LIKELIHOOD RATIO DECODING AND CONFIDENCE MEASURES FOR CONTINUOUS SPEECH RECOGNITION

Eduardo Lleida * *Richard C. Rose*

University of Zaragoza, Spain AT&T Research, Murray Hill, USA

ABSTRACT

Automatic speech recognition (ASR) systems are being integrated into a wider variety of tasks involving human–machine interaction. In evaluating these systems, however, it has become clear that more accurate means must be developed for detecting when portions of the decoded recognition hypotheses are either incorrect or represent out–of–vocabulary utterances. This paper describes the use of confidence measures based on likelihood ratio based optimization procedures for decoding and rescoring word hypotheses in an HMM based speech recognizer. These techniques are applied to spontaneous utterances obtained from a "movie locator" based dialog task.

1. INTRODUCTION

It is well known that speech recognition performance degrades rapidly when presented with utterances that either contain out-of-vocabulary words or are not well modeled by predefined language models. It is very difficult to anticipate all possible input utterances in the process of configuring an automatic speech recognition (ASR) system. As a result, it is often desirable to accept only the portions of a decoded utterance that have been decoded with sufficient "confidence". This implies the existence of an utterance verification (UV) procedure. The goal of UV is to verify whether a hypothesized word or string of words corresponds to actual occurrences of those words or to word substitutions or insertions, referred to here collectively as false alarms.

HMM based continuous speech recognition systems are based on decoding the input speech by maximizing the likelihood of the input data with respect to a set of HMM models. Depending on the application, the decoder produces the word sequence corresponding to the single best path, a list of the N-best word sequences, or a lattice of word hypotheses. In any case, some measure of confidence can be assigned to each hypothesis that may be used to determine whether it should be passed on to later stages of processing.

Utterance verification using confidence measures has generally been applied as a two pass procedure. First, the input speech is segmented using a maximum likelihood decoder, and, second, a confidence measure is computed for each segment. This procedure has the drawback that it must verify a hypothesized lexical item given the possibly errorful segmentation produced by the decoder. A different approach to utterance verification where decoding and UV are integrated into a single procedure is described in Section 3. The decoder is modified so that the decoded string is that which obtains the highest confidence score with respect to all possible string hypotheses. This implicitly allows an evaluation of an entire ensemble of hypotheses simultaneously.

The organization of the paper is as follows. The spontaneous spoken utterances from the "movie locator" spoken dialog task are described in section 2 [4]. A discussion of the two pass UV procedure, the new one pass UV procedure, and likelihood ratio tests as applied to UV is provided in section 3. Several definitions of confidence measures for utterance verification are presented in Section 4. Finally, in Section 5, a set of experiments are described where these techniques are applied to utterances from the movie locator task.

2. MOVIE LOCATOR TASK

The "movie locator" is a telephone service based on speech technology which answers users' queries about movies that are currently showing at theaters in their area [4]. The users are not constrained to follow any menu or prespecified pattern in formulating their queries. During a trial of the system over the public switched telephone network, a collection of 4777 spontaneous spoken utterances were recorded. A total of 3025 sentences were used for training acoustic models and 752 utterances were used for testing. The subword models used in the recognizer consisted of 43 context independent units. Recognition was performed using a finite state grammar built from the specification of the service, with a lexicon of 570 different words. The total number of words in the test set was 4864, where 134 of them were out-of-vocabulary. There were 275 sentences in the test set that could not be parsed by the finite state grammar, where 85 of these sentences contained out-of-vocabulary words. Recognition performance of 91.2% word error rate was obtained on the "in–grammar" utterances. The feature set used for recognition included 12 Mel-cepstrum, 12 delta Mel-cepstrum, 12 delta-delta Mel-cepstrum, energy, delta energy and delta-delta energy coefficients. Cepstral mean normalization was applied to the cepstral coefficients to compensate for linear channel distortions.

