Confidence Measures for Speech Recognition: A Survey

Hui Jiang

Technical Report CS-2003-06

May 30, 2003

Department of Computer Science

4700 Keele Street North York, Ontario M3J 1P3 Canada

Confidence Measures for Speech Recognition: A Survey

Hui Jiang

Department of Computer Science, York University,

4700 Keele Street, Toronto, Ontario M3J 1P3, CANADA

Correspondence:

Prof. Hui Jiang, Department of Computer Science, York University,

4700 Keele Street, Toronto, Ontario M3J 1P3, CANADA

Phone: (416)736-2100 x33346 (Hui Jiang) **Fax:** (416)736-5872 (Hui Jiang)

Email: hj@cs.yorku.ca

Abstract

In speech recognition, confidence measures (CM) are used to evaluate reliability of recognition results. A good confidence measure can largely benefit speech recognition systems in many practical applications. In this survey, I summarize most research works related to confidence measures which have been done during the past 10-12 years. I will present all these approaches as three major categories, namely CM as a combination of predictor features, CM as a posterior probability, and CM as utterance verification. Then, I also introduce some recent advances in the area. Moreover, I will discuss capabilities and limitations of the current CM techniques and generally comment on today's CM approaches. Based on the discussion, I will conclude the paper with some clues for future works.

1 Introduction

Automatic speech recognition (ASR) has achieved some substantial successes in past few decades mostly attributing to two prevalent technologies in the field, namely hidden Markov modeling (HMM) of speech signals and efficient dynamic programming search (also known as *decoding*) techniques for very-large-scale networks. Today, in many aspects, it has become a standard routine to build a state-of-the-art speech recognition system for any particular task if sufficient training data is provided for the target domain. However, when we migrate speech recognition systems from laboratory demonstrations to real-world applications, even the best ASR systems available today still encounter some serious difficulties. First of all, system performance usually dramatically degrades in the real fields because of ambient noises, speaker variations, channel distortions and many other mismatches. How to maintain and/or improve ASR performance in real-field conditions has been extensively studied in speech community under the topic of *robust speech recognition*. Many good tutorial and overview papers, such as Juang (1991), Gong (1995), Lee (1998) and many others, can be easily found in the literature with regard to this topic. Secondly, since every speech recognizer inevitably will make some mistakes during recognition, outputs from any ASR system are always fraught with a variety of errors. Thus, in any real-world application, it is extremely important to be able to make an appropriate and reliable judgement based on the error-prone ASR results. This requires the system to automatically assess reliability or probability of correctness for every decision made by ASR systems. Nowaday, to certain degree, the capability to evaluate reliability of speech recognition results has been regarded as a crucial technique to increase usefulness and "intelligence" of an ASR system in many practical applications. In this area, researchers have proposed to compute a score (preferably between 0 and 1), called *Confidence Measure* (CM), to indicate reliability of any recognition decision made by ASR systems. For example, a CM can be computed for every recognized word to indicate how likely it was correctly recognized or for an utterance to indicate how much we can trust the results for the utterance as a whole. Despite a large amount of research efforts in the past, we still believe that robust speech recognition and confidence measure will remain as two most active and influential research topics in speech community for a foreseeable future. Due to importance of CM in ASR systems, it has attracted considerable research attentions from most major speech research groups all over the world and an excessive amount of research works have been reported in the past decade. But, unlike robust speech recognition, so far we have not seen too many overview papers in the literature to survey this active topic. This largely motivates me to write a comprehensive survey to summarize the CM-related research works reported mostly in the past 10-12 years. In the survey, I will mainly highlight the major progresses we have achieved in the CM area during the past decade. And I will stress some promising CM computation approaches which are theoretically sound and experimentally superior, and also discuss their capabilities and limitations. Finally, I will present some comparative discussions with respect to all reported CM computation methods and conclude the paper with some clues for possible future works from my personal perspective. Throughout the paper, I will attempt to present the CM techniques from a fairly high level and avoid technical and experimental details as much as possible, for which readers may wish to refer to the original papers. At the end of this paper, I also compose a comprehensive list of reference papers for readers' convenience. To my best knowledge, Lee (2001) seems to be the only CM-related overview paper which gives some good tutorials on statistical nature of confidence measure problems and also enumerate many potential CM applications for ASR.

First of all, we can backtrack some early research works on confidence measure (CM) to non-keyword rejection in word-spotting systems which were proposed to handle unconstrained speech inputs, such as Wilpon et al (1990), Mathan & Miclet (1991), Chigier (1992), Rose (1992), Sukkar & Wilpon (1993), etc. In these works, they first adopted the so-called *garbage* or *sink* models to explicitly model non-keywords, extraneous speech and background noises in unconstrained input utterances, with which key-word spotting systems first recognize speech inputs to detect all embedded keywords as well as other speech segments corresponding to non-keywords or noises. Besides all of these, they all noticed a need to build additional rejection module to effectively distinguish non-keywords from the detected keywords to reduce false alarms in non-keyword rejection. Apparently, the rejection module can be viewed as a stage to investigate reliability or confidence measures for the decisions made by word-spotters. Secondly, other early CM-related works lie in automatic detection of new words (out of the current lexicon) in a large vocabulary speech recognition, such as Asadi et al (1990), Young & Ward (1993) and Young (1994), etc. In addition to modeling out-of-vocabulary(OOV) words with a (or a set of) generic hidden Markov model, Young & Ward (1993) proposed to use word score normalization to detect mis-recognition and out-of-vocabulary words for continuous speech recognition. Young (1994) first elucidated how to use posterior probability as a confidence measure for speech recognition, where she used the acoustic score normalization based on a separate all-phone recognition to approximate such posterior probability. Young (1994) also tried to combine the normalized acoustic score with other high-level knowledge sources, e.g. semantic, pragmatic and discourse analysis, to improve quality of confidence measures. Thirdly, Sukkar & Wilpon (1993), Sukkar (1994) and others realized that confidence measures for ASR become extremely important when speech recognition technology is applied to any practical applications or services for end-users. Some intensive research efforts at the former AT&T Bell Labs resulted in lots of fruitful works related to confidence measure for ASR, but under a different name, namely *utterance verification*. Rose, Juang & Lee (1995a) first formally cast the confidence measure problem in speech recognition as a statistical hypothesis testing problem as in classical statistics and proposed to use likelihood ratio testing (LRT) to solve the problem. Then, the LRT-based formulation has become the basic theoretical foundation for the follow-up works, e.g. Sukkar & Lee (1996), Rahim, Lee & Juang (1997a), Rahim & Lee (1997b, 1997c), Sukkar et al (1997), etc. In these works, they discovered that some discriminative training techniques, such as minimum verification error (MVE) estimation, can significantly improve performance of modeling both the *null* and *alternative* hypotheses in utterance verification. More recently, a tremendous amount of research activities have been carried out in this area to seek for some reliable confidence measures for ASR, mainly driven by an increasing number of dialogue applications. Based on confidence measures, spoken language systems will be able to handle error-prone ASR outputs more intelligently in those post-recognition modules, such as language understanding and dialogue management. Some representative works include Eide et al (1995), Cox & Rose (1996), Chase (1997), Gillick et al (1997), Neti et al (1997), Kemp & Schaaf (1997), Schaaf & Kemp (1997), Weintraub et al (1997), Rueber (1997), Jiang & Huang (1998), Willett et al (1998), Siu & Gish (1999), Benitez et al (2000), Wessel et al. (2000), Kamppari & Hazen (2000), Kamppari & Hazen (2000), Jiang et al (2001b, 2002, 2003), San-Segundo (2001), Zhang & Rudnicky (2001) and many others.

