USING COMPARISON OF PARALLEL PHONEME PROBABILITY STREAMS FOR OOV WORD DETECTION

Tamara Tošić[1,2], *Mathew Magimai.-Doss*[1], *Hynek Hermansky*[1,2]

[1] IDIAP Research Institute
Centre du Parc, Av. des Prés-Beudin 20, 1920 Martigny, Switzerland
[2]École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland
phone: + (41) 27 721 77 11, fax: + (41) 27 721 77 12, email:{ttosic, mathew, hynek}@idiap.ch
web: www.idiap.ch

ABSTRACT

In this paper, we investigate the approach of comparing two different parallel streams of phoneme posterior probability estimates for OOV word detection. The first phoneme posterior probability stream is estimated using only the knowledge of short-term acoustic observation. In our work we refer this stream as "out-of-context posteriors". The second posterior probability stream, referred also as "in-context posteriors" is estimated using the knowledge of the whole acoustic observation sequence: the acoustic model and the language model of an ASR system. In particular, we focus our study on different types of distance measures, namely KL-divergence and Euclidean distance, to compare the two phoneme posterior probability streams. Our experiments on large vocabulary automatic speech recognition task shows that using KL-divergence measure estimated with the in-context posteriors as reference distribution consistently yields a better OOV word detection system.

1. INTRODUCTION

State-of-the-art statistical speech recognition systems combine the acoustic likelihood estimates for the input acoustic observation with prior knowledge (dependency between words) from the language model to find the most likely word sequence. One of the main limitations of Automatic Speech Recognition (ASR) systems is that it can only recognize those words which are present in its lexicon (vocabulary). Therefore, any occurrence of *out-of-vocabulary* (OOV) words, i.e., the words that are not included in the lexicon, results in a recognition error. This error can spread to neighboring words as a result of constraints introduced by a language model. Since language is a continuously evolving system, new words tend to be created or borrowed from other languages, while existing words tend to be associated with different meanings. Therefore, the occurrence of OOV words cannot be avoided.

In the literature, different approaches have been proposed to detect OOV words. The most simple and common approach which bears resemblance to a keyword spotting is to estimate a confidence value for each word in the ASR output hypothesis and decide if they are OOV words or not [1]. Another common approach is to use garbage models [2]. This method is also widely used for keyword spotting [4]. The garbage model is a generic model (for instance, phoneme models connected in ergodic fashion) or explicit filler model trained off-line on a database of OOV word utterances. The garbage model tries to model/match any word that can not be modelled/matched by the words in the lexicon. The main drawback of this technique is the fact that OOV word database has to be defined in advance. More recently, a lattice comparison approach has been proposed [3]. Lattices are formed during the recognition process and they depict possible recognition results. By observing the areas of the word mismatch amongst different lattice paths, it is possible to detect the regions of potential OOV word occurrences. Boundaries and words which differ in lattice are further analyzed using word confidence measures. However, the effectiveness of this method heavily relies on the recognizer accuracy.

In a recent work, an approach for detecting OOV words by comparing two parallel phoneme posterior probability streams was proposed [5]. In this approach, OOV words are detected by finding the discrepancies between two estimates of phoneme posterior probabilities. First estimate is obtained from the short-term acoustic feature knowledge and it is referred to as *out-of-context posteriors*, while the second estimate is obtained using the knowledge of the whole acoustic observation sequence, as well as the ASR models (i.e. acoustic model and language model) and it is referred to as *in-context posteriors*. OOV words are detected as the regions where in-context and out-of-context estimates of phoneme posterior probabilities diverge significantly. It is interesting to note that this approach is in spirit close to the lattice comparison approach which was described earlier. This approach was studied on a small vocabulary digit recognition task using a *Kullback-Leibler* (KL) divergence [6] as the distance measure for comparing two phoneme posterior probability streams, and it was shown that this approach performs better than the usual approach of detecting OOV words using word level confidence measure.

