INITIAL EVALUATION OF A PRESELECTION MODULE FOR A FLEXIBLE LARGE VOCABULARY SPEECH RECOGNITION SYSTEM IN TELEPHONE ENVIRONMENT

*J. Macías-Guarasa, A. Gallardo, J. Ferreiros, J.M. Pardo and *L. Villarrubia*

Grupo de Tecnología del Habla. Dept. de Ingeniería Electrónica. Universidad Politécnica de Madrid. Spain
*Grupo de Tecnología del Habla. Telefónica Investigación y Desarrollo. Spain
macias@die.upm.es

ABSTRACT

We are improving a flexible, large vocabulary, speaker independent, isolated-word recognition system in a telephone environment, originally designed as an integrated system doing all the recognition process in one step. We have transformed it, by adopting the hypothesis-verification paradigm.

In this paper, we will describe the architecture and results of the hypothesis subsystem. We will show the system evolution and the modifications adopted to face such a difficult task, achieving significant improvements using automatically clustered phoneme-like units, semi-continuous HMMs, and multiple models per unit. Also, system behavior for vocabulary dependent and independent tasks and vocabularies up to 10000 words will be tested.

1. INTRODUCTION

At Telefónica I+D, a speech recognition system over the telephone network has been developed, handling about one thousand words in real time with dedicated hardware [1].

We wanted to design and implement a flexible large vocabulary speech preselection module to be run before their system, to allow increasing dictionary size without loosing recognition accuracy; or increasing the number of recognizers per board. So, The main goal is achieving high inclusion rate at minimum cost.

Flexibility is taken in two different senses: easy change of vocabulary and flexibility in the development and testing of different technological alternatives and algorithmic approaches.

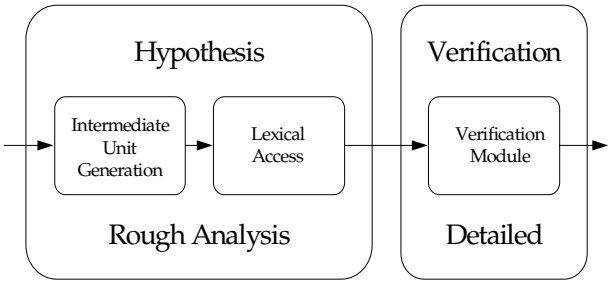

Figure 1: General System Architecture

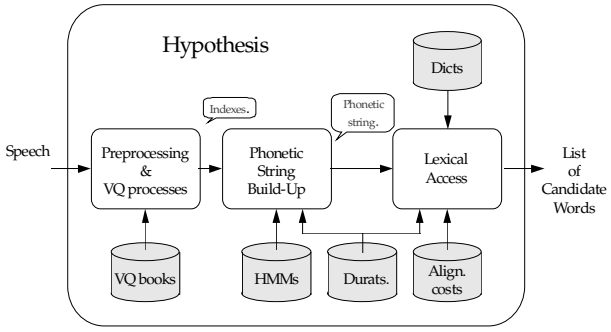

Figure 2: Hypothesis Module Architecture

2. SYSTEM ARCHITECTURE

2.1. Overview

The general system architecture is shown in **Figure 1**. The hypothesis module divides the recognition process in two: the first one generates an intermediate structure (phonetic lattices, in general), which is taken by a lexical access module to give a list of candidate words to the verification stage.

2.2. Preselection module: Detailed architecture

The current implementation of the hypothesis module follows a bottom-up, two stage strategy, as shown in **Figure 2**

Acoustic Processing (AP): The input speech signal is preprocessed (8 MFCCs, 8 delta-MFCCs, cepstral energy and its first derivative) and quantized for discrete HMMs (DHMMs) or soft quantized if semi-continuous HMMs (SCHMMs) are used, with up to 2 codebooks and 256 centroids each).

Phonetic String Build-Up (PSBU): the resulting indexes are passed to the phonetic string build up module which generates a string of alphabet units. We have used the One-Pass algorithm [4] with minor modifications.

