# A NEW WORD-CONFIDENCE THRESHOLD TECHNIQUE TO ENHANCE THE PERFORMANCE OF SPOKEN DIALOGUE SYSTEMS

*R. López-Cózar, A. J. Rubio, P. García, J. C. Segura*
Dpto. Electrónica y Tecnología de Computadores
Universidad de Granada, 18071 Granada, España (Spain)
Tel.: +34-958-243193, FAX: +34-958-243230, *E*-mail: {ramon,rubio,pedro,segura}@hal.ugr.es

## ABSTRACT

Spoken dialogue systems generally use one or two confidence thresholds during speech recognition. A confidence value assigned to a word represents the recognizer's confidence in the correct recognition of the word. If the confidence value is under a threshold then the word is considered a recognition error and the system must ask the user to re-enter it. Alternatively, the system can ask for a confirmation from the user. Environmental conditions and peculiarities of the speaker's voice can change from one dialogue to another, so that it is necessary to decide the most appropriate value for the confidence threshold. If the selected value is too low, the words that are wrongly inserted by the recognizer may be considered correctly recognized. On the other hand, if the selected value is too high, even the words actually uttered by the user can be considered recognition errors, or words that must be confirmed. In this paper we present an experimental strategy to automatically select the most appropriate value for the confidence threshold. This strategy has been applied to the dialogue system we have developed, which aims to deal with telephone-based fast food queries and orders. We present the results obtained and indicate possibilities for future work .

## 1. INTRODUCTION

Applications of speech technology include all kinds of systems in which a part of the communication process is carried out by voice. Real-world applications involve a human being trying to communicate with a machine to get some information or service. Simple automatic speech recognition systems can be used to achieve some goals when the dialogue is heavily limited, as is usually the case of isolated-word speech recognition systems. The dialogue restrictions imply training and collaboration on the part of the users. Unrestricted dialogue applications are much more appealing as users do not need to be trained and collaboration requirements are minimal. These systems include a dialogue module that is an important part of the whole system.

Recognition of continuous speech is much harder than recognition of isolated words. On the one hand, the ends of words and sentences are often not clear. This makes it difficult to segment the discourse into units that can be handled separately (words, for instance). As a result, continuous-speech recognition systems are more complex. Additionally, articulation effects are more pronounced in continuous speech. Prosody is also a problem as it depends on the position of words in sentences. Consequently, recognition error rates are greater in continuous speech [1].

Recognition systems generally produce as output a list of hypotheses ordered on the basis of a probability measure. However, this measure does not consider the quality of the recognition process or the system's confidence in the correct recognition of the words. The performance of spoken dialogue systems is enhanced if they use confidence measures generated by recognizers, as these can omit words with a low confidence value [2]. Typically, one or two confidence thresholds can be used. Let us suppose the system uses just one confidence threshold $t$, $0 < t < 1$. Let $conf(w)$ be the confidence value associated with a word $w$, $0 < conf(w) <1$. If $conf(w) \geq t$ then $w$ is considered correctly recognized. Otherwise, it is considered a recognition error and then the system must either confirm it or ask the user to reenter it. Let us now suppose that the system uses two confidence thresholds, $t_1$ and $t_2$, $t_1 < t_2$ . In this case, the system will ask the user to reenter $w$ if $conf(w) < t_1$. It will ask the user to confirm it if $t_1 \leq conf(w) < t_2$. Finally, $w$ will be considered correctly recognized if $conf(w) \geq t_2$. It is possible to use confidence measures based on the conjunction of acoustic and language models, or based on either of them separately. For example, [3] presents a confidence measure based on a trigram. The measure considers that the trigrams' probabilities are more reliable than those of a bigram, and that those of the latter are more reliable than the probabilities of unigrams. This measure relies on acoustic models, which are just a part of the knowledge used during recognition.

## 2. THE DIALOGUE SYSTEM

Spoken dialogue systems are a relatively new technology that was introduced in the late 1980s and mainly promoted by the projects DARPA SLS (Spoken Language Systems) in the United States and SUNDIAL (Speech UNderstanding and DIALog) in Europe. The goal of both programs is to develop computer programs that can provide travel information to the user via speech. The SLS project focuses on flight information in English, while the SUNDIAL project deals with both plane and train information in several languages. Theses systems are currently used for database querying, ticket-reservation, weather information, search for information through the Internet, etc. [4]. They make use of speech recognition, comprehension and speech synthesis technologies. They must be robust against a wide range of acoustic and language variabilities that may considerably degrade their performance.