*This work has been supported by the CICYT grant TIC95-0884-C04-04

3. UTTERANCE VERIFICATION STRATEGIES

3.1. Two-pass recognition/verification

Utterance verification (UV) is considered here as a hypothesis testing procedure which determines whether subword, word, or sentence level recognition hypotheses correspond to actual occurrences of these events in the utterance. It is assumed that the input to the speech recognizer is a sequence of feature vectors $Y = \{\vec{y}_1, ..., \vec{y}_T\}$ representing a speech utterance containing both within vocabulary (*target*) and out-of-vocabulary (*imposter*) words. It is also assumed that the output of the recognizer is a single *word* string hypothesis $\mathcal{W} = W_1, ..., W_K$ of length K.

The likelihood ratio $\mathcal{LR}$ test is designed to determine whether or not a sequence of feature vectors Y were generated by a given family of probability densities, defining the following test:
$\mathcal{H}_0$: null hypothesis, Y generated by target model λ_c
$\mathcal{H}_1$: alternative hypothesis, Y generated by alternative model λ_a

$$\mathcal{LR}(Y, \lambda^c, \lambda^a) = \frac{P(Y|\lambda^c)}{P(Y|\lambda^a)} \underset{\mathcal{H}_1}{\overset{\mathcal{H}_0}{\gtrless}} \tau \qquad (1)$$

where τ is a decision threshold. For utterance verification in ASR, λ^c and λ^a are HMM's corresponding to the *correct* or *target* hypothesis and the *alternative* hypothesis respectively.

As with any hypothesis testing procedure, an utterance verification procedure is evaluated with respect to two types of errors. *Type I errors* correspond to the correctly decoded vocabulary words being rejected by the utterance verification procedure. *Type II errors* or *false alarms* correspond to incorrectly decoded word insertions and substitutions being accepted by the UV process.

The simplest way of incorporating an utterance verification procedure in a speech recognition system is to implement UV as a post-processor, assigning a confidence measure to each hypothesis produced by the speech recognizer. This is referred to as *two-pass* recognition/verification. Figure 1 shows the block diagram of the *two-pass* procedure. The first pass gives a segmentation of the input speech based on a set of HMM's called here *recognition* HMM models. Based on this maximum likelihood segmentation, the second pass computes the likelihood ratio between the null and alternate hypothesis for each word hypothesis segment by using a set of target and alternative HMM models. This ratio could be used to accept or reject single words or could be combined to give a ratio score for the whole sentence. The target and alternative models could be trained in a discriminative way to reduce the type I + type II errors as in [3]. However, as the recognition step is performed with the *recognition* HMM set, no improvements could be expected in the word hypothesis and segmentation performances.

3.2. One-pass recognition/verification

Some increase in the word hypothesis and segmentation performances could be expected when including the alternative

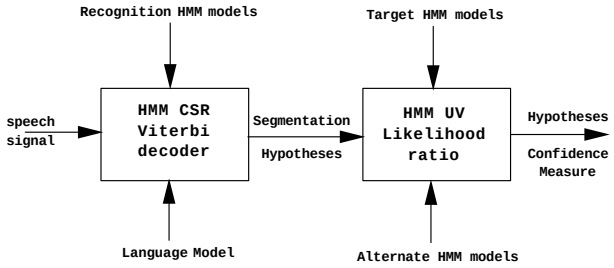

Figure 1: Two-pass utterance verification procedure

models in the recognition step. For this purpose, the speech recognition decoding criterion is modified so that the decoded string is that which obtains the highest confidence score with respect to all possible string hypotheses. This is referred to here as a *one-pass* recognition/verification strategy. The principal advantage of the *one-pass* strategy is that the hypothesis test is not applied to a single string. In performing search by directly optimizing the confidence measure or hypothesis test criterion, an entire ensemble of hypotheses are simultaneously being evaluated. Figure 2 shows the block diagram of this procedure.

