



# Use of Acoustic Prior Information for Confidence Measure in ASR Applications

Erhan Mengusoglu, Christophe Ris

Faculté Polytechnique de Mons, TCTS Lab, Mons, Belgium

{mengus,r.is}@tcts.fpms.ac.be

## Abstract

In this paper, we propose a new acoustic confidence measure of ASR hypothesis and compare it to approaches proposed in the literature. This approach takes into account prior information on the acoustic model performance specific to each phoneme. The new method is tested on two types of recognition errors: the out-of-vocabulary words and the errors due to additive noise. We then propose an efficient way to interpret the raw confidence measure as a correctness prior probability.

## 1. Introduction

The use of acoustic confidence measures can be very useful for most Automatic Speech Recognition (ASR) applications. Indeed, it could be of great help to be able to predict whether a hypothesis provided by an ASR system is correct or not. For instance, high level dialogue systems can be significantly improved if we have a good idea of the recognition accuracy [3, 4], detection of out-of-vocabulary words is required for keyword spotting systems [11, 12]. Let's cite also the selection of reliable acoustic for unsupervised model adaptation [8, 14], alternative pruning methods in efficient search techniques [9], use of confidence measure in diagnostic tools to evaluate performance of ASR components [6] ...

Recognition errors can roughly be classified into two groups: a) the Out-Of-Vocabulary words (OOV) which occur when words that are not in the recognizer lexicon have been pronounced and b) recognition in mismatched conditions, that is when use conditions strongly differ from training conditions that is that acoustic models are not well suited anymore, this can be due to ambient noise, disfluencies, reverberation, ... The goal of acoustic confidence measure is to extract from the acoustic information only, an indicator of the confidence we can have in the word hypothesis. As we will see in the following, many approaches can be investigated, unfortunately few of them are well suited for both types of error (OOV and mismatched condition). In this paper, we will propose a new acoustic confidence measure and compare it to other approaches in the two specific conditions of OOV words and noisy speech signal. All the methods have been developed in the framework of hybrid HMM/ANN system which are well suited to confidence measure calculation as they provide local phone class posterior probability estimations. Finally, confidence measures need to be interpreted in order to decide whether a word is probably correct or incorrect. Ideally, the value should be interpreted as the prior probability that the word hypothesis is correct. We then propose a mapping of the raw confidence measure on a prob-like scale. Experiments have been carried out on PHONEBOOK, an American English isolated word database.

## 2. Definition of the Problem

The problem of confidence measure can be seen as a process of statistical hypothesis testing [5] in which we want to decide to accept or reject the hypothesis that the most probable sequence of words provided by the recognizer is correct. In this particular case, the acceptance region is delimited by a single threshold value. Therefore, a value of the test statistic that falls on one side of the threshold will result in the hypothesis being accepted, while a value falling on the other side will result in the hypothesis being rejected. Two types of error can occur in such a test: a) a *type I error* if the hypothesis is rejected when it is true, we will call it *false rejection error* (FR) in the following and b) a *type II error* if the hypothesis is accepted when it is false, we will call it *false acceptance error* (FA). We then define the unconditional classification error rate (CER) as the metric for the hypothesis test evaluation:

$$CER = \frac{N(\text{type I errors}) + N(\text{type II errors})}{N \text{ tested hypothesis}}$$

Of course, this metric is dependent on the global performance of the system. This is the reason why in the experiments we carried out, we have defined test conditions for which the word error rate is set to 50%. In this case, extreme decisions (accept/reject every hypothesis) will lead to a classification error rate of 50%. The confidence measure can therefore be used as the value used to perform this test statistic.

Note that this paper is concerned by acoustic confidence measure which is derived from the acoustic model only, no higher level information such as language model, semantic analysis, ... is investigated here. As stated before, two types of recognition errors must be detected: the out-of-vocabulary words and the mismatched test conditions. It is important to note that from the acoustic point of view these two kinds of error must be examined differently. Indeed, in case of mismatched conditions, we can expect the acoustic model to poorly classify acoustic data and provide rather smooth likelihoods with high confusion. On the other side, in case of out-of-vocabulary words, the model matches the acoustic data correctly but leads to sequences of phonetic units that are not covered by the lexicon so that part of the word hypothesis (some phonemes for instance) will not be matched the acoustic data. As we will see in the next section, some confidence measure are well suited for one type of error and not the other. For this reason, we defined two evaluation sets, one specific to out-of-vocabulary words by modifying the lexicon so that 50% of the set are not recognized anymore, the other one specific to mismatched condition by adding white noise to the half database.



