CONFIDENCE MEASURES FOR HMM-BASED SPEECH RECOGNITION

Daniel Willett, Andreas Worm, Christoph Neukirchen, Gerhard Rigoll

Department of Computer Science
Faculty of Electrical Engineering
Gerhard-Mercator-University Duisburg, Germany
e-mail: willett@fb9-ti.uni-duisburg.de

ABSTRACT

In this paper, we describe our work on the field of confidence measures for HMM-based speech recognition. Confidence measures are a means of estimating the recognition reliability for single words of the recognizer output. The possible applications of such measures are manifold. We present our experiments with well known approaches and propose some new ones. Particularly, we propose to combine the mere acoustical measures with language model-based ones for continuous speech recognition that involves a stochastic language model. This slightly improves the acoustical measures and preserves their advantage of being computationally very cheap. Experiments are carried out on a German isolated word recognition system and on continuous speech recognition systems for the Resource Management database and the Wall Street Journal WSJ0 task.

1. INTRODUCTION

Word-based confidence measures for speech recognition based on hidden Markov models (HMMs) have for some years now been an important research topic. While in the beginning the main interest was to detect out-of-vocabulary (OOV) words and recognition errors in isolated word recognition and to use them as diagnostic tool [2], several new applications of these measures arose in the more recent years. Such as adjusting the degree of unsupervised online adaptation or guiding the decoding search by scaling the language model with the previous words' confidence for a reduced word error rate [5].

For a word w within the boundaries a and b (utterance of w is hypothesized to have caused the acoustic observation $X_{ab} := (x_a, ..., x_b)$), a word-based confidence measure can be defined as function $c(w, a, b)$ with a usual domain of $[0, 1]$. The higher this function is for a hypothesized word and its hypothesized boundaries, the more confident one can be that it really has been uttered within the interval.

Often, the confidence measure of a word w, hypothesized for an acoustic observation sequence $X = (x_1, x_2, ..., x_n)$, is directly interpreted as the word's posterior probability $P(w|X)$. However, especially in continuous speech recognition, the confidence of a word whose position has been hypothesized as the interval $[a, b]$, is often understood as an estimate of the probability that the word starts somewhere around a and ends somewhere around b. In this case, the interpretation as mere posterior probability $P(w|X_{ab})$ is inadequate. Nevertheless, confidence measure and posterior word probability are strongly related.

This paper first presents the basic approaches for measuring the word-based recognition reliability and then outlines some of the special techniques that we use, discusses how to measure the quality of confidence measures and concludes with several experiments for isolated word and continuous speech recognition.

2. CONFIDENCE MEASURES BASED ON THE ACOUSTIC MODELS

Conventional HMM-based speech recognizers choose the word or word-sequences W with the highest posterior probability estimate $P(W|X)$ for an observed acoustic observation X. $P(W|X)$ is split up into $\frac{P(X|W)P(W)}{P(X)}$ where $P(X|W)$ is modeled by sequences of HMMs, $P(W)$ by stochastical language models or finite grammars and $P(X)$ is a scaling factor that can be omitted because it does not depend on W. In order to transform the HMM-based likelihoods into posterior probabilities that can be interpreted as confidence measures, Bayes' rule can be applied. For a word w hypothesized for the interval [a,b] this yields:

$$c(w, a, b) = P(w|X_{ab}) = \frac{P(X_{ab}|w)P(w)}{P(X_{ab})} \qquad (1)$$

Neglecting the word priors $P(w)$ (in case that we have a model for them, we will discuss how to incorporate this later) the confidence measure becomes the observation likelihood for the hypothesized word (the score estimated by the recognizer) weighted by an unconditioned observation likelihood.

2.1. Estimation of the unconditioned observation likelihood

The unconditioned likelihood $P(X_{ab})$ can be modeled in several ways. In ordinary continuous or tied continuous systems there is no dedicated model for this likelihood. However, with estimates $P(s)$ for the HMM states' priors, the unconditioned probabilistic distribution function $p(x)$ can be estimated as the weighted sum over all the S states' probabilistic distribution functions according to

$$p(x) \approx \sum_{i=1}^{S} p(x|s_i)P(s_i) \qquad (2)$$

The state priors $P(s)$ can be easily estimated on the training data or on the vocabulary. In [7], we showed how Eq. 2 simplifies for several kinds of systems and how to approximate it in others. Especially in discrete and tied continuous systems the additional computation of $p(x)$ turns out to be computationally very cheap. The unconditional likelihood of a sequence of observation vectors can then be estimated as

$$P(X_{ab}) \approx \left(\prod_{i=a}^{b} p(x_i) \right) t_{\text{av}}^{(b-a)} \qquad (3)$$

with t_{av} representing an average transition probability.

