A SENONE BASED CONFIDENCE MEASURE FOR SPEECH RECOGNITION

Z. Bergen
Berdy Medical Systems
4909 Pearl East Circle, Suite 202
Boulder, Colorado, USA 80301
Tel. 303-417-1603, FAX 303-417-1662, E-mail: zbergen@berdy.com

W. Ward
Carnegie Mellon University
Pittsburgh, PA, USA 15213
E-mail: whw@cs.cmu.edu

ABSTRACT

This paper describes three experiments in using frame level observation probabilities as the basis for word confidence annotation in an HMM speech recognition system. One experiment is at the word level, one uses word classes, and the other uses phone classes. In each experiment we categorize hypotheses into correct and incorrect categories by aligning a best recognition hypothesis with the known transcript. The confidence of error prediction for each class is a measure of the resolvability between the correct and incorrect histograms.

1. INTRODUCTION

Speech recognition systems generally rank order hypotheses by computing scores for utterance hypotheses. These scores are useful for preference ordering the hypotheses, but do not give a good indication of the quality of the recognition or how confident the system is that the decoding is correct. For applications to act on speech input, they must be able to assess the confidence that the input has been decoded correctly. This work combines and extends the work described in [1], [2], and is related to extending one feature of [3] for providing confidence annotation of speech recognizer output. The idea is to normalize decoded word strings and phone acoustic scores by scores produced by a less constrained search. [1] used an all-phone recognition to normalize the scores of the hypotheses, followed by Bayesian updating. Among other things, [3] also used the best matching observation for each frame (senone) to normalize the acoustic score for the hypothesis. This paper describes further experiments with this measure.

For our acoustic measure we use 10ms frame-level observation scores as the basis for the normalization. We use the Sphinx-II system [4] as our speech recognizer. It is a Semi-Continuous HMM recognizer using a trigram language model. Acoustic observations are modeled in this system by senones [5]. Senones are tied hmm-state specific mixture weights for the Gaussian distributions used by the semi-continuous HMM system. For each 10ms frame of input, the recognizer compares the input feature vector to all senones in the system. The best scoring senone for that frame is recorded, this being the unconstrained match. After the recognizer has produced the best word string (using a Viterbi search), these scores are used to normalize the scores of the words and phones in the hypothesis. For each frame, the score of the senone used by the hypothesis for that frame is subtracted from the best scoring senone for the frame. The average of this normalized score is then computed for each word and for each phone of each word.

Chase [3] used this measure as one predictor feature in a decision tree for confidence annotation. The acoustic scores of both words and phones were normalized by the best senone path. Used directly as a predictor feature, this measure seemed to have relatively little predictive power. We investigate the further classification into word and phone classes respectively, in hopes of improving the discrimination power of this measure.

2. EXPERIMENTS

Three experiments were performed to determine the utility of using these normalized acoustic scores for word and phone level confidence measures. The categories for the three cases are:

- all words
- word classes
- phone classes

Each class is divided into a correct and incorrect set so the distributions for each can be compared.

2.1 Experiment 1

We tested the measure by computing histograms of correct and incorrect words from a development corpus. The recognizer was run on 1000 utterances from the Wall Street Journal Corpus and the confidence measure computed for each word. The distributions of the confidence scores were computed for correct words and incorrect words as determined from the reference

