Hybrid word-subword decoding for spoken term detection.

Igor Szöke
szoke@fit.vutbr.cz

Michal Fapšo
ifapso@fit.vutbr.cz

Lukáš Burget
burget@fit.vutbr.cz

Jan Černocký
Speech@FIT, Brno University
of Technology
Božetěchova 2, 612 66
Brno, Czech Republic
cernocky@fit.vutbr.cz

ABSTRACT

This paper deals with a hybrid word-subword recognition system for spoken term detection. The decoding is driven by a hybrid recognition network and the decoder directly produces hybrid word-subword lattices. One phone and two multigram models were tested to represent sub-word units. The systems were evaluated in terms of spoken term detection accuracy and the size of index. We concluded that the best subword model for hybrid word-subword recognition is the multigram model trained on the word recognizer vocabulary. We achieved an improvement in word recognition accuracy, and in spoken term detection accuracy when in-vocabulary and out-of-vocabulary terms are searched separately. Spoken term detection accuracy with the full (in-vocabulary and out-of-vocabulary) term set was slightly worse but the required index size was significantly reduced.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*

General Terms

spoken term detection, hybrid word-subword recognition

Keywords

word, subword, recognition, speech, decoding, indexing, term

1. INTRODUCTION

Spoken term detection (STD) is an important part of speech processing. Its goal is to detect terms in spoken documents, such as broadcast news, telephone conversations, or meetings. The most common way to perform STD is to use the output of large vocabulary continuous speech recognizer (LVCSR). Rather than using the 1-best output of LVCSR, the state-of-the-art STD systems search terms in lattices – acyclic oriented graphs of parallel hypothesis. In addition to better chances to find the searched term, the lattices also offer an easy way to estimate the confidence of the given query [6].

A drawback of the LVCSR system is, that it recognizes only words which are in an LVCSR vocabulary, so that the following STD system can not detect out-of-vocabulary words (OOVs) although OOVs usually carry a lot of the information (named entities, etc.). Common way to search OOVs is to use subword units – a search term is converted into a sequence of such units when it is entered, either using a dictionary (which can be much larger than that of LVCSR) or by a grapheme-to-phoneme (G2P) converter. Such sequence is then searched in the output of subword recognizer.

In our prior work [2], we have studied the combination of words (LVCSR) and subwords (phones). Both systems were run separately, and the outputs were indexed in two indices: word unigrams and phone trigrams. In the search phase, input term was split into in-vocabulary and out-of-vocabulary parts and these were searched in the indexes. Finally, the outputs were combined and term candidates were produced. The drawbacks were the impossibility to search an OOV word shorter than 3 phones and the complexity: word and subword decoding had to be done separately and two separate indices had to be maintained. Finally, word and subword systems had to be calibrated separately.

Our previous paper [9] deals with phone multigrams for subword recognition and indexing instead of phone trigrams. We concluded that the multigrams increase subword spoken term detection accuracy by 10% relative and decrease the index size to 1/5 in comparison to phone trigrams.

In this paper, we investigate into the use of hybrid word-subword recognizer to simplify the spoken term detection system. Our goal is to produce **word-subword lattices**. These lattices should be indexed in **one**, as small as possible index. Terms should be easily searched and it should not matter if a term contains OOVs or not. It is also important to preserve the accuracy of the simplified system.

2. HYBRID WORD-SUBWORD DECODING

Combination of word and subword STD can be done on several levels:

SIGIR The 31st Annual International ACM SIGIR Conference 20-24 July 2008, Singapore

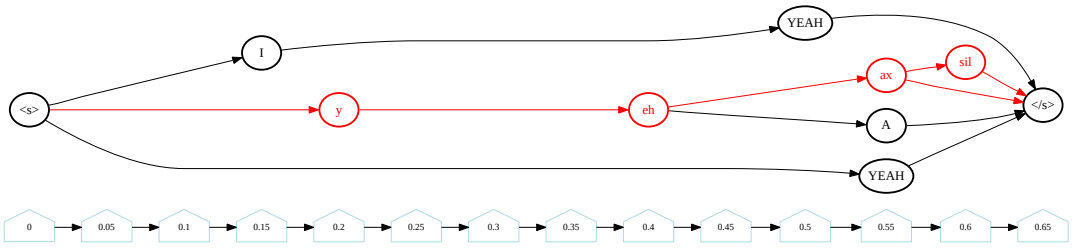

Figure 1: An example of word-subword lattice.

