INCLUDING DURATION INFORMATION IN A THRESHOLD-BASED REJECTOR FOR HMM SPEECH RECOGNITION

Antonio M. Peinado, Antonio J. Rubio, Juan M. Lopez
Jose C. Segura, and Victoria E. Sanchez

Dpto. de Electronica y Tcno. de Computadores
Facultad de Ciencias, Univ. de Granada
Granada (SPAIN)

ABSTRACT

State duration has been shown as a useful information for recognition. This work tries two ways of including this information in a postprocessing stage, and the effect of incorporating word duration is investigated in each one. In order to diminish the error rate, those utterances that are not clearly recognized can be rejected. Both inclusion ways are tested in a threshold-based rejector. Finally, this rejector is tested with a list of confusing words with those of the vocabulary.

1. INTRODUCTION

In the last years, the HMM technique has reached a very high performance for isolated and connected word recognition and for continuous speech recognition [1]. Applications such as voice dialing of telephone numbers and automatic credit card entry require a high level of safety. In order to improve the recognition systems accuracy, the three basic stages of such systems (signal analysis, recognition and postprocessing) must be improved. This work is concerned with the study of the postprocessing stage on a speaker-independent isolated word recognition system.

State duration densities can be explicitly incorporated in the HMM algorithms, but the computational cost is quite high. An alternative is to include the duration information in the postprocessing stage, as an additional score to that provided by the HMMs. This solution has been shown to be as efficient as the explicit inclusion [1]. In this work, we study two ways of including the state duration in the postprocessing, along with word duration.

For applications that require special safety, a rejection technique can be added in the postprocessing. By means of this technique, those utterances that may yield a misrecognition are rejected. The problem is to decide *a priori* which utterance may yield a misrecognition. We propose a rejection method that consists on defining a score threshold for each HMM of the vocabulary, and find the best way of including the duration information in this threshold-based rejector.

This work has been supported by CICYT (project TIC 88-0774).

2. THE HMM-BASED RECOGNITION SYSTEM

The data were sampled at 8.091 KHz, and preemphasized with a preemphasis factor μ=0.95. Hamming windows were applied to blocks of 256 samples, with an overlapping of 64 samples. Liftered Cesptrum is computed for each frame (with 10 cepstral coefficients and length 12 for the liftering window) and Delta Cepstrum is approximated by linear regression on a ±3 frames environment. Frame energy is normalized to the peak of energy in the word and expressed in the dB scale. Delta Energy is computed from the normalized dB-scaled values of Energy. Finally, an average of all of these parameters is performed every other consecutive frames to compose the feature vectors. The final result is as we had 256-samples frames overlapped 128 samples.

The utterances were coded with a 64-centroid codebook in all the experiences, using the MWDM distance measure [2]. We used one model per word, and, when linear segmentation is used for HMM initialization, 7 states per model.

The vocabulary consists of the ten Spanish digits and the Spanish words {CUERPO, HOMBRO, CODO, MUNECA, MANO, DEDOS}, thought for controlling each motor of a Robot.

The database consists of 40 speakers and 3 utterances per speaker and per word (1920 words), and it was recorded under the normal conditions of work rooms, so certain level of noise, such as the computer noise, is included. The conditions of recording (echo and noise conditions) were variable along the time, from the first speaker up to the last speaker. Two subsets of this database were considered for our experiments: a) the first 20 speakers (DB1) for training, and b) the last 20 speakers (DB2) for testing. With this choice, the error rate is not near 0%, because of the variable conditions of the recording, but we can better observe the variations of error rates and rejections in our experiments, and simulate a real situation of environment change of the recognition system.

3. INCLUDING DURATION INFORMATION

It is usual to use some additional information, as energy and duration, in a postprocessing stage. Since we already use energy information in the feature vectors, we develop our postprocessing only with duration.

State duration and word duration can be included in the postprocessing as two new scores, P_{sd} and P_{wd}, respectively, where,

$$P_{sd} = \sum_{i=1}^{N} \log(p_i(d_i)) \qquad (1)$$

$$P_{wd} = \log(p_w(T)) \qquad (2)$$

where $p_i(d_i)$ is the duration distribution of state i, N is the number of states, T is the utterance duration, $p_w(T)$ is a gaussian distribution of word duration. We consider three ways of calculating the state duration distribution: a) histograms (SD1), b) histograms with normalized duration $d_i \cdot \overline{T}/T$ ($\overline{T}$ is the mean word duration) (SD2), and c) gaussian distributions with normalized duration (SD3). All of them are tested in the experimental results section. Word duration is easily modeled by a gaussian density, considering that the word duration process is a gaussian process (what is basically true). P_{sd} and P_{wd} are incorporated to the word log-score using experimental weights.

