rates (WER) were computed on an annotated test set every 60 seconds during the acquisition process. Detected non-labeled supernumerary models were treated as recognition errors. Nevertheless, it is not yet clear to us whether WER is suitable to reflect the quality of the emerged AM.

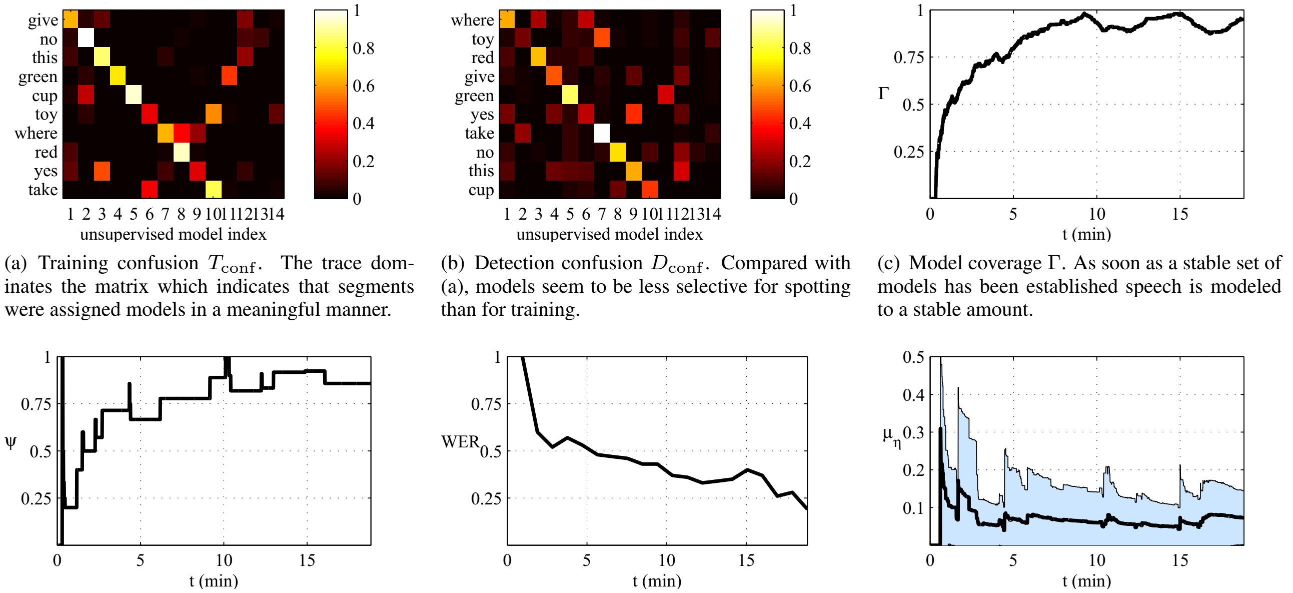
4.1. Evaluation results

The speech acquisition system was evaluated on subsets of a single-speaker speech database containing subsets of 10, 20 and 30 monosyllabic uniformly distributed isolated words (0.7 words/second). Subsets of different size were used to evaluate whether regulation takes place as expected, or whether the system parameterization accounts for the emerging AM. For the same reason the length of the speech input (cf. table 1) was chosen to be much longer than required to bootstrap a stable AM.

The models-words-ratio in table 1, AM stability and coverage are always close to 1, which shows that the emerging AMs are appropriate to model the underlying speech corpora. Because of the self-referential properties of our method, slight under- and over-representation is not avoidable.

The increase of WER for larger corpora depicts the current limits of our method. One possibility to improve the recognition rate would be to reuse training segments of pruned models. Additionally, it might be beneficial to perform a reassignment of training segments in a manner similar to [5].

As shown in figure 3 for the 10 words scenario, the proposed online bootstrapping method leads to a stable representation of the underlying speech entities. The achieved word error rate is 19% which is larger than for the supervised case, but results from the implemented unsupervised bootstrapping framework only. Because D_{conf}



(a) Training confusion T_{conf}. The trace dominates the matrix which indicates that segments were assigned models in a meaningful manner.

(b) Detection confusion D_{conf}. Compared with (a), models seem to be less selective for spotting than for training.

(c) Model coverage Γ. As soon as a stable set of models has been established speech is modeled to a stable amount.

(d) Pool stability ψ. Short drops of stability are due to the creation of new models.

(e) Word error rate WER. WER drops from 100% (because of an initially empty AM) to 19%. WER was computed on an additional test set every 60 seconds.

(f) Model coactivity summarized by a Gaussian with mean $\mu_\eta(t)$ of all $\eta(i, j, t)$ and its variance $\sigma^2_\eta(t)$. Sudden increases are due to the derivation of new models from existing ones.

Fig. 3. Acquisiton process properties for the 10 words case. The used setting was $\theta_0 = 0.05$, $\beta = 0.2$ and $\alpha = 0.1$

lacks of the orthogonality amount found for T_{conf} it seems reasonable to conclude that the acquired models are sufficient to classify input speech but not sufficient to be used as keyword spotting models. Cross-model correlations are inherent to our approach, but were always canceled out after some additional training.

5. DISCUSSION

We proposed a method for unsupervised online word clustering 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 and low model correlation. Additionally the number of models was bounded by using model pruning based on model activity 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 be modeled. Because the approach is based on time-continuous keyword spotting and time-incremental training the method is suitable for online learning. Currently we're already working on using visual percepts to provide the necessary semantic grounding for the acquired acoustical word models.

In this work we restricted ourselves to unsupervised word clustering. This step was necessary to gain deeper insights into ongoing processes during unsupervised word model acquisition. Our next step towards unsupervised speech acquisition will be to also derive new training segments within continuous speech by combining voice activity and spotted word segments. Such a 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 more aspects of the early speech acquisition process of children.

6. REFERENCES

[1] Thomas Kemp and Alex Waibel, "Unsupervised training of a speech recognizer: Recent experiments," in *Proc. Eurospeech*, 1999, pp. 2725–2728.

[2] Lori Lamel, Jean luc Gauvain, and Gilles Adda, "Unsupervised acoustic model training," in *Proc. of ICASSP*, Orlando, May 2002, vol. 1, pp. 877–880.

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

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

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

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

[8] Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*, Wiley-Interscience, 2nd edition, October 2000.

[9] Jean-Luc Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," *IEEE Transactions SAP*, vol. 2, pp. 291–298, 1994.

[10] Gebhard Banko, "A review of assessing the accuracy of classifications," Tech. Rep., International Institute for Applied Systems Analysis, 1998.