More recently, this approach was investigated in combination with different approaches/confidence measures for large vocabulary ASR task at John Hopkins Workshop 2007, where a neural network was trained to integrate the input (different confidence measures, divergence from parallel probability streams etc.) and classify a word as OOV, non-OOV or silence [7]. The integrated system was shown to perform best among all the systems that were studied.

In this paper, we expand the work done in [5] by extending the study to large vocabulary task with a particular focus on a distance/divergence measures that can be used to compare out-of-context posteriors with in-context posteriors. Our experimental studies on Wall Street Journal (WSJ) 5k task shows that KL-divergence measure with in-context posteriors as reference distribution yields consistently better OOV

word detection system.

The rest of the paper is organized as follows: Section 2 motivates the approach of comparing parallel phoneme posterior streams in the spirits of top-down and bottom-up processing in human speech recognition and briefly describe the OOV detection system investigated in this paper. Section 3 describes in detail different components of the OOV detection system. Section 4 presents the experimental setup and results. Finally, Section 5 concludes the paper.

2. MOTIVATION AND DESCRIPTION

The approach of comparing parallel stream of phoneme posterior estimates can be seen as comparison of *top-down* prior context information (in our case in-context posteriors) and *bottom-up* acoustic information (in our case out-of-context posteriors). There is physiological evidence [9] and psychophysical evidence [12, 10, 11, 8] that human speech recognition integrates the top-down prior context information and bottom-up acoustic information.

The work of van Petten et al [9] deals with the issue of timing of the triggering of the negative peak N400 event in human EEG signal by incongruous words that differ from the expected congruous word (e.g. "dollars" in the sentence "Pay with ...") either at their beginning (e.g. "scholars") or at their ends (e.g. "dolphins"). They demonstrate that the rhyming words ("scholars") trigger the N400 *earlier* than the incongruous words with the correct first syllable ("dolphins"). This supports the notion of integration of top-down context information and bottom-up acoustic information.

In psychophysical study by Boothroyd et al [10], it has been shown that the error of human speech recognition in the context of other meaningful and related words is product of error of hypothetical "context" channel that contributes to the correct word recognition and error of the acoustic channel. Furthermore, as discussed by Allen [8], the data of Miller et al [12] and Boothroyd et al [10, 11] support the existence of two statistically independent channels: bottom-up acoustic channel and top-down context channel. When an expected word is uttered and the acoustics is reliable, both channels indicate the same word. In a case when the uttered word is unexpected (e.g. OOV) or the acoustics is unreliable, the channels may disagree and indicate different word. Still, the words will be recognized as long as the evidence in either the acoustic or the context channel is strong enough.

Finally, in statistical sense, comparison of two phoneme posterior streams for OOV word detection can be interpreted in terms of Bayesian surprise [15], where the prior and posterior distribution of events differ in the presence of unexpected events.

Figure 1 depicts the general scheme of OOV word detection approach using two parallel stream of phoneme posterior estimates, where the out-of-context posteriors are estimated by using a phoneme classifier and the in-context posteriors are estimated by using an ASR system.

Procedure of OOV word detection, given the acoustic feature observation sequence $X = \{x_1, x_2, \cdots, x_t, \cdots, x_T\}$, parameters of phoneme classifier Θ_{pc} and parameters of ASR system Θ_{asr}, can be decomposed in three main steps:

- Estimation of out-of-context posteriors $P(c_t = k|x_t, \Theta_{pc})$ and in-context posteriors $P(c_t = k|X, \Theta_{asr})$, where $c_t \in \{1, \cdots, k, \cdots, K\}$ and K is the number of phonemes.

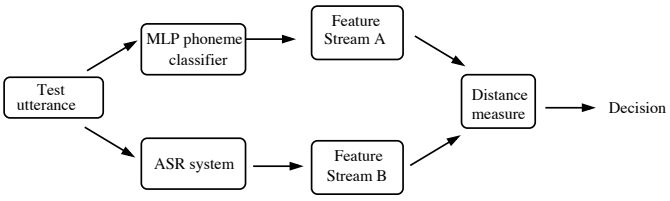

Figure 1: Schematic of the stream comparison approach for OOV word detection.

