Detection and Identification of Rare Audiovisual Cues

Insperata accident magis saepe quam quae speres.
(Things you do not expect happen more often than
things you do expect) Plautus (ca 200(B.C.)

Project no: 027787

DIRAC

Detection and Identification of Rare Audio-visual Cues

Integrated Project
IST - Priority 2

DELIVERABLE NO: D 2.6
Detection of Unexpected Words in Machine Recognition of Speech

Date of deliverable: 30.06.2007
Actual submission date: 31.03.2007

Start date of project: 01.01.2006 Duration: 60 months

Organization name of lead contractor for this deliverable: **IDIAP Research Institute**

Revision [1]

Project co-funded by the European Commission within the Sixth Framework Program (2002-2006)		
Dissemination Level		
PU	Public	X
PP	Restricted to other program participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Detection and Identification of Rare Audiovisual Cues

Insperata accident magis saepe quam quae speres.
(Things you do not expect happen more often than
things you do expect) Plautus (ca 200(B.C.)

D2.6 Detection of Unexpected Words in Machine Recognition of Speech

IDIAP Research Institute (IDIAP)

Abstract:
In the first year of the DIRAC project, partners in WP5 worked on general framework for information fusion in detection of unexpected audio-visual events (deliverable D5.2). One particular strategy emerged as a promising one, and has been pursued in WP2 as a technique for identification of unexpected words in machine recognition of speech. This Deliverable D2.6 reports on early and very promising results obtained by using the proposed strategy. The results are reported on a rather limited recognition task of recognizing ten American English digits with the eleventh one introduced as an unexpected word unknown to the machine. However, the applied technique is in principle extendable to larger more realistic speech recognition tasks, as well as to identification of audio-visual events.

Table of Content

1. The Motivation

In speech communication, the unexpected items (words) carry more information than the expected ones [6]. 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 one very undesirable effect: Unexpected lexical items (words) in the phrase are typically replaced by acoustically acceptable in-vocabulary items [3]. This is the major source of error [7, 20]. Improving the machine ability to handle these unexpected words would considerably increase the utility of speech recognition technology.

2. Some Relevant Knowledge about Use of Context in Human Speech Recognition

2.1 The Context of the Message

No doubt that what we believe that we hear is heavily influenced by what we expect to hear. In other words, the **context** of the message in speech is an important element that contributes to the decoding of the message. It is the context that limits the number of possible alternative words that could most likely occur in a given part of the message.

A simple experiment carried out more than a half century ago by Miller, Heise and Lichten [16] illustrates the point. In this experiment, listeners were given a list of words that will be presented and their task was to recognize the spoken words. The number of possible words in the lists varied between 2 and 1000. The noise of varying levels was added to the spoken material to make the task harder. As expected and as discussed by Miller et al., it was easier to discriminate between 2 words than to discriminate among 1000 words.

As pointed out by Allen [22] there is a possible model that the collected data follow reasonably well. The model 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 decoding of the message in addition to the sensory (acoustic) channel*. Indeed, such a model has been proposed and tested earlier by Boothroyd and Nittrouer [17], where 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

$$p_w = 1-(1-p_N)(1-p_C),$$

where p_C indicates the contribution of the probability due to the context channel. Since in the presence of distortions such as noise, both p_C and p_N are degraded equally (the target words get degraded in the same way as the words that provide the context), the equation above simplifies to

$$p_w = 1-(1-p_N)^K, K>1.$$

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.

2.2 Unexpected Words

Further, Allen also discusses physiological data of van Petten et al [19] that are very relevant in our quest for discovery of unexpected words. The data come from the EEG experiments

dealing with negative swing N_{400} in EEG potential observed about 400 ms after encountering an unexpected word. The careful design of the experiment allows for discovery that the human hearing provides *instantaneous* indication when encountering the unexpected word, thus further supporting the parallel character of both the sensory and the context channels.

3. How ASR Works

Current machines for automatic speech recognition (ASR) work differently than human system does.