### 3. Confidence Measures

All the confidence measures we have tested in this study are based on (possibly scaled) posterior probabilities provided by ANN [2]. As stated in [13], these posterior probabilities are well suited to confidence measure as they are independent on what has been uttered, so that no explicit normalization is required as in [11].

So, if we denote  $W = \{q_k^1, \dots, q_k^N\}$  the best state sequence as provided by the Viterbi decoding, the basic measure denoted PCM is defined as

$$PCM(W) = \frac{1}{N} \sum_{n=1}^N \log(P(q_k^n | X^n))$$

where  $P(q_k^n | X^n)$  is the posterior probability of being in state  $q_k$  for acoustic vector  $X^n$ , and  $N$  is the number of frames of the hypothesized word. Confidence measure proposed here after are basically different normalization procedures of the PCM. Note that some of these approaches can be combined.

#### 3.1. Use of acoustic prior information (PPCM)

A source of disparity between different phonemes comes from the acoustic model itself which intrinsically better matches some acoustics than others. This will lead to average posterior probabilities higher for certain phonemes than for others<sup>1</sup>. So the idea is to compensate for this effect by normalizing the posterior phoneme probability by the mean posterior probability of this phoneme when it is actually pronounced.

Practically during the training phase, we compute for each phoneme and according to a phonetic alignment, the average posterior probability provided by the ANN. This value gives an idea of the acoustic score we can expect for a given phoneme when it is correctly recognized. Normalizing the posterior probabilities by this value comes to give the same importance to each phoneme. In the following, this confidence measure will be denoted PPCM for *prior phone probability normalized posteriors*.

$$\bar{P}(q_k) = \frac{1}{T} \sum_{t=1}^T P(q_k^t | X^t)$$

Where  $X_t$  are the frames corresponding to the phoneme  $q_k$  as provided by a phonetic alignment.  $\bar{P}(q_k)$  are computed during the training phase.

$$PPCM(W) = \frac{1}{N} \sum_{n=1}^N \log(P(q_k^n | X^n) / \bar{P}(q_k))$$

where  $P(q_k^n | X^n)$  is the posterior probability of being in state  $q_k$  for feature vector  $X^n$ ,  $N$  is the number of frames of the current word.

#### 3.2. Relative posterior probability (RPCM)

This measure is computed by dividing each posterior probability of a frame by the best posterior probability for this frame, and normalizing over the length of the word. This comes to compare to the best acoustic score we can expect for a word hypothesis in the same time segment. As we will see in the results, this confidence measure is very efficient for OOV words

<sup>1</sup>we, of course, talk about phonemes matching the acoustics as given by a phoneme alignment

but degrades performance for mismatched conditions.

$$RPCM(W) = \frac{1}{N} \sum_{n=1}^N \log(P(q_k^n | X^n) / P(q_{best}^n | X^n))$$

where  $P(q_k^n | X^n)$  is the posterior probability of being in state  $q_k$  for feature vector  $X^n$ ,  $P(q_{best}^n | X^n)$  is the best posterior probability for current frame,  $N$  is the number of frame of the current word.

#### 3.3. Entropy (ECM)

The entropy is calculated for each frame and is independent of the optimal state sequence. Therefore, entropy should rather be seen as a measure of the acoustic model adequacy. The lower the entropy is, the better the model matches the acoustic data.

$$ECM(W) = -\frac{1}{N} \sum_{n=1}^N \sum_{k=1}^K p(q_k^n | X^n) \log(p(q_k^n | X^n))$$

where  $K$  is the number of phones and  $p(q_k^n | X^n)$  is the probability estimated by ANN for the current phone.

#### 3.4. Phone based normalization [1]

This normalization can be applied to any of the above described measures. It is computed by normalizing the posterior probabilities of a hypothesis word, first by the phoneme length, then over the word. This comes to give the same importance to each phoneme of the word whatever its length is. As we will see in the results, this normalization leads to very good performance. Indeed, due to their poor acoustic score, mismatched phonemes will be kept as short as possible by the Viterbi decoding. Without normalization, they would have very little influence on the global confidence measure.

$$PCM_{PN}(W) = \frac{1}{M} \sum_{m=1}^M \frac{1}{n_s - n_e} \sum_{n=n_s}^{n_e} \log(P(q_k^n | X^n))$$

where  $P(q_k^n | X^n)$  is the posterior probability of being in state  $q_k$  for feature vector  $X^n$ ,  $M$  is the number of phonemes in the current word,  $n_s$  and  $n_e$  are the beginning and ending time index of the current phoneme in the word.