- The first level of word-subword combination is on the recognition (decoding) level (denoted as *prior combination*). The output is a hybrid word-subword lattice (Figure 1) which is searched for terms [1, 11, 12].

- The second level is combination after the decoding (denoted as *posterior combination*). Word and subword outputs are generated separately and then combined together to hybrid lattice. In both approaches, in-vocabulary (IV) and out-of-vocabulary (OOV) terms are directly searched in the lattices. These two approaches were compared in [12]. The authors concluded that the word level posterior combination achieved better accuracy on IV keywords than the prior combination. The deterioration was caused by the mismatching score levels of the word and phonetic language model. On the other hand, they did not use word language model with a special symbol for OOV.

- The last level is combination of searched results. Decoding and search is done separately for words and subwords. The term's word and subword parts are searched separately in the appropriate lattice [2]. Lastly, the candidates are combined together. The drawback of this approach is that we need two standalone systems.

Doing the combination of word and subword STD at the first level (during the decoding) is the most straightforward approach. A hybrid word-subword language model is the only thing which is needed for the decoding.

The word recognizer is considered as strong recognizer. It has strong acoustic model (words) and language model (word bigrams). The subword recognizer is considered as a weak recognizer. It has weak phone or multigram units and no or unigram language model. Combination of the word and subword recognizer should allow to traverse between words and subwords at any time. If traversing penalties and other parameters are set correctly, the word part should well represent in-vocabulary speech. On the other hand, increased "resistance" of the word part should lead to activation of subword part for an OOV segment of speech.

We decided to use the approach similar to [1], which is based on a word language model containing a symbol for unseen words. The unseen word is modeled by the OOV (subword) model. In [1], the author investigated the OOV detection and its impacts on word recognition. On contrary, we aimed at an investigation of STD accuracy and practical application for searching in spoken documents.

2.1 Building the hybrid recognition network

We used our static decoder *SVite* for hybrid recognition experiments. The only one modification was realized in the **network** for hybrid recognition/decoding.

The network can be seen as a weighted finite state transducer (*WFST*) which maps a sequence of HMM models to a sequence of word labels which are accepted by a language model (weighted finite state acceptor).

The WFST is a finite state device that encodes a mapping between input and output symbol sequences. A weighted transducer associates weights such as probabilities, durations, penalties or any other quantity that accumulates linearly along paths, to each pair of input and output symbol sequence. WFST provides a natural representation of HMM models, pronunciation dictionary and language model [8]. Weighted determinization and minimization algorithms optimize their time and space requirements, and a weight pushing algorithm distributes the weights along the paths of a weighted transducer optimally for speech recognition.

Consider a pronunciation lexicon L and take its Kleene closure by connecting an ϵ-transition from each final state to the initial state. The resulting pronunciation lexicon can transcribe any sequence of words from the vocabulary to the corresponding phoneme sequence.

Consider a language model G. The composition of these two WFSTs,

$$L \circ G, \tag{1}$$

gives a transducer that maps from phones to word sequences while assigning a language model score to each such sequence of words. Incorporating context-dependent triphone models is a simple matter of composing

$$C \circ L \circ G, \tag{2}$$

where C represents the mapping from context-dependent to context-independent phonetic units. Then, incorporating HMM models H:

$$H \circ C \circ L \circ G, \tag{3}$$

results in a transducer capable of mapping distributions to word sequences restricted to the language model G. The hybrid word-subword recognition network can be built by

$$H \circ C \circ (L_{word} \cup L_{subword}) \circ G_{subword} \circ G_{word}, \tag{4}$$

where H and C are the same as in Eq. 3, L_{word} is the pronunciation dictionary mapping phones to words, $L_{subword}$ maps phones to subword units (eg. syllables, multigrams or phones). $G_{subword}$ is a weighted transducer created from the subword language model and G_{word} represents the word language model.

