A NEW DECODER
BASED ON A GENERALIZED CONFIDENCE SCORE

Myoung-Wan Koo[*], *Chin-Hui Lee and Biing-Hwang Juang*

Multimedia Communications Research Lab
Bell Labs, Lucent Technologies
Murray Hill, NJ 07974-0636, USA
mwkoo@smm.kotel.co.kr, {chl,bhj}@research.bell-labs.com

ABSTRACT

We propose a new decoder based on a generalized confidence score. The generalized confidence score is defined as a product of confidence scores obtained from confidence information sources such as likelihood, likelihood ratio, duration, duration ratio, language model probabilities, supra-segmental information etc. All confidence information sources are converted into confidence scores by a confidence pre-processor. We show an extended hybrid decoder as an example of the decoder based on the generalized confidence score. The extended hybrid decoder uses multi-level confidence scores such as frame-level, phone-level, and word-level likelihood ratios, while the conventional hybrid decoder uses the frame-level confidence score. Experimental result shows that the extended decoder gives better result than the conventional hybrid decoder, particularly in dealing with out-of-vocabulary words or out-of-task sentences.

1. INTRODUCTION

In the area of decoding techniques for hidden Markov model(HMM), two kinds of search strategies have been studied[1]. The first one is the integrated approach in which recognition decision is made by jointly considering all the knowledge source. The second one is the modular approach proposed for modular design in which final decision is obtained by performing each modular in sequential manner.

One example of modular approach is utterance verification(e. g.[2]) consisting of two modules: recognition and verification. In the stage of recognition, knowledge sources such as likelihood and duration have been used for finding multiple hypotheses. In the stage of verification, knowledge source such as likelihood ratio has been used as a post-processor for rejecting unlikely hypotheses.

We propose a family of new decoders that is capable of handling a wide range of confidence scores generated from different knowledge sources. These scores help improve speech recognition in various ways. However, because their properties are often different, it is not easy to integrate these scores during decoding. Some examples of the scores being used in conventional decoding are *frame acoustic likelihood, frame acoustic likelihood ratio, phone and word duration penalties, word language probabilities, word insertion penalties, frame energy penalties, prosodic confidence scores,* etc.

It is obvious that there are three problems to utilize all the above confidence scores fully. First, the scores exhibit different levels of confidence in decision making in speech recognition. For example, the frame acoustic likelihoods are often more reliable that the duration statistics. Second, some confidence scores have a large dynamic range which often abruptly change search decisions locally in a beam search strategy. For example, the frame acoustic likelihood ratios vary much more rapidly than the frame acoustic likelihoods. Third, the confidence scores often need to be incorporated frame asynchronously. For example, the language model probabilities are typically folded into the overall confidence at the end of a word. Prosodic confidence scores will have to be incorporated differently from the frame energy scores, although they all represent supra-segmental information.

In this paper, we introduce a *generalized confidence score (G CS) function* that enables a framework to integrate different confidence scores. Three soultions are proposed to deal with the above three problems respectively. First the generalized confidence score combines various scores by exponential weighting. Second, we propose the use of a confidence pre-processor to transform some raw scores into manageble terms easier to integrate with other scores. Specifically, we use a sigmoid function limit the value of the log likelihood ratios. Third, the GCS function makes it easy to integrate confidence scores at all levels, including frame, state, phone, syllable, word, phrase and others.

2. GENERALIZED CONFIDENCE SCORE

2.1. Definition

The generalized confidence score $\Gamma_i(o_t)$ that an observation vector o_t is generated at state i of frame t, is defined to be a product of confidence scores as

$$\Gamma_i(o_t) = \prod_k \gamma_{ik}(o_t), \qquad (1)$$

where $\gamma_{ik}(o_t)$ is a k-th confidence score that the observation vector at frame t is shown at state i. The confidence scores come from many knowledge sources such as likelihood, likelihood ratio, duration, duration ratio, word language probabilities, word insertion penalties and prosodic confidence scores etc. We divide all knowledge into two kinds of knowledges:non-ratio and ratio. The non-ratio knowledge sources can be likelihood, duration, language probabilities and word insertion penalities etc., and the ratio knowledge sources can be likelihood ratio and duration ratio etc.