Applied to other confidence measures, we obtain:

$$PPCM_{PN}(W) = \frac{1}{M} \sum_{m=1}^M \frac{1}{n_s - n_e} \sum_{n=n_s}^{n_e} \log\left(\frac{P(q_k^n | X^n)}{\bar{P}(q_k)}\right)$$

$$RPCM_{PN}(W) = \frac{1}{M} \sum_{m=1}^M \frac{1}{n_s - n_e} \sum_{n=n_s}^{n_e} \log\left(\frac{P(q_k^n | X^n)}{P(q_{best}^n | X^n)}\right)$$

$$ECM_{PN}(W) = \frac{1}{M} \sum_{m=1}^M \frac{1}{n_s - n_e} \sum_{n=n_s}^{n_e} \sum_{k=1}^K p(q_k^n | X^n) \log(p(q_k^n | X^n))$$

## 4. Experiments

Experiments have been carried out on the PHONEBOOK database [10], American English, telephone speech, isolated word database. To evaluate the efficiency of the confidence measures described above, we plot the classification error rate (CER) as a function of the word rejection rate (WRR) [1]. The word rejection rate is of course directly dependent on the decision threshold applied to confidence level.

Three tests were defined in such a way that the initial recognition accuracy was 50%.

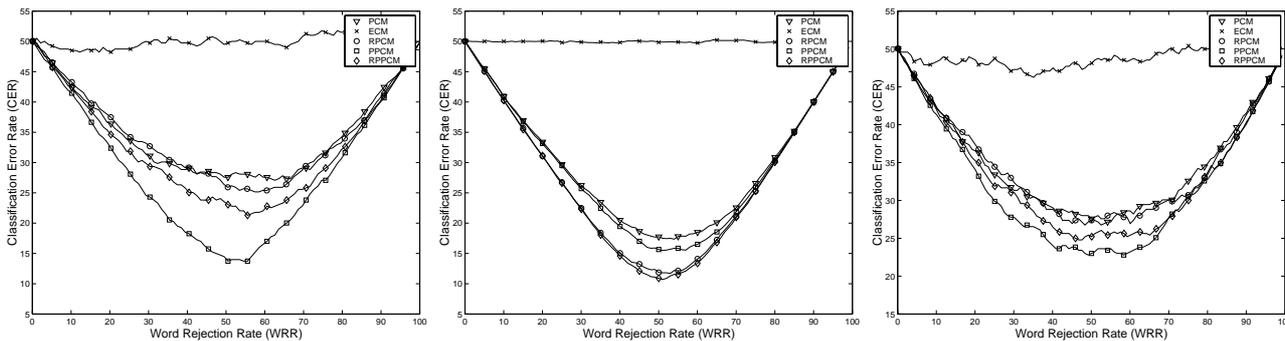


Figure 1: Performance curve of confidence measures using word level normalization. (Left) test for noise effect, (middle) test for OOV cases, (right) test for PHONEBOOK database test set.

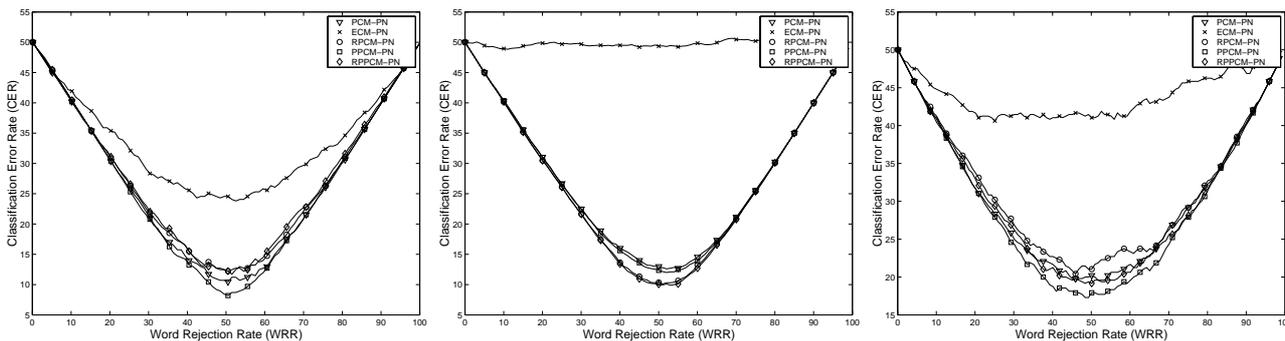


Figure 2: Performance curve of confidence measures using phone based normalization. (Left) test for noise effect, (middle) test for OOV cases, (right) test for Phonebook database test set.