- Using a suitable measure to estimate the distance/divergence d_t between two phoneme posterior estimates $P(c_t = k|x_t, \Theta_{pc})$ and $P(c_t = k|X, \Theta_{asr})$
- Estimation of confidence score S_{w_i} for each word w_i in the hypothesis using d_t, followed by a classification of a word to OOV or non-OOV group based on this score.

These three steps are described in the following sections.

3. COMPONENTS OF OOV WORD DETECTION SYSTEM

3.1 Estimation of out-of-context and in-context phoneme posteriors

We estimate the two phoneme posterior probabilities in the following manner:

- Estimation of out-of-context posteriors $P(c_t = k|x_t, \Theta_{pc})$: Similarly to earlier work, this values are estimated using a multilayer perceptron (MLP) trained to classify phonemes given acoustic observation x_t[1] as input.
- Estimation of in-context posteriors $P(c_t = k|X, \Theta_{asr})$: Estimation of this probabilities is not straightforward. The common approach is to use the word lattice generated using the ASR system. In [5], the words in the word lattice were connected in ergodic fashion and $P(c_t = k|X, \Theta_{asr})$ was estimated by Baum-Welch forward-backward algorithm [13] with $P(c_t = k|x_t, \Theta_{pc})$ as emission probability [17]. In [7], the word lattice was expanded into phoneme lattices, and then $P(c_t = k|X, \Theta_{asr})$ was estimated by Baum-Welch forward-backward algorithm with acoustic model scores and language model scores in the lattice.
  In this study we take a slightly different approach. Given a set of N-best hypothesis $\{W_1, \cdots W_i, \cdots W_N\}$ obtained from the word lattice, $P(c_t = k|X, \Theta_{asr})$ is approximated as:

$$P(c_t = k|X, \Theta_{asr}) = \sum_{i=1}^{N} P(c_t = k, W_i|X, \Theta_{asr}) \quad (1)$$

$$= \sum_{i=1}^{N} P(c_t = k|W_i, X, \Theta_{asr}) \cdot P(W_i|X, \Theta_{asr}) \quad (2)$$

  The first term $P(c_t = k|W_i, X, \Theta_{asr})$ for phoneme k is either 0 or 1 given the alignment of hypothesis W_i with respect to acoustic observation sequence X. The second term $P(W_i|X, \Theta_{asr})$ is approximated by normalizing the

[1] Note that in practice the MLP takes acoustic feature at current time t with preceding and following time context. Typically, 4 time frames that precede and follow the current frame are taken into account.

joint likelihoods $P(W_i, X|\Theta_{asr})$ of each hypothesis W_i by likelihood of acoustic observation sequence $p(X|\Theta_{asr})$ as follows:

$$P(W_i|X,\Theta_{asr}) = \frac{p(W_i,X|\Theta_{asr})}{p(X|\Theta_{asr})} \quad (3)$$

$$= \frac{p(W_i,X|\Theta_{asr})}{\sum_{j=1}^{N} p(W_j,X|\Theta_{asr})} \quad (4)$$

3.2 Distance/Divergence measures

We investigate two different distance measures to compare out-of-context $P(c_t = k|x_t,\Theta_{pc})$ and in-context posteriors $P(c_t = k|X,\Theta_{asr})$:

- Euclidean distance: Previous work in the area of template matching, where posterior probability estimates were also used as a speech features, investigates this distance measure for capturing the similarity amongst two templates [14]. Indeed, squared Euclidean distance is one of the common distance measures used in speech recognition, since it is related to a Gaussian distribution, which is widely used for signal modeling. It is easy to show that it can be obtained from the logarithm of a Gaussian distribution with unity covariance matrix, where a constant represents a residue of this operation. Furthermore, Euclidean measure is a symmetric measure which in general represents the shortest distance between the two points. Here, it depicts the similarity of two posterior probability streams. If bottom-up and top-down posterior estimation vectors in a particular time frame are similar, Euclidean distance measure will be small (indicates the correctly recognized phoneme), and reverse. At each time frame, the Euclidean distance value is given with:

$$d_t = \sum_{k=1}^{K} [P(c_t = k|x_t,\theta_{pc}) - P(c_t = k|X,\theta_{asr})]^2 \quad (5)$$