We have developed a spoken dialogue system for Spanish, named SAPLEN, which aims to deal with the product orders and queries of fast-food restaurants' clients using the telephone [5]. The system uses a continuous-speech recognition module developed at our laboratory that uses context-independent phoneme-like units modelled by SCHMM (Semi-Continuous Hidden Markov Models). The vocabulary size is about 2.000 words, including restaurant-product names, names of streets, avenues, squares, etc. The language is modelled by bigrams.

In our preliminary experiments we used a simulator which can include, change or remove words in the sentences uttered by the users, depending upon nine parameters which determine its performance [6]. A *noise level* parameter (N) represents the negative effect upon the user's voice signal of extraneous noise. Four parameters decide how many words uttered by the user are made unrecognizable because of noise. The system then processes sentences containing words that might have been inserted, changed or removed. Three parameters are used to calculate the confidence value associated with every word in a sentence. The system uses expectations about what the user will probably say in his/her interaction [7]. Finally, a *confidence threshold* (T) decides whether every word *w* in a sentence is considered as having been correctly recognized. This is the case when *conf(w)*≥ T.

Initially, we carried out a preliminary evaluation of the system at a fast-food restaurant, using different confidence threshold values (FT) [6]. The evaluation was carried out using objective and subjective methods [8]. The metrics used for the objective evaluation were: word accuracy (WA), key-word accuracy (KWA), implicit recovery (IR), sentence recognition (SR), sentence understanding (SU), turn correction ratio (TCR), contextual appropriateness (CA), transaction success (TS) and dialogue-abandonment on the part of the users (AB). Table 1 sets out the objective results obtained.

| | FT=0 | FT=0.6 | FT=0.7 | FT=0.8 | FT=0.9 |
|---|---|---|---|---|---|
| WA | 100 | 90.47 | 90.24 | 79.59 | 33.33 |
| KWA | 100 | 90.99 | 91.63 | 77.17 | 41.15 |
| IR | 0 | 46.66 | 46.87 | 37.5 | 18.98 |
| SR | 100 | 70.0 | 68.0 | 52.0 | 21.0 |
| SU | 85.71 | 82.25 | 69.08 | 55.97 | 24.75 |
| TCR | 4.27 | 10.46 | 15.79 | 26.45 | 56.5 |
| CA | 85 | 79.77 | 73.72 | 54.8 | 37.87 |
| TS | 84.1 | 56.41 | 46.92 | 31.42 | 13.63 |
| AB | 0 | 0 | 0 | 0.25 | 0.75 |

*Table 1.* Objective evaluation of SAPLEN system using several values for a fixed confidence threshold FT

As we can see in the table, the performance of the system decreases as the confidence threshold increases. This happens because the percentage of words that are considered recognition errors increases when the threshold increases. As a result, 75% of users who tested the system when FT=0.9 abandoned the dialogue.

The 100 test users (restaurant clients) were asked to rank from 1 (*very bad*) to 5 (*very good*) the following measures: sentence understanding (SU), error recovery (ER), natural language generation (NLG), naturalness (NA), transaction success (TS), task completion (TC), speed (SP), and overall satisfaction (SAT). Table 2 sets out the averaged subjective results obtained.

| | FT=0 | FT=0.6 | FT=0.7 | FT=0.8 | FT=0.9 |
|---|---|---|---|---|---|
| SU | 4,05 | 3,7 | 4 | 3,25 | 2,1 |
| ER | 4,3 | 4,1 | 4,1 | 3,55 | 1,95 |
| NLG | 4,4 | 4,3 | 4,6 | 4,4 | 4,15 |
| NA | 3,7 | 3,95 | 4,1 | 3,35 | 3,05 |
| TS | 4,55 | 4,55 | 4,8 | 3,85 | 2,1 |
| TC | 4,55 | 4,8 | 4,75 | 3,35 | 2 |
| SP | 4,35 | 4 | 4,15 | 3,65 | 3,65 |
| SAT | 3,95 | 4,1 | 4,1 | 3,15 | 2,1 |
| Total: | 4,23 | 4,19 | 4,33 | 3,57 | 2,64 |

*Table 2.* Subjective evaluation of SAPLEN system using several values for a fixed confidence threshold FT

As we can see in the table, the performance of the system generally decreases as the confidence threshold increases. However, from the point of view of the users, the system achieves its best performance when FT=0.7. There are three reasons for this fact. First, the same performance of the system is not always equally-ranked by all users, as their requirements are different. Second, many users think the system is more natural if it asks them to re-enter some words sometimes. Thirdly, as the implicit recovery capability of the system attains its highest value when FT=0.7, many recognition errors can be recovered without being noticed by users.

