HYPOTHESIS DEPENDENT THRESHOLD SETTING FOR IMPROVED OUT-OF-VOCABULARY DATA REJECTION

D. Jouvet, K. Bartkova & G. Mercier

France Télécom, CNET/DIH/DIPS
2 Av. Pierre Marzin
22307 Lannion, France

ABSTRACT

An efficient rejection procedure is necessary to reject out-of-vocabulary words and noise tokens that occur in voice activated vocal services. Garbage or filler models are very useful for such a task. However, a post-processing of the recognized hypothesis, based on a likelihood ratio statistic test, can refine the decision and improve performance. These tests can be applied either on acoustic parameters or on phonetic or prosodic parameters that are not taken into account by the HMM-based decoder.

This paper focuses on the post-processing procedure and shows that making the likelihood ratio decision threshold dependent on the recognized hypothesis largely improves the efficiency of the rejection procedure. Models and anti-models are one of the key-points of such an approach. Their training and usage are also discussed, as well as the contextual modeling involved. Finally results are reported on a field database collected from a 2000-word directory task using various phonetic and prosodic parameters.

1. INTRODUCTION

An efficient rejection procedure is necessary to reject out-of-vocabulary words and noise tokens that occur in voice activated vocal services. Garbage or filler models are very useful for such a task. However they also reject vocabulary words and a trade-off is to be found between false alarm rate and false rejection rate. The aim of the post-processing procedures is either to decide if the recognized hypothesis should be accepted or rejected, or to provide a confidence score for the recognized hypothesis that will be delivered to and treated by the dialog module. Only the first aspect (acceptation or rejection) is considered in this paper.

As the post-processing procedure is applied after the HMM-based decoding, the full hypothesis is known, including the segmentation of the signal. Thus segmental information can easily be used, and have largely been used for re-ordering the N-best hypotheses [1,2]. The formalism developed in [2] lead to using for each segment a model of correct events and a model of incorrect events (anti-model), and computing likelihood ratios according to these 2 sets of models. Various features have been used, among which phonetic and prosodic parameters such as duration [3], energy and voicing, and also some segmental phonetic features estimated by means of neural fuzzy networks [4]. As these features are not used in the HMM decoding, they provide complementary information. Comparing the likelihood ratio value to a threshold allows to reject out-of-vocabulary data. Phonetic [5] and acoustic [6] parameters have already been used on small vocabulary recognition tasks.

All these approaches rely on likelihood ratio test statistics, which are very similar between speaker and utterance verification [7]. The choice and training of the anti-models necessary for the alternate hypothesis is one of the key-points of the approach. Cohorts of speaker models are often used in speaker verification, and sets of phoneme models are used in utterance verification [8]. Here a set of anti-models is proposed. They are trained on various data corresponding to different types of errors. Another key-point, the threshold setting, was not much investigated so far, except a few studies for speaker verification. Here, it is shown that the likelihood ratio statistics differ according to various classes of hypothesis, thus it is necessary to make the decision threshold dependent on the hypothesis to improve the rejection performance.

The paper is organized as follows. Section 2 describes the post-processing procedure. Two aspects are emphasized, namely the choice and training of the anti-models, and the threshold setting according to its dependence on the recognized hypothesis. Section 3 presents and discusses some experiments conducted on a field database collected from a 2000 word- directory task using various phonetic and prosodic parameters.

2. POST-PROCESSING

The aim of the post-processing described below is to decide if, for a given test utterance X, the recognized sequence of words W should be accepted or rejected. A likelihood ratio test statistic is performed, and the answer W is accepted if the resulting value is above a given threshold; otherwise, the answer W is rejected. Although the following description concerns the whole sequence W, the same approach can be applied for each individual word.

2.1 Likelihood Ratio Statistic

Let $\Phi = (\varphi_1, \varphi_2, \cdots, \varphi_I)$ denotes the sequence of phonemes φ_i corresponding to the recognized sequence of words W (the approach can be applied to any set of units, they do not need to be phoneme-like). Let S_i denotes the segment associated to phoneme φ_i and X_i a feature vector measured on segment S_i. The likelihood ratio test statistic $LR(X,W)$ is defined as:

$$LR(X|W) = \frac{\prod_i P(X_i | M_{\varphi_i})}{\prod_i P(X_i | M_{\overline{\varphi}_i})} \qquad (1)$$

where M_{φ_i} and $M_{\overline{\varphi}_i}$ are respectively the model and the anti-model associated to the phoneme φ_i.