Generally speaking, all methods proposed for computing confidence measures (CM's) in speech recognition can be roughly classified into three major categories. Firstly, a large portion of works aims to compute confidence measures based on a combination of the so-called *predictor features*, which are collected during decoding procedure and may include acoustic as well as language information about recognition decisions. Then all predictor features are combined in a certain way to generate a single score to indicate correctness of the recognition decision. We will briefly summarize these methods in section 2. Secondly, it is well known that the *posterior* probability in the standard maximum *a posterior* (MAP) decision rule is a good candidate as CM for speech recognition since it is an absolute measure of how well the decision is. However, it is very hard to estimate the posterior probability in a precise manner due to its normalization term in the denominator. In practice, many different approaches have been proposed to approximate it, ranging from simple *filler*-based methods to complex word-graph-based approaches. We will introduce these methods in section 3. Next, as already mentioned above, under the name of *utterance verification* (UV), lots of works have been conducted to verify the claimed content of a spoken utterance. The content can be hypothesized by a speech recognizer or keyword detector or human transcriber. Under the framework of utterance verification (UV), the CM problem can be formulated as a statistical hypothesis testing. In section 4, we will briefly present all proposed methods in this category, ranging from the LRT-based non-Bayesian approach (based on Neyman-Pearson Lemma) to the Bayes-Factors-based Bayesian approach. We also introduce how to use some discriminative training methods to improve modeling in UV. In the remainder of the paper, in section 5, I will first mention several samples of very recent research advances regarding CM computation, including an in-search data selection for improving verification models in UV, and a novel idea to compute confidence measures based on neighborhood information in model space, and how to annotate confidence measures based on semantic information measured by latent semantic analysis (LDA). In section 6, I will discuss about performance comparison issues among all different CM methods and focus on capabilities and limitations of the current CM techniques in a variety of potential applications. Finally, I will make some general comments on CM methods in ASR and conclude this paper with some clues for future works.

2 CM as Combination of Predictor Features

In the literature, a very large portion of CM-related works aims to search for a predictor feature (or a set of features) which is informative to distinguish correctly recognized results from other possible recognition errors. Any feature can be called a predictor if its probabilistic distribution (e.g., p.d.f.) of correctly recognized words is clearly distinct from that of mis-recognized words. Usually, the predictor features may have to be collected within the recognition process at levels of acoustics, language model, syntax, and semantics. Some common predictor features reported in the literature may include:

- *Pure normalized likelihood score* related: acoustic score per frame.
- *N-Best* related: count in the N-best list, N-Best homogeneity score (the weighted ratio of all paths passing through the hypothesized word in N-best list), top N recognition scores, top N-1 difference in adjacently ranked recognition scores, etc.
- *acoustic stability*: a number of alternative hypotheses are generated based on different language model weights in decoding and acoustic stability of any given word is defined as the number of times the word occurs in the list divided by the number of alternatives in the list.
- *hypothesis density*: the number of alternative arcs spanning the time segment of the recognized word in word graph.
- *Duration* related: HMM state duration, phoneme duration, word duration.
- *Language model (LM)* related: LM score, LM back-off behavior, etc.

- *Posterior probability* related: see section 3 for details.
- *Log-likelihood-ratio* related: see section 4 for details.

For more predictor features, please refer Cox & Rose (1996), Schaaf & Kemp (1997), Chase (1997), Benitez et al (2000), San-Segundo et al (2001), etc. An ideal predictor feature should provide strong information to separate the correctly recognized words from other misrecognitions and the distribution overlap between the two classes should be minor. However, none of the above predictor features is ideal in this sense. As reported in many papers, the overlap is actually quite large even for the best predictor feature. Therefore, some people attempt to combine several different predictor features for a better performance. Many different combinational models have been reported in literature, including linear discriminant function (Sukkar, 1994, Sukkar & Lee, 1996), generalized linear model (Gillick et al, 1997, Siu & Gish, 1999), single or mixture Gaussian classifier (Chigier, 1992), neural networks (Mathan & Miclet, 1991, Weintraub et al., 1997, San-Segundo et al 2001), decision tree (Eide et al, 1995, Neti et al, 1997), support vector machine (Zhang et al, 2001), boosting (Moreno et al, 2001) and others. In most cases, parameters of combinational models are estimated from some discriminative training procedures based on some criteria such as cross-Entropy, classification error rate (see Weintraub et al., 1997 for more details about this).

A combination approach can improve the overall performance only when all individual components are statistically independent. Obviously, this is not the case for the above predictor features. It has been observed in many experiments that all these predictor features are highly correlated[1]. Usually the combination methods can not significantly improve over the best predictor feature. So far, we have not seen any compelling results by combining various predictor features in confidence measure estimation.

3 CM as Posterior Probability

It is well known that the conventional ASR algorithms are usually formulated as a pattern classification problem using the *maximum a posterior* (MAP) decision rule to find the most

[1]Refer to final remarks in section 7 for a possible explanation for this.

likely sequence of words $\hat{W}$ which achieves the maximum posterior probability $p(W|X)$ given any acoustic observation X, i.e.,

$$\hat{W} = \arg\max_{W\in\Sigma} p(W|X) = \arg\max_{W\in\Sigma} \frac{p(X|W)\cdot p(W)}{p(X)} = \arg\max_{W\in\Sigma} p(X|W)\cdot p(W) \quad (1)$$

where Σ denotes the set of all permissible sentences, $p(W)$ is the probability of W evaluated with a language model, $p(X)$ is the probability of observing X, and $p(X|W)$ is the probability of observing X by assuming that W is the underlying word sequence for X. In theory, the posterior probability $p(W|X)$ is a good confidence measure for the recognition decision that X is recognized as W. However, as shown in the above eq.(1), most practical ASR systems simply ignore the term $p(X)$ in decision-making because it is constant across different words W. This explains why the raw ASR scores are inadequate as confidence measures to judge recognition reliability. However, after being normalized by $p(X)$, the posterior probability $p(W|X)$ can serve as a good confidence measure since it represents the absolute quantitative measure of the match between X and W. In theory, we should compute $p(X)$ as follows:

$$p(X) = \sum_{H} p(X,H) = \sum_{H} p(H)\cdot p(X|H) \quad (2)$$

where H denotes any a hypothesis for X, and the above summation must be done over all possible hypotheses for X, including all combinations of words, phonemes, noises and other events. Obviously, without any further constraint, it is impossible to enumerate and model all these hypotheses so that it is extremely difficult to estimate $p(X)$ in a precise manner. In practice, we have to either impose certain assumptions or adopt some approximate methods when estimating $p(X)$ for the posterior probability.

