Combination of Word and Phoneme Approach for Spoken Term Detection

Igor Szöke, Lukáš Burget, Martin Karafiát

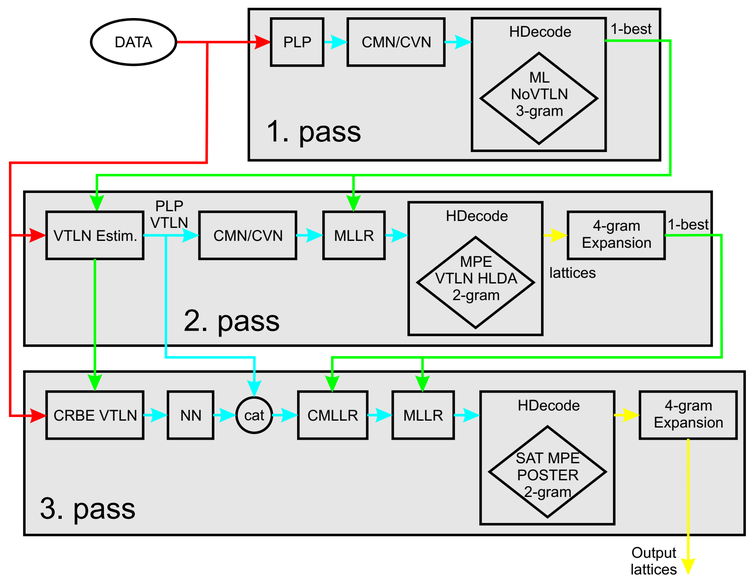
Speech@FIT group, Brno University of Technology, Czech Republic
e-mail: {szoke,burget,karafiat}@fit.vutbr.cz

Abstract

The goal is to provide a technique for search for an out-of-vocabulary (OOV) term or keyword in a speech. The spoken term detection is based on lattices generated from word and phoneme recognizers. The term is a sequence of words (quoted query). The results are evaluated on the Broadcast news subset of NIST STD 2006 development set. Our goal is to propose an upper-bound word/phoneme combination technique for Spoken Term Detection (STD). We want to estimate the upper bound results for OOV terms, which can be achieved using rescoring by strong (and slow) LVCSR system. Then we can decide how much worse is an simple (and fast) sub-word method.
We also evaluated different approaches for term confidence estimation and two differen approaches for phoneme lattice generation.

1 Evaluation

We generate lattices using our LVCSR system from NIST STD 2006 evals.



Broadcast news subset of the STD 2006 development set was taken as test data. We removed 880 words from the LVCSR vocabulary (original size 50k). The term list had 1099 terms. 123 terms contain true OOV word(s). These were omitted from term list and also one 1-phoneme term (A.) was omitted. Final term set had **975 terms** of which were **481 IV** terms and **494 OOV** terms.

We use posterior probability as the term confidence in lattice STD.

$$p^{latt}(term) = \frac{L^{\alpha}(N_b(A_1(term))) \cdot \prod_{a=A_1(term)}^{A_M(term)} L(a) \cdot L^{\beta}(N_e(A_M(term)))}{L^{\alpha}(N_e(latt))}$$

The posterior probability can be evaluated in BaumWelch or Viterbi style:

$$L^{\alpha BW}(N) = \sum_{p \in P_b^N} \prod_{A_i=A_1^p}^{A_n^p} L(A_i)$$

$$L^{\alpha VIT}(N) = \max_{p \in P_b^N} \prod_{A_i=A_1^p}^{A_n^p} L(A_i)$$

NIST proposed new metric for STD 2006 evaluations. The metric is called Term Weighted Value. We propose and evaluate using upper bound TVW to overcome TVW's one global threshold.

$$p_{MISS}(thr) = \underset{term}{average}\{p_{MISS}(term, thr)\}$$
$$p_{FA}(thr) = \underset{term}{average}\{p_{FA}(term, thr)\}$$
$$TWV(thr) = 1 - \underset{term}{average}\{p_{MISS}(term, thr) + 999.9 p_{FA}(term, thr)\}$$

