The use of speech recognition confidence scores in dialogue systems

GABRIEL SKANTZE
gabriel@speech.kth.se
Department of Speech, Music and Hearing, KTH

This paper discusses the interpretation of speech recognition confidence scores and how they can be used in spoken dialogue systems to improve robustness and efficiency. An evaluation of confidence accuracy from a commercial speech recognition system is presented and discussed. It is argued that the common use of confidence scores against a threshold will not give much error detection improvement compared to a baseline, even if the confidence accuracy is good. The confidence score should instead be used as a probability measure among others.

1 INTRODUCTION

Confidence scores are provided by most speech recognisers to provide a measure of how confident it is in the hypothesised recognitions. This is important, since errors occur frequently in speech recognition and threatens to make applications such as spoken dialogue systems too cumbersome to use. An accurate confidence measure can help us to determine when the recognition is incorrect, in order to take appropriate actions. One important question here is what the confidence score really is supposed to reflect and how it should be interpreted. This question has consequences for how we should evaluate the accuracy of confidence measures.

In this paper, I will first discuss the possible interpretations of confidence scores and confidence accuracy. I will then discuss how confidence scores can be used in dialogue systems. One application is for the interactive error handling that systems must engage the user in. A second use is automatic repair of speech recognition errors. Third, confidence scores can be used to detect if the user is using words that are not in the system's vocabulary.

An evaluation of confidence accuracy in a commercial speech recognition system will also be presented. Recognition and confidence accuracy will be analysed based on word length and how they differ for specific words. Based on the results from this evaluation, we will discuss how speech recognition confidence scores should be interpreted, and how they could be applied in dialogue systems to increase the robustness and efficiency of the dialogue.

2 SPEECH RECOGNITION CONFIDENCE SCORES

A speech recognition confidence score should reflect how confident the speech recogniser is in the recognition of an utterance. Confidence scores are often given in the range 0-100 or 0-1. One question here is what this figure really means. Often, the only thing to know for sure when using a commercial speech recogniser is that high scores mean "confident" and low scores mean "not confident". Different speech recognisers use different methods for calculating the score. According to Hazen et al. (2002), the confidence score is usually based on the probability of the acoustic observation (x) given the speech segment (u), normalized by the general probability of the acoustic observation:

$$confidence(u \mid x) = \log \frac{p(x \mid u)}{p(x)}$$

The problem for the speech recogniser is that words that are out of vocabulary will corrupt the theoretical model. Many competing hypotheses will also make the estimation harder. The scores from the different hypothesis are used differently in different recognisers. Often, different methods are tried and tuned against empirical data. Probabilities from the language models could also be used to increase performance.

Thus, there seem to be some relation between the probability of the observation and the confidence score. If the confidence score would have a perfect correspondence to how probable it is that the recognition is correct, recognitions with a confidence score of 0.75 would be correct 3 times out of 4. This paper will present an empirical study of a commercial speech recogniser where this relation is explored.

In a dialogue system there are of course other modules than the speech recognition, such as the parser, which could also provide confidence scores. However, we will focus on the speech recognition confidence score in this paper. The speech recogniser could assign confidences to the whole recognised string (and other hypotheses). Some recognisers can also assign confidences to each word in the recognised string.

2.1 Confidence accuracy

Speech recognition can be evaluated by measuring *recognition accuracy*. For single words, we can simply measure the proportion of correctly recognised words. For whole strings, a common measure is the *word error rate*. This is calculated by aligning the words in the recognised string with those in the actual utterance, using the minimum edit distance, and then divide the number of words that have been added, deleted or substituted, with the total number of word in the spoken utterance (Jurafsky & Martin, 2000).

The speech recognition accuracy must be separated from *confidence accuracy*. The confidence accuracy does not tell us how well the speech recogniser did in the recognition, but instead how well it did in assigning a confidence to this recognition. Thus, a speech recogniser could have very poor recognition accuracy, but still very high confidence accuracy (assigning low confidences to the recognition results). The question is how this confidence accuracy should be calculated. This is, of course, dependent on how it should be defined.