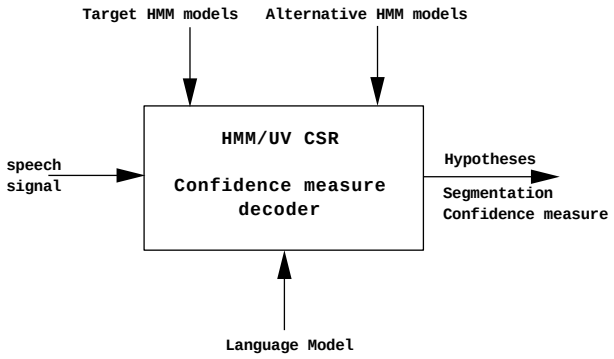

Figure 2: One-pass recognition/verification procedure

A practical implementation of the *one-pass* strategy was presented in [1]. It is described as a method for obtaining a state sequence that maximizes a likelihood ratio. Specifically, the optimum string maximizes the ratio of the likelihood with respect to a null hypothesis HMM model and dedicated alternate hypothesis HMM model. The decoding procedure is referred to here as *Likelihood Ratio Decoder* (LRD). The likelihood ratio decoder corresponds to a modified Viterbi algorithm which obtains the path which maximizes the likelihood ratio

$$\tilde{Q} = \arg\max_{Q} \mathcal{LR}(Y, \lambda^c, \lambda^a) \qquad (2)$$

As show in [1] this path could be found using a 3-D Viterbi search over the target hypothesis model states and alternate hypothesis states. As a result, a likelihood ratio based confidence measure could be obtained for every subword, word, or sentence hypotheses directly from the output of the LRD.

3.3. Alternative Hypothesis Models

Dedicated alternate hypothesis models are assigned to each context independent subword acoustic unit. In practice, the alternative hypothesis model has two roles in the likelihood ratio used for utterance verification. The first is to reduce the effect of sources of variability on the confidence measure. If the probabilities of the null hypothesis model and the alternative hypothesis model are similarly affected by some systematic variation in the observations, then forming the likelihood ratio should cancel the effects of the variability. For this purpose, a so called *background* HMM model, λ^a_{bg}, is defined. This is shared amongst all the subword units. The second role of the alternate model is more specifically to represent the incorrectly decoded hypotheses that are frequently confused with a given lexical item. In a subword model based CSR system, an *imposter* HMM model, λ^a_{im}, is defined for each subword unit to represent the false alarm space. The new alternative probability is computed as a linear combination of the background model probability and the imposter model probability.

$$P(Y|\lambda^a) = \alpha_{bg} P(Y \mid \lambda^a_{bg}) + \alpha_{im} P(Y \mid \lambda^a_{im}) \qquad (3)$$

where α_{bg} is the background HMM model and $\alpha_{im} = 1 - \alpha_{bg}$ is the imposter HMM model.

3.4. Likelihood ratio training

The training procedure for the target and alternative models has the goal of increasing the value of the likelihood ratio for correctly hypothesized words and decreasing the value for the false alarms. The reader is referred elsewhere for a more detailed discussion of the likelihood ratio training [1, 2].

The initialization of the HMM's is based on a maximum likelihood estimation. The background model is trained from the entire set of training utterances, and the imposter model for each subword unit is trained from the false alarms decoded for that unit. After this initialization, the likelihood ratio training is used to update the HMM parameters of the correct and imposter models.

4. WORD CONFIDENCE MEASURES FOR UTTERANCE VERIFICATION

Let CMU_u and CMW_w be the confidence measure for phonetic unit u and word w. A sentence S is composed of a sequence of words $W = w_1, ..., w_{N_S}$, where w_n is the n^{th} word in S and a word w_n is composed by a phonetic baseform $U_n = u_{n,1}, ..., u_{n,N_n}$ where $u_{n,k}$ is the k^{th} subword unit in U_n.