4. THRESHOLD-BASED REJECTION

In the postprocessing stage, one possibility to diminish the error rate is to reject those utterances that are not clearly recognized.

Our rejection method consists on defining a score threshold for each HMM λ of the vocabulary, so when the score x of a test utterance O is under the threshold of the recognized HMM, the utterance is rejected. This is possible thanks to a temporal normalization of the HMM score $p(O|\lambda)$ by the word duration T, that extracts the temporal dependence of the HMM score, and, thus, we can compare scores from different utterances (with different durations). The threshold is $\bar{x}-\alpha\sigma_x$, where $\bar{x}$ and σ_x are the log-score mean and the log-score standard deviation, obtained from the training data of a given word. The use of σ_x in the score yields a different threshold for each model λ. Moving this threshold (by the factor α) it is possible to get several rejection percentages ($RDB2$) on the testing database ($RDB2=RDB2(\alpha)$). In the experimental results section, several experiments are performed to find the best rejection.

5. EXPERIMENTAL RESULTS

As reference, we use a system that provides an error rate of 5.52%, using the HMM score $p(O|\lambda)$ only. We develop 4 experiments with 4 new types

of score that include duration information. The inclusion of this information is performed in two steps: first, only state duration is included (experiments 1 and 2), and second, state and word durations are included (experiments 3 and 4).These experiments are:

1) Experiment 1: the log-score used for the utterance O in model λ is as follows:

$$x=\frac{\log(p(O|\lambda))+\alpha_{sd}P_{sd}}{T} \quad (3)$$

In this case, the mean log-score per symbol includes the state duration log-score. State duration is included by the experimental weight α_{sd}. The optimal error rates for the different $p_i(d_i)$ distributions are: SD1) 4.58% (α_{sd}=0.7), SD2) 4.68% (α_{sd}=0.7), and SD3) 5.20 % (α_{sd}=1.7).

2) Experiment 2: the duration information is simply added to the mean symbol score

$$x=\frac{\log(p(O|\lambda))}{T}+\alpha_{sd}P_{sd} \quad (4)$$

The optimal error rates for the different $p_i(d_i)$ distributions are: SD1) 4.79% (α_{sd}=0.03), SD2) 4.79% (α_{sd}=0.03), and SD3) 5.20 % (α_{sd}=0.03).

3) Experiment 3: the same as experiment 1, but including word duration information,

$$x=\frac{\log(p(O|\lambda))+\alpha_{sd}P_{sd}+\alpha_{wd}P_{wd}}{T} \quad (5)$$

Word duration information is included as state duration in exp. 1, using an experimental weight α_{wd}. An experiment (using SD1, α_{sd}=0.7) was developed, obtaining that the error rate is an increasing function of α_{wd}.

4) Experiment 4: the same as experiment 2, but including word duration,

$$x=\frac{\log(p(O|\lambda))}{T}+\alpha_{sd}P_{sd}+\alpha_{wd}P_{wd} \quad (6)$$

The optimal error rate is 4.58% for α_{wd}=0.05 (using SD1, α_{sd}=0.03).

These results show that it is better to include the state duration as in experiment 1 than as in experiment 2. The word duration is slightly useful in experiment 4 but not in experiment 3, but, in general, it does not imply any significant improvement. There are no important differences between SD1 and SD2, but SD3 yields the worst results in all the cases. This can be easily understood since state duration is not a gaussian process.

The rejection results of experiments 1 and 2 are depicted in Fig. 1, along with a rejection curve using a non-normalized log-score (all of them with SD1, α_{sd}=0.7,0.03),

$$x=\log(p(O|\lambda))+\alpha_{sd}P_{sd} \quad (7)$$

We can observe that the best rejection is obtained when the duration information is included in the mean symbol log-score (eq. (3)), and that the threshold-based rejection works better for low rejections (where the curve slope is higher). Also, the necessity of the temporal normalization for the threshold-based rejection is observed.

