Detection of Out-of-Vocabulary Words in Posterior Based ASR

Hamed Ketabdar[1,2], *Mirko Hannemann*[1], *Hynek Hermansky*[1,2]

[1]IDIAP Research Institute, Martigny, Switzerland
[2]Swiss Federal Institute of Technology at Lausanne, Switzerland
{hamed.ketabdar, mirko.hannemann, hynek.hermansky} @idiap.ch

Abstract

Over the years, sophisticated techniques for utilizing the prior knowledge in the form of text-derived language model and in pronunciation lexicon evolved. However, their use has an undesirable effect: unexpected lexical items (words) in the phrase are replaced by acoustically acceptable in-vocabulary items [1]. This is the major source of error since the replacement often introduces additional errors [2, 3]. Improving the machine ability to handle these unexpected words would considerably increase the utility of speech recognition technology.

Index Terms: out-of-vocabulary words, confidence measures, out-of-context phone posterior, in-context phone posterior, posteriors comparison

1. Introduction

1.1. Machine speech recognition

The speech signal X is assumed to be produced by a model M. The probability $p(M|X)$ that the model M produced the data X is proportional to

$$p(M|X) \sim p(X|M) \cdot p(M) \quad (1)$$

where $p(X|M)$ represents so called acoustic likelihood that indicated how likely is the data X given the model M (this likelihood can be derived from training acoustic data) and $p(M)$ indicates a prior probability of the model M. Efficient use of $p(M)$ is essential. Without its use, almost an order-of-magnitude increase in word errors has been observed [4].

In the current continuous recognition of speech, just as in the human speech recognition, the individual words are recognized in context of other words. The model M typically represents a sequence of words that are then represented by a sequences of phoneme models. Thus, the words are not recognized one-by-one as they come but the decision about the best matching string of words is delayed until the last element of the recognized phrase is processed. This principle of the delayed decision is one of the most powerful principles of the HMM-based ASR.

The prior probability of the model $P(M)$ is given by a product of prior probabilities of the individual words in the model (typically derived from large amounts of text data), together with an expected lexicon that lists the words that are to be recognized. When the word is not in the lexicon, it's prior probability is zero and the prior probability of any model M that would contain such a word is also zero. Thus, the out-of-vocabulary (OOV) word that is not in the lexicon can never be recognized and is in recognition replaced by another acoustically similar word or a sequence of shorter words. Quite likely, due to the global match of the whole utterance X with the sequence of words represented by the model M, other words in such an utterance may be also recognized in error.

In communication, the unexpected items (words) carry more information than the expected ones [5]. Thus, misrecognizing the unexpected out-of-vocabulary (OOV) words is one of the most serious problems of the current ASR.

1.2. The context in human speech recognition

No doubt that what we believe we hear is heavily influenced by what we expect to hear. Thus in human speech recognition, just as in the ASR, the context of the message that limits the number of possible alternative words, which could occur in a given part of the message in speech, is an important element that contributes to the decoding of the message.

However, the similarity between human and machine recognition of speech ends here. As pointed out by Allen [6], a possible model that accounts reasonably well for the effect of the context in human recognition of words and speech sounds has been proposed by Boothroyd and Nittrouer [7] who show on human recognition data that the probability of correct recognition of the words in context p_W relates to probabilities of the correct recognition of words without context p_N trough

$$p_W = 1 - (1 - p_N) \cdot (1 - p_C) \quad (2)$$

where p_C indicates the contribution of the probability due to the context channel. This model is different from (1) and implies that error from the acoustic channel and the context channel multiply, i.e. *the context contributes an independent parallel channel of information, which contributes to the decoding of the message in addition to the sensory (acoustic) channel.*

The parallel architecture of the model is intuitively appealing. It implies that it is not necessary for the both channels to be correct. When either of the channels, the acoustic one or the context one, is providing the correct evidence, the system makes the correct decision. Further, Allen also discusses some relevant physiological data of van Petten et al. [8], which indicate that the human cognitive system provides *instantaneous* response when encountering the incongruous word, thus further supporting the parallel character of both the sensory and the context channels. More recent work [9] shows that similar but even stronger reaction is observed for non-words.