The most common use of confidence is to compare it to a threshold. Words (or utterances) with confidence scores below the threshold are rejected, and those with a score higher than the threshold are accepted. Alternatively, one could define a grey zone where the utterance is implicitly or explicitly confirmed (see section 3.1.1).

The threshold used should be tuned according to some empirical data. There is a trade-off that has to be made between the number of *correct rejections* and *correct acceptances*. These terms are explained in Table 1:

Table 1: Classification of acceptances and rejections.

	Correctly recognized	**Falsely recognized**
Above threshold	Correct acceptance	False acceptance
Below threshold	False rejection	Correct rejection

A high threshold will give a large number of rejections, but also a low number of acceptances. A low threshold will instead give a low number of rejections and a high number of acceptances.

The tuning should aim at finding the lowest total number of false acceptances and false rejections. This break-off is often close to the so-called *equal error rate*, where the number of false acceptances equals the number of false rejections.

An intuitive measure of confidence accuracy is how low the equal error rate can get. However, this measure is not in line with the assumption that the confidence score should reflect probability of correctness, as can be seen in Figure 1. If the confidence should reflect the probability as perfectly as possible, recognitions with confidences around 50% would be correctly judged in 50% of the cases, and those with a confidence of 0% or 100% would be correctly judged in 100% of the cases. Thus, the equal error rate would be as high as 25%, given the hypothetical case of an even distribution of confidence scores.

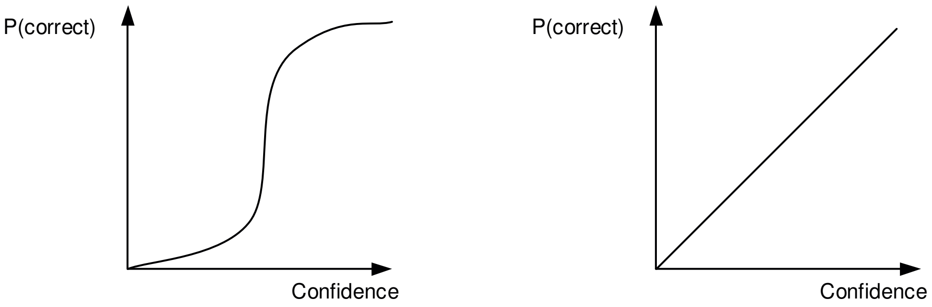

Figure 1: The graph to the left shows a relation between confidence and probability of correctness that would give a low equal-error rate. The right graph shows instead an ideal correlation, which would not give as low equal-error rate.

To get a zero equal error rate, all confidence scores above the threshold should reflect correct recognition and all confidence scores below threshold should reflect false recognition. Thus, if the confidence score reflects probability, the use of confidence against a threshold is not good for detecting errors in the recognition result, even if the confidence accuracy is 100%.

2.2 Augmenting the confidence score

The speech recognition confidence score is not the only measure that could be used to detect errors in the recognition result. Litman et al. (2000) used machine learning to test how different features could be used for detection of speech recognition errors. One of these features was the recognition confidence score, but they also used prosodic features: the maximum and minimum F_0 and RMS values, the total duration of the utterance, the length of the pause preceding the turn, the speaking rate and the amount of silence within the turn. The algorithm was trained to learn if-then-else rules against the set of features, in order to classify if a given recognition had a word error rate (WER) of more than 0 or not. If the algorithm had guessed that all recognition results were incorrect, it would have had an accuracy of 51.34%. This gives us a baseline to compare the results with. Training the algorithm on the ASR confidence score as the only feature, gave an accuracy of 77.77%. Adding the prosodic features increased the accuracy to 87.24%. To increase the accuracy further, the words from the recognized string and the grammar used were added as features. Thus, the context of the recognition and certain recognized keywords could be considered by the algorithm. This increased the classification accuracy further to 93.47%. This means that the adding of prosodic and contextual features give much higher error detection accuracy than just the recognition confidence score. There have also been studies of error machine learning using features from natural language processing components, dialogue management and discourse history (Walker et al., 2000).

3 USE OF CONFIDENCE SCORES IN DIALOGUE SYSTEMS

As we have seen, the confidence score gives us information about how probable it is that the recognition is correct, so that we can detect recognition errors. Now, what should we do with this information? We will look at three applications for dialogue systems: error handling (involving the user in the error detection and recovery), robust understanding of utterances that may contain recognition errors, and detection of out of vocabulary words.