Lexical Access (LA): The phonetic string is matched against the dictionary, using a dynamic programming algorithm and alignment costs for unit substitution, insertion and deletion errors [5]

Additional tools were designed for the training stages, dictionary handling, grapheme-to-phoneme conversion, automatic HMM clustering, database analysis, etc.

3. IMPLEMENTATION ALTERNATIVES

We wanted the system to be simple and modular, to allow quick development and testing of different technological alternatives. Actually, this architecture has been already used in our Group with different string generation and lexical access algorithms [3,7]).

We also wanted to be able to easily change any parameter related to the recognition process (number and size of codebooks, units in the alphabet, string generation process control, refinement level in the acoustic model, etc.).

We started from the simplest baseline system (identical to the one used in easier tasks in our Group before [6]), and took decisions based on performance and prior experiences with the architecture.

4. DATABASES AND DICTIONARIES

For our tasks we used part of the VESTEL database [2] (a telephone speech corpus collected over commercial telephone lines, composed of digits, numbers, commands, city names, etc.; and designed to support research in speaker independent automatic speech recognition based on word and sub-word units),

The training set is composed of 5820 utterances, corresponding to 3011 different speakers

The recognition set is further divided in two different parts:

Vocabulary dependent (VD) data: Containing 2536 utterances corresponding to 2255 speakers and graphemes already present in the training set.

Vocabulary independent (VI) data: Containing 1434 utterances corresponding to 1351 speakers. Corresponding graphemes have never been seen in the training data.

Two different sets of dictionaries have been created, for the VD and VI tasks. In the VD case, a dictionary of 1200 was available from the application domain. In the VI case, 2000 word were extracted from the same domain. Additional 5000 and 10000 words dictionaries were extracted, adding words to the available ones from the ONOMASTICA dictionaries.

5. SYSTEM ALTERNATIVES

In the following figures, we will give either recognition or error rates versus the size of the preselection list needed to get those rates. The preselection list size will be given either as “number of words” or as a “percentage” of words calculated over the whole dictionary size (i.e. for a 10000 words task, a 10% in the figures would mean we used a preselection list composed of 1000 words). Using percentages instead of number of words, we will even be able to compare experiments in which dictionary sizes are different.

We wanted to achieve 2% error rate for the tasks under study, using different preselection list sizes depending on the dictionary size (in order for the hypothesis module to be really useful in the overall system). Initial requirements were, for example, for the 1200VD task: 2% error rate for a preselection list 4% of the dictionary size. In the case of the 5000 and 10000 words dictionaries, we estimated a preselection list 10% of dictionary size would be reasonable. Dictionary reduction needed is not proportional to dictionary size due to verification module characteristics.

5.1. Previous Work

We had previous experience with this architecture [6,7], but in an easier task: clean speech, speaker dependent and a vocabulary of 2000 words, achieving the results shown in **Figure 3**, with a DHMM system with 1 codebook and 128 centroids.

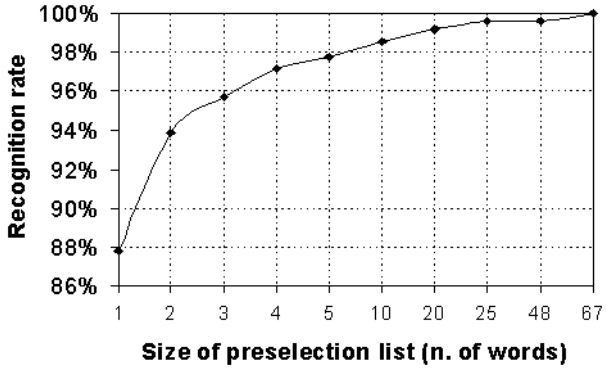

Figure 3: Recognition rate for a clean speech, speaker dependent task with 2000 words dictionary size

5.2. System Evolution

In the task under study, we began doing a preliminary experiment to estimate the top inclusion rate achievable with the acoustic models used, for both DHMM and SCHMM We ran a Viterbi algorithm over full-word models, built by concatenating sub-word units.