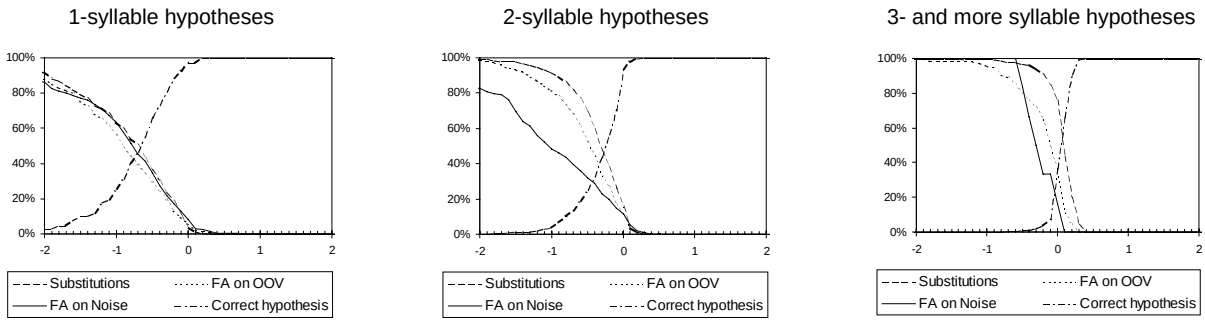

Figure 1 – Log Likelihood Ratio Cumulated Histograms According to the Hypothesis Length in Syllables on Different Types of Data

2.2 Modeling Issues

The same way contextual units are used for acoustic decoding, contextual models are also used for computing the likelihood ratio test statistic. Here again, a compromise is required between a rough modeling (leading to a small amount of parameters that will be correctly estimated) and a detailed modeling (leading to a large amount of parameters that might not be correctly estimated because of a lack of data). Moreover, it is important to note that the contexts, that need to be taken into account, are not necessarily the same for post-processing modeling as for acoustic modeling. Phonetic knowledge is useful to define pertinent models that will provide a good modeling of the post-processing features used. For small vocabularies, a detailed but efficient modeling can be achieved using word- and position-dependent models [9]. However, such a detailed modeling is not manageable for large vocabularies.

Hence, M_{φ_i} and $M_{\overline{\varphi_i}}$ stand for a more precise notation $M_{\kappa(\varphi_i,\Phi,W)}$ and $M_{\overline{\kappa(\varphi_i,\Phi,W)}}$, where $\kappa(\varphi_i,\Phi,W)$ refers to the contextual model index that should be used for the phoneme φ_i according to the fact that it belongs to the sequence of phonemes Φ and that the recognized sequence of words is W.

When phonetic or prosodic parameters are used, a model can be shared by a set of phonemes having the same behavior for what concern the feature(s) under study. For the experiments described in section 3, the classification used depends on the feature considered. For example, for duration based post-processing, the models depend on the length of hypothesis in syllables, and also on the relevant position (last syllable or not) and context (followed by a pause, by a lengthening consonant or not, etc.). For the energy and voicing based post-processing, a smaller set of models was considered which takes into account the fact that the left and right contexts are voiced or not.

The other modeling issue concerns the anti-model $M_{\overline{\varphi_i}}$ associated to phoneme φ_i (in the adequate context). Previous experiments [5] showed that having several anti-models trained on specific sets of data leads to better performances than a single anti-model. Consequently, a set of anti-models will be used $\{M_{\overline{\varphi_i}(k)}; k=1,\cdots,K\}$, each one being trained on a specific set of data corresponding to a specific type of errors. 3 anti-models are considered: one trained from data that are mis-recognized (substitution errors); one trained on out-of-vocabulary data (thus generating false alarms), and an other trained on noise tokens (generating also false alarms). These sets of anti-models can be handled in different ways. For example one can take the best anti-model for each segment feature X_i:

$$P\left(X_i\middle|M_{\overline{\varphi}_i}\right)=\max_k P\left(X_i\middle|M_{\overline{\varphi}_i(k)}\right) \quad (2)$$

However, another approach is possible that takes into account the fact that each type of model comes from a particular set of data:

$$LR\left(X\middle|W\right)=\frac{\prod_i P\left(X_i\middle|M_{\varphi_i}\right)}{\max_k \prod_i P\left(X_i\middle|M_{\overline{\varphi}_i(k)}\right)} \quad (3)$$

This is like handling a multi-model instead of a single model with mixtures. By doing this, a possible correlation between the features of a given hypothesis is taken into account. It is this approach which is used in the following experiments.