In the first category, it includes the so-called *filler*-based methods which try to calculate $p(X)$ from a set of general filler or background models, i.e., all-phone recognition (Young, 1994), catch-all model (Kamppari & Hazen, 2000), the highest score in recognizing the word from decoder (Cox & Rose, 1996), etc. These approaches are very straightforward and usually can achieve an reasonable performance in many cases. In another category, there are the so-called *lattice*-based methods which attempts to calculate $p(X)$, then the posterior probability $p(W|X)$ in turn, from a word lattice or graph based on the forward-backward algorithm, such as Kemp & Schaaf (1997) and Wessel et al (1998, 1999, 2000, 2001). Usually, one word lattice or graph is generated by the ASR decoder for every utterance. Then the posterior probability of each recognized word or the entire hypothesized sentence can be calculated based on the word-graph from an additional post-processing stage. Since word graph is a compact and fairly accurate representation of all alternative competing hypotheses of the recognition result which usually dominate the summation when computing $p(X)$ over a variety of hypotheses in eq.(2), the posterior probability calculated from a word graph can approximate the true $p(W|X)$ pretty well. Therefore, the resultant confidence measures generally achieve better performance than all other CM's mentioned in the above. However, generating word graphs and scoring word-graphs for posterior probabilities are relatively complicated and quite demanding in computation, especially in large vocabulary ASR systems. Thus, for the sake of simplicity, an N-Best list can also be used in place of word graph for this purpose, such as Rueber (1997), Wessel et al (2000), etc. Due to its superior performance as a CM for ASR, in the following, I will review some details about how to compute posterior probabilities from a word graph as originally reported by Wessel et al. (2001).

3.1 Word Graph Notations

Usually the ASR decoder generates a word graph $\mathcal{X}$ for each utterance X. Here, the word graph is represented as a directed, acyclic, weighted graph. All its nodes represent discrete points in time. Each arc is labeled with three variables, i.e. $[w]_s^e$, where w is the hypothesized word attached to the arc, and s and e denote the starting and ending time instances of the arc. Also, each arc is associated with a weight, $\mathcal{B}(w)_s^e$, which is actually acoustic score of generating acoustic feature vectors from time s to e from the HMM of word w. In every word graph, there are two special notes: one is called START note which corresponds to the beginning of the utterance and one END note for the end of the utterance. Any path from START node to END note is called a complete path which represents a sentence (a sequence of words) hypothesis for the underlying utterance. Let's assume a complete path in word graph $\mathcal{X}$ of an utterance X, which consists of n different arcs as $C = \{ \ [w_1]_{s_1}^{e_1}, [w_2]_{s_2}^{e_2}, \cdots, [w_n]_{s_n}^{e_n} \ \}$. Obviously, it is straightforward to compute the probability of this complete path given the word graph $\mathcal{X}$ as follows:

$$p(C|\mathcal{X}) = \prod_{i=1}^{n} \mathcal{B}(w_i)_{s_i}^{e_i} \cdot p(w_i \mid h_i) \tag{3}$$

where h_i denotes the history of word w_i and $p(w_i \mid h_i)$ is the language model score computed with n-gram language models.

3.2 Posterior Probability of an Arc

Based on the above notations, it is easy to compute the posterior probability of any arc $a = [w]_s^e$ given the word graph $\mathcal{X}$, namely $p(a|\mathcal{X})$.[2] Normally, $p(a|\mathcal{X})$ is calculated as a ratio between the total probability of all complete paths passing through the arc a to that of all complete paths in $\mathcal{X}$, i.e.,

$$p(a \mid \mathcal{X}) = \frac{\sum_{C\in\mathcal{X}\ a\supset C} \quad p(C|\mathcal{X})}{\sum_{C\in\mathcal{X}} \quad p(C|\mathcal{X})} \tag{4}$$

where $C \in \mathcal{X}$ denotes C is a complete path in word graph $\mathcal{X}$ and $a \supset C$ denotes that the complete path C passes through the arc a. The posterior probability $p(a \mid \mathcal{X})$ can be efficiently computed based on a forward-backward algorithm. A forward probability $\alpha_s(a)$ is recursively computed from the start note of the arc a backward until the START note of the word graph as:

$$\alpha_s(a) = \mathcal{B}(w)_s^e \cdot \sum_{a'} \alpha_{s'}(a') \cdot p(w|h') \tag{5}$$

where the summation is conducted for all arcs a' (s' is start time of a') merging into the start node of a and h' is word history of w in language model computation. Analogously, a backward probability $\beta_e(a)$ is computed from the end node of a forward until END node of the word graph as:

$$\beta_e(a) = \sum_{a''} \beta_{e''}(a'') \cdot \mathcal{B}(w'')_{s''}^{e''} \cdot p(w''|h'') \tag{6}$$

where the summation is conducted over all arcs a'' (e'' is ending time of a'' and w'' is word id in a'') leaving the end node of a and h'' is word history of w'' when calculating language

[2] Note that the posterior probability of an arc given the word graph $\mathcal{X}$ differs from the original posterior probability given the utterance X.

model score. Obviously, the numerator in eq.(4) can be computed as the product of $\alpha_s(a)$ and $\beta_e(a)$. And the denominator in eq.(4) can be recursively computed as forward probability $\alpha_s(a)$ in eq.(5) staring from START node until END node of the word graph or backward probability $\beta_e(a)$ in eq.(6) from END node backward to START node of word graph.

3.3 Posterior Probability of a Recognized Word

We can directly use the posterior probability, $p(a \mid \mathcal{X})$, of the arc, $a = [w]_s^e$, as confidence measure for the recognized word w. But it has been shown that it does not perform very well as a confidence measure for w. We know that except the arc $a = [w]_s^e$, there are usually lots of other arcs in word graph that have the same word id w but slightly different starting time s and ending time e. It will underestimate confidence measure of w if we only count $p(a \mid \mathcal{X})$ for w. Thus, it is very important to take into account other arcs which have the same word id w but slightly different s and e. Wessel et al (2001) proposes three different ways to solve this problem. In the first method, called C_{sec}, when calculating confidence measure for the word w in an arc $a = [w]_s^e$, we sum over all arcs in word graph which have the same word id w and intersect with the current arc $a = [w]_s^e$ in time domain. In the second method, called C_{med}, we only accumulate posterior probability for all arcs with the same word id which intersects the median time frame of the arc under consideration. In the third approach, called C_{max}, we determine a best-case probability for word w in an arc $a = [w]_s^e$. We accumulate posterior probability for all arcs (with the same word id) which not only intersect the median time frame but also all other time frames between s and e, and then choose the maximum one from these sums as the confidence measure for the word w in the underlying arc. Based on Wessel et al (2001), the third method, namely C_{max}, yields the best performance.

There are many other implementation issues to consider when computing posterior probability in word graph, e.g., scaling of probabilities in summation, elimination of redundant silence edges, etc. Readers are referred to Wessel et al (2001) for more details.

