# WORD-BASED ACOUSTIC CONFIDENCE MEASURES FOR LARGE-VOCABULARY SPEECH RECOGNITION

*Asela Gunawardana*[1], *Hsiao-Wuen Hon and Li Jiang*

Microsoft Research
Redmond, Washington 98052, USA

## ABSTRACT

Word level confidence measures are of use in many areas of speech recognition. Comparing the hypothesized word score to the score of a 'filler' model has been the most popular confidence measure because it is highly efficient, and does not require a large amount of training data. This paper explores an extension of this technique which also compares the hypothesized word score to the scores of words that are commonly confused for it, while maintaining efficiency and the low demand for training data. The proposed method gives a 39% relative false accept rate reduction over the 'filler'- model baseline, at a false reject rate of 5%.

## 1. Introduction

Confidence measures are useful in many aspects of speech recognition, including supervised and unsupervised adaptation, recognition error rejection, out-of-vocabulary word detection, and keyword spotting. A method that has been popular for word-based confidence modeling is the comparison of the score of the hypothesized word with the score of a 'filler' model [1,2,3,4]. It was later demonstrated that the use of a large-vocabulary speech recognizer improves confidence modeling [5]. The approach taken in this paper is to extend this by also considering the scores of words which are commonly confused with the hypothesized word, thus detecting such confusions. The set of the commonly confused words is constructed from the N-best list for the hypothesized word. Weintraub [6] has used "N-best list homogeneity"' for keyword spotting where the percentage of all recognized N-best sentence hypotheses (weighted by the probability of the different sentences) in which a keyword appears is computed for use as a confidence measure for that keyword. However, word-level N-best scores have not been directly used for confidence measure purposes. The advantage of this method over others such as decision trees [7] and building models of $P(\text{Correct} \mid \text{Evidence})$ through Bayesian methods is that this method requires very little data. Particularly, it requires no examples of incorrect hypotheses (i.e. examples which should be rejected).

In this paper, the problem of word-based confidence measures for the supervised adaptation problem is considered. I.e. the user of an ASR system is given a `training' utterance to read, and the problem is to determine if this is indeed what the user spoke. Since the words to be read are fixed, only acoustic model scores are used in the confidence measure developed. This method could readily be extended to other applications such as unsupervised adaptation, error-rejection and keyword spotting.

The paper is organized as follows. In section 2 we will describe the database and the Whisper speech recognition system used for this study. In section 3 we will describe in detail our proposed word-based confidence measure. In section 4, we will discuss how we evaluate our word-based confidence measure and in section 5, we will report the results. Finally we will present a summary of this work and a discussion of future work in section 6.

## 2. Database

To avoid effects (such as alignment) which could influence confidence model performance, we decided to use an isolated database (where words are well separated by silence and the alignment of misrecognized words is likely to be correct) for this work. However, our method could easily be applied to continuous speech, as described in section 6, and in fact, there could be some definite advantages to applying this confidence model to a continuous database.

We used the Microsoft large-vocabulary isolated database (MS-LVID) for this study. MS-LVID is a speaker-independent database where each speaker spoke many lexical words in isolation. The lexicon contains about 60,000 words. A subset (about 210,000 words) was selected to train the acoustic Hidden Markov models (HMMs) using Microsoft Whisper speech recognition system (please see [8] for a full description of Whisper). An independent subset (about 120,000 words) was selected to build the confidence models which will be described later. Another small, completely independent subset (about 10,000 words) was reserved as an independent test set. Note that all three sets have no speaker overlap at all, i.e., all the confidence measure experiments are strictly speaker-independent.

## 3. The New Confidence Measure

The approach taken was to generate an exemplar N-best list (which should represent the set of the commonly confused words) for each word, and then compare this with the N-best list for each test occurrence of that word to get a confidence score. Section 3.1 contains a description of how the exemplars can be generated. Section 3.2 will describe how the comparison was performed. In the following descriptions, the letter $v$ is used to denote a word for which an exemplar is being built, as opposed to words which occur in an N-best list, which will be denoted by $w$. Thus, an N-best list for a word $v$ would include words $w_1$ through $w_N$, one of which would usually be $v$ itself.

## 3.1 Exemplar Generation

Exemplars were generated for all words in the vocabulary of which there were at least four occurrences in the training set. This gave 984 words for which exemplars were built. The acoustic scores in each N-best list for a given word $v$ were normalized first by utterance length and then by the score of $v$ to give the per-frame acoustic probability normalized by the per-frame probability of $v$. $v$ and the $M-1$ other words in the N-best lists which had the highest mean normalized score were retained. The scores of the other words were pooled to obtain an estimate of the score of words which occur infrequently in the N-best lists for word $v$ (this will be referred to as the score of $w_{\varepsilon}$ in the sequel). $M$ is chosen so that this score is small. In addition to the $M$ words and mean scores, the exemplar also contains the normalized score from a fully connected network of context independent phone models (henceforth referred to as $w_{CI}$), which serves as the 'filler' model. The variances of the normalized scores for each word $w_1$ through $w_M$, for $w_{CI}$, and for $w_{\varepsilon}$ were also calculated at this stage, as these are used in the comparison metric between the exemplar and the test N-best lists. Owing to the sparseness of the training data, these variances were smoothed by ensuring they were greater than a constant which was determined experimentally.

