DISCRETE-UTTERANCE RECOGNITION WITH A FAST MATCH BASED ON TOTAL DATA REDUCTION

Jan Nouza

Department of Electronics and Signal Processing, Technical University of Liberec
Halkova 6, 461 17 Liberec, Czech republic
email: jan.nouza@vslib.cz

ABSTRACT

In the paper, a two-level classification scheme applicable to practical discrete-utterance recognition systems is presented. Both the fast and fine match employ CDHMM whole-word models. The fast match is based on total data reduction, which includes both the minimalization of the acoustic data flow (the numbers of speech frames and features) and the reduction of the basic HMM parameters (the numbers of states and mixtures). The optimal choice of the fast match parameters is a subject of the procedure that aims at minimizing the total classification time while preserving the maximum available recognition accuracy. On a medium-size vocabulary task (121 city names) the fast match reduced recognition time to approx. 20% (compared with the original one-level system) with a negligible loss of accuracy. The time savings were even more considerable in case of a system with multi-mixture HMMs.

1. INTRODUCTION

Continuous density hidden Markov models (CDHMM) have become a widely used technique that is recently employed not only in laboratory prototypes but also in many practically oriented speech recognition systems. It is prefered mainly due to its ability to capture from many points of view complex and highly variable speech signals in a relatively simple parametric form that can be used for modelling both discrete and continuous speech.

The well-known drawback of the CDHMM recognition systems is the high computational load of the classification algorithms. In a standard classification scheme, the number of word-to-model matches and score evaluations increases proportionally with the volume of the vocabulary. When the vocabulary size exceeds a certain critical value it may be difficult or even impossible to perform recognition in real time.

In order to overcome this problem, many practical systems adopt a two-level classification scheme. On the first level, a fast match (FM) makes a preselection of the most likely candidates that proceed to the second, computation more expensive, accurate match (AM). In literature, many different approaches applied in the fast match design can be found; for example, a fast match employing a phoneme based [1,2] or acoustic information based [3] search, vector quantization [4] or rough HMMs [5]. The architecture proposed in [6] utilizes even three hierarchically structured decision stages.

In this paper we present a two-level classification scheme that is applicable in the design of discrete-utterance recognition systems operating with medium-size vocabularies. Both the high accuracy and minimum response time are taken into account.. To achieve high recognition rates we prefer to use whole-word continuous density HMMs. Unlike many other multi-level techniques, our scheme employs the CDHMMs also on the first, fast match, level.

We build our approach up on the results of our previous studies on speech feature selection methods [7,8] and variable frame rate analysis [10]. Both of them have proved the extremely good modelling ability of the CDHMMs, that remains high even if the number of model and speech parameters is dramatically reduced. This is demonstrated in Table 1, where results of several recognition tests conducted on the BUS database are shown (for details see section 5). Though the recognition rate itself, i.e. the Top 1 score, falls down in the case of a low-parameter speech representation, the probability of the correct candidate being on the Top 5 or Top 10 lists is still quite high.

The concept of the proposed fast match scheme is a subject of the next section. In sections 3 and 4 we describe the techniques used for data reduction and the procedure for the fast match optimalization. Experimental results are discussed in section 5.

Table 1. *Classification results from a 121-utterance test task to compare recognition scores and times for different speech and model parameters. The Top N score indicates the probability of the correct model being among the first N candidates. The last column values are relative recognition times in comparison with the baseline experiment.*

Model and speech parameters				Classification scores [%]			Time
Features	States	Mixtures	Frames	Top 1	Top 5	Top 10	[%]
18	14	1	all	97.41	99.43	99.76	100.0
18	14	2	all	97.75	99.51	99.78	248.3
10	8	1	all	95.08	99.06	99.50	37.4
6	8	1	1:2	93.02	98.22	99.09	13.5
5	6	1	1:3	87.65	96.96	98.11	6.4

2. THE FAST MATCH CONCEPT

In a whole-word CDHMM system, both the recognition accuracy and the classification time depend on basic signal processing parameters, namely on the frame rate and the number of features,

and on the model parameters. While the recognition rate becomes saturated for certain parameter values, the time increases nearly proportionally with these parameters. This is illustrated on a practical task (the BUS database) in Fig.1. The plots demostrate that in a standard one-level scheme any attempt to accelerate the recognition significantly (by a factor greater than 2) through a parameter reduction would be paid by a non-negligible loss of accuracy. The only solution is a multi-level classification scheme.