4 CM as Utterance Verification

Mainly motivated by speaker verification problem, Rose, Juang & Lee (1995a), Sukkar & Lee (1996), Rahim, Lee & Juang (1997a) have proposed to tackle confidence measure problems from a different perspective. Under the framework of utterance verification (UV), the confidence measure problem in ASR is formulated as a statistical hypothesis testing problem. For a given speech segment X, assume that an ASR system recognizes it as word W which is represented by an HMM λ_W. Utterance verification is a post-processing stage to examine the reliability of the hypothesized recognition results. Under the framework of UV, we first propose two complementary hypotheses, namely the *null* hypothesis H_0 and the alternative hypothesis H_1 as follows:

$$\begin{aligned} H_0 &: \quad X \text{ is correctly recognized and truly comes from model } \lambda_W \\ H_1 &: \quad X \text{ is wrongly classified and is NOT from model } \lambda_W \end{aligned} \tag{7}$$

Then we test H_0 against H_1 to determine whether we should accept the recognition result or reject it. According to Neyman-Pearson Lemma, under some conditions, the optimal solution to the above testing is based on a likelihood ratio testing (LRT), i.e.,

$$LRT = \frac{p(X|H_0)}{p(X|X_1)} \underset{H_1}{\overset{H_0}{\gtrless}} \tau. \tag{8}$$

The LRT-based utterance verification provides a good theoretical formulation to address confidence measure problem in ASR. As pointed out by Lee (2001), the above LRT can be transformed to a confidence measure based on a monotonic one-to-one mapping function. The major difficulty with LRT is how to model the alternative hypothesis which usually represents a very complex and composite event, where the true distribution of data is unknown. In practice, as in Rose, Juang & Lee (1995a), Sukkar & Lee (1996), Rahim, Lee & Juang (1997a), the same HMM structure is adopted to model the alternative hypothesis, which can be a general background model, or hypothesis-specific *anti-model*, or a set of competing models, or a combination of all the above. In these works, a variety of training methods have been used to estimate HMM's for the alternative hypothesis. It is generally agreed that a discriminative training procedure plays a crucial role in improving modeling performance for the alternative hypothesis. In Sukkar & Lee (1996), a GPD (generalized probabilistic descent) based discriminative training procedure is used to estimate parameters of a linear discriminant function based on a criterion to minimize sub-word level verification error counts represented by a sigmoid function. In Rahim, Lee & Juang (1997a), it is found that the minimum classification error (MCE) training, which is originally proposed to reduce recognition errors, can contribute to improving performance of UV. In Rahim & Lee (1997b) and Sukkar et al (1997), a GPD-based training algorithm is proposed to achieve minimum verification error (MVE) estimation for utterance verification with respect to optimizing verification HMM parameters. In MVE, the string-level verification errors are approximated by using a sigmoid function embedded with a mis-verification function, which actually is negative log-likelihood ratio used in verification. Then the total empirical verification errors can be minimized over all training data by optimizing the verification HMM parameters corresponding to the both null and alternative hypotheses. The optimization can be iteratively achieved by using a GPD algorithm. Experiments clearly show all these discriminative training methods can largely improve performance of the LRT-based utterance verification.

Alternatively, if we consider the above UV problem from a Bayesian viewpoint, the final solution ends up with calculating and evaluating the so-called *Bayes factors* as in Jiang & Deng (2001a). Bayes factors has its solid foundation from Bayesian theory. Given the speech data X along with the above two hypotheses H_0 and H_1, Bayes factors is computed as:

$$BF = \frac{\hat{p}(X \mid H_0)}{\hat{p}(X \mid H_1)} = \frac{\int f(X \mid \lambda_0, H_0) \cdot p(\lambda_0 \mid H_0) \, \mathrm{d}\lambda_0}{\int f(X \mid \lambda_1, H_1) \cdot p(\lambda_1 \mid H_1) \, \mathrm{d}\lambda_1} \tag{9}$$

where, for $k = 0, 1$, λ_k is the model parameter under H_k, $p(\lambda_k \mid H_k)$ is its prior density, and $f(X \mid \lambda_k, H_k)$ is the likelihood function of λ_k under H_k.

Bayes factors offers a way to evaluate evidence in favor of the *null* hypothesis H_0 because Bayes factors is the ratio of the posterior odds of H_0 to its prior odds, regardless of the value of the prior odds.[3] Therefore, Bayes factors can be used to compare with a threshold, just like the likelihood ratio in Neyman-Pearson lemma, to make a decision with regard to H_0. In other words, if $BF > \tau$, where τ is a pre-set critical threshold,

[3]Any probability can be converted to the odds scale, i.e., odds=probability/(1-probability). Thus, $\frac{Pr(H_0|y)}{Pr(H_1|y)}$ is called the posterior odds in favor of H_0, and $\frac{Pr(H_0)}{Pr(H_1)}$ is prior odds in favor of H_0.

then we accept H_0, otherwise reject it. Like LLR, the BF value can also be transformed or formulated as a confidence measure for ASR.

As shown in Jiang & Deng (2001a), Bayes factors is a powerful statistical tool to model composite hypotheses and can be used to solve many different verification problems. The same formulation proposed for speaker verification in Jiang & Deng (2001a) is also equally applicable to the above UV problem though no research work has been reported about this. The key issues are what role the prior distributions $p(\lambda_0 \mid H_0)$ and $p(\lambda_1 \mid H_1)$ will play in utterance verification and how to use them as a flexible tool to incorporate a variety of information sources useful for UV.

5 Some Recent Efforts

Confidence measures or utterance verification aims to verify reliability of speech recognition outputs, which significantly differs from other typical verification problems in statistics, such as test for goodness-of-fit, and outlier detection in statistical data analysis. We believe that it is beneficial not to isolate confidence measures (or utterance verification) from its prior recognition stage. In acoustic level, it is very important to know the distribution properties of competing sources in recognition phase in order to optimize performance of CM or UV. In the following, I will first present two pieces of recent research works along this direction. Besides, I will also briefly summary some other research works to integrate some high-level knowledge (beyond acoustic information) for CM or UV.

5.1 In-search Data Selection for Accurate Competing Models

Under the UV framework, it is not an easy job to model the alternative hypothesis. Juang & Lee (1995a), Sukkar & Lee (1996), Rahim, Lee & Juang (1997a) propose to use the so-called *anti-models* for this purpose. However, it is still unclear what data should be used to estimate these *anti-models*. In their works, some heuristic methods are adopted, such as performing forced-alignment against a wrong or random transcript to generate training data for each anti-model. More recently, Jiang et al (2001b) propose a well-defined in-search data selection procedure to collect the most representative competing tokens for each HMM in the system. Then the selected tokens can be used to estimate highly accurate competing models for the utterance verification purpose.

In Jiang et al (2001b), we first define competing tokens (CT) of any a given HMM model as data segments which are mis-recognized to this model during recognition. A dynamic in-search data selection method is proposed to collect competing tokens for every HMM automatically from training data set. In the method, every utterance in training set is recognized with the Viterbi beam search algorithm just as in regular recognition phase. During the Viterbi search, all potential segments located in all active partial paths within the search beam width are compared with the reference segmentation generated from a forced-alignment procedure to determine whether each segment should be a competing token or true token of the model. The procedure is carried out for all training data to collect two token sets, namely the competing token set $\mathcal{S}_C(a)$ and the true token set $\mathcal{S}_T(a)$ for every HMM a in the system. The competing information collected in this way is very valuable for utterance verification. Given that a speech observation X is recognized as W by the decoder, the original hypotheses in eq.(7) can be re-phrased as follows:

$$\begin{aligned} H_0 &: \quad X \text{ belongs to } W\text{'s true token set } \mathcal{S}_T(W), \text{i.e., } X \in \mathcal{S}_T(W) \\ H_1 &: \quad X \text{ belongs to } W\text{'s competing token set } \mathcal{S}_C(W), \text{i.e., } X \in \mathcal{S}_C(W) \end{aligned} \tag{10}$$

Comparing with the original hypotheses, both the null hypothesis H_0 and the alternative hypothesis H_1 in the above are well-defined from available data, which in turn make our modeling problem easier. The simplest way to model them is to estimate two different models Λ_T and Λ_C for $\mathcal{S}_T(W)$ and $\mathcal{S}_C(W)$ respectively, based on all tokens collected from training data. Then the LRT-based utterance verification is operated as follows:

$$\eta = \frac{p(X \mid H_0)}{p(X \mid H_1)} = \frac{\Pr(X \in \mathcal{S}_T(W))}{\Pr(X \in \mathcal{S}_C(W))} = \frac{p(X \mid \Lambda_T)}{p(X \mid \Lambda_C)} \underset{H_1}{\overset{H_0}{\gtrless}} \tau \tag{11}$$

where τ is the critical decision threshold. The above models Λ_T and Λ_C can be estimated based on different criteria, such as maximum likelihood (ML), or minimum verification error (MVE), etc. Jiang et al (2001b) shows the ML-trained models already significantly surpass the conventional UV methods, such as in Sukkar & Lee (1996), Sukkar et al (1997), Rahim et al (1997a, 1997b).