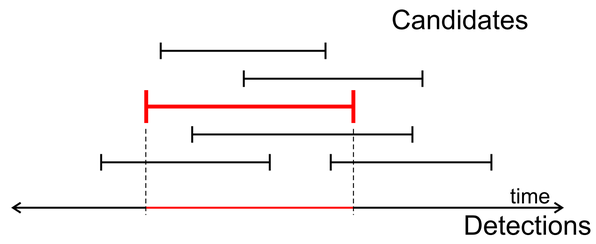
$$thr_{ideal}(term) = arg \max_{thr} TWV(term, thr)$$
$$UBTWV = 1 - \underset{term}{average}\{p_{MISS}(term, thr_{ideal}(term)) + 999.9 p_{FA}(term, thr_{ideal}(term))\}$$

Word, phoneme and phoneme from word lattices were generated. The baseline results using "Viterbi" style posterior probability are:

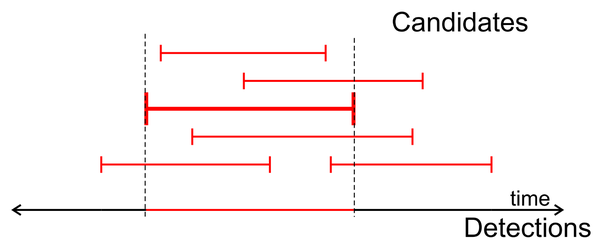
Description UBTWV	ALL	IV	OOV
WRD	0.721	0.740	0.701
WRD REDUCED	0.370	0.716	0.000
WRD-PHN	0.653	0.631	0.676
WRD-PHN REDUCED	0.449	0.612	0.274
PHN	0.542	0.532	0.569

2 Term confidence estimation

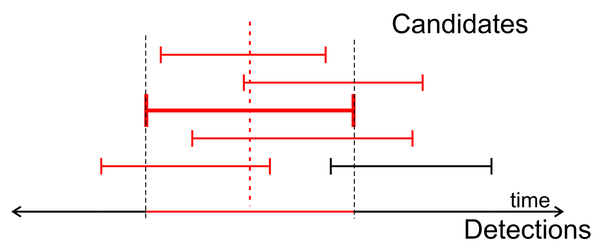
Term confidence is the term posterior probability.



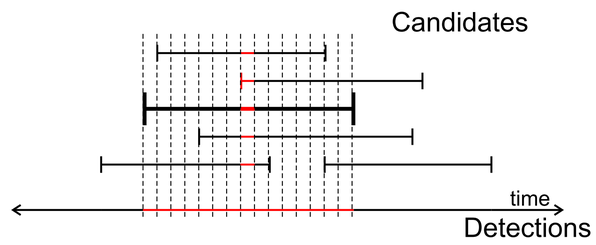
Term confidence is the sum of posterior probabilities of the same overlapped terms.



Term confidence is the sum of posterior probabilities of the same terms overlapping the term center.



Term confidence is the best frame-by-frame sum of posterior probabilities of the same overlapped terms.



The accuracies of different term confidence evaluation approaches:

Description UBTWV	ALL Viterbi	ALL BaumWelch
Post. Prob.	0.721	0.726
Sum P.P.	0.700	0.731
Center sum P.P.	0.720	0.726
Frame-by-frame	0.700	0.731

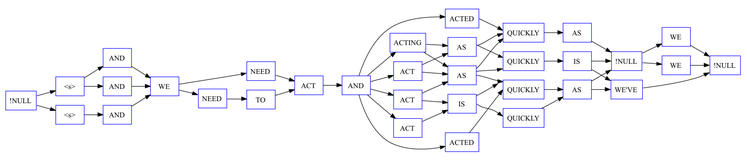
3 OOV term rescoring using LVCSR

Estimation of the keyword (term) confidence is in terms of likelihood ratio. The confidence is the posterior probability in case of lattice base STD.