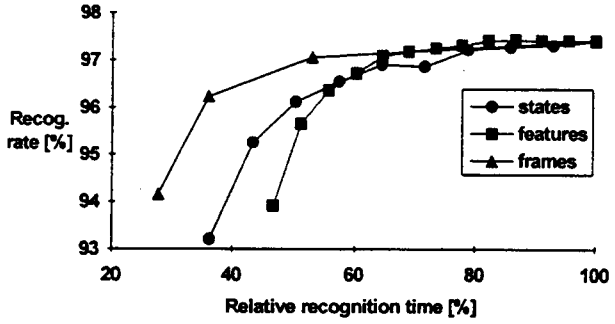

Figure 1: *Recognition rates versus recognition times plotted for varying numbers of HMM states (5 - 14), features (5 - 18) and frames (linear frame undersampling in range 1/4 to 1/1). The 100% recognition time corresponds to 14-state, 1-mixture, 18-feature HMMs and no frame undersampling.*

In a one-level scheme, the time t to classify a speech token depends on these parameters: N - the number of models to be matched, S and M - the numbers of model states and mixtures, P - the feature vector size and F - the number of the token's frames:

$$t = N \cdot T = N \cdot f(S, M, P, F) \tag{1}$$

where function f is determined by the implementation of the HMM classifier.

In our scheme we consider possibilities of reducing all the above parameters: N, S, M, P and F. That is why we speak about *a total data reduction*. The N will be reduced by a two-level scheme, in which models with minimized values of S and M are matched to speech represented by a lower number of features and frames. In such a scheme the recognition time will be a combination of two contributions: $t_R = t_F + t_A$. The first term is the time of the fast match between a reduced token representation (P_F features and F_F frames) and N simplified models (S_F states and M_F mixtures). The t_A corresponds to the accurate match performed with the best N_A candidates using the standard parameter values:

$$t_R = N \cdot f(S_F, M_F, P_F, F_F) + N_A \cdot f(S, M, P, F) = NT_F + N_A T \tag{2}$$

The central point of the fast match design is the search for the optimal values of parameters P_F, F_F, S_F, M_F and N_A, with the aim to minimize the time reduction ratio δ_R defined as:

$$\delta_R = \frac{t_R}{t} = \frac{NT_F + N_A T}{NT} = \frac{T_F}{T} + \frac{N_A}{N} \tag{3}$$

3. DATA REDUCTION TECHNIQUES

The FM parameters are task dependent and must be optimized individually for each task using the available training and testing material. In the following text, several techniques that have been investigated as eligible for the data reduction are described.

Model parameters. Two parameters of the FM models are to be set, the S_F and M_F. The latter can be always chosen equal to 1 because in practice many whole-word CDHMM systems perform succesfully in a single-mixture mode. The choice of the S_F will be a subject of the optimizing procedure described in section 4. Essentially, there are two alternatives:

1. All models have the same number of states.
2. Models have different state numbers that are related, e.g. to the average token length, to the number of acoustic events, etc.

In our investigations the latter alternative was represented by a scheme that assigns a model the number of states according to the following formula:

$$S_F = \text{int}[k \cdot \sqrt{\bar{F}}] \tag{4}$$

where k is a (control) constant and $\bar{F}$ is the average token length.