2.2 Word model

The WFST L is generated from standard pronunciation lexicon. The word LM must be open vocabulary, so it must contain an "<unk>" word. The "<unk>" is considered as the OOV word which will be modelled by the subword model (see Figure 2).

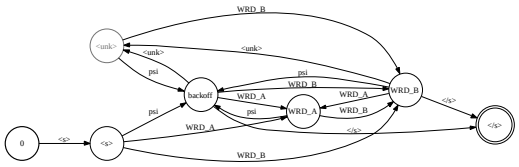

Figure 2: An example of open vocabulary language model. The <unk> states for the out-of-vocabulary words.

2.3 Subword model

The second input is a subword model. Simple phone bigram language model is shown as an example in Figure 3. The <unk> symbol is replaced by this subword model.

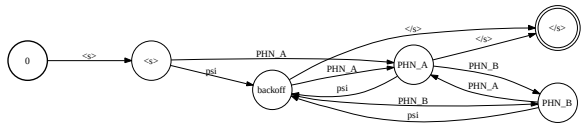

Figure 3: An example of a subword (phone) language model.

The substitution is illustrated in the Figure 4. The gray part of network is <unk> substituted by the subword model.

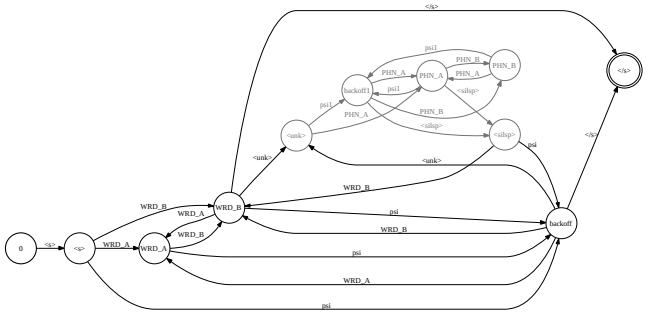

Figure 4: An example of hybrid word-phone language model where the <unk> symbol was substituted by the phone model.

Parameters such as word insertion penalty and acoustic or language model scaling factors can be tuned to control the recognition accuracy and output of the LVCSR system. But the hybrid network is considered as one unit by the decoder. The same penalty and scaling factor apply for both word and subword parts. That is why three different parameters were incorporated into the combined network during its building. The first parameter is subword language model scaling factor *SLMSF*. This parameter multiplies the log likelihoods assigned to the subword LM transitions. The second parameter is the subword word insertion penalty *SWIP*. It is a constant which is added to each transition's log likelihood value leading to a word node. The last parameter is the subword cost *SC*. It is a constant which is added to the <unk> symbol and represents a simple cost of going to the whole subword model.

We decided to use three different subword models. The first is a phone loop. The second and the third are multigram based units.

2.4 Multigrams

The multigram language model was proposed by Deligne et al. [3]. Multigram model is a statistical model having sequences with variable number of units. We implemented the multigram estimation according to [3] and we used the Viterbi approach. Multigram units which occur less than 5 time (multigram pruning factor) are omitted from the inventory.

3. SYSTEM DESCRIPTION

3.1 Recognition system

During the pre-processing, the acoustic data was split into shorter segments in silences (output of speech/nonspeech detector) longer than 0.5s. The data was also split if the speaker changed (based on the output of diarization). Segments longer than 1 minute were split into 2 parts in silence the closest to the center of the segment. This was done to overcome long segments and accompanying problems during decoding.

Acoustic models from an LVCSR system were used for subword recognition. The used LVCSR [10] is a state-of-the-art system derived from AMIDA LVCSR [5]. The system uses standard cross-word tied states triphone models and works in three passes of recognition.

The acoustic models are trained on *ctstrain04* [7] corpora which is a subset of *h5train03* set defined at Cambridge. Total amount of data is 277 hours. A bigram word language model was trained on 977M words of a mix of 9 corpora. The corpora contain mainly conversation speech and round table meeting transcripts.

The same *ctstrain04* corpora was used as base phone corpora for our experiments. The size is 11.5M phones.

3.2 Subword training data

Phone language model and multigrams were trained on phone strings. The *ctstrain04* was searched for utterances containing "out-of-vocabulary" words defined in section 4. These utterances were omitted and the set was denoted *LnoOOV*.