ASR uses context in the form of the lexicon and the so called language model. The use of the language model is quite essential. Without its use, almost an order-of-magnitude increase in word errors has been observed [21]. The acoustic (sensory) and the language model (prior knowledge) information sources are used serially since the recognized message w_{best} is chosen as the most probable message through a search over all possible messages using the Bayes rule

$$w_{best} = \arg\max_i \{p(x \mid M(w_i)P(M(w_i))\} \quad (1)$$

where $p(x|M(w_i)$ denotes the likelihood of the data x given the model of the i-th message $M(w_i)$ multiplied by the probability of the model $P(m(w_i))$. Since the probabilities (likelihood) of both information sources multiply, both need to be high for the final result to be high.

Also, the words (represented by sequence of sub-models forming the model $M(w_i)$) 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 is because the likelihood $p(x|M(w_i))$ of the whole sequence of words must be evaluated and the product with the prior probability of a given sequence $P(M(w_i))$ must be formed to yield the final decision about the found sequence of words w_{best}. This principle of the delayed decision is one of the most fundamental and most powerful principles of the HMM-based ASR.

4. An Engineering System for Discovery of Unexpected Words

An unexpected input to the conventional HMM-based ASR for which $P(M(w_i))$ goes to zero, (e.g. the word that is not in the lexicon of the recognizer) is a significant source of error since, as long as the Eq. (1) is applied, the particular message containing this word can never be chosen. The recognizer in this case must substitute another, acoustically similar, word (or a sequence of shorter words) from its lexicon which has in the particular context a reasonable prior probability of occurrence.

When a human listener encounters 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 to be able to identify the out-of-vocabulary word in ASR too and efforts in this direction are ongoing (see e.g. [20] for a review of recent efforts).

We approach this problem in a principled way, following the general strategy illustrated in Fig. 1. This strategy assumes the existence of two parallel processing streams, one - the more powerful, employing the prior knowledge in addition to the information from the sensory data, the other one evaluating only the sensory data. When both streams yield the results in

the same form, the results can be compared. As long as the sensory data are consistent with the prior knowledge, results from both streams are similar. However, when the sensory data deviate from the prior knowledge (e.g. in the case when the prior knowledge disallows an existence of a certain word), the results in the streams may be different. This is indicated in the comparison module and a proper action might be taken. In the case of the unexpected word, an attempt may be made to describe the word in terms of its acoustic elements (phonemes). Such a phonetic description could be then used for updating the lexicon of the available models so when this word is encountered the next time, the upper (prior knowledge employing) stream may be able to recognize it.

In the current work described in this report we are addressing the blocks with the bold lettering. Clearly, this scheme assumes the existence of the processing stream that could reliably describe the information in the input signal. As mentioned above, the prior knowledge (language model) constraints and the principle of the delayed decisions, both applied in the conventional (prior knowledge employing) system, are responsible for its reasonable performance. So it is a challenge to obtain useable recognition results without the use of the prior knowledge. Subsequently, the majority of the work is devoted to obtaining reasonable evaluation of the acoustic input without the use of the prior knowledge. The adaptation, the description and the update of the lexicon schemes will be addressed in the future.

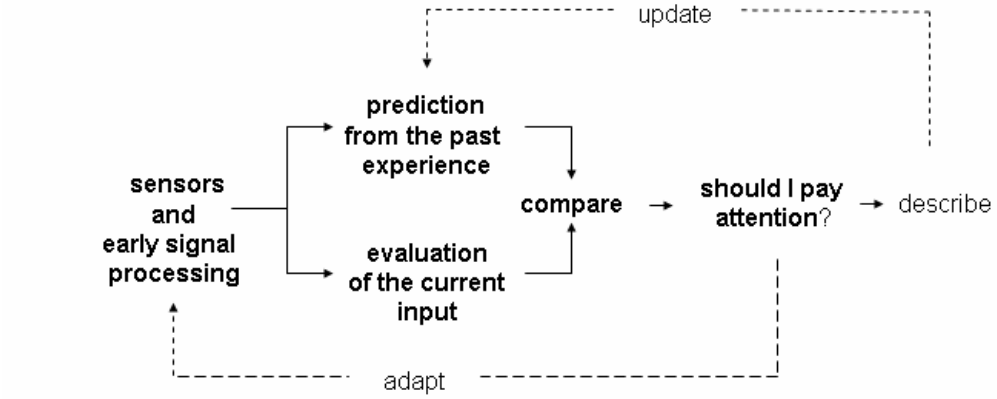

Fig. 1 General scheme of discovery and dealing with unexpected low prior probability stimuli

