OOV DETECTION IN LVCSR USING NEURAL NETWORKS

Stefan Kombrink
Erasmus Exchange Programme, Stuttgart University - FIT BUT
E-mail: katakombi@gmail.com

Supervised by: Lukas Burget, Pavel Matejka
E-mail: burget@fit.vutbr.cz,matejka@fit.vutbr.cz

ABSTRACT

Confidence measures and classifying techniques are widely used for the recognition error detection task in LVCSR (Large Vocabulary Continuous Speech Recognition).
But in many recognition scenarios the amount of words not included in the dictionary (e.g. real names, neologisms) lead to so-called OOV (Out Of Vocabulary) errors which increase the WER (Word Error Rate) even more.
The hereby described work acknowledges and investigates further improvements of an OOV detection task performed by combining strong and weak phone posterior features using neural networks based on [ICASSP08] and the use of phone context.

1 INTRODUCTION

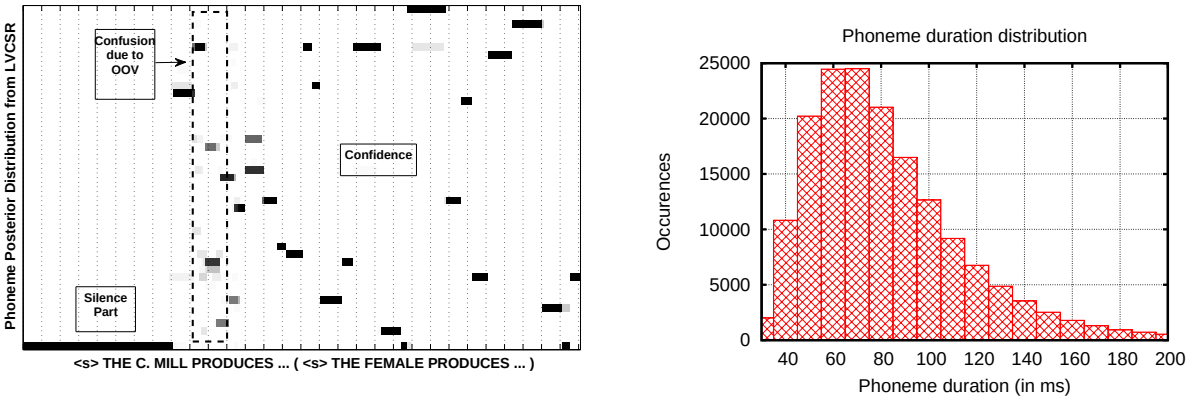

Figure 1: posterior distribution (*left*) , phone length distribution (training set) (*right*)

The reason why OOV words usually decrease overall performance so badly is that LVCSR systems commonly operate on N-Gram based Language Models. Those spread the error due to their contextual nature to previous and following words. Therefore a well-performing OOV detection seems promising for helping to improve existing LVCSR systems.
One approach to detect OOV words can be done by watching the distribution of frame-by-frame phone posteriors [ICASSP08].

Two different types are generated: One from a strongly constrained system using a language model (the CTS LVCSR derived from AMIDA, later on referred to as *lvcsr*) and one from a weakly constrained system (the FIT BUT NN-based phone recognizer, later on referred to as *phnrec*).
Figure 1 (left) shows strong posteriors created for one utterance over all frames representing the phone set (44 phones + silence). Probabilities are expressed by blackening degree and sum up to one in any given frame of 10ms duration. In case of OOV words confusion can be found in the distribution whereas a weakly constrained system will remain unaffected.
Posteriors had been extracted from the phone lattices of both systems. Training and testing lattices consisted of misrecognized words both caused by OOV and IV (In Vocabulary) words. To perform the OOV detection task a neural net has been trained on such posteriors to determine the frame-by-frame probabilities for silence, IV (in vocabulary) and OOV. For a detailed description of how lattices with posteriors can be generated see [Wessel01] and [ICASSP08].

2 NEURAL NET PARAMETERS

To determine the optimal parameters for classifying a particular frame a Hidden Layer MLP (Multi Layer Perceptron) is trained. Several adjustments have been tried, namely

- **number of neurons** (no hidden layer, 4,7,13,25,50,100,200 neurons in one hidden layer)
- **posteriors** (*lvcsr w/o silence*, *lvcsr*, *lvcsr+phnrec*, same with word entropy [ICASSP08])
- **context** (no context, frame context)

The ideal choice for the number of neurons was the MLP with 50 neurons.
Best choice for posteriors turned out to be *lvcsr+phnrec* posteriors.
The use of context was found to improve the most with best results for a 130ms to 150ms window. Instead of a single frame (no context) three frames within a window were used as input while omitting a certain amount of frames.
Trying out partially and shuffled training sets or forced realigns of input labels clearly revealed optimal settings while adjusting learning rate did not affect performance at all.
These results are consistent with the ones in [ICASSP08].
Noticeably the average duration of a phone is about 75 ms which means the best performing frame context tend to catch posteriors from adjacent phones.