According to the size of *LnoOOV* and the iterative multigram training procedure, the data used for estimation of multigram dictionary was reduced to 3.75M phones to achieve reasonable training time (several hours). This corpus was denoted as *MnoOOV*. The multigram training has 2 steps. Multigram dictionary and unit probabilities are estimated in the first step (on *MnoOOV*). Standard n-gram language model is then estimated (on *LnoOOV*) in the second step.

The sizes of above mentioned corpora are summarized in Table 1.

3.3 Confidence of terms

Link in a lattice represents one word or subword. Multiword terms or terms consisting of a sequence of subword units are represented by sequence of links in a lattice. The

System	Word accuracy		Word UBTWV			WrdSIZE
	1-best	lattice	ALL	IV	OOV	
WRD	69.38	83.11	73.3	73.5	72.8	0.200M
WRDRED	66.50	80.39	48.9	69.8	-	0.200M

Table 2: Comparison of baseline word recognizers with full (*WRD*) and reduced (*WRDRED*) vocabulary.

Notation	# of utters.	# of phones (incl. sil)	# of phones (w/o sil)
LnoOOV	237.2K	6.40M	5.60M
MnoOOV	143.5K	3.82M	3.35M

Table 1: Comparison of corpora used for multigram dictionary (MnoOOV) and language model (LnoOOV) training.

confidence measure, which is produced by the term detector, is the posterior probability of the term (link or sequence of links in a lattice).

4. EVALUATION

Conversational Telephone Speech (CTS) data from 2006 NIST Spoken Term Detection evaluations (NIST STD06) [4] were used in our experiments. For our tests, they are however not representative as the original NIST STD06 development term set for CTS contains low number of OOVs. Therefore, first of all, all terms containing true OOVs were omitted. Then, a set containing "artificial" OOV was defined. A limited LVCSR system was created (denoted by *WRDRED* which means "reduced vocabulary") where 880 words were omitted from the vocabulary. We selected 440 words from the term set and other 440 words from the LVCSR vocabulary. This system had reasonably high OOV rate on the NIST STD06 DevSet. The term set has 975 terms of which 481 are in vocabulary (IV) and are 494 OOV (terms containing at least one OOV) for the reduced system. The number of occurrences is 4737 and 196 for IV and OOV terms respectively. We can detect all the "artificial" OOV terms by the original full vocabulary LVCSR (denoted as *WRD*). All results are reported on the DevSet as NIST did not provide reference transcriptions for the EvalSet. System parameters (decoder insertion penalties and scaling factors) are tuned also on the DevSet. We evaluate word 1-best accuracy (word accuracy), word lattice accuracy (oracle), upper bound TWV and lattice size.

4.1 UBTWV - Upper Bound TWV

We used Term Weighted Value (TWV) for evaluation of spoken term detection (STD) accuracy of our experiments. The TVW was defined by NIST for STD2006 evaluation [4]

$$TWV(thr) = 1 - \underset{term}{average}\{p_{MISS}(term, thr) + \beta p_{FA}(term, thr)\}, \quad (5)$$

where β is 999.9. The $p_{MISS}(term, thr)$ is miss probability of the $term$ and given threshold thr. The $p_{FA}(term, thr)$ is the term false alarm probability.

One drawback of TWV metric is its one global threshold for all terms. This is good for evaluation for end-user environment, but leads to uncertainty in comparison of different experimental setups, as we do not know if the difference is caused by different systems or different normalization and global threshold estimation. This is a reason for *Upper Bound TWV* (UBTWV) definition which differs from TWV in individual threshold for each term. The ideal threshold for each term is found to maximize the term's TWV:

$$thr_{ideal}(term) = \arg\max_{thr} TWV(term, thr) \quad (6)$$

and UBTWV is then defined as

$$UBTWV = 1 - \underset{term}{average}\{p_{MISS}(term, thr_{ideal}(term)) + \beta p_{FA}(term, thr_{ideal}(term))\}. \quad (7)$$

It is equivalent to a shift of each term to have the maximum $TWV(term)$ at threshold 0. Two systems can be compared by UBTWV without any influence of normalization and threshold estimation. The $UBTWV$ was evaluated for the whole set of terms (denoted *UBTWV-ALL*), only for in-vocabulary subset (denoted *UBTWV-IV*) and only for out-of-vocabulary subset (denoted *UBTWV-OOV*).

