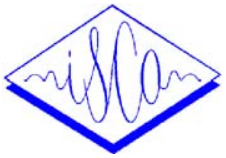

5th International Conference on Spoken Language Processing (ICSLP 98)
Sydney, Australia
November 30 - December 4, 1998

A NEW CONFIDENCE MEASURE BASED ON RANK-ORDERING SUBPHONE SCORES

Qiguang Lin, Subrata Das, David Lubensky, and Michael Picheny

IBM T.J. Watson Research Center
Computer Science Department
P. O. Box 218, Yorktown Heights, NY 10598, USA

10.21437/ICSLP.1998-818

ABSTRACT

This paper describes a new confidence measure based on rank-ordering subphone likelihood scores. The approach consists of three major steps. The first step is standard decoding which hypothesizes a word string given an utterance. Then forced Viterbi alignment is made at the second step. (This step is of course not needed for a Viterbi decoder.) From the aligned sentence, the third step computes likelihood scores of the hypothesized subphone and all other competing subphones to generate a list of subphones in the descending order of the likelihood scores. A rank is assigned to the hypothesized subphone according to its positioning in the list. The rank value is then merged to obtain the corresponding rank at the phone level. After the merge, selective weighting is applied such that contribution of phones having large acoustic variations is de-emphasized. Additional upper-bound limiting is also made to guarantee a rank computation not to be contaminated by a very bad segment. The new confidence measure has been favorably evaluated on word confirmation/rejection experiments with a small vocabulary of many confusable and short words. More specifically, experimental results show that the new approach outperforms other measures such as whole-word scores by reducing the equal error rate from 32% to 20%.

1. INTRODUCTION

Current speech recognition systems are not flawless. They are always associated with a certain amount of error rates. Although the errors cannot be eliminated completely, for many applications it is desired to know when an error is likely to have occurred so that subsequent action for the uncertain words may be taken. Depending upon actual applications, the action may include: (1) to reject the word in a voice-operated system (such as command/control navigation) or to ask the user to repeat/confirm such that unpredicted behavior of the system can be minimized; (2) to exclude the uncertain words from being utilized in unsupervised speaker/environment adaptation; (3) to highlight uncertain words in a dictation scenario for easy location and correction of the errors.

Because of the aforementioned advantages, increasing attempts have been made to develop approaches for measuring how confidently a word has been correctly recognized, or confidence measures. Various confidence measures have been reported in the literature. Most of them belong to the category of post-processing. Namely, the confidence is usually measured for hypothesized words following standard decoding [4, 6, 7, 8]. In [5], however, confidence measure is incorporated into search strategy and better recognition performance is reported.

In this paper, a new confidence measure is presented. It is based on rank-ordering subphone likelihood scores and it involves three steps. The first step is standard decoding which hypothesizes a word string for a given utterance. The second step is a forced Viterbi alignment, which is, of course, unnecessary for a Viterbi decoder. From the aligned sentence, the third step computes likelihood scores of the hypothesized subphone and all other competing subphones. A list of the subphones is generated in the descending order of the scores and a rank is assigned to the hypothesized subphone according to its positioning. The rank value is next merged to obtain the corresponding rank at the phone level. Selective weighting is then applied to de-emphasize contributions of phones known to have large acoustic variations. Additional upper-bound limiting is also made to help a rank computation not to be dramatically affected by a very bad segment. The resultant rank is compared with a preset threshold. If it is smaller than the threshold, the decoded word is considered to be confidently correctly recognized.

Subphone scores have been previously used in different ways. For example, the word likelihood score normalized by HMM state score was used in [8] for keyword spotting. Sukkar and Lee [7] used the difference between the likelihood score of the decoded subphone and the geometric mean of anti-subphones as the criterion function. As noted in the paper, the major improvement of utterance verification is due to the inclusion of spectral and duration information, in addition to HMM likelihood scores, and due to discriminative training ([7], p. 427). In the present method, no spectral/duration information nor discriminative training is resorted. Enhancement of verification performance is instead achieved via use of ranks and subsequent selective weighting.

2. THE ALGORITHM

As mentioned previously, the present algorithm involves regular decoding, Viterbi alignment, and rank computation. A brief description of the decoder is therefore useful to facilitate understanding of the new algorithm.

2.1. Stack Decoder And Viterbi Alignment

The recognizer used in this study is the IBM stack decoder [1]. Its front-end process consists of (i) preemphasis; (ii) computing FFT spectra every 10 ms using a 25 ms Hamming window; (iii) converting the mel-band output of the spectra to 12-dimensional cepstral coefficients, MFCC's; (iv) removing sentence-wise means of MFCC's (C_1

to C_{12}) and normalizing the energy term C_0; (v) computing first-order and second-order derivatives of MFCC's. Thus, the final acoustic vector for each frame constitutes 39 elements. The acoustic modeling includes 52 context-independent (CI) phones. Each of the phones are modeled with 3 left-to-right HMM arcs which are context-dependently (CD) trained. These arcs, representing subphone units, are actually terminal leaves of a decision tree. A speaker independent, large vocabulary speech recognition system typically contains 2,000 to 4,000 leaves. Each of the leaves is associated with a rank distribution histogram estimated from the training speech data. During recognition likelihood scores are computed using the histograms (cf. [1]).

From a stack decoder, word boundaries and word scores (e.g., fast match score, detail match score, and overall likelihood) are readily available. Word detail-match (DM) scores have previously been used as confidence measure in [4]. To obtain boundaries and corresponding scores of subword units, a forced Viterbi alignment against the decoded string needs to be performed.

2.2. Rank Computation

Given an utterance O, the task of speech recognition is to maximize the conditional probability, $Pr(W|O)$, where W denotes a word sequence. By Bayes' rule, we have

$$Pr(W|O) = \frac{Pr(O|W)Pr(W)}{Pr(O)}. \qquad (1)$$

However, because the denominator $Pr(O)$ is usually assumed to be a constant for any sequence of words (the so called non-informative prior), only the numerator is calculated in speech recognition. Hence the recognizer likelihood $Pr(O|W)Pr(W)$ is associated with W which may be wrong for the given utterance. To convert the likelihood to confidence measure, normalization is often resorted, such as duration normalization (cf. [4]), normalization using HMM state scores [8] or individual frame scores, and normalization using antimodels [6, 7, 3].

Different variants of antimodels have been used, for example, at word levels or subword levels. In the present work, competing HMM arcs are used as antimodels to accommodate rank computations. To illustrate, let us consider a single-phone example, /AA/. As stated above, it has 3 context-dependent, left-to-right HMM arcs: AA_1 AA_2 and AA_3. From Viterbi alignment, their respective segments S_i are known, so are their likelihoods, $Pr(O_{s_i}|AA\_i)$, where $i = 1, 2, 3$. For AA_1, that is the first arc of the HMM, the first arc of all other HMM's can be used as its antimodel and their likelihood score over the corresponding segment S_1 is computed and inserted into an array. The array is next sorted in the descending order and a rank is assigned to the hypothesized subphone, AA_1, according to its positioning in the sorted array.

There are at least two ways to create the array depending on whether the context is preserved or not. Assume that we have a set of CD subphones for another vowel AE_1. If they are all inserted into the array, a CD array is attained. On average the size of the array is a third of the number of terminal leaves of the decision tree and the array may amount to many hundreds of elements. But if a maximum operation is applied and only the CD subphone with the highest likelihood score is chosen and inserted to the array, a CI array is achieved. The CI array has a fixed size and is independent of training material. For the IBM system, the size is usually set to 156. A shortened CI array expedites sorting and at the same time we have found that it leads to a slightly better performance than using the full-size array. A possible explanation for the improved performance may be due to the fact that erroneous context hurts confidence measuring.

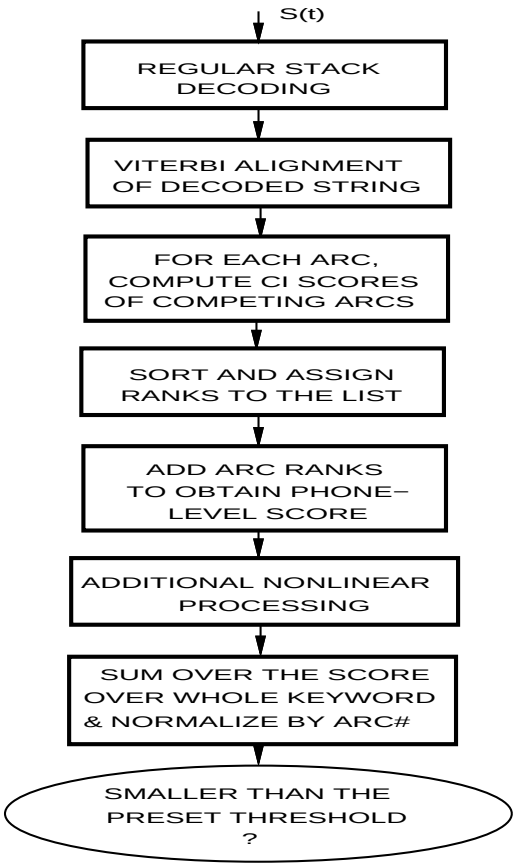

Figure 1. *Overall design of the new approach for computing subphone score-based confidence measure.*

Ranks for the second arc and the third arc can be determined in a similar manner. To ensure that a particularly bad segment will not dramatically contaminate the rank computation, individual ranks are examined to determine whether they exceed a preset value, R_0. As a result, individual rank values are within the range of 1 to R_0.

By adding up the arc ranks, the corresponding rank value at the phone level is attained. At this stage, selective weighting is carried out to de-emphasize those phones which are known to have a wide variability in articulation. These weighting rules are empirically designed. One of the rules, for example, is to reduce ranks of weak fricative

sounds of /f/, /v/, and /th/ if their ranks exceed a certain range. Ranks of stop consonants are found to vary considerably and they are weighted to have a narrowed variation.

The weighted ranks, R_{phone} are finally summed up and then normalized by the number of phones of the word, to produce the rank value of the word, R_{word}, namely the new confidence measure. The lower R_{word} is, the more confident it is that the word has been correctly decoded. When comparing the derived R_{word} with some preset threshold, a decision can be made to reject or accept the word hypothesis.

Figure 1 depicts the block diagram of the present algorithm. It should be noted at this point that the rank used in the confidence measure and the rank used for probability estimation described in [1] have similarities as well as differences. The reader is referred to [1] for more details on probability estimation using rank distribution histograms.

It is also noted that Sukkar and Lee [7] suggested a similar antimodels scheme. But their method differs from the present one at least in two aspects. One of the primary differences is that in their method all antimodel scores are pooled to a single score and that the difference between the hypothesized subword score and the averaged antimodels score is used. In the present algorithm, individual scores of competing antimodels are not pooled. On the contrary, they are maintained such that a ranking list can be attained. The second major difference is that they also used cohort models to minimize likelihood computation. By doing so, it is inherently assumed that the underlying subphone has been correctly recognized. This assumption is, of course, not always true. In the present algorithm, contextual information is ignored as much as possible (such as CI antimodels), because we want to make confidence estimate as much independent of decoding as possible. In other words, no assumption is made regarding the correctness of previous, current, and following subphones.

3. SPEECH DATABASE

The new confidence measure has been evaluated using an in-house speech database collected locally at IBM Watson Research Center. There are a total of 903 utterances from 8 adult speakers (both male and female) and each utterance is a single word. An Andrea microphone is used for data acquisition, and the sampling frequency is 22 kHz. The database comprises a small vocabulary of approximately 70 words. However, most words are short in duration and many are acoustically highly confusable. Table 1 shows a partial list of the words with confusing words being given inside the parenthesis.

4. EXPERIMENTAL RESULTS

In this section, two experiments are described. The first one pertains to regular decoding, and the other pertains to utterance verification. Because each utterance is known to contain only a single word, a simple finite-state-grammar with one word per path is used in the experiments. It is important to note that all vocabulary words are equally likely to occur and there is no additional prior language model knowledge applied. Consequently, the decoding is entirely based on acoustic scores.

I"ll	a	all
and	are (all)	at
by	cancel	come
eat	end (and)	go
has	have	here
in	is	it (eat)
left	let"s	likes
my (by)	of	okay
on (of)	that (at)	the
there (the)	they (a)	this (is)
to	who (to)	yes (mess)

Table 1. *Partial list of the vocabulary words.*

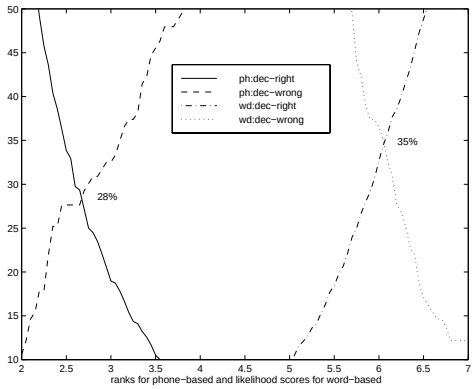

Figure 2. *ROC for regular decoding. The left-hand side is for the rank-based approach and the right-hand side is for the word score-based approach.*

4.1. Regular Decoding

In the first experiment, a word for each utterance is recognized with regular decoding. A word error rate of 16.1% is achieved for this task. Depending on whether the recognized word is correct or wrong, all utterances are divided into two groups. For each group, the new confidence measure is estimated. Figure 2 shows distribution of the new measure for each of the groups. The resulting plots are termed receiver's operating characteristic curves (ROC). The intersection point between the curves denotes "equal error rate."

For comparison, Figure 2 also shows the ROC curves for another confidence measure based on word DM scores [4]. It can be easily seen that the present rank-based approach outperforms the word score-based approach.

4.2. Utterance Verification

This utterance verification experiment is similar to more familiar speaker verification experiments (see e.g., [2]). An utterance is compared with the true word, and then with all other vocabulary words by rotation. The ROC

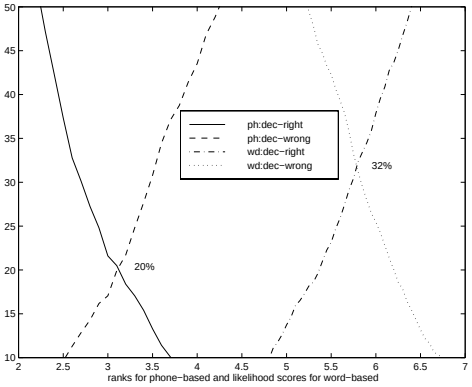

Figure 3. *ROC for utterance verification. The left-hand side is for the rank-based approach and the right-hand side is for the word score-based approach.*

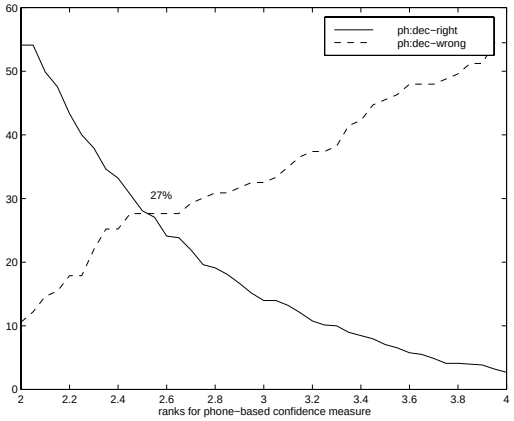

Figure 4. *ROC for utterance verification: Effect of nonlinear processing.*

curves for utterance verification are depicted in Figure 3. Again, it is clear that the present method is superior to the word score-based method. The equal error rate reduces from 32% to 20% for the new method, which corresponds to a relative reduction of 37%.

4.3. Effect of Nonlinear Processing

Figure 4 gives the ROC curves when all additional, nonlinear processes of upper-bound limiting and selective weighting are removed. By comparing Figures 3 and 4 for the case of utterance verification, it can be seen that the suggested nonlinear processing successfully drives down the equal error rate from 27% to 20%, or a reduction of 25%. For the regular decoding experiment, similar reduction in the equal error rate has also been obtained by the nonlinear processing.

The results in Figures 3 and 4 also reveal the advantages of the present rank-ordering approach over the word DM score based approach, even when the rank is directly utilized to compute the confidence measure (i.e., with no additional processing).

5. CONCLUSION

A new confidence measure has been proposed in the above and it has been favorably evaluated experimentally. The new measure is characterized by the following new features. It uses competing HMM arcs as the antimodels of the hypothesized subphone to generate a sorted array from which a rank is assigned to the hypothesized subphone according to its location in the array. Competing arcs can either (1) include all other arcs of the decision tree or (2) include only those that can appear in the same position of the given HMM topology, although only the second case has been treated here. Because correct context is not warranted when estimating confidence, context-independent (CI) subphone scores are utilized. The number of CI antimodels is usually much smaller than that of CD antimodels and hence, faster sorting can be achieved. At the same time we have found that use of CI antimodels is slightly superior to use of CD antimodels, suggesting that mistaken context may impair confidence estimation. It also suggests that one vote from each CI subphone unit is more effective than many votes from its CD variants.

Another new feature pertains to nonlinear processing of the obtained rank values to achieve more robust performance. The processes include upper-bound limiting and selective weighting of rank values. At present, the processes are based on several intuitive rules. In the future we would like to develop algorithms such as neural network computing for learning/adapting the weights.

REFERENCES

[1] Bahl, L., de Souza, P., Gopalakrishnan, P., Nahamoo, D., and Picheny, M., "Robust methods for using context-dependent features and models in a continuous speech recognizer," *IEEE-ICASSP94*, pp. 533-536.

[2] Che, C. and Lin, Q.: "Speaker recognition using HMM with experiments on the YOHO database," *Eurospeech 1995*, pp. 625-628, Spain.

[3] Dharanipragada, S. and Roukos, S.: "A fast vocabulary independent algorithm for spotting words in speech," *IEEE-ICASSP98,* pp. 233-236, Seattle.

[4] Lin, Q., Lubensky, D., Picheny, M., and Rao, S.: "Key-phrase spotting using an integrated language model of N-grams and finite-state grammar", *Eurospeech 1997,* pp. 255-258 Greece.

[5] Neti, C., Roukos, S., and Eide, E. "Confidence measures as a guide for stack search in speech recognition," *IEEE-ICASSP96*, pp. 883-887, Germany.

[6] Rahim, M., Lee, C.-H., Juang, B.-H.: "Discriminative utterance verification for connected digits recognition", *IEEE-Trans Speech Audio Proc. 5*, pp. 266-277, 1997.

[7] Sukkar, R. and Lee, C.-H.: "Vocabulary independent discriminative utterance verification for nonkeyword rejection in subword based speech recognition," *IEEE-Trans Speech Audio Proc 4*, pp. 420-429, 1996.

[8] Wilpon, J., Lee, C.-H., and Rabiner, L.: "Application of Hidden Markov models for recognition of a limited set of words in unconstrained speech," *IEEE-ICASSP89*, pp. 254-257. Scotland.