To avoid any hypothesis on the shape of the densities, discrete densities are used for models M_{φ_i} and $M_{\overline{\varphi}_i(k)}$.

2.3 Threshold Setting

Figure 1 shows some statistics of the likelihood ratio defined before. These statistics are computed on the training set, and are plotted according to length of the hypothesis (recognized sequence of words) in syllables. 3 cases are considered: 1 syllable, 2 syllables, and finally 3 or more syllables. Statistics are plotted for each type of data: correct hypotheses (i.e. correctly recognized sequences of words), substitution errors, out-of-vocabulary words and noise tokens. Cumulated histograms are reported. For substitutions and false alarms, each point represents the percentage of utterances (of this type) that have a log likelihood ratio greater than a given value (abscise). For correct hypotheses, each point reports the percentage of utterances that have a log likelihood ratio smaller than a given value. The point at which curves cross indicates that for this threshold (abscise) percentage of false alarms accepted is the same as the percentage of correct answers rejected (thus yielding false rejections).

Comparing the 3 graphs, it is clear that the threshold (abscise) at which the curves cross, depends on the length of the recognized hypothesis in syllables. The threshold is around –0.7 for 1-syllable hypotheses, around –0.3 and –0.5 for 2-syllable hypotheses depending on the type of data, and greater than 0 for

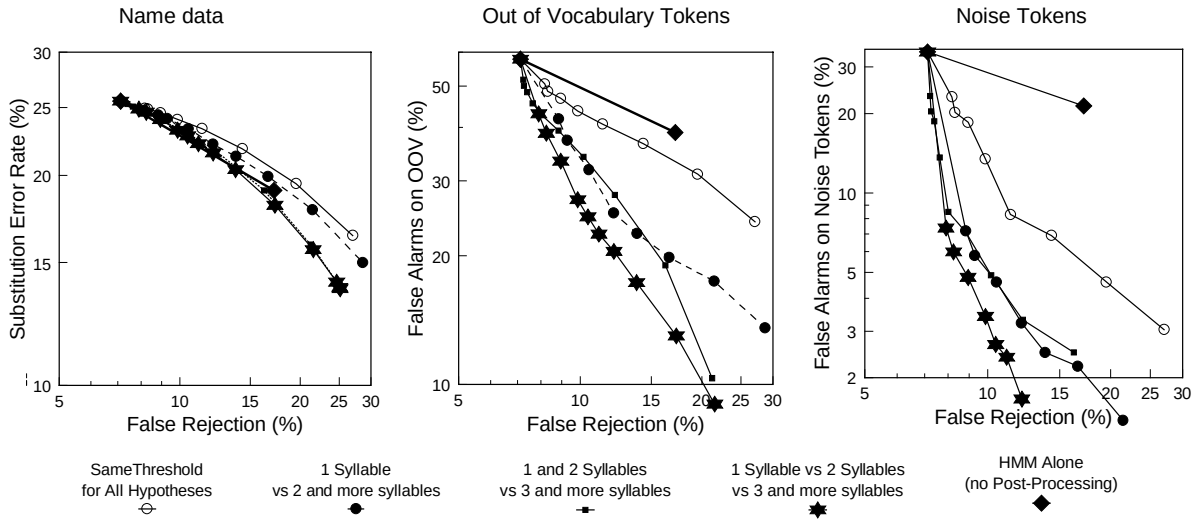

Figure 2 – Error Rates after Post-Processing with Different Threshold Settings.

3- and more syllable hypotheses. From these graphs, it seems quite obvious that the optimal decision threshold should depend on the hypothesis. Thus the post-processing test for accepting the hypothesis (answer) will be:

$$LR(X|W) > \theta(W) \qquad (4)$$

A simple dependency just consists in taking into account the hypothesis length in syllables: $\theta(W) = \theta(length(W))$. The results obtained with such an approach are reported on figure 2 and discussed later.

2.4 Features

In the experiments presented in section 3, phonetic and prosodic parameters are used: phoneme duration, phoneme energy, voicing degree [9] and a consonant/vowel phonetic feature estimated by means of a fuzzy neural network [4]. These features are computed on the segments S_i associated to the phonemes φ_i and resulting from the Viterbi alignment. As the duration measure is a normalized duration, models are estimated separately for various hypothesis lengths in syllables. The energy of the segments is also normalized.