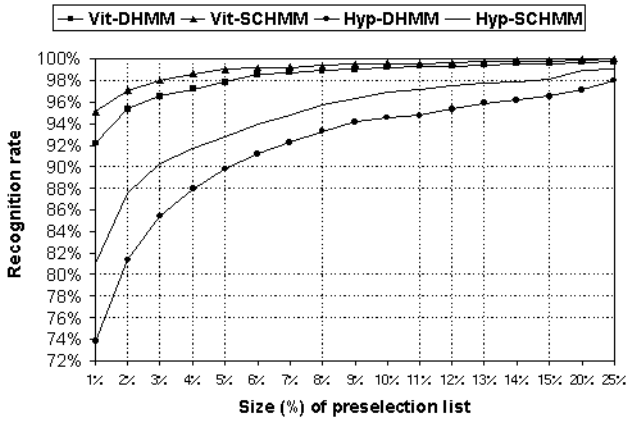

Figure 4: Comparison of Viterbi results for DHMM and SCHMM, along with the equivalent hypothesis subsystems.

In **Figure 4** we compare the Viterbi VD-1200 experiment for the DHMMs and SCHMMs, along with the results for the equivalent

hypothesis subsystems (in all cases, 2 codebooks, 256 centroids each and 25 manually selected units were used, grouping the initial grapheme-to-phoneme units (51) according to linguistic criteria). As shown, the Viterbi DHMM-based system is unable to reach 98% inclusion rate for the required 4% list length. Taken into account that the Viterbi-based is a guided-system, results will suffer a severe further degradation in the non-integrated approach, so that DHMM seems not to be able to model such a variability in the speech data, to get the desired rates, even more in the hypothesis subsystem case.

To assess the system evolution and modification, we measured both error rate and phonetic strings quality, as the PSBU performance is hidden by the LA mechanisms when measuring error rate.

We evaluated the hypothesis subsystem starting with the simplest baseline approach (DHMM, 1 codebook, 128 centroids and 25 units in the alphabet), with very poor results. Progressively incrementing system complexity: using 256 centroids, increasing number of codebooks, modifying the PSBU module in several ways (inter-unit penalty scores, phonotactic restrictions with different grammars, phonetic strings post-processing, etc.), little improvements in overall performance was observed. Some of the modifications included had been successfully applied in the task described in 5.1 above. From the figures we obtained, it was clear that a much more robust acoustic modeling was needed. The LA algorithm proved to be efficient to handle loss of information in the phonetic strings, but the task was difficult enough to require better acoustic resolution (percent correct was preferred to phoneme accuracy, as the LA module handled insertion errors much better than deletions). So, SCHMMs were implemented to improve the acoustic modeling.

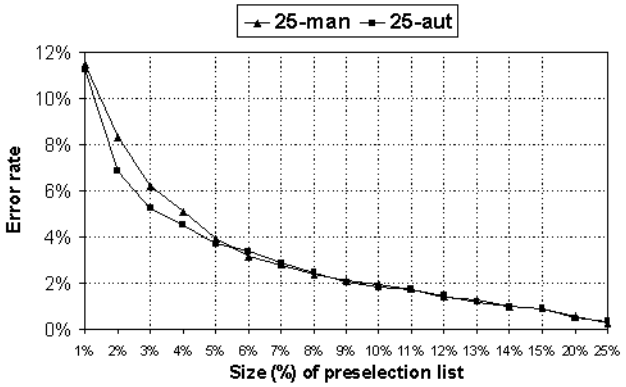

Figure 5: Comparison of manual vs. automatic selection of units.

Results improved considerably with the SCHMM as shown in **Figure 4**, but still didn't meet our requirements. So, we decided to use multiple models per unit. Basically, models were generated for male and female speakers, and several approaches were studied in the PSBU and LA modules integration to use them.

In the same direction, we decided to apply automatic clustering techniques to get the optimum alphabet units instead of the theoretically selected ones, based on an entropy measure.

In the following sections we will describe the most relevant experimental results. In all cases, if nothing else is indicated, we will refer to the "base experiment" using 2 codebooks, 256 centroids each, 25 units and SCHMMs and the 10000 words dictionary for the VI task.

