IDIAP RESEARCH REPORT

Identifying unexpected words using in-context and out-of-context phoneme posteriors

Hamed Ketabdar [a,b]   Hynek Hermansky [a,b]

IDIAP–RR 06-68

September 2006

[a] IDIAP Research Institute
[b] EPFL

IDIAP Research Institute **www.idiap.ch**
Rue du Simplon 4 P.O. Box 592 1920 Martigny – Switzerland
Tel: +41 27 721 77 11 Fax: +41 27 721 77 12 Email: info@idiap.ch

IDIAP Research Report 06-68

Identifying unexpected words using in-context and out-of-context phoneme posteriors

Hamed Ketabdar Hynek Hermansky

September 2006

Abstract. The paper proposes and discusses a machine approach for identification of unexpected (zero or low probability) words. The approach is based on use of two parallel recognition channels, one channel employing sensory information from the speech signal together with a prior context information provided by the pronunciation dictionary and grammatical constraints, to estimate 'in-context' posterior probabilities of phonemes, the other channel being independent of the context information and entirely driven by the sensory data to deliver estimates of 'out-of-context' posterior probabilities of phonemes. A significant mismatch between the information from these two channels indicates unexpected word. The viability of this concept is demonstrated on identification of out-of-vocabulary digits in continuous digit streams. The comparison of these two channels provides a confidence measure on the output of the recognizer. Unlike conventional confidence measures, this measure is not relying on phone and word segmentation (boundary detection), thus it is not affected by possibly imperfect segment boundary detection. In addition, being a relative measure, it is more discriminative than the conventional posterior based measures.

1 Introduction

One of the most serious problems of the current automatic recognition of speech (ASR) is its poor ability in dealing with unexpected sounds [1, 2]. Be it the word that is not in the dictionary of the recognizer or the word which prior probability of occurrence is low, such an item is likely to be replaced in the output of the recognizer by the high prior probability word that is in the dictionary of the recognizer and is emphasized by the language model. This undesirable property could have disastrous consequences on the utility of the recognizer in applications such as speech data mining or information summarization from the spoken input, since the low probability words could have very high information value. Unexpected words are not necessarily rare words in general. A word can be unexpected for an specific small vocabulary task, scenario or conversation situation but can be common in general. Unexpected word detection can be essential for small vocabulary tasks (specific applications), as well as large vocabulary.

One approach to address the unexpected word problem is using some form of a garbage model (i.e. the word model that allows for arbitrary sub-word sequences) that accommodates the unexpected word [1, 3]. The good match with the garbage model then can indicate the unexpected word. The use of the garbage model requires ad hoc setting of the garbage entry penalty which is a critically important parameter in this approach. For the high entry penalty, many unexpected words are misidentified as the words in the vocabulary, for the low penalty, many in-vocabulary words are treated as unexpected words, thus increasing the WER. In some applications, it could be preferable not to change the existing recognizer model configuration because the in-vocabulary recognition can possibly be degraded by introducing the unexpected word garbage model.

The other alternative approach is to identify potentially misrecognized words from the low confidence of the recognition result [4, 5, 6, 7]. One indicator of confidence is derived by detecting word and phone segments (usually by back-tracking alignment of the recognized utterance), and evaluating a normalized average likelihood or posterior measure inside the detected segments. Relying explicitly on the recognition and segmentation results of the recognizer is the main disadvantage of these measures. The effectiveness of these measures is sensitive to correct and precise detection of segment boundaries [6, 7].

To address this problem, this paper presents an alternative approach that does not require explicit recognition or segmentation (decisions about phone and word boundaries) in the utterance. Instead, two streams of frame-level probabilities of phonemes are compared based on the measure of similarity between their distribution. One stream of probabilities is derived solely from the acoustic evidence by trained Artificial Neural Net (ANN), called here the ‘out of context posteriors’ or ‘sensory channel’, the other is derived from acoustic evidence together with higher level prior knowledge (e.g. lexical and grammar knowledge as available for the existing recognizer) and long acoustic context, called here the ‘in context posteriors’ or ‘context channel’ [8, 9]. The comparison of these two frame level posteriors provide a frame level confidence measure on the match between the acoustic information and prior knowledge. A significant mismatch can indicate an unexpected word. This way, one can identify unexpected words in parallel with the existing conventional ASR process, to mark suspect part of the decoded sequence that can contain unexpected words. Unlike conventional measures, the new measure does not use explicit phone and word segment boundary detection, thus it is not affected by imperfect segmentation. Moreover, as compared to the conventional posterior based confidence measures [6, 7], it is a relative measure obtained by comparing two posteriors estimated with different prior knowledge, therefore it is expected to be more discriminative.