3. EXPERIMENTS

Previous experiments on small vocabularies were reported in [5] and a single decision threshold was used. Here the threshold setting is studied, and a much larger vocabulary is considered.

3.1 System and Database Overview

The speaker-independent speech recognition system is HMM based and relies on continuous densities. Mel frequency cepstral coefficients are computed every 16 ms, as well as their first and second order derivatives estimated over 5 frame windows. A flexible modeling relying on contextual models of the phonemes is used [10]. The acoustic HMM parameters are trained on a specific database design so as to exhibit as many phonetic contexts as possible.

The HMM modeling also includes a garbage model made of a loop of context-independent phonemes, noise models and a silence model. By modifying the penalty (cost) associated to the loop, one can modify the tradeoff between false rejections and false alarms and substitutions. On Figures 2 and 3, the 2 points corresponding to the "HMM alone" are obtained by running the decoder with 2 different weights. The rightmost point allows to measure the benefit of the post-processing procedures. The post processing is applied from the leftmost point, that is from a compromise leading to a smaller false rejection rate but a higher false alarm rate. The post-processing procedure checks each non-rejected answer and decides either to keep it or reject it. If the system rejects a correct answer or a substitution error, that increases the false rejection rate; if it rejects a false alarm on out-of-vocabulary data or on noise tokens, that reduces the false alarm rate. In order to obtain various compromises after post-processing, various decision thresholds are considered.

The database was collected from a vocal service in operation, which allows to obtain the phone number of, and to get connected to, any CNET Lannion employee simply by pronouncing its name. The name can be pronounced isolated or can be preceded by the first name. This leads to a 2000-word vocabulary. The data collected from several thousands calls over several months is used here. Part of the data is used to estimate the parameters of the post-processing models. The remaining part is used to evaluate the performances.

3.2 Threshold Setting

The graphs on Figure 2 report the results obtained with the voicing based post-processing using different thresholds. 3 error rates are reported: substitutions on vocabulary data (left), false alarms (FA) on out-of-vocabulary (OOV) data (middle), and false alarms on noise tokens (right). The first curve (circles) corresponds to a single threshold whatever the length of the hypothesis is, the second curve (dots) relies on 2 thresholds, one for the 1-syllable hypothesis and one for the other hypotheses. The third curve (squares) uses a threshold for the 1- and 2-syllable hypotheses and one for the others, finally, the fourth

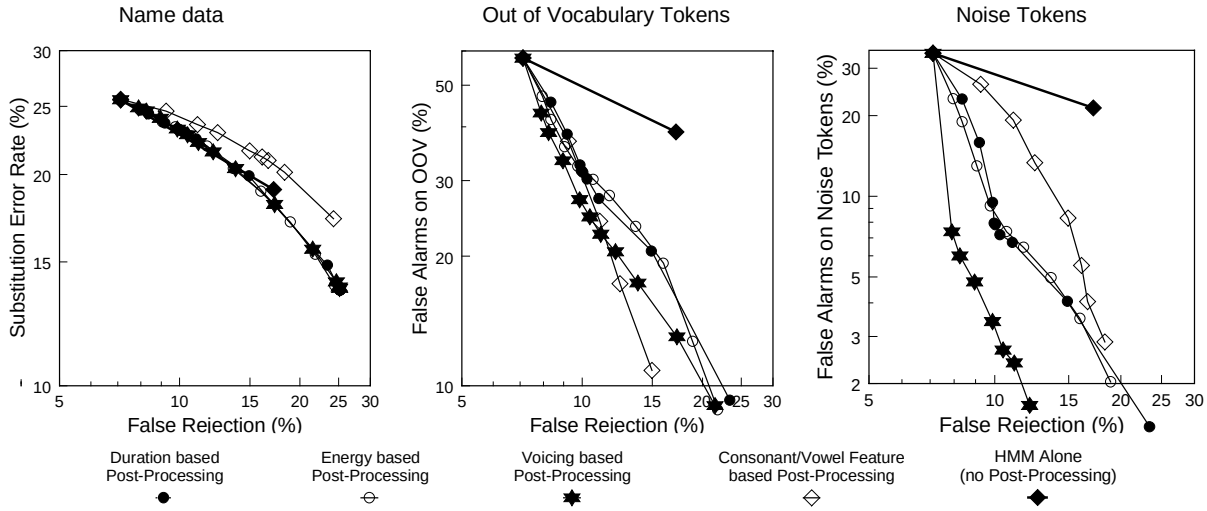

Figure 3 – Error Rates after Post-Processing with Various Features

curve (stars) uses 3 thresholds. In order to obtain a curve, the log likelihood ratio is compared to $\theta(length(W))+\rho$, where ρ is independent on the hypothesis and was varied in order to obtain different false rejection rates and associated false alarm and substitution rates. These figures clearly show that checking all the hypotheses with a same threshold does not yield very good results (top curve – circles) . On the opposite, adjusting the threshold according to the hypothesis length provides good results, and allows to reject many false alarms.