2.2. Phoneme-based measures

For the detection of OOV words, Asadi et al. [1] proposed to weight the hypothesized word's likelihood against the one of an unconstrained model sequence. The idea is that for OOV words there must be a better fitting sequence of phones p^* than the one found, that simply is not part of the dictionary. Hence, they used the quotient

$$c(w,a,b) := \frac{P(X_{ab}|w)}{P(X_{ab}|p^*)} \tag{4}$$

to decide, whether the hypothesized word w is correct or not. Young [10] extended this approach by applying additional prior probabilities for phone sequences estimated using a (tri)phone-trigram.

The likelihood $p(X_{ab}|p^*)$ of an arbitrary phone sequence can be estimated with ordinary speech recognizers using a so-called 2+-Model [1], a network that allows any sequence of at least two HMMs without any constraints on the sequence priors.

Observations showed that often it is only one or at most two phone models within an incorrectly hypothesized word, that produce an extremely bad likelihood score. They get squeezed inbetween phones, where they simply don't occur, in order to let at least the other models fit well. In order to cope with this observation, several approaches have been followed. In [9], each phone's confidence is estimated separately and the words' confidence is computed as the average over all the phones. This results in a normalization over the phone duration. It puts more emphasis on short phones than the previously presented measures. Another possibility to compensate for the described observation is to set the words' confidence to the minimum confidence among its phones.

$$c(w,a,b) := \min_{p \in w} c_{p_a p_b}(p) \tag{5}$$

In the equation above, p_a and p_b represent the phone boundaries of phone p hypothesized by the speech recognizer. Some experiments using this technique are presented in Section 5.

3. CONFIDENCE MEASURES FOR CONTINUOUS SPEECH RECOGNITION

In continuous speech recognition measuring word-based confidence mainly faces two additional problems compared to isolated word recognition. On the one hand, the hypothesized word boundaries are often incorrect. Ideal substitutions with correct word boundaries (but an incorrect word hypothesis) are rather rare. On the other hand, recognition is based not only on the Markov models' probabilistic distribution functions, but also on a language model that limits the possible word sequences (word-pair grammars, finite grammars) or estimates each word sequence's prior probability (stochastic language models, n-grams).

3.1. Measures based on the response of the recognizer

A very straightforward approach for measuring confidence in continuous speech recognition is to consider the speech recognizer as a black box, that we cannot or want not look inside, but to let it generate multiple hypothesis and to take the words' 'emission' probabilities as their confidence measure. The multiple hypothesis can be set up in various ways. In [3], Finke et al. proposed to perform multiple recognition procedures applying different scaling factors for weighing the language model based likelihoods against those based on the acoustic models. In [4], this was compared to the somewhat cheaper alternative of simply taking the N-best or lattice-output of only one recognition procedure. No severe differences to the scaling factor approach have been measured. The multiple recognizer outputs are usually stored in word-lattices [4] or N-best lists. This approach provides very useful confidence estimates. (Often, these estimates are even considered as reference for other approaches.) However, this method that is only based on the output of the speech recognizer is extremely expensive with respect to computational time needed for decoding. Thus, for real-time applications, such as dictation or dialogue-systems, methods which do not require additional decoding computations are desirable.

3.2. Model-based measures

Confidence measures that only need little additional computation consider the statistical models themselves. In the following, measures that use the acoustic hidden Markov models, the language model and those that aim to combine them are discussed separately.

3.2.1. Measures based on the acoustic models

Confidence measures for continuous speech recognition merely based on the acoustic hidden Markov models have been investigated in [9] for hybrid speech recognizers. It turned out that there is a noticeable degradation of these measures compared to the lattice-based ones described in the previous section. We experienced the same when applying the acoustical confidence measures described in Section 2 for hypothesized words of a continuous speech recognizer. Experiments can be found in Section 5. These measures neglect the language model and can hardly cope with the fact that often the acoustic match is fine but the hypothesized word boundaries are wrong. Therefore, in the following we describe our approach to improve the acoustic model-based measures by the incorporation of language model information.