- KL-divergence: This measure has its origins in the information theory field. It represents the average number of extra bits used to code an information source using the output probability p with a code q. In [15], it has been also shown that this measure can be used to measure the surprise quantitatively.
  Given two discrete probability distributions, reference distribution p and test distribution q, the KL-divergence is defined as:

$$d_t = KL(p||q) = \sum_{k=1}^{K} p_k \log \frac{p_k}{q_k}. \quad (6)$$

  It is known that KL-divergence is not a symmetric measure. Thus, the choice of reference distribution will make a difference. In our study we consider two conditions, KL_A where $P(c_t = k|x_t,\Theta_{pc})$ is the reference distribution p, and KL_B where $P(c_t = k|X,\Theta_{asr})$ is the reference distribution p. In the previous studies, the KL_A measure was used for comparing the two streams of phoneme posterior probabilities. The distance/divergence d_t at each time frame t is given with KL_A or KL_B depending upon the choice of reference distribution.

3.3 Confidence Measures

In [19], a double normalization approach for word confidence measure was proposed. The double normalization automatically normalizes the effect of sub-word unit duration. Given a word w_i, its alignment with begin time frame t_b^s and end time frame t_e^s for each phoneme $s \in \{1,2,\cdots ph\}$ in the word, and the phoneme probability $P(c_t = s|x_t,\theta)$, the confidence S_{w_i} is estimated as:

$$S_{w_i} = \frac{1}{ph} \sum_{s=1}^{ph} \frac{1}{t_e^s - t_b^s + 1} \sum_{t=t_b^s}^{t_e^s} -\log(P(c_t = s|x_t,\theta)) \quad (7)$$

where ph is the number of phonemes in the word w_i. In [19], the parameter θ was corresponding to the parameters of phoneme MLP of the hybrid system. In this work, the baseline system consists of estimating S_{w_i} with the phoneme MLP of out-of-context posteriors and deciding the word w_i in the ASR output hypothesis as OOV or non-OOV by comparing it to a threshold.

While using the OOV word detection approach described in this paper, the word confidence S_{w_i} is estimated using (7) except that $-\log(P(c_t = s|x_t,\theta))$ is replaced by distance/divergence score d_t. Then each word is classified as OOV or non-OOV by comparing the word confidence to a threshold. Threshold values were evaluated based on the current data word scores for all the test data: minimum and maximum threshold values represent minimum and maximum word scores, while other thresholds take values amongst them, with the equal pace.

It is important to note that the distance/divergence measures studied in this paper will yield high value for not only the occurrence of OOV word but also for misrecognized and mispronounced words. Similar to previous work [5, 7], we make no distinction between OOV word and misrecognized word in a process of classification each of the word in the hypothesis as OOV or non-OOV.

4. EXPERIMENTAL SETUP AND RESULTS

We have evaluated the OOV detection on Wall Street Journal (WSJ) 5k task. The size of the training data is 80 hours. The test set consisted of 913 utterances, out of which 376 of them contain at least one OOV word.

The acoustic observation x_t is given with a speaker-level mean and variance (normalized) and it is represented with 13 dimensional PLP cepstral coefficients [16], its first order derivative and second order derivative, resulting in a 39 dimensional vector per time frame t.

The MLP trained to estimate $P(c_t = k|x_t,\Theta_{pc})$ had 9 frames of acoustic observation vector as input and 45 phoneme classes as output. Θ_{pc} here refers to the parameters of the trained MLP and the number of parameters were 5% of the total number of training feature vectors. The MLP was trained with one-hot-encoding and yielded 80.9% frame accuracy on cross-validation data.

