USE OF A CONFIDENCE MEASURE BASED ON FRAME LEVEL LIKELIHOOD RATIOS FOR THE REJECTION OF INCORRECT DATA

N. Moreau & D. Jouvet
France Télécom, CNET/DIH/DIPS
2, Av. Pierre Marzin, 22307 Lannion, France
{nicolas.moreau;denis.jouvet}@cnet.francetelecom.fr

ABSTRACT

Interactive vocal services are based on speech recognition systems which must be able to reject efficiently incorrect utterances such as out-of-vocabulary or noise tokens. A possible approach is a post-processing of the hypotheses delivered by the recogniser, based on the computation of a confidence measure (CM). A recognition hypothesis is rejected if its CM is below a chosen threshold. This paper presents a new way of computing a CM on a recognition hypothesis, based on the calculation of a likelihood ratio for each acoustic frame of the utterance. Promising results are reported on a large vocabulary of a telephone directory task. Significant falls in the error rates are observed, compared to a reference system which include a garbage model, with no post-processing of the recognised words.

Keywords: post-processing, confidence measure, rejection of incorrect data, likelihood ratio.

1. INTRODUCTION

One of the major problems that most interactive vocal services have to cope with is the rejection of incorrect data. Indeed, the users of such services are generally not aware of the constraints of the systems and may be talking in a noisy environment. Thus, automatic speech recognition systems must be able to reject incorrect utterances such as out-of-vocabulary words, speaker’s hesitations or noise tokens. The problem is then to find an acceptable trade-off (depending on the particular constraints of the service) between the different possible types of errors: substitution errors, false rejection errors (when a vocabulary word is rejected) and false alarm errors (when an incorrect utterance is recognised as part of the vocabulary).

Among the directions explored to improve the rejection of incorrect utterances is the computation, for any hypothesis delivered by the recogniser, of a confidence measure (CM) which is supposed to represent the reliability of the hypothesis. This information can be used in a post-processing procedure that rejects the less reliable hypotheses by thresholding the computed CMs.

Many methods of computation have been proposed, such as the use of decision trees [1] or generalised linear models [2]. However, the present work focuses on a different approach based on the likelihood ratio testing theory, largely used for speaker verification and utterance verification [3, 4]. It consists in computing a likelihood ratio (LR) between two likelihood scores representing the probability that the tested hypothesis is correct on the one hand, and its probability of being incorrect, on the other hand.

The study reported in [5] proposes, for a recognised word $W$, a ratio between two acoustic scores. The first one is the score of the utterance on the Markov model of $W$ given by the decoder. The second score is either a combination of scores on a set of competing models or the score on an anti-model associated with $W$. This approach is very similar to the methods used for speaker verification. Other studies have used some segmental information (phonetic and prosodic parameters) to calculate LRs at the phoneme level and combine them to obtain the confidence score of the whole utterance [6, 7]. The likelihood scores for a segment decoded as phoneme $\varphi$ are obtained on the model and anti-model (trained in the chosen features’ space) associated with $\varphi$.

This paper presents another way of computing a confidence measure on a recognition hypothesis. The formalism is very close to the one of the segmental approach [7] where LRs are computed at a sub-word level; here they are computed for each acoustic frame using the same acoustic features as the HMM decoder. The idea is to get a likelihood ratio for the whole utterance by combining the LRs computed for each of its frames. The main issue was to estimate (with distinct training data corpora) two sets of probability densities for each acoustic state of the HMM model: one for the correct events, the other for the incorrect ones.

The following section describes the general principles of the method. We then focus, in section 3, on the different strategies we have investigated to estimate the models used in the post-processing procedure. Finally, section 4 presents some of the results that have been achieved so far on a large vocabulary telephone directory task.
2. POST-PROCESSING PROCEDURE

In this study, post-processing is used to decide if an hypothesis \( W \) (word or sequence of words) delivered by an HMM decoder is to be accepted or rejected. This decision process is based on the theory of statistical hypothesis testing, and more particularly on likelihood ratio tests [3, 4].

2.1 Likelihood Ratio Test

If \( X \) denotes the utterance decoded as \( W \) by the HMM based recogniser, the likelihood ratio statistic, \( LR(X | W) \) is given by:

\[
LR(X | W) = \frac{P(X | W \text{ correct})}{P(X | W \text{ incorrect})} \begin{cases} \geq \omega & \Rightarrow W \text{ accepted} \\ < \omega & \Rightarrow W \text{ rejected} \end{cases}
\]

The likelihood ratio test consists in comparing this quantity to a chosen threshold \( \omega \). If the hypothesis \( W \) being rejected if the ratio is below \( \omega \). In this study, \( P(X|W \text{ correct}) \) and \( P(X|W \text{ incorrect}) \) are approximated by the scores of \( X \) on two sets of distributions modelling respectively the correct events (hypothesis \( W \) correct) and the incorrect events (\( W \) incorrect).