Principles of our approach [5], shown schematically in Fig. 2, are briefly described below.
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 frame-level posterior probabilities of phonemes are derived from two levels of processing in a hybrid Hidden Markov Model (HMM) recognizer utilizing an artificial neural network (ANN) probability estimation (HMM/ANN ASR) [2]. In this technique, the context-unconstrained phoneme probabilities are estimated by the trained ANN. These are subsequently being used in the search for the most likely stochastic model of the input utterance. Thus, one set of posterior probabilities is obtained directly from the ANN (context-unconstrained posteriors), another set comes from the Baum-Welch estimation procedure that, as outlined e.g. in [8,9], provides 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 (context-constrained posteriors).

Comparing ‘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 posterior probability distributions in the sensory and context channels. When encountering an unexpected word, the context-constrained posteriors significantly deviate from the context-unconstrained posteriors because the unexpected word is not supported by the prior knowledge.

Conventional confidence estimation techniques [10,11,12,13] 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. Our technique does not require any explicit segmentation and subsequently it is not affected by the problems that may be encountered while doing so.

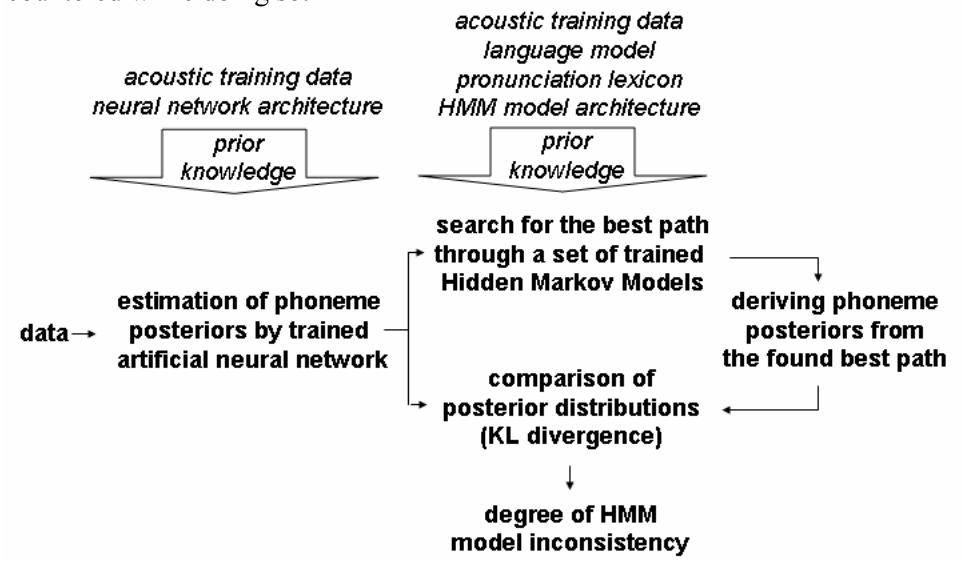

Fig. 2 Discovery of out-of-vocabulary words using hybrid HMM-NN ASR system, in which the out-of-context posterior probabilities estimated by the ANN are also directly used in the constrained search for the best model sequence.

4.1 Context-unconstrained Posteriors

The basic principles of deriving context-unconstrained posterior probabilities of phonemes are illustrated in Figs. 3 and 4. Feed-forward ANN is trained on phoneme labeled speech data. It uses as an input a segment xi of the data X that carries the local information about the identity of the underlying phoneme at the instant i. This segment is projected on 448 time-spectral basis. The output from the ANN represents a vector of context-unconstrained posterior probabilities of phonemes p(qti | xt). As seen in the middle part of Fig. 5, estimate from the ANN can be different that the estimate from the context-constrained stream since it is not dependent of the constraints L.

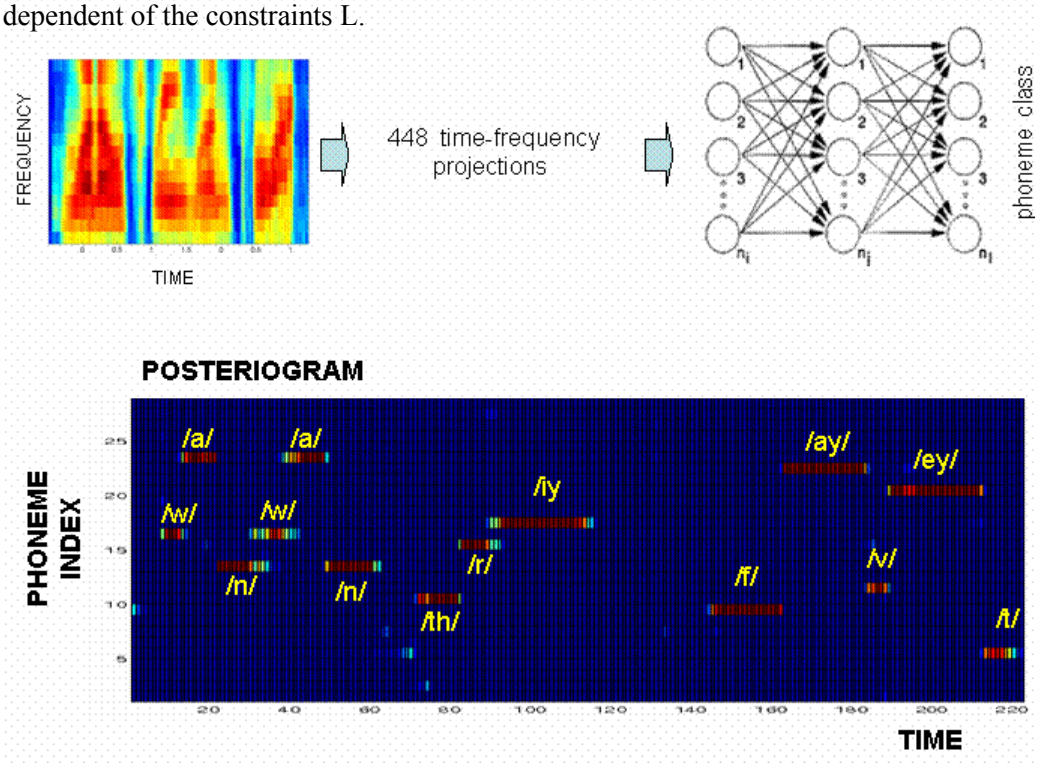

Fig. 3 Illustration of the technique for obtaining reliable estimate of posterior probability density functions pi(Q|X) without the use of top-down constraints L. Short-term critical band spectrogram, illustrated in the left part of the figure, is derived by weighted summation of appropriate components of the short-term spectrum of speech. A segment of this spectrogram is projected on 448 different time-frequency bases (shown in Fig. 3), centred at the time instant i, yielding 448 point vector that forms the input to the MLP neural net, trained on about 2 hours of hand-labeled telephone-quality speech to estimate a vector of posterior probabilities pi(Q|X). A set of pi(Q|X) for all time instants forms the so-called posteriogram, shown for the utterance "one-one-three-five-eight" in the lower part of the figure, higher posterior probabilities being indicated by warmer colors (see [4] for more details).

