RECENT WORK ON A PRESELECTION MODULE FOR A FLEXIBLE LARGE VOCABULARY SPEECH RECOGNITION SYSTEM IN TELEPHONE ENVIRONMENT

*J. Ferreiros, J. Macías-Guarasa, A. Gallardo, J. Colás, R. de Córdoba, 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
{jfl,macias,gallardo,colas,cordoba,pardo}@die.upm.es,luigi@craso.tid.es

ABSTRACT

At ICSLP'96 we presented a flexible, large vocabulary, speaker independent, isolated-word preselection system in a telephone environment, using a two stage, bottom-up strategy [6]. We achieved reasonable performance in large and very large vocabulary tasks, ranging from 1200 to 10000 words.

In this paper, we describe recent studies we have carried out on the system, aimed at two directions: handling of non speech sounds in the speech signal (we consider lips, respiration and click noises); and making the preselection lists dynamic in length, to reduce computational load, in the average. In the first case, we want to model non speech sounds, as these effects are crucial in real-life situations, leading to wrong endpointing and increasing error rates. In the second, we are interested in integrating any available system parameter to calculate the preselection list length to use, having applied both parametric and non parametric methods.

1. INTRODUCTION

When facing the design and implementation of real-world public information services using the telephone network and working in real time, important aspects arise, as opposed to the conditions found in laboratory environments and in laboratory recorded speech databases.

First of all, we are interested in avoiding recognition performance degradation due to extraneous noises made by the service users. In real situations, users are prone to embed non-speech sounds in actual speech utterances. The typical examples in these cases are tongue clicks and lips and respiration noises. Additionally, the telephone network switching devices may add non wanted clicks to the speech signal. All these effects lower recognition performance due to two facts: errors in the endpointing and confusions in the search. In this paper we show a simple strategy to ameliorate them based in specific training of these non-speech sounds.

Additionally, we want to lower computational demands, in order to allow real time execution of the algorithms involved in the recognition process. Automatic speech recognition systems, being based in complex pattern matching techniques are computationaly expensive, so that a lot of techniques have been studied in order to reduