THE TUNING OF SPEECH DETECTION IN THE CONTEXT OF A GLOBAL EVALUATION OF A VOICE RESPONSE SYSTEM

Laurent MAUUARY and Lamia KARRAY
e-mail: mauuary@lannion.cnet.fr
France Télécom, Centre National d'études des télécommunications, CNET/DIH/RCP,
Technopole Anticipa, 2, avenue Pierre Marzin, 22307 LANNION, FRANCE

ABSTRACT

Field evaluations of automatic speech recognition (ASR) systems clearly demonstrate the importance of efficient rejection procedures for filtering out-of-vocabulary tokens. High performance speech recognition systems also require efficient speech detection. This paper presents an original framework for a global evaluation of speech recognition systems allowing to tune the speech detection module of an ASR system. A global evaluation allows to measure the performances of the speech recognition system from the user point of view and to identify the weak modules of an ASR system. Global evaluations are carried out on PSN (Public Switch Network) and GSM (Global System Mobile) databases. On the PSN database, global evaluation is used to choose the best value for the speech detector threshold. The results also show, that for this optimal value, the rejection of out-of-vocabulary words is currently the main problem to be solved for building high performance speech recognition systems for large public telecommunication applications. On GSM database, global evaluation is used to evaluate the benefits of speech enhancement before speech detection. Results show that the use of spectral subtraction as the speech enhancement technique before the detection drastically improves the speech detection, and consequently the global speech recognition.

1. INTRODUCTION

Voice Response Systems (VRSs) now allow speaker-independent recognition of isolated words over the telephone network. Even in the case of limited-size vocabulary (about a hundred words in order to maintain high recognition performances), two major problems appear in real-life applications. Firstly, the speech detector, which is responsible for conveying the detected speech signal to the recognition module, may transmit noises or truncate words. Secondly, VRSs users may utter non-vocabulary words or sentences. Rejection of extraneous speech and adaptive version of the speech detector to the working environment have been proposed to compensate for these problems [2] [3]. Many field evaluations of ASR systems have been carried out, but none of them takes the errors due to the speech detection module into account [1] [4]. This paper proposes an original framework for a global evaluation of ASR systems that allows to evaluate and tune the speech detection module.

In subsequent sections, the following terms are used:

- "Utterance": any acoustic stream segmented by the speech detector of the ASR system.
- "Correct utterance": utterance including at least one word of the vocabulary (isolated correct words and embedded keywords).
- "Incorrect utterance": noise or speech utterance that does not contain any word of the recognition vocabulary.
- "Substitution error": misrecognition of a correct utterance.
- "False rejection error": rejection of a correct utterance or non detection of an acoustic stream corresponding to a correct utterance.
- "False acceptance error": non rejection of an incorrect utterance.

The speech detector can be responsible for some false rejection errors (e.g. non detection of a correct utterance), for some false acceptance errors (e.g. transmitting noise to the recognition module) and for some substitution errors (e.g. transmitting truncated words to the recognition module). The rejection module is responsible for most of the false rejection errors and also for most of the false acceptance errors (e.g. non rejection of a noise input). The recognition module is responsible for most of the substitution errors. It appears that the tuning of these three modules must be done globally since some errors made by a module may be recovered by another module. For example, the rejection module can reject a noise input, thus recovering the error made by the speech detector.

We call "global evaluation" the evaluation of the speech recognition system, including the speech/non-speech detection module, the non-keyword rejection module and the recognition module.

This paper focuses on the use of global evaluation for the tuning of the speech detection module of an ASR system. Section 2 presents a framework for a global evaluation of automatic speech recognition system. Section 3 describes the speech detection module. The two databases used in this work are described in section 4. The HMM training and recognition processes are presented in section 5. Finally, section 6 gives the results of global evaluations conducted on the different databases.

2. GLOBAL EVALUATION OF ASR SYSTEM

Two specific speech field databases have been elaborated for the global evaluation of different versions of the CNET speech recognition system. Each database is continuously recorded, including the words of the database vocabulary and silence or noise between the words.

The global evaluation of a speech recognition system is based upon the comparison of the reference and the test segments. The reference segments correspond to the hand-segmentation and labeling of these databases. The test segments correspond to the automatic segmentation (by the speech detector) and labeling (by the recognition module) of these databases.

A two step procedure is developed for counting the errors. In the first step, reference and test segments are aligned and tied if their temporal intersection exceeds half the duration of the shortest segment. An illustration is given in Fig. 1, where six reference segments (r1 to r6) and six recognized segments (t1 to t6) are aligned. Some of them are tied since they correspond to the same location (on Fig. 1 a vertical arrow shows the tying).

