ROBUST AND EFFICIENT CONFIDENCE MEASURE FOR ISOLATED COMMAND RECOGNITION

Gustavo Hernández-Ábrego, Xavier Menéndez-Pidal, Lex Olorenshaw

Spoken Language Technology, Sony NSCA
3300 Zanker Road MS SJ1B5, San Jose, CA. USA
tel: +408-955-5374; E-mail:gustavo@slt.sel.sony.com

ABSTRACT

A new confidence measure for isolated command recognition is presented. It is versatile and efficient in two ways. First, it is based exclusively on the speech recognizer's output. In addition, it is robust to changes in the vocabulary, acoustic model and parameter settings. Its calculation is very simple and it is based on the computation of a *pseudo-filler* score from an N-best list. Performance is tested in two different command recognition applications. Finally, it is efficient to separate the correct results both from the incorrect ones and from the false alarms caused by out-of-vocabulary elements and noises.

1. INTRODUCTION

In the real-life application of speech recognition systems, it is unavoidable to have portions of speech that do not represent meaningful information to the system. Such speech instances, such as noises, hesitations, and words not related to the application, may not transmit any real message to the system and instead may only disturb the recognition performance. Generically these potentially irrelevant portions of speech are referred to as out of vocabulary instances (OOVs). OOV handling represents an important part of the speech recognition process for real-life speech recognition systems. OOV handling can be addressed by evaluating the recognition reliability. By building confidence measures (CMs), every recognition result is scrutinized and its reliability assessed.

Confidence measures and their application to OOV handling and verification of recognition results have been a matter of study for some time [1]. Several CM proposals are based on the computation of features related to the recognition process which are later combined into a unified CM. Even though some of these approaches have demonstrated their value [2], some of them represent a substantial computational burden for the recognition system, or require a lot of additional information to be computed. Moreover, many earlier CMs are based on the particular characteristics of a given recognition system [3] and are not easy to port to other recognition architectures or applications. The CM we propose tries to avoid the drawbacks previously mentioned. We have conceived it to be as simple and robust as possible. In terms of simplicity, our CM is only based on the recognizer's output and does not require an extensive scrutiny of the recognition process. Furthermore, it does not require external knowledge sources (such as previous recognition statistics) for its computation. In terms of robustness, the CM has been formulated to present a stable behavior when the application (vocabulary) is changed or the characteristics of the recognizer are modified (for instance, when the acoustic model is changed). Though these goals may appear ambitious, we have fulfilled them as we have focused our work on isolated command recognition which still represents a useful, albeit limited, method of human-machine interaction.

2. CONFIDENCE MEASURE FORMULATION

The formulation of our confidence measure is based on a recognizer with an N-best list output. Such an N-best list contains the top N best word candidates for a given speech input. Speech recognition is a cumbersome process even for isolated word recognition, but it can be accelerated by time-reduction techniques such as beam search. In the N-best recognizer, the number of candidates generated depends largely on the width of the beam used in the search. Each recognition candidate has a recognition score associated with it. In order to find out whether the top hypothesis is correct or not, the recognition score can be compared to the score of a background filler model. This method, known as likelihood ratio test, has been used for keyword recognition verification in [1]. The likelihood ratio is formulated as

$$S_{LR} = S_{KW} - S_B \tag{1}$$

where S_{LR}, S_{KW} and S_B are the log scores for the likelihood ratio, keyword and background respectively. Instead of using the score of the filler model, we propose to build a *pseudo-filler* score taken as the average of the scores of the items in the N-best list. As such, equation (1) turns into

$$S_{LR} = S_1 - \frac{1}{N-1}\sum_{i=2}^{N} S_i \tag{2}$$

where S_i is the score of the ith candidate in the N-best list. On the other hand, some recognition verification systems are based on comparing the recognition score against a second alternative [4], taken from an alternative recognizer. In order to keep the simplicity of our current formulation, we may use the second-best recognition hypothesis as an alternative hypothesis. If this approach is combined with the approximated likelihood ratio test of (2), the CM can be re-formulated as