3.1 Error handling

3.1.1 *Explicit and implicit confirmation*

The most common use of confidence scores in error handling in dialogue systems is to choose between *rejection* (low confidence), levels of *confirmation* (medium confidence) and *acceptance* (high confidence) (c.f. Larsson, 2002). The rejection is normally expressed as a request for repetition, e.g. "sorry, I didn't understand, please repeat". Confirmations are often divided into *implicit* and *explicit* confirmations (c.f. McTear, 2002). In explicit confirmation, the system asks a yes-no question, like:

(1) Do you want to go to Stockholm?

The user could then reply with a "yes" or "no". The system should also support answers like "no, Gothenburg", since this is a natural reply. If the system asks several questions, one could also choose to confirm all slots in one question:

(2) Do you want to go from Malmö to Stockholm on the 4th of November?

This can speed up the interaction, at least if all slots were correct. On the other hand, if one of the slots was incorrect, the error recovery can be very cumbersome (just imagine what such an answer could look like).

With implicit confirmation, the system instead embeds the understanding in the next contribution, like:

(3) To Stockholm. At what time do you want to leave?

Or even more implicit:

(4) At what time do you want to leave for Stockholm?

The user can then just continue if it was correct, or correct the system if it was wrong. Implicit confirmation can give rise to a wider range of possible responses. More implicit confirmation (as in example 4) also puts further demands on the natural language generation.

The choice between rejection, confirmation and acceptance relies on the possibility to compare the confidence score against a set of thresholds (which we have seen could be problematic). One could add different levels of explicitness or implicitness to this scale, depending on the confidence, in order to make the dialogue more natural. It is also possible to confirm parts of an utterance, based on the confidence scores for the individual words.

3.1.2 *Grounding in dialogue*

Different levels of confirmation actually reflect different *signals of understanding* (or *linguistic feedback*), used in human-human dialogue in order to make sure that we share a common understanding of the things we say. According to Clark (1996), a dialogue consists of "contributions", where each contribution consists of a *presentation* phase and an *acceptance* phase. In the presentation phase, A presents a *signal* for B to understand. In the acceptance phase, B accepts A's signal by giving him some sort of *evidence* that she believes she understands what A means by it. She assumes that, once A gets the evidence, he too will believe that she understands. Clark means that every presentation needs some sort of closure or acceptance in order to be judged as common ground for the participants. The evidence can be more or less explicit, depending on the *grounding criterion*, i.e. how important it is that the presented signal is perceived correctly.

A common signal of understanding is some sort of *assertion*, like "uh huh" or a nod. Another way of signalling understanding is to *presuppose* understanding, by initiating a relevant next turn. These signals do not give the speaker any real evidence of understanding; they just mean that the addressee thinks that he understands what was intended. Stronger evidence can be signalled by *displaying* parts of what was said, i.e. giving an answer to a question that in parts show how the question was understood. In some situations, the addressee could also *exemplify* what she perceived, for example give a paraphrase or repetition of what was said.

These signals give the speaker of the original utterance a chance to correct if it was not perceived as intended. The addressee must choose the right kind of evidence for the situation. Too strong evidence would not be in line with what Clark calls the *principle of least effort*: "All things being equal, agents try to minimize their effort in doing what they intend to do." The process by which the speakers make sure that their contributions are added to the common ground as accurately and efficiently as possible is called *grounding*.

In a spoken dialogue system the confidence score could be used in determining how strong evidence of understanding the system should show (just like with different levels of confirmation). However, as we have seen, there are also other factors that determine the strength of evidence. The grounding criterion is dependent on the *consequence of task failure*, i.e. how hard it is to correct the misunderstanding in the specific situation. Another factor is the user's *need for feedback*: The fact that the system is confident in its understanding does not mean that the user has to be confident that the utterance was understood correctly. The same signals can also have different costs depending on the situation (if it is a common task for the user, if the user is in a hurry, etc). This implies that the traditional distinction between explicit and implicit confirmation based on a set of thresholds should be updated with a more complex model, in which the confidence score is treated as a probability value among others, which affects how the system should signal understanding.

