Posterior-based Out of Vocabulary Word Detection in Telephone Speech

Stefan Kombrink[1], *Lukáš Burget*[1], *Pavel Matějka*[1], *Martin Karafiát*[1], *Hynek Hermansky*[2,1]

[1]Speech@FIT, Brno University of Technology, Czech Republic
[2]Johns Hopkins University, USA

{kombrink,burget,matejkap,karafiat}@fit.vutbr.cz, hynek@jhu.edu

Abstract

In this paper we present an out-of-vocabulary word detector suitable for English conversational and read speech. We use an approach based on phone posteriors created by a Large Vocabulary Continuous Speech Recognition system and an additional phone recognizer, that allows detection of OOV and misrecognized words. In addition, the recognized word output can be transcribed more detailed using several classes. Reported results are on CallHome English and Wall Street Journal data.

Index Terms: confidence measures, out-of-vocabulary word detection, phone posteriors, neural net, OOV

1. Introduction

Current speech recognition systems are customized to operate with a limited vocabulary on a restricted domain. Therefore, acoustic models are being trained on a target language, and the language model is designed to cover the most frequently expected words and multigrams. Under real conditions, however, such constraints of a restricted domain are violated very easily: Systems still have difficulties to deal with open vocabulary (foreign words, proper names) or accented speech and mispronunciations all of which are common in human speech. If a word is missing in the dictionary, the corresponding speech will be misrecognized in any case, and its semantical information is lost. Due to the contextual nature of the language model, the estimates for the surrounding words also tend to be invalid.

Given a speech signal x, the process of finding the most-likely uttered word sequence w can be considered as a search for the most appropriate model $M(w_i)$:

$$w = \arg\max_i P(M(w_i)|x) \tag{1}$$

Applying Bayes rule yields

$$w \propto \arg\max_i p(x|M(w_i)) \cdot P(M(w_i)) \tag{2}$$

where $P(M(w_i))$ describes the prior probability of word sequence w_i determined by the use of a language model. $p(x|M(w_i))$ describes the conditional probability of the speech input determined by the use of the acoustic model. As long as *in-vocabulary* (IV) speech (i.e. speech of the restricted domain) is concerned, misrecognized words in w are due to deficient modeling or difficult acoustic conditions. But since the prior probability of any OOV word sequence $P(M(w^{OOV}))$ according to such a model is zero, the system will forcibly map OOV speech input to an acoustically similar sequence of IV words w^{IV} with prior probability $P(M(w^{IV})) > 0$.

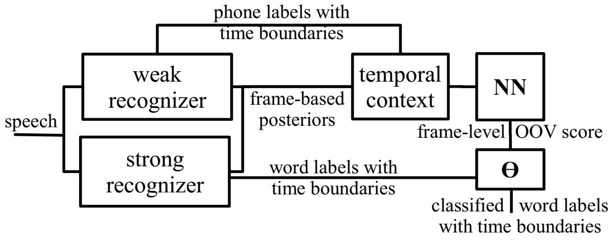

Figure 1: *Posterior-based OOV Word Detection.*

An attribute of current systems is apparently, that they still treat OOV words and IV words in the same way. Therefore, even though a standard speech recognizer is not able to deliver the correct sequence of words, a classification of the word output is still desirable - e.g. in order to detect speech aberrating from a strictly limited domain or repeatedly occuring OOV content. Furthermore, OOV words are considered valuable in information retrieval because they tend to carry semantics.

In order to identify OOV regions in speech, various confidence measures have been used to detect misrecognitions due to OOV content rather than just misrecognitions in general [2], [1]. Early work done in [1] and more recently by others in [2], [3] aims to detect and model OOV content on the level of language modeling.

2. Method

We proposed a new approach to detect OOV words in read speech (Wall Street Journal) in [4]. A single score was obtained by a neural net using two types of phone posterior input features from concurrent recognizers. Among all other tested single confidence measures, our posterior-based score estimated by the neural net performed the best.

