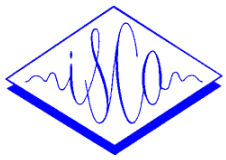
ISCA Archive
http://www.isca-speech.org/archive



6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

THE USE OF DYNAMIC RELIABILITY SCORING IN SPEECH RECOGNITION[1]

Xiaolong Mou and Victor Zue

Spoken Language Systems Group
Laboratory for Computer Science
Massachusetts Institute of Technology
200 Technology Square,Cambridge, Massachusetts 02139, USA
{mou, zue}@sls.lcs.mit.edu

10.21437/ICSLP.2000-856

ABSTRACT

Typically, along a recognizer's search path, some acoustic units are modeled more reliably than others, due to differences in their acoustic-phonetic features and many other factors. This paper presents a dynamic reliability scoring scheme which can help adjust the partial path scores while the recognizer searches through the composed lexical and acoustic-phonetic network. The reliability models are trained on the acoustic scores of the correct arc and its immediate competing arcs extending the current partial path. During recognition, if, according to the trained reliability models, an arc can be more easily distinguished from the competing alternatives, that arc is more likely to be in the right path, and the partial path score can be adjusted accordingly on the fly to have a more accurate path hypothesis. We have applied this reliability scoring mechanism in two weather related domains, JUPITER [6] (for English) and PANDA (a predecessor of MUXING [5] for Mandarin Chinese). We get 9.8% word error rate (WER) reduction in the JUPITER domain and 12.4% WER reduction in the PANDA domain, thus demonstrating the effectiveness of this approach.

1. INTRODUCTION

Speech recognition is formulated as a problem of searching for the best string of symbols, subject to the constraints imposed by the acoustic and language models. In implementing such a formulation, systems typically apply the constraints uniformly across the entire utterance. This does not take into account the fact that some units along the search path may be modeled and recognized more reliably than others, perhaps due to differences in their acoustic-phonetic characteristics, the particular feature extraction and modeling approaches the recognizer chooses, and the amount and quality of available training data. One possible way to incorporate reliability information is through word- and utterance-level rejection [4]. However, this approach generally provides confidence information after the recognition phase, and as such the confidence score is usually measured from a set of chosen features [2], most of which are obtained after the recognition is done. In contrast, this work attempts to incorporate reliability information directly into the search phase in order to help the recognizer find the correct path.

In this paper, we introduce the notion of dynamic reliability scoring that adjusts the path score according to the trained reliability models while the recognizer searches through the composed lexical and acoustic-phonetic network. In our scheme, the recognizer evaluates the reliability of a hypothesized arc extending the current path by adding a weighted reliability score to the current path score. With more accurate path evaluation, we can derive two immediate benefits. First, the overall path score now reflects a more realistic probability measurement of the whole path; thus a path with higher score is more likely to be a correct path. Second, with necessary pruning to balance accuracy and complexity, unpromising partial paths are pruned according to their current scores. However, it is crucial to have a good estimation of partial paths, because pruning errors are not recoverable once they happen.

In the next sections, we elaborate on the details of constructing, training and applying the reliability models. We also describe some related issues such as back-off models and iterative training. The evaluation of reliability scoring is conducted on two weather information domains, JUPITER [6], which is in English, and PANDA, which is a predecessor of MUXING [5] in Mandarin Chinese. The experimental results, conclusions and future work are also presented.

2. RELIABILITY MODELS

In this section we will describe the reliability models and the training procedures we use. The notion of reliability here refers to how confident we are while choosing a hypothesized arc to extend the current partial path. It could happen that a high scoring arc is actually not in the correct path, particularly if its immediate competitors have similar high scores. In this case we are less confident to choose it, even though it has a high score. On the other hand, an arc with relatively low acoustic score is very likely to be the right one to extend the current path if its competitors have much lower scores. We build separate Gaussian models to describe the correct arc scores and the competing arc scores for each arc in the lexical network, and then use these models to help the recognizer decide which arc to choose during recognition time.

2.1. Lexical and Acoustic-Phonetic Networks

The recognizer we use in this work is the MIT SUMMIT [1] segment based recognizer. A segment-based recognizer usually has a lexical network and an acoustic-phonetic network. The lexical network is constructed from the recognizer's vocabulary. Each word in the vocabulary is represented by a pronunciation network and these networks are combined into a single lexical net-

[1]This research was supported by a contract from Industrial Technology Research Institute.

work by connecting word end nodes and word start nodes that satisfy the inter-word pronunciation rules. This network provides strong constraints on the phone sequences the recognizer can choose from. The acoustic-phonetic network is built from the input speech signal. It provides possible segmentations of the speech signal and the corresponding segment or boundary features to obtain acoustic scores from the acoustic models. The recognizer is a finite-state transducer based recognition system; its search space is define by composing two finite state transducers:

$$P \circ L \tag{1}$$

Where P is the acoustic-phonetic transducer that maps speech signals to phones with acoustic scores, and L is the lexical transducer that maps phones to words. The recognizer will search through the composed network, find the best path, and give the resulting word sequence. Language model scores are also applied during the search by composing another finite state transducer G, which specifies the probability of word transitions. However, since we are focusing on acoustic model reliabilities, we separate language model scores from the acoustic model scores while training the reliability models.