4.2 Lattice Size

Using STD in large scale implies using an indexing technique where the size of index is important. That is why we do not calculate the size of lattice as the number of nodes or links. In contrary, we calculate lattice size as the number of indexed units.

Groups of the same overlapped words are found in the word or multigram lattice. Each group is substituted by one candidate and the count of such candidates is denoted *WrdSIZE*. Phone lattices are not processed phone-by-phone, but by indexing phone trigrams: phone trigrams are generated first, then the same procedure is applied as for the word lattices: groups of the same phone trigrams are identified and each group is substituted by one candidate. The count of such candidates is denoted *PhnSIZE*.

5. BASELINE SYSTEMS

Comparison of baseline LVCSR systems in the Table 2. The *WRD* system is LVCSR with the full 50k vocabulary. The *WRDRED* LVCSR system has reduced vocabulary as defined in section 4. Decoding parameters (word insertion penalty, language scaling factor and pruning coefficient) were tuned for the best STD accuracy (UBTWV) and fixed for further experiments.

5.1 Subword systems

We compared several subword systems. The first one is a simple phone loop and the others are multigram systems. Language models are applied on the phone or multigram units. The baseline accuracies are summarized in the subsections below. The phone accuracy for multigram systems is evaluated by switching the decoder from producing "word" labels (multigrams) to model labels (phones).

5.1.1 Phone loop system

The phone based STD accuracies are summarized in Table 3. The first cluster of systems *phn* has light pruning which is the same for all three language model orders. The second cluster (denoted as *phncs*) is an example of phone systems having reasonably large and comparable sizes of the index. This was achieved by severe pruning tuned separately for each language model order. The best UBTWV for OOV terms is achieved for bigram language model. Notice the size of the index (phone trigrams are indexed in this case) needed to achieve these relatively good results.

Unit	LM	Phn. UBTWV			PhnSIZE	prun.
	ngram	ALL	IV	OOV		
phn	1	37.6	35.7	42.2	33.7M	light
phn	2	49.7	45.4	59.7	37.4M	light
phn	3	49.4	45.3	59.0	9.6M	light
phncs	1	29.0	27.8	31.6	2.9M	severe
phncs	2	43.9	41.6	49.3	4.0M	severe
phncs	3	46.9	45.0	51.3	3.1M	severe

Table 3: Comparison of phone based system with different order of language model (trained on *LnoOOV*) and pruning.

5.1.2 Multigram systems

Our previous work compares several multigrams systems for phone recognition and STD tasks. We concluded that the best accuracy was achieved by *Non Cross Word Multigram* system with maximal length of unit 5. In this modification of multigram training, word boundaries were marked in the training corpus. Then a rule was incorporated into the training algorithm to not allow the word boundary symbol inside multigram units. Results of the *Non Cross Word Multigram* are summarized in Table 4. The best UBTWV-OOV accuracy is achieved with unigram language model. This system is denoted as *noxwrd*.

System	LM	Multigram UBTWV			WrdSIZE
	ngram	ALL	IV	OOV	
noxwrd	1	53.3	50.5	59.8	5.7M
noxwrd	2	62.2	64.7	56.3	2.3M
noxwrd	3	61.1	64.4	53.4	1.7M

Table 4: Comparison of multigram based system (trained on *MnoOOV*) with different order of language model (trained on *LnoOOV*).

To compare with state-of-the-art OOV detection systems, we also trained multigrams on the LVCSR pronunciation dictionary. As was shown in [1], training the OOV language model on a dictionary of words improves performance over just using the training corpus. This is because training the language model on phone/multigram transcriptions of sentences in the training corpus will favor more frequent units and the resulting OOV model then prefers these frequent units. Since OOV words are often unseen, training the language model on a dictionary with a weak language model leads to better performance.

The dictionary based system was built using only the LVCSR dictionary. Each pronunciation was taken as an utterance. Then multigram system was trained over these utterances and the language model over the multigrams was estimated. The baseline results of this simple system are summarized in Table 5. The best UBTWV-OOV accuracy is achieved also with the unigram language model. This system is denoted as *dict*.