3.2.2. Measures based on the language model

As a mere language model based confidence measure we propose to use an n-gram score weighted by the previous words' confidence. In the bigram case, this is formulated by

$$C_{\text{lm}}(w_2) = C_{\text{lm}}(w_1) p_{bi}(w_2|w_1) + (1 - C_{\text{lm}}(w_1)) p_{uni}(w_2) \tag{6}$$

for the hypothesized word w_2 succeeding the hypothesized word w_1. Another possible measure that we made use of is the product of the specific word's likelihood and its reverse likelihood (the language model likelihood of the succeeding word). In the bigram case, this yields

$$C_{\text{lm}}(w_2) = p(w_2|w_1) p(w_3|w_2) \tag{7}$$

for the hypothesized word w_2 between the words w_1 and w_3.

The results of experiments with these measures (Eqs. 6 and 7) without any acoustical information are given in Section 5. They show that the measures contain little but at least some information on whether a hypothesized word is correct or not and motivate the combination with an acoustical measure, described in the following paragraph.

3.2.3. A combination of acoustic and language model-based measures

Combining multiple features to result in only one metric can be accomplished in many ways. Neural Networks are a common tool for deriving such a function from training data. In [4, 6], several features were combined this way for measuring the word-based confidence. As we only want to combine two measures, an acoustic and a language model-based one, the product of these two is sufficient. Unfortunately though, just as in continuous speech recognition when combining language model and acoustic model likelihoods, a scaling factor s has to be introduced to cope with the different scale and quality of these measures. Thus, the combined confidence measure becomes

$$C(w) = C_{\text{ac}}(w) C_{\text{lm}}(w)^s \quad (8)$$

with $C_{\text{ac}}(w)$ representing one of the acoustical measures defined in Section 2.

As the measure based on the language model that was defined in Eq. 6 uses the previous words' confidence, the language model based measure C_{lm} can be refined by referring to the combined measure recursively so that in the bigram case Eq. 6 becomes

$$C_{\text{lm}}(w_2) = C(w_1)p_{bi}(w_2|w_1) + (1 - C(w_1))p_{uni}(w_2). \quad (9)$$

We measured some improvement using this combined measure over the acoustic ones, but there still is a remarkable gap to the measures derived from the recognizers lattice response (see Section 5). However, it has to be considered that the proposed measure can be computed very efficiently and thus allows real-time applications.

4. ASSESSMENT OF CONFIDENCE MEASURES

The quality of confidence measures largely depends on the task that they are set up for. A metric proposed by NIST is the relative entropy as described for example in [4]. It measures the amount of information that the confidence measure contains on the correctness of the hypothesized words. The disadvantage of this measure is the need for a transformation of the confidence measure to the interval $[0, 1]$ and to an average of the recognizer's correctness in order to be interpreted as posterior word probabilities. This affords the knowledge and incorporation of this correctness figure and allows an additional degree of freedom in scaling that directly effects the metric.

Therefore, we concentrated our experimental work and evaluation on the classic confidence measure application of detecting recognition errors. A meaningful diagram plots the amount of correctly hypothesized words that are falsely rejected (false alarms) against the amount of undetected errors depending on a specific rejection threshold. An interesting figure is the minimum percentage of falsely rated words with an ideal rejection threshold. In [6], this metric is called the classification error (CER). In the following experiments section we mainly use these types of evaluation. It has to be noticed, that the absolute value of the CER largely depends on the evaluated system's recognition performance. Hence, the CER may not be compared over different recognition systems.

5. EXPERIMENTS AND RESULTS

A first set of experiments was run on a 1000 word German isolated word recognition system. The acoustic models of this system have been trained on the Verbmobil spontaneous speech database. Figure 1 displays the ratio of false alarms and undetected errors for different detection thresholds for the various confidence measures presented in Section 2. It is obvious that the phone-based