$$C_1 = \frac{S_1 - S_2}{S_1 - \frac{1}{N-2}\sum_{i=3}^{N} S_i} \tag{3}$$

Some detail inspection of equation (3) shows that C_1, because of the normalization ratio, is a positive number with values between 0 and 1. This formulation comprises some concepts which are worthwhile noticing. First, the recognition score is compared against the second hypothesis in order to determine how likely it is that recognition gets confused with elements within the vocabulary. Second, recognition is compared against the pseudo-filler in order to find out how close it is to something that does not exist in the vocabulary. Both comparisons are compared again in order to compute the CM. Some experimentation on (3) shows that the CM values are vocabulary dependent. In a recognition system with closed vocabulary, when two commands are acoustically close, their scores also tend to be very close. Thus, the numerator of (3) tends to remain constant when there are several confusable commands in the vocabulary. To compensate for the effect of confusable vocabulary, we propose to split the pseudo-filler score between the numerator and denominator of (3) according to

$$C_2 = \frac{S_1 - \frac{1}{m-1}\sum_{i=2}^{m} S_i}{S_1 - \frac{1}{N-m}\sum_{i=m+1}^{N} S_i} \qquad (4)$$

where m is the parameter that determines the splitting point and may range from 2 to N-1. If the upper bound of m is considered, equation (4) turns into

$$C_3 = \frac{S_1 - \frac{1}{N-2}\sum_{i=2}^{N-1} S_i}{S_1 - S_N}. \qquad (5)$$

which no longer depends on the parameter m anymore. Equation (5) still retains the spirit of equation (3) because the likelihood ratio test is computed between the recognition result and a sort of background score (S_N). Nevertheless, the second alternative comparison of (5) does not depend on the vocabulary anymore.

3. EXPERIMENTAL DEVELOPMENT

The principal purpose of the proposed CM is command verification by comparing the CM value against a given threshold. As in every classification problem, the performance of a system like this can be evaluated through the Receiver Operating Characteristic (ROC) curve which plots the rate of false alarms against the correct detections. An ideal classifier would verify all the correct results while rejecting all false alarms. We consider that a good quantitative summary of ROC curves is the area below them. We present the performance of our system in terms of ROC curves and their areas, which also includes an estimation of the error margin. However, since ROCs are not meant to present threshold values, it is hard to set the operating point for a verification system from these curves. That is why we also present plots of the total verification error (false alarm and false detection rates) for our results.

This research was primarily aimed at finding an efficient recognition verification system for Sony's AIBO entertainment robot [5] shown in figure 1. Many of our experiments were based on this application. We tried two vocabularies of AIBO. The first is called "AIBO Life" and is comprised of 48 commands for basic interaction with AIBO. It includes movement commands and praise words such as *go forward, lie down* and *good boy*. A second AIBO vocabulary is more extended and includes 290 commands with several variations of the same command such as *lie down* and *lay down*. In order to test the system in a completely different vocabulary, we include a third test using a car navigation speech

Fig. 1. Sony's AIBO entertaiment robot

Database	Tokens	Description
AIBO Life	1.6k	standard AIBO set
AIBO 290	6.2k	extended AIBO set
City 125	2.2k	car navigation commands
atr00	1k	travel domain sentences
trdv00	1k	travel domain sentences
noise	1.7k	pulse and human noises

Table 1. Vocabulary and OOV database configuration

database called "City125" which contains 125 words related to car commands and town names of the San Francisco Bay Area.

The the goal of the CM is to help to distinguish correct recognition results from incorrect ones as well as OOVs. Therefore, we tested the capabilities of our system when dealing with speech input not related to the recognition application by adding to the test a large number of speech tokens with varying characteristics. Even though our system is equipped with a *filler* model, it is not very reliable to detect OOVs by itself because it only detects 35% of them in the best case. So an OOV input typically produces an incorrect recognition result that should be rejected. The extraneous speech which we used as OOV input includes two different continuous speech databases from the travel domain (named "atr00" and "trdv00"). Also, for the AIBO Life vocabulary, we introduced the rest of the AIBO 290 testing as OOV. The City 125 application was used as OOV for both AIBO applications. Conversely, AIBO was used as OOV for the City application. To complement the OOV configuration, a large number of typical room noises were used. This included common pulse noises such as door slams and surface impacts, as well as human noises such as lip smacking and speaker hesitations. The number of speech tokens used in every part of the testing is shown in table 1.

