IMPROVING POSTERIOR BASED CONFIDENCE MEASURES USING ENHANCED LOCAL POSTERIORS

Hamed Ketabdar

Deutsche Telekom Laboratories, Technical University of Berlin
Ernst-Reuter-Platz 7, 10587, Berlin, Germany
phone: + (49) 15116135305, email: hamed.ketabdar@telekom.de

ABSTRACT

In this paper, we propose a new technique for enhancing posterior probability based confidence measures in ASR systems. We propose to enhanced local posterior estimates used in confidence measurement process, in order to improve the overal confidence score in terms of ability to accept/reject a hypothesis. Posterior based confidence measures are global scores obtained by accumulating local evidences. These local evidences are usually phone posterior probabilities estimated in frame basis from speech signal. Having better (more informative) local evidences can potentially lead to better confidence measures. In [1, 2, 3], a method for enhancing local phone posterior estimates (evidences) has been proposed. This method is based on integrating prior knowledge (such as phone duration, lexical knowledge) and temporal context in the local posterior estimation. We show that using enhanced local posteriors in the confidence measurement process significantly and constantly improves their ability to predict whether a hypothesis (at word or phone level) is correct or incorrect, as compared to using regular local posterior estimates.

1. INTRODUCTION

A confidence measure is a score that is applied to the speech recognition output. It gives an indication of how confident we are that the unit to which it has been applied (e.g. a phrase, word, phone) is correct. A word may be hypothesized with low confidence when the word model is matched against unclear acoustics caused by disfluencies or noise, or when an out-of-vocabulary (OOV) word is encountered. Confidence measures can be used to reject those hypotheses which are likely to be erroneous (i.e., have a low confidence) in a hypothesis test. Over the last two decades, considerable research has been devoted to the development of confidence scores associated with the outputs of ASR systems [4, 5, 6, 7]. A reliable measure for the confidence of a speech recognizer output is useful in many applications. These measures have been used mostly to help spot keywords in spontaneous or read texts, and to provide a basis for the rejection of OOV words. Many other ASR applications could also benefit from knowing the level of confidence for a recognized word. For example, text-dependent speaker recognition systems could put more emphasis on words recognized with higher confidence; unsupervised adaptation algorithms could adapt the acoustic model only when the confidence level is high, human-made transcriptions could be verified by ASR systems outputting their confidence in the transcribed word sequence, etc.

In this work, our preliminary concern is confidence measures for posterior based ASR systems such as hybrid HMM/Artificial Neural Networks (ANNs) [8] and Tandem [9]. Several confidence measures have been proposed for posterior based ASR, particularly hybrid HMM/ANN systems [10, 11, 12, 13]. Artificial Neural Networks (ANNs) are capable of providing good estimates of local posterior probability $p(q_t^i|x_t)$ of an HMM state/phone q^i at time t given an acoustic feature vector x_t. Hybrid HMM/ANN systems thus seem particularly well suited to generate confidence measures since, by definition posterior probabilities measure the probability of being correct. Usually Multi-Layer Perceptrons (MLPs) are used for the local phone posterior estimation. The posterior based confidence measures (PCMs) [10, 11] are existing at the word and at the phone levels. They are estimated based on accumulating local phone posteriors (estimated by MLP) within a phone or word hypothesis boundary, followed by normalization with respect to the length of the hypothesis.

In [1, 2, 3], a method for enhancing the estimation of local posterior probabilities is proposed. According to this method, phone posterior estimation is enhanced by integrating prior phonetic and lexical knowledge, as well as long temporal context. This is achieved by post-processing regular phone posteriors (estimated by MLP) through an HMM. The prior knowledge is encoded in the topology of this HMM. The outcome of this process is what we call as enhanced or more informative phone posteriors. A more detailed explanation of the posterior enhancement is given in Section 3.

In this work, we study the use of mentioned enhanced posteriors as local phone posteriors, replacing the regular MLP posteriors in confidence measurement methodologies. Since the enhanced posteriors are expected to be more informative than the regular MLP posteriors (due to integrating prior and contextual knowledge), they can potentially lead to better (more reliable) confidence measures. We show that enhanced posteriors used in confidence measurement consistently outperform regular posterior performance for predicting whether a hypothesis is correct or incorrect.