3.2 Robust spoken language understanding

Confidence scores can also be used to automatically correct speech recognition errors in spoken language understanding. Wang & Lin (2002) describes a system where the recognized string is first parsed into a concept sequence. Figure 2 (F) shows an example where the utterance "tell me forecast in Taipei tonight" has been misrecognised as "Miami forecast in Taipei tonight". The system then compares this concept sequence with a set of possible concept sequences, which are derived from a training corpus and stored in a database. If none of the hypothetical concept sequences match the recognized sequence, the best matching sequence is chosen (K). The matching is made by calculating the minimum edit distance for each hypothetical sequence and choosing the one with the least edit cost (E). The word confidence scores are used to calculate a confidence score for each concept, which is used to calculate the edit cost. Thus, a deletion is more costly for a concept that has a higher confidence. The exact costs for different edit operations at different confidence scores are empirically derived from the training corpus. This method takes into account that the confidence score reflects probability, instead of using a threshold to discard concepts.

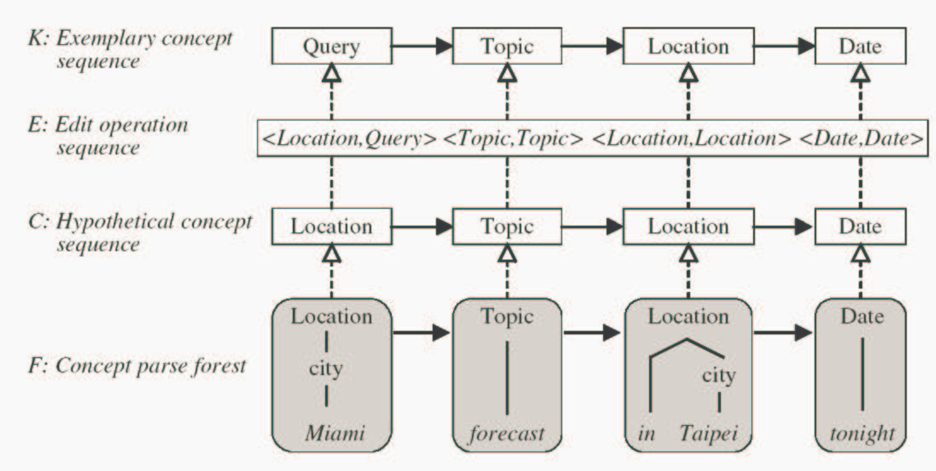
K: Exemplary concept sequence
Query
Topic
Location
Date
E: Edit operation sequence
<Location,Query> <Topic,Topic> <Location,Location> <Date,Date>
C: Hypothetical concept sequence
Location
Topic
Location
Date
F: Concept parse forest
Location
city
Miami
Topic
forecast
Location
city
in
Taipei
Date
tonight

Figure 2: An example of recognition correction. (Taken from Wang & Lin, 2002)

3.3 Detection of OOV words

A big problem with spoken language understanding in dialogue systems is that the speech recogniser only can recognise words that are in the predefined vocabulary. If the user uses a word that is out of vocabulary (OOV), the recogniser will still make a "guess", using in vocabulary words, which will give erroneous input to other components. One solution is to use "filler" or "garbage" models in the speech recognition that should catch OOV words. Hazen & Bazzi (2001) shows how a generic word model can be used in this way. Another option is to use the confidence scores (compared to a threshold) to identify the recognised words that represent OOV words. The two methods are compared by Haze & Bazzi (2001). The comparison showed that the OOV model performed better than the confidence scoring model. However, the best result was gained by combining the two methods.

4 AN EMPIRICAL EVALUATION OF CONFIDENCE ACCURACY

What the confidence scores can be used for is dependent on what they really reflect, and on how good the accuracy really is. To test this, the confidence scores from a commercial speech recogniser were analysed.

4.1 Data collection and analysis

The data for the evaluation was taken from a study on human error handling strategies, which will only be described briefly. A more detailed description of the experiment will be provided in a subsequent paper. This paper will only concern the speech recognition confidence accuracy, which should be more dependent on the specific speech recogniser than the experimental set-up.