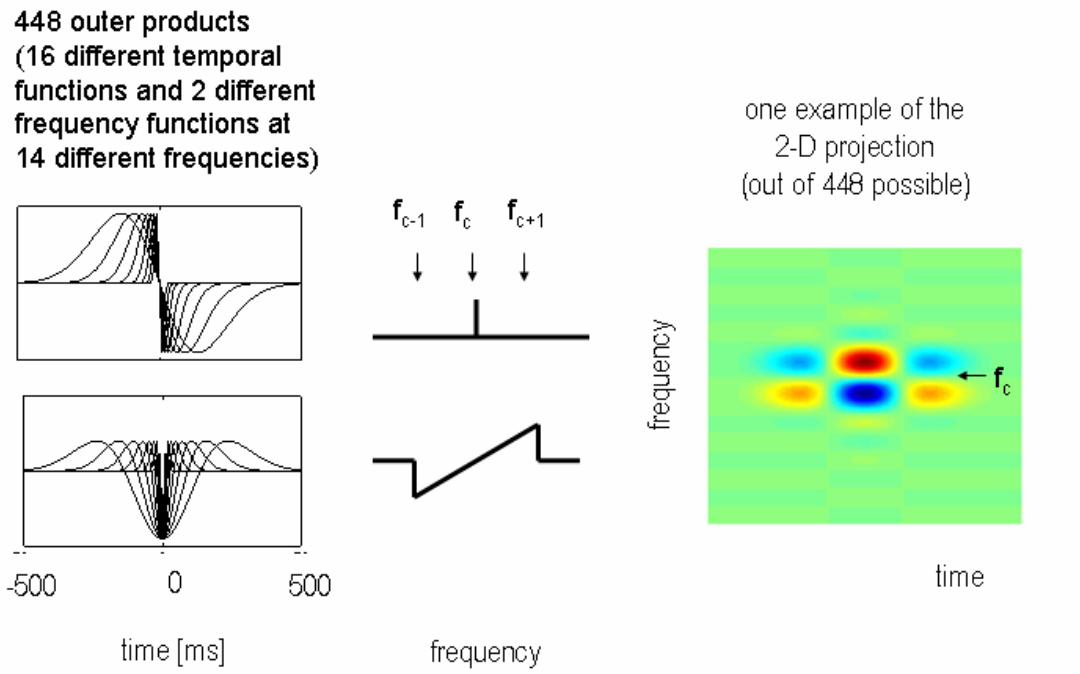

Fig. 4 The time-frequency bases that attempt to emulate some very basic properties of auditory cortical receptive fields. They are formed as outer products of first and second derivatives of truncated Gaussian functions of 8 different widths in the time domain, and by summation and differentiation over three frequency components (3 critical bands), centred at 14 different frequencies in the frequency domain (see [4] for more details).

4.2 Context-Constrained Posteriors

The context-unconstrained phoneme probabilities are used in a search for the most likely Hidden Markov Model (HMM) sequence that could have produced the given speech phrase. As a side product, the HMM can also yield, for any given instant *i* of the message, its estimates of posterior probabilities of the hypothesized phonemes

$$\gamma_i(i,t) = p(q_t^i \mid x_{1:T}, M)$$

given the whole observation sequence $x_{1:T}$, and constrained by a set of constraints *M* implied by the training speech data, model architecture, pronunciation lexicon, and the applied language model. Here x_t denotes a feature vector at time *t*, $x_{1:T} = \{x_1, \ldots, x_T\}$, is an acoustic observation sequence, q_i indicates the *i*-th HMM state, and q_t^i indicates that the model is in the state q_i at the time *t*. In the following, we drop the indication *M,* however, bearing in mind that all the operations are carried under the constrains implied by the model *M*.

The in-context state posteriors $\gamma(i,t)$ can be estimated by using HMM forwards and backward recursions using the local emission likelihood $p(x_t|q_t^i)$ derived from GMM model in the case

of the conventional HMM or probability $p(q^i_t| x_t)$ derived from ANN in our case of the hybrid ANN/HMM system.

$$\alpha(i,t) = p(x_{1:t}, q_t^i) = p(x_t \mid q_t^i) \sum_j p(q_t^i \mid q_{t-1}^j)\alpha(j,t-1)$$

$$\beta(i,t) = p(x_{t+1:T}, q_t^i) = \sum_j p(x_{t+1} \mid q_{t+1}^j) p(q_{t+1}^j \mid q_t^i)\beta(j,t+1)$$

$$\gamma(i,tr) = p(q_t^i \mid x_{1:T} M) = \frac{\alpha(i,t)\beta(i,t)}{\sum_j \alpha(j,T)}$$

Assuming that a phoneme is represented by one state *q* in the HMM architecture, then $\gamma(i,t) = p(q^i_t|x_{1:T},M)$ is the in-context phoneme posterior for the phone *i* at the time *t*. Otherwise, when a phoneme is represented by more than one state of the HMM model, the in-context posterior can be obtained by adding posteriors from all states that form the phoneme.