A phone or subword based confidence measure is computed from the likelihood ratio between the correct and alternative hypothesis models. The simplest phonetic confidence measure for a sequence of feature vectors $Y = \{\vec{y}_{t-N}, ..., \vec{y}_t\}$ is defined as

$$CMU_{u_i} = \mathcal{LR}(Y, \lambda^c, \lambda^a) = \frac{P(Y|\lambda^c)}{P(Y|\lambda^a)} \qquad (4)$$

where the likelihood ratio could be computed by using the *two-pass* or the *one-pass* UV strategies.

Given a definition of the subword level confidence measure, a word level confidence measure is computed as a non–uniform weighting of the subword level scores. It is important to prevent extreme values of the subword level likelihood ratio scores from dominating the word level confidence measures. These extreme values can often occur when a good *local* match has occurred in the decoding process even when the overall utterance has been incorrectly recognized. As a result, the best word level confidence scores tend to assign reduced weight to those extremes in subword level scores. Figure 3 shows the receiver operating characteristic (ROC) curves obtained when word level confidence scores are obtained using the arithmetic mean and the geometric mean of the subword level scores on the movie locator database. Since the geometric mean tends to emphasize the smaller valued subword level scores, the UV performance is better than that obtained using the arithmetic mean. The phonetic confidence measure defined in Equation 4 can still exhibit a wide dynamic range, and, therefore, be subject to the influence of out–liers. A way to limit the dynamic range of the subword based confidence scores is to use a sigmoid function

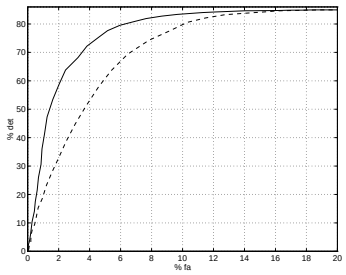

Figure 3: Word detection operating curve using a geometric mean combination (solid line) and arithmetic mean(dash line).

$$CMU_{u_i} = \frac{1}{1 + \exp(-\gamma_{u_i}(\log \mathcal{LR}_{u_i} - \tau_{u_i}))} \; . \qquad (5)$$

Figure 4 shows both the ROC curves and the type I + type II error curves when a sigmoidal weighting function is applied to the subword level confidence scores, and the word level confidence scores are obtained from the subword level scores using the arithmetic or the geometric mean. It is clear from Figure 4 that when the sigmoidal weighting is applied, the ROC curves are similar. However, the geometric mean is less sensitive to the setting of the confidence threshold.

5. DISCUSSION

A set of experiments was performed to provide a comparative evaluation of the one-pass and two-pass decoding procedures and to evaluate the effect of the likelihood ratio based training procedure. The arithmetic mean of sigmoidal phonetic confidence measures was used as the word level confidence measure in this set of experiments. Figure 5 shows a comparison between the one-pass procedure and the two-pass procedures. UV is performed either by using the likelihood ratio decoder to compute the subword level confidence

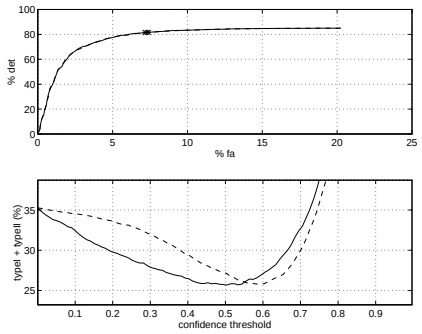

Figure 4: Word detection operating and type I + type II error curves using a geometric mean combination (solid line) and arithmetic mean(dash line) of the sigmoidal phonetic confidence measure with $\tau_{u_i} = \tau = 0$ and $\gamma_{u_i} = \gamma = 0.5$.

score (TPFR) or by computing the likelihoods $P(Y, Q^c|\lambda^c)$ and $P(Y, Q^a|\lambda^a)$ separately and taking their ratios (TPMR). Clearly, the one-pass procedure gives better performance than the two-pass procedure over the entire range of confidence threshold settings. Also, the use of the likelihood ratio decoder for "re-scoring" word hypotheses in the two-pass procedure (TPFR) outperforms the TPMR. The recognition