5.3. Units Generation: Manual selection vs. Automatic clustering

In **Figure 5**, we compare results using single-SCHMMs, 25 manually selected units and 25 automatically clustered ones.

As can be clearly seen, results are even better for a wide range of preselection list sizes in the case of the automatic clustering. Error rate reduction is low, but consistent for all the other experiments done. So, automatic clustering is preferred as it is fully automatic.

5.4. Single-SCHMM vs. Multiple-SCHMM acoustic modeling

The introduction of SCHMMs improved system performance significantly, but even better acoustic modeling was needed. So, we decided to use multiple models per unit. As a first approach, we trained two sets of models: for male and female speakers. The comparison between single-SCHMMs and multiple-SCHMM is shown in **Figure 6** for the base experiment. An average error rate reduction of 40% in the range of interest was achieved.

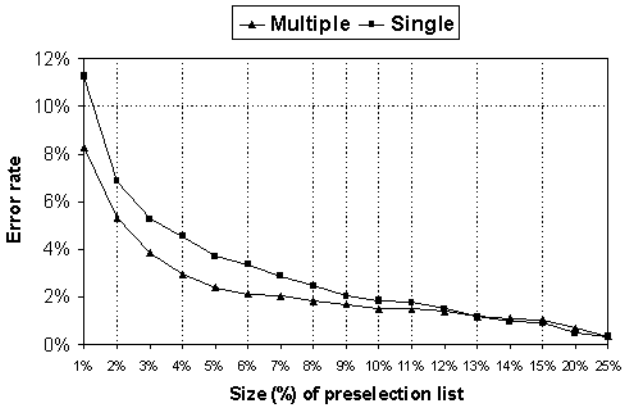

Figure 6: Multiple-SCHMM vs. Single-SCHMM.

5.5. Vocabulary Dependent vs. Independent Tasks

A comparison between VD and VI tasks was also done. Surprisingly, as **Figure 7** shows, the VI results are better, although the complexity is supposed to increase. We think this is due to the lower acoustic similarity in the VI dictionaries (mainly because, in these, words are longer, in average length): Average error rate reduction is 36% for the 5000 words task and 27% for the 10000 one. Same behavior has been observed in the VD-1200 vs. VI-2000 comparison, with an average error rate reduction of 16%.

The VD-1200 and VI-2000 dictionaries were extracted from the application domain and already showed this difference in the relative average word length. Getting the 5000 and 10000 words

dictionaries was done by simply adding words from the ONOMASTICA dictionaries, in a random fashion, so that no control was enforced during the process, apart from trying to add longer words to the VI dictionaries and shorter to the VD ones, to follow the tendency of the application domain dictionaries.

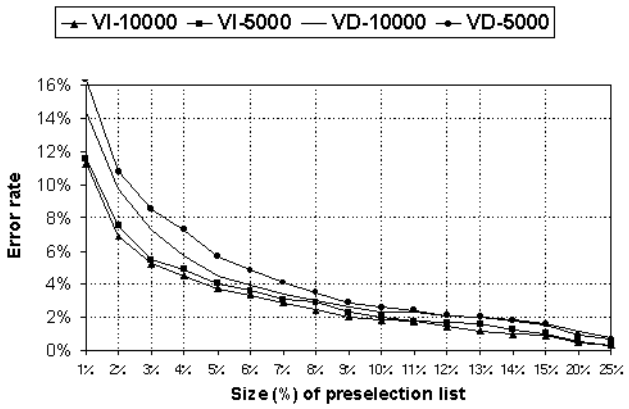

Figure 7: Comparison between VD and VI tasks for 5000 and 10000 words dictionaries. SCHMMs, 25 clustered units.

5.6. Effect of Dictionary Size

We also studied the effect of dictionary size in recognition rate (i.e., behavior of performance vs. the number of words in the dictionary).