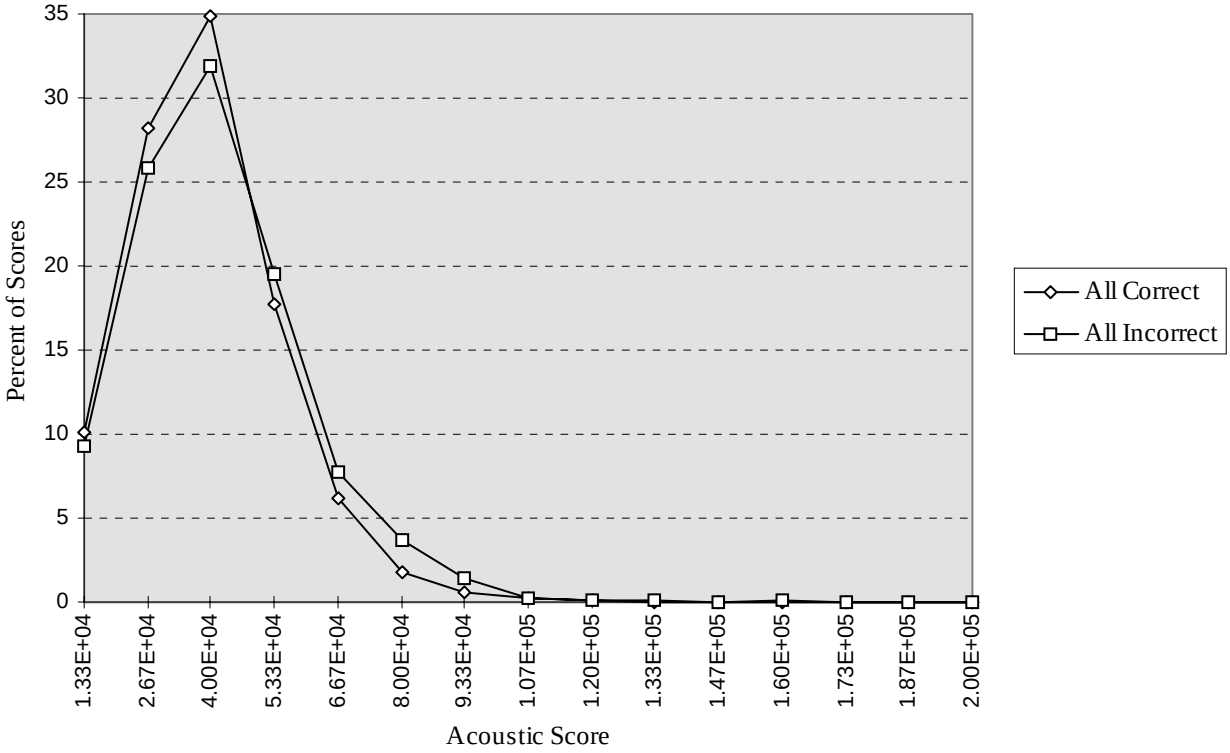

Figure 1. Correct and incorrect distributions for all words.

transcripts. An alignment program was used to flag incorrect words where the hypothesis decoding differed from the transcript.

Figure 1 shows the distributions and illustrates the high degree of overlap of the two distributions. These results are consistent with results for the similar measure described in [3]. Much accuracy is probably lost in our confidence measure by averaging across all words.

2.2 Experiment 2

The results of the first experiment led us to cluster words into classes and evaluate using our acoustic measure. It was hoped that clustering words would uncover variations hidden by averaging across all words. We formed the following classes of phones:

- Vowels (AE, EH, IH, IX, IY, UW, OW, UH, AH, AX, AA, AO, ER, AXR)
- Dipthongs (AW, AY, EY, OY)
- Orals (B, D, G, DX, BD, DD, GD, P, T, K, PD, TD, KD)
- Fricatives (DH, Z, ZH, V, S, TH, SH, F)
- Affricates (CH, TS, JH)
- Nasals (M, N, NG)
- Aspirates (HH)
- Approximants (W, R, L, Y)

Using these phone groups we formed word classes by looking at the beginning phone and total number of phones for each word. While not optimal, this classification results in distributions that exhibit areas of error prediction. Phone level differences are averaged out over the length of the word and their effects may not appear as predominantly as in the single phone case described in the next section. Normalization was performed independently for each of the classes.

Figure 2. Shows the distributions for words starting with a dipthong. We can see that there is some variation in the incorrect distribution as compared to the correct distribution which remains similar in shape to the all word case from Figure 1. In general, separation of the correct and incorrect distributions did improve slightly with the more specific statistics.