5.2 UV based on neighborhood Information in Model Space

In Jiang & Lee (2002), a novel approach is proposed for utterance verification based on competing information in model space. First of all, let's look at the model space $\mathcal{T}$ of HMM. Each HMM λ in the system can be viewed as a point in the model space $\mathcal{T}$. Intuitively, we can imagine two nested neighborhoods surrounding the underlying model λ, namely a small neighborhood Λ_1 and a medium neighborhood Λ_2. The small neighborhood Λ_1 is a tiny neighborhood which tightly surrounds the underlying model λ. As indicated in Jiang et al (1999), a neighborhood with a relatively small size contains all variants of the original model due to estimation errors and possible mismatches in testing. It serves as a robust representation of the original model. On the other hand, the medium neighborhood Λ_2 is significantly larger than Λ_1. As the neighborhood size increases, it starts to cover all of its competing models in the model space, which by definition should be close to the original one in some sense. Based on the concept, we can translate the original hypotheses in eq.(7) in another way.

Once again, assume a speech observation X is recognized as W which is represented by the model λ_W. We are interested in verifying the reliability of the decision. Given the decision that X is recognized as model λ_W, if X is not from the model λ_W (as stated in the alternative hypothesis), it is reasonable to consider that X probably comes from some competing model of λ_W. Therefore, we can translate the original hypotheses in eq.(7) as:

$$\begin{aligned} H_0: &\quad \text{The true model of } X \text{ locates in the small neighborhood } \Lambda_1 \\ H_1: &\quad \text{The true model of } X \text{ locates in the region } \Lambda_2 - \Lambda_1 \end{aligned} \tag{12}$$

where $\Lambda_2 - \Lambda_1$ denotes the holed region inside the the medium neighborhood by excluding the small neighborhood, as shown in Figure 1.

In Jiang & Lee (2002), an approach based on Bayes factors is proposed to solve the above hypothesis testing problem.

$$\eta = \frac{\int_{\Lambda_1} f(X|\lambda) \cdot p_0(\lambda)\, \mathrm{d}\lambda}{\int_{\Lambda_2 - \Lambda_1} f(X|\lambda) \cdot p_1(\lambda)\, \mathrm{d}\lambda} \underset{H_1}{\overset{H_0}{\gtrless}} \tau \tag{13}$$

where $f(X|\lambda)$ is likelihood function, and $p_0(\lambda)$ and $p_1(\lambda)$ represent the prior distribution of model parameters under the hypothesis H_0 and H_1 respectively. Furthermore, Jiang & Lee

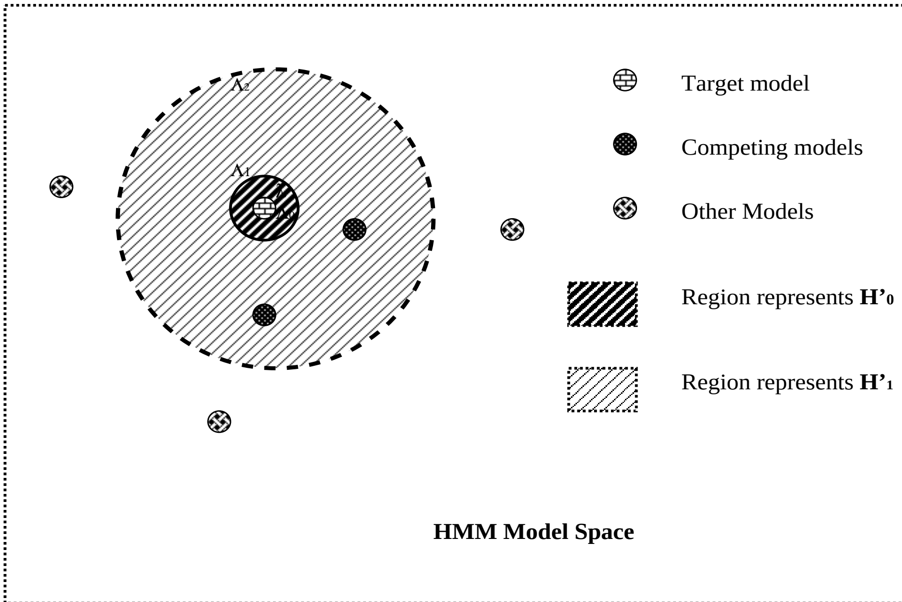

Figure 1: Illustration of the hypothesis testing based on the neighborhood information in model space.

(2002) propose two simple methods, i.e. a parametric definition and a nonparametric one, to quantitively define the neighborhoods as well as the prior distributions for the above formulation. Some preliminary experiments show some promising results for utterance verification based on the above framework. Obviously, much more research efforts are needed to define the neighborhoods in a more precise and controllable manner.

5.3 Incorporation of High-Level Information for CM

So far we have concentrated on confidence measures which solely rely on acoustic information. However, other syntactical or sematic information is also reported to provide certain clues for the purpose of confidence measure, such as Young (1994), Pao et al (1998), Zhang et al (2001), etc. More recently, Cox & Dasmahapatra (2002) report that human can clearly identify a certain portion of recognition errors in recognizer outputs on purely semantic grounds. They also propose to use latent semantic analysis (LSA) to annotate confidence scores for recognized words. Latent sematic analysis (LSA) is a technique for associating words that are "semantically coherent". The semantic coherence between any two words is computed as the cosine of the angle between the two vectors corresponding to these two words in a reduced subspace. Thus, confidence measure of a recognized word is calculated as an average of coherence of this word with all other recognized words in a close context. Although CM's based on this kind of semantic information is generally not as good as the best CM's in the acoustic level, a combination probably will yield a better performance due to their clear independence.

6 Performance and Applications of CM: Capabilities and Limitations

It is well known that good confidence measures will largely benefit a variety of ASR applications, e.g., to smartly reject non-speech noises, detect/reject out-of-vocabulary words, detect/correct some potential recognition mistakes, clean up human transcription errors in large training corpus, guide the system to perform un-supervised learning, provide side information to assist high-level speech understanding and dialogue management, and so on and so forth. However, confidence measures for ASR is an extremely difficult problem. Even today's best available CM's are generally not good enough to effectively support most of the above-mentioned applications. In this section, we first briefly talk about the assessment problem of confidence measures in ASR. Then, based on my personal understanding, I will discuss on performance issues of various CM's. At last, I will point out several promising applications for the current CM's in ASR even though many other applications are apparently beyond the capability of today's techniques.

