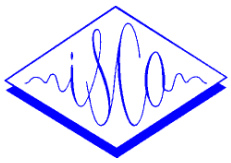

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

Incremental Training of a Speech Recognizer for Voice Dialling-by-Name

L. Fissore ◇ G. Micca ◇ F. Ravera ◇

◇ CSELT - Centro Studi e Laboratori Telecomunicazioni
Via G. Reiss Romoli 274 - 10148 Torino, Italy

Abstract

The present paper describes an incremental approach to the optimization of an HMM recognizer for a dialling-by-name task in a telephone environment. A Bayesian adaptive learning scheme was used to obtain domain-adapted models of sub-word speech units deriving from general, application-independent models. The Maximum A Posteriori (MAP) theory was exploited in order to combine two sources of information, the Domain-Independent (DI) and the Domain-Dependent (DD) speech corpora. This approach was tested on a Voice Dialling-by-Name recognition task over the telephone in a speaker-independent mode. A nearly 92% recognition score was obtained for the best hypothesis using the optimal model adaptation procedure, whereas a 97% score has resulted by inclusion of the second best hypothesis.

1 Introduction

The optimization to the task of HMM parameters presented in the paper belongs to a more general class of adaptation problems in HMM-based speech recognition, which originate from the two following considerations:

a) universal models are unavailable due to the difficulty of collecting enough data accounting for all types of variability encountered in real world environments (new speakers, different microphones, channels, application lexicons, noise levels, etc.);

b) some information about a specific task or environment in which the recognizer is put in operation is usually available. Mostly, this information derives from a certain acoustic data collected in the application environment.

The question therefore is: how can we tackle the "adaptation problem" and optimally combine the statistical robustness of universal models and the task-specificity of speech samples collected in the application environment ? Furthermore, we ideally look for adaptation procedures which allow *incremental* optimization of HMM parameters, since re-processing from scratch of the overall acoustic data base would imply heavy and unacceptable computational loads.

In our experiments, we use an incremental training approach for updating "prior" Domain-Independent DDHMMs through new observations from task-specific and partially Vocabulary-Dependent (VD) data. The scheme is outlined in Fig. 1.

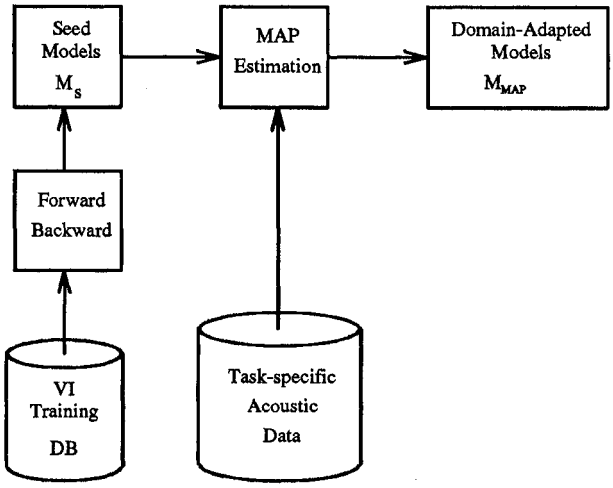

Figure 1: Scheme of the incremental model training

The theoretical background for the MAP estimation of the parameters of the updated HMM models has been outlined in [1, 2, 3]. The "prior" DI parameters were estimated through the Forward/Backward algorithm running on a "universal" database consisting of a few thousands of continuous utterances, partially phonetically balanced and partially belonging to a lexical domain different from the task domain. Successively, the seed models M_S were updated through the MAP estimation algorithm running on DD data, consisting of isolated word utterances in the application domain (Italian surnames). In this second step, only task-specific data was processed, therefore a large saving in computational effort was obtained.

The MAP estimated model M_{MAP} corresponds to the solution to the equation

$$\begin{aligned} M_{MAP} &= argmax_M \; f(\overline{x}/M) \cdot G(M) \\ &= argmax_M \; G(M/\overline{x}) \end{aligned}$$

where $G(M)$ carries "prior" information, $\overline{x} = x_1, x_2, \ldots, x_N$ is the acoustic evidence of N samples in the task domain, $f(\cdot)$ is the likelihood function and $G(M/\overline{x})$ is the "a posteriori" estimate of $G(\cdot)$ given $\overline{x}$.

In turn, prior information on M is represented by the ML estimate

$$M_{ML} = argmax_M \ f(\overline{x_T}/M)$$

where $\overline{x_T}$ is the acoustic evidence from which the seed models are derived by means of ML estimation. In the limit when $N \rightarrow \infty$, the influence of the seed component reduces to zero and the MAP estimation asymptotically converges to the ML estimation.

