DIRAC Training – Report

01.12.2006 – 31.05.2007

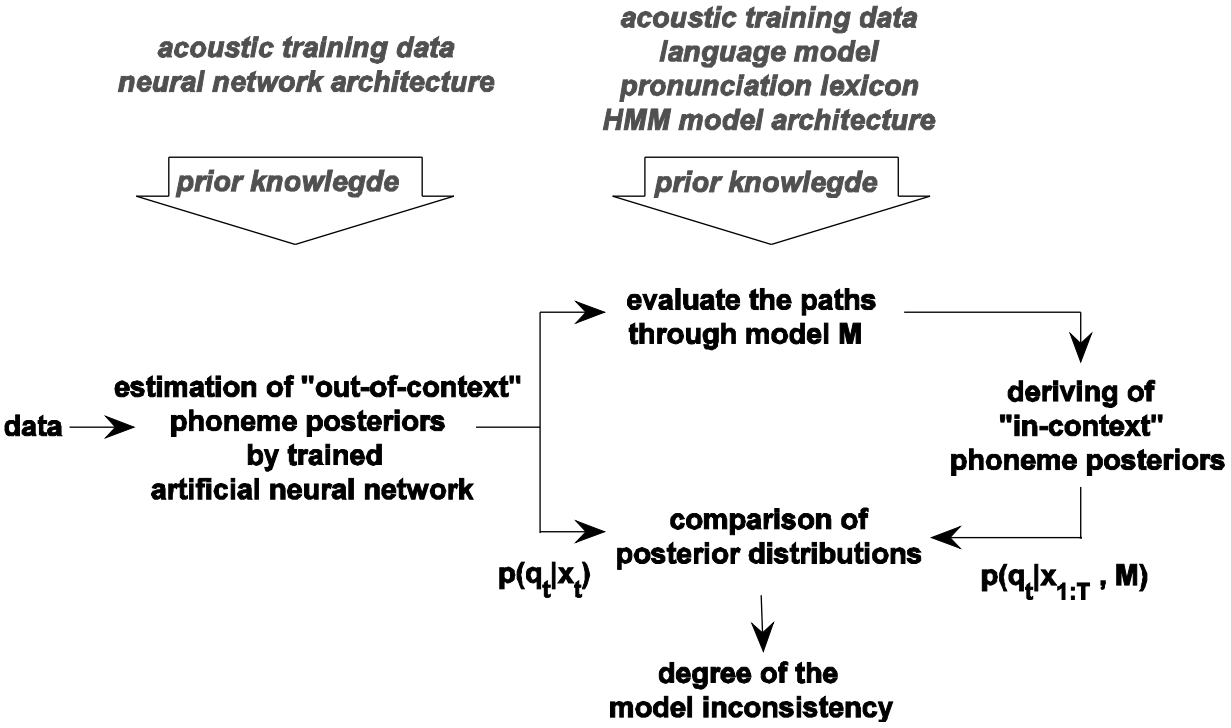
IDIAP Research Institute

Mirko Hannemann

Otto-von-Guericke Universität Magdeburg
Universitätsplatz 2, 39106 Magdeburg
mirko.hannemann@student.uni-magdeburg.de

Introduction

In Speech Recognition, out-of-vocabulary words (OOV) are unexpected events, which can never be recognized correctly, using state-of-the-art technology, since they will be replaced by in-vocabulary items. At IDIAP, a new technique for detection out-of-vocabulary words has been developed, which is based on the comparison of two phoneme posterior streams derived from identical acoustic evidence while using two different sets of prior constraints. This technique has the advantage of not requiring any segment boundary decisions, and was successfully evaluated on a small vocabulary task [1], where it lead to better performance than some earlier reported posterior based confidence measures.



The figure explains the basic concept: Two streams of phoneme posterior probabilities are compared to detect any kind of model inconsistency. One path ("out-of-context") uses only a minimum of prior knowledge (bottom-up recognition), while the other one ("in-context") is also derived from the same acoustical evidences, but uses the full prior knowledge, incorporated in the model M, as well as the full context (top-down recognition).

My task was to apply and extend this approach to Large Vocabulary Continuous Speech Recognition (LVCSR). This was challenging in different ways: (1) The derivation of out-of-context posterior probabilities turns out to be more difficult, since the resulting posteriors are more noisy and less stable. (2) The derivation of in-context posterior probabilities computationally is more complex and therefore different approximation techniques have to be examined. (3) For Comparison, different divergence measures with different parameters and windows have to be compared.
(4) The large vocabulary implies smaller phonetic distances between out-of-vocabulary and in-vocabulary words, and also such phenomena like mispronunciations of in-vocabulary words have to be considered. During my stay, we tried to address each of these, but there are still many questions open. An introductory paper is to appear at Interspeech 2007 [2].

We chose to work with the Wall Street Journal Corpus, because it is well known, we had state of the art recognizers, and we wanted to have high quality speech to concentrate on the OOV problems.

Improving Out-of-Context Posterior Probabilities

Our technique is critically depending on the quality of the out-of-context posteriors, since we compute relative divergences, which rely on these posteriors as acoustic observations.
On the small vocabulary task, the out-of-context posteriors have been derived using neural networks (MLP), using 9 frame context PLP features. For this task, we also derived features using the MRASTA [3] technique (long context modulation spectrum), and we tried different combinations of these two posterior streams, since they contain complimentary information. This research is still going on, also examining hierarchical posterior estimation.
A second possibility to improve posteriors for our task is to smooth the time trajectory of the posteriors by learning nets with long contexts of posteriors or by using Hidden Markov Models (HMM) with only the minimum-duration-assumption as used in phoneme recognizers.
It turned out to be an important question how to measure the quality of posteriors for our task, since it was observed, that better frame error rates do not necessarily lead to better recognition results, and do neither imply any smoothness. Also a confidence measures for this posterior stream are needed, to identify regions, where we cannot rely on this acoustic observation.