2.2 Modelling Level

The following formalism is close to the one developed for phoneme level modelling [7]. The present work proposes to compute likelihood ratios at the frame level. That requires a set of distributions (modelling correct and incorrect events) for each acoustic state of the Markov models used by the HMM decoder.

At the end of the decoding process, each frame \( x_i \) of utterance \( X \) is associated with a state \( q_i \) of a Markov model. The likelihood ratio of frame \( x_i \) is given by:

\[
LR(x_i | q_i) = \frac{P(x_i | M_{q_i})}{P(x_i | M_{\bar{q}_i})} \quad (2)
\]

where \( P(x_i|M_{q_i}) \) and \( P(x_i|M_{\bar{q}_i}) \) are the scores of \( x_i \) on the model \( M_{q_i} \) and anti-model \( M_{\bar{q}_i} \) associated with state \( q_i \).

The estimation of these models will be discussed in the following section. The likelihood ratio of \( X \) can then be obtained by combining these frame ratios.

Let \( Q = (q_1,q_2,...,q_T) \) be the sequence of acoustic states corresponding to the utterance \( X = (x_1,x_2,...,x_T) \). The likelihood ratio of the whole utterance is computed as:

\[
LR(X | W) = \prod_{i=1}^{T} \frac{P(x_i | M_{q_i})}{P(x_i | M_{\bar{q}_i})} \quad (3)
\]

Finally, we normalised the logarithm of this ratio by the number of acoustic frames of the utterance and defined our confidence measure on hypothesis \( W \) as follows.

\[
CM(W) = \frac{1}{T} \log \left[ LR(X | W) \right] \quad (4)
\]

3. POST-PROCESSING MODELS

The major issue of a likelihood ratio testing approach is the definition of the anti-model since it must model a large range of different events. This section explains what approach has been chosen for this work and how probability densities were estimated.

3.1 Set of Anti-Models

The anti-model \( M_{\bar{q}_i} \) associated to HMM state \( q_i \) consists actually of a set of probability distributions, each one trained on a specific corpus corresponding to a particular type of error. Three distributions have been considered for each state \( q_i \): \( f_{1,q_i} \) for substitution errors, \( f_{2,q_i} \) for false alarm errors on out-of-vocabulary (OOV) tokens and \( f_{3,q_i} \) for false alarm errors on noise tokens. These distributions were trained on acoustic frames associated to \( q_i \) by the decoder within wrongly recognised vocabulary utterances, OOV utterances and noise utterances, respectively.

The model \( M_{q_i} \) for correct events is a single distribution, trained on correctly recognised vocabulary tokens.

3.2 Score Combination

One can imagine many ways of combining the anti-model distribution values in order to get a single anti-model score. For example, the anti-model scores at the frame level, (i.e. terms \( P(x_i|M_{q_i}) \) in the denominator of ratio (3)), can be written as:

\[
P(x_i | M_{\bar{q}_i}) = \left[ f_{1,q_i}(x_i) \cdot f_{2,q_i}(x_i) \cdot f_{3,q_i}(x_i) \right]^{\frac{1}{3}} \quad (5)
\]

if the distribution values are averaged (a geometric average here), or:

\[
P(x_i | M_{\bar{q}_i}) = \max_k \left[ f_{k,q_i}(x_i) \right] \quad (6)
\]

if only the best value is kept.

But the combination of scores may also be processed at a higher level than the frame level. For instance, one can take into account only the type of anti-model distributions (associated with a particular type of error) that gives the best score at the whole utterance level, as shown in the denominator of ratio (7). The hypothesis likelihood ratio (3) then becomes:

\[
LR(X | W) = \frac{\prod_{i=1}^{T} P(x_i | M_{q_i})}{\max_k \prod_{i=1}^{T} f_{k,q_i}(x_i)} \quad (7)
\]

It is also possible to apply the same combination methods with only two of the three types of anti-model distributions, as it will be reported later.

3.3 Distributions

Post-processing distributions have been estimated in the acoustic space used by the HMM decoder. It consists of Mel frequency cepstral coefficients (MFCC) and the energy plus their first and second order derivatives.

Two types of distributions have been evaluated. On the one hand, simple continuous distributions (gaussian densities) were estimated. They offer, in addition to easy computation, the possibility to model correct events with
the observation densities (gaussian densities too) of the HMM model. On the other hand, in order to get more precise models and to avoid any assumption on the shape of the densities, discrete distributions were trained. However, some distributions may not be trained properly because the corresponding HMM states do not have enough training frames. In this case, frames associated to such states within the test utterances were ignored in the post-processing procedure (i.e. they were not taken into account in the product (Eq. 3) and in the normalisation term $1/T$ of Eq. 4).

4. EXPERIMENTS

This section reports post-processing results obtained on a large vocabulary with different sets of densities.