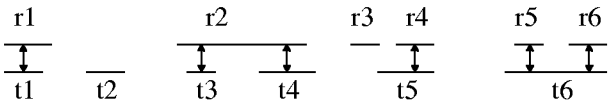

Figure 1: Tying of reference and test segments.

In the second step, the errors resulting from the comparison of tied segments are counted. The different kinds of errors are summarized in table 1.

Table 1: Different Kinds of Errors in the Global Evaluation.

Reference ⇨ / Test ⇩	Inside vocabulary word	Outside vocabulary word or noise	Silence
Inside vocabulary word	Correct recognition or ***substitution error***	***False acceptance error***	***False acceptance error***
Rejection	***False rejection error***	Correct rejection	Correct rejection
Non detection	***False rejection error***	Non detection, but no error induced	Correct non detection

3. SPEECH DETECTION MODULE

An adaptive 5-state automaton is considered in this paper [3]. The 5 states are: *silence*, *speech presumption*, *speech*, *silence or plosive*, and *possible speech continuation.*

The transition from a given state to another one is conditioned by the frame energy and constrained by speech duration. The transition between the different states determines the segment boundaries. In the case of an adaptive detector, the energy requirements are based on an estimate of the signal-to-noise ratio of the observed speech signal.

This technique relies on the comparison between short-term and long-term estimates of the signal energy. The short-term estimate is the mean energy computed over the last K frames ; K specifies the short-term span.

The long-term energy is recursively estimated when the automaton is in the *silence* state, as follows:

$$LTEE \leftarrow (1.0 - \alpha)\,Frame\_Energy + \alpha\,LTEE \qquad (1)$$

where LTEE denotes the long-term energy estimate and α the forgetting factor (we use α=0.99).

A simple detection is described to illustrate how the automaton works.

The initial state of the automaton is *silence*. The automaton remains in this state as long as the frame energy level is not high enough. At the first energetic frame, the automaton changes to the *speech presumption* state.

After having stayed Minimum Speech Duration frames in the *speech presumption* state, the detector changes to the *speech* state.

The automaton remains in the *speech* state as long as the frame energy level is high enough. At the first non-energetic frame, the automaton changes to the *silence or plosive* state.

After having stayed Maximum Stop Closure Duration frames in the *silence or plosive* state, the end of the speech detection is confirmed and the automaton returns to the *silence* state.

4. DATABASES

Two specific speech databases have been developed for global evaluation of speech recognition systems.

The first one, named BAL1000, is a PSN field speech database. 1000 calls to a VRS, currently available to the general public in France, have been continuously recorded, hand-segmented and labeled. These 1000 calls represent 32 hours 25 minutes of signal (speech, noise and silence). The vocabulary consists of 26 command-words. Six classes have been distinguished for labeling: *speech*, *aside* (speech item not adressed to the VRS), *third* (third person speech), *LCBS* (Laughing, Coughing, Breathing, Shouting), *noise*, *echo* (although an echo canceller was implanted, some echo remains from the VRS speech output). We tried to correctly apply this labelling, but there were some cases of ambiguity.

This resulted in a database of 9383 segments including 69 % of vocabulary words, 11 % of out-of-vocabulary (OOV) words and 20 % of noise.

The second database is a GSM laboratory database. The vocabulary of this database consists of 50 words (digits and command words). About 500 speakers calling from different regions of France over a GSM network

repeated the utterances. They were 15 to 75-year old male and female speakers. Several call conditions were distinguished. There were calls from stopped cars (29 %), running cars (23 %), indoors (26 %) and outdoors (22 %).

This database was hand-segmented and labeled. Some extra labels of noise and OOV words were added to the initial vocabulary words. This resulted in a database of 35995 segments including 64 % of vocabulary words, 7 % of OOV words and 29 % of noise (16 % of ambient noise, 9 % of GSM channel distorsion and 4 % of remaining echo).

5. HMM MODELLING

The feature vectors used for our experiments contain 27 coefficients. First, the energy on a logarithmic scale and the first 8 Mel Frequency Cepstrum Coefficients (MFCC) are computed on 32 ms frames, with a frame shift of 16 ms. First and second derivatives of these 9 coefficient vectors are then estimated using 5-frame windows.

The recognition module used is the CNET HMM based system PHIL90 [5] and the rejection of OOV words and noises is made with garbage models [6][7].