## 3.2 Exemplar - Test Occurrence Comparison

The scores in a test N-best list (including the score of $w_{CI}$) are normalized by length and the probability of $v$, and the top $L$ are retained to form a 'test vector' of scores, $s(w_1')$ through $s(w_L')$. The confidence score $C$ is then given by

(1) $$C = \exp\left[ -\tfrac{1}{2} \sum_{i=1}^{L} \frac{\left(s(w_i') - \mu_v(w_i')\right)^2}{\sigma_v^2(w_i')} \right]$$

where $\mu_v(w_i')$ and $\sigma_v^2(w_i')$ are the mean and variance of $w_i$' in the exemplar for $v$, and the mean and variance of $w_{\varepsilon}$ is used if $w_i$' is not one of the $M$ words $w_1$ through $w_M$ in the exemplar. Note that although this has the same form as an unnormalized normal density (assuming independence of the components), the scores can only take non-negative values, and cannot truly be distributed as such. However, this form of the confidence measure has the intuitively satisfying property that it allows differences in factors that usually vary little to contribute heavily to the confidence score, and discounts the contribution of factors which vary more in the training set. The (lack of) normalization gives us a confidence score between zero and one.

## 3.3 Rejection

The confidence scores calculated are used to reject words with low confidence. Experiments were performed to compare performance when this was done using a single word independent threshold and a word dependent threshold. Word dependent thresholds were trained only for very frequent words, while other words used a common threshold. All thresholds are calculated on an independent threshold tuning set, which is separate from the data used in training and testing the confidence models.

A novel variation on the thresholding is the use of a two-level threshold. If a test instance passes the thresholding described above, the ratio between the confidence score of the expected word $v$ and the confidence score of the recognized (one-best) word $w1$' is compared to a second threshold. This rejection scheme is in essence seeing whether the expected or recognized word has higher confidence.

## 3.4 Other Variations

Other variations on the comparison metric were explored, but did not perform as well as the system described above. These include

- Normalizing so that the scores in each test vector sum up to one. Similarly, the scores in exemplar generation are also normalized so that each N-best score total sums up to one. Since this type of normalization tends to be arbitrary (depending on how many entries in the N-best list were kept), the resulting confidence model is inferior to the proposed one where the normalization based on the score of hypothesized word $v$ is guaranteed to be consistent across exemplar generation and testing occurrence comparison.
- When computing the Gaussian confidence score $C$, one can sum over the words $w_1$ through $w_M$ in the exemplar instead of the words $W_1$' through $W_L$' in the test vector in equation 1 to gives the following form for the confidence score:

  $$C = \exp\left[ -\tfrac{1}{2} \sum_{i=1}^{M} \frac{\left(s(w_i) - \mu_v(w_i)\right)^2}{\sigma_v^2(w_i')} \right]$$

  Since our exemplar was generated by pooling multiple instances of N-best lists for $v$ from the training set, this variation requires positive testing tokens to match high scoring words in all the training instances. This is an overly strong requirement, so this variation generates less desirable results.

# 4. EVALUATION

To evaluate our proposed confidence model, three test sets were created. The first is the positive set where the expected word was spoken, and the instances should all be accepted. This is known as the 'accept set.' It is relative tricky to come up negative sets for supervised adaptation, where all the instances should all be rejected. We decided to artificially create two negative sets. The first one is a 'random reject set' where an error transcription $v$' is randomly created for each testing token $v$. Since $v$' is randomly created, the chances are that $v$ and $v'$ are so different that detecting this substitution is trivial. What we are more interested in is the cases where people misread the correct transcription by uttering similar (therefore confusable) words. Therefore we need another 'confusable reject set' where we deliberately choose a $v'$ that often appears near the top of the N-best lists for word $v$. This ensures that $v$ and $v'$ are confusable. Table 1 shows examples from these sets.

| Word Spoken | Accept Set | Random Reject Set | Confusable Reject Set |
|---|---|---|---|
| The(2) | The(2) | Felt | D. |
| All | All | And(2) | Ball |
| Dollar | Dollar | Approved | Dollars |
| He | He | Tender | She |
| Is | Is | Foundation | His |

**Table 1:** The word actually spoken compared with the expected word according to the transcripts of the accept, random reject, and confusable reject sets.

