I D I A P   R E S E A R C H   R E P O R T

IDIAP

Martigny - Valais - Suisse

IMPROVING POSTERIOR BASED CONFIDENCE MEASURES IN HYBRID HMM/ANN SPEECH RECOGNITION SYSTEMS

Giulia Bernardis [a]   Hervé Bourlard [a b]

IDIAP–RR 98-11

AUGUST 1998

TO APPEAR IN
Proceedings of ICSLP'98

Dalle Molle Institute for Perceptual Artificial Intelligence • P.O.Box 592 • Martigny • Valais • Switzerland

phone +41 − 27 − 721 77 11
fax +41 − 27 − 721 77 12
e-mail secretariat@idiap.ch
internet http://www.idiap.ch

[a] Dalle Molle Institute for Perceptual Artificial Intelligence (IDIAP), Martigny, Switzerland
[b] Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland

IDIAP Research Report 98-11

IMPROVING POSTERIOR BASED CONFIDENCE MEASURES IN HYBRID HMM/ANN SPEECH RECOGNITION SYSTEMS

Giulia Bernardis Hervé Bourlard

AUGUST 1998

TO APPEAR IN
Proceedings of ICSLP'98

Abstract. In this paper, building upon previous work by others [7], we define and investigate a set of confidence measures based on hybrid Hidden Markov Model/Artificial Neural Network (HMM/ANN) acoustic models. All these measures are using the neural network to estimate the local phone posterior probabilities, which are then combined and normalized in different ways. Experimental results will indeed show that the use of an appropriate duration normalization is very important to obtain good estimates of the phone and word confidences.
The different measures are evaluated at the phone and word levels on both an isolated word task (PHONEBOOK) and a continuous speech recognition task (BREF). It will be shown that one of those confidence measures is well suited for utterance verification, and that (as one could expect) confidence measures at the word level perform better than those at the phone level. Finally, using the resulting approach on PHONEBOOK to rescore the N-best list is shown to yield a 34% decrease in word error rate.

1 INTRODUCTION

In hybrid HMM/ANN speech recognition systems, ANNs are used to estimate the K local posterior probabilities $p(q_k|x_n)$ of phone classes q_k (with $k \in \{1, \ldots, K\}$) given the current acoustic vector x_n (usually including some additional contextual information) [1, 3]. Hybrid HMM/ANN systems thus seem particularly well suited to generate confidence measures since, by definition, posterior probabilities also measure the probability of being correct. In [7], different confidence measures were compared in the framework of hybrid HMM/ANN systems and it was shown that a posterior based measure directly using the ANN outputs was best performing at the phone level. At the word level though, although still being among the best methods, this conclusion was overruled by an approach based on the word lattice density. Unfortunately, it was recently shown in [6] that this conclusion did not hold for all databases and that in some cases (broadcast news), the lattice approach was yielding the worst performance while the posterior based measure was more consistent.

In this paper, we investigate further HMM/ANN confidence measures based on posterior probabilities (as well as on entropies of the local phone posterior probabilities) and we show that the posterior based measures are consistently yielding very good performance not only at the phone level (as done in [7]), but also at the word (and possibly sentence) level, provided that the right normalization techniques are used during HMM integration. Indeed, since the hypothesized phones and words (eventually sentences) do not necessarily have the same length, it is necessary to normalize them to have a reliable measure. As a matter of fact, we show here that a double normalization, involving the number of frames in each phone and the number of phones in each word, significantly improves the approach proposed in [7].

2 CONFIDENCE MEASURES

ANNs are capable of providing good estimates of the posterior probability $p(q_k^n|x_n)$ of an HMM state/phone q_k at frame n given an acoustic feature vector x_n (possibly including context). These local posterior probabilities may be combined in different ways to produce an estimate of the global posterior probability (i.e., a confidence measure) of a recognized unit given the acoustic observations. Alternatively, these local posteriors are also well suited to compute the entropy of the ANN output distribution averaged over the segment for which we want to estimate the confidence level. We thus start this paper by describing the confidence measures at the phone level and at the word level that were used in the present study.

2.1 Posterior Based Confidence Measures

Normalized Posterior based Confidence Measures (NPCMs) are defined in three different ways: one at the phone level and two at the word level using two different kinds of duration normalization.