The acoustic evidence that obeys these constraints is emphasized and the evidence that does not support it is suppressed. Thus, the search, when e.g. encountering an unknown item in the phoneme string (e.g. the word `three' in Fig. 5), it assumes it is one of the well known items. Note that these 'in context' posterior probabilities, even when wrong, are estimated with high confidence.

4.3 Comparing In-Context and Out-of-Context Posteriors

To detect unexpected words, the difference between the two channels is evaluated. The 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.

$$KL(S_t, C_t) = \sum_i S_t^i \log_2 \frac{S_t^i}{C_t^i} = \sum_i p(q_t^i \mid x_t) \log_2 \frac{p(q_t^i \mid x_t)}{p(q_t^i \mid M, x_{1:T})}$$

Here S_t indicates the out-of-context posterior vector at the time *t*, and C_t the in-context posterior vector at the time *t*, and S^i_t and C^i_t are *i*-th elements of the respective vectors of posteriors.

Here we somehow arbitrarily have chosen $KL(S_t,C_t)$. The $KL(C_t, S_t)$ or the symmetrized version of the KL divergence could have been probably chosen with a similar effect.

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.

An example of in context and out of context posteriors and the smoothed divergence as a function of time is shown in Fig. 5. The utterance contains 'five three zero' where the word 'three' represents an unexpected word, not present in the vocabulary. The upper part shows the out-of-context posteriors, the middle part the in-context posteriors, and the lower part shows the smoothed KL divergence between 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).