[*] This work has been carried out while visiting from Multimedia Technology Research Lab, Korea Telecom, Seoul, Korea during 1996-1997.

Each knowledge source can be changed into the confidence score by a confidence pre-processor.

We use two kinds of confidence pre-processors. The first one is for the non-ratio knowledge sources and uses a simple linear function. And the second one for the ratio knowledge sources is based on a sigmoid function. The reason that we use the sigmoid function is to reduce the dynamic range caused by the ratio knowledge source[3].

2.2. Decoding Algorithm

Our decoding algorithm is based on the generalized confidence score. If we define $\delta_j(t)$ as the best score along a single path at frame t, which accounts for the first t observations and ends in state j, induction rule gives us

$$\delta_j(t) = \max_i \{\delta_i(t-1) \cdot \Gamma_i(o_t)\}, \tag{2}$$

where $\Gamma_i(o_t)$ is the generalized confidence score at state i of frame t. If we take logarithm of the generalized confidence score, the generalized confidence score would be a sum of log confidence scores as

$$\log \Gamma_i(o_t) = \sum_k \log \gamma_{ik}(o_t). \tag{3}$$

There can be many kinds of knowledge sources for log confidence score. However, here we consider only three kinds of knowledge sources such as the likelihood, the duration, the likelihood ratio. If we consider the likelihood, the duration and the likelihood ratio sequently in Eq.(3), Then $\log \gamma_{i1}$ for the likelihood, $\log \gamma_{i2}$ for the duration, and $\log \gamma_{i3}$ for the likelihood ratio are, respectively,

$$\log \gamma_{i1}(o_t) = w_1[\log a_{i,j} + \log b_j(o_t)], \tag{4}$$

$$\log \gamma_{i2}(o_t) = w_2 \log \delta(d), \tag{5}$$

$$\log \gamma_{i3}(o_t) = w_3 \log \delta(CM), \tag{6}$$

where $\log \delta(d)$ is the log duration score at the end of a duration unit, and $\log \delta(CM)$ is the log confidence score by the log likelihood ratio, and w_1, w_2 and w_3 are weighting parameters for log confidence scores.

3. LIKELIHOOD RATIO CONFIDENCE

3.1. Multi-Level Confidence Scores

In the conventional hybrid decoder[3], we proposed the hybrid decoder using the confidence score by the likelihood ratio as

$$\log \delta(CM) = \log \frac{1}{1 + \exp(-\alpha \cdot (LLR - \beta))} \tag{7}$$

where β and α were location and weighting parameters. And LLR, which means log likelihood ratio, is defined as

$$LLR = \log \frac{a_{ij}^c b_j^c(o_l)}{a_{ij}^a b_j^a(o_l)}, \tag{8}$$

where a_{ij}^c, b_j^c are probabilities for the HMM model for the unit, and a_{ij}^a, b_j^a are probabilities for the anti-models[2]-[3] for the corresponding unit. However, this confidence score is a sinle-level confidence score calculated at every frame, which may be too sensitive to the change of LLR. We can find other confidence scores at the end of phone and word, respectively.

Here, we show three kinds of confidence scores based on the likelihood ratio. The first one is a frame-level confidence score, which is as same as in the conventional hybrid decoder[3]. The second one is a phone-level confidence score in which the confidence score is calculated at the end of each phone. Let the phone-level log likelihood ratio of phone n, LLR_n, be defined as

$$LLR_n = 1/\tau \sum_{t-\tau < l \le t} LLR_l, \tag{9}$$

where τ is the frame duration for phone n and LLR_l is the LLR at frame l. The phone-level confidence score can be obtained by replacing LLR in Eq.(7) with LLR_n in Eq.(9) as

$$\log \delta(CM_p) = \log \frac{1}{1 + \exp(-\alpha \cdot (LLR_n - \beta))}. \tag{10}$$

The final one is a word-level confidence score in which the confidence score is obtained at the end of each word. The word-level confidence score is obtained as

$$\log \delta(CM_w) = 1/N \sum_n \delta(CM_p(n)) \tag{11}$$

where N is the number of phones consisting of word w and $\delta(CM_p(n))$ is the phone-level confidence score of phone n.

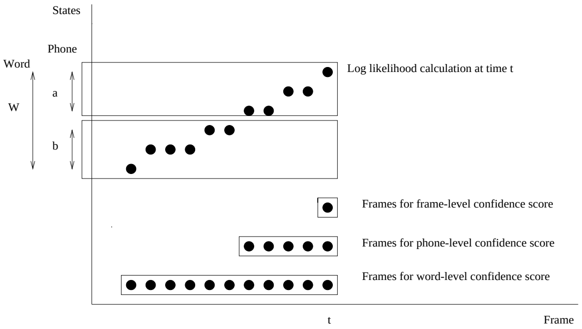
States
Phone
Word
W
a
b
Log likelihood calculation at time t
Frames for frame-level confidence score
Frames for phone-level confidence score
Frames for word-level confidence score
t
Frame

Figure 1: How to calculate log confidence scores

Figure 1 shows how to calculate the log confidence scores in three ways. Supposed that word ”W” has two phones: ”a” and ”b”, we could obtain a sequence of previous frames at frame t. Then we can calculate the multi-level log confidence scores:frame-level, phone-level and word-level as shown in Figure 1. The frame-level confidence score gives us a short-term information which result in frequent change every frame, while the word-level confidence score gives us the smoothed effect of information caused by anti-models. However, the word-level confidence score causes a big change of ordering in beam search at the end of word.

3.2. Extended Hybrid Decoder

The extended hybrid decoder uses the multi-level confidence scores instead of the frame-level confidence score which the conventional hybrid decoder has used. If we use the multi-level confidence scores in our decoder, Eq.(6) would be

$$\begin{aligned} \log \gamma_{i3}(o_t) &= w_{31} \log \delta(CM_f) + w_{32} \log \delta(CM_p) \\ &\quad + w_{33} \log \delta(CM_w), \end{aligned} \tag{12}$$

where w_{31}, w_{32} and w_{33} are the parameters for frame-level, phone-level and word-level log confidence scores, respectively. If w_{32} and w_{33} are set to be zero, Eq.(12) is as same as the conventional hybrid decoder[3]. The reason that we use the multi-level confidence scores is that we can exploit both the short-term and the long-term information.

4. SYSTEM OVERVIEW

4.1. Overview

Our system is a continuous speech recognition system based on continuous density hidden Markov models. Mixture Gaussian state observation density has a maximum of 8 mixture components per state. Each subword unit except the silence unit is modeled by a 3-state left-to-right HMM with no state skips. The silence unit has only one state. It differentiates from the previous system[3] in that we used the generalized confidence score during the recognition process. A brief explanation of each module is described in the following. Detection strategies[2] can also be incorporated.

4.2. Verification as Part of Recognition

Recognition is done by a frame synchronous beam search algorithm. The beam search algorithm evaluates the generalized score in Eq.(3) at every frame. The conventional likelihood decoder and the hybrid decoder are examples of the extended hybrid decoder, respectively. And the extended hybrid decoder is also one example of the decoder based on the generalized confidence score. Our grammar is based on a finite state grammar and each element at the grammar node is a key-phrase pattern[2].

We use two kinds of grammars: the rigid grammar and the loosed grammar[2]. The rigid grammar is built for decoding and the loosed grammar is built for detection. The main idea of detection is to use the loosed grammar for wider coverge during the recognition stage and to make semantic constraints in sentence parsing of key-phrase candidates to get a higer recognition rate[2].

4.3. Verification as a Post-Processor

Even though we use the generalized confidence score for verification in the recognition process, we can still add a verification module as a post-processor to our system[3]. This verification module use only the word-level log confidence score as a confidence measure. Phrase candidates which come from the recognition procedure can be rejected according to the confidence measure in the post-process. Accepted phrase candidates are rescored to be merged into the sentence hypotheses according to the sentence parsing algorithm[2].

5. EXPERIMENTAL RESULTS

5.1. Task-Independent Training

we use the set of right context-dependent(RCD) phone units as a universal phone set[4]. The total units we used are 1075 RCD + 42 context-independent(CI) phones. We also use the 40 CI anti-models.

5.2. Evaluation Results

We have evaluated our algorithms in a spoken dialogue system for a car reservation task. All the data were collected via telephone lines and spoken by the general public. For evaluation, we define the semantic accuracy in the same way as the word accuracy[2].

