COMBINATION OF STRONGLY AND WEAKLY CONSTRAINED RECOGNIZERS FOR RELIABLE DETECTION OF OOVS

Lukáš Burget [1], *Petr Schwarz* [1], *Pavel Matějka* [1], *Mirko Hannemann* [2], *Ariya Rastrow* [3], *Christopher White* [3], *Sanjeev Khudanpur* [3], *Hynek Hermansky* [4,1,5] *and Jan Černocký* [1]

(1) Speech@FIT, Brno University of Technology, Czech Republic,
{burget,schwarzp,matejkap,cernocky}@fit.vutbr.cz
(2) Magdeburg University, Germany, mirko.hannemann@student.uni-magdeburg.de
(3) Johns Hopkins University, USA, {ariya,cmileswhite,khudanpur}@jhu.edu
(4) IDIAP Research Institute, Switzerland, hynek@idiap.ch
(5) EPFL Lausanne, Switzerland.

ABSTRACT

This paper addresses the detection of OOV segments in the output of large vocabulary continuous speech recognition (LVCSR) system. First, standard confidence measures based on frame-based word- and phone- posteriors are investigated. Substantial improvement was however obtained when posteriors from two systems - strongly constrained (LVCSR) and weakly constrained (phone posterior estimator) were combined. We show that this approach is suitable also for the detection of general recognition errors. All the results are presented on WSJ task with reduced recognition vocabulary.

***Index Terms*—** LVCSR, OOV, confidence measures.

1. INTRODUCTION

Out of vocabulary words (OOVs) are important source of errors in current large vocabulary continuous speech recognition systems (LVCSR). They are *unavoidable*, as human speech contains proper names, out-of-language and invented words, and also *damaging*, as it is known, that one OOV in input speech generates about 2 recognition errors. On the word error rate of LVCSR, OOVs usually do not have large impact, as they are rare. On the other hand, the information theory tells us that rare and unexpected events tend to be information rich. The working group "Recovery from Model Inconsistency in Multilingual Speech Recognition" (informally "WHAZWRONG?") of 2007 JHU summer workshop concentrated on the detection of OOVs. Reliable detection of OOVs can lead to potential automatic update of recognizer's vocabulary or help subsequent open vocabulary recognition [1, 3].

Confidence measures (CM) [11] are being routinely used for detecting incorrectly recognized words. Our goal is to find confidence measures that would be suitable for the detection of OOVs. We are actually comparing our results to the C_{max} measure computed from word lattices, that was evaluated as the best performing in [11]. In this work, the use of frame-based word- and phone- posterior probabilities (shortly "posteriors") is investigated. Frame-based posteriors have been used as CM too, for example in [2], they served to estimate confidence of words from a hybrid system.

By comparing posteriors from *two* systems: *strongly constrained* (word-based, with language model) and *weakly constrained* (only phones) (Fig. 1), we however aim at not only detecting where the recognizer is not sure (which is the task for confidence estimation) but also to detect places where the recognizer is sure about wrong thing. The mismatch in LVCSR-posteriors and posteriors generated by a weakly constrained system has a chance to reveal the OOV, although the LVCSR itself is quite sure of its output. Preliminary work in this direction was done by Ketabdar and Hermansky [7], the results were however obtained only on a small connected-digit recognition task.

The paper is organized as follows: the following section 2 presents the posteriors and their comparison. Section 3 defines the experimental setup and 4 follows with the results. Section 5 concludes the paper.

2. POSTERIORS AND THEIR COMPARISON

All posteriors used in our work are **frame-based** and are denoted $p(u|t)$, where u is the respective unit (word, phone) and t is time in frames.

2.1. Posteriors from strongly constrained system

LVCSR output is represented as recognition lattice with arcs representing hypothesized words w_i^j, where w_i is the word identity and j is the occurrence of word w_i in the lattice. Each w_i^j spans certain time interval and has associated acoustic and LM scores. Note that occurrences of several w_i^j for the same word w_i can overlap in time. Lattice arc posteriors $p(w_i^j)$ are estimated from the lattice by standard forward-backward algorithm. *Frame-based word-posterior*

This research was conducted under the auspices of the 2007 Johns Hopkins University Summer Workshop, and partially supported by the US Department of Defense (Contract No H-98230-07-C-0365), National Science Foundation (Grant No IIS-0121285) and Defense Advance Research Projects Agency (GALE Program).
Speech@FIT researchers were partly supported by European projects AMIDA and DIRAC, by Grant Agency of Czech Republic under project No. 102/08/0707 and by Czech Ministry of Education under projects No. MSM0021630528. Lukáš Burget was supported by Grant Agency of Czech Republic under project No. GP102/06/383. The hardware used in this work was partially provided by CESNET under projects Nos. 162/2005 and 201/2006.
Thanks to Martin Karafiát and Franta Grézl who provided valuable support to the JHU team at Brno base-station. Thanks also to AMIDA LVCSR team, especially to Thomas Hain from University of Sheffield.
We are grateful to Milind Mahajan from MSR for the MaxEnt code.