System	LM	Multigram UBTWV			WrdSIZE
	ngram	ALL	IV	OOV	
dict	1	36.1	33.0	43.6	3.3M
dict	2	34.0	30.8	41.4	2.0M
dict	3	33.6	30.6	40.6	1.7M

Table 5: Comparison of accuracy of *dict* multigram system and different language model order.

5.2 Conclusion

The best performances on the OOV STD task are summarized in Table 6. The conclusion is that the best UBTWV for out-of-vocabulary words is achieved by the *Non Cross Word Multigram* (*noxwrd*) system and the worst accuracy by the *dict* based multigram system. All multigram systems have reasonably small sizes of the index. Note, that the index size of the phone system (trigram LM) is 2 times larger than the multigram one for the same UBTWV-OOV accuracy.

System	LM	Multigram UBTWV			SIZE
	ngram	ALL	IV	OOV	
phn	3	46.9	45.0	51.3	3.1Mp
noxwrd	1	53.3	50.5	59.8	5.7Mw
dict	1	36.1	33.0	43.6	3.3Mw

Table 6: Comparison of our baseline subword systems. Mp – milions of indexed phone prigrams, Mw – milions of indexed word unigrams.

6. RESULTS OF WORD-SUBWORD RECOGNITION

The first set of experiments (Table 7) compares the word accuracies (1-best and lattice) for in-vocabulary words only. It was confirmed that the modeling of OOV parts of speech in `<unk>` model positively influenced the in-vocabulary word accuracy. We obtained 0.85% absolute improvement on 1-best word accuracy and 5% absolute improvement on lattice word accuracy. The UBTWV for in-vocabulary words searched as word forms also slightly increases from 69.8% to 70.4%.

System	WORD acc.		UBTWV	WrdSIZE
	1-best	latt.	IV	
WRDRED	66.50	80.39	69.8	0.20M
WRDRED&phn	67.22	85.16	70.2	0.20M
WRDRED&dict	67.35	84.31	70.4	0.20M
WRDRED&noxwrd	66.78	85.30	70.2	0.20M

Table 7: Comparison of baseline and hybrid systems on word accuracy (1-best), word lattice accuracy (oracle) and UBTWV for in-vocabulary words.

We found that the best gain on the STD task was achieved by the *WRDRED&dict* system. So the *WRDRED&dict* system will be used for the following analysis of STD task with hybrid word-subword recognizer.

We have to tune all three parameters for scaling the subword LM in word LM. The subword language model scaling factor and the subword word insertion penalty had the greatest effect. The following experiments are done only tuning the *SLMSF* parameter to show what is happening inside.

We evaluate the STD accuracy on the in-vocabulary terms. The dependency of UBTWV-IV on the subword language model scaling factor is plotted in Figure 5. The best UBTWV-IV was achieved for the $SLMSF = 0.9$. The accuracy of

terms detected by the word art of the lattice (terms are in word forms) increases by 0.6% absolute. When the subword language model weight increases, the accuracy of in-vocabularies detected by subword part (terms are in multigram form) also rises. Note however that this is not really wanted, as the word and subword models compete in case of IV terms. If the word and subword detections are combined, we got another 0.5% improvement over the baseline *WRDRED* system.

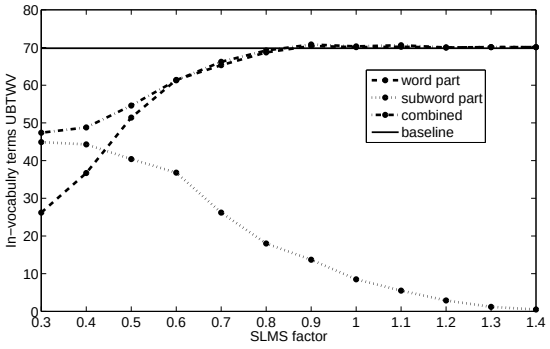

Figure 5: Dependency of the in-vocabulary terms UBTWV on the SLMS factor. Dotted: the subword detection accuracy (terms are in multigram form), Dashed: the word detection accuracy (terms are in word form), Dash-dotted: detection accuracy of combined word-subword detections, Solid: the baseline *WRDRED*.