In the experiment, two subjects were placed in separate rooms. One person was given the task of describing the route on a simulated campus to the other person. The other person could only see a small fraction of the map, just where he was standing. Both subjects could speak to each other, however, the speech from the subject who was supposed to find the way was processed by a speech recogniser. The other subject could only read the results from the speech recognition, and was not able to hear anything. To help the person who was reading the recognition result, the words in the result string were coloured, depending on the confidence score. Words with high confidence scores were coloured in dark grey, and words with low confidence were coloured in light grey. In post-interviews, the subjects told that this colouring was of great help, suggesting that the confidence scoring carried some information.

For speech recognition, Nuance speech recogniser 8.0 with built-in acoustic models for Swedish was used (see htttp://www.nuance.com). Stochastic trigram grammars were used, trained on handcrafted example utterances and a corpus from some pilot studies. The sparse data gave a relatively poor recognition performance. The vocabulary consisted of 352 words.

8 pairs of users were used, and each pair was given 5 scenarios, which resulted in 40 dialogues. This gave a total of 822 recognised utterances (from one of the subjects) with a total of 4481 recognised words (approx. 5 words per utterance on average).

Each utterance was transcribed, so that the accuracy of the recognition could be calculated. The minimum edit distance was used for each utterance (Jurafsky & Martin, 2000), in order to align the words. Each recognised word could then be given the tag "correct" or "incorrect". It is important to note that this method only gives a measure of insertions and substitutions, not deletions. This is because deletions cannot be related to confidence scores.

4.2 Accuracy

Only 71.9% of the words were correctly recognised. This figure does only reflect substitutions and insertions (not deletions), which means that a word-error-rate measure would give an even worse result.

The words were divided into ten intervals, depending on the word confidence score (0-0.1, 0.1-0.2, etc). This distribution was not even, as can be seen in Figure 3.

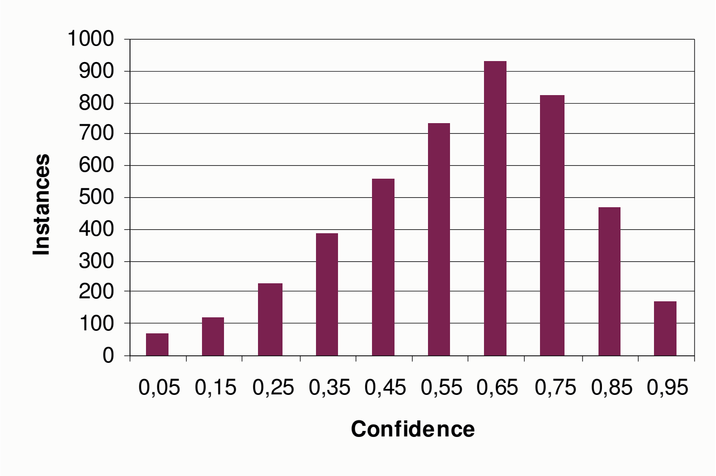

Figure 3: The distribution of words in the different confidence intervals.

This, of course, reflects the fact that most recognitions were correct (71.9%). It is interesting to note that the peak in the histogram is close to this figure.

The number of correct recognitions in each interval was divided with the total number of words in each interval, which gave the probability P(Correct) for each interval. The result can be seen in Figure 4.

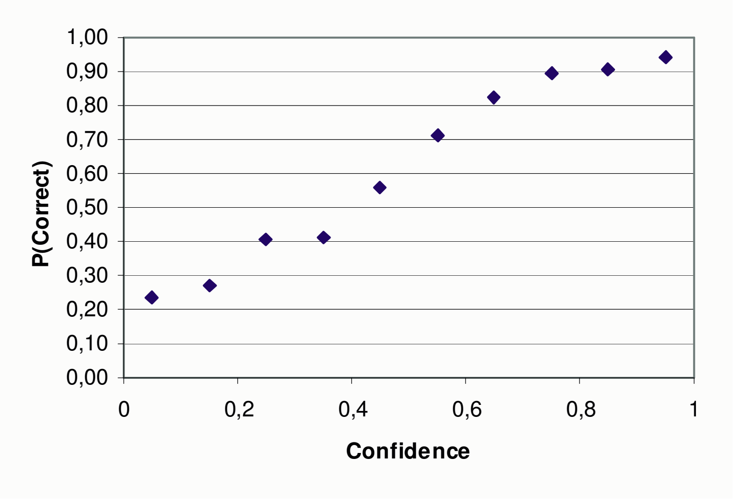

Figure 4: The proportion of correctly recognised words in the different confidence intervals.