## 3. THE NEW TECHNIQUE

The technique we present in this paper uses an *Adaptive Confidence Threshold* (ACT). The advantage of this technique is that the confidence threshold can be adapted automatically to different environmental conditions existing during the conversations between the system and the users. The ACT technique uses a buffer that stores the confidence value assigned to the last *n* interactions of the user (we have used *n*=10 in our experiments). The confidence value assigned to an interaction is computed as the average of the confidence values of the words in the interaction. At the beginning of the dialogue the system clears the buffer, and stores a confidence value which is considered the minimum confidence threshold. During the dialogue, the value for the confidence threshold is calculated as the average of the values in the buffer. If the value obtained is smaller than the minimum confidence threshold then the new confidence threshold is the minimum confidence threshold. This procedure tends to slowly raise the confidence threshold. If the system understands sentences correctly, the confidence threshold is considered to be properly set, otherwise it is considered wrongly set. In our experiments, we have noticed that many misunderstandings occur because the confidence threshold is set too high. In such a situation it is necessary to decrease its value. To do so, the values in the buffer that are greater or equal to the threshold value are removed. From this moment on, we consider the threshold is *fixed*; it cannot be updated again unless another misunderstanding occurs, in which case its value is decreased as described above.

Figure 1 shows the updating procedure of a dynamic threshold (DT) as described above. The dialogue comprises 13 interactions. Five types of noise (N) are considered.

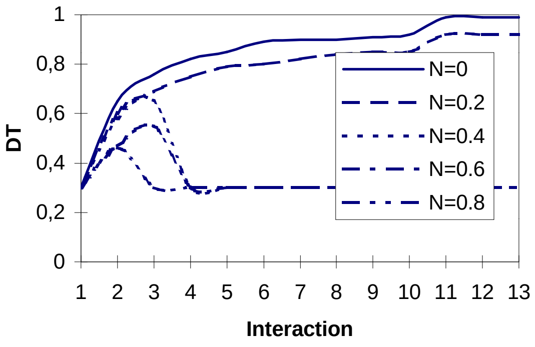

*Figure 1*. Dynamic threshold (DT) updating for different types of noise (N)

N=0 represents the ideal case, i.e., it is assumed the voice signal is not affected by any noise. We consider no recognition errors occur in this case. As we can see in the figure, when N=0 the threshold increases slowly until it is fixed at its maximum value DT=0.99 at interaction number eleven. N=0.2 means that 20% of the energy of the voice signal corresponds to noise. In this case, the increment of the threshold is slower because of the presence of noise. The threshold is fixed at 0.92. N=0.4 means that 40% of the voice signal corresponds to noise. In this case there are only two increments of the threshold because at the third interaction a misunderstanding occurs, which means that the threshold is then fixed at its minimum value DT=0.33. N=0.6 means that 60% of the voice signal corresponds to noise. Again, in this case there are only two increments of the threshold, as at the third interaction a misunderstanding occurs, and so the threshold is fixed at its minimum value DT=0.33. Finally, N=0.8 means that 80% of the voice signal corresponds to noise. In this case there is only one increment of the threshold, as at the second interaction a misunderstanding occurs, and the threshold is thus fixed at is minimum value DT=0.33.

## 4. EXPERIMENTAL RESULTS

To measure the effect of the dynamic threshold in dialogues, we used 43 dialogues taken previously from test users of the dialogue system in a fast-food restaurant [6]. These dialogues were taken using different fixed threshold values (FT). Dialogues 1-10 were taken using FT=0, dialogues 11-20 were taken using FT=0.6, dialogues 21-30 were taken using FT=0.8, dialogues 31-40 were taken using FT=0.8, and dialogues 41-43 were taken using FT=0.9. There are only three dialogues in the last group because most test-users abandoned the dialogue as the performance of the system was unacceptable for them. We simulated these dialogues in the laboratory using a dynamic threshold that is updated following the procedure described above. We measured the duration of dialogues (in terms of *turns of dialogue*) and the sentence-understanding rate. Our goal was to evaluate the performance of the dialogue system when fixed and dynamic thresholds were used. Figure 2 shows the results obtained with respect to the duration of dialogues.

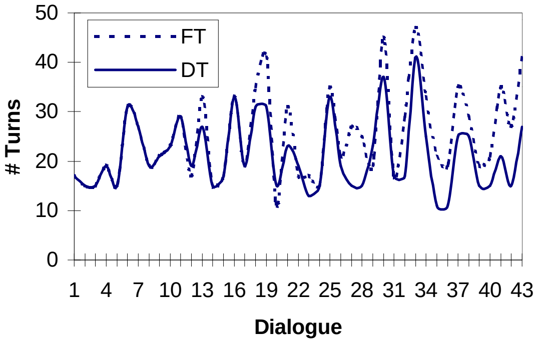

*Figure 2*. Length of dialogues for fixed (FT) and dynamic (DT) confidence thresholds

As we can see in the figure, the duration is similar for the first ten dialogues because they correspond to the ideal case, i.e., it is assumed that no recognition errors occur. The differences start at dialogue 12. The duration tends to be lower when the dynamic threshold is used. Differences are most relevant from dialogue 41, as in this case many recognition errors occur. Figure 3 shows the results obtained with respect to the sentence-understanding rate.