The paper is organized as follows: Section 2 provides an introduction on posterior based confidence measures. Section 3 explains the method for enhancing local posterior estimates and their usage in confidence measurement. Section 4 presents experiments for comparing regular and enhanced posteriors in confidence measurement and corresponding results. Finally, Section 5 concludes the paper.

2. POSTERIOR BASED CONFIDENCE MEASURES

As already mentioned in the Introduction, posterior based confidence measures are global scores measured usually at the word or phone hypothesis level based on accumulating local phone evidences (posteriors). These measures are then normalized with respect to the length of the hypothesis. The local phone evidences are usually in the form of posterior probability estimated for one or a few speech frames, $p(q_t^i|x_t)$, where q_t^i is phone i at time (frame) t, and x_t is acoustic feature vector at time t. This posterior probability is estimated using a Multi-Layer Perceptron (MLP). Each MLP output is associated with one phone class. Acoustic feature vector(s) are presented at MLP input, and the MLP estimates phone posterior probabilities for the current frame at the output.

2.1 Phone Confidence Measures

At the phone hypothesis level, the normalized posterior based confidence measure, denoted *NPCM* is defined as the logarithm of a global phone posterior probability computed as the product of the local phone posteriors along the optimal state sequence, and normalized by the duration of the phone hypothesis [10, 11]. For a phone hypothesis q^i, starting at frame b and ending at frame e, the

© EURASIP, 2010 ISSN 2076-1465

confidence measure is defined as:

$$NPCM(i) = \frac{1}{e-b+1}\sum_{t=b}^{e}\log p(q_t^i|x_t) \tag{1}$$

The normalization is necessary due to different phone durations, as otherwise short phones would be favored.

2.2 Word Confidence Measures

The word confidence measures are defined in a similar manner. For a word hypothesis w, composed of a sequence of L phone hypotheses $(q^1, ..., q^l, ..., q^L)$, the $frame-basedNPCM(w)$ is defined as:

$$frame-basedNPCM(w) = \frac{1}{\sum_{l=1}^{L}(e_l - b_l + 1)}\sum_{l=1}^{L}\sum_{t=b_l}^{e_l}\log p(q_t^l|x_t) \tag{2}$$

where b_l and e_l are respectively the beginning and end frames of phone hypothesis q^l in the considered word. A second word confidence measure can be defined by doing a secondary normalization with respect to the number of phones in the hypothesized word. This measure is called $phone-basedNPCM(w)$, and defined as follows:

$$phone-basedNPCM(w) = \frac{1}{L}\sum_{l=1}^{L}\Big(\frac{1}{e_l - b_l + 1}\sum_{t=b_l}^{e_l}\log p(q_t^l|x_t)\Big) \tag{3}$$

There are also other alternatives to these confidence measures such as mean posterior confidence measures (MPCMs). MPCMs at phone and word levels are computed as NPCMs in (1), (2) and (3), except that we compute the average of local posteriors before taking the logarithm.

For all these measures, phone and word hypothesis boundaries (b_l and e_l) and optimal state sequence are obtained using Viterbi decoding by back tracking the decoded state sequence.

3. ENHANCED POSTERIORS IN CONFIDENCE MEASUREMENT: MORE INFORMATIVE LOCAL EVIDENCES

Confidences measures defined in the previous section are global scores obtained by accumulating local evidences (phone posteriors). Having better (more informative) local evidences can potentially lead to better confidence measures. In [1, 2, 3] a new method for enhancing estimation of local posteriors based on integrating prior phonetic and contextual knowledge, as well as long temporal context have been proposed. In this approach, regular MLP phone posteriors are used as local scores (emission probabilities) in HMM forward-backward recursions. The outcome of these recursions are the so called “HMM state posterior probability”. Assuming that each phone is modeled with one HMM state, this state posterior probability is a phone posterior probability and can be considered as enhanced (more informative) version of regular MLP posterior. This is because it additionally integrates prior knowledge (phonetic, lexical knowledge) encoded in HMM topological constraints as well as temporal context. There are two terms contributing in HMM forward-backward recursions: (1) emission likelihood which is obtained from regular MLP posteriors, and (2) HMM topological constraints (transition probabilities). The outcome of the recursions (enhanced phone posteriors) has contribution of regular phone posteriors, as well as HMM topological constraints encoding prior knowledge. We denote the enhanced posteriors as $p(q_t^i|M, X_T)$ which is posterior probability of a certain phone i at time t, q_t^i, taking into account prior knowledge encoded in HMM topology (M), as well as temporal context as available in the whole utterance (X_T). $p(q_t^i|M, X_T)$ is estimated through HMM forward-backward recursions as follows.