2 Experimental set-up

The seed models used in the experiments were obtained from a speech corpus (G1) of nearly 12,000 utterances spoken from 150 training speakers on a telephone connected to a local PBX at CSELT premises in Turin. Around 1/3 of the corpus is phonetically balanced, another 1/3 belongs to an E-mail enquire domain and the final 1/3 is related to a railway information retrieval domain. G1 has been used to train the continuous speech recognition component of a speech understanding system [6].

Speech was filtered in telephone bandwidth and sampled at 8 KHz, with 16 bit quantization. Signal was preemphasized with coefficient 0.95. 13 MEL-based spectral features were computed along 256 sample intervals with 10 ms frame shift. A Hamming window of equal length (256) was applied before FFT computation. Finally, a cosine transform was computed to derive 12 Cepstral coefficients. Dynamic features were included by means of a moving window of 5 frames with weights (-2, -1, 0, +1, +2). Two 256-vector codebooks were generated for Cepstrum and ΔCepstrum, a third 32-vector codebook carried energy and Δenergy information. 3 state DDHMMs with loop and forward arcs were trained through F/B iterations computed in two successive steps [5]. In the first step, Context-Independent (CI) phonemes were derived with uniform probability initialization. In the second step, Context-Dependent (CD) models were obtained with Context-Independent initialization. 3 sets of CD models were experimented (Tab. 1): the first one (S1) consisted of 203 speech units (28 phonemes and 175 right biphones) selected basing on frequency counts computed on the G1 corpus. S2 and S3 were generated in DD mode: S2 consisted of 200 units (172 biphones), S3 had 156 biphones and 132 triphones. Both S2 and S3 were selected on the basis of occurrences counted on the application domain lexicon.

Due to the limited context dependency of biphones, S1 and S2 were quite similar, since they differ by only 12%. A silence and a garbage model were jointly trained with all the 3 sets of speech units.

The DD acoustic data consisted of three components, collected by means of an automatic procedure [7]:

D1: 950 Italian surnames (CSELT employees) for a total of nearly 12,000 tokens

D2: 600 surnames covering a large range of phonotactic phenomena in Italian, for a total of nearly 12,000 tokens;

D3: 600 surnames (the same as D2), uttered by different speakers at the IRI-STET TLC Master school in L'Aquila, for a total of nearly 13,000 tokens.

D1 and D2 were used for model updating; they were collected at CSELT premises in Turin on the internal PBX network. D3 was only used for test purposes. The overall DD data were collected in less controlled conditions than the continuous utterances database (G1), and they include a certain amount of noise and extra-linguistic phenomena. Moreover, the D3 component was collected at a different site with respect to D1 and D2, so we suspect that some spectral tilts existed between the frequency characteristics of the corresponding channels. The real-time recognizer for the voice dialling-by-name application is described in [4].

3 Recognition experiments

The first set of experiments was carried out with the D2 test database and the S1 Acoustic-Phonetic Unit (APU) set. The D2 database was used in the incremental training step. Four models were compared (Tab. 2):

E1a , obtained from MLE on the Domain-Independent G1 training database;

E1b , obtained from MLE on the DD, Vocabulary-Independent (VI) D1 database. The procedure was the same as **E1a**, but with only 1/5 of speech data;

E1c , obtained from MAP estimation of incremental training parameters computed from the D1 database;

E1d , the same as **E1c**, but with a more flexible topology for HMMs.

APU set	No.	Type	Structure
S1	203	DI	phones + biphs
S2	200	DD	phones + biphs
S3	416	DD	phones + biphs + triphs

Table 1: Different types of APUs

Mods	APU	Training	WR (%)	ER (%)
E1a	S1	G1	87.86	-
E1b	S1	D1	88.10	2.0
E1c	S1	incr. G1 + D1	90.53	20.4
E1d	S1	incr. G1 + D1 (relax. topology)	91.61	11.4

Table 2: Tests on D2 (Fig. 2). ER = Error Reduction

Mods	APU	Training	WR (%)	ER (%)
E2a	S1	D1	88.11	-
E2b	S1	incr. G1+D1+D2	88.89	6.5
E2c	S1	incr. D1 + D2	90.03	11.4
E2d	S2	incr. D1 + D2	90.53	14.8
E2e	S3	incr. D1 + D2	90.30	12.7

Table 3: Tests on D3 (Fig. 3)

E1b models (88.10% Word Recognition (WR)) performed slightly better than **E1a** models (87.66%) even if far less training data were used, probably because the lexical learning effect of acoustic-phonetic models was encountered, since APUs were now trained in Domain-Dependent mode. Recognition performance benefitted of a 20.4% Error Reduction (ER) by including the DD-VI component in the MAP estimation of HMM parameters during the incremental training step to derive **E1c** models. A further 11.3% ER was gained by relaxing the topological structure of HMMs; in this test, models were given 6 states with skip arcs, differently from the standard structure of 3 states without skip arcs.