2. An engineering system for discovery of unexpected words

When a human listener encounters a clearly pronounced and uncorrupted unknown word, she is typically immediately aware of its novelty and can choose an appropriate action. It would be very desirable if a machine could that too. This problem is well recognized and efforts to address it are ongoing (see e.g.

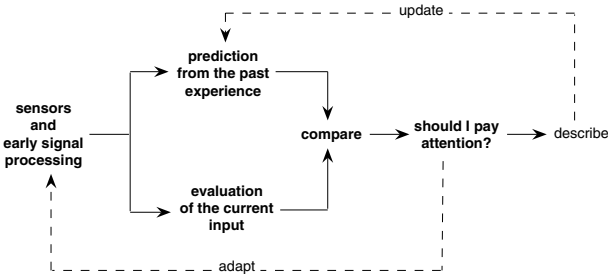

Figure 1: *General scheme of discovery and dealing with unexpected low prior probability stimuli.*

[3] for a review). Many of these techniques [10, 13, 14, 15] are based on segmenting the utterance into phones and words and evaluating a likelihood or posterior based measure for the hypothesized word inside the detected segments.

We approach this problem following the general strategy illustrated in Fig. 1. This strategy of comparing results in different processing streams is conceptually similar to earlier work [2] that proposes use of variable weighting γ in (1), where significant changes in the final best match with changes in the γ may indicate presence of an OOV word.

The similarity is in the assumption of the existence of several parallel processing streams with different levels of top-down influence, all triggered by the input sensory data, while the difference is that in our case one stream employs the prior knowledge fully and estimates what we call *in-context* posterior probabilities of speech sounds, and the other stream uses only the sensory data in estimating the *out-of-context* posteriors. Further, rather than comparing final results of the search in the *ad hoc* fashion as in [2], we compare frame-by-frame phoneme posterior estimates using relative entropy (KL divergence) measure.

The technique is especially simple to apply in systems that already use ANN estimated posteriors, as in the HMM/ANN hybrid ASR [12] or in TANDEM-based systems [18]. Its use in conjunction with the HMM/ANN hybrid ASR is illustrated in Fig. 2. Comparisons of posterior probabilities of phonemes for each individual speech frame (i.e. in equally spaced intervals of 10 ms) are utilized to identify the unexpected words.

The upper path from the Fig. 1 (estimating the in-context posteriors) represents the conventional HMM/ANN hybrid recognizer and provides in-context phoneme posteriors derived with the use of the prior constraints such as the knowledge of the expected lexicon and prior word probabilities provided by the applied language model. The lower path from the Fig. 1 that estimates the out-of-context posteriors is represented by the ANN-based classifier, trained to estimate posterior probabilities of context-independent phoneme classes [12].

Comparison of in-context posteriors and out-of-context posteriors provides an indication of the effect of the context. The comparison is done based on measuring Kullback-Leibler (KL) divergence between the two sets of estimates of posterior probabilities. When encountering an unexpected word, the in-context posteriors significantly deviate from the out-of-context posteriors because the unexpected word is not supported by the prior knowledge.

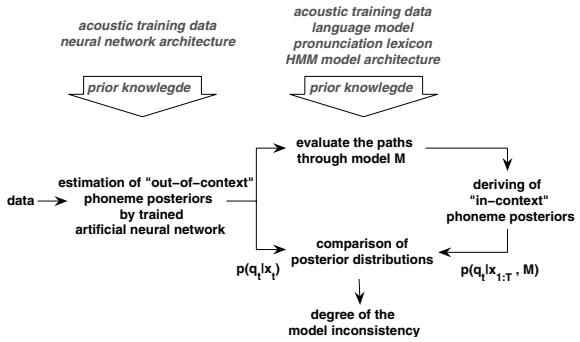

Figure 2: *Discovery of out-of-vocabulary words using hybrid HMM/ANN ASR system, in which the 'out-of-context' posterior probabilities estimated by the ANN are used in combination with the model constraints to derive the 'in-context' posteriors*

2.1. Out-of-context posteriors

Our scheme assumes the existence of the processing stream that could reliably describe the out-of-context posteriors with a minimal use of the prior knowledge. For this we apply our recently developed MRASTA technique [16] that employs multiple multi-resolution projection of relatively long segments of the auditory-like time-frequency plane.