For a careful analysis, we classified the sample utterances into three categories. In-grammar(ING) sentences consists of defined phrases only and are covered by the conventional finite-state sentence grammars. Out-of-grammar(OOG) sentences have out-of-vocabulary or fragmental words. They can be interpreted for proper action but are usually not accepted by the sentence grammars. Out-of-task(OOT) sentences contain no key-phrases and should be rejected. We use the TIME sub-task of Car Reservation for the preliminary evaluation. The TIME sub-task has 51 key words and many out-of-vocabulary words.

Figure 2 shows a comparative result with regard to various methods to generate the generalized confidence score. The followings are some methods we have considered. All methods are based on the selective use of anti-models we proposed[3].

1. Hybrid 1(Frame-level); This is as same as the conventional hybrid decoder we proposed[3] (w_1=0.8 ,w_2=1, w_{32}=w_{33} =0, w_{31}=1). The log confidence score is added to the log likelihood score at every frame.
2. Hybrid 2(Phone-level); This is the hybrid decoder using phone-level confidence score(w_1=w_2=1, w_{31}=w_{33}=0, w_{32}= 1). The phone-level log confidence score is added to the log likelihood score at the end of each phone.
3. Hybrid 3(Word-level); This is the hybrid decoder using the word-level confidence score(w_1=w_2=1, w_{31}=w_{32}=0, w_{33}= 1). The word-level log confidence score is added to the log likelihood score at the end of each word.
4. Extended hybrid 1(Frame-level + Word-level); This is the hybrid decoder using both the frame-level and the word-level confidence scores simultaneously (w_1=0.8, w_2=1, w_{33} =0, w_{31}=w_{33}=0.6). The frame-level and the word-level log confidence scores are added to the log likelihood score.
5. Extended hybrid 2(Frame-level + Word-level + Phone-level); This is the hybrid decoder using the frame-level, the word-level and the phone-level confidence scores simultaneously (w_1=0.8, w_2=1, w_{31}=w_{33}=0.4, w_{32}=0.3). The frame-level, the phone-level and the word-level log confidence scores are added to the log likelihood score.

The results show that the extended hybrid decoder gives better recognition rate than the conventional hybrid decoder and that the extended hybrid 2 gives the best recognition rate when the post-processor is added. However, we choose the extended hybrid 2 because of fast computation and optimization. We also compared the proposed extended hybrid decoder with the likelihood decoder (conventional Viterbi decoder). Table 1 shows the comparative results of the likelihood decoder(w_1=w_2=1, w_{31}=w_{32}=w_{33}=0), the likelihood ratio decoder[5](LLR: w_{31}=1, w_1=w_2=0, w_{32}=w_{33}=0) and conventional hybrid decoder(Hybrid 1: w_1=0.8, w_2=1, w_{31}=1, w_{32}=w_{33}=0) with the proposed hybrid decoder(Extended hybrid 1:w_1=0.8, w_2=1, w_{31}=0.6, w_{33}=0.6, w_{32}=0) in detail.

We also compared the results when we used the loosed grammar for detection. Table 2 shows the results for DATE when we used the loosed grammar. And Table 3 shows the results for the LOCATION sub-task when the rigid grammar and loosed grammar are respectively used. DATE sub-task has 99 key words and

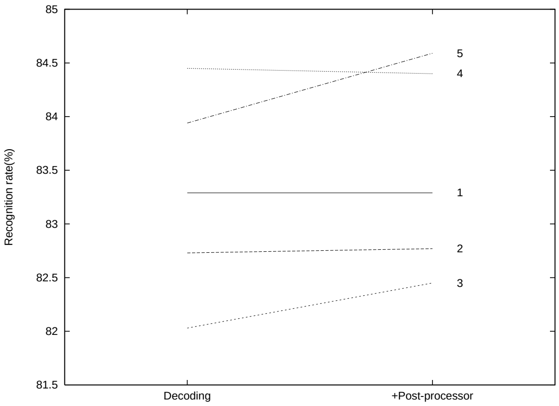

Figure 2: Comparative results with regard to various methods

Table 1: Recognition rate for TIME sub-task

	ING	OOG	OOT	Total
Sentences (Semantic slots)	818 (1895)	110 (190)	63 (63)	991 (2148)
Decoding				
Likelihood	87.81%	11.05%	26.98%	79.23%
LLR	82.16%	35.26%	41.26%	76.82%
Hybrid 1	87.65%	43.68%	71.43%	83.29%
Extended Hybrid	88.71%	45.26%	74.60%	84.45%
+Post-processor				
Likelihood	87.70%	25.79%	47.62%	81.05%
Hybrid 1	87.34%	45.79%	74.60%	83.29%
Extended Hybrid	88.44%	46.32%	77.78%	84.40%

LOCATION sub-task has 371 key words. In addition, many out-of-vocabulary words are also observed.