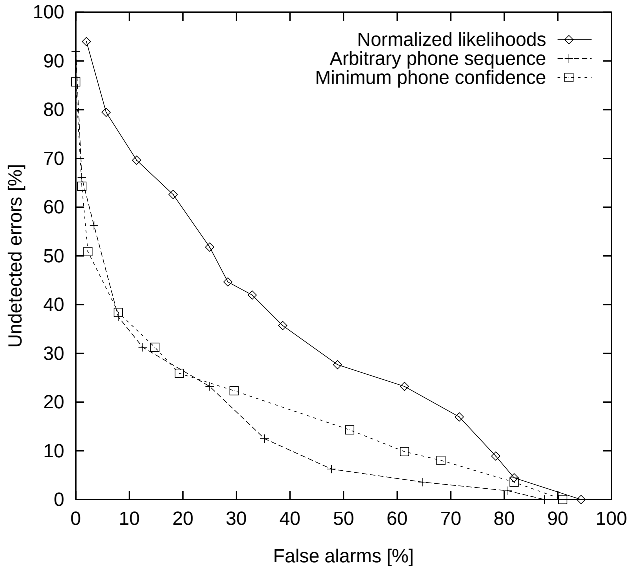

Figure 1. Confidence measures for isolated word recognition

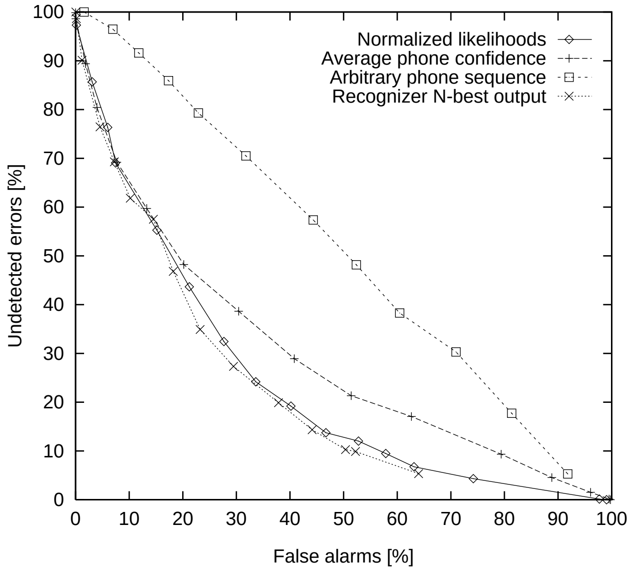

Figure 2. Continuous speech (RM) without grammar

measures (Eqs. 4 and 5) outperform the mere normalized likelihood (Eq. 1 with Eqs. 2 and 3). The strong degradation of the normalized likelihood measure is probably mainly due to the large amount of OOV words in the test set the results are based upon. Half of the words in the test set are OOV which causes an error rate of about 55%. The CER of the two phone-based measures is around 24%. Confidence measures applied on continuous speech recognition can be seen in Figures 2, 3 and 4. They were set up on tied continuous (tied according to [8]) context-dependent (triphone) recognition systems for the Resource Management task (2, 3) and the Wall Street Journal WSJ0 task (4). In Figure 2, recognition was performed without any constraints on the sentence priors. The recognition correctness of this system is at about 65%. It is interesting to see that the measure based on Eq. 4 performs just as bad as a random confidence measure would. It seems that hardly no phones are misclassified, but that it is mainly the segmentation into words that is often wrong. Furthermore, it is remarkable that in this case of having an unconstrained recognition grammar