2.2. In-context posteriors

The out-of-context posteriors are used in a search for the most likely Hidden Markov Model (HMM) sequence that could have produced the given speech phrase. As well known, for any given instant of the message, the HMM can also yield estimates of posterior probabilities of the hypothesized phonemes [10, 11] using the Baum-Welch estimation procedure. As shown later, the Viterbi approximate estimates can also be used.

2.3. Comparing in-context and out-of-context posteriors

To detect unexpected words, the difference between the two channels is evaluated. A large difference may indicate the unexpected word. In this work, we use Kullback-Leibler (KL) divergence to evaluate the difference between the two vectors of posteriors. The frame level KL divergence as a function of time is then smoothed by a moving average filter to emphasize word-level mismatch between two posterior streams. An unexpected word is indicated by smoothed KL divergence being above the pre-set threshold.

3. An example

An example adopted from [19] is shown in Fig. 3. The utterance contains 'five three zero', where the word 'three' represents an unexpected word, not present in the vocabulary. The upper part shows the in-context posteriors, the middle part the out-of-context posteriors, and the lower part shows the smoothed KL divergence between the two. As it can be seen, there is a region with major divergence corresponding to the word 'three' (which is marked roughly by dashed lines). An inconsistency between these two information streams, reflected in the increase in the KL divergence, could indicate an unexpected out-of-vocabulary word.

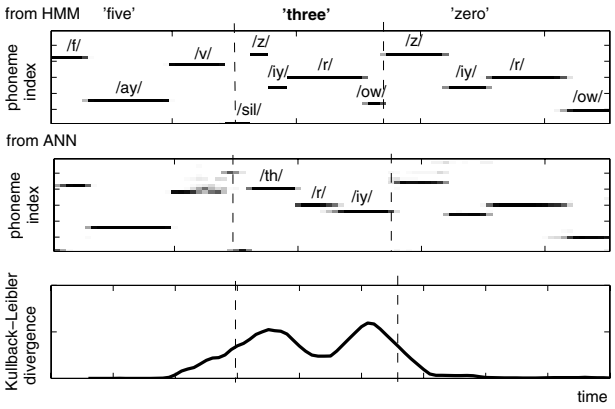

Figure 3: *Posterior probabilities (posteriograms) of phonemes estimated by an HMM-based system (the upper part of the Figure), and by ANN (the middle part of the Figure). KL divergence between the two estimates is shown smoothed by 100 ms square time window as a function of time in the lower part of the figure.*

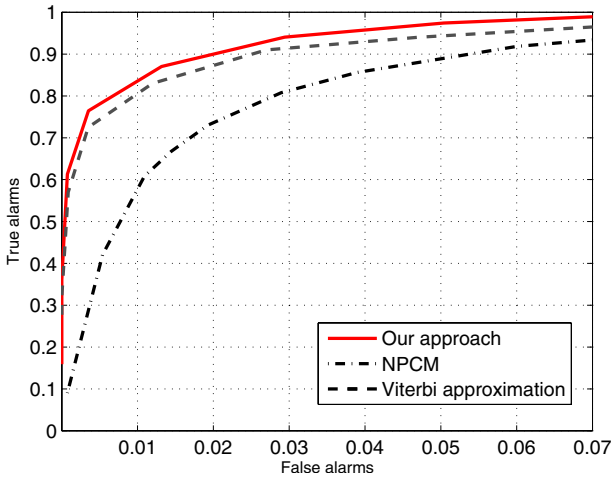

Figure 4: *Receiver operating curves (ROC curves) for our confidence measurement approach and conventional methods (phone-based and frame-based NPCM). The y axis is showing the percentage of true alarms and the x axis is showing the percentage of false alarms. Our approach shows a significantly better trade off (larger area under the ROC curve).*

4. Experiments and results

In this section, we report the initial results in detecting unexpected words. In the first set of experiments, we have used the OGI digits database [17]. The digits contain only 29 context-independent phones (monophones). We have introduced each of the words individually as an unexpected word by removing it from the vocabulary.

There are 2169 utterances in the test set and 2547 utterances in the training set. Prior probabilities of the occurrence of digits are all equal and the prior knowledge is in this case represented by a lexicon that contains 10 digits, with the remaining 11th digit representing the OOV word. All digits took their turn to represent the OOV word.