1. Test for noise effect: 794 correctly recognized words selected from test set, noise is added to half of the utterances. Recognition error is caused by noise only. All noisy utterances lead to recognition errors. The noise is white gaussian noise with  $SNR = 13$ .
2. Test for OOV words: 2000 correctly recognized words are selected from test set. Confidence measures for first and second hypothesis in the N-best list is calculated. All the scores (for first and second hypothesis) are used to plot the performance curve. This test can be considered as OOV test, because in case of absence of the best word hypothesis in the vocabulary, recognizer will choose the second word as output. Confidence scores for second word of N-best list is calculated after re-alignment of phones for the acoustic data.
3. Test on all of the PHONEBOOK test sets: eight test sets from the PHONEBOOK database are combined as one test set of 6599 utterances. The number of incorrectly recognized words is 480. To obtain 50% accuracy, only 480 of correct word hypotheses were selected.

Figure 1 displays results for the confidence measures introduced in the section 3 on the three tests set. Figure 2 displays results with the phone-based normalization. As can be seen from the two figures, PPCM outperforms all the other methods in the case of test for noise effect. In OOV case, RPPCM which is the mixed version of RPCM and PPCM, is the best, that means normalization of the decoded phone probability by the best phone probability, improve the efficiency of the confidence measure in OOV cases. Indeed, this normalization comes to give the value

zero as an upper bound to the confidence score, so we can expect that correctly recognized words will have confidence score very close to zero while OOV words will have lower scores leading to better discrimination between correct and incorrect hypotheses. Unfortunately, this normalization degrades the performance in case of additive noise. Indeed, in this case, we can expect smoothed posteriors so that every hypotheses will have confidence score close to zero with no more discrimination. For the normal test set it is PPCM which is still the best. Figure 2 shows that the use of phone-based normalization systematically improves the efficiency of confidence measures. This normalization is very efficient but requires to keep phone level backtracking information during the decoding.

### 5. Decision Threshold

The confidence scores computed as above must be used to take the final decision of accepting or rejecting a hypothesis. Of course, we would like to have a value that can be directly interpretable so that the decision threshold can be easily fixed. A smart interpretation of such a value could be the probability of a word to be correct. Indeed, in such a case, a confidence score of 0.8 would mean that the word is statistically correctly recognized with 80% chance. During the training phase, we can build the histogram of word recognition rate according to their confidence score. We propose to match a sigmoid on this histogram. This sigmoid can be interpreted as a mapping function from the raw confidence score to probability-like values.

The procedure can be described as follows:



- for each confidence score <sup>2</sup>, compute the word recognition rate as the ratio of the number of correct word on the total number of word, that is for each score  $i$ :

$$score(i) = \frac{h_{correct}(i)}{h_{correct}(i) + h_{incorrect}(i)}$$

- The sigmoid to be matched is as follows:

$$y = \frac{1}{1 + e^{-\beta(x-\alpha)}}$$

- For  $(x - \alpha) = 0$  we find  $y = 0.5$ . This point can either be immediately taken from the histogram or preferably computed from the distributions of correct and incorrect words. Indeed, if we assume these distributions can be approximated by gaussians, we can find  $\alpha$  as the point where the probability of a word to be correct is equal to the probability to be incorrect.

$$\alpha = \frac{\mu_{correct} * \sigma_{incorrect} + \mu_{incorrect} * \sigma_{correct}}{\sigma_{correct} + \sigma_{incorrect}}$$

Where  $\mu$  and  $\sigma$  are the mean and standard deviation of the gaussian distributions.

- The last unknown parameter is  $\beta$  which can be approximated by *golden section search* algorithm [7]. This algorithm finds a polynomial interpolation for a function that minimize a criteria. In our case, we want to minimize the distance between the histogram points and the sigmoid.