In order to use regular MLP posteriors in HMM recursions, they are first turned to the so called “scaled likelihoods” [8] by dividing MLP phone posteriors by their respective class priors $p(q_t^i)$:

$$\frac{p(q_t^i|x_t)}{p(q_t^i)} = \frac{p(x_t|q_t^i)}{p(x_t)} \tag{4}$$

The scaled likelihoods are then used instead of emission likelihoods in “scaled forward-backward” recursions [14]. Scaled forward-backward recursions, $\alpha_s(i,t)$ and $\beta_s(i,t)$, are defined as[1]

$$\begin{aligned}\alpha_s(i,t) &= \frac{p(x_{1:t}, q_t^i)}{\prod_{\tau=1}^{t} p(x_\tau)}\\ \beta_s(i,t) &= \frac{p(x_{t+1:T}|q_t^i)}{\prod_{\tau=t+1}^{T} p(x_\tau)}\end{aligned}$$

and it can be shown that they can be expanded as follows [14]:

$$\begin{aligned}\alpha_s(i,t) &= \frac{p(q_t^i|x_t)}{p(q_t^i)}\sum_j p(q_t^i|q_{t-1}^j)\alpha_s(j,t-1)\\ \beta_s(i,t) &= \sum_j \frac{p(q_{t+1}^j|x_{t+1})}{p(q_{t+1}^j)}p(q_{t+1}^j|q_t^i)\beta_s(j,t+1)\end{aligned}$$

Finally, the enhanced phone posterior is estimated as:

$$p(q_t^i|X_T) = \frac{\alpha_s(i,t)\beta_s(i,t)}{\sum_j \alpha_s(j,t)\beta_s(j,t)} \tag{5}$$

It has been shown that using enhanced posteriors lead to better recognition performance at the frame, phone and word levels, indicating that they are better (more precise) local estimators than the regular MLP posteriors [1, 2, 3]. Therefore, they can provide better (more informative) local evidences for phones in the confidence measurement process. This means that using the enhanced posteriors instead of the regular MLP posteriors can potentially improve the confidence measures previously defined. In order to evaluate this idea, the local posterior estimates (MLP outputs) in the definitions of Section 2 (Equations 1-3) are simply replaced with the enhanced posterior estimates. In the following, the performance of the two types of posteriors (regular and enhanced) for confidence estimation is compared.

4. EXPERIMENTS AND RESULTS

The confidence measures are evaluated in terms of their ability to predict whether a particular phone or word hypothesis is correct or incorrect. A hypothesis is rejected if its confidence score falls below a threshold. Two types of error can occur: Type I error corresponding to the rejection of a correct hypothesis, and type II error corresponding to the acceptation of an incorrect hypothesis. The performance of confidence measures is then evaluated in terms of type I and type II errors, and the classification error rate (CER) is defined as:

$$CER = \frac{\text{Type I errors + Type II errors}}{\text{Total number of hypotheses in the test set}} \tag{6}$$

CER has been conventionally used in related posterior based confidence measure studies to evaluate the performance.

For the experiments, we have used a partition of Wall Street Journal (WSJ) Database [15]. There are 45 phones and 5k words in this database. The training set size is about 70 hours and the test size is about 1.1 hours. The test set is recognized using Viterbi decoding through the best trained ASR model available for the task. The ASR model is a HMM/GMM using context-dependent models for phone acoustic modeling and a bi-gram language model for decoding. The decoding generates word and phone level hypotheses

[1] In all the presented HMM recursions, we assume that a phone is modeled with one state, thus we can use the same notation for phones and states.

and segmentations. For the evaluation, the decoding results and reference word and phone sequences were aligned so that each hypothesis could be marked as correct or incorrect, allowing the evaluation of the performance of each of the confidence measures as hypothesis test statistics. In order to make the performance differences clear between the different confidence measures, the number of true and false word hypotheses in the test set were equalized for each condition. This was done by counting the number of false hypotheses for a condition and randomly selecting the same number from the set of true hypotheses for that condition[2]. Equalizing the number of true and false hypotheses had the effect of artificially raising the recognizer error rate close to 0.5 for each condition.