2.2. Reliability Models

Our phonetic reliability measurement is obtained from a reliability model that gives the likelihood of extending the current candidate path using one specific arc as opposed to using its immediate competing alternatives in the composed network. The current partial path score is then adjusted according to this reliability measurement. The reliability models are trained from transcribed speech data. First a forced alignment search is conducted using the known orthography and current acoustic models, and the results are used as references to correct paths. Then, for each partial path along the forced path, the score of the arc extending the forced path, denoted s, and the scores of the arcs that are not in the forced path, denoted $t_1, t_2, \ldots, t_n$, are collected. After that, for each arc in the lexical network, Gaussian models for the correct scoring (i.e., scores of corresponding arcs that are in the forced path from the composed acoustic-phonetic network), M_s, and incorrect scoring (i.e., scores of corresponding arcs not in the forced path), M_t, are trained.

An important aspect of our approach is that all the acoustic scores used to train the reliability models are normalized by a "catch-all" model. We use the normalized log-likelihood (NLL) scoring, and the acoustic score is given by:

$$\log(p(x|\omega_i)/p(x)) \tag{2}$$

where x is the feature observation, $p(x|\omega_i)$ is the probability density of x given the class model ω_i, and $p(x)$ is the catch-all normalization model defined as:

$$p(x) = \sum_{j=1}^{N} p(x|\omega_j)P(\omega_j) \tag{3}$$

The NLL score is expressed in the log domain and can be viewed as a zero-centered score. An acoustic score greater than zero represents a greater than average possibility that the feature observation belongs to the hypothesized class. The use of normalized acoustic scores ensures that the reliability models are built from acoustic scores that are comparable across different acoustic models.

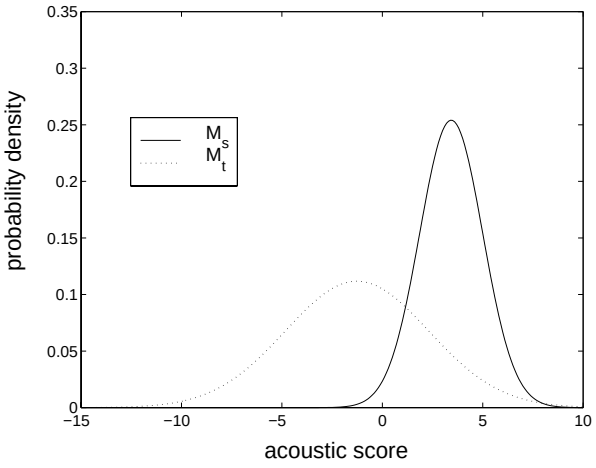


Figure 1: The reliability models M_s (correct scoring) and M_t (incorrect scoring) trained for the arc labeled [t] in the lexical network of the word "want".

Figure 1 shows example reliability models M_s and M_t trained for a given arc in the lexical network. We can see that M_s is generally centered at a score greater than zero with smaller variance, while M_t is centered at a score less than zero with greater variance. The further apart the two models and the smaller their variances are, the easier it would be to distinguish the correct arc from its immediate competing alternatives while searching.

3. DYNAMIC RELIABILITY SCORING

In this section we will describe the application of the trained reliability models and some related issues. After training the reliability models for each arc in the lexical network, we can use these models to obtain the probability of extending the current path using a hypothesized arc with acoustic score s. This probability is then used as the reliability score to adjust the current partial path score. For an arc with high acoustic score, if its immediate competitors usually have similar high scores, the reliability scoring will give a less confident result. On the other hand, for an arc with a low acoustic score, if its immediate competitors usually have even lower scores, the reliability scoring will give a more confident result.

3.1. Application of Reliability Models

After all the models are trained, we can obtain reliability measurements on the fly while searching through the network. The reliability measurement is essentially the likelihood that a particular arc in the network with acoustic score s is in the right path while its immediate competitors with acoustic scores $t_1, t_2, \ldots, t_n$ are not in the right path. This probability is given by the following formula, assuming the correct path and the com-

peting alternative paths are independent of each other:

$$\begin{aligned} & p(s|M_s)p(t_1,t_2,\ldots,t_n|M_t) \\ &= p(s|M_s)\prod_{i=1}^{n} p(t_i|M_t) \\ &= \frac{p(s|M_s)p(s|M_t)\prod_{i=1}^{n} p(t_i|M_t)}{p(s|M_t)} \\ &= \frac{p(s|M_s)}{p(s|M_t)} p(s,t_1,t_2,\ldots,t_n|M_t) \end{aligned} \quad (4)$$

Because we use the reliability score to help the recognizer choose an arc hypothesis from its *immediate* competing arcs, $p(s,t_1,t_2,\ldots,t_n|M_t)$ is a constant factor in this case, and we can just use the log domain score $\log(p(s|M_s)/p(s|M_t))$ as the reliability measurement, which saves a lot of computation effort during search.