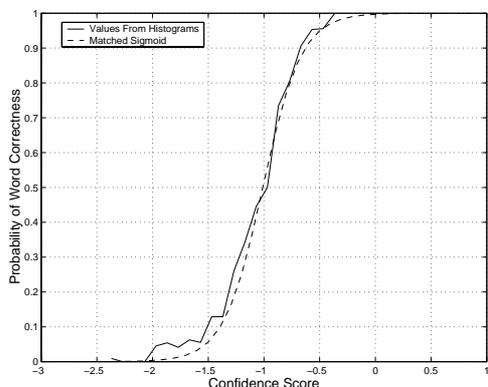


Figure 3: Mapping function (sigmoid) for the confidence measure  $RPPCM_{PN}$  calculated over the test set prepared for OOV test

## 6. Conclusion

We can conclude that, if there is a presence of additive noise in test data, use of acoustic prior information obtained from training data improve the efficiency of confidence measure. For OOV tests the best confidence measure was obtained after normalizing the decoded phone posterior probabilities by the best posterior probability of each frame. It was seen that using phone based normalization improve the efficiency for all methods. It is therefore interesting to note that ideally different confidence measure should be used for different types of error. Note also

<sup>2</sup>actually, confidence score interval of the histogram

that while the entropy could not be used as a confidence measure (totally inefficient in case of OOV words), it could possibly be used to identify portion of the signal where the model mismatches the acoustic data and therefore predict which kind of error we can expect and eventually which kind of confidence measure should be used.

## 7. References

- [1] H. Bourlard, G. Bernardis "Improving Posterior Based Confidence Measures in Hybrid HMM/ANN Speech Recognition Systems", proc. of ICSLP'98, pp. 318-321, 1998.
- [2] H. Bourlard and N. Morgan, "Connectionist Speech Recognition: A Hybrid Approach", Kluwer, 1994.
- [3] G. Bouwman and J. Hulstijn, "Dialogue Strategy Redesign with Reliability Measures", in proc. of the First International Conference on Language Resources and Evaluation, Granada, Spain, pp. 191-198, 1998.
- [4] G. Bouwman, J. Sturm and L. Boves, "Incorporating Confidence Measures in the Dutch Train Timetable Information System Developed in the Arise Project", in proc. of ICASPP'99, Phoenix, pp. 493-496, 1999.
- [5] T.M. Cover and J.A. Thomas, "Elements of Information Theory", Wiley Series in Telecommunications. Wiley-Interscience, New-York, 1991.
- [6] E. Eide, H. Gish, P. Jeanrenaud and A. Mielke, "Understanding and Improving Speech Recognition Performance Through the Use of Diagnostic Tools", in proc. of ICASSP'95, Detroit, pp. 221-224, 1995.
- [7] Forsythe, Malcolm and Moler, "Computer Methods for Mathematical Computations", Prentice-Hall, 1976.
- [8] T. Kemp and A. Waibel, "Unsupervised Training of a Speech Recognizer Using TV Broadcasts", in proc. of ICSLP'98, Sydney, pp. 2207-2210, 1998.
- [9] C.V. Neti, S. Roukos and E. Eide, "Word-Based Confidence Measures as a Guide for stack Search in Speech Recognition", in proc. of ICASPP'97, Munich, pp. 883-886, 1997.
- [10] J. Pitrelli, C. Fong, S. Wong, J. Spitz and H. Leung, "Phonebook: a Phonetically Rich Isolated Word Telephone Speech Database", proc. of ICASSP'95, Detroit, 1995.
- [11] R.C. Rose, "Word Spotting from Continuous Speech Utterances", in C.-H. Lee, F.K. Soong and K.K. Paliwal, editors, Automatic Speech and Speaker Recognition, pp. 303-329.
- [12] R.A. Sukkar, A.R. Setlur, M.G. Rahim and C.-H. Lee, "Utterance Verification of Keyword Strings Using Word-Based Minimum Verification Error Training", in proc. of ICASPP'96, Atlanta, pp. 518-521, 1996.
- [13] G. Williams and S. Renals, "Confidence Measures for Hybrid HMM/ANN Speech Recognition", in Proc. of Eurospeech'97, Rhodes, pp. 1955-1958, 1997.
- [14] T. Zeppenfeld, M. Finke, K. Ries, M. Westphal and A. Waibel, "Recognition of Conversational Telephone Speech Using the Janus Speech Engine", in proc. of ICASSP'97, Munich, pp. 1815-1818, 1997.