In this work, we extended our technique and applied it to lower quality telephone speech (CallHome English), which is our main target data. We compare results to those obtained on read speech and show, that posterior-based OOV word detection generalizes to a reasonable extent across data and across the language model of the speech recognizer. In addition, our solution offers an alternative to binary OOV detection: full classification of the word output.

Figure 1 shows our current system combining the output of two recognizers: a large vocabulary continuous speech recognizer (strong recognizer) which is constrained by a language model, and a phone recognizer (weak recognizer).

The output of both recognizers consists of posteriors and

This work was partly supported by European projects DIRAC (FP6-027787) and AMIDA (FP6-033812), by Grant Agency of Czech Republic project No. 102/08/0707, and by Czech Ministry of Education project No. MSM0021630528.

Thanks to Daphna Weinshall for fruitful feedback.

Time	Class Label	Recognition	Reference
0.49	IV incorrect	THAT'S	LET'S
0.75	IV incorrect	A	SAY
0.93	IV correct	BACK	BACK
1.17	IV correct	TO	TO
1.31	Silence	<pau>	<pau>
1.33	OOV	BALANCE	BELGIUM
1.68	OOV	THEM	
1.87	IV correct	TO	TO

Figure 2: *Reference Classification Labels (IV=in-vocabulary).*

labels with time boundaries. Phone posteriors of both systems serve as input features to a neural net classifier. For any given input vector in time, the net estimates the probability of being out-of-vocabulary. We used the phone labels with time boundaries to preprocess the input features to contain temporal context, which improved the accuracy of the neural net considerably.

The strong recognizer also provides the recognized word sequence with timing information - the actual speech recognition output. By averaging the probability estimate we create a word-level score, and by thresholding it we classify the recognized words.

2.1. System Operation

Figure 2 shows an example of how the OOV detection system should ideally classify a recognized word sequence. The desired word classification is determined by an alignment of recognition (left) and reference (right) labels. We decided to distinguish between

- *sil* - no speech at all (both labels suggest silence)
- *ivcorr* - correctly recognized speech (word labels equal)
- *ivincorr* - misrecognized IV speech (word labels differ)
- *oov* - misrecognized speech due to OOV input

We prepared data with reference and recognition labels for classification. In cases, where OOV words in the speech partially or completely overlap (see “BELGIUM” in figure 3) we defined the desired classification as *oov*. Hence, we trained our neural net using a combination of frame-level phone posterior features extracted from both recognizers and their corresponding desired labeling. The neural net can now be used to estimate per-frame class probabilities from phone posteriors produced by the two recognizers using any speech data.

2.2. Patterns in Posteriors

Certain reoccuring patterns contained in the phone posteriors of both systems allow the neural net to learn the different word classes. In correctly recognized segments we find:

- agreement between strong and weak recognizer
- strong only: certainty about predicted phones

In the part covered by OOV input we find:

- disagreement between strong and weak recognizer
- strong only: confusion about predicted phones

In [5], error patterns in phone posteriors and the ideas behind our approach are examined thoroughly.

2.3. Posterior-based OOV Detection

Figure 3 shows the OOV word detection on a transcribed example from the evaluation set. On the top, phone posteriors from the strong and the weak recognizer are shown. They represent probability distributions over the phone set sampled with 10 ms frame length, and show how likely a certain phone has been uttered at a given time (the best path is pointed out by phone labels). Below are given reference and recognition labels showing the actual overlap of the OOV word “BELGIUM”. The estimated frame-by-frame OOV probability is plotted by the red curve.

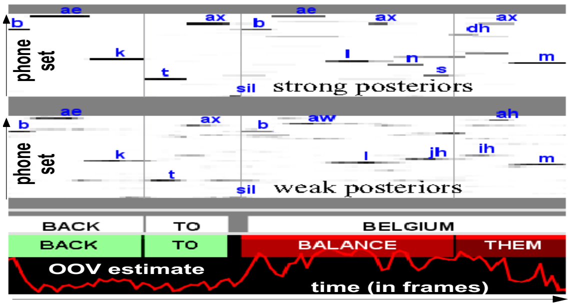

Figure 3: *Input and Output for the OOV Word “BELGIUM”.*

Vocabulary	Word Error Rate	OOV Word Rate
38385 words	24.9%	1.52%
2860 words	29.5%	5.74%

Table 1: *Hub5 eval01 Recognition Performance.*

3. System Setup

3.1. Weak Recognizer

Our phone posterior estimator based on a neural net processing long temporal trajectories of Mel-filter bank energies as previously used in [4] served as *weak recognizer* for the immediate estimation of phone posteriors with a sample length of 10 ms.