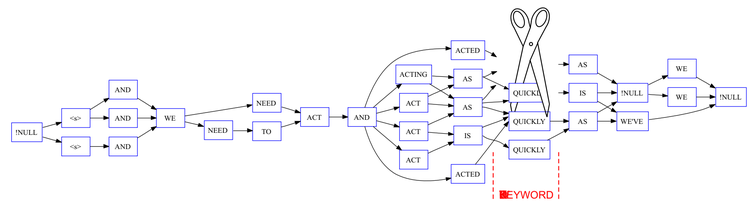
Confidence of the keyword is given by:

$$Conf(kw) = \frac{L(keyword)}{L(background)}$$
$$Conf(kw) = \frac{L^{\alpha}(N_{beg}(kw)) L(kw) L^{\beta}(N_{end}(kw))}{L^{\alpha}(N_{end}(lattorig))}$$

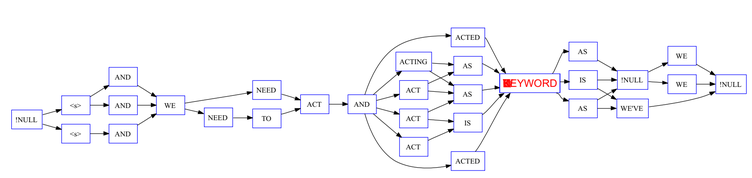
Original LVCSR lattice. The keyword candidate is searched in phoneme lattice. The original lattice is the background model.



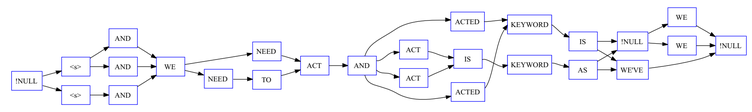
Corresponding part of the lattice is cut-out.



The keyword is inserted and linked to all neighbour nodes.



The lattice is rescored. The rescored lattice is the keyword model with filler models.



The results:

Description UBTWV	ALL	IV	OOV
WRD	0.721	0.740	0.701
WRD INS	0.714	0.728	0.700
PHN	0.542	0.532	0.569
PHN INS 0fr	-	-	0.349
PHN INS 1fr	-	-	0.410
PHN INS 3fr	-	-	0.518
PHN INS 5fr	-	-	0.563
PHN INS 6fr	-	-	0.577
PHN INS 13fr	-	-	0.592

Context [fr]	0	1	3	5	6	13
PHN INS	0.349	0.410	0.518	0.563	0.577	0.592
PHN baseline	0.340	0.400	0.466	0.533	0.545	0.548

4 Conclusions and discussion

- Phoneme lattices generated from word lattices have lower accuracy for OOV terms.
- Term confidence computed as sum of posterior probabilities of the same overlapped terms as the same accuracy as the frame-by-frame approach. This increases accuracy for "BaumWelch" style posterior probability, but decrease accuracy for "Viterbi" style posterior probability.
- Rescoring of term candidates generated from phoneme lattice by LVCSR system improves the accuracy.
- Experiments with different language models (open, class-based) must be done in case of the term rescoring task.

5 References

Wessel2001: F. Wessel and R. Schluter and K. Macherey and H. Ney,Confidence Measures for Large Vocabulary Continuous Speech Recognition, IEEE Trans. Speech and Audio Processing, vol 9, No. 3, 2001

Yu2004:P. Yu and F. Seide, A Hybrid Word/Phoneme-Based Approach For Improved Vocabulary-Independent Search in Spontaneous Speech, ICSLP 2004

Jiang2005: H. Jiang, Confidence Measures for Speech Recognition: A Survey, Speech Communication, vol 45, 2005

Chen2006: T. H. Chen and B. Chen and H. M. Wang,On Using Entropy Information to Improve Posterior Probability-based Confidence Measures, ISCSLP, 2006

Acknowledgement

This work was partly supported by the European Union 6th FWP IST Integrated Project AMI (Augmented Multi-party Interaction, FP6-506811, publication) and Caretaker project (FP6-027231), by Grant Agency of Czech Republic under project No. 102/05/0278 and by Czech Ministry of Education under project No. MSM0021630528. The hardware used in this work was partially provided by CESNET under project No. 119/2004, No. 162/2005 and No. 201/2006