The size of the word and subword parts of the lattice depending on the *SLMSF* factor are plotted in Figure 6. The lattice size of word baseline system (*WRDRED*) was $0.20M$ and the size of subword baseline system (*dict*) was $3.26M$.

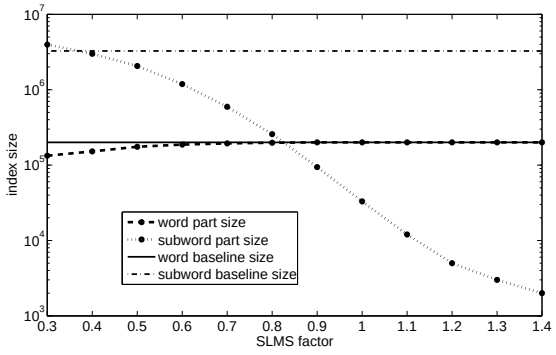

Figure 6: Dependency of the word-subword system's index size on SLMS factor. Dashed: word index size, Dotted: subword index size, Solid: word index size of the baseline (*WRDRED*), Dash-dotted: subword index size of the baseline *dict*.

Figure 7 compares the UBTWV accuracies of in-vocabulary and out-of-vocabulary term detection to the baseline of *dict* system. If the subword system is combined with the word system, the subword accuracy significantly improves (from 45.4% up to 62.3%). It is important to note that the accuracy of hybrid system on the OOVs is 2.5% higher than the accuracy of the best single *noxwrd* multigram system (59.8). This was achieved with only about 1/3 size of the index.

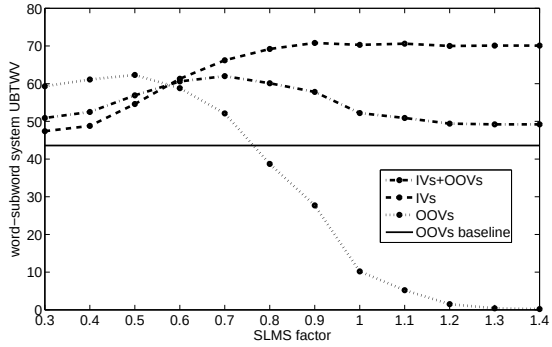

Figure 7: Dependency of the word-subword system UBTWV on the SLMS factor. Dotted: OOVs detection accuracy (by subwords), Dashed: IVs detection accuracy (by combined words-subwords), Dash-dotted: all terms detection accuracy of combined word-subword system, Solid: the OOV baseline of system *dict*.

A summary of the best STD accuracies for all, IV and OOV terms is in Table 8. The conclusion is that the *dict* subword model is the best for out-of-vocabulary words. The last set of experiments aimed to find the optimal values of parameters *SLMSF*, *SWIP* and *SC*. The goal was to find such values to maximize the accuracy of OOV terms and to maintain the baseline accuracy for IV terms. We achieved the best STD performance for $SLMSF = 1.0$, $SWIP = 1.5$ and $SC = 0.5$. The overall UBTWV was 62.7%. The UBTWV-IV was 69.6%, which is close enough to the baseline. The UBTWV for OOV terms was 44.7%. The OOV term detection accuracy was improved by 1% absolutely. The index size of OOV subpart was $0.40M$.

On the other hand, we can achieve higher accuracy if the word and subword systems are combined at the level of term detections. The OOV term detection accuracy of *noxwrd* system was 59.8%, which is about 13% better than the word-subword OOV detection accuracy, but the complexity of such combination must be taken into account. The decoding must be run two times, and the size of the index only for OOVs is $5.7M$.

7. CONCLUSIONS

Hybrid word-subword spoken term detection system proposed in this paper is a good alternative to the the combination of standalone word and subword systems. The system can achieve slightly worse accuracy (6.1% relative deterioration) than the combined standalone word and subword systems. The hybrid system is however simpler and has only 1/10 of the merged *WRD* and *noxwrd* system index size. Also the decoding is faster because it is run only once. The accuracies and index sizes are summarized in the Table 9.