In a second set of experiments, D2 database for training and D3 database for testing were used. Furthermore, the speech units were selected in DD mode. 5 types of models were computed (Tab. 3):

E2a , obtained through MLE on DD-VI (D2) data; these models were equivalent to the **E1b** ones;

E2b , obtained from a first MAP estimation on the DD-VI component (D2) with G1-derived seed models and from a second MAP estimation on the DD-VD component (D3);

E2c , these were the same as **E2b**, but with DD-VI seed models;

E2d , these were the same as **E2c**, but with DD speech units;

E2e , these were the same as **E2d**, but with triphones.

An advantage of the E2 series of experiments with respect the E1 series resulted in the vocabulary dependence of the acoustic data of the adaptation database component; on the other hand, the worse recording conditions of the speech signal (different PBX link, higher S/N ratio, naif speakers, etc.) turned out to be a drawback.

The baseline models **E2a** yielded the same recognition performance (88.11%) as in the corresponding **E1b** test (88.10%). The incremental training with VI data yielded models **E2b** that benefitted of a lower improvement (6.5% ER; it was 20.4% in **E1c**), probably due to the mismatching between training and test environment, even if the VD D2 component has been added in the second MAP estimation step; in fact, the incremental training component, D2, still belonged to the **E1** experimental set-up. This statement was strengthened by the successive test **E2c**: the further ER of 11.4% could be fairly explained by the higher relevance that was now assigned to the VD component D2 included in the MAP estimation of the updated HMM models: it amounted to nearly 50% of the overall acoustic data exploited in the HMM parameter estimation (D1+D2), while in the **E2b** models the VD data were 15% only. An even higher improvement (14.8% ER) was observed by adding the VD feature at the APU level in **E2d** models; this result was noticeable since the S1 and S2 APU inventories share almost 88% of APUs. Then, we tried to further specialize the APU set to the task domain. In the **E2c** test, 132 triphones were added with a DD criterion, but recognition performance were not brilliant (12.7% ER with respect to E2b models), probably because we doubled the number of speech units while mantaining the same amount of speech to estimate HMM parameters, and therefore we obtained less robust models. Figs. 2 and 3 show the recognition results from the first best hypothesis (Word Recognition) up to the first 9 hypotheses, for E1 and E2 tests respectively. The distribution of recognition performance across the speakers' population for the E2b and E2d models is given in Fig. 4 (density distribution) and Fig. 5 (cumulative distribution). In Fig. 4 the number of speakers is plotted versus the attained recognition performance. E2d models determine a sharper distribution than E3b models, and yield a shift leftwards, towards higher recognition scores. The same result is described by the cumulative distribution of Fig. 5.