Speech features. A considerable amount of the computational costs can be saved by minimizing the size of the feature vectors applied in the FM. For any efficient reduction in the speech parameter space it is necessary to order the features according to their discriminative power. Our previous works [7,8] indicated that the best results in the speech feature ordering task are provided by sequential selection methods. We have also shown that speech can be effectivelly represented, particularly for the FM purpose, by a small number of dynamic (time derivative) features.

Speech frames. Another important source of the computational savings is the frame flow. Traditionally, the frame rate is chosen to meet the phonetic and acoustic characteristics of speech. (In practice, 15-25 ms long, half overlapped frames have become almost a standard.) However, it has been demonstrated in several studies [9,10] that a lower or even a variable frame rate can be used in whole-word CDHMM systems without a greater impact on the recognition results. Hence, for the fast match we have considered the application of either of the following techniques:

1. Linear frame undersampling.
2. Nonlinear frame selection techniques based, e.g. on spectral variation functions (SVF).

Our study on the spectral variation functions [10] identified the following type as an appropriate one:

$$SVF(f) = \sum_{p=1}^{P} \left(\frac{1}{L}\sum_{i=1}^{L} c_{f-1}^{p} - \sum_{i=1}^{L} c_{f+1}^{p}\right)^2 \tag{5}$$

where c_f^p is the p-th cepstral coefficient in the f-th frame and L is

a constant (usually in range 2 to 4). As shown in [10], the peaks in the SVF waveform define quasistationary segments that can be represented by a limited (fixed or variable) number of frames.

Candidate lists. The key-point in the FM design is to determine the number of the candidates that are to proceed to the accurate match. The right choice of the number N_A is crucial both from the recognition rate and time points of view. Since we consider the accuracy as the primary goal, we will always adjust the N_A so that the accuracy of the recognition system is not degraded.

In practice we may accept a minor accuracy loss constrained by the condition that $(R - R_R)/R < \varepsilon$, where R and R_R are recognition rates of the one-level and the two-level systems, respectively, and ε is the relative loss (e.g. 0.001). The optimal value of the N_A is then searched as a function of the other FM parameters (P_F, F_F, S_F, M_F) with the aim to achieve the minimum value of the δ_R (defined by eq. (3)) provided the corresponding recognition rate R_R fulfills the above condition.

4. THE FAST MATCH OPTIMIZATION

The complete design of the proposed FM scheme is a subject of a series of experiments conducted on the target task speech material. However, the design process can be automated and run as an unsupervised, system training and optimizing, procedure.

The procedure uses a gradient search in the space of the FM parameters. The search is controlled by the δ_R factor. It can proceed quite fast because it is limited to the FM tests only. The results and the duration of the complete two-level classification are estimated from the available candidate lists and from eq. (3). The procedure consists of the following steps:

1. Split the available speech material into a training and testing part. Find the optimal parameters (*P, S* and *M*) for the one-level classification, and hence also for the accurate match, and evaluate the following parameters: R (recognition rate) and $\overline{T}$ (average time to match one model). Make a subset C of those testing tokens that were recognised correctly.

2. Order the given feature set according the their discriminative power (by utilizing any convenient method). Choose an apropriate frame reduction method such that the reduction rate can be controlled parametrically (towards increasing or decreasing rates). Choose a method for selecting the number of model states, controlled again by an external parameter (e.g. the k in eq. (4)). Define the allowed recognition loss ε.

3. Set the initial values of the parameters that will control the choice of the FM parameters P_F, F_F, S_F, M_F. Start, for example, with: $P_F = P/2$ $F_F = F/2$, $S_F = S/2$, $M_F = 1$.

4. For the given parameter settings: train the FM models and test them on set C to evaluate the overall Top n (n = 1 ... N) scores and time $\overline{T}_F$ (average time to match a model). Find such minimum N_A so that the Top N_A score will meet the allowed loss ε. Estimate the reduction factor δ_R by applying values $N, N_A, \overline{T}$ and $\overline{T}_F$ in eq. (3).

