Out-Of-Vocabulary Detection and Confidence Measures for Speech Recognition Using Phone Models

Arlindo Veiga [†], Cláudio Neves [†], Fernando Perdigão [†,¥] and Luís Sá [†,¥]
[†] Instituto de Telecomunicações, Pólo II da Universidade de Coimbra, 3030-290 Coimbra, Portugal
Phone: +351-239796236, Fax: +351-239796293, e-mails: {aveiga, claudiorneves, fp, luis}@co.it.pt
[¥] Universidade de Coimbra, Dep. Eng.ª Electrotécnica e de Computadores, 3030-290 Coimbra, Portugal
Phone +351-239796200, Fax +351-239796247

***Abstract* — This paper describes a fast and efficient method to detect out-of-vocabulary words and compute confidence measures in a command-based speech recognition system. The method uses a phone-loop model to reject out-of-vocabulary words and a filler model to compute a confidence measure for each accepted word present in the recognizer output. Tests with this method show that it achieves a good trade-off between false-acceptance versus false-rejection rate. The system runs in real time in a platform with low computational resources and operates in noisy environment conditions (industrial environments and inside vehicles).**

I. INTRODUCTION

A reliable measure of confidence in the output of a speech recognizer is an essential task in real-world speech recognition applications. Confidence measures can be employed for detecting possible errors due to out-of-vocabulary (OOV) words or confusions between vocabulary words caused by noise or unclear pronunciation. In these cases the recognition system should reject OOV and assign a confidence measure for the accepted words.

Previous works on confidence measures can be classified in three types: a) feature based; b) posterior probability based and c) hypothesis testing. Feature based confidence measures use one or more features computed in the decoding processes e.g., log-likelihoods, word duration or word graph density. The work presented in [1] illustrates this approach using normalized log-likelihood scores to detect OOV. Posterior probability based confidence measures try to estimate the probability of the word sequence given the acoustic observations, $P(W|X)$, using word-graphs, as proposed in [2], [3] and [4]. Hypothesis testing formulates the confidence measure problem as a statistical hypothesis testing. This approach is used in [5] and [6]. A likelihood ratio is taken between the word likelihood score and another score from an alternative hypothesis, usually taken from a model: a filler (background model), a specific "anti-model" or a competing model, as in the case of the present paper.

This work describe an implementation of confidence measures in a command based speech recognizer working on specific embedded hardware and therefore with limited computational resources. This restriction prevents using many proposed solutions. For example, in a previous work we have found word-graphs very reliable [7], however we do not consider it here because it would practically duplicate our system response time.

In this paper we propose a fast and efficient method to detect OOV words and compute confidence measures based on a filler model and a generic model made out of phones models as described in the following section.

This paper is organized as follows: section II introduces our confidence measures algorithm; section III describes the speech recognition system and database used: section IV presents experiments and results and finally section V presents our conclusions.

II. CONFIDENCE MEASURES

The speech recognition system used in this work is a command recognizer, and then a grammar is defined to restrict recognizer words - the sequence of commands. Therefore, if an utterance contains OOV words, the decoder will always return a word sequence according to the defined grammar. A Voice Activity Detector (VAD) system is also used in our system to detect utterances boundaries and restrict the number of observations sent to the decoder. In addition, garbage models were used to eliminate clicks and short noises.

After the decoding process we have a word sequence that must be validated. The confidence measures can be computed at various levels, e.g., sentence, word, or phone level. In our case, it is more adequate to use word-level measures, and then the word boundaries defined by decoder are not changed.

For each word to validate, we need first to decide if this word was really uttered or if it could be an OOV, in order to accept or reject it. If the word is accepted, a confidence measure is then computed.

A. Phone-Loop Model

To decide whether a given word is an OOV or not, a generic model is used to supply a comparison term. This model is defined in terms of the phone grammar show in the Figure 1. We call this model Phone-Loop-Model (PLM). This model is used after the decoding process to compute a likelihood score. This score serve as a normalisation coefficient to compare with the recognized word score. The same observations associated with the recognized word are used to compute likelihood of the PML model (Viterbi alignment). As there are no restrictions in the phone sequence, any word can be adequately modelled. A vocabulary word model would produce similar likelihood as the PML model. However, an OOV should produce very

different likelihoods. Thus, a likelihood ratio can assess whether or not an OOV word occur. A similar idea, applied to utterance verification, is presented in [8].