Recognition is carried out using a Sony-built isolated word recognizer capable of producing an N-best list. Two kinds of acoustic models (AMs) were used on our testing. They are Gaussian-mixture Hidden Markov Models (HMMs) based on triphones of American English. One of the AMs has 4 Gaussians (G4) in the mixture while the other has only 1 Gaussian (G1). Different widths of beam search were used. Recognition performance, in terms of

	AIBO Life		AIBO 290		City 125	
Beam	1G	4G	1G	4G	1G	4G
150	3.02	1.35	11.38	6.04	11.06	6.19
300	2.77	1.35	6.96	3.47	3.89	2.57
600	2.70	1.09	6.13	3.29	1.81	1.11
900	2.57	1.09	5.96	3.21	1.19	0.75

Table 2. Recognition results (WER) for three vocabularies

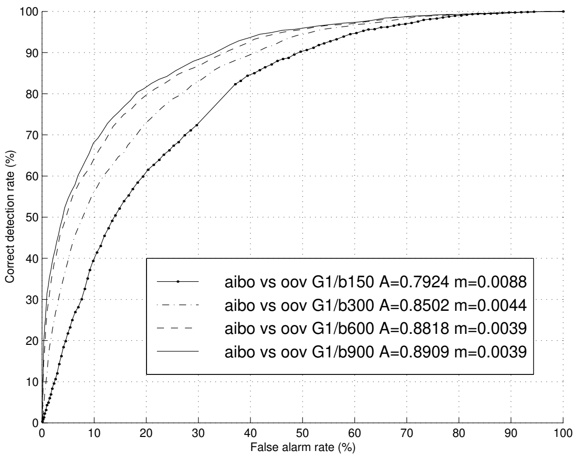

Fig. 2. ROCs and areas with margins for AIBO Life and OOV for the 1-Gaussian HMM

word error rate (WER), are shown in table 2 for the two different AMs using various beams with the three different test vocabularies.

All the results reported are related to a CM computed using equation (5). This formulation, which theoretically is the most advantageous one, has demonstrated higher performance than the rest of the CM formulations introduced in section 2, which were devised during the preliminary experimentation phase of this work.

In figure 2 the ROC curves for different beams for the HMM with 1 Gaussian are presented. The vocabulary tested is AIBO Life and the OOV set is formed by combining all the other databases. Figure 3 shows the ROC curves for several beams for the HMM of 4 Gaussians tested under the same vocabulary/OOV configuration.

4. DISCUSSION

Figure 2 shows how dependent the CM are to the beam used in the recognition. This is understandable because a wider beam allows more alternative candidates in the N-best list and, according to (2), the pseudo-filler score is better estimated when more candidates are considered. In figure 2, for a beam of 150, the discriminative performance of the confidence measure is not satisfactory, but for a beam of 300 it presents fair performance. For instance, at a 20% rate of false alarms, the correct detection rate is above 70%. This value grows to nearly 80% for a beam of 600. The maximum beam tested here, 900, surpasses the 80% detection level at a 20% rate of false alarms. If more false alarms are allowed, for instance a 30% rate, almost 90% of the correctly recognized results are verified by the system. It is well known that incrementing the beam in the recognition raises the accuracy rates but it also increases recognition time. The performance of the CM also rises with the beam. However, for real-life applications, the beam should be defined considering the trade-off between accuracy and speed.