3.2. Strong Recognizer

The setup of the strong recognizer was derived from the 3-pass Large Vocabulary Continuous Speech Recognition (LVCSR) system used in [4]. We kept the acoustic models (trained on 250 hrs of Switchboard data) and replaced the language model with a 38k closed-vocabulary language model trained on more than 2000 hrs of conversational telephone speech (Switchboard 1+2, Fisher). Without changing the processing during the passes, we achieved a very decent performance using NIST scoring (see table 1).

Next, we substantially reduced the vocabulary size to 2860 words by removing words considered being rare according to their low unigram probabilities. As table 1 shows, the percentage of word errors increased slightly more than the percentage of OOV words contained in the eval01 data set.

The output of our strong recognizer were lattices containing both word and phone arcs attached with acoustic and language model scores. As acoustic features we used the posterior features estimated by the weak recognizer. We extracted the most likely path using the Viterbi algorithm and used it as the actual word output of the complete OOV detection system. Phone posteriors were extracted using the Forward-Backward algorithm on the lattices as explained in [8]. In both cases we used word insertion penalty -10 and language model scaling factor 32.

Class	CHE	WSJ
sil	17.6%	19.1%
ivcorr	53.2%	63.8%
ivincorr	20.6%	3.6%
oov	8.6%	13.5%

Table 2: *Time per Word Class.*

3.3. Data

A major criticism of the work done in [4] was the small amount of in-vocabulary misrecognitions in the data because misrecognitions had been mainly introduced by OOV words. By switching to conversational speech, both misrecognition classes were more balanced and reflect properties of more realistic data. Also, the inversion of the ratio between IV and OOV misrecognitions would show whether the approach is actually capable of discriminating between both classes of misrecognitions.

Table 2 compares the time per class in CallHome English (CHE) and Wall Street Journal (WSJ) data, after we introduced OOV words by manually reducing the sizes of the dictionaries (2k8 words CHE, 5k words WSJ). Information is given in time rather than word domain because the neural net operates on posterior frames with unity length. The amount of words in CHE was 26k in the evaluation and 159k in the training set and 7394 word types in total.

The CHE data consisted of three subsets. We inherited the given partitioning excluding about 10% of all sentences because they contained OOV words wrt. the 38k dictionary used to create the reference labels. The purpose and the statistics of the reduced subsets are shown in table 3.

3.4. Training and Evaluation Labels

We compared two types of word labels, which were provided with time boundaries, to create a reference classification for the training and evaluation set of the neural net. The *reference labels* were created using force-alignment of the speech given the reference transcription and the *large* dictionary (38k words). The *recognition labels* were created by the strong recognizer during speech decoding using the *small* dictionary (2k8 words). Thus, words contained in the 38k dictionary but excluded from the 2k8 dictionary represented OOV content. Finally, we combined it with the preprocessed phone posterior features of the weak and strong recognizer in order to train and evaluate the neural net.

3.5. The Neural Net

We used a 1-hidden layer multi layer perceptron (MLP) with 200 hidden neurons on preprocessed phone posterior features sampled per 10 ms frames. The 270-dimensional input layer accepted the current frame (45+45 phone posterior features) plus one left and one right context frame. Initially, we used a 3-dimensional output layer to distinguish between the word classes *sil*, *ivcorr*, *misrec* (= *ivincorr* ∪ *oov*) [4]. In this work we extended it to four classes in favor of better results.

The 4-dimensional output layer of the neural net assigns probabilities to a given frame of being in one of the following classes: *sil*, *ivcorr*, *ivincorr*, *oov*. A softmax function in the output layer yielded posterior probabilities of the output classes.