5. Do the same as in step 4 for all combinations of parameter setting $P_F^-, P_F, P_F^+, F_F^-, F_F, F_F^+, S_F^-, S_F$ and S_F^+, where the upper indeces „-" and „+" indicate next lower and higher values of the parameters.

6. If the factor δ_R approaches minimum for the combination P_F, F_F, S_F and M_F stop the procedure, otherwise set these parameters to the combination that achieved the lowest δ_R in step 5 and go to step 4.

Note: Since a lot of evaluations are common for succeeding sweeps through steps 4 and 5, they can be eliminated by utilizing the previously achieved and stored values.

5. EXPERIMENT RESULTS

The proposed scheme has been experimentally evaluated on several speech databases containing either isolated words or multi-word utterances. From all the databases, the most appropriate one, with respect to the purpose of this study, was BUS database. It consists of 121 items, mostly Czech city names, spoken by 48 (male and female) speakers in two repetitions. The database belongs to a real project and includes both very short words (*Ne*) and very long, multi-word, names (*Hodkovice nad Mohelkou*) as well as confusable pairs (*Trutnov - Turnov, Decin - Jicin*, etc.).

The database has been recorded via a telephone set with average SNR being aprox. 20 dB. The 8 kHz/16 bit signal was represented by a vector of 18 features (8 cepstrum coefficients + 8 delta cepstrum + delta energy + delta-delta energy) using 20 ms long frames with 10 ms frame rate.

All the experiments were organized as speaker-independent tests with a half of the database employed in training and the other half involved in testing (approximatelly 6000 testing tokens). The evaluation system used CDHMMs trained by the Baum-Welch reestimation method and tested by the Viterbi algorithm.

In preliminary experiments we focused on determining the optimal model parameters for the standard one-level classification scheme. Best results have been achieved with 14-state HMMs; 97.41 % recognition rate for 1-mixture models and 97.75 % for 2-mixture ones - see also Table 1. For the purpose of the further comparisons, the former settings, i.e. S=14, M=1, P=18 were used as the baseline system parameters.

In the next series of experiments we evaluated the data reduction techniques proposed in section 3. As displayed in Fig. 1, the best recognition rate/time ratio is provided, surprisingly, by the frame reduction. We tested both linear and several non-linear, on the SVF based, frame selection techniques. The results of quite an extensive investigation showed that none of the proposed variable frame rate methods got over the simple frame undersampling technique. The 1:2 and 1:3 frame reduction was found optimal for the fast match. The set of 18 features was ordered by means of the

sequential forward selection method with results nearly identical to those published in [7]. Among the first 10 features eligible for the fast match mostly the dynamic ones were identified. The investigations on the number of model states showed that in the case of the standard AM models the equal number was the optimal choice. For speech with undersampled frames, however, different, word-dependent, state numbers were necessary to assure the correct Viterbi allignment both for short and long utterances.

After these preliminary experiments, we could apply the fast match optimization procedure that was described in section 4. The procedure was provided by the data (R, $\bar{T}$ and C) of the baseline system and run with the parameters that controlled the linear frame undersampling rate, the varying number of model states according to eq. (4) and the number of features. On the given database, the procedure found these optimal settings for the fast match: $P_F = 18, F_F = F/3, \bar{S}_F = 8, M_F = 1$.

How much the FM based on these parameters influenced the total recognition score and time is shown in Table 2. The table displays data as a function of the loss factor ε (that determines the number of the AM candidates, i.e. the N_A). Compared are the results estimated by the optimization procedure with those achieved in the real two-level test. We can observe that the fast match helped to reduce the computation time to less than 20% compared with that of the standard scheme while causing a loss of accuracy that was negligible (about 0.1 %). We may also notice a certain difference between the estimated and real losses. It is a positive byproduct of the fast match that may push a potentially best, but wrong, AM candidate out of the Top N_A list.