Previous results [8] showed that HMM trained using a mixture of field and laboratory data yield 30 % less errors than models trained exclusively using laboratory data. Furthermore, improved non-keyword rejection was obtained [9] when training garbage models using non-speech signals and OOV words from field data.

Experiments using the PSN database are therefore carried out with HMM trained with mixed data and OOV tokens from field data. For the GSM database, vocabulary and garbage HMM are trained with GSM data.

Left-right word HMM with 30 states (Bakis models) are used to model the vocabulary words, and silence models are placed on both sides of the vocabulary models to avoid precise detection of the words to be recognized. A one-mixture Gaussian with a diagonal covariance matrix is associated with each HMM transition.

Since recognition concerns isolated words, no grammar is used.

6. RESULTS OF GLOBAL EVALUATIONS

The first experiment is conducted on the PSN database. The different versions of the evaluated speech recognition system correspond to different values of the threshold used in the speech detector. Results are reported in figure 2.

Results show that false rejection errors increase as the value of the speech detector threshold raises. This is mainly due to the increase of the non detection of correct utterances. On the opposite, false acceptance errors decrease as the speech detector threshold raises. The decrease of OOV words and noise detections explains this fact. The substitution error rate is the lowest error rate and stays almost constant. The best trade-off corresponds to an optimal value of 20 dB for the speech detector threshold.

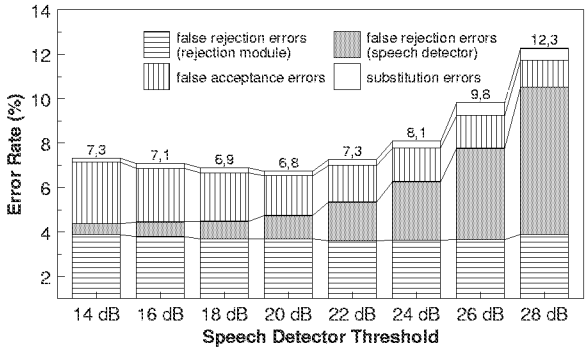

Figure 2: Global Evaluation with the PSN Database as a Function of the Speech Detector Threshold.

For this value, the global error rate is 6.8 % (0.2 % of substitution errors, 1.8 % of false acceptance errors and 4.8 % of false rejection errors). These percentages are computed by dividing the different counts of errors by the total amount of segments (vocabulary and OOV tokens). In order to obtain the performance of the system from the user point of view, the total of the different kinds of errors must be divided by the number of vocabulary words. Evaluation of the system from the user point of view, for the optimal value of the speech detector threshold, is presented in Fig. 3.

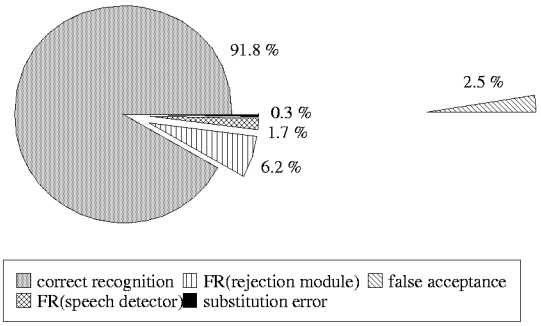

Figure 3: Evaluation of the System from the User Point of View, for the Optimal Speech Detector Threshold.

Looking at Fig. 3, it appears that the 7.9 % of false rejection errors are mainly due to the rejection module (6.2 % compared to 1.7 % of speech non-detection errors). A more detailed analysis of the results shows that noise inputs are correctly rejected (99 %) but that only 90 % of the OOV words are correctly rejected. The false acceptance errors due to OOV words account for 86 % of the total false acceptance errors.

The second experiment is conducted on the GSM database. The optimal value of the speech detector threshold for the GSM database is 14 dB. The global evaluation of the system is presented in Fig. 4 for this best value of the threshold. Fig. 4 shows that the false rejection errors due to the speech detection module drastically increase from 1,7 % on the previous PSN data

to 18,3 % for this GSM data. This is due to the fact that the GSM calls are noisier than the PSN ones.

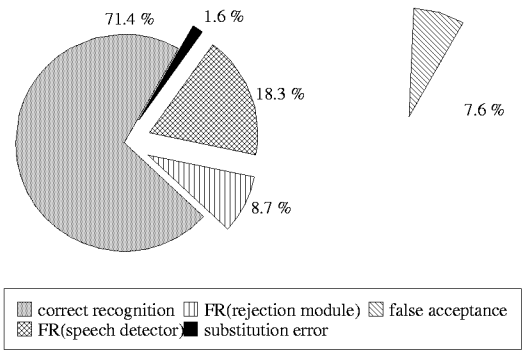

Figure 4: Global Evaluation of the ASR System, on the GSM Database.