We used an on-the-shelf 32 Gaussian mixtures context-dependent phoneme based acoustic models trained on the WSJ data to estimate $P(c_t = k|X,\Theta_{asr})$. We used HTK system to generate word lattices using bigram language model with back-off, and for aligning the N-best hypotheses were utilized.

In our experiments, we varied the number of hypothesis obtained from word lattice to 1, 2, 5, 10 while estimating

$P(c_t = k|X, \Theta_{asr})$ to see the effect of it on the OOV word detection performance with different distance/divergence measures.

The evaluation of the system performance is carried out by calculation of the receiver operating characteristic (ROC) curve, as well as the area under the curve - figure-of-merit (FOM).

ROC curve corresponding to the different distance/divergence measures and the number of hypotheses 1 and 5 used to estimate $P(c_t = k|X, \Theta_{asr})$ are shown in Fig. 2. and Fig. 3, respectively.

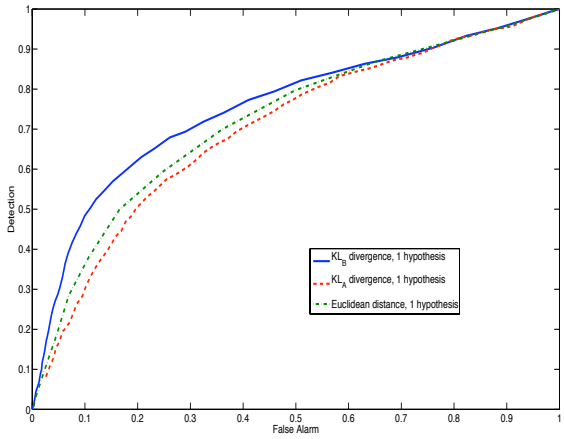

Figure 2: ROC curve corresponding to a different distance/divergence measures when the number of hypotheses used to estimate $P(c_t = k|X, \Theta_{asr})$ is 1.

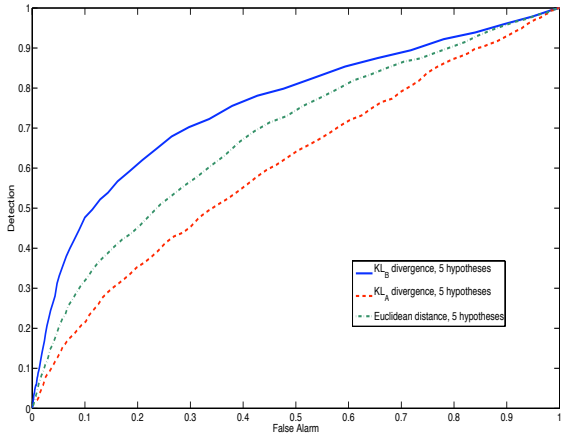

Figure 3: ROC curve corresponding to a different distance/divergence measures when 5 hypotheses were used to estimate $P(c_t = k|X, \Theta_{asr})$.

Table 1 gives the FOM for ROC curve corresponding to the different distance measures and the number of hypothesis that are used to estimate $P(c_t = k|X, \Theta_{asr})$. The performance of system given in the row KL_B *divergence* for number of hypothesis equal to 1 is the baseline performance. The reason for this case is the following: given a single hypothesis $P(c_t = k|X, \Theta_{asr})$ is either 1 or 0. Using this fact, it can be seen that $d_t = KL_B$ is nothing but $-\log(P(c_t = k|x_t, \Theta_{pc}))$.

Number of hypotheses	1	2	5	10
KL_A divergence	0.70	0.65	0.60	0.59
KL_B divergence	0.75	0.75	0.76	0.75
Euclidean distance	0.72	0.69	0.68	0.68

Table 1: Area under the ROC curve: results obtained by varying the number of hypotheses to estimate $P(c_t = k|X, \Theta_{asr})$ and using different distance/divergence measures.