The frame-by-frame KL divergence trajectory is smoothed by a moving average filter with the length of 10 frames. The smoothed divergence measures are used as confidence measures and compared with a threshold to make a decision on detecting the unexpected word.

Comparison with two other conventional posterior based confidence measures [10, 15], that were proposed for use in conjunction with the MLP-estimated posteriors (Normalized Posterior based Confidence Measure, NPCM) and are based (as also many similar ones [13, 14]) on evaluating ANN-derived posterior estimates inside the detected phoneme segments for the hypothesized word, is illustrated in Fig. 4. This Figure shows the Receiver Operating Characteristic (ROC) curves obtained by our method and by the conventional posterior based methods. Our approach shows a noticeably larger area under the ROC curve (better trade off between true and false alarms).

Finally, Fig. 4 also shows what happens when the Baum-Welch posterior probability estimates are replaced by the computationally much less expensive Viterbi approximation. Then the posterior estimates degenerate to one for the phoneme class on the best path and zero for all other phoneme classes, but the rest of the procedure remains unchanged. As evident, the Viterbi approximation is in this case almost equally effective. Since the Viterbi back-tracking alignment can be obtained during the recognition at almost no additional cost, this may turn into a major advantage for the applications of the proposed technique.

5. Discussion and conclusion

A new technique for discovery of unexpected out-of-vocabulary words, which is based on comparison of two phoneme posterior streams derived from the identical acoustic evidence while using two different sets of prior constraints, and which does not require any segment boundary decisions, has been proposed and evaluated on a small vocabulary task, where it leads to better performance than some earlier reported posterior based confidence measures.

In this technique, the out-of-context phone posteriors are estimated by an MLP. The phone posteriors in the context channel are estimated using an HMM that employs prior contextual knowledge. This HMM layer uses the MLP estimate out-of-context posteriors as the state emission probabilities. The posterior estimates in the two channels are compared by evaluating the KL divergence at each frame. The frame-by-frame divergence track is then smoothed and compared with a threshold to decide if there is an unexpected word.

5.1. Towards larger vocabularies

The digit recognition reported above used context-independent phoneme classes and single state phoneme HMM models. Therefore, obtaining the in-context posteriors was straightforward. A Large Vocabulary Continuous Speech ASR (LVCSR) yields a set of additional challenges. However, we see no conceptual problem with applications of our technique in LVCSR.

First, the LVCSR typically employs context-dependent phoneme classes and multi-state phoneme models. This can be dealt with since the context-dependent phonemes could be mapped on the context-independent ones, the multi-state models can be merged into one-state, and the respective posteriors summed.

Further, the Baum-Welch estimation of in-context posterior probabilities of phonemes $\gamma(i,t)$ requires going through all possible paths into and out-from the current phoneme state. This operation can quickly grow in difficulty with increase of the anticipated lexicon.

One way of approaching this problem could be in breaking the process into two steps. In the first step, the full recognition of the utterance is carried out, yielding the K-best word lattice. In the second step the vocabulary for the OOV detection could be limited to only those words that appear in a K-best word lattice from the first step. This is carried out individually for each utterance in the test.

When dealing with the OOV word, the first ASR step obviously yields a number of erroneous hypothesis. However, the assumption is that all correct within-vocabulary words that appear in the tested utterance would appear in the K-best paths of the word lattice. In this way, a large vocabulary task encountered in the first step can be turned into a smaller-vocabulary utterance-specific problem for the OOV detection search.

Additionally, it appears that the Viterbi approximation can be applied for the estimation of the in-context posteriors. If this would be the case, the estimation of both the in-context and the out-of-context posteriors in the posterior-based ASR could be trivial. This assumption still remains to be fully tested for more difficult tasks than the digit task reported here but our initial experiments with LVCSR, though not reported here, are so far positive.

5.2. Several additional thoughts

When the unexpected OOV word is detected, an attempt may be made to describe the word in terms of its acoustic elements (phonemes). Such a phonetic description could then be used for updating the lexicon of the available models. This could lead to an ASR system that would be able to improve its performance as being used over time, i.e. to learn.

The inconsistency between in-context and out-of-context probability streams does not only have to indicate the presence of unexpected lexical items but can also indicate any other inadequacy of the model that could possibly be alleviated.