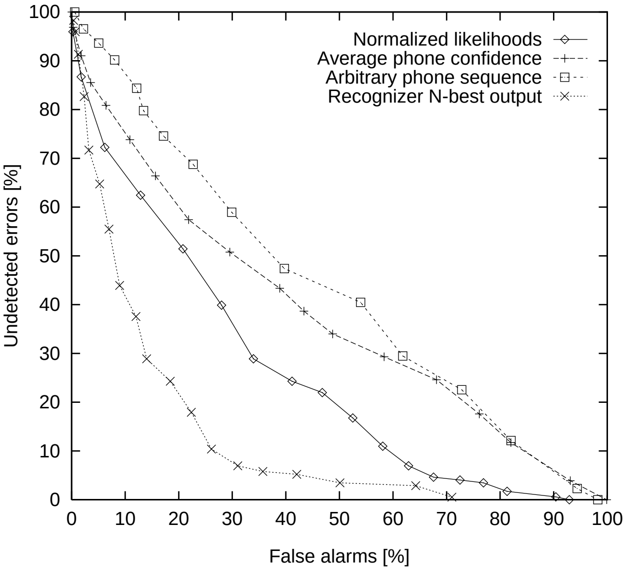

Figure 3. Continuous speech (RM) with word-pair grammar

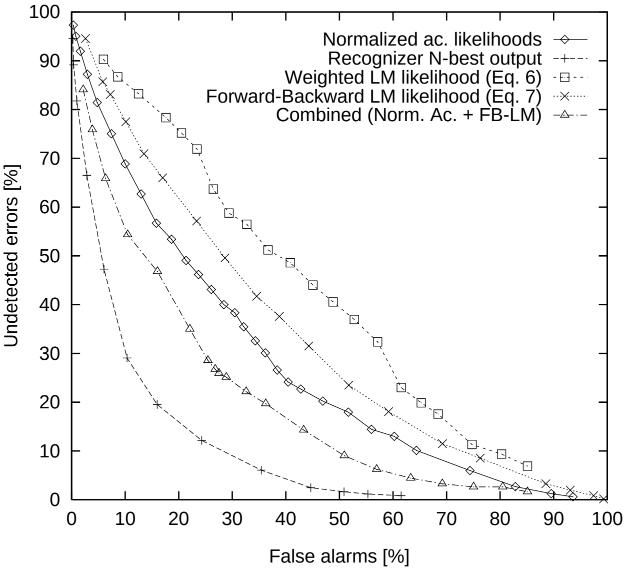

Figure 4. Confidence measures on the WSJ0 database

the normalized likelihood measure nearly achieves the same performance as the measure that is based on the multiple recognizer outputs. The CER of these two measures is at about 27%.

Figure 3 displays the same measures for a recognition system that is enhanced by the standard RM word-pair grammar. The system achieves a correctness of 93% (the test set from September '92 was used). It can be noticed that the quality of the diverse measures strongly differs from the no-grammar case. As expected, the N-best list based measure now outperforms the other measures by far (CER $\approx$ 6%). It is the only measure that (indirectly) takes both statistical models (acoustic HMMs and language model) into account. According to the observation in the no-grammar case the best acoustic confidence measure is the mere normalized likelihood. Phoneme-based measures only seem to be of good use in isolated word recognition.

The combination of language model and acoustic model-based measures, as proposed in Section 4, was evaluated on the WSJ0 database. Figure 4 shows the chart. Is is obvious that the language model-based measures are weak but hold some information at least. The Forward-Backward measure of Eq. 7 outperforms the weighted likelihood of Eq. 6. When combined with an acoustical measure, it remarkably improves this measure (CER $\approx$ 9%). However, the measure based on multiple recognizer outputs (CER $\approx$ 8%) cannot be outperformed.

6. CONCLUSION

The experiments have shown that confidence measures that only rely on a portion of the statistical models will always degrade against others that involve all of them. While for isolated word recognition phone-based measures give best results, in continuous speech recognition we measured the best performance with a simply weighted likelihood measure enhanced by the language model confidence measure. However, this still degrades strongly against the lattice-based measure. Nevertheless, it is lots cheaper to compute and thus useful even for real-time applications.

7. REFERENCES

1. A. Asadi, R. Schwartz, J. Maakhoul: "Automatic Detection of New Words in a Large-Vocabulary Continuous Speech Recognition System", ICASSP'90, pp. 125–128.

2. E. Eide, H. Gish, P. Jeanrenaud, A. Mielke: "Understanding and Improving Speech Recognition Performance through the use of Diagnostic Tools", ICASSP'95, pp. 221–224.

3. M. Finke, T. Zeppenfeld, M. Maier, L. Mayfield, K. Ries, P. Zhan, J. Lafferty, A. Waibel: "Switchboard April 1996 Evaluation Report", DARPA 1996.

4. T. Kemp, T. Schaaf: "Confidence Measures for Spontaneous Speech Recognition", ICASSP'97, Munich, pp. 875–878.

5. C. V. Neti, S. Roukos, E. Eide: "Word-based Confidence Measures as a Guide for Stack Search in Speech Recognition", ICASSP'96, Munich, pp. 883–886.

6. M. Weintraub, F. Beaufays, Z. Rivlin, Y. Konig, A. Stolcke: "Neural-Network Based Measures of Confidence for Word Recognition", ICASSP'96, Munich, pp. 887–890.

7. D. Willett, C. Neukirchen, G. Rigoll: "Efficient Search with Posterior Probability Estimates in HMM-based Speech Recognition", ICASSP'98, Seattle, pp. 821–824.

8. D. Willett, G. Rigoll: "A New Approach to Generalized Mixture Tying for Continuous HMM-Based Speech Recognition", EUROSPEECH '97, Rhodes, pp. 1175–1178.

9. G. Williams, S. Renals: "Confidence Measures for Hybrid HMM/ANN Speech Recognition", EUROSPEECH '97, Rhodes, pp. 1955–1958.

10. Sheryl R. Young: "Detecting Misrecognitions and Out-Of-Vocabulary Words", ICASSP'94, Adelaide, pp. II 21–24.