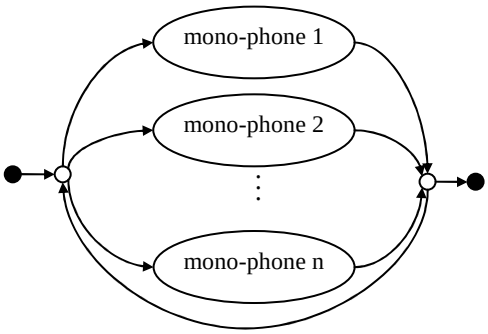

Fig. 1. Unconstrained phones grammar

Instead of a likelihood ratio, we use, as a confidence measures to detect OOV words, a sigmoid as a function of the log-likelihoods difference (likelihoods ratio) between the two hypotheses. Specifically, being P_{cmd} the command model likelihood (defined by decoder) and P_{PLM}, the PLM model likelihood, we define a confidence measure as:

$$CM_{OOV} = \frac{1}{1+e^{-(\log P_{cmd} - \log P_{PLM})}} = \frac{P_{cmd}}{P_{cmd} + P_{PLM}} \qquad (1)$$

Applying a threshold to this confidence measure for each command, allows to accept or to reject the decoder results.

B. Filler Model

To compute confidence measures for accepted commands, we used another model - a filler model - to make a new likelihood comparison. The filler model is trained with all database utterances, commands and phrases, in order to represent a generic speech segment. If the command and filler likelihoods are almost the same, then there is a low confidence on the result. On the contrary, if the likelihoods are very different, the confidence on the recognition result should be high. After several tests it was found that a good confidence measure should use also CM_{OOV} as follows:

$$CM_{cmd} = \frac{e^{-\log P_F}}{1+e^{-(\log P_{cmd} - \log P_{PLM})}} = \frac{CM_{OOV}}{P_F} \qquad (2)$$

where P_F is the filler model likelihood.

The evaluation of these two confidence measures is given in section IV.

III. SPEECH RECOGNITION SYSTEM

A. Overview

Our system was developed to operate on-live in an embedded platform to perform command sequence recognition in noisy environments. The vocabulary has 254 Portuguese commands and the recognition system is based on continuous density Hidden Markov Models (HMM). Each command model is defined by three different ways: whole-word models, a sequence of phone models (mono-phones) or a sequence of right-left context phone models (tri-phones). To increase the system robustness a modified Advanced Front-End ETSI standard for Distributed Speech Recognition is used as well as robust voice activity detection [9].

B. Database

The Tecnovoz corpus has 232,000 Portuguese utterances (around 11,040 minutes) with 254 commands and 408 phrases. For model training and testing only command utterances with signal-to-noise ratio (SNR) higher than 15 dB were used, resulting in 137,237 utterances as shown in Table I.

Table I
Database command utterances

Train	103,001 (75 %)
Test	27,382 (20 %)
Development	6,854 (5 %)

To perform confidence measure tests and train the filler model, 25,069 phrase utterances with SNR higher 15 dB were used.

C. Training

The model training was carried out using HTK tools [10]. There are three sets of models: 254 whole-word models, 40 mono-phone models and 872 tri-phone models. Each model set was trained separately. For the present study we used whole-word models with 8 mixtures, mono-phone models with 16 mixtures and tri-phone models with 8 mixtures. All these models have left-to-right topologies and equal number of states (3 emitting states per phone).

D. Testing

The performance of each model set is shown in the following Table.

Table II
Command model's performance

Whole-word	96.61 %
Mono-phone	91.41 %
Tri-phone	97.03 %

Theses results are archived using HTK test tool with command utterance boundaries manually defined and a simple command grammar. However, a more realistic test should use our speech recognition system. In this case, the system performance decreases, as shows in Table III, because the full utterance is given to the recognition engine, letting the VAD detect command boundaries. Also, a different command grammar is used with background noise and silence models, as described in Figure 2.

Tri-phone models archived, globally, the best performance. Whenever whole-word models have better performance than tri-phone models, those models were selected for the system dictionary, arising in a "mixed-model" set.

Table III
Models performance

Whole-word	94.31 %
Mono-phone	84.46 %
Trip-phone	94.67 %
Mixed-model	95.84 %

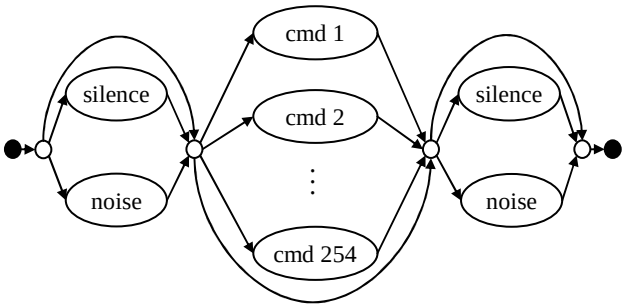

Fig. 2. Command grammar

With the mixed model set, which has 94 whole-word models and 160 tri-phone models, we archived a performance of 95.84 %.