Deriving In-Context Posterior Probabilities

We derive the in-context posteriors using the Baum-Welsh forward-backward recursion, given the HMM state posterior probabilities, with the HMM containing prior knowledge like lexicon, pronunciation and grammar[4]. In the small vocabulary case, the lexicon was very limited and the grammar was just an ergodic word loop. Therefore a hybrid (HMM/ANN) system was used, taking the local emission probabilities of the HMM from the MLP, and the full forward-backward recursion could be computed.
For large vocabulary, and especially using higher order language models, this is not feasible, and therefore approximation techniques had to be applied, and also the hybrid recognizer had poor performance.
Therefore we performed a two stage recognition with the first one being our state of the art HMM/GMM recognizer which we used to generate word lattices.
These lattices, we can be used in several ways: (1) We take the vocabulary contained in the lattice to form a dynamic vocabulary, and now the task can be solved as in the small vocabulary case.
(2) We also tried to take only the best (viterbi) path and use this as a rough ("digital" with posteriors close to zero and one) estimate of the posteriors, or to use only the words from the viterbi path in the in-context posteriors estimation, which is giving smoother boundaries. (3) Also we can use multiple (n-best) paths through the lattice and sum up their divergences, and (4) finally, we can also

compute the forward-backward recursion using the whole lattice.
In our experience, the viterbi approximation (2) performed equally to (1) in most cases, but obtaining the phoneme boundaries is an important issue. We obtained the best results deriving them from context-independent models, and using the out-of-context posteriors as features (hybrid system).
An alternative for using lattices would be to constrain the forward-backward recursion, e.g. by beam searches. Also there is still a comparison needed between using HMM/GMM systems for local emission probabilities and using the emission probabilities from the MLP.

Comparing Posterior Streams for OOV Detection

To compare the two posterior streams, we compute the Kullback-Leibler (KL) divergence frame by frame, and for the small vocabulary task it was sufficient to just integrate that over several frames (on syllable level) and set a threshold to detect OOVs. This technique was compared against the Normalized Posterior based Confidence Measures (NPCM) [5].
In our case, we had to carefully review all these decisions, and to examine in more detail the divergence curves to explain the different sources of divergence peaks.
We tried different divergences from the Jensen-Shannon family (inverse, symmetric, bounded,...) and integrated them over different window sizes, or filtered them using different filters.
It turned out that when having worse out-of-context posteriors it is sometimes better to rely on the model than on the acoustic observation.
We also examined some measures to estimate the confidence of the single streams, and to use it to exclude areas which are not reliable, which resulted in an improvement of the true-to-false alarms ratio. Using this as further normalizations for the NPCM techniques and computing new segmented measures based on KL was also tested.
So far we obtain the best results using inverse KL divergence with window sizes on phoneme level, but the research is still going on.
When using any approach which relies on segmentation, but also on all other levels of our system, the boundaries are one major source for divergence. Therefore it is important to compute them coherently (using the same features for segmentation and for confidence measures, using context-independent models). It is also possible to ignore the regions around boundaries, but they also contain information about misalignment in OOV regions.
In the small vocabulary task, we had quite short utterances, and so the evaluation was only carried out on utterance level - so one question is also how to obtain the OOV boundaries after having identified the peak of divergence. This is non-trivial since the neighboring words are also often misrecognized, and the difference between OOV and replacement might only be a single phoneme.

Mispronunciations and Context

Depending on the definition of OOV words, there isn't a clear distinction between OOV and mispronunciations. Therefore, we got exactly the same divergence patterns when a mispronunciation was occuring. Since these are not included in the pronunciation dictionary and in our case not labeled in the corpus, they were reported as false alarms.
The context plays an important role for the detection of OOV and also for the distinction between OOV and mispronunciations, and so our experience was, that using higher order language models is important. If not, an OOV word can easily be replaced by short in-vocabulary words, but this will produce unlikely sequences of words. A mispronunciation could be identified when there is an in-vocabulary word, which has a small phonetic distance to the acoustic observation, and a high(er) probability from the context.

References

[1] Hamed Ketabdar, Hynek Hermansky, “Identifying unexpected words using in-context and out-of-context phoneme posteriors”, IDIAP–RR 06-68, September 2006

[2] Hamed Ketabdar, Mirko Hannemann, Hynek Hermansky, “Detection of Out-of-Vocabulary Words in Posterior Based ASR”, to appear in Proc. Interspeech 2007, Antwerp

[3] H.Hermansky, P.Fousek, “Multi-resolution RASTA filtering for TANDEM-based ASR”, Proc. Interspeech 2005, Lisbon, September 2005

[4] Bourlard, Bengio, Doss, Zhu, Mesot, Morgan, “Towards using hierarchical posteriors for flexible automatic speech recognition systems”, DARPA RT-04 Workshop, Nov. 2004

[5] Guilia Bernardis, Hervé Bourlard, “Improving posterior based confidence measures in hybrid HMM/ANN speech recognition systems”, Proceedings ICSLP'98, August 1998