When evaluating confidence measure annotation, we usually encounter two types of errors, namely false alarm errors and false rejection errors. Obviously, receiver operating characteristic (ROC) curve gives a full picture of verification performance at all operating points. In many cases, it is convenient to use a single-number metric for CM assessment. Some widely used metrics include equal error rate (EER), confidence error rate, normalized cross entropy, etc. Refer to Kemp & Schaaf (1997), Siu & Gish (1999), Maison & Gopinath (2001) and Wessel et al (2001) for details. Another important issue in CM evaluation is to take recognition boundaries into account. For example, a correctly recognized word may have a very low confidence measure because its boundary is wrong (though its identity is correct). Thus, it is helpful to use the concept of *"word-correctness"* proposed by Weintraub et al (1997) in evaluating CM's.

As far as CM's performance issues are concerned, it has been widely reported that N-Best related feature predictors perform much better than other predictors introduced in section 2 (see Chase, 1997, Rueber, 1997, Williams & Renals, 1999, etc). Moreover, Wessel et al (1998, 1999, 2001) clearly shows that posterior probabilities calculated from word graphs significantly outperform N-best-related confidence measures. On the other hand, along a totally different line, Sukkar et al (1996, 1997) and Rahim et al (1997a, 1997b) demonstrate that MVE-based discriminative training significantly improve performance of utterance verification. Furthermore, Jiang et al (2001b) shows the performance of utterance verification is largely improved over the previous UV approaches by using an in-search data selection method to train some highly accurate competing models. In Garcia et al (1999), the conventional LRT-based utterance verification is compared with posterior probabilities in word graph albeit their implementation is an approximate one. The results show word-graph-based posterior probabilities outperform the LRT-based utterance verification methods. However, it still remains unclear how the approach in Jiang et al (2001b) compares with word-graph-based posterior probabilities in Wessel et al (2001). Moreover, it will be more informative if all CM's are evaluated in a common corpus for several well-designed verification tasks. Generally speaking, the CM's based on posterior probabilities derived from word graphs are advantageous since language model scores can be naturally incorporated in CM computation in addition to acoustic information. But once being strictly implemented as in Wessel et al (2001), the performance of CM's can not be easily improved within the same paradigm because word graphs with various sizes usually generate CM's with similar performance. On the other hand, performance of utterance verification can be progressively improved by estimating better and better verification models. And the hypothesis testing paradigm, as formulated in LRT- or Bayes-factors-based testing, provides a flexible framework to incorporate a variety of knowledge sources which may be useful for CM computation.

As already mentioned above, the overall performance of CM's (even the best ones) remains fairly poor, which largely limits their applications. Some early research on CM aimed to detect out-of-vocabulary (OOV) words in large-vocabulary ASR system. However, even by today, an effective detection of OOV words in continuous speech recognition remains as an open question. It seems only feasible to use CM's to reject OOV words in some constrained small vocabulary applications, such as isolated voice command controlling, etc. Besides, a large number of works have been conducted to improve ASR performance with assistance of various CM's, e.g., Neti et al (1997), Jitsuhiro et al (1998), Vergyri (2000), Wessel et al (2000), Tan et al (2000), Lleida & Rose (2000), Koo et al (2001) and others. A consistent and significant error reduction over the state-of-the-art performance is still not an easy goal to achieve[4] unless the performance of CM's is enhanced further. Moreover, CM's have been included in many spoken dialogue systems to provide certain level of support for language understanding and dialogue management. But the CM's themselves are found not robust and reliable enough to be a solid basis for decision-making in many cases. In spite of these, the current CM techniques still have a chance to shine if they are applied to a proper place. Although it is hard to detect or correct errors made by ASR systems by using CM's, it seems much easier to use CM's to detect human-made errors. Thus, it is promising to use CM's to clean-up or verify transcription in a large corpus, such as some preliminary studies in Arslan & Hansen (1999), Li et al (2002). In addition, there are two other successful stories to apply CM's to verify some decisions not hypothesized by ASR systems. One is verbal information verification (VIV) in Li et al (2000) and another one is Liu et al (2001) where they demonstrate effectiveness of using CM's for search space pruning prior to recognition stage. Moreover, another interesting area to apply CM's is un-supervised adaptation where CM's are used to select more reliable speech segments from recognizer's outputs to self-improve recognition models, e.g. Wallhoff et al (2000) and many others. One important issue here is that the operating point in verification stage should be set up to guarantee a low false acceptance rate.

[4] If counting the correct recognition results which are mistakenly rejected.

7 Final Remarks

Although there are various types of CM's reported for ASR, almost all CM's in acoustic level fundamentally rely almost entirely on a single information source, namely how much the underlying decision can overtake other possible competitors. The larger the difference is the more confident we will believe the decision to be. This explains why most research works to combine a variety of CM's usually do not yield better results. The various CM or UV methods mentioned in this paper attempt to explore this discrepancy in different ways (direct or indirect). For example, in the posterior probability method based on a word graph, if the recognition result significantly surpasses other competing choices in the word graph, the contribution of the recognized path will dominate the total posteriori probability computed based on the forward-backward algorithm. In this case, the derived CM will be large (close to 1). If other competing paths in the word graph come very close to the recognized results, the contribution of the recognized path will be relatively small when computing the posterior probability. Thus, the derived CM will be small (close to 0). Similarly in UV, if the recognized result largely surpasses other competitors, the likelihood under the null hypothesis will be significantly larger than that of the alternative hypothesis. As a result, the likelihood ratio will be large. On the other hand, the likelihood ratio will be small if the competing sources from the alternative hypothesis gives comparable results with the recognized one in the null hypothesis. This also explains why it is very important to model distribution properties of competing hypotheses when deriving CM's for ASR. Apparently, it is a real challenge to compute any effective CM's beyond this sole source. Besides, one major drawback of almost all CM or UV methods is that we only verify segment identities but never question the correctness of segmentation hypothesized by ASR systems. It is common that most recognition errors accompany with segmentation mistakes in continuous speech recognition. A preliminary study on boundary adjustment for UV can be found in Matsui et al (2001). We believe it is critical to improve performance of CM's by taking this segmentation issue into account. How to consider it effectively in any formal way in CM estimation still remains unclear. Finally, despite a large number of research activities in the past, confidence measure estimation for ASR still remains unsolved in so many aspects. Due to its importance in practice and its difficulty in theory, we expect much more research efforts will be devoted into this topic in coming years.

References

1. Arslan, L. & Hansen, J. (1999). Selective training for hidden Markov model with applications to speech classification. *IEEE Trans. on Speech and Audio Processing 7(1)*, 46-54.

2. Asadi, A., Schwartz, R. & Makhoul, J. (1990). Automatic detection of new words in a large vocabulary continuous speech recognition system. *Proc. of International Conference on Acoustics, Speech and Signal Processing*, 125-128.

3. Benitez, M. C., Rubio, A. & Torre, A. (2000). Different confidence measures for word verification in speech recognition. *Speech Communication 32*, 79-94.

4. Bourlard, H., D'hoore, B & Boite, J.-M. (1994). Optimizing recognition and rejection performance in word-spotting systems. *Proc. of International Conference on Acoustics, Speech and Signal Processing*, I-373-376.