Table 2. *Recognition results of the system with a fast match displayed for different values of the loss factor ε that determines the number of the accurate match candidates. Compared are estimated results with those of real tests.*

Estimated results			Real test results		
Loss ε [%]	AM candidates N_A	Time reduction δ_R [%]	Time reduction δ_R [%]	Score [%]	Real test score loss [%]
baseline system:			-	97.41	-
0.1	12	22.9	23.1	97.41	0
0.2	8	19.6	19.8	97.31	0.10
0.5	5	17.1	17.4	97.28	0.13
1.0	3	15.4	15.9	96.89	0.53

In order to verify the robustness of the method we repeated the experiments on the same database but with different train/test data splitting. In all cases the results were very similar. We have also tested the FM scheme together with multi-mixture AM models. The application of 2-mixture HMMs on the same database cost 2.5 times more time while the score approached 97.75%. Using the FM we reached the same score in a time ten times shorter.

6. CONCLUSIONS

A two-level classification scheme applicable to discrete-utterance recognition has been proposed. It offers a considerable reduction of computational costs even if employing continuous density HMMs at both the first and second level. It can be easily adapted for a practical task by means of the fast match optimization procedure described in section 4. The scheme, that is depicted in Fig.2, is well suited, particularly, for applications with a medium-size vocabulary that are to run on a common hardware, like a PC.

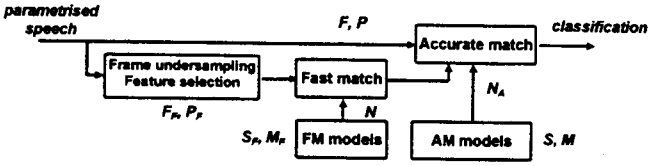

Figure 2. *Overview of the two-level classification scheme*

Acknowledgments

The research was partly supported by Czech grant agency GACR.

8. REFERENCES

1. Colton, D., Fanty, M., Cole, R.: *Utterance Verification Improves Closed-set Recognition and Out-of-vocabulary Rejection.* Proc. EUROSPEECH'95, Madrid, pp. 1067-1070.
2. Psutka, J.: *Desicion Techniques in Large Vocabulary Speech Recognition System.* 13th Europen Meeting on Cybernetics and System Research, Vienna, 1996, pp.1206-1210.
3. Waast, C., Bahl, L., El-Beze, M.: *Fast Match Based on Decision Tree.* Proc. EUROSPEECH'95, pp. 909-912.
4. Bocchieri E.: *Vector quantization for efficient computation of continuous density likelihoods.* Proc. ICASSP'93, pp.692-696.
5. Komori, Y., Yamada M., Yamamoto H., Ohora, Y.: *An Efficient Output Probability Computation for Continuous HMM Using Rough and Detailed Models.* Proc. EUROSPEECH'95, Madrid, Sept. 1995, pp. 1087-1090.
6. Morin, P., Applebaum T.H.: *Word Hypothesizer Based on Reliably Detected Phoneme Similarity Regions.* Proc. EUROSPEECH'95, Madrid, Sept. 1995, pp. 897-900.
7. Nouza, J.: *On the Speech Feature Selection Problem: Are Dynamic Features More Important Than the Static Ones?* Proc. of EUROSPEECH'95, Madrid, pp. 919-922.
8. Nouza, J.: *Feature Selection Methods Applicable to HMM Based Speech Recognition.* Proc. of 13th Int. Conference on Pattern Recognition, Vienna, August 1996.
9. Flammia, G., Dalsgaard, P., Andersen, O., Lindberg, B.: *Segment Based Variable Frame Rate Speech Analysis and Recognition Using Spectral Variation Functions.* Proc. ICSLP'92, Banff 1992, pp.627-630.
10. Nouza, J.: *A Study on Spectral Variation Functions Ap-plied to Speech Signals.* Final report on EU Project No.4678. Aalborg University, Denmark, June 1994.