Section 2 introduces the new concept. Section 3 discusses more about the sensory and context channels, and the way the information in each channel is obtained. Section 4 deals with the way the two channels are compared to yield a measure of confidence and detect unexpected words. Section 5 presents the results and compares the new confidence measure with the conventional ones. Section 6 gives final discussions and conclusions.

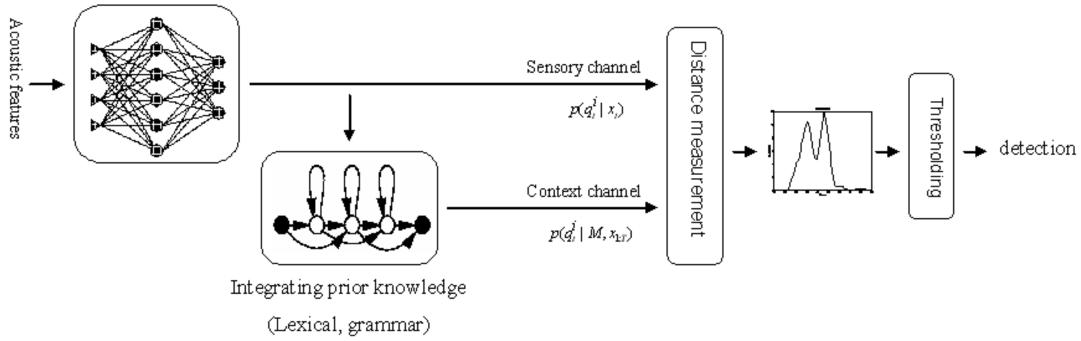

Figure 1: *The configuration for our confidence measurement and unexpected word detection method. 'Out of context posteriors' in the sensory channel are estimated by an MLP. 'In context posteriors' in the context channel are estimated using HMM/ANN layer integrating prior and contextual knowledge. The two channels are compared by measuring the distance (KL divergence) between posterior vectors at each frame. This distance is considered as a confidence measure. The distances are then compared with a threshold to decide on having unexpected word.*

2 Identifying unexpected words using frame-level phoneme posterior streams

In our approach, phoneme classification results on speech frame level (i.e. the classifications available in equally spaced intervals of about 10 ms) are utilized to identify the unexpected words. This is possible by employing the frame-level posterior probabilities of phonemes derived from two levels of the recognition process in hybrid HMM/ANN ASR [10], one from the feature level that provides pure input-based posteriors ('out of context posteriors'/'sensory channel'), and from the Baum-Welch process that provides phoneme posteriors derived with the use of the prior knowledge such as lexical and grammar knowledge as available for existing recognizer ('in context posteriors'/'context channel'). Similarly as in [8, 9], we estimate these 'in context phone posteriors' based on HMM state posterior probability definition, estimated using Baum Welch method, integrating prior and contextual knowledge. Comparing 'in context posteriors' obtained this way, and the 'out of context posteriors' in the sensory channel provides a measure of confidence on the output of the recognizer. The comparison is done based on measuring Kullback-Leibler (KL) divergence between the posterior probability distributions in the sensory and context channels. When encountering an unexpected word, the context channel significantly deviates from the sensory channel because the unexpected word is not supported by the prior knowledge. Fig. 1 shows a digram of our unexpected word detection method.

Conventional confidence estimation techniques [4, 5, 6, 7] 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. Unlike them, our method does not require explicitly recognition results or phone or word segment boundary detection, thus it is not affected by imperfect segmentation and boundary detection. Moreover, as compared to the conventional posterior based confidence measures [6, 7], it is a relative measure obtained by comparing two posteriors estimated with different prior knowledge, therefore it is expected to be more discriminative.