5. Charlet, D., Mercier, G. & Jouvet, D. (2001). On combining confidence measures for improved rejection of incorrect data. *Proc. of European Conference on Speech Communication Technology.*

6. Chase, L. (1997). Word and acoustic confidence annotation for large vocabulary speech recognition. *Proc. of European Conference on Speech Communication Technology*, 815-818.

7. Chigier, B. (1992). Rejection and keyword spotting algorithms for a directory assistance city name recognition application. *Proc. of International Conference on Acoustics, Speech and Signal Processing*, II-93-96.

8. Cox, S. & Rose, R. (1996). Confidence measures for the Switchboard database. *Proc. of International Conference on Acoustics, Speech and Signal Processing*, 511-514.

9. Cox, S. & Dasmahapatra, S. (2002). High-level approaches to confidence estimation in speech recognition. *IEEE Trans. on Speech and Audio Processing 10(7)*, 460-471.

10. Dolfing, J. G. A. & Wendemuth, A. (1998). Combination of confidence measures in isolated word recognition. *Proc. of International Conference on Spoken Language Processing.*

11. Eide, E., Gish, H., Jeanrenaud, P. & Mielke, A. (1995). Understanding and improving speech recognition performance through the use of diagnostic tools. *Proc. of International Conference on Acoustics, Speech and Signal Processing*, 221-224.

12. Garcia, M. C., Reichl, W. & Ortmanns, S. (1999). On combining confidence measures in HMM-based speech recognizers. *Proc. of Automatic Speech Recognition and Understanding Workshop.*

13. Gillick, L. Ito, Y. & Young, J. (1997). A probabilistic approach to confidence estimation and evaluation. *Proc. of International Conference on Acoustics, Speech and Signal Processing*, 879-882.

14. Gong, Y.-F. (1995). Speech recognition in noisy environments: a survey. *Speech Communication 16*, 261-291.

15. Gunawardana, A., Hon, H.-W. & Jiang, L. (1998). Word-based acoustic confidence measures for large-vocabulary speech recognition. *Proc. of International Conference on Spoken Language Processing*, 791-794.

16. Gupta, S. K. & Soong, F. K. (1998). Improved utterance rejection using length dependent thresholds. *Proc. of International Conference on Spoken Language Processing*, 795-798.

17. Hacioglu, K. & Ward, W. (2002). A concept graph based confidence measure, *Proc. of International Conference on Acoustics, Speech and Signal Processing*, I-225-228.

18. Hernandez, A. G. & Marino, J. B. (2000). Contextual confidence measures for continuous speech recognition. *Proc. of International Conference on Acoustics, Speech and Signal Processing*, 1803-1806.

19. Jiang, H., Hirose, K. & Huo, Q. (1999) Robust speech recognition based on Bayesian prediction approach. *IEEE Trans. on Speech and Audio Processing 7(4)*, 426-440.

20. Jiang, H. & Deng, L. (2001a). A Bayesian approach to the verification problem: applications to speaker verification. *IEEE Trans. on Speech and Audio Processing 9(8)*, 874-884.

21. Jiang, H., Soong, F. K. & Lee, C.-H. (2001b). A data selection strategy for utterance verification in continuous speech recognition. *Proc. of European Conference on Speech Communication and Technology*, 2573-2576.

22. Jiang, H. & Lee, C.-H. (2002). Utterance verification based on neighborhood information and Bayes Factors. *Proc. of International Conference on Spoken Language Processing.*

23. Jiang, H. & Lee, C.-H. (2003). A new approach to utterance verification based on neighborhood information in model space. *to appear in IEEE Trans. on Speech and Audio Processing.*

24. Jiang, L. & Huang, X.-D. (1998). Vocabulary-independent word confidence measure using subword features. *Proc. of International Conference on Spoken Language Processing.*

25. Jitsuhiro, T., Takahashi, S. & Aikawa, K. (1998). Rejection of out-of-vocabulary words using phoneme confidence likelihood. *Proc. of International Conference on Acoustics, Speech and Signal Processing*, 217-220.

26. Juang, B.-H. (1991). Speech recognition in adverse environments. *Computer Speech and Language 5*, 275-294.

27. Kamppari, S. O. & Hazen, T. J. (2000). Word and phone level acoustic confidence scoring. *Proc. of International Conference on Acoustics, Speech and Signal Processing*, 1799-1802.

28. Kemp, T. & Schaaf, T. (1997). Estimating confidence using word lattices. *Proc. of European Conference on Speech Communication Technology*, 827-830.

29. Koo, M.-W., Lee, C.-H., & Juang, B.-H. (2001) Speech recognition and utterance verification based on a generalized confidence score. *IEEE Trans. on Speech and Audio Processing 9(8)*, 821-832.

30. Lee, C.-H. (1998). A tutorial on speaker and utterance verification. *Proc. of NORSIG*, 9-16.

31. Lee, C.-H. (1998b). On stochastic feature and model compensation approaches to robust speech recognition. *Speech Communication 25*, 29-47.

32. Lee, C.-H. (2001). Statistical confidence measures and their applications. *Proc. of ICSP*, August.

33. Li, Q, Juang, B.-H., Zhou, Q. & Lee, C.-H. (2000). Automatic verbal information verification for user authentication. *IEEE Trans. on Speech and Audio Processing 8(5)*, 585-596.

34. Li, Q., Jiang, H., Zhou Q., & Zheng, J. (2002). Automatic Enrollment for Speaker Authentication. *Proc. of International Conference on Spoken Language Processing.*

35. Lin, Q., Das, S., Lubensky, D., Picheny, M. (1998). A new confidence measure based on rank-ordering subphone scores. *Proc. of International Conference on Spoken Language Processing.*

36. Lin, Q., Lubensky, D., Roukos, S. (1999). Use of recursive mumble models for confidence measuring. *Proc. of European Conference on Speech Communication Technology.*

37. Liu, F., Afify, M., Jiang, H. & Siohan, O. (2001). A new verification-based fast match approach to large vocabulary speech recognition. *Proc. of European Conference on Speech Communication and Technology*, 851-854.

38. Lleida, E. & Rose, R. C. (2000). Utterance verification in continuous speech recognition: decoding and training procedures. *IEEE Trans. on Speech and Audio Processing 6*, 558-568.

39. Maison, B. & Gopinath, R. (2001). Robust confidence annotation and rejection for continuous speech recognition. *Proc. of International Conference on Acoustics, Speech and Signal Processing.*

40. Matsui, T., Soong, F. K. & Juang, B.-H. (2001). Verification of multi-class recognition decision using classification approach. *Proc. of Automatic Speech Recognition and Understanding Workshop.*

41. Mathan, L. & Miclet, L. (1991). Rejection of extraneous input in speech recognition applications, using multi-layer perceptrons and the trace of HMMs. *Proc. of International Conference on Acoustics, Speech and Signal Processing*, 93-96.

42. Mengusoglu, E. & Ris, C. (2001). Use of acoustic prior information for confidence measure in ASR applications. *Proc. of European Conference on Speech Communication Technology.*

43. Modi, P. & Rahim, M. (1997) Discriminative utterance verification using multiple confidence measures. *Proc. of European Conference on Speech Communication Technology*, 103-106.