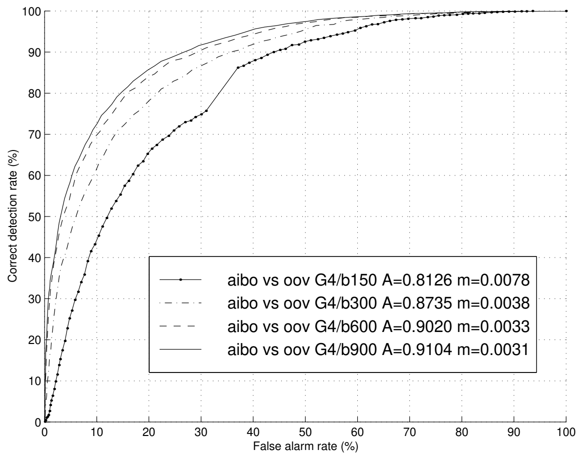

Fig. 3. ROCs and areas with margins for AIBO Life and OOV for the 4-Gaussians HMM

When higher recognition accuracy is desired, an acoustic model with higher resolution can be used. We tested the CM based on equation (5) using an HMM with 4 Gaussians in the recognizer. Results are shown in figure 3.

As expected, using a 4-Gaussians HMM improves the performance of both the recognizer (table 2) and the confidence measuring system. According to the plots in figure 3, even when narrow beams are used, the CM of the 4-Gaussians HMM clearly surpasses the performance of the CM calculated with the 1-Gaussian HMM. Better results for the 4-Gaussians HMMs are seen when the beam is broadened. Very good discrimination characteristics at beams of 600 and 900, for instance, get around 85% of detection at a 20% false alarm rate. Also, in these curves, accuracy at very low false alarm rates is very high reaching the 70% detection at a low 10% of false alarms.

From figures 2 and 3 it can be said that our CMs depend on the accuracy of the recognizer, which in this case is very related to the kind of HMMs used and the beam used during recognition. It ultimately depends on the number of scores used in the pseudo-filler computation of (5). Experimental results show that the more scores are considered, the more accurate the CMs are. Moreover, it should be noticed that our CMs require at least 3 candidates in the N-best list. Otherwise, the pseudo-filler would be ill-defined. Thus, we have decided to include exception values for the CM for recognition results that cannot provide more than 2 candidates in the N-best list, having CM=0 for the no-candidate N-best list and CM=0.5 for the lists with 1 or 2 candidates.

It is important to note that the calculation of the CM can be considered as a post-process of the recognition engine which does not require any previous and/or external knowledge. In such a way it can be considered as a plug-in process to the recognizer. However, in any classification problem, the setting of the decision threshold is a delicate question. We would like to have a plug-in whose parameters are the least dependent on the application as possible. Therefore, we must examine how the threshold varies when the recognition architecture is changed.

First, in figure 4, the behavior of the CM thresholds is shown

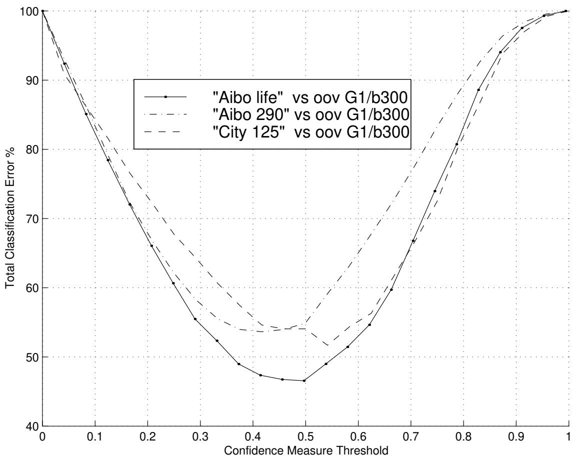

Fig. 4. Total error rates and thresholds for three different vocabularies using the same HMM and beam