3 Sensory and context channels

3.1 Sensory channel

This channel provides phoneme posterior probabilities entirely driven by sensory data and independent of the long context or prior knowledge (Fig. 1). The phone posteriors are estimated only from a limited span of acoustic feature frames and without taking into account any contextual or prior knowledge about words and the way they form utterances. Among different approaches for estimating phone posteriors, ANNs and more specifically Multi Layer Perceptrons (MLPs) provide a discriminative way of estimating phoneme posteriors. In this work we use MRASTA phone posterior estimation method [11]. The MLP, trained on the training part of the database, estimates the posterior probabilities of phoneme classes at each frame $p(q_t^i|x_t)$. We call these posteriors 'out of context posteriors'.

3.2 Context channel

This channel provides phoneme posteriors that are derived not only from the acoustic input but also by integrating prior knowledge (e.g. lexical knowledge, grammar, etc.) and the long context of the whole utterance that is being recognized. Subsequently, the acoustic evidence that match the prior and contextual knowledge is emphasized and the evidence that does not support it is suppressed.

These 'in-context posteriors, as studied in [8, 9], are given based on HMM state posterior probabilities derived using Baum-Welch method. This posterior probability is defined as the probability of being in state i at time t, given the whole observation sequence $x_{1:T}$ and model M encoding specific prior knowledge (lexical and grammatical constraints):

$$\gamma(i,t) = p(q_t^i|x_{1:T}, M) \tag{1}$$

where, x_t is a feature vector at time t, $x_{1:T} = \{x_1, \ldots, x_T\}$ is an acoustic observation sequence, q_t is the HMM state at time t, q_t^i is the event "$q_t = i$". In the following, we often drop the M, keeping in mind that all recursions are processed through some prior (Markov) model M.

The state posteriors $\gamma(i,t)$ can be estimated by using HMM forward and backward recursions using local emission probability or likelihoods $p(x_t|q_t^i)$ or $p(q_t^i|x_t)$ (modeled by GMMs or ANNs):

$$\begin{aligned} \alpha(i,t) &= p(x_{1:t}, q_t^i) \\ &= p(x_t|q_t^i)\sum_j p(q_t^i|q_{t-1}^j)\alpha(j,t-1) \qquad (2) \\ \beta(i,t) &= p(x_{t+1:T}|q_t^i) \\ &= \sum_j p(x_{t+1}|q_{t+1}^j)p(q_{t+1}^j|q_t^i)\beta(j,t+1) \qquad (3) \\ \gamma(i,t) &= p(q_t^i|x_{1:T}, M) = \frac{\alpha(i,t)\beta(i,t)}{\sum_j \alpha(j,T)} \qquad (4) \end{aligned}$$

If we assume that a phoneme is represented by one state (q) in our HMM configuration, then $\gamma(i,t) = p(q_t^i|x_{1:T}, M)$ is the 'in context phone posterior' for phone i at time t. Otherwise if a phoneme is modeled with more than one HMM state, the 'in context phoneme posterior' can be simply estimated by adding up posteriors of all states composing the phone in the HMM (for more details refer to [8, 9]).

As shown in Fig. 1, we use 'out of context posteriors' (the sensory channel content estimated by MLP) as emission probabilities for the HMM/ANN layer which integrates prior and contextual knowledge. This layer can be considered as a filter which enhances the acoustic evidence matching the prior and contextual knowledge and suppresses the evidence which does not match it. As a consequence, when encountering an unexpected word, the evidence representing the unexpected word is significantly suppressed, because of no match with the prior knowledge. Therefore, the context channel deviates from the sensory channel, this deviation indicating the unexpected word.

4 Comparing sensory and context channels

In order to detect unexpected words, the difference between the two channels is measured. This difference then yields an estimate of a confidence in correctness of the recognizer output. In this work, we use Kullback-Leibler (KL) divergence (KL) to evaluate the difference between the two channels. KL divergence is suitable for measuring similarity of two probability distributions.

$$
\begin{aligned}
KL(\overline{S_t}, \overline{C_t}) &= \sum_i S_t^i log_2 \frac{S_t^i}{C_t^i} \\
&= \sum_i p(q_t^i|x_t) log_2 \frac{p(q_t^i|x_t)}{p(q_t^i|M, x_{1:T})}
\end{aligned} \tag{5}
$$

Where $\overline{S_t}$ is the posterior vector in the sensory channel at frame t, and $\overline{C_t}$ is the posterior vector in the context channel at frame t. S_t^i and C_t^i show the ith element of the posterior vectors at frame t.