3 PHONE CONTEXT

However, figure 1 (right) shows clearly that using fixed length context is an approximation only. This led to the assumption that using true phone context would improve even more.
Furthermore the phone lengths of the lvcsr phone lattices were supposed to be not as good as reference as the one from the phone recognizer. In case of misrecognitions those phone lattices are forced to align their best path through certain phones which never have occurred and thus yielding in minimum possible duration which is 30 ms because of the 3 state HMMs.
Posteriors of previous, current and following phones were concatenated to form one feature vector. For every central frame the prorated frame for the preceding and following phone is taken as context, e.g. as seen in figure 2 (left) the frames corresponding to 33% of `/h/` and 33% of `/m/`. Taking the central frames of the adjacent phones had also been tried but performed worse. Pseudo phones with zero duration and `sils` have been omitted.
The following combinations have been tried for building the new feature vector:

- **phnrec aligned** (*lvcsr/phnrec* with context both aligned according to *phnrec* phone labels)
- **lvcsr aligned** (*lvcsr/phnrec* with context both aligned according to *lvcsr* phone labels)
- **aligned** (*lvcsr/phnrec* context with aligned according to *lvcsr/phnrec* phone labels)
- **cross aligned** (*lvcsr/phnrec* with context aligned according to *phnrec/lvcsr* phone labels)

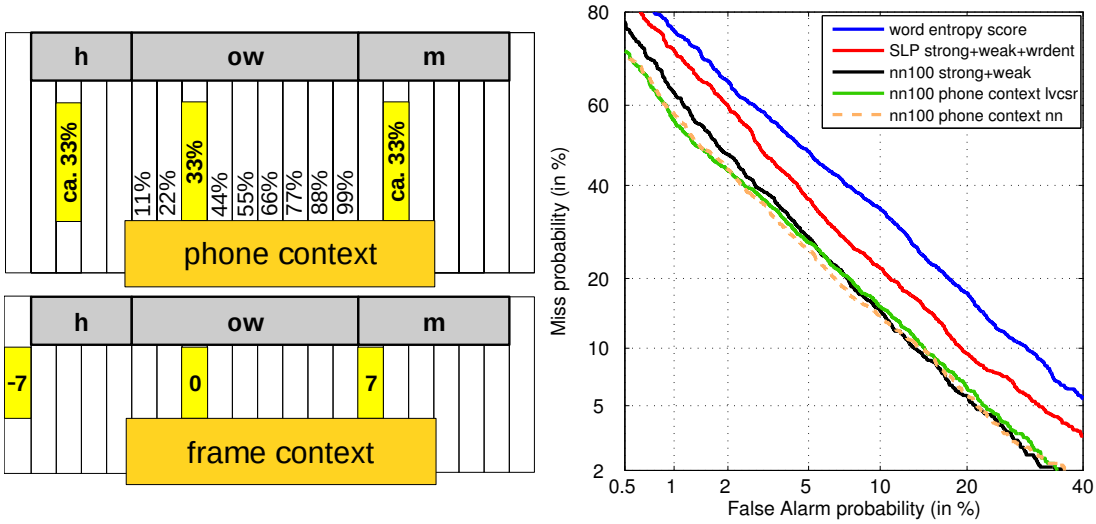

Figure 2: frame context vs. phone context (*left*), OOV detection with phone context (*right*)

The first combination performed the best as seen in figure 2 (right) which seems reasonable since it should reflect the actual phone boundaries the best. Adding word entropy score as an additional input to the neural net gave worse results only.

4 FURTHER PROCEEDINGS

In [ICASSP08] several additional scores had been merged with the neural net output using a maximum entropy model. Phone Context should be merged with scores such as Word Entropy in the complete setup to see whether it improves OOV detection.

Furthermore the environment used in [ICASSP08] had been degraded artificially to behave like a LVCSR with OOVs by limiting the dictionary to the most frequent words. In spontaneous speech (e.g. the CallHome database) word accuracy is expected to be much lower and containing words that are OOV when an arbitrary dictionary is being used.
A phenomenon specifically for spontaneous speech are unfinished words (or restarts) which usually can be treated like OOVs: About 0.8% of the words contained in the CH English speech were found to be restarts. Instead of building a language model that can work around these sequences somehow, it might be possible to even train a language model without them and try to detect those restarts using a derivation of the previously built OOV detection framework.

REFERENCES

[ICASSP08] Burget, L. et al: Brno University of Technology, accepted to ICASSP 2008: "Combination of Strongly and Weakly Constrained Recognizers for reliable detection of OOVs", Las Vegas, NV, USA

[Wessel01] 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, Aachen, DE, 2001.