The figure shows a very high correlation (r = 0,98) between P(Correct) and confidence score, suggesting a very high accuracy if we interpret the confidence score in this way. We can also see that the confidence score is constantly underestimating the real probability of correctness, and that the trend line does not cross the origin.

If we apply the threshold principle on this data, we can see that we get the best trade-off between correct acceptance and correct rejection with a threshold of 0,4. However, the accuracy (the proportion of correctly accepted or rejected words) is only 76.4 %. This should be compared to a baseline of 71.9%, which we get if we accept all words (since 71.9% of the words were correct). This confirms the hypothesis that the confidence scores reflect the probability of correctness, and is not good for using against a threshold.

4.3 Word length

Another interesting result was that both recognition and confidence accuracy seemed to be dependent on word length. To systematise this observation, the words were categorized by the number of syllables. The distribution is shown in Table 2.

Table 2: The distribution of words in the different word length intervals.

Syllables	Instances
1	2708
2	1221
3	191
4	175
5	69
6	110
7	3
8	3

The proportion of correctly recognised words was calculated for each category. This constituted the baseline. The accuracy of confidence when compared to a threshold of 0.4 (the proportion of correctly accepted or rejected words) was also calculated. The result is shown in Figure 5.

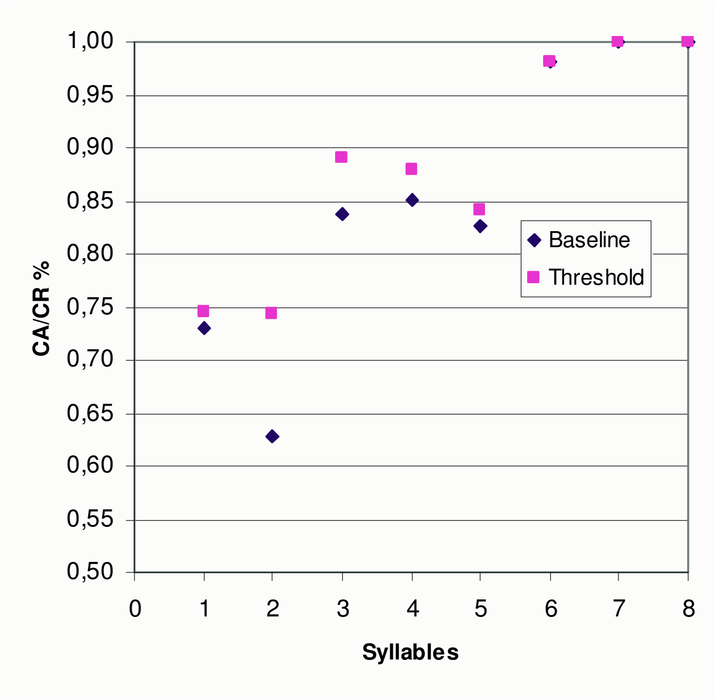

Figure 5: The proportion of correctly recognised words in the different syllable intervals (baseline), and the confidence accuracy compared to a threshold.

As can be seen in the figure, both the recognition accuracy and the confidence accuracy are better for longer words than for shorter words. The use of threshold seems to be most meaningful for words with two syllables, which are also most often incorrect.

4.4 Sensitivity to specific words

The accuracy of recognition and confidence also seemed to be different for different words. Words that occurred in the speech recognition results at least ten times were counted for recognition accuracy. The mean recognition accuracy and confidence for the worst and the best words are shown in Table 3.

Table 3: Mean recognition accuracy and confidence scores for the worst and the best words.