Confidence levels are then estimated at the phone and word levels for each hypothesis using the described measures. For estimating regular phone posteriors, we have used an MLP with 351 (corresponding to 9 frames of 39 dimension PLP features) input, 2000 hidden, and 45 output nodes (corresponding to the number of phones). The MLP is trained with the 70 hours of training set data, and then used to estimate local phone posteriors for the test data set.

In order to estimate enhanced posteriors, phone duration information was integrated in the regular MLP posterior estimation. This was achieved by using an HMM composed of ergodic connection of all phone models. In this HMM, each phone is modeled with 3 states implying a minimum phone duration of 3 frames as prior knowledge.

All the NPCM and MPCM confidence measures defined in (1-3) are estimated using both regular and enhanced posteriors. The confidence measures are then compared with a range of thresholds to decide about acceptance/rejection of hypotheses. Finally, CER values are computed as previously described.

4.1 Phone Confidence Measures

Figures 1 and 2 are showing performance curves for NPCM and MPCM phone level confidence measures obtained using regular and enhanced posteriors. Regular posterior results are plotted in blue and enhanced posterior results are plotted in red. The horizontal axis shows the percentage of hypotheses that were rejected and is a function of the confidence threshold. The vertical axis shows the CER percentage. The area under the error curves corresponding to the enhanced posteriors is smaller (i.e. better trade-offs) compared to the ones corresponding to the regular posteriors. This is consistent for both NPCM and MPCM measures.

4.2 Word Confidence Measures

The same study is repeated for the NPCM and MPCM word confidence measures defined in (2, 3). Figures 3 and 4 are showing the results for different word confidence measures estimated using regular and enhanced posteriors. The results corresponding to regular posteriors are plotted in blue, and results corresponding to enhanced posteriors are plotted in red. Again, it can be observed that the enhanced posteriors are consistently performing better than the regular posteriors for confidence measurement. For all the measures (frame and phone-based NPCM, frame and phone-based MPCM), the area under the error curves corresponding to enhanced posteriors is smaller.

The experiments confirm that using enhanced posteriors instead of regular MLP posteriors in confidence measurement consistently improves their ability for accepting/rejecting a hypothesis at phone or word levels.

5. SUMMARY AND CONCLUSIONS

In this paper, we investigated a new method for enhancing posterior probability based confidence measures in ASR systems. We presented the conventional confidence measures defined for hybrid HMM/ANN ASR. We proposed to use the so called "enhanced phone posteriors" instead of the regular posteriors in the confidence measurement. Enhanced posteriors are enriched by integrating prior and contextual knowledge in the posterior estimation. As confidence measures are calculated based on local posteriors, and enhanced posteriors provide better (more informative) local evidences of phones, they are expected to improve the performance of confidence measures. The experiments showed that using enhanced posteriors, the confidence measures are consistently performing better for predicting whether a hypothesis is correct or incorrect at word and phone levels, as compared to using regular phone posteriors.

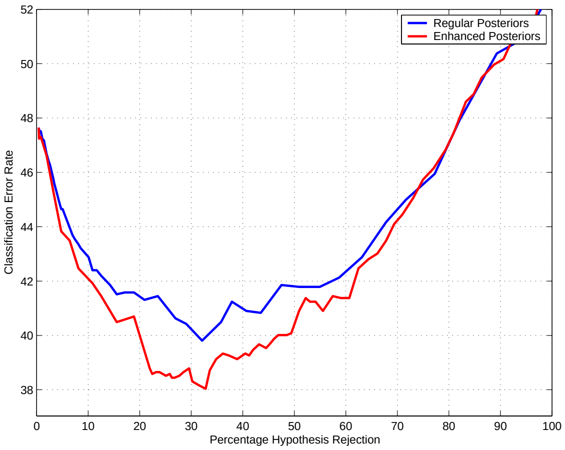

Figure 1: CER curves for NPCM phone hypothesis confidence measure. The y axis is showing CER percentage and the x axis is showing phone hypothesis rejection percentage. The blue curve is obtained using regular posteriors and the red curve is obtained using enhanced posteriors.