At the phone level, the normalized posterior based confidence measure, denoted $\mathbf{NPCM}(k)$, is defined as the logarithm of the 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 in frames. For a phone q_k, beginning at frame b and ending at frame e, this *frame-based* confidence level is defined as

$$\mathbf{NPCM}(k) = \frac{1}{e-b+1} \sum_{n=b}^{e} \log p(q_k^n|x_n) \tag{1}$$

This normalization compensates for different phone durations, as otherwise short phones would be favored.

Similarly, we can also define a word level confidence measure. For a word w, consisting of a sequence of J phone segments $(q_1, \ldots, q_j, \ldots, q_J)$, the *frame-based*$\mathbf{NPCM}(w)$ will thus be defined according to:

$$\textit{frame-based}\mathbf{NPCM}(w) = \frac{1}{\sum_{j=1}^{J}(e_j - b_j + 1)} \sum_{j=1}^{J} \sum_{n=b_j}^{e_j} \log p(q_j^n|x_n) \tag{2}$$

where b_j and e_j are respectively the first and last frame of phone segment q_j of the considered word.

At the word level, we also define another confidence measure called *phone-based*$\mathbf{NPCM}(w)$, and involving a double normalization taking into account the number of frames in each phone and the number of phones in each word, yielding the following estimate:

$$\textit{phone-based}\mathbf{NPCM}(w) = \frac{1}{J}\sum_{j=1}^{J}\left(\frac{1}{e_j - b_j + 1}\sum_{n=b_j}^{e_j}\log p(q_j^n|x_n)\right) \quad (3)$$

For comparison purposes, those confidence level estimates will be assessed against other alternatives 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 (and phone normalization, at the word level) before taking the logarithm.
- **Posterior Probability Confidence Measures (PPCMs)**: the confidence measures equivalent to (1) and (2), but resulting from a straight accumulation of local posteriors, without any normalization.
- *Scaled Likelihood Confidence Measures (SLCMs)*: we also tested all the above measures by dividing the local posterior probabilities by the a priori probabilities $p(q_k)$ as estimated on the training data, thus yielding estimates of local and global scaled likelihoods $p(x_n|q_k)/p(x_n)$. However, since this method never led to satisfactory performance compared to the above approaches, results will not be reported here.

2.2 Entropy Based Confidence Measure

At the phone level, the **Entropy Based Confidence Measure (EBCM)** is defined as the per frame entropy of the K phone class posterior probabilities estimated by the ANN, averaged over the phone segment (from frame b to e):

$$\mathbf{EBCM} = \frac{1}{e - b + 1}\sum_{n=b}^{e}\sum_{k=1}^{K} p(q_k^n|x_n)\log p(q_k^n|x_n) \quad (4)$$

This entropy measure differs from the other confidence measures in that it does not make use of the optimal state sequence (by the Viterbi decoding) and provides a confidence for a segment, rather than a $(segment, label)$ pair. It will be shown that this measure never provided us with relevant confidence measures and, consequently, results at the word level will not be reported. [1]

3 EXPERIMENTAL SET UP

The confidence measures were evaluated in terms of their ability to predict whether a particular hypothesis (phone or word) is correct or incorrect, and experiments were carried out on two databases:

- PHONEBOOK [2], a large isolated word telephone speech database. A training set of 21 lists of 75 words each pronounced by 11 speakers was used, together with 5 test sets each with a disjoint set of 75 words and 11 speakers (making this task speaker and vocabulary independent).
- BREF [4], a continuous, read speech microphone database. $3,736$ utterances were used for training and 242 utterances with a $2,300$ word lexicon for testing.

The phone models were repeated state HMMs (with self-loops and minimum duration equal to half the average duration) with an ANN (multilayer perceptron, in our case) output associated with each state/phone. Nine

[1] It is believed that this entropy based measure will be more appropriate when measuring the reliability of different features, independently of the recognized hypotheses, as opposed to assessing the confidence of recognized hypotheses.

feature vectors, $X_{n-4}^{n+4} = (x_{n-4}, \ldots, x_{n-1}, x_n, x_{n+1}, \ldots, x_{n+4})$, were used as input to the ANN. Each feature vector consisted of 12 RASTA-PLP coefficients, 12 Δ RASTA-PLPs, Δ log energy, and $\Delta\Delta$ log energy.