These three sets enable the calculation of a false reject rate (on the accept set) and two false accept rates (one on each reject set) at each threshold value, giving two ROC curves, one corresponding to each reject set. These curves are used to compare the performance of the baseline system and the different variations of the confidence measure described above.

As mentioned earlier, the baseline used is the comparison of the length normalized acoustic score of $v$ with that of $w_{CI}$, which corresponds to limiting the test vector to contain only the score of $w_{CI}$ (which is normalized by the score of $v$).

## 5. Results

Figure 1 clearly demonstrates that the proposed system outperforms the baseline on both the random and confusable word rejection tasks. The difference in the false accept rates in Figures 1a and 1b illustrates that while detecting random substitutions is trivial, the problem of detecting confusable substitutions is a harder one. A practical rejection system must address both problems.

The performance of the system with changing test vector size ($L$ in the description above) is shown in Figure 2. As expected, the performance improves as more and more words are considered in computing the confidence, and then drops off as the additional words used in the computation become less reliable indicators of possible confusion

Finally, Table 2 shows the effect of the various thresholding schemes used on the performance of both the proposed model and the baseline. The results show the false accept rate of the baseline compared with that of the proposed model with one-level and two-level thresholds, when used with word independent and word dependent thresholds, while keeping the false reject rate at 5%. These results are on the confusable reject set. Trends on the random reject set and at other false reject levels are similar.

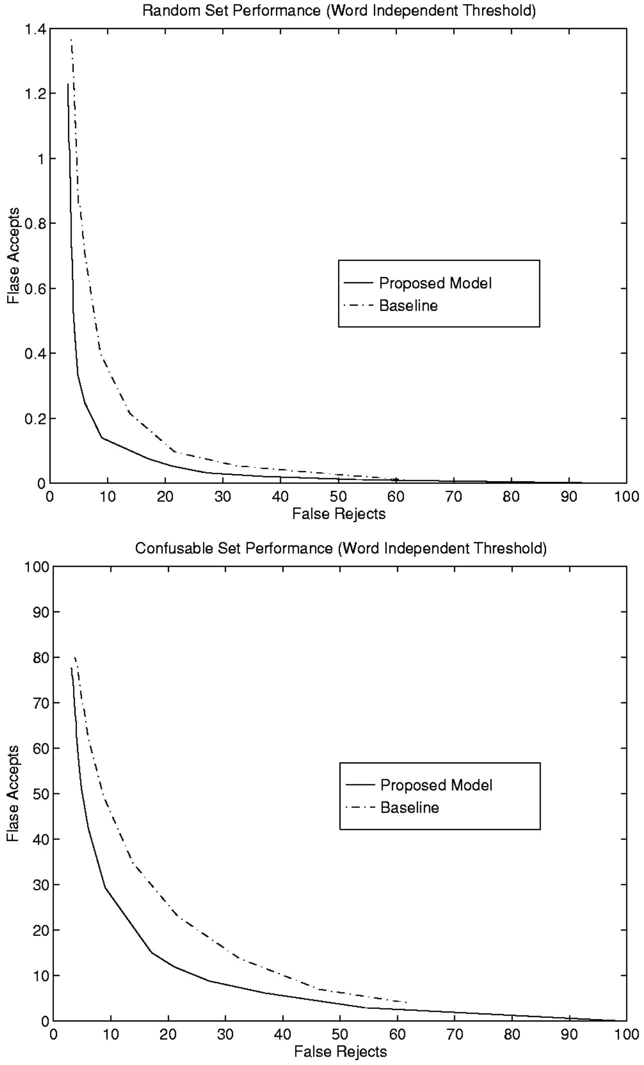

**Figure 1:** ROC curves comparing the performance of the proposed model with that of the baseline. These curves were obtained using word independent threshold systems.

| | Baseline | Proposed Model | |
|---|---|---|---|
| | | One-Level | Two-Level |
| Word Independent | 70% | 50% | 43% |
| Word Dependent | 67% | 47% | 41% |

**Table 2:** The false accept rate of the baseline compared with that of the proposed model with one-level and two-level thresholds, when used with word independent and word dependent thresholds, while keeping the false reject rate at 5%.