8. ACKNOWLEDGMENTS

This work was partly supported by European project AMIDA (FP6-033812), by Czech Ministry of Interior (project No.

System	UBTWV								
	ALL	IV	OOV	ALL	**IV**	OOV	ALL	IV	**OOV**
WRDRED&phn	58.9	61.6	52.6	49.1	70.2	0.0	51.6	47.8	60.6
WRDRED&dict	62.0	66.2	52.1	60.8	71.0	37.2	56.9	54.6	62.3
WRDRED&noxwrd	56.6	57.3	54.9	50.0	70.7	2.0	51.8	48.8	58.7

Table 8: Comparison of hybrid systems with the SLMS constant tuned to obtain (1) the best UBTWV-ALL (the first cluster) – SLMS=, UBTWV-IV (middle cluster) and UBTWV-OOV (the last cluster).

System	UBTWV			WrdSIZE
	ALL	IV	OOV	
WRD&dict	62.7	69.6	46.7	0.6M
WRD and noxwrd	66.8	69.8	59.8	5.9M

Table 9: Comparison of the best hybrid system (*WRDRED&dict*) and combination of standalone word and subword (*WRD and noxwrd*) systems.

VD20072010B16), by Grant Agency of Czech Republic under project No. 102/08/0707 and by Czech Ministry of Education under project No. MSM0021630528. The hardware used in this work was partially provided by CESNET under project No. 201/2006. Lukáš Burget was supported by Grant Agency of Czech Republic, project No. GP102/06/383.

9. REFERENCES

[1] I. Bazzi. Modelling Out-of-vocabulary Words for Robust Speech Recognition, Ph.D. Thesis, MIT, 2002.

[2] J. Černocký et al. Search in Speech for Public Security and Defense. In *Proc. IEEE Workshop on Signal Processing Applications for Public Security and Forensics, 2007 (SAFE '07)*, pages 1–7. IEEE Signal Processing Society, 2007.

[3] S. Deligne and F. Bimbot. Language Modeling by Variable Length Sequences: Theoretical Formulation and Evaluation of Multigrams. In *Proceedings of ICASSP 1995*, pages 169–172, Detroit, MI, 1995.

[4] J. Fiscus, J. Ajot, and G. Doddington. The Spoken Term Detection (STD) 2006 Evaluation Plan, NIST USA, Sep 2006.

[5] T. Hain et al. The AMI Meeting Transcription System. In *roc. NIST Rich Transcription 2006 Spring Meeting Recognition Evaluation Worskhop*, page 12. National Institute of Standards and Technology, 2006.

[6] H. Jiang. Confidence Measures for Speech Recognition: A Survey. In *Speech Communication*, volume 45, pages 455–470. Science Direct, 2005.

[7] M. Karafiát, L. Burget, and J. Černocký. Using Smoothed Heteroscedastic Linear Discriminant Analysis in Large Vocabulary Continuous Speech Recognition System. In *2nd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms*, page 8, 2005.

[8] M. Mohri, F. Pereira, and M. Riley. Weighted Finite State Transducers in Speech Recognition, in ISCA ITRW Automatic Speech Recognition: Challenges for the Millenium, 2000 Paris, 2000.

[9] I. Szöke, L. Burget, J. Černocký, and M. Fapšo. Sub-word Modeling of Out of Vocabulary Words in Spoken Term Detection. In *submitted to Proceedings of Interspeech 2008*, 2008.

[10] I. Szöke et al. BUT System for NIST STD 2006 - English available from http://www.fit.vutbr.cz/speech/std/2006/, file but_06_std_eval06_eng_all_spch_p-but-stbu-merged_1.txt, but, dec 2006., 2006.

[11] A. Yazgan and M. Saraclar. Hybrid Language Models for Out of Vocabulary Word Detection in Large Vocabulary Conversational Speech Recognition. In *Proceedings of ICASSP 2004*, volume 1, pages 745–748, May 2004.

[12] P. Yu and F. Seide. A Hybrid Word / Phoneme-Based Approach for Improved Vocabulary-Independent Search in Spontaneous Speech. In *Proceedings of ICSLP 2004*, volume 1, pages 895–898, 2004.