The log domain reliability score is combined to the current partial path score to adjust the current path ranking. This will help reduce the pruning errors, and the path with better overall score is more likely to be the correct path.

3.2. Back-off Models

Generally, the models M_s and M_t are trained for each arc in the lexical network. However, due to sparse data problem, some arcs in the lexical network may not have enough data to train these models. To avoid this problem, we have established two-level back-off models, namely the phonetic back-off model and the generic back-off model. If the original arc-specific model is not well trained, we will use the corresponding phonetic back-off model, which is trained by combining the data for all the arcs bearing the same phone label. If this phonetic back-off model is still not well trained, we will use the generic back-off model, which is trained from all the data available regardless of their corresponding lexical arcs or phone labels.

Currently the weights for the original lexical arc specific models, the corresponding phonetic back-off models and the generic back-off model are controlled by smoothing factors r_1 and r_2, according to the amount of data available for training:

$$r_1 \frac{p(s|M_s)}{p(s|M_t)} + (1-r_1)\left(r_2 \frac{p(s|M_s^P)}{p(s|M_t^P)} + (1-r_2)\frac{p(s|M_s^G)}{p(s|M_t^G)}\right) \quad (5)$$

where M, M^P and M^G are the original arc-specific model, phonetic back-off model and generic back-off model, respectively; r_1 is the ratio between the amount of training data available for an arc-specific model and its corresponding phonetic back-off model; r_2 is the ratio between the amount of training data available for a phonetic back-off model and the generic back-off model.

3.3. Iterative Training

Since the reliability scores can be used to adjust the partial path scores and obtain more accurate phonetic transcription results for the training data, we can improve the acoustic and reliability models iteratively. Given the training data orthography, a forced search is first conducted to transcribe the data according to current acoustic and reliability models. Then the newly labeled data are used to re-train the acoustic and reliability models.

In practice, we find that the recognition performance converges quickly after a few iterations. More details are given in the experimental results in section 5.

4. CORPUS

The recognizer's acoustic models and reliability models are trained and evaluated in an English weather information domain called JUPITER and a Mandarin Chinese weather information domain called PANDA, a predecessor of MUXING [5]. For the JUPITER domain, the training set consists of 24,182 live utterances recorded over the phone and the test set consists of 1,806 utterances randomly selected from the data collection independent of the training set. Both boundary and segment models are used, and the reliability models are built on the normalized and combined boundary and segment acoustic scores. For the PANDA domain, the training set consists of 1455 utterances and the test set consists of 244 utterances. Due to insufficient data for training boundary models, only segment models are used. The reliability models are built on the normalized segment scores. There are no out-of-vocabulary (OOV) words in the training or test set in either the JUPITER or the PANDA domain.

5. EXPERIMENTAL RESULTS

We have incorporated the reliability scoring scheme into the segment-based, SUMMIT [1] speech recognition system, which can use both boundary acoustic models and segment acoustic models when enough training data are available.

Iteration Number	*WER without Reliability Models(%)*	*WER with Reliability Models(%)*	*Relative WER Reduction (%)*
1	12.1	10.1	16.5
2	10.3	9.2	10.7
3	10.2	9.2	9.8

Table 1: The recognition results in the JUPITER domain on a 1,806 utterance test set.

Iteration Number	*WER without Reliability Models(%)*	*WER with Reliability Models(%)*	*Relative WER Reduction (%)*
1	11.9	9.8	17.6
2	10.4	8.9	14.4
3	9.7	8.5	12.4

Table 2: The recognition results in the PANDA domain on a 244 utterance test set.

Tables 1 and 2 show the recognizer's performance before and after applying the reliability models for 3 iterations of training.