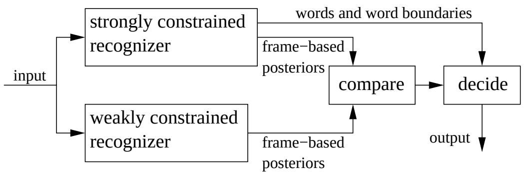

Fig. 1. General scheme.

$p(w_i|t)$ is given by summing all $p(w_i^j)$ active at the given time t. *Word entropy* for time t is estimated as:

$$H(t) = -\sum_i p(w_i|t) \log_2 p(w_i|t), \tag{1}$$

and, in the case of C_{max} confidence measure, the confidence of hypothesized word w_i spanning time (t_s, t_e) is[1]

$$C_{max}(w_i, t_s, t_e) = \max_{t \in (t_s, t_e)} p(w_i|t). \tag{2}$$

The second set of posteriors from strongly constrained system are *LVCSR-phone posteriors*. In our decoder, phones are parts of recognition lattices [8]. It is straightforward to run the forward-backward algorithm on the level of phones and obtain $p(g_i^j)$, where g_i^j denotes jth occurrence of ith phone from the alphabet. Note that there is still a possibility to have concurrent hypothesis of the same phone at the same time. Similarly to words, frame-based phone-posterior $p(g_i|t)$ is given by summing all $p(g_i^j)$ active at the given time t.

2.2. Phone posteriors from weakly constrained system

First set of "weak" posteriors was obtained from a system having the same front-end and acoustic models as LVCSR, but with phones populating the vocabulary and a simple bigram phonotactic model. The resulting phone lattices were processed in the same way as above. We will call these *Phone recognizer posteriors*.

The second set of "weak" posteriors is generated by phone posterior estimator based on a neural net (NN). The NN contains the soft-max non-linearity in the output layer, so that its outputs can be directly considered as frame-based posteriors. These will be denoted *NN phone posteriors.*

Weak posteriors of any kind will be further denoted $p(f_i|t)$. Note that we expect lower entropy for *phone recognizer posteriors*, because of use of 3-state HMMs and phonotactic LM.

2.3. Comparison of posteriors from strong and weak systems

To come up with frame-based confidence measures based on comparison of posteriors from strong and weak systems, we have investigated the following three approaches:

1. **fPCM:** frame-by-frame posterior-based confidence measures [2] are phone posteriors from weakly constrained system found for the phones hypothesized by the strongly constrained system:

$$fPCM(t) = p(f_{i^\star(t)}|t), \tag{3}$$

where $f_{i^\star(t)}$ is the phone recognized by strongly constrained system at time t.

2. **KL:** Kullback-Leibler divergence between the posteriors from the strong and weak system was evaluated. The classical formula:

$$KL(t) = \sum_i p(g_i|t) \log \frac{p(g_i|t)}{p(f_i|t)} \tag{4}$$

was not sufficient and some engineering was needed. First, some posteriors (especially from LVCSR) tend to have zero values, so that thresholding is necessary. Second, there is a temporal alignment problem between the phones generated by the strong and weak systems. We solved it by a soft-alignment: first, for time t, the strongest phone posterior from LVCSR was detected: $s^\star(t) = \arg\max_i p(g_i|t)$. A context of $2N+1$ frames ($t_1 = t - N, t_2 = t + N$) from the weak system was taken and a weighting corresponding to the posterior of $s^\star(t)$ in its output was applied:

$$KL_{avg}(t) = \frac{\sum_{t' \in (t_1,t_2)} p(f_{s^\star(t)}|t') \sum_i p(g_i|t) \log \frac{p(g_i|t)}{p(f_i|t')}}{\sum_{t' \in (t_1,t_2)} p(f_{s^\star(t)}|t')}$$

3. **NN:** The third and most successful approach relied directly on the estimated posteriors. A neural network was trained to combine posterior vectors from strong and weak systems and come up with frame-based confidence measure.