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 increase in smoothed KL divergence above the pre-set threshold.

A sample of 'in context' and 'out of context' posteriors in the sensory and context channels, their difference and KL divergence over time is shown in Fig. 2. The utterance contains 'five three zero' where the word 'three' represents an unexpected word, not present in the vocabulary. Fig. 2.a shows the posteriors in the sensory channel, 2.b shows the posteriors in the context channel, and 2.c shows the difference between 2.a and 2.b. As it can be seen, there is a region with major difference corresponding to the word 'three' (which is marked roughly by dashed lines). Fig. 2.d shows the KL divergence between the two posteriors. Again peak corresponding to the word 'three' can be observed.

Figure 1 shows a diagram of the whole system: The phone posteriors in the sensory channel are estimated by an MLP. The phone posteriors in the context channel are estimated using an HMM integrating prior and contextual knowledge. This HMM layer uses the MLP posteriors in the sensory channel as the state emission probabilities. The content of the two channels ('in context' and 'out of context' posteriors) are compared based on measuring KL divergence at each frame. The divergence measure is considered as a frame level confidence measures for the correctness of the recognizer output. The divergence measures are then smoothed and compared with a threshold to decide if there is an unexpected word.

5 Experiments and results

In this section, we report the initial results in detecting unexpected words. We have used OGI digits database [12] for the experiments. There are 29 context-independent phones (monophones). We have introduced each of the words individually as an unexpected word by removing it from the vocabulary. The MLP based MRASTA method [11] was used to estimated phone posteriors for the sensory channel. There are 2169 utterances in the test set and 2547 utterances in the training set.

For the context channel, the phone posteriors in the sensory channel are used as emission probabilities for an HMM/ANN block. The role of this block is to integrate prior and contextual knowledge to estimate 'in context posteriors'. The topology of this HMM/ANN block contains all the words in the vocabulary except the one that was removed. The phone posterior vectors in the two channels are compared frame by frame by measuring the KL divergence. The divergence measures are then 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.

We have compared our posterior based confidence measure with a group of conventional posterior based confidence measures presented in the literature [6, 7]. These confidence measures (and many basically similar ones [4, 5]) are based on recognition and segmentation of the utterance into phonemes and words (by back-tracking alignment of the recognized utterance), and evaluating a posterior based measure inside the detected segments for the hypothesized word [6, 7]. The most typical ones, Normalized Posterior Based Confidence Measures (NPCMs), are defined as follows:

$$phone - basedNPCM(w)$$

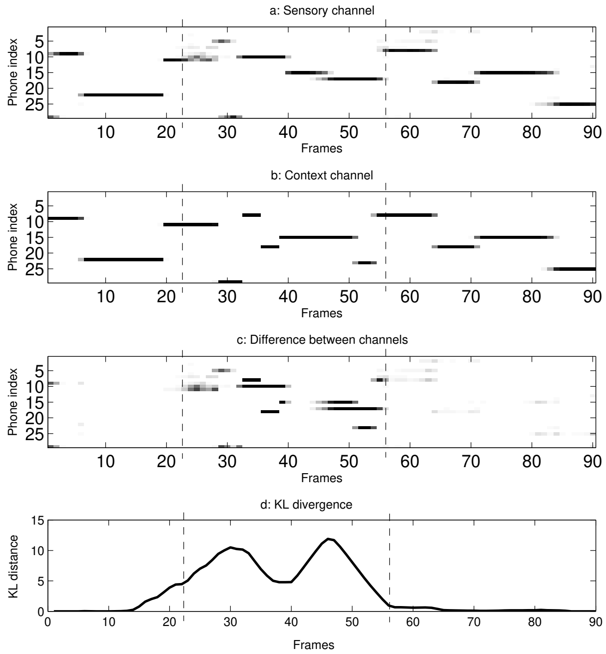

Figure 2: *'Out of context' and 'in context' posteriors and their difference and divergence for the utterance 'five three zero', where 'three' has considered as unexpected word.*

$$= \frac{1}{J}\sum_{j=1}^{J}\left(\frac{1}{e_j - b_j + 1}\sum_{n=b_j}^{e_j} logp(q_j^n|x_n)\right) \tag{6}$$

$$frame - basedNPCM(w)$$

$$= \frac{1}{\sum_{j=1}^{J}(e_j - b_j + 1)}\sum_{j=1}^{J}\sum_{n=b_j}^{e_j} logp(q_j^n|x_n)) \tag{7}$$

where J in number of phones in the hypothesized word, and e_j and b_j are the beginning and the end of each phoneme.