The test sets were recognized using Viterbi decoding, generating word and phone level hypotheses and segmentations. Confidence levels were then estimated at the phone and word level for each hypothesis using each of the measures described above.

For the evaluation, the decoding results and reference words (word sequences) 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. To do this, and to allow fair comparisons between different systems, the number of correct and incorrect hypotheses in the test sets were equalized by counting the number of incorrect hypotheses and selecting the same number of correct hypotheses from the test sets. This has the effect of artificially raising the recognizer's error rate to 50%. Evaluation sets at both phone and word levels were constructed accordingly.

Finally, and following [7], we considered this confidence based hypothesis test: "a recognition hypothesis is rejected whenever its confidence score falls below a threshold". Thus, two types of errors may occur: *TypeI error* corresponding to the rejection of a correct hypothesis and *TypeII error* corresponding to the acceptation of an incorrect hypothesis. The performance of confidence measures is then evaluated in terms of *TypeI* and *TypeII errors*, and the Classification Error Rate may be defined as:

$$\mathbf{CER} = \frac{\textit{TypeI errors} + \textit{TypeII errors}}{\textit{total hypotheses in evaluation set}} \tag{5}$$

4 EXPERIMENTAL RESULTS

4.1 PHONEBOOK

First we investigated the phone confidence measures on the evaluation set defined for phones ($2,400$ hypotheses). Figure 1 shows a comparison of the performance of the confidence measures $\mathbf{NPCM}(k)$, $\mathbf{MPCM}(k)$,

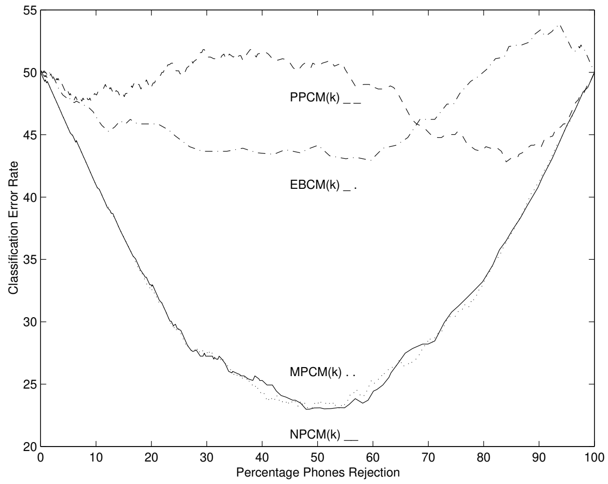

Figure 1: Performance of the confidence measures at the phone level.

$\mathbf{PPCM}(k)$, and $\mathbf{EBCM}(k)$. The figure reads like this: 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 which is 50% when *no* hypotheses are rejected (50% *TypeII errors*) and when *all* hypotheses are rejected (50% *TypeI errors*).

In Figure 1, it is shown that the $\mathbf{NPCM}(k)$ and the $\mathbf{MPCM}(k)$ significantly outperform the $\mathbf{PPCM}(k)$ and the $\mathbf{EBCM}(k)$ at the phone level. $\mathbf{PPCM}(k)$ probably fails because of its lack of duration normalization, which biases it towards accepting short phone hypotheses and rejecting long ones. $\mathbf{EBCM}(k)$ is a less powerful confidence measure when used for utterance verification: since it does not make explicit use of label information it is independent of the recognition results. Our investigations showed that the EBCM is more suited to assessing the quality of the recognition model (as opposed to assessing the confidence of the output results). From Figure 1, we conclude that $\mathbf{NPCM}(k)$ and $\mathbf{MPCM}(k)$ are basically yielding the same good performance and are both able to detect many of the incorrect hypotheses.