IV. Experiments and Results

A. OOV Experiments

To test the performance of the confidence measures a large number of well-recognized as well as misrecognized results are needed. For this purpose, we selected 25,069 database phrases, in a total of 3,141 minutes, to use as OOV examples. The VAD system accepted around 2,489 minutes of speech, resulting in 42,532 misrecognized commands. This corresponds roughly to 14 commands per minute, in average.

As expected, analysing CM_{OOV} for well- and misrecognized commands, we found a clear difference between their values, enabling OOV word detection.

B. Confidence Measures

From the 27,382 commands in the test database 1,139 were misrecognized, which gives less than 5 misrecognized samples per command, in average. Many of these errors are confusions between similar commands, so the likelihood comparison does not guarantee a reliable confidence measure. However, more serious errors are the 1,298 insertions which were detected. In these cases it is possible to use likelihoods to discard these errors or give a weak confidence measure to them. Using CM_{cmd} we note a clear separation between insertions errors and well-recognized commands as illustrated in Figure 3.

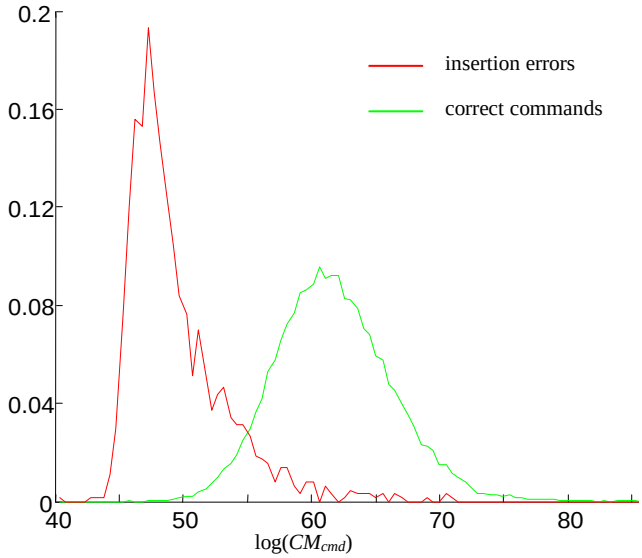

Fig. 3. Estimated Pdf's for log(CM_{cmd}) for insertion errors and for correct commands.

C. Results

To evaluate confidence measures, it is common to use the so called ROC (Receiver Operation Characteristics) curves [11] or DET (Detection Error Trade-off) curves [12]. A DET curve is a plot of false rejection error (type I error – FR) rate against false acceptance error (type II error – FA) rate, by varying confidence measure threshold.

For the OOV case, the DET curve allows us to determine the optimum operation point from which we can set an optimum threshold value of the confidence measure, CM_{OOV}. There are several ways to define an "optimum" operation point, such as "equal error rate" (point where FA rate is equal to FR rate), minimum distance to the DET plane origin point and minimum sum of FA and FR rates [13]. We use the last one because it minimizes the total error given by the OOV detection system.

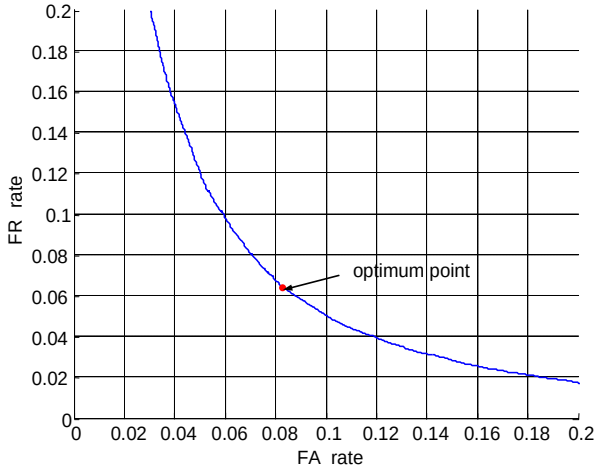

Fig. 4. DET curve for CM_{OOV}

The optimum operation point corresponds to 8.22 % of FA rate and 6.51 % of FR rate. At this point the CM_{OOV} threshold is 0.2324. Any recognized command with a CM_{OOV} below this threshold is automatically rejected. Otherwise, it is accepted and CM_{cmd} value is computed. Using CM_{OOV} and the optimum threshold, the system rejects 39,034 of the 42,532 misrecognized commands, decreasing FA error from 14 to around 1 command per minute. However, this improvement implies to reject 6.51% of well-recognized commands in a normal situation.

The DET curve for $\log(CM_{cmd})$ is presented in Figure 5. The optimum operation point, using the same criterion, is 8.24 % of FA rate and 5.51 % of FR rate. The corresponding threshold is 54.98. In order to present a normalized confidence measure to the speech recognizer applications, we convert this measure to a value between 0 and 1, using a sigmoid function in such a way that the optimum point corresponds to a 0.5 value. The application can use this confidence measure, for example, to decide if it should prompt the user or not with a confirmation. The obtained results show that 94.49 % of well-recognized commands that were not rejected by OOV detector have a confidence value higher than 0.5.