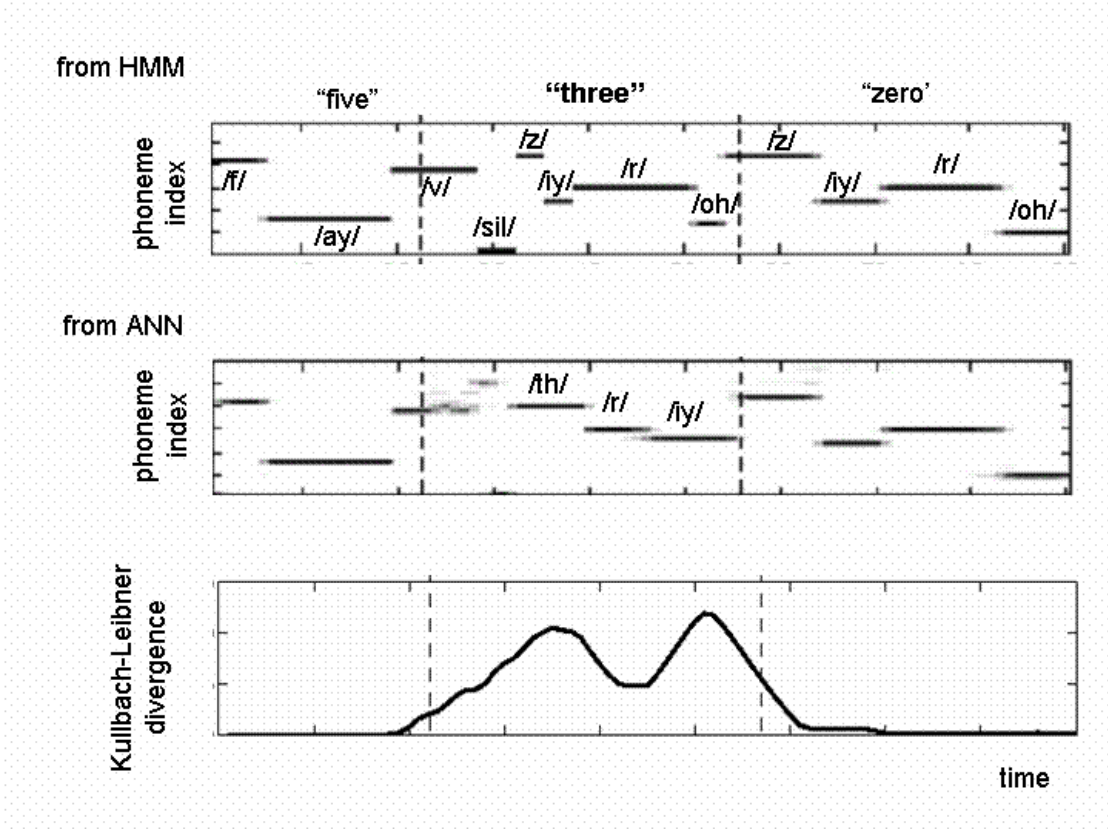

Fig. 5 Posterior probabilities (posteriograms) of phonemes estimated by an HMM-based system (the upper part of the Figure), and by ANN (the middle part of the Figure). In this example, the HMM model inconsistency was introduced by removing the word `three' from the recognizer vocabulary. The correct phoneme sequence for the word `three' is misrepresented in the HMM-derived posteriogram (replaced by a sequence /z/iy//r//oh/ of the in-vocabulary word `zero'). The ANN derived probabilities indicate in this case the correct sequence /th//r//iy/ for the out-of-vocabulary word `three'. Comparison of the respective posterior probability density functions by evaluating their relative entropy (KL divergence), shown smoothed by 100 ms square time window as a function of time in the lower part of the figure, indicates HMM model inconsistency in the neighbourhood of the out-of-vocabulary word `there' (The figure is adopted from [5] that also gives more details).

An example of a typical result is shown in Fig. 5. As seen in the lower part of the Fig. 5, an inconsistency between these two information streams could indicate unexpected out-of-vocabulary word.

5. Experiments and Results

In this section, we report the initial results in detecting unexpected words. We have used OGI digits database [15] for the experiments. 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. The MLP based MRASTA method [4] 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.