We also investigated confidence measures at the word level, on the evaluation set that was previously defined for words (586 hypotheses). Figure 2 shows the performance of the *frame-based* $\mathbf{PPCM}(w)$ and both

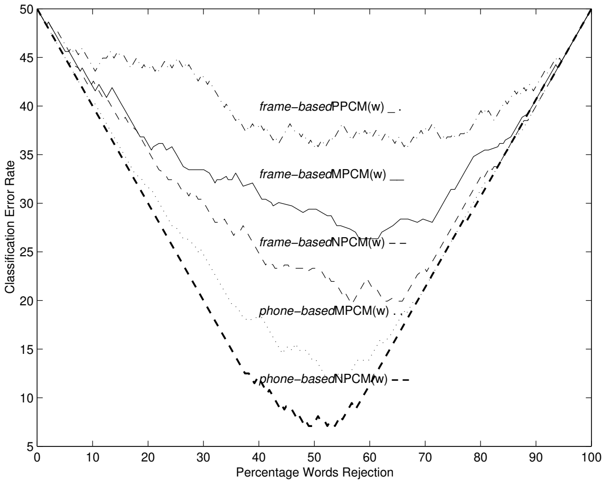

Figure 2: Performance of the posterior based confidence measures at the word level.

frame-based and *phone-based* $\mathbf{NPCM}(w)$ and $\mathbf{MPCM}(w)$ word confidence measures. As for the phone level, the $\mathbf{PPCM}(w)$ performs poorly because of its lack of duration normalization. With regard to both $\mathbf{NPCM}(w)$ and $\mathbf{MPCM}(w)$, it is clear from the figure that the *phone-based* measures (involving the double normalization in terms of number of phones and number of frames per phone) perform significantly better than the *frame-based* measures (with the normalization involving only the number of frames in each word). One possible reason for this could be that when a hypothesized word shares many phones with the correct word, but has several extra phones, those extra phone HMMs are forced to match poor acoustic segments. Very often, in order to get the best recognition match, those phones will have minimal duration in the Viterbi backtrace. Furthermore, since those recognized phones are incorrect, they will usually have poor posterior probability scores. Consequently, a confidence measure that is more sensitive to those few frames of poor posterior probability should be better to identify the unreliable words. Since the *frame-based* confidence measure weights frames equally, while the *phone-based* confidence measure weights phones equally, incorrect phones have a better chance to have an impact on the confidence measure with the latter measure.

The *phone-based* $\mathbf{NPCM}(w)$ is thus the best of all the considered confidence measures, and actually exhibits a very interesting behaviour. In Figure 2, it can be seen that its CER between $(0, 50)$ and $(37.5, 12.5)$ is basically a straight line, meaning that for rejection rates of up to 37.5% only incorrect hypotheses are rejected (straight line going from $(0, 50)$ to $(50, 0)$ and corresponding to the optimal solution). In other words, for this evaluation set, $3/4$ of the errors can be detected without any false rejection.

Figure 3 shows histograms of the *frame-based* $\mathbf{NPCM}(w)$ and *phone-based* $\mathbf{NPCM}(w)$ confidence scores. It can be observed that the distributions for correct and incorrect words are indeed much better separated for the *phone-based* $\mathbf{NPCM}(w)$ measure. The doubly normalized *phone-based* confidence measure is more discriminant than the simple *frame-based* measure, which corroborates further our previous discussion.

Finally, we tested our best confidence measure, i.e., *phone-based* $\mathbf{NPCM}(w)$, to rescore the N-best list resulting from a Viterbi search providing us with the N-best recognition hypotheses, as well as the phone segmentation associated with each hypothesis. Rescoring the 5-best hypotheses reduced the error rate from 7.1% down to 4.7% (see [2] for comparative results), i.e., a 34% relative improvement.

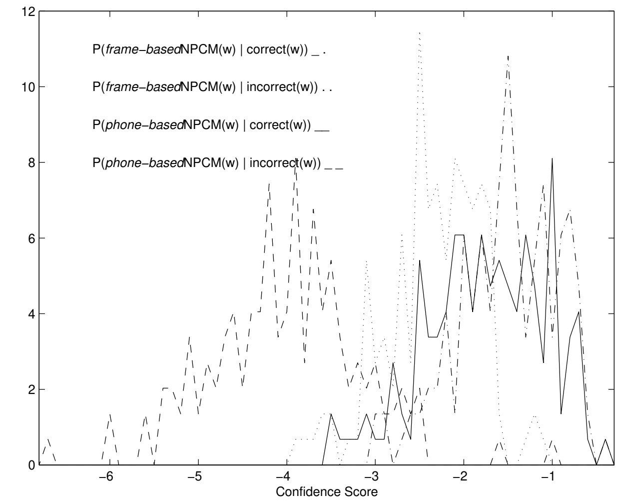

Figure 3: Distribution of the *frame-based* and *phone-based* $\mathbf{NPCM}(w)$ measures for correct and incorrect word hypotheses.