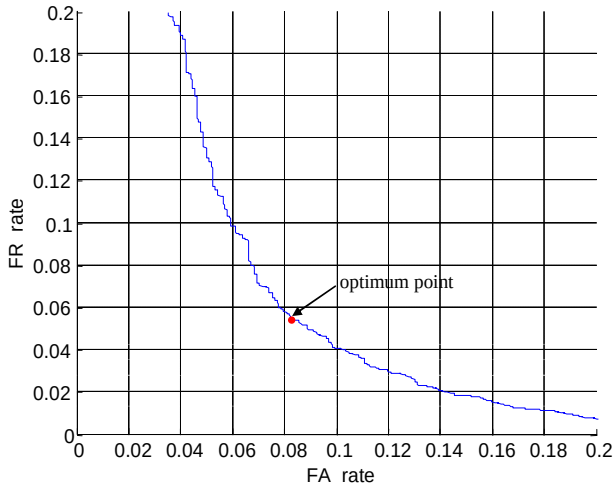

Fig. 5. DET curve for $\log(CM_{cmd})$

V. CONCLUSIONS

In this paper we propose a fast and efficient method to detect out-of-vocabulary words and compute confidence measures for a command-based speech recognition system. Two measures are defined, one for OOV, CM_{OOV}, and the other to estimate a confidence on the recognized commands, CM_{cmd}. A good trade-off between false-acceptance versus false-rejection rate is achieved for OOV. The recognizer provides a normalized value computed with CM_{cmd} as a final confidence measure for the accepted commands.

Confidence measures increases the system robustness but also increases the system response time. More reliable solutions are available but require more computational recourses. We propose a way to compute efficient confidence measures (just two more log-likelihood computations) with a small increase of the computational load, allowing the operation of our embedded system in real time.

REFERENCES

[1] S. Young, "Detecting Misrecognitions and Out-Of-Vocabulary Words", in *ICASSP 1994*, Adelaide, Australia, April 1994, pp. II-21–II-24.

[2] F. Wessel, R. Schluter, K. Macherey and H. Ney, "Confidence Measures for Large Vocabulary Continuous Speech Recognition", *IEEE Transactions on Speech and Audio Processing*, vol. 9, pp. 288-298, March 2001.

[3] T. Kemp.and T. Schaaf "Estimating Confidence Using Word Lattices", in *EuroSpeech 1997*, Rhodes, Greece, September 1997, pp. 827-830.

[4] J. Razik, O. Mella, D. Fohr and J.-P. Haton, "Local Word Confidence Measure Using Word Graph and N-Best List", in *InterSpeech/EuroSpeech 2005*, Lisbon, Portugal, September 2005, pp. 3369-3372.

[5] R. Sukkar and C.-H. Lee, "Vocabulary Independent Discriminative Utterance Verification for Nonkeyword Rejection in Subword Based Speech Recognition", *IEEE Transactions on Speech and Audio Processing*, vol. 4, pp. 420–429, November 1996.

[6] H. Jiang, F. Soong and C.-H. Lee, "A Dynamic In-Search Data Selection Method with its Applications to Acoustic Modeling and Utterance Verification", *IEEE Transactions on Speech and Audio Processing*, vol. 13, pp. 945–955, September 2005.

[7] A. Veiga, F. Perdigão, "An Efficient Word Confidence Measure Using Likelihood Ratio Scores", in *Proc. International Conf. on Language Resources and Evaluation*, Lisbon, Portugal, May 2004, pp. IV-1525–IV-1528.

[8] G. Bouwman and L. Boves, "Utterance Verification Based on the Likelihood Distance to Alternative Paths", in *International Conf. on Text, Speech and Dialogue*, Brno, Czech Republic, September 2002, pp. 213-220.

[9] C. Neves, A. Veiga, L. Sá, F. Perdigão, "Efficient Noise-Robust Speech Recognition Front-End Based on the ETSI Standard", in *ICSP 2008*, Beijing, China, October 2008, pp. 609-612.

[10] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book Version 3.0*, Cambridge, Cambridge University Press, 2000.

[11] J. Egan, *Signal Detection Theory and ROC Analysis*, New York, Academic Press, 1975.

[12] A. Martin, G. Doddington, T. Kamm, M. Ordowski and M. Przybocki, "The DET Curve in Assessment of Detection Task Performance", in *EuroSpeech 1997*, Rhodes, Greece, September 1997, pp. 1895-1898.

[13] D. Falavigna, R. Gretter and G. Riccardi, "Acoustic and Word Lattice Based Algorithms for Confidence Scores", in *ICSLP 2002*, Colorado, USA, September 2002, pp. 1621-1624.