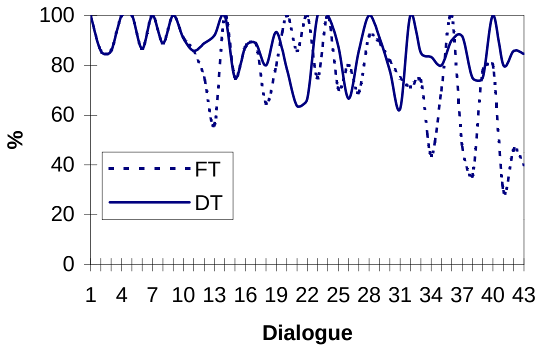

*Figure 3*. Sentence understanding rate for fixed (FT) and dynamic (DT) confidence thresholds

As we can see in the figure, the sentence-understanding rate is similar for the first ten dialogues. As before, differences start at dialogue number 12. Generally, the comprehension rate is greater when the dynamic threshold is used. The most relevant difference appears for the dialogues of the last two groups (starting at dialogue number 31) when many recognition errors occur.

## 5. CONCLUSIONS AND FUTURE WORK

Our experiments have proved that the dialogues take less time if the ACT technique is used. The average duration for the corpus of dialogues was 25.04 dialogue-turns when we used the fixed confidence threshold FT=0.6 (the ideal case FT=0 was not considered), which was reduced to 21.34 dialogue-turns when the dynamic threshold was used. The sentence-understanding rate under noise-simulated conditions was also greater when the dynamic threshold was used. The average sentence-understanding rate for the same corpus of dialogues was 82.25% when we used a fixed threshold FT=0.6 (the ideal case was not considered), and it increased to 86.96% when the dynamic threshold was used. Although the experiments produced good results, it must be remembered that they were carried out under simulated conditions, and so it is necessary to experiment under real-world conditions to obtain more reliable results.

At the moment, the dialogue system we have developed uses only one confidence threshold. It would present a more natural performance if it used two confidence thresholds instead of just one. As we said before, if we set up the 2-threshold strategy, the system could ask the user to re-enter words or to confirm them. Additionally, on some occasions it might be more reliable to recognize confirmations, as the set of possible answers would be smaller.

The strategy we used for updating the confidence threshold is arbitrary. We could consider other ways of updating and compare the results obtained. The performance of the word-recognition simulator depends on several parameters. Some of them decide how many words are inserted, removed or changed, while others decide how to assign confidence values to words. Thus, the overall performance of the system depends on these parameters. More studies are necessary to determine, if possible, the values these parameters must be set to, in order to model the behaviour of a given recognition system.

The recognizer we are using at the moment does not assign confidence values to the words it provides as output, but only uses probability measures to prune the search space when decoding sentences uttered by users [9]. Therefore, we need to modify this recognizer to include confidence measures that can be used by the control module of the system.

## 6. REFERENCES

[1] L. Rabiner, B.H. Juang. "Fundamentals of Speech Recognition". Prentice-Hall, 1993.

[2] Thomas Kemp, Thomas Schaaf, “Estimating Confidence Using Word Lattices”, Eurospeech ‘97, pp. 827-830

[3] C. Uhrik, “Confidence Metrics Based on N-Gram Language Model Backoff Behaviors”, Eurospeech ‘97, pp. 2771-2774

[4] Victor Zue. “Conversational interfaces: Advances and challenges”. Keynote 2, Eurospeech’97

[5] Ramón López-Cózar, Pedro García, J. Díaz, Antonio J. Rubio. "A Voice Activated Dialog System for Fast-Food Restaurant Applications", Eurospeech ' 97, pp. 1783-1786

[6] R. López-Cózar, A. J. Rubio, P. García, J. C. Segura, “A Spoken Dialogue System Based on a Dialogue Corpus Analysis”, First International Conference on Language Resources and Evaluation (LREC‘98), pag. 55-58

[7] Morena Danieli. “On the use of expectations for detecting and repairing human-machine miscommunication”, Working notes of the AAAI-96, pp. 87-93

[8] Morena Danieli y Elisabetta Gerbino. “Metrics for evaluating dialogue strategies in a spoken language system”. Proceedings of the 1995 AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation, pp. 34-39

[9] Lee K. F., Alleva F., “Advances in Speech Signal Processing”, Continuous Speech Recognition, pp. 623-650, Dekker