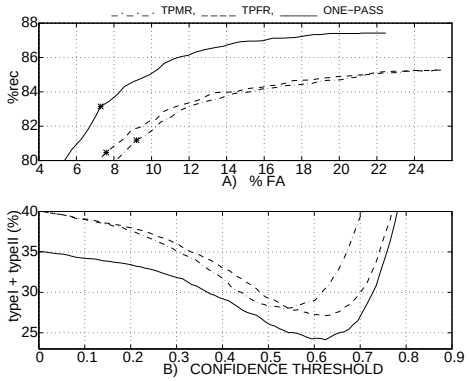

Figure 5: a,b) Comparison of UV scenarios

rate increases significantly when the likelihood ratio training procedure is used. Figure 6 shows the ROC curves for the in-grammar and out-of-grammar sentences respectively when using the likelihood ratio decoder and likelihood ratio training. In both cases, there is a significant increase in the performance of the recognizer in terms of word detection and false alarm rejection after applying the likelihood ratio training for training the target and alternative models. There is more than 40% improvement in performance at the "operating point" (minimum type I + type II errors) in the in-grammar sentences and more than 15% improvement in the out-of-grammar sentences.

6. SUMMARY

Likelihood ratio based confidence measures for HMM utterance verification and recognition have been presented. A single pass recognition and utterance verification strategy was evaluated with respect to a more traditional two-pass

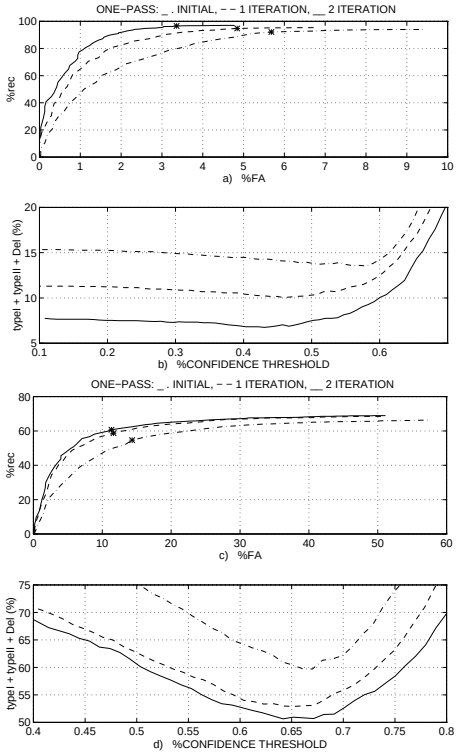

Figure 6: a,b) Likelihood ratio decoder results within the in-grammar sentences c,d) Likelihood ratio decoder results within the out-of-grammar sentences

approach to utterance verification. This single pass strategy was implemented using a "likelihood ratio decoder". A likelihood ratio based training procedure for estimating both target and alternate hypothesis HMM parameters was also evaluated. In both cases, significant improvement in the overall system accuracy was obtained for the spontaneous utterances of a movie locator task. Additional applications of the likelihood ratio training and decoding procedures have been reported elsewhere [5].

7. REFERENCES

[1] E. Lleida, R.C. Rose, "Efficient Decoding and Training Procedures for Utterance Verification in Continuous Speech Recognition", Proc. ICASSP-96.

[2] R.C. Rose, B.H. Juang, C.H. Lee, "Verifying String Hypothesis in Continuous Speech Recognition", Proc. ICASSP-95.

[3] R.A. Sukkar, C.H. Lee, B.H. Juang,, "A Vocabulary Independent Discriminatively Trained Method for Rejection of Non-Keywords in Subword Based Speech Recognition", EUROSPEECH-95, 1995.

[4] J.J. Wisowaty, "Continuous Speech Interface for a Movie Locator Service", Proc. of the Human Factors and Ergonomics Society, 1995.

[5] R.C. Rose et al, "A user-configurable system for voice label recognition", in these proceedings.