2.3 Experiment 3

In this experiment we investigated the phone level for a more specific model of the behavior of acoustic scores. To prepare the data we used the Sphinx-II decoder to produce phone level segmentations and scores for the best path hypothesis. We form 50 classes comprised of the individual phones from our previous phone classes (see section 2.2). Normalization is done by averaging the difference between the constrained and unconstrained path over each phone.

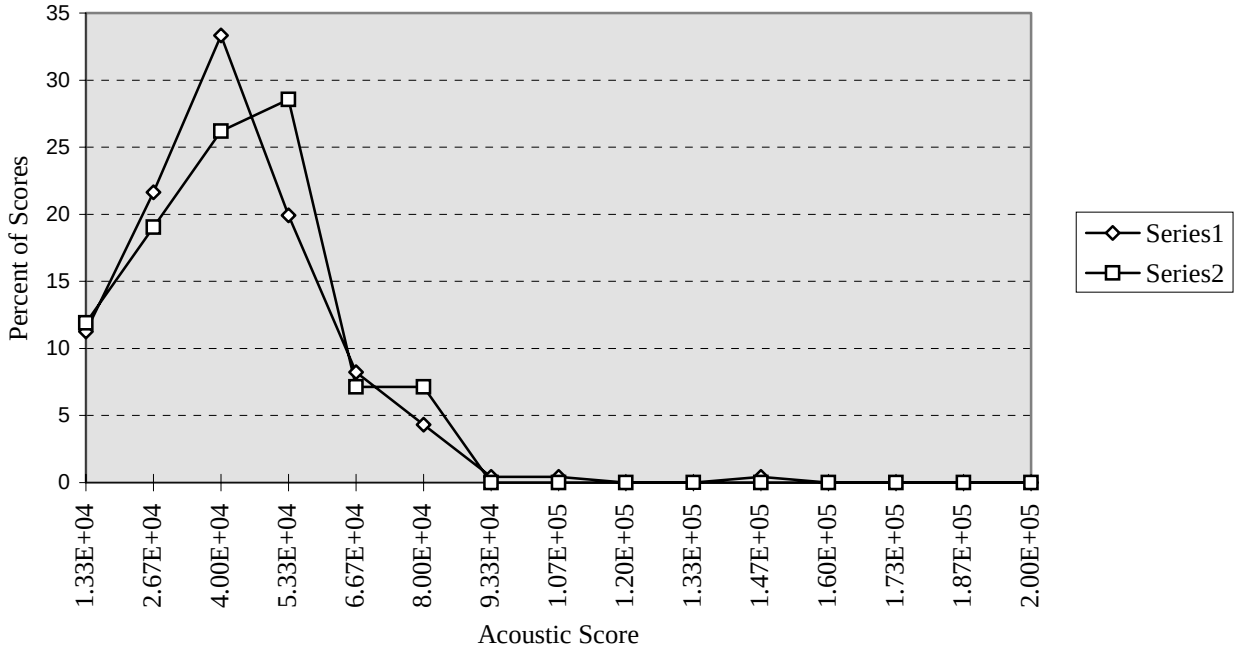

Figure 2. Correct and incorrect distributions for words beginning with dipthongs (Series 1:correct, Series 2: incorrect).

For some phones, the measure shows a significant degree of separation between the correct and incorrect distributions. Figure 3 is an example of how the distributions look for one of the phone classes: UW. Comparing Figure 3 with Figures 1 and 2 note that the distribution for correct scores remains fairly constant while the distribution for incorrect scores spreads over the range of scores providing a distinct region of separation between the distributions. For the more general classes of the prior experiments, the overlap in the distributions is due to large localized differences in the single phone classes that get averaged out in the word level classes. While hidden in the general statistics, for the single phone case, it is possible that a misrecognized phone may cause the recognizer to traverse the lexical tree along the wrong word path and cause a word level error.