## 6. Discussion

The proposed confidence measure results in improvement over the baseline because the baseline only uses the likelihood of the expected word relative to a `filler' model score, while the proposed system compares it to the likelihood of words

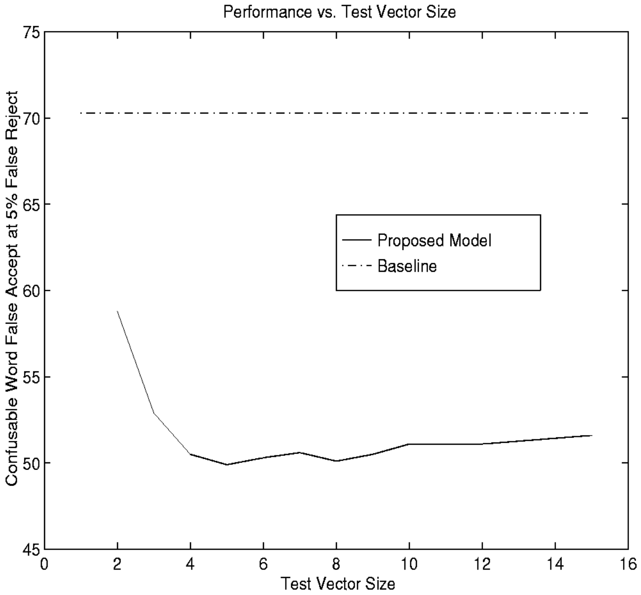

**Figure 2:** The false accept rate on the confusable reject set, at a false reject rate of 5%, as a function of test vector size. The results at other false reject rates, and on the random reject set, are similar.

commonly confused for it. The hypothesis behind the proposed system is that each word is characterized by the words it sounds similar to, and by the degree of similarity with which it matches each of these words. The N-best list exemplars attempt to capture this characteristic pattern of similarity for each word. Using N-best scores also allows the constrained use of context dependent acoustic model scores in making confidence assessments at no additional computational cost.

We have demonstrated that with only 4 occurrences of training tokens, we are able to construct an efficient Gaussian based confidence models which could give us 39% relative false accept rate reduction at the 5% false rejection point. A natural direction for future work is in generating data for words with less than 4 occurrences by concatenative synthesis, using the real acoustic segments (which could be either senones or triphones), as is done in Microsoft's Whistler TTS system [9]. The use of the real acoustic segments in the synthesis may ensure that the synthesized data exhibits the same similarity patterns as the real data.

Although our confidence measure is being measured in the supervised adaptation scenario, it should be feasible to extend this work to unsupervised adaptation, recognition error rejection, out-of-vocabulary word detection, and keyword spotting. This system can also be adapted for use with a continuous recognizer in the following ways. For the supervised adaptation task, the expected transcript can be used to align the input utterance first, and the N-best word recognizer can then be applied on each word segment in order to generate the confidence exemplars. For the unsupervised adaptation or recognition error rejection tasks, the recognition alignment can be used in a similar way. In either case, our model may benefit from using a continuous recognizer because misspoken or misrecognized words often create alignment errors in surrounding words, so that these errors could be more easily identified due to significant mismatch between the N-best list of confidence exemplars and the testing tokens.

This confidence model could also be improved by adding other features to the exemplars. For example, the use of language model scores could be useful in unsupervised adaptation or recognition error-rejection, though these particular features may not be useful in the supervised adaptation case. Other features such as word length may also be of value.

Another direction for investigation is the use of negative training examples when they are available. For example, the similarity measure in equation 1 may be modified to penalize similarity to features derived from negative examples. This would allow for the use of negative training examples, but not require that they exist for every word $v$.

## 7. REFERENCES

1. Rohlicek, J.R., Russel, W., Roukos, S., and Gish, H. “Continuous Hidden Markov Modeling for Speaker-Independent Word Spotting,” *ICASSP-89* 1:627–630, 1989.

2. Rose, R.C. and Paul, D.B. “A Hidden Markov Model Based Keyword Recognition System,” *ICASSP-90* 1:129–132, 1990.

3. Alleva, F., Beeferman, D., and Huang, X.D. “Confidence Measures and Their Application to Speech Recognition,” *IEEE Auto. Spch. Recog. Wkshp.,* 173–174, 1995.

4. Cox, S. and Rose R., “Confidence Measures for the Switchboard Database,” *ICASSP-96* 1:511–514, 1996.

5. Jeanrenaud, P., Siu, M., and Gish, H. “Large Vocabulary Word Scoring as a Basis for Transcription Generation,” *Eurospeech-95,* 3:2149–2152, 1995.

6. Weintraub, M. “LVCSR Log-Likelihood Ration Scoring for Keyword Spotting,” *ICASSP-95* 1:297–300, 1995.

7. Neti, C.V., Roukos, S., Eide, E. “Word-Based Confidence Measures as a Guide for Stack Search in Speech Recognition,” *ICASSP-97* 2:883–886, 1997.

8. Huang, X., Acero, A., Alleva, F., Hwang, M.-Y., Jiang, L., and Mahajan, M. “Microsoft Windows Highly Intelligent Speech Recognizer: Whisper,” *ICASSP-95* 1:93–96, 1995.

9. Huang, X., Acero, A., Adcock, J., Hon, H.-W., Goldsmith, J., Liu, J., and Plumpe, M., “Whistler: A Trainable Text-to-Speech System,” *ICASSP-96* 4:2397–2390, 1996.