The experimental results show that KL_B divergence yields consistently better system compared to systems using KL_A divergence and Euclidean distance. It is interesting to note that for the one best hypothesis all systems yield best performance except KL_B. In order to understand the reason behind this, we computed the average value of the posterior probabilities $P(W_i|X, \Theta_{asr})$ across the test utterances for 50 hypotheses. The values of average $P(W_i|X, \Theta_{asr})$ for the first four hypotheses are given in Table 2.

Ordinal number of hypotheses	Average hypotheses score
1	0.945
2	0.047
3	0.006
4	0.002

Table 2: Average value of $P(W_i|X, \Theta_{asr})$ computed over all the test utterances for the first four hypotheses.

It can be seen that $P(W_i|X, \Theta_{asr})$ is highest when W_i is the first hypothesis i.e. $i = 1$ and it is low for subsequent hypotheses. This means that the phonemes present in the first hypothesis get higher weight values while estimating $P(W_i|X, \Theta_{asr})$.

5. DISCUSSION OF THE METHOD AND RESULTS

In this paper we focused our attention on finding the distance measure amongst the out-of-context and in-context estimates of posterior probabilities for localizing the area where OOV word occurred. We want to point out different observations confirmed by experiments, which directly influenced the results:

- *Bottom-up system*: For a small vocabulary task it is shown in [5] that is more convenient to use out-of-context posterior estimates as a reference posteriors for computing a KL divergence measure. It is a intuitive result, since for this system (digit recognition) recognition accuracy was very high, while the number of considered classes was small (29 phoneme classes). In this case bottom-up recognition was more accurate then a top-down approach. In contrary, for large vocabulary task recognition accuracy is lower, while the number of phoneme classes is higher (we used 45 phoneme classes), so it is easier to confuse phonemes during a recognition process. Therefore, in our case it was more convenient to "trust" to the recognizer which uses the acoustic and lexical knowledge (in-context posteriors), which explains the results in the Table 1. Better phoneme recognizer is needed in order to intrinsically improve described model for OOV word detection.
- *Top-down system*: From the Table 2 is obvious that the

posterior estimates from the top-down approach in our experiments depend only on a few number of hypotheses. When the OOV word occurs, recognizer should be able to allow more variations in these utterance regions as the number of considered hypotheses grows.

Also, we noticed that some phonemes do not occur in test sentences, which causes their probability value to be set to zero. These values in a realistic case would be close to zero, but never equal to it. This case introduces difficulties, since KL divergence defined with equation (6) is not defined for division with zero values. One possible way to avoid explicit thresholding of a zero values, which further on influences on the distance amongst posterior estimates, is integration of a nonparametric interpolation method (e.g. weighted interpolation of in-context posteriors with out-of-context posteriors, with higher weight for the in-context posteriors). This method can overcome the problem of zero values existence and it can also decrease the mismatch effect between two recognizers which occurs on a word and phoneme boundaries.

6. CONCLUSIONS

In this paper, we investigated different distance/divergence measures for the comparison of the out-of-context and in-context phoneme posterior probabilities. Experimental studies conducted on WSJ 5k task show that KL-divergence estimated between in-context and out-of-context phoneme posterior probabilities with in-context phoneme posterior probabilities as reference distribution yields consistently better system.

7. ACKNOWLEDGMENTS

This work is supported by the European IST Programme Project FP6-0027787 under the integrated project DIRAC, Detection and Identification of Audio-Visual Cues, and by the DARPA GALE program. This paper only reflects the authors' views and funding agencies are not liable for any use that may be made of the information contained herein.

The authors would like to thank Hamed Ketabdar and Mirko Hannemann, as well as Lukas Burget and Martin Karafiat (from Brno University) for their help. First author expresses special thanks to Sriram Ganapathy and Guillermo Aradilla Zapata for helpful discussions.

REFERENCES

[1] F. Wessel, R. Schlüter, K. Macherey and H. Ney, "Confidence measures for large vocabulary continuous speech recognition", in *IEEE Trans. Speech and Audio Processing*, vol. 9, no. 3, pp. 288-298, 2001.