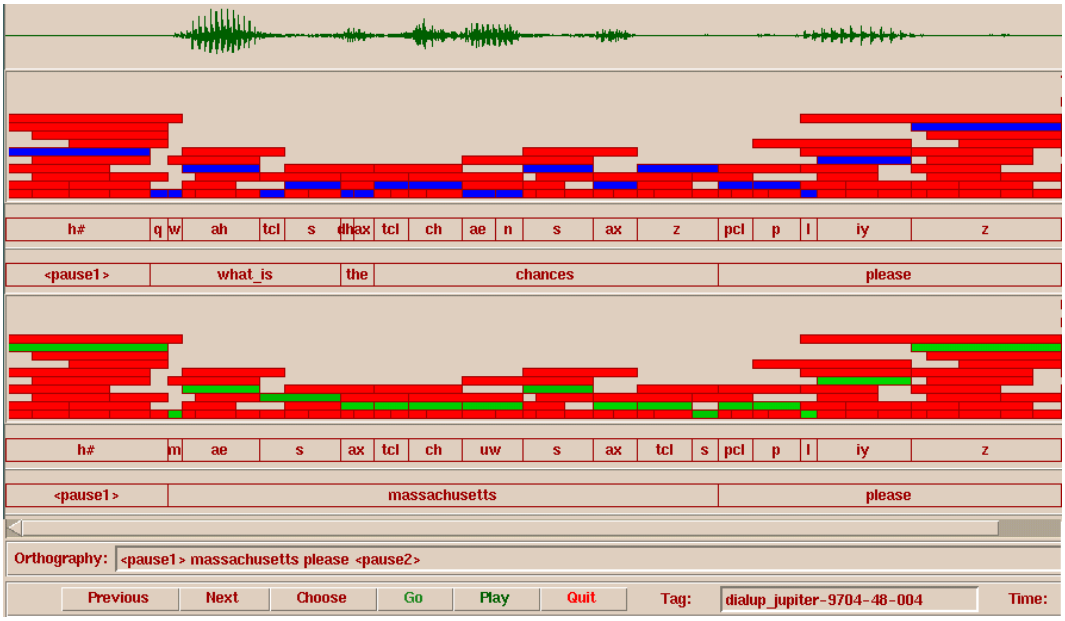


Figure 2: Example of the recognition result with and without reliability models in the JUPITER domain. The upper panel shows that, without reliability models, the utterance is incorrectly recognized as "what is the chances please". The lower panel shows the result using the reliability models, and the utterance is correctly recognized as "Massachusetts please".

In the JUPITER domain, after 3 iterations of training, 9.8% word error rate (WER) reduction is achieved on a 1,806 utterance test set using the reliability models. In the PANDA domain, 12.4% WER reduction is achieved on a 244 utterance test set using the reliability models, also after 3 iterations of training.

Figure 2 shows an example comparing the search results with and without the reliability models. As can be seen, the reliability models have corrected the hypothesized path and correctly identified the word sequence "Massachusetts please".

6. CONCLUSIONS AND FUTURE WORK

The work described in this paper demonstrates that reliability models can be used to address the fact that acoustic units along the search path are generally modeled and recognized with different reliability, and we can use the reliability score given by the reliability models during the search to help early recovery of search errors.

Currently the reliability models are used for the immediate future of the current path, helping the recognizer choose the best-hypothesized arc from its competitors. Ideally, we would like to provide the reliability measurements based on all the possible future arcs extending the current path. Since this may lead to combinatorial explosion, a compromise may be to use a combination of reliability scores within a certain interval beyond the current node. This would help eliminate the bias of considering the immediate future context only, thus giving better reliability adjustments.

With a given recognition lexicon, it is usually not necessary to accurately recognize every acoustic unit in the word to get correct recognition results. It is possible to use only several reliable pieces of the word to distinguish it from other words. Future work includes trying to obtain such reliable pieces with the guidance of the reliability measurements, and changing the lexical access scheme from precise matching to reliable-island matching. This has the advantage of modeling complex phonological variations implicitly, and can potentially deal with the non-native speech [3] problem as well as the out-of-vocabulary word problem better.

7. REFERENCES

1. J. Glass, J. Chang, and M. McCandless, "A probabilistic framework for feature-based speech recognition," in *Proc. ICSLP'96*, Philadelphia, 1996.

2. S. O. Kamppari and T. J. Hazen, "Word and phone level acoustic confidence scoring," in *Proc. ICASSP'00*, Istanbul, Turkey, 2000.

3. K. Livescu and J. Glass, "Lexical modeling of non-native speech for automatic speech recognition," in *Proc. ICASSP'00*, Istanbul, Turkey, 2000.

4. C. Pao, P. Schmid, and J. Glass, "Confidence scoring for speech understanding systems," in *Proc. ICSLP'98*, Sydney, 1998.

5. C. Wang, S. Cyphers, X. Mou, J. Polifroni, S. Seneff, J. Yi, and V. Zue, "A telephone-access mandarin conversational system in the weather domain," in *these proceedings*.

6. V. Zue, S. Seneff, J. Glass, J. Polifroni, C. Pao, T. Hazen, and L. Hetherington, "JUPITER: A telephone-based conversational interface for weather information," *IEEE Trans. on Speech and Audio Processing*, vol. 8, no. 1, pp. 85–96, Jan. 2000.