2.4. Post-processing of frame-based values into scores

To convert the described frame-based CM to word-based CM (or simply "confidence measures"), several techniques were investigated. Averaging over hypothesized phones normalized by the number of phones worked well for most of the measures described above. By averaging frame-based word-entropy from Eq. 1, we obtain word-based CM that will be referred to as *mean word entropy* in the following text. Similarly, *mean posterior-based confidence measure (MPCM)* [2] can be obtained by averaging fPCMs (Eq. 3).

In some cases, it was however advantageous to convert frame-based CM to word-based CM in a different way. For example, variance over hypothesized word boundary worked the best for KL divergences. For the following combination, we have selected few well performing post-processing methods for each frame-based CM.

2.5. Combination of word scores

The combinations of word-scores generated by the individual techniques were post-processed by conditional models trained using the maximum entropy (MaxEnt) criterion [12]. Conditional maximum entropy models were chosen based on their history of good performance for speech and language related tasks including language modeling, parsing, etc. Besides MaxEnt classifier, we have experimented also with NN- and SVM-fusing, with similar results, so that we stick with MaxEnt.

2.6. Evaluation

The results are reported for two detection tasks:

- Detecting mis-recognized words overlapping with OOV words
- Detecting mis-recognized words

False alarm probability and miss probability are evaluated on a test set and are shown in standard detection error trade-off (DET) curves. We are primarily interested in the area with low number of false

[1]Wessel et al. in [11] describe special processing of silence arcs. In our case, silences are considered as final parts of words so that no special treatment is necessary.

alarms which is more relevant to practical applications. No one-number metrics such as EER or CER are used in the paper as they are dependent on the ratio of correct targets to overall number of tokens. We leave the choice of the operating point open by reporting the whole DET curve.

3. EXPERIMENTAL SETUP

3.1. Data

The Wall Street Journal corpus (WSJ) was used for both evaluation and development sets. The evaluation set consists of 1243 utterances (2.5 hours), composed from the November 1992, Hub2 5k closed test set and the WSJ1 5k open vocabulary development test set. To train the MaxEnt and the NN for frame-by-frame scores, we defined a development set, consisting of 4088 utterances (7.7 hrs.) of WSJ0 si_tr_s/c. To introduce OOVs, we limited our vocabulary to the 4968 most frequent words from the LM training texts. We decoded the 8 kHz down-sampled utterances with our CTS LVCSR system, and OOVs and recognition errors were labeled. The evaluation set has an OOV token rate of 4.95% in the reference, and in the ASR output we had 13.95% tokens marked as mis-recognized, out of them 8.51% OOV tokens (recognized words overlapping with OOV words in the reference).

3.2. LVCSR and NN-phone posterior estimator

The **NN phone-posterior estimator** was based on NN processing long (300 ms) temporal trajectories of Mel-filter bank energies. On contrary to [10], we used a simple system with only one 3-layer NN with 500 neurons in the hidden layer. The output layer of NN represents phone-state posteriors, but these were summed for each phone to form phone-posteriors. In [10], we have shown that phone-states in the final layer of the NN always greatly improve the accuracy so that we kept this scheme also in this work.

The **LVCSR** was a CTS system derived from AMI[DA] LVCSR [5]. It was trained on 250 hours of Switchboard data. The decoding was done in three passes, always with a simple bigram language model. In the *first pass*, PLP+Δ+$\Delta\Delta$+$\Delta\Delta\Delta$ features were used, they were processed by Heteroscedastic Linear Discriminant Analysis (HLDA), and the models were Minimum-Phone Error (MPE) trained. In the *second pass*, vocal-tract length normalization (VTLN) was applied on the same PLP+Δ+$\Delta\Delta$+$\Delta\Delta\Delta$ features, HLDA and MPE were used, and in addition, constrained maximum likelihood linear regression (CMLLR) and speaker adaptive training (SAT) were used for speaker adaptation. Finally, the *third pass* was the same as pass 2, but PLP-based features were replaced by posterior-features generated by the system described in the previous paragraph, along with their deltas [4].

On WSJ0, Hub2 test from November 92, this system reached word error rate (WER) of 2.9% using a trigram LM, on this closed-set 5k word task.

3.3. Score estimators

When NN was used for direct estimation of frame-based scores, the network was directly fed by posteriors from strong and weak systems. The NN was a 3-layer perceptron with 100 neurons in the hidden layer and the final layer having 3 outputs: OOV, non-OOV and silence. Different schemes of frame-labeling for NN training were devised, the best was to label all frames of an ASR word overlapping with an OOV as "OOV".