[2] I. Bazzi and J. R. Glass, "Modeling Out-of-Vocabulary Words for Robust Speech Recognition", in *Proc. ICLSP 2000*, Beijing, China, October 16-20. 2000.

[3] H. Lin, J. Bilmes, D. Vergyri and K. Kirchhoff, "OOV Detection by Joint Word/Phone Lattice Alignment ", in *IEEE Automatic Recognition and Understanding (ASRU) 2007*, Japan, pp. 478–483, Dec. 9-13 2007.

[4] I. Szöke, P. Schwarz, P. Matejka, L. Burget, M. Karafiát, M. Fapšo and J. Černocký, "Comparison of Keyword Spotting Aproaches for Informal Continuous Speech ", in *Interspeech'05*, Lisboa, Portugal, pp. 633–636, Sep. 4–8 2005.

[5] H. Ketabdar, M. Hanneman and H. Hermansky,"Detection of Out-of-Vocabulary Words in Posterior Based ASR ", in *Proc. of Interspeech'07*, Antwerp, Belgium, Aug. 27-31 2007.

[6] Solomon Kullback, *Information Theory and Statistics*, Dover Publications

[7] L. Burget, P. Schwarz, P. Matejka, M. Hannemann, A. Rasrtow, C. White, S. Khudanpur, H. Hermansky and J. Černocký", "Combination of strongly and weekly constrained classifiers for reliable detection of OOVs", in *To appear in ICASSP'08*

[8] J. B. Allen , "Articullatiion and Intelligibility ", Morgan & Claypool, 2005.

[9] C. Van Petten, S. Coulson, S. Rubin, E. Plante and M. Parks, "Time course of word identification and semantic integration in spoken language", *Journal of Experimental Psychology: Learning, Memory and Cognition*, vol.25, no. 2, 394–417, March 1999.

[10] A. Boothroyd, "Speech perception and sensorineural hearing loss in Auditory Management of Hearing - Impaired Children", M. Ross and G. Giolas, Eds., University Park, Baltimore, MD, 1978.

[11] A. Boothroyd and S. Nittrouer, "Mathematical treatment of context effects in phoneme and word recognition", *Acoustical Society of America Journal 1988*, vol. 84 (1), 101–114, 1988.

[12] G.A. Miller, G. A. Heise and W. Lichten, "The intelligibility of speech as a function of the context of the test material", *Journal of Experimental Physchology*, vol. 41, pp.329–335, 1951.

[13] L. Rabiner and B.-H. Juang, "An Introduction to Hidden Markov Models ", *IEEE Acoustics, Speech and Signal Processing Magazine 1986*", vol 3(1), 4–16, Jan. 1986

[14] Guillermo Aradilla and Hervé Bourlard, "Posterior-Based Features and Distances in Template Matching for Speech Recognition ", in *Proc. Multimodal Interaction and Related Machine Learning Algorithms (MLMI)*, Prague, June 2007

[15] Pierre Baldi and Laurent Itti "Bayesian Surprise Attracts Human Attention", in Proc. Neural Information Processing Systems NIPS 2005 Vancouver, British Columbia, Canada, Dec. 5–8 2005.

[16] Hynek Hermansky, "Perceptual linear predictive (PLP) analysis of speech", *The Journal of the Acoustic Society of America*, vol. 87, Issue 4, pp.1738–1752, April 1990

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

[18] H. Bourlard, S. Bengio, M. Magimai.- Doss, Q. Zhu, B. Mesot and N.Morgan, "Towards using hierarchical posteriors for flexible automatic speech recognition system", DARPA RT-04 Workshop, Nov. 2004.

[19] G. Bernardis and H. Bourlard, "Improving Posterior Based Confidence Measures in Hybrid HMM/ANN Speech Recognition Systems ", in *Proc. of ICSLP'98*, Sydney, Australia, pp. 775–778, 1998.