The objective function during the neural net training was the overall classification accuracy on frame level determined by choosing the class with maximum probability as estimate and comparing it to the reference classification. During each iteration, the learning rate was either kept or halved depending on the accuracy of the training data. Training was done up to ten iterations by using early stopping.

Subset	Length	OOV Word Rate	Used for
evltest	1.33 hrs	5.84%	Evaluation
devtest	2.13 hrs	4.92%	Cross Validation
train	8.54 hrs	4.93%	Training

Table 3: *Subsets and their Usage.*

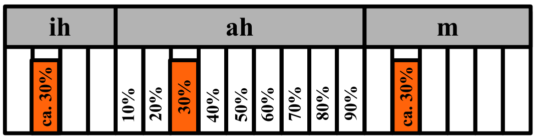

Figure 4: *Context (left+right) for a given Frame (middle).*

3.6. Phone Posterior Preprocessing

In [4] the use of temporal context was found to improve the performance of the neural net. Taking one preceding and one succeeding frame at a fixed distance of 7 frames and concatenating them with the original frame into a new feature vector of three times the original size showed optimal improvement. We expected this optimum to be correlated with the average phone duration (about 75ms).

In this work, we also preprocessed posterior features using *phone context*. Hence, for each frame we take into account dynamic context based on actual phone labels extracted from the phone posteriors of the weak system. Figure 4 shows the temporal context preprocessing for the recognized word “THEM” as second part of the OOV input “BELGIUM”. For a frame located at 30% within the “ah” vowel, phone context is created by adding one left and right frame accordingly located at approximately 30% within the adjacent phones. Similar to using fixed distance frames, we concatenated them into a single input feature vector of three times the original size. By doing this, we could drop the ad-hoc determination of the optimal fixed distance. Furthermore, phone context showed improvements in OOV word detection in cases of difficult OOV words on WSJ data [9].

4. Experiment and Results

4.1. Classification

Since all classifications were performed on word level, we averaged the frame-level scores within the word boundaries using arithmetic mean.

A *binary classification* on the recognized words was already sufficient for detecting OOV words. In this case, a threshold on the score determined an operating point and allowed to balance between the number of misses and false alarms. The two following binary classifications were performed commonly:

- *misrec* - $w \in$ *ivincorr* ∪ *oov* vs. $w \in$ *sil* ∪ *ivcorr*?
- *oov* - $w \in$ *oov* vs. $w \in$ *sil* ∪ *ivcorr* ∪ *ivincorr*?

In the misrecognition task we scored on $P(ivincorr) + P(oov)$ estimates and in the OOV detection task on $P(oov)$ estimates solely.

We also performed a *full classification* of the recognized word output by choosing the word class with the maximum estimated probability and comparing it to the reference class in

	sil	*ivcorr*	*ivincorr*	*oov*
Precision	94%	77%	52%	61%
Recall	91%	91%	45%	21%

Table 4: *Performance of Full Classification.*

Test	Train	3 Classes		4 Classes	
Data	Data	*misrec*	*oov*	*misrec*	*oov*
CHE	WSJ	26.80	**27.07**	25.83	25.80
CHE	CHE	23.56	**27.60**	23.48	**21.73**
WSJ	WSJ	17.19	11.90	17.52	**11.41**
WSJ	CHE	18.36	13.63	17.56	14.35

Table 5: *EER (in %) of NN-based MISREC/OOV Detection.*

our evaluation data. In this case, the overall word class accuracy was 78%. We observed similar accuracies on the frame level during neural net training. Precision and recall for each class is shown in table 4.

4.2. From three to four Classes

In our previous work, word entropy [4] was found to be the best lattice-based confidence score for OOV word detection. By using this measure, the EER for OOV detection raised from 18.62% (WSJ) to 35.08% (CHE) and made us expect a considerable performance drop-down for the neural net based score as well.