3. CONCLUSION

Senone based acoustic normalization seems to provide only very slight information for confidence when averaged across all words. However, the performance begins to improve as statistics are computed over finer categories, word classes or phones.

We intend to investigate better clustering of word classes, and the estimation of phone class reliability, similar to the updating technique of [1]. We believe this will further improve the predictive capability of senone normalization.

4. ACKNOWLEDGMENT

This project is supported in part by an ATP Cooperative Agreement, Number 70NANB5H1184, from the National Institute of Standards and Technology.

5. REFERENCES

[1] Young, S. and Ward, W., *Recognition Confidence Measures for Spontaneous Spoken Dialog*, EUROSPEECH'93, September 1993.

[2] Chase, L., Rosenfeld, R., and Ward, W., *Error-Responsive Modifications to Speech Recognizers: Negative N-grams*, ICSLP 1994.

[3] Chase, L., *Error-Responsive Feedback Mechanisms for Speech Recognizers*, Unpublished Ph.D. Dissertation, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, April, 1997.

[4] Ravishankar, M.K., *Efficient Algorithms for Speech Recognition*, Unpublished Ph.D. Dissertation, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, May 1996.

[5] Hwang, M. Y. and Huang, X. D., *Subphonetic Modeling With Markov States - Senone*, ICASSP'92, March 1992.

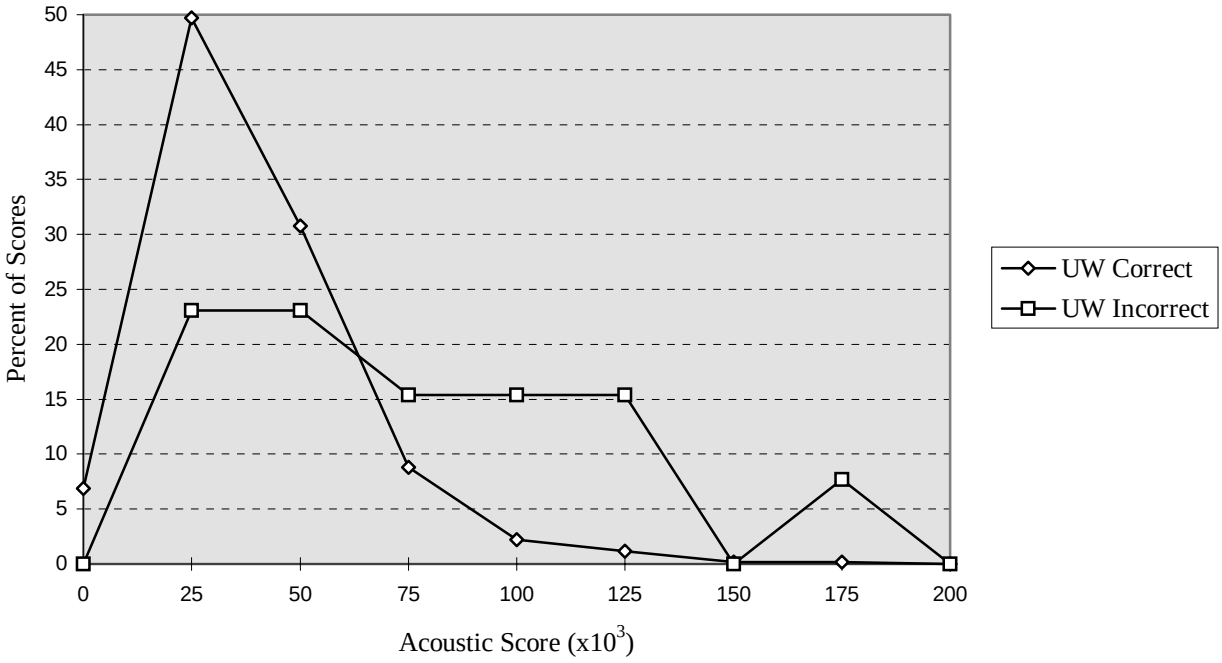

Figure 3. Correct and incorrect distributions for the phone: UW.