4.1 Experimental Settings

Experiments have been conducted on a field database collected from a telephone directory task including a vocabulary of more than 2000 words (surnames preceded or not by the first name). Two distinct corpora were used: one for the estimation of post-processing distributions, the other for the evaluation tests. Actually, each corpus contains three sub-corpus: vocabulary data, out-of-vocabulary (OOV) data and noise data. The recogniser is speaker independent, HMM based, and observation functions are continuous gaussian densities. The acoustic modelling of words is flexible, based on allophones and silences models. Rejection of out-of-vocabulary and noise tokens is performed only by the post-processing procedure presented before. Results are compared to the ones of a reference system in which, classically, rejection is carried out only by a garbage model (with no post-processing at all).

4.2 Results

Performance tests were performed with the different configurations of the post-processing procedure reported before, by thresholding the confidence measure (Eq. 4) computed for each recognition hypothesis provided by the decoder. As described in Eq. 1, hypotheses whose CM is below a chosen threshold value are rejected. The number of rejected and accepted data in each the three test corpora (vocabulary words, OOV words and noise tokens) gives, respectively, the false rejection rate, the false alarm rate (FA) on OOV words and the false alarm rate on noise tokens.

In order to study the evolution of the trade-off between these different kinds of errors, we varied the rejection threshold and computed the error rates for each of its values. Three types of curves were plotted: substitution error rate, FA on OOV words and FA on noise tokens versus the false rejection rate. The same way, the reference curves were obtained by changing the cost associated to the garbage model of the reference system.

4.2.1 Different Combination Methods

The graphs in Figure 1 show the error rate curves obtained with discrete distributions (using 64 quantization points) and the anti-model scores combination methods proposed in section (3.1.2): geometric average of scores at the frame level (squares), best score at the frame level (circles) and best score at the utterance level (black and white triangles). Dotted curves give the error rates of the reference system. All tests were performed with the three types of anti-model distributions on the one hand, and only two, on the other hand: in that case, only the distributions trained on false alarms (on OOV and on noise tokens) are kept. Figure 1 gives the results with three sets (white triangles) and two sets (black triangles) of distributions for the third combination method. It appears that ignoring the distributions trained on substitution errors improve significantly the rejection of OOV and noise tokens (graphs b and c) while degrading very slightly the substitution error rate (graph a). That’s why we have only reported, in the following, results obtained from using two sets of anti-models (the distributions trained on OOV and noise tokens) for each HMM state.

In any case, the use of a post-processing procedure improves largely the rejection of incorrect utterances compared to a reference system with no post-processing. Thus, at a given false rejection rate of 10%, the false alarm rate is decreased by more than 60% (20% against 60% with the reference system) for OOV tokens (graph b), and up to 90% (3% against 40%) for noise tokens (graph c), while the substitution error rate (graph a) remains stable, around 21%.
4.2.2 Discrete or Continuous Distributions

The post-processing procedure was also evaluated with different types of distributions. Figure 2 gives some results with the average combination method only. The curves concerning the reference system (dotted line) and the post-processing procedure with discrete distributions (squares) are the same as in Figure 1. Continuous gaussian distributions (black diamonds) were tested. Moreover, as the HMM models of the decoder use gaussian observation functions trained on correct vocabulary data, one can imagine that these distributions could be used to model the correct events. Thus, post-processing was evaluated with HMM observation functions of states $q_i$ being taken as correct events models, i.e. the models $M_{q_i}$. Corresponding curves are plotted with white diamonds.

It appears that single gaussian distributions, used in these experiments, do not achieve the rejection performances of the discrete system (except in a small region, around a false rejection rate of 4%). That may seem normal since our simple continuous models are not as accurate as the discrete ones. However the rejection of OOV and noise tokens is still better than with the reference system. In any case, the substitution error rate remains the same as the one of the reference system for false rejection rates below 10%.

Finally, it’s interesting to compare the results obtained when correct events are modelled by gaussians trained on vocabulary tokens (black diamonds) and by vocabulary independent HMM densities (white diamonds). It should be noticed that although the rejection performances are lower in the second case, they still remain significantly better than the ones of the reference system. For a false rejection rate of 10%, we still have a fall of about 30% in the false alarm error rate on OOV tokens (graph b), and about 75% on noise tokens (graph c).

5. CONCLUSION

This paper presented a new post-processing procedure to reject incorrect utterances based on likelihood ratio computation at the acoustic frame level. It requires, for each HMM state, the estimation of a correct event density and several anti-model distributions trained on distinct types of error. Experimental results on a large vocabulary task are reported. They show, for different types of densities and anti-model scores combination methods, that this rejection procedure outperforms a garbage model based system. The best results are obtained with discrete models and anti-models. But even a vocabulary independent procedure using simple gaussian densities and HMM observation functions do much better than the reference system. However, further work is necessary to refine the procedure. Task independent post-processing with discrete distributions as well as the use of multi-gaussian densities for anti-models should be experimented. It would also be interesting to optimise the set of acoustic features used for post-processing.

REFERENCES