False rejection errors due to the rejection module, false acceptance errors and substitution errors slightly increase. Hence the speech recognition module must be first considered for improving GSM speech recognition. This is done by using spectral subtraction as a speech enhancement technique before speech detection. However, recognition is carried out without spectral subtraction. Results are presented in Fig. 5.

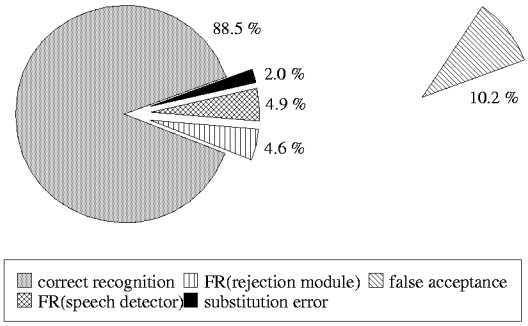

Figure 5: Global Evaluation of the ASR System, using Spectral Subtraction before Speech Detection, on the GSM Database.

Fig. 5 shows that the use of spectral subtraction before speech detection drastically improves speech detection. False rejection errors due to the speech detector drop from 18.3 % to 4.9 %. Since more OOV tokens are also detected, false acceptance errors slightly increase. False rejection errors due to the rejection module fall from 8.7 % to 4.6 %. This is explained by the fact that the speech detection is improved, so less detections are truncated. A previous study [1] showed that half of truncated detections leads to a recognition error (substitution or rejection).

7. CONCLUSION

Global evaluation is useful and necessary for the tuning of the speech detection module of an ASR system. It allows to measure the performances of an ASR system from the user point of view and to identify the weak modules of a system.

For PSN speech, most of the errors are false acceptance and false rejection errors, therefore the rejection of OOV words is currently the main problem to be solved for building high performance speech recognition systems for large public telecommunication applications. The optimal value of 20 dB for the speech detector threshold results in an efficient speech detector for public switch network applications.

Nevertheless, this speech detector shows poor performance for GSM speech. The results of global evaluation show that using spectral subtraction as a speech enhancement technique before speech detection drastically improves the speech detector and consequently the global speech recognition.

8. REFERENCES

[1] C. Sorin, D. Jouvet, M. Toularhoat, D. Dubois, B. Cherbonnel, D. Bigorgne, & C. Gagnoulet. *CNET Speech Recognition and Text-to-Speech in Telecommunications Applications*. IEEE - Workshop on IVTTA, Piscattaway, New Jersey, USA, October 1992.

[2] L. Mauuary. *Improvements of the Performances of Interactive Voice Response Services*. Doctoral Thesis, Université de Rennes, January 1994 (in French).

[3] L. Mauuary and J. Monné. *Speech/non-Speech Detection for Voice Response Systems*. Eurospeech 93, Berlin, Germany, pp. 1097-1100, September 1993.

[4] W.E. Longenbaker. *Successful Applications of Speech Recognition Technology in the Automation of Network-Based Services*. Proceedings of the 7th World Telecommunication Forum TELECOM'95, Geneva, Switzerland, pp. 21-25, October 1995.

[5] D. Jouvet, K. Bartkova and J. Monné. *On the Modelization of Allophones in a HMM based Speech Recognition System*. Eurospeech 91, Genova, Italy, pp. 923-926, September 1991.

[6] J.G. Wilpon, L.G. Miller, P. Modi. *Improvements and Applications for Key Word Recognition using Hidden Markov Modeling Techniques*. Int. Conf. ASSP, Toronto, Canada, pp. 309- 312, May 1991.

[7] R.C. Rose, D.B. Paul. *A Hidden Markov Model Based Keyword Recognition System*. Int. Conf. ASSP, Albuquerque, USA, pp. 129-132, May 1990.

[8] D. Morin. *Influence of Field Data in HMM training for a Vocal Server*. Eurospeech 91, Genova, Italy, pp. 735-738, September 1991.

[9] K. Bartkova, D. Dubois, D. Jouvet, and J. Monné. *Error Analysis on Field Data and Improved Garbage HMM modeling*. Eurospeech 95, Madrid, Spain, pp. 1275-1278, 1995.