The two columns in the center of table 5 compare the performance of the *original 3-class* neural net on both data sets. Surprisingly, the net trained on WSJ data still performed better in OOV word detection on CHE than its CHE-trained counterpart (compare bold numbers). And while on WSJ data the detection of OOV words performed better than the detection of misrecognized words, it was the other way around on CHE data. We assumed that the data properties previously shown in table 2 forced the neural net to specialize too much on the subclass with the higher prior while learning the *misrec* class, and a proper distrimination of *ivincorr* and *oov* could be beneficial.

Thus, we retrained the neural net *using 4 classes* in the output layer. As a matter of fact, this solved the unwanted behavior we observed before. While the EER for misrecognitions did not change significantly, it improved for OOV word detection (see the bold numbers in the two right columns in table 5).

Generalization across data improved for detection of misrecognitions, whereas for OOV detection it degraded. Obviously, the net learned some properties of OOV words which are unique to WSJ and CHE data. We suggest training on WSJ and CHE data together to see if this again improves the generalization of OOV word detection.

4.3. Overall System Performance

Finally, figure 5 shows the *detection-error tradeoff* (DET) curve of *misrec*, *oov* and a third classification task performed on CHE using the 4-class net scores. During the third task, an oracle told whether a recognized word was in *misrec* or *corr*. Thus, all correctly recognized words were omitted from the OOV scoring. This task shows that the neural net was actually able to distinguish between *oov* and *ivincorr*. Full classification OOV detection results can be observed on this DET curve with the operating point set to the false alarm probability of 1.2% yielding a miss probability of 79%.

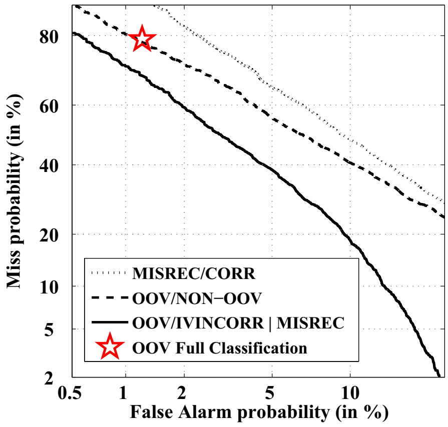

Figure 5: *Performance of Binary Classification Tasks.*

5. Conclusion

The posterior-based OOV word detection approach generalizes across data (clean speech, 16kHz vs. noisy speech, 8kHz) and across varied language models (read speech, 5k words vs. spontaneous speech 2k8 words) with some performance degradation. The 4-class neural net improves classification performance and allows scoring with a single class or any conjuncted class probability. Evaluation can be performed either on binary detection or on a classification with two up to four classes.

6. References

[1] Fetter, P., "Detection and Transcription of Out-of-Vocabulary Words in Continuous Speech Recognition", Ph.D.thesis, TU Berlin, 1998

[2] Hazen, T. J. et al, "A comparison and combination of methods for OOV word detection and word confidence scoring", ICASSP, pp. 397–400, 2001

[3] Yazgan, A. et al, "Hybrid language models for out of vocabulary word detection in large vocabulary conversational speech recognition", ICASSP, 2004

[4] Burget, L. et al, "Combination of strongly and weakly constrained recognizers for reliable detection of OOVs", ICASSP, 2008

[5] Hannemann, M. "Combinations of Confidence Measures for the Detection of Out-of-Vocabulary Segments in Large Vocabulary Continuous Speech using differently constrained Recognizers", Studies thesis, Otto-von-Guericke University Magdeburg, 2008, `http://www.iesk.ovgu.de/iniesk_media/bilder/ks/publications/theses/student_paper/studar_mh.pdf`

[6] Fiscus, J. G., "A post-processing system to yield reduced error rates: ROVER", Proceedings of the IEEE ASRU Workshop, pp. 347–354, 1997

[7] F. Wessel et al, "Confidence measures for large vocabulary speech recognition", IEEE Trans. Speech and Audio Processing, vol. 9, no. 3, pp. 288–298, 2001

[8] Evermann G. et al, "Posterior probability decoding, confidence estimation and system combination", Proc. Speech Transcription Workshop, 2000

[9] Kombrink S. et al, "Out-of-vocabulary detection in LVCSR using neural networks", Student EEICT, BUT Brno, 2008