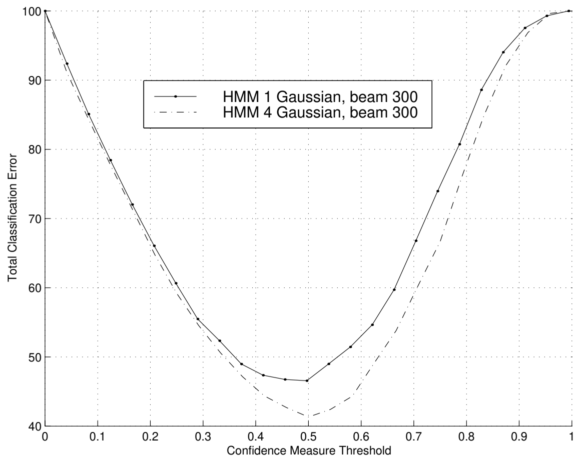

Fig. 5. Total error rates and thresholds for the AIBO Life vocabulary for two different AMs.

when three different vocabularies are tested using the same HMM and beam. This figure shows the total error rate, which is the false alarm rate plus the false rejection rate, as a function of the value of the CM threshold used to separate correct results from incorrect ones. Of course, the optimal operation point for the system is located where the error is minimum. Even though vocabularies are very different, their curves present minimum errors located at similar threshold points. As the vocabulary grows, recognition accuracy drops (see AIBO 290 results in table 2) and it becomes harder to use the CM to distinguish between correct and incorrect results. On the other hand, the number of alternative hypotheses in the N-best list grows and the pseudo-filler can be estimated more accurately.

In figure 4, all curves belong to systems that use the same HMM at the same beam. If these conditions change, the optimum threshold point may also be different because, according to (5), when more scores are considered, the CM tends toward lower values. In order to find out how the thresholds change when different acoustic models are used during recognition, figure 5 shows the total error curves for the CM calculated after the recognition results for the HMMs of 1 and 4 Gaussians.

From figure 5 it can be observed that, in spite of the drastic changes in the recognition architecture, the behavior of the CM is very stable. The total error rates of the 4-Gaussians curve are lower than those of the 1-Gaussian curve. Still, both of them present optimal points at very similar CM values.

These observations about robustness indicate that with this method of calculating confidence measures, the recognition configuration or the vocabulary of the application can be changed without adjusting the CM module at all.

5. CONCLUDING REMARKS

Here we have presented a confidence measuring method whose computation is very simple. It does not require any external knowledge sources and it only considers the recognition results for its computation. The computation requires, however, a recognizer equipped with an N-best list output. The performance of such CMs depends on the number of candidates found in the N-best list and, in consequence, it depends on the width of the beam used during recognition. Also, the CM performance is related to the length of the vocabulary. These, however, are not major drawbacks because the accuracy of any isolated command recognizer depends on these same conditions. Nevertheless, we have shown that our CM have good performance when used as the basis of a recognition verification system. A confidence measuring system such as this can be used as a module that does not depend on the recognizer's configuration and that does not have any parameters of its own to be adjusted. The CMs calculated with this method present stable behavior when the recognition vocabulary is changed. Furthermore, they are robust to changes in the recognition architecture. Even though the CM method introduced here is far from being an ideal recognition verification system, its simplicity and robustness make it appealing for real-life command recognition applications.

In the future, we will consider the behavior of this CM under noise and environment disturbances as well as its extension to continuous speech recognition.

6. REFERENCES

[1] R. C. Rose and D. B. Paul, "A Hidden Markov Model based keyword recognition system," in *Proceedings of 1990 ICASSP*, Albuquerque, April 1990, vol. I, pp. 129–132.

[2] T. Schaaf and T. Kemp, "Confidence measures for spontaneous speech recognition," in *Proceedings of 1997 ICASSP*, Munich, April 1997, vol. II, pp. 875–878.

[3] D. Wu, X. Menéndez-Pidal, and et at, "Speech and word detection algorithms for hands-free applications," in *Proceedings of Sony Research Forum '99*, Tokyo, October 1999, p. 224.

[4] G. Hernández-Ábrego and J. B. Mariño, "A second opinion approach for speech recognition verification," in *Proceedings of the VIII SNRFAI*, Bilbao, May 1999, vol. I, pp. 85–92.

[5] M. Fujita, "Digital creatures for future entertainment robotics," in *Proceedings of International Conference on Robotics and Automation*, San Francisco, April 2000.