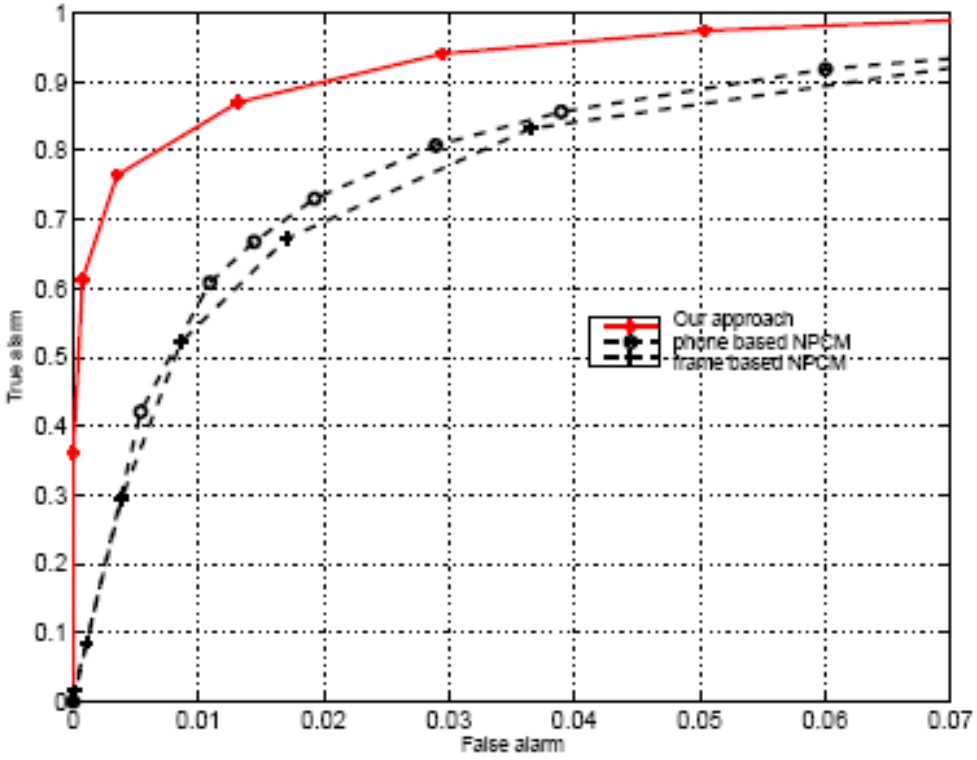

Fig. 6 Receiver operating curves (ROC curves) for our confidence measurement approach and conventional methods (phone-based and frame-based NPCM). The y axis is showing the percentage of true alarms and the x axis is showing the percentage of false alarms. Our approach shows significantly better trade off (larger area under the ROC curve).

We have compared our posterior based confidence measure with a group of conventional posterior based confidence measures presented in the literature [12, 13]. These confidence measures (and many basically similar ones [10, 11]) 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 [12, 13]. The most typical ones, word-based and frame-based Normalized Posterior Based Confidence Measures), are defined as follows:

Word based

$$NPCM(w) = \frac{1}{J}\sum_{j=1}^{J}\left(\frac{1}{(e_j - b_j + 1)}\sum_{n=b_j}^{e_j}\log p(q_j^n \mid x_n)\right)$$

Frame-based

$$NPCM(w) = \frac{1}{\sum_{j=1}^{J}(e_j - b_j + 1)}\sum_{j=1}^{J}\sum_{n=b_j}^{e_j}\log p(q_j^n \mid x_n)$$

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. 6 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 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 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

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 likelihood or posterior based measure for the expected words inside the detected segment boundaries. The accuracy of these measures are very sensitive to 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.

6.1 Several Additional Thoughts

Being able to identify which words are not in the lexicon of the recognizer, and being able to provide an estimate of their pronunciation, may allow for inclusion of these new words in the pronunciation dictionary, thus leading 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 have to indicate only the presence of unexpected lexical item but could also indicate any other inadequacy of the model.

Further, it may also indicate corrupted input data when the in-context probability estimation using the prior L 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.

References

[1] H. Bourlard, C.J. Wellekens, *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.

[2] Bourlard, H. and Morgan, N., Connectionist Speech Recognition - A Hybrid Approach, Kluwer Academic Publishers, 1994

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

[4] H. Hermansky and P. Fousek, *Multi-resolution RASTA filtering for TANDEM-based ASR*, , in Proceedings of Interspeech 2005, 2005.

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

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

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

[8] 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.

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

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

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

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

[13`] 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, November 2004, also IDIAP-RR 04-58.

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

[15] 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.

[16]Miller, G. A., Heise, G. A., and Lichten, W. (1951). The intelligibility of speech as a function of the context of the test material, J. Exp. Psychol. 41:329–335.

[17]Boothroyd, A. and Nittrouer, S. (1988). Mathematical treatment of context effects in phoneme and word recognition, J. Acoust. Soc. Am. 84(1):101–114.

[18] Fletcher, H. (1995). Speech and hearing in communication, in Allen, J. B., editor, The ASA edition of Speech and Hearing in Communication. Acoustical Society of America, New York.

[19] Van Petten, Cyma; Coulson, Seana; Rubin, Susan; Plante, Elena; Parks, Marjorie, Time course of word identification and semantic integration in spoken language, Journal of Experimental Psychology: Learning, Memory, and Cognition. 25(2), Mar 1999, 394-417.

[20] Bazzi Issam (2002) Modelling Out-of-Vocabulary Words for Robust Speech Recognition, MIT PhD. Thesis, Department of Electrical Engineering and Computer Science, 2002

[21] Lee, Kai-Fu, Hsiao-Wuen Hon, Mei-Yuh Hwang, Recent Progress in the SPHINX Speech Recognition System, Proceedings DARPA Speech and Natural Language Workshop, 1989
[22] J.B. Allen, Articulation and Intelligibility, Morgan & Claypool 2005