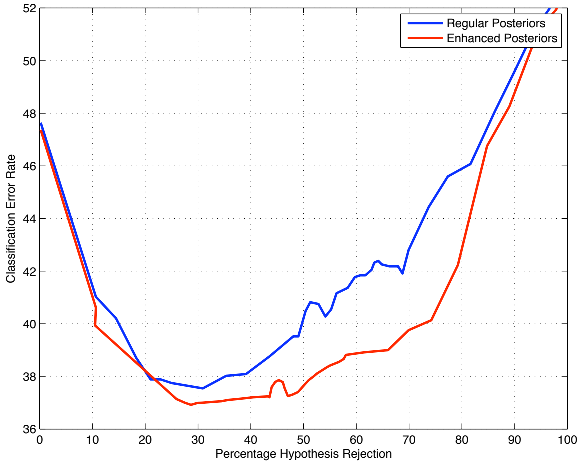

Figure 2: CER curves for MPCM phone hypothesis confidence measure. The conditions are the same as Fig. 1.

6. ACKNOWLEDGEMENTS

The author would like to highly acknowledge contribution of Herve Bourlard in initiating and supervising the work on enhancing local posterior probabilities and their applications, and advising this work. The author also thanks Mathew Magimai Doss and Joel Pinto for helpful discussions.

REFERENCES

[1] Hamed Ketabdar, Jithendra Vepa, Samy Bengio and Herv Bourlard, "Using More Informative Posterior Probabilities

[2] In practice, since the equalization is done at the utterance level, the number of true and false hypotheses are not exactly equal, but very close.

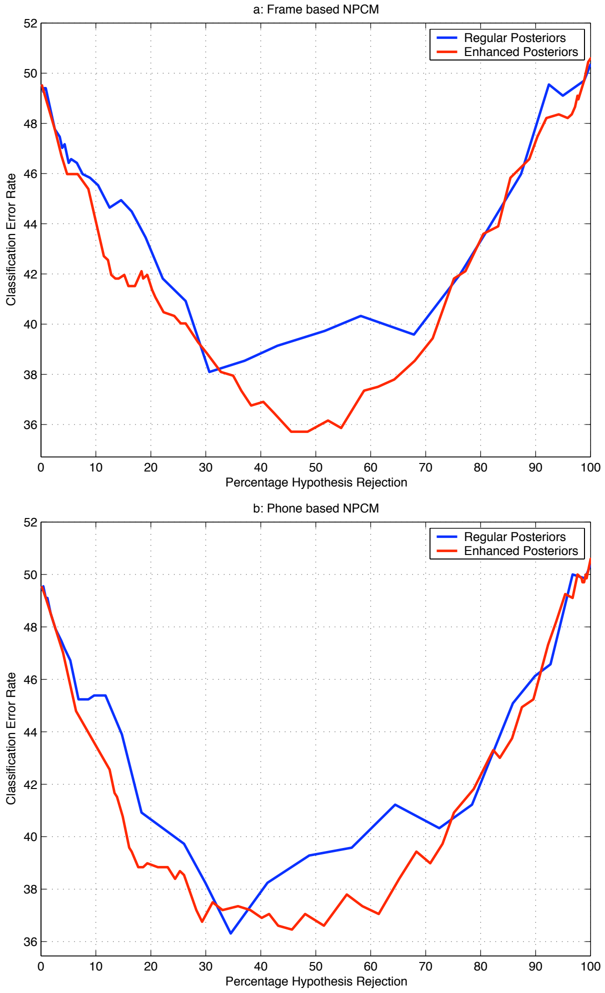

Figure 3: CER curves for NPCM word hypothesis confidence measures. (a) The error curves for $frame - basedNPCM$ measures, and (b) the curves for $phone - basedNPCM$ measures. The y axis is CER percentage and the x axis is word hypothesis rejection percentage. The blue curves are obtained using regular posteriors and the red curves are obtained using enhanced posteriors.