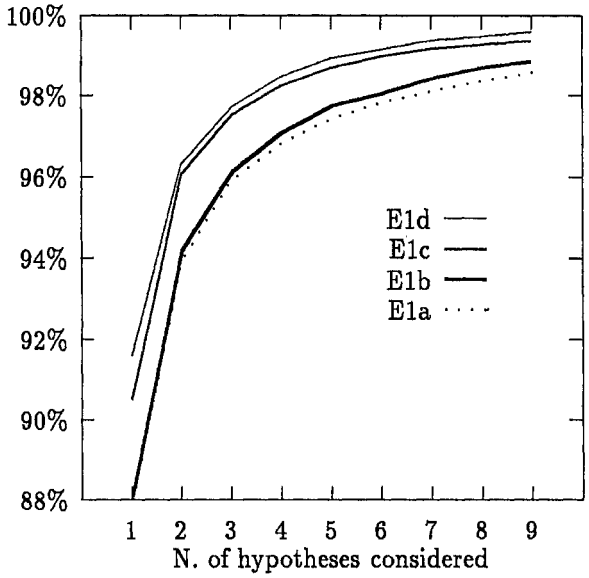

Figure 2: Recognition performance for E1 tests

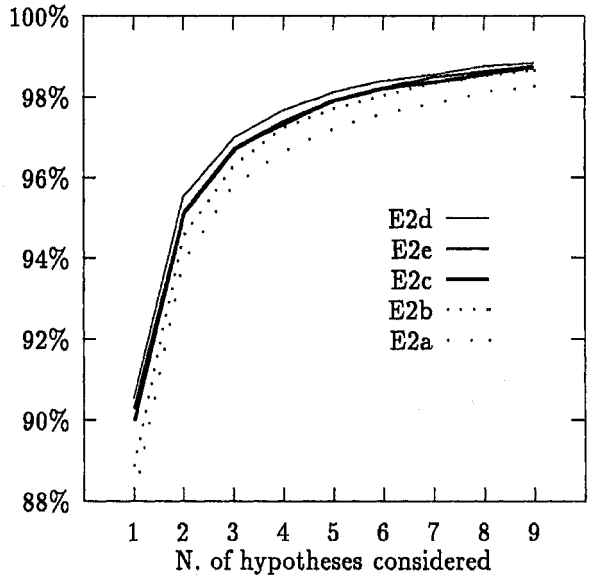

Figure 3: Recognition performance for E2 tests

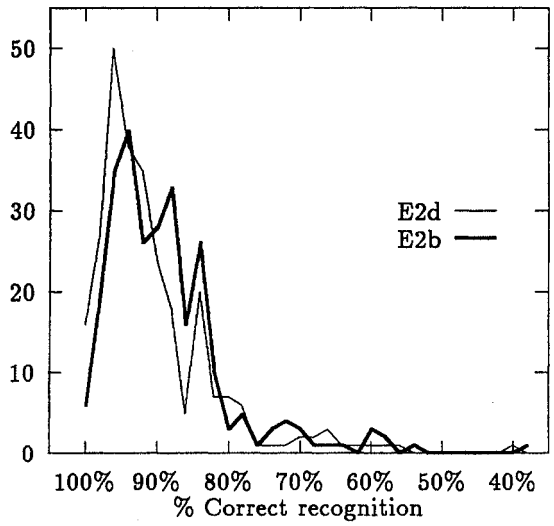

Figure 4: Speakers' performance density distribution

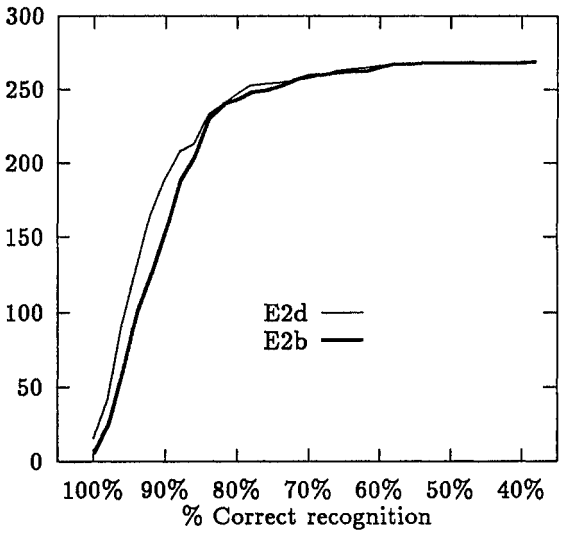

Figure 5: Speakers' performance cumulative distribution

4 Conclusions

An incremental training procedure has been tested, based on MAP estimation of HMM parameters, in order to improve existing HMMs by adaptation to specific tasks. Domain-independent DDHMM parameters were specialized to a surname recognition task for a voice dialling-by-name application. Our tests confirmed that MAP theory constitutes a suitable theoretical framework for adaptation of HMM-based speech recognizers to new testing environments. Furthermore, MAP estimation by incremental training induced a remarkable decrease in processing time and complexity, and allowed a rather unexpensive specialization of Hidden Markov Models to different application tasks.

References

[1] J.-L.Gauvain, C.-H.Lee "Bayesian Learning of Gaussian Mixture Densities for Hidden Markov Models", Proc. DARPA Speech and Natural Language Workshop, Pacific Grove, February 1991.

[2] C.H.Lee, J.-L.Gauvain "Speaker Adaptation Based on MAP Estimation of HMM Parameters", ICASSP 1993, April 27-30 1993, Minneapolis (USA), Vol II, pp. 558-561.

[3] Q.Huo, C. Chan, C.-H. Lee "Bayesian learning of the Parameters of Discrete and Tied Mixture HMMs for Speech recognition", EUROSPEECH 1993, 21-23 September 1993, Berlin, Germany, pp. 1567-1570.

[4] A. Ciaramella, D. Clementino, L. Fissore, R. Pacifici, S. Sperti "Voice Dialling by name in a PBX environment", Joint ESCA-NATO/RSG 10 Workshop *Applications of Speech Technology*, Bavaria, Germany, 16-17 September 1993, pp. 179-182.

[5] L. Fissore, E. Giachin, P. Laface, G. Micca "Selection of Speech Units for a Speaker-Independent CSR Task", EUROSPEECH 1992, Genova (IT), 24-26 September 1991, pp. 1389-1392.

[6] P. Baggia, L.Fissore, E. Giachin, G. Micca, C. Rullent, P. Laface "A Speech Understanding System for Information Retrieval", Int. Jou. of Pattern Recogn. and AI, Vol. 8, N. 1, 1994.

[7] F. Canavesio, G. Castagneri, G. Di Fabbrizio, A. Massone "TESCOS - An integrated workstation to collect large speech databases on the telephone network", AVIOS '93, San Jose', California, 28-30 September 1993.