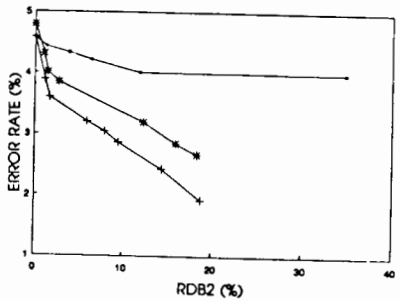

Figure 1.- Error rate vs. RDB 2 for the log-scores of experiments 1 (+) and 2 (*), and a non-normalized log-score (·).

We perform a last trial on the rejector using score (3), SD1 and α_{sd}=0.7. It consists on testing the ability of the system on rejecting words that do not belong to the vocabulary. For that, we apply a database (DB3) containing 3 confusing words with every word of the vocabulary (3×16=48 words total). These words are divided in 3 types, according to the number and type of phonemes in which the word differs:

Type-1) It differs on one or two consonants.
Type-2) It differs on a vowel.
Type-3) It differs on a vowel plus something else (vowels and/or consonants).

Types 1 and 3 correspond to the closest and farthest words to those of the vocabulary, respectively. Figure 2 shows a plot of the word error rate on DB2 and the mean rejection rate on DB3 (*RDB* 3) as function of *RDB* 2(α). We can observe that *RDB* 3 has a good behavior for the same values of *RDB* 2 (the small ones) as the error rate. We can use this graphic to fix a work point of rejection. Table 1 shows, for each type of words on DB3, the percentages of the rejected (R), recognized as correct (C) and recognized as uncorrect (U) words (α=3.9, *RDB* 2=5.93). As we could expect, the lowest percentage R corresponds to type 1 words, and the highest one to type 3. Also note that the percentages C and U diminish from type 1 to 3. A important point of this results is that words that do not clearly belong to the vocabulary are rejected quit right. In figure 3 is depicted a plot of *RDB* 2, *RDB* 3 and $RDB\,3_3$ (rejection on the type 3 subset of DB3) as function of parameter α. *RDB* 2(α) has a exponential behavior, while *RDB* 3(α) has a linear one. $RDB\,3_3(\alpha)$ keeps high in any case.

	R	C	U
Type-1	38.4	46.1	15.3
Type-2	55.5	33.3	11.1
Type-3	84.6	7.6	7.6

Table 1.- Percentages of R, C and U words for each type of words of DB3.

6. SUMMARY

Several HMM log-scores, including temporal normalization and duration information, for utterance evaluation were tested. Among all of them, the best result was obtained using only state duration, including it in the mean log-score per symbol (eq. (3)). No significant differences were found between using normalized state duration or not.

A threshold-based rejector (using the proposed log-scores) was used to diminish the error rate in a simple way. It was shown that the temporal normalization of score is basic to perform this rejection. This rejector can be efficiently used to also reject utterances that do not belong to the vocabulary. Logically, the performance of the rejection of a confusing word is better as more different is that word to any of the vocabulary.

REFERENCES

[1] L.R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition". *Proceedings of the IEEE*, vol. 77, no. 2, Feb. 1989.

[2] A.M. Peinado, P. Ramesh, D.B. Roe, "On the use of energy information for speech recognition using HMM". *Proceedings of EUSIPCO-90*, vol. 2, pp. 1243-1246, Sept. 1990.

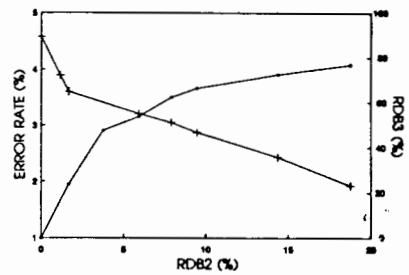

Figure 2.- Error rate (+) and RDB 3 (·) vs. RDB 2.

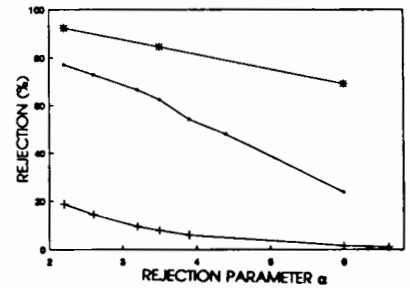

Figure 3.- RDB 2 (+), RDB 3 (·) and $RDB\,3_3$ (*) vs. α.