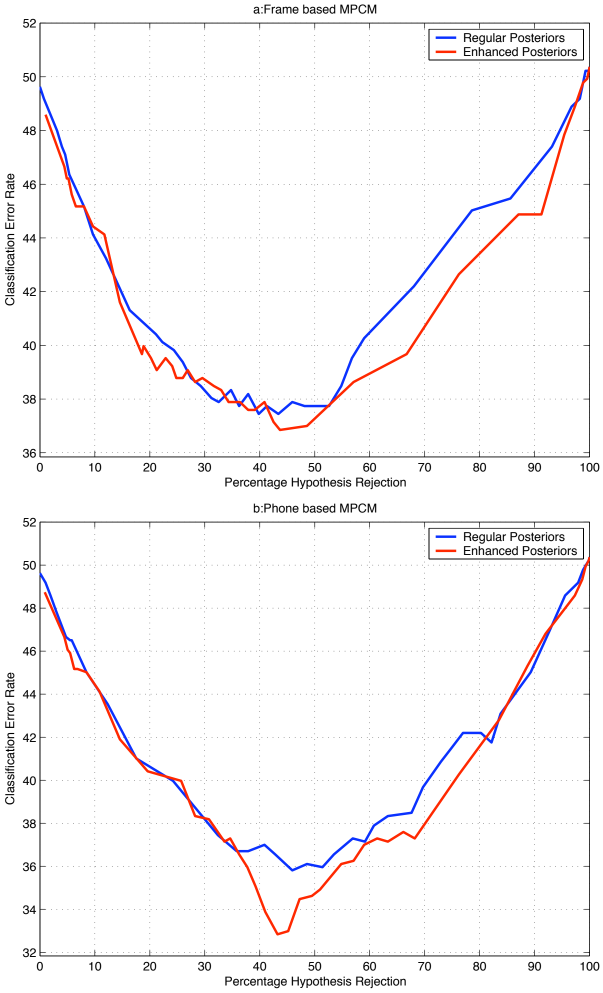

Figure 4: CER curves for MPCM word hypothesis confidence measures. (a) The error curves for $frame - basedMPCM$ measure, and (b) the curves for $phone - basedMPCM$ measure. The y axis is CER percentage and the x axis is word hypothesis rejection percentage. The blue curves are obtained using regular posteriors and the red curves are obtained using enhanced posteriors.

for Speech Recognition", In Proceedings of ICASSP 2006, Toulouse, France.

[2] Hamed Ketabdar and Herve Bourlard, "In-context Posteriors as Complementary Features for TANDEM ASR", In Proceedings of Interspeech07, Belgium, August 2007.

[3] Hamed Ketabdar and Hervé Bourlard, "Enhanced Phone Posteriors for Improving Speech Recognition Systems", to appear in IEEE Transactions on Speech and Audio Processing.

[4] F. Wessel, R. Schlueter, K. Macherey, and H. Ney, Confidence Measures for Large Vocabulary Continuous Speech Recognition, IEEE Trans. Speech and Audio Processing, vol. 9, no. 3, pp. 288-298, March 2001.

[5] Z. Rivlin, M. Cohen, V. Abrash, Th. Chung, A Phone Dependent Confidence Measure for Utterance Rejection, Proc. IEEE Intl. Conf. on Acoustic, Speech and Signal Processing, pp. 515-518, Atlanta, USA, 1996.

[6] R. C. Rose and D. B. Paul, "A Hidden Markov Model Based Keyword Recognition System," Proc. IEEE Intl. Conf. on Acoustic, Speech and Signal Processing, pp. 129- 132, 1990.

[7] R. A. Sukkar and J. G. Wilpon, A Two Pass Classier Utterance Rejection in Keyword Spotting, Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, vol. 2, pp. 451-454, 1993.

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

[9] Hermansky, H., Ellis, D.P.W., and Sharma, S., "Connectionist Feature Extraction for Conventional HMM Systems", *Proc. ICASSP*, 2000.

[10] G. Bernardis and H. Bourlard, "Improving Posterior Confidence Measures in Hybrid HMM/ANN Speech Recognition System," Proceedings of the Intl. Conference on Spoken Language Processing, pp. 775-778, Australia, 1998.

[11] G. Williams and S. Renals, "Confidence Measures for Hybrid HMM/ANN Speech Recognition," Proceedings of Eurospeech97, pp. 1955-1958, Greece, 1997.

[12] G. William and S. Renals, "Confidence Measures from Local Posterior Probability Estimates", Computer, Speech and Language, vol. 13, pp. 395-411, 1999.

[13] G. Williams and S. Renals, "Confidence Measures for Evalu-

ating Pronunciation Models," In ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition, pp. 151-155, Kerkrade, Netherlands, 1998.

[14] J. Hennebert, C. Ris, H. Bourlard, S. Renals, and N. Morgan, "Estimation of Global Posteriors and Forward-Backward Training of Hybrid HMM/ANN Systems," Proceed- ings of EUROSPEECH97, pp. 1951-1954, Rhodes, Greece, 1997.

[15] J. Fransen, D. Pye, T. Robinson, P. Woodland, and S. Young, "WSJCAM0 Corpus and Recording Description," Technical Report 192, Cambridge University Engineering De- partment, 1994.