44. Moreau, N. & Jouvet, D. (1999). Use of a confidence measure based on frame level likelihood ratios for the rejection of incorrect data. *Proc. of European Conference on Speech Communication Technology.*

45. Moreno, P. J., Logan, B. & Raj, B. (2001). A boosting approach for confidence scoring. *Proc. of European Conference on Speech Communication Technology.*

46. Neti, C. V., Roukos, S. & Eide, E. (1997). Word-based confidence measures as a guide for stack search in speech recognition. *Proc. of International Conference on Acoustics, Speech and Signal Processing*, 883-886.

47. Rahim, M. G., Lee, C.-H., Juang, B.-H. (1997a). Discriminative utterance verification for connected digits recognition. *IEEE Trans. on Speech and Audio Processing 5(3)*, 266-277.

48. Rahim, M. G. & Lee, C.-H. (1997b). String-based minimum verification error (SB-MVE) training for speech recognition. *Computer Speech and Language 11*, 147-160.

49. Rahim, M. G. & Lee, C.-H. (1997c). A study on robust utterance verification for connected digits recognition. *The Journal of the Acoustical Society of America 101(5)*, 2892-2902.

50. Palmer, D. D., Ostendorf, M. (2001). Improved word confidence estimation using long range features. *Proc. of European Conference on Speech Communication Technology.*

51. Pao, C., Schmid, P. & Glass, J. (1998). Confidence scoring for speech understanding systems. *Proc. of International Conference on Spoken Language Processing.*

52. Rohlicek, J. R., Jeanrenaud, P., Ng, K., Gish, H., Musicus, B. & Siu, M. (1993). Phonetic training and language modeling for word spotting. *Proc. of International Conference on Acoustics, Speech and Signal Processing*, II459-462.

53. Rose, R. C. (1992). Discriminant wordspotting techniques for rejecting non-vocabulary utterances in unconstrained speech. *Proc. of International Conference on Acoustics, Speech and Signal Processing*, II-105-108.

54. Rose, R. C., Juang, B. H. & Lee, C. H. (1995a). A training procedure for verifying string hypothesis in continuous speech recognition. *Proc. of International Conference on Acoustics, Speech and Signal Processing*, 281-284.

55. Rose, R. C. (1995b). Keyword detection in conversational speech utterances using hidden Markov model based continuous speech recognition. *Computer Speech and Language 9*, 309-333.

56. Rose, R. C. & Riccardi, G. (1999). Automatic speech recognition using acoustic confidence conditioned language models. *Proc. of European Conference on Speech Communication Technology.*

57. Rose, R. C., Yao, H., Riccardi, G. & Wright, J. (2001). Integration of utterance verification with statistical language modeling and spoken language understanding. *Speech Communication 34*, 321-331.

58. Rueber, B. (1997). Obtaining confidence measures from sentence probabilities. *Proc. of European Conference on Speech Communication Technology.*

59. Sankar, A. & Wu, S.-L. (2003). Utterance verification based on statistics of phone-level confidence scores. *Proc. of International Conference on Acoustics, Speech and Signal Processing*, I-584-587.

60. San-Segundo, R., Pellom, B., Hacioglu, K. & Ward, W. (2001). Confidence measures for spoken dialogue systems. *Proc. of International Conference on Acoustics, Speech and Signal Processing.*

61. Schaaf, T. & Kemp, T. (1997). Confidence measures for spontaneous speech recognition. *Proc. of International Conference on Acoustics, Speech and Signal Processing*, 875-878.

62. Siu, M. & Gish, H. (1999). Evaluation of word confidence for speech recognition systems. *Computer Speech and Language 13*, 299-319.

63. Sukkar, R. A. & Wilpon J. G. (1993). A two pass classification for utterance rejection in keyword spotting. *Proc. of International Conference on Acoustics, Speech and Signal Processing*, II-451-454.

64. Sukkar, R. A. (1994). Rejection for connected digit recognition based on GPD segmental discrimination. *Proc. of International Conference on Acoustics, Speech and Signal Processing*, I-393-396.

65. Sukkar, R. A. & Lee, C.-H. (1996). Vocabulary independent discriminative utterance verification for nonkeyword rejection in subword based speech recognition. *IEEE Trans. on Speech and Audio Processing 4(6)*, 420-429.

66. Sukkar, R. A., Setlur, A. R., Lee, C.-H. & Jacob, J. (1997) Verifying and correcting recognition string hypotheses using discriminative utterance verification. *Speech Communication 22*, 333-342.

67. Tan, B. T., Gu, Y. & Thomas, T. (2000). Utterance verification based speech recognition system. *Proc. of International Conference on Spoken Language Processing.*

68. Tan, B. T., Gu, Y. & Thomas, T. (2001). Word level confidence measures using N-best sub-hypotheses likelihood ratio. *Proc. of European Conference on Speech Communication Technology.*

69. Vergyri, D. (2000). Use of word level side information to improve speech recognition. *Proc. of International Conference on Acoustics, Speech and Signal Processing*, 1823-1826.

70. Wallhoff, F. Willett, D. & Rigoll, G. (2000). Frame-discriminative and confidence-driven adaptation for LVCSR. *Proc. of International Conference on Acoustics, Speech and Signal Processing*, 1835-1838.

71. Weintraub, M., Beaufays, F., Rivlin, Z., Konig, Y. & Stolcke, A. (1997). Neural-network based measures of confidence for word recognition. *Proc. of International Conference on Acoustics, Speech and Signal Processing*, 887-890.

72. Wessel F., Macherey, K. & Schluter R. (1998). Using word probabilities as confidence measures. *Proc. of International Conference on Acoustics, Speech and Signal Processing*, 225-228.

73. Wessel F., Macherey, K. & Ney., H. (1999). A comparison of word graph and N-best list based confidence measures. *Proc. of European Conference on Speech Communication Technology*, 315-318.

74. Wessel, F., Schluter, R. & Ney, H. (2000). Using posterior word probabilities for improved speech recognition. *Proc. of International Conference on Acoustics, Speech and Signal Processing*, 1587-1590.

75. Wessel, F., Schluter, R., Macherey, K. & Ney, H. (2001). Confidence measures for large vocabulary continuous speech recognition. *IEEE Trans. on Speech and Audio Processing 9(3)*, 288-298.

76. Willett, D., Worm, A., Neukirchen, C. & Rigoll, G. (1998). Confidence measures for HMM-based speech recognition. *Proc. of International Conference on Spoken Language Processing.*

77. Williams, G. & Renals, S. (1999). Confidence measures from local posterior probability estimates. *Computer Speech and Language 13*, 395-411.

78. Wilpon, J. G., Rabiner, L. R., Lee, C.-H. & Goldman, R. (1990). Automatic recognition of keywords in unconstrained speech using hidden Markov models. *IEEE Trans. on Acoustics, Speech, and Signal Processing 38(11)*, 1870-1878.

79. Young, S. R. & Ward, W. (1993). Learning new words from spontaneous speech. emphProc. of International Conference on Acoustics, Speech and Signal Processing, II-590-591.

80. Young, S. R. (1994). Detecting misrecognitions and out-of-vocabulary words. *Proc. of International Conference on Acoustics, Speech and Signal Processing*, II-21-24.

81. Zhang, R. & Rudnicky, A. I. (2001). Word level confidence annotation using combinations of features. *Proc. of European Conference on Speech Communication Technology.*