A lot of improvement was obtained when temporal context was used in the NN input (see the following section).

4. RESULTS

The first set of DET curves in Fig. 2 shows the results for OOV detection (detection of mis-recognized words overlapping with OOVs) without the use of NN. Mean word entropy significantly outperformed standard C_{max} confidence measure and was found to be the best single score for this task (not considering NN based scores).

Two remaining curves show performance obtained with MaxEnt combination of groups of confidence measures[2]: *LVCSR-based confidence measures* include C_{max}, mean word posterior (related to fWER defined in [6]), mean word entropy, word posterior and entropy from confusion networks [9] and measures related to acoustic stability [11], as well as lattice link entropy, number of different active words, word lattice width and acoustic score, LM-score and duration measures from 1-best word string. Mean posterior-based confidence measure (MPCM) [2] based only on LVCSR posteriors (no combination of strong and weak systems) and mean phone entropy based on lattice from LVCSR were also among *LVCSR-based confidence measures*.

The group of *"weak" confidence measures* consisted of phone entropy based on lattice from phone recognizer, phone entropy based on NN output (both weak recognizers only) and a group of confidence measures comparing LVCSR and weak: KL-divergence between LVCSR and NN posteriors, KL-divergence between LVCSR and phone recognizer posteriors, MPCM based on NN posteriors, MPCM based on phone recognizer posteriors, and several variations of the KL-divergence. The weak confidence measures themselves had poor results, but they provide a nice improvement when combined with LVCSR-based confidence measures.

The second set of results in Fig. 3 shows the results for the NN detecting OOVs from the combination of strong (LVCSR-phone) and weak (NN-phone) posteriors. Note that even the simplest NN-based method taking into account only 1 frame of **phone** posteriors without any context has performance comparable to above mentioned techniques based on **word** posteriors.

Several experiments were done regarding the context for NN. We found that it was optimal to take the strong and weak posteriors from the current frame t, 1 frame in past: $t - 6$ and 1 frame in future: $t + 6$. We attribute this improvement to actually sampling neighboring phonemes, but it deserves further investigation. The last DET curve in Fig. 3 shows that this is the best single technique for OOV detection.

Finally, MaxEnt classifier was used to fuse the results from LVCSR+weak confidence measures and NN – see Fig. 4. In Fig. 5, we present the performance of the same systems in the detection of **all** recognition errors. We see that in both tasks, the NN combined with LVCSR+weak confidence measures performs excellently.

5. CONCLUSIONS

We have shown that combination of parallel strong and weak posterior streams is efficient for detection of OOVs and also for the detection of recognition errors. Different scores perform differently for the two tasks; NN seems especially suitable for the OOV detection. We are however aware of the simplicity of the defined task, and in future we plan to test the outlined approaches on more representative spontaneous speech data.

[2]Some CMs were not described in the previous text, the meaning is either obvious, or the reader is referred to the citations.

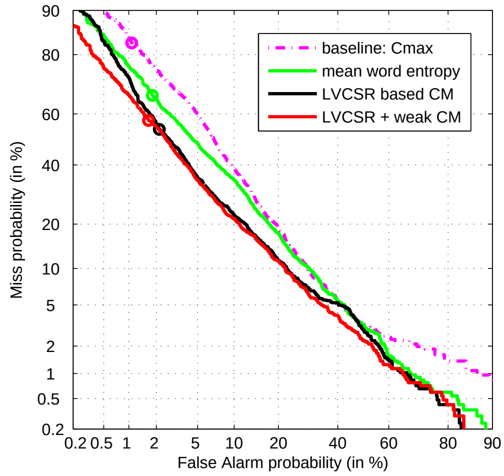

Fig. 2. OOV detection using strong system only and combination of strong and weak systems.

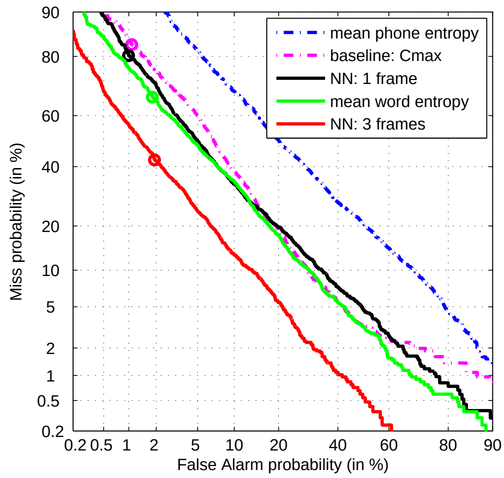

Fig. 3. OOV detection using NN with 1-frame and 3-frame input $(t, t-6, t+6)$.