4.2 BREF

Posterior based confidence measures at the word level were also investigated on the BREF database on an evaluation set of $1,600$ word hypotheses and similar results and conclusions were obtained.

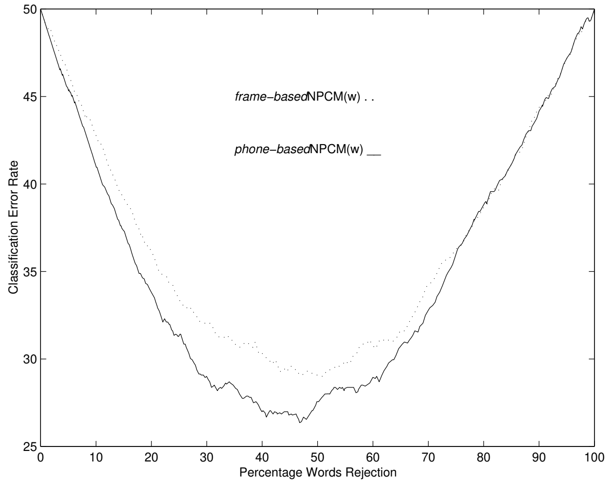

Figure 4: Performance of the *frame-based* and *phone-based* $\mathbf{NPCM}(w)$ measures on the word evaluation set from BREF.

Figure 4 shows the CER against the percentage of rejected words for the *frame-based*$\mathbf{NPCM}(w)$ and the *phone-based*$\mathbf{NPCM}(w)$ (the best confidence measures for isolated words). It can be observed that the previous conclusions still hold for continuous speech and that the confidence measure based on the double normalization is still yielding the best performance.

Unfortunately, it was also observed that the use of multiple pronunciations tends to reduce the difference between the different approaches, as well as the overall performance of the confidence measures. This can be explained by the fact that, when adding alternative pronunciations, all word likelihoods, including those for wrong hypotheses, are increased, resulting in poor confidence scores.

5 CONCLUSIONS

In this paper, we investigated HMM/ANN confidence measures based on posterior probabilities and showed that posterior based measures are yielding very good performance not only at the phone level, as initially shown in [7], but also at the word level, provided that the appropriate normalization is used. These conclusions were tested on both isolated word (PHONEBOOK) and continuous speech (BREF) tasks.

In the same framework, we are now investigating the possibility to automatically adapt the language model scaling factor as a function of our best confidence measure (e.g., computed for the previous hypothesized word), as initially proposed in [5].

Acknowledgements

We acknowlege the support of the Swiss Federal Office for Education and Science, in the framework of the COST249 and THISL European projects. We also thank Gethin Williams and Steve Renals for helpful discussions.

References

[1] H. Bourlard and N. Morgan. *Connectionist Speech Recognition - A Hybrid Approach*. Kluwer Academic Publishers, ISBN 0-7923-9396-1, 1994.

[2] S. Dupont, H. Bourlard, O. Deroo, V. Fontaine, and J.-M. Boite. Hybrid HMM/ANN systems for training indipendent tasks: Experiments on PHONEBOOK and related improvements. *Proceedings of ICASSP'97*, pages 1767–1770, 1997.

[3] J. Hennebert, C. Ris, H. Bourlard, S. Renals, and N. Morgan. Estimation of global posteriors and forward-backward training of hybrid HMM/ANN systems. *Proceedings of EuroSpeech'97*, pages 1951–1954, 1997.

[4] L.-F. Lamel, J.-L. Gauvain, and M. Eskénazi. BREF, a large vocabulary spoken corpus for French. *Proceedings of EuroSpeech'91*, pages 505–508, 1991.

[5] C. Neti, S. Roukos, and E. Eide. Word-based confidence measures as a guide for stack search in speech recognition. *Proceedings of ICASSP'97*, pages 883–886, 1997.

[6] S. Renals. Private communication.

[7] G. Williams and S. Renals. Confidence measures for hybrid HMM/ANN speech recognition. *Proceedings of EuroSpeech'97*, pages 1955–1958, 1997.