Worst words	Accuracy	Confidence	Best words	Accuracy	Confidence
matsal	0,00	0,32	gräsmatta	0,92	0,69
trädet	0,05	0,43	vänster	0,93	0,57
gata	0,13	0,46	mig	0,93	0,65
var	0,17	0,45	nummer	0,94	0,57
fem	0,26	0,40	institutionen	0,99	0,72
åtta	0,27	0,45	befinner	1,00	0,68
bara	0,28	0,43	industriell	1,00	0,75

att	0,30	0,44	system	1,00	0,72
av	0,33	0,66	tegelbyggnaden	1,00	0,68
huset	0,33	0,40	träbyggnad	1,00	0,70

As can be seen in the table, some words are almost always correct, while others are almost never correct. It is interesting to note that the confidence scores do not seem to reflect this fact very well. The correlation between confidence and correctness for the specific words was only 0,58. Thus, knowledge about the recognition accuracy of specific words can give information for rejecting or accepting words, which is not given in the confidence score. Such figures, taken from the evaluation of the recognition performance of a dialogue system, could be used to make the system more suspicious or forgiving against specific words.

5 CONCLUSIONS & DISCUSSION

In the commercial speech recognition software evaluated here, there is a strong linear correlation between confidence and probability of correctness and thereby high confidence accuracy, suggesting that confidence scores are definitely useful in spoken dialogue systems. Still, the use of a threshold bears little improvement on the detection of recognition errors (compared to a baseline). The results are of course dependent on the speech recognition software used here (however common), but the results at least demonstrates that such an evaluation is important in order to understand how the confidence scores should be used in a spoken dialogue system. There seems to be two extremes in the attitudes towards confidence scores. They are either naively seen as a simple means for error detection or as totally worthless. This is perhaps because due to the misconception that they could easily be compared against a threshold. This standard approach to confidence scores needs to be replaced with models where confidence scores are treated as probability values, among others. Approaches that use the confidence score combined with other measures (e.g. Litman et al., 2000; Hazen & Bazzi, 2001) or treat it as a probability measure (e.g. Wang & Lin, 2002) seem to be successful.

Using empirical data, we can see that we can add more knowledge to the error detection process, other than just using confidence scores. The accuracy is dependent on word length as well as specific words. In this example, words with two syllables are most often incorrect. This seems to be one case where a confidence threshold would significantly increase error detection accuracy. Words with six syllables or more are almost always correct (however uncommon). Using knowledge from an empirical evaluation, we can enhance the system to be more suspicious towards some specific words that occur in the recognised string, and almost always accept other specific words.

In order to improve the performance of interactive error detection and recovery, which involves the user, we must build more complex models of the grounding process. The confidence score is important as one variable that should affect the system's grounding behaviour. However, there are also other variables (such as consequence of task failure) that must be taken into consideration, in order to get a robust and efficient dialogue that the user does not perceive as burdened with errors and confirmations.

6 REFERENCES

Clark, H. H. (1996). *Using language*. Cambridge University Press.

Hazen, T. J., & Bazzi, I. (2001). A comparison and combination of methods for OOV word detection and word confidence scoring. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*.

Hazen, T.J, Polifroni, J., & Seneff, S. (2002). Recognition confidence scoring for use in speech understanding systems. *Computer Speech and Language*, 16(1), 49-67.

Jurafsky, D. S., & J. H. Martin. (2000). *Speech and Language Processing*. Prentice Hall, Inc., Englewood, N.J.

Larsson, S. (2002). *Issue-based Dialogue Management*. PhD Thesis, Göteborg University.

Litman, D. J., Hirschberg, J., Swertz, M. (2000). Predicting Automatic Speech Recognition Performance Using Prosodic Cues. In *Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics*. 218-225.

McTear, M. F. (2002). Spoken Dialogue Technology: Enabling the Conversational User Interface. *ACM Computing Surveys*, 34(1): 90-169.

Walker, M. A., Wright, J., Langkilde, I. (2000). Using Natural Language Processing and Discourse Features to Identify Understanding Errors in a Spoken Dialogue System. In *Proceedings of the 17th International Conference on Machine Learning.*

Wang, H. & Lin, Y. (2002). Error-tolerant spoken language understanding with confidence measuring. *Proceedings of the International Conference on Spoken Language Processing*.