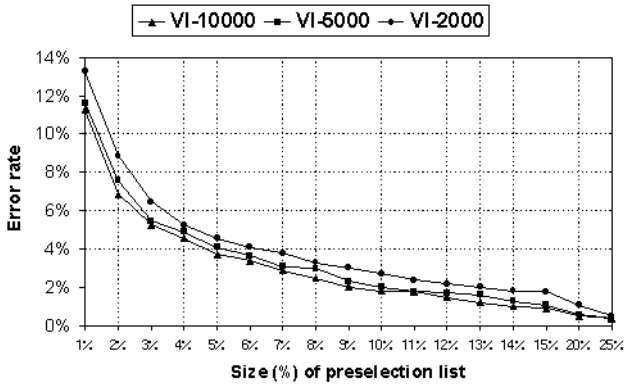

Figure 8: Effect of dictionary size in error rate. SCHMM 25 automatic units, VI data.

In **Figure 8**, we show the results for the VI task using 2000, 5000 and 10000 words dictionaries for the 25 automatically selected units, single-SCHMM). As shown, a decrease in error rate is obtained as we raise vocabulary size for the same relative preselection list length (average error rate reduction is 18% between 2000 and 5000 words dictionaries, 10% between 5000 and 10000, and about 30% between 2000 and 10000). We consider two effects are affecting this result: the degradation is not linear (i.e. increasing vocabulary size does not increase word confusability at the same pace); and a lower acoustic similarity in the 5000 and 10000 words dictionaries, similar to what we commented in section 5.5, in the case of VD vs. VI tasks.

6. CONCLUSIONS AND FUTURE WORK

The DHMM modeling is absolutely insufficient for even the easiest task. As the system lacks any guiding mechanism, the PSBU module is more affected by acoustic variability while trying to concatenate the context-independent models. The LA algorithm needs more powerful improvements in the phonetic strings generation than the ones we had used in other tasks successfully. So, further work is to be done in the following directions: multiple pronunciations, continuous HMM modeling, and using phonetic *lattices* instead of strings.

On the other hand, automatic HMM clustering has proved to be, at least, as efficient as manually generated units, but we also need to analyze different alphabet sizes and their effect in performance.

The effects of average dictionary word length variation in recognition performance will also be studied in detail.

Our main objective right now is integrating the system in the real-time application, so that we will first choose the best system to use in terms of performance and computational requirements; synchronization mechanisms (to allow parallel execution of the PSBU, LA and verification stage); and work will be done in applying beam-searching strategies in the tree-based LA module. In this sense, we are also studying the possibility of using dynamic preselection list sizes, depending on information extracted from the current hypothesizing process.

7. REFERENCES

1. Villarrubia, L., Gómez, L.H., Elvira, J.M., Torrecilla, J.C. *"Context-dependent units for Vocabulary-independent Spanish Speech Recognition"*. ICASSP 96:. 451-454. 1996.

2. Tapias, D., Acero, A., Esteve, J., and Torrecilla, J.C. *"The VESTEL Telephone Speech Database"*. ICSLP 94: 1811-1814. 1994

3. Leandro, M.A., and Pardo, J.M. *"Low Cost Speaker Dependent Isolated Word Speech Preselection System Using Static Phoneme Pattern Recognition"*. Eurospeech 93, Vol. 1, 117-120. 1993

4. Ney, H. *"The use of a One-Stage Dynamic Programming Algorithm for Connected Word Recognition"* IEEE Trans. ASSP Vol. 32, n. 2. 1984

5. Fissore, L., Laface, P., Micca, G. and Pieraccini, R. *"Lexical Access to Large Vocabularies for Speech Recognition"*. IEEE Trans. ASSP Vol. 17, n. 8. 1197-1213. 1989

6. Macías-Guarasa, J. Leandro, M.A., Colás, J., Villegas, A. Aguilera, S. and Pardo, J.M. *"On the Development of a Dictation Machine for Spanish: DIVO"*. ICLSP 94, S22-26, 1343-1346. 1994.

7. Macías-Guarasa, J., Leandro, M.A., Menéndez-Pidal, X., Colás, J., Gallardo, A., Pardo, J.M. and Aguilera, S. *"Comparison of Three Approaches to Phonetic String Generation for Large Vocabulary Speech Recognition"*. ICSLP 94, S36-22, 2211-2214. 1994