3.3 Modeling Issues and Features

Figure 3 (same axis meaning and data sets as for Figure 2) compares the performances achieved with the different features. The consonant/vowel phonetic feature yields a larger false rejection rate than the other features. This may be due to the use of a single threshold. On out-of-vocabulary data, all features behave similarly. The main differences are observed on the noise tokens. This might be due to the fact that for the consonant/vowel feature a single discrete anti-model was used. Thus training set statistics on noise tokens may have vanished when combined with statistics on substitutions and out-of-vocabulary tokens as noise tokens are less frequent. This problem is overcome when separate anti-models are used.

4. CONCLUSION

It is shown that post-processing the HMM based hypothesis with various phonetic and prosodic parameters largely reduces the false alarms rate on out-of-vocabulary data and noise tokens for a rather difficult task, namely a 2000-word directory access task. A study of the likelihood ratio values showed that the decision threshold has to be dependent on the hypothesis under test. Experiments have confirmed this point, the best post-processing results are indeed obtained using several decision thresholds that depend on the hypothesis length in syllables. Using several anti-models trained on separate types of errors was also efficient and middle part of Figure 1 indicates that the threshold should also depend on the optimal index k of the anti-model in equation (3). Further work is still necessary to fusion the various post-processing likelihood scores based on different features in order to further improve performances and to select, if relevant, the best features for each segment. Moreover, an automatic procedure should help in determining the optimal set of thresholds according to the cost of the various types of errors.

REFERENCES

[1] R. Schwartz, S. Austin, F. Kubala, J. Makhoul, L. Nguyen, P. Placeway & G. Zavaliagkos: "New uses for the N-best sentence hypothesis within the Byblos speech recognition system"; *ICASSP*, San Francisco, USA, 1992.

[2] M. Lokbani, D. Jouvet & J. Monné: "Segmental post-processing of the N-best solutions in a speech recognition system"; *Eurospeech*, Berlin, Germany, 1993, pp. 811-814.

[3] K. Bartkova, D. Jouvet & T. Moudenc: "Using segmental duration prediction for rescoring the N-best solutions in speech recognition"; *ICPhS*, Stockholm, Suede, 1995, Vol. 4, pp. 248-251.

[4] T. Moudenc, R. Sokol & G. Mercier: "Segmental phonetic features recognition by means of neural-fuzzy networks and integration in an N-best solutions port-processing"; *ICSLP*, Philadelphia, USA, 1996, Vol.1, pp. 338-341.

[5] K. Bartkova & D. Jouvet: "Usefulness of phonetic parameters in a rejection procedure of an HMM based speech recognition system"; *Eurospeech*, Rhodes, Greece, 1997.

[6] M. Rahim, C. H. Lee & B. H. Juang: "Discriminative utterance verification for connected digit recognition", *IEEE Trans. on Speech and Audio Processing*, 1997, Vol. 5, No. 3, pp. 266-277.

[7] C. H. Lee: "A unified statistical Hypothesis testing approach to speaker verification and verbal information verification"; *COST Workshop on Speech Technology in the Public Telephone Network*, Rhodes, 1997, pp. 63-72.

[8] R.A. Sukkar & C. H. Lee: "Vocabulary independent discriminative utterance verification for non-keyword in subword based speech recognition"; *IEEE Trans. on Speech and Audio Processing*, Vol. 4, No. 6, pp. 420-429, 1996.

[9] K. Bartkova: "Some experiments about the use of prosodic parameters in a speech recognition system"; *Proc. ESCA Workshop on Intonation*, Athens, Greece, 1997.

[10] D. Jouvet, K. Bartkova & J. Monné: "On the Modelization of Allophones in an HMM-Based Speech Recognition System"; *Eurospeech*, Genova, Italy, 1991, pp. 923-926.