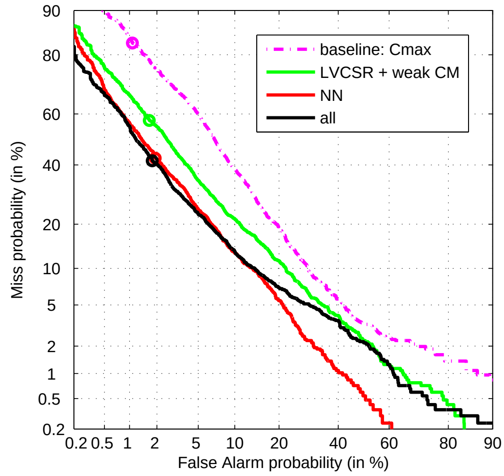

Fig. 4. OOV detection using combination of LVCSR+weak confidence measures and NN.

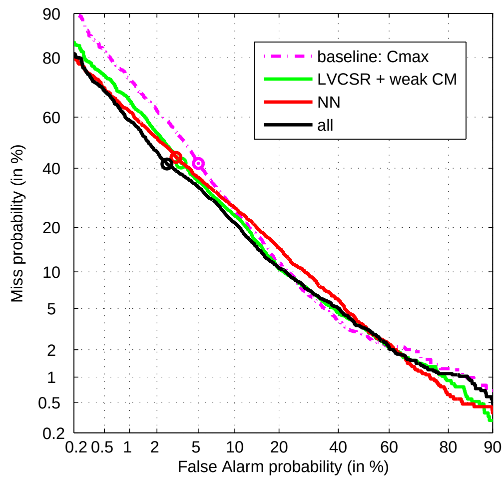

Fig. 5. Recognition error detection using combination of LVCSR+weak confidence measures and NN.

6. REFERENCES

[1] Issam Bazzi: *Modelling out-of-vocabulary words for robust speech recognition*, Ph.D. thesis, MIT, 2002.

[2] G. Bernardis and H. Bourlard: "Improving Posterior Based Confidence Measures in Hybrid HMM/ANN Speech Recognition Systems", in *Proc. ICSLP'98*, Sydney, Australia, 1998.

[3] M. Bisani and H. Ney: "Open vocabulary speech recognition with flat hybrid models", In *Proc. Interspeech-2005*.

[4] P. Fousek and F. Grzl: "Optimizing Bottle-Neck Features for LVCSR", in *Proc. ICASSP 2008*, Las Vegas, 2008.

[5] T. Hain, et al: "The AMI System for the Transcription of Speech in Meetings", In *Proc. ICASSP 2007*, Hawaii, 2007, pp. 357-360.

[6] B. Hoffmeister, D. Hillard, S. Hahn, R. Schluter, M. Ostendorf and H. Ney: Cross-site and intra-site ASR system combination: comparisons on lattice and 1-best methods, in *Proc. ICASSP 2007*, Hawaii, 2007.

[7] H. Ketabdar, M. Hannemann and H. Hermansky: "Detection of Out-of-Vocabulary Words in Posterior Based ASR", in *Proc. Interspeech 2007*, Antwerp, 2007.

[8] A. Ljolje, F. Pereira, and M. Riley: "Efficient General lattice Generation and Rescoring". In *Proc. Eurospeech '99*, Budapest, 1999.

[9] L. Mangu, E. Brill and A. Stolcke: "Finding Consensus Among Words: Lattice-Based Word Error Minimization". In *Proc. Eurospeech'99*, Budapest, 1999.

[10] P. Schwarz, P. Matějka, and J. Černocký: "Hierarchical structures of neural networks for phoneme recognition", in *Proc. ICASSP 2006*, Toulouse, 2006.

[11] F. Wessel, R. Schlüter, K. Macherey and H. Ney: "Confidence measures for large vocabulary continuous speech recognition", *IEEE Trans. Speech and Audio Processing*, vol. 9, no. 3, pp. 288–298, 2001.

[12] C. White, J. Droppo, A. Acero and Julian Odell: "Maximum entropy confidence estimation for speech recognition", in *Proc. ICASSP 2007*, Hawaii, 2007.