The performance of the individual systems is measured in terms of the trade off between true and false alarms for detecting unexpected words. Fig. 3 shows the Receiver Operating Characteristic (ROC) curves obtained by our method, and conventional posterior based methods. Our approach shows noticeably larger area under the ROC curve (much better trade off between true and false alarms).

6 Discussion and conclusion

A new confidence measure, 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. Unexpected word detection can be essential for small vocabulary tasks (specific applications), as well as large vocabulary.

The conventional confidence measurement methods usually explicitly segment the utterance into phonemes and words, then they evaluate a likelihood or posterior based measure for the expected words inside the detected segment boundaries. The accuracy of these measures are very sensitive to

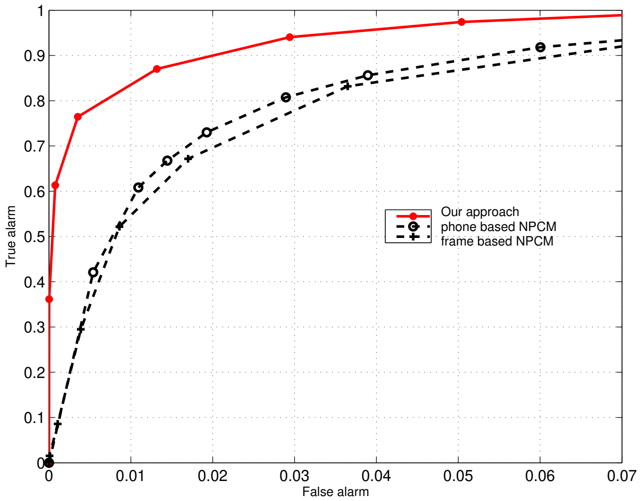

Figure 3: *ROC curves for our confidence measurement approach and conventional methods (phone-based and frame-based NPCM). The y axis is showing the percentage of true alarms and the x axis is showing the percentage of false alarms. Our approach shows significantly better trade off (larger area under the ROC curve).*

correct and precise detection of segment boundaries. In contrast, in our approach, there is no need for explicit segmentation and boundary detection. This is one of advantages which could lead to the observed better performance of our system. The other possible advantage is that our technique compares two phoneme posterior streams derived using different prior constrains but using identical acoustic evidence. This could alleviate inherent inconsistency of confidence estimates based on absolute posterior or likelihood measures.

Another interesting consequence of comparing the results from two parallel posterior streams is that the large divergence between the two streams could be also an indication of the correct decision in the context-constrained stream and the incorrect one in the sensory stream. Thus, one possibly fruitful extension of the current technique would be to investigate it as a general confidence measure technique.

7 Acknowledgment

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). Thanks to Petr Fousek for providing MRASTA phone posteriors.

References

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

[2] Hazen, T., and Bazzi, I., "A Comparison and Combination of Methods for OOV Word Detection and Word Confidence Scoring", ICASSP'01, Salte Lake City, Utah.

[3] Bazzi, I., and Glass, J., "Modeling out-of-vocabulary words for robust speech recognition", Proc. of ICSLP, Beijing, 2000.

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

[5] Kamppari, S., and Hazen, T., "Word and phone level acoustic confidence scoring", Proc. of ICASSP, Istanbul, 2000.

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

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

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

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

[10] Bourlard, H. and Morgan, N., "Connectionist Speech Recognition – A Hybrid Approach", Kluwer Academic Publishers, 1994.

[11] Hermansky, H., Fousek, F., "Multi-Resolution RASTA Filtering for TANDEM-Based ASR", Interspeech'05, Lisbon, Portugal, September 4-8, 2005.

[12] Cole, R., Noel, M., Lander T. and Durham T."New Telephone Speech Corpora at CSLU", In Proc. of EUROSPEECH (Madrid, Spain, 1995), pp. 821-824.