All results indicate that the extended hybrid decoder results in a better recognition rate than the conventional hybrid decoder or likelihood decoder alone. The addition of verification as post-processors also improves performance in all system configurations. The main advantage of the extended hybrid decoder is that the recognition rate of ING is not so much deteriorated by including the post-processor.

6. CONCLUSION

In this paper, we have presented a new decoder based on the generalized confidence score which is the product of confidence scores obtained from confidence information sources such as likelihood, likelihood ratio, duration, duration ratio, word language probabilities, prosodic confidence scores, etc. The extended hybrid decoder has been presented as one example of the proposed decoder, which use the multi-level confidence score obtained from the likelihood ratio. We have also shown that the conventional likelihood decoder and the conventional hybrid decoder are respectively some examples of the extended hybrid decoder. The experimental result shows that the extended hybrid decoder with the post-processor gives better results than the likelihood decoder or the hybrid decoder alone.

Table 2: Recognition rate for DATE sub-task

	ING	OOG	OOT	Total
Sentences (Semantic slots)	1123 (2444)	154 (310)	91 (91)	1368 (2845)
Detection				
Likelihood	92.31%	57.42%	18.68%	86.15%
Hybrid 1	92.06%	60.00%	27.47%	86.50%
Extended Hybrid	92.14%	58.71%	29.67%	86.50%
+Post-processor				
Likelihood	92.02%	70.65%	43.96%	88.15%
Hybrid 1	92.06%	72.26%	50.55%	88.58%
Extended Hybrid	92.27%	71.94%	48.35%	88.65%

Table 3: Recognition rate for LOCATION sub-task

	ING	OOG	OOT	Total
Sentences (Semantic slots)	681 (1025)	99 (137)	131 (131)	911 (1293)
Decoding				
Likelihood	94.24%	16.06%	25.95%	79.04%
Hybrid 1	94.05%	21.90%	26.72%	79.58%
Extended Hybrid	93.95%	22.63%	25.19%	79.43%
+Post-processor				
Likelihood	93.76%	23.36%	45.80%	81.44%
Hybrid 1	93.66%	30.66%	43.51%	81.90%
Extended Hybrid	93.56%	29.20%	47.33%	82.06%
Detection				
Likelihood	93.07%	38.68%	20.61%	79.97%
Hybrid 1	93.27%	43.07%	22.90%	80.82%
Extended Hybrid	92.88%	43.80%	21.37%	80.43%
+Post-processor				
Likelihood	92.39%	59.85%	28.24%	82.44%
Hybrid 1	92.68%	63.50%	29.77%	83.22%
Extended Hybrid	92.29%	64.23%	32.82%	83.29%

7. REFERENCES

[1] C.-H. Lee, F.K. Soong and K.K. Paliwal, *Automatic Speech and Speaker Recognition;Advanced Topics, Chapter 1* Kluwer Academic Publishers, MA, 1996.

[2] T. Kawahara, C-H Lee, and B-H. Juang. “Combining key-phrase detection and subword-based verification for flexible speech understanding,” *Proc. IEEE-ICASSP*, volume 1, pp 193–196, 1997.

[3] M.-W. Koo, C.-H. Lee and B.H. Juang, “A new hybrid decoding algorithm for speech recognition and utterance verification,” *IEEE Workshop on Speech Recognition and Understanding*, To be published, Dec. 1997.

[4] C.-H Lee, B-H Juang, W. Chou and J.J. Molina-Perez. “A study on task-independent subword selection and modeling for speech recognition,” *Proc. ICSLP*, pp 1816–1819, 1996.

[5] E. Lleida and R.C. Rose. “Efficient decoding and training procedures for utterance verification in continuous speech recognition,” *Proc. IEEE-ICASSP*, pp 507–510, 1996.