A large divergence could also result from corrupted input data when the in-context probability estimation using the prior knowledge could yield more reliable estimate than the unconstrained out-of-context stream. Providing and using a measure of confidence in the estimates from the two individual information streams would allow for a disambiguation. Such a confidence measure is a topic of our current research interest.

6. Acknowledgements

This work was supported by the European Community 6th FWP IST integrated projects DIRAC and AMI, by the Swiss National Science Foundation through the National Center of Competence in Research (NCCR) on Interactive Multimodal Information Management (IM2).

7. References

[1] Hermansky, H. and Morgan, N., "Automatic Speech Recognition", in "Encyclopedia of Cognitive Science", L. Nadel, Ed., Nature Publishing Group, Macmilian Publishers, 2002.

[2] Chase, L., "Error-Responsive Feed Back Mechanisms for Speech Recognizers", PhD Thesis, Carnegie Mellon University, April 11, 1997.

[3] Bazzi, I., "Modelling Out-of-Vocabulary Words for Robust Speech Recognition", PhD. Thesis, MIT, Department of Electrical Engineering and Computer Science, 2002.

[4] Lee, K.-F., Hon, H.-W. and Hwang, M.-Y., "Recent Progress in the SPHINX Speech Recognition System", Proc. DARPA Speech and Natural Language Workshop, Philadelphia, Pennsylvania, 1989.

[5] MacKay, D.J.C., "Information Theory, Inference, and Learning Algorithms", Cambridge University Press, 2004.

[6] Allen, J.B., "Articulation and Intelligibility", Morgan & Claypool Publishers, 2005.

[7] Boothroyd, A. and Nittrouer, S., "Mathematical treatment of context effects in phoneme and word recognition", J. Acoust. Soc. Am., Vol. 84(1), 1988, pp. 101-114.

[8] Van Petten, C., Coulson, S., Rubin, S., Plante, E., Parks, M., "Time course of word identification and semantic integration in spoken language", J. Exp. Psychol.: Learning, Memory, and Cognition., Vol. 25(2), Mar 1999, pp. 394-417.

[9] McLaughlin, J., Osterhout, L., Kim, A., "Neural correlates of second-language word learning: minimal instruction produces rapid change", Nature (Neuroscience), Vol. 7, 2004, pp. 703-704.

[10] Bourlard, H., Bengio, S., Magimai Doss, M., Zhu, Q., Mesot, B., and Morgan, N., "Towards using hierarchical posteriors for flexible automatic speech recognition system", DARPA RT-04 Workshop, Palisades NY, November 2004, also IDIAP-RR 04-58.

[11] Ketabdar, H., Vepa, J., Bengio, S., and Bourlard, H., "Using More Informative Posterior Probabilities For Speech Recognition", Proc. of ICASSP'06, Toulouse, France, 2006.

[12] Bourlard, H. and Wellekens, C.J., "Links between Markov Models and Multilayer Perceptrons", IEEE Conference on Neural Information Processing Systems, 1988, Denver, CO, Ed. D. Touretzky, Morgan-Kaufmann Publishers, pp. 502-510, 1989.

[13] Hazen, T., et al., "Recognition confidence scoring for use in speech understanding systems", Proc. of ISCA ASR2000 Tutorial and Research Workshop, Paris, 2000.

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

[15] Williams, G., and Renals, S., "Confidence Measures for Hybrid HMM/ANN Speech Recognition", Proc. of Eurospeech'97, Rhodes, 1997.

[16] Hermansky, H. and Fousek, P., "Multi-resolution RASTA filtering for TANDEM-based ASR", Proc. of Interspeech 2005, 2005.

[17] Cole, R., Noel, M., Lander T. and Durham T., "New Telephone Speech Corpora at CSLU", Proc. of EUROSPEECH'95, Madrid, 1995, pp. 821-824.

[18] Hermansky, H. and D.P.W. Ellis and S. Sharma, "Connectionist Feature Extraction for Conventional HMM Systems", in Proceedings of the International Conference on Acoustics, Speech and Signal Processing, Istanbul, Turkey, 2000

[19] Ketabdar, H. and Hermansky, H., "Identifying and dealing with unexpected words using in-context and out-of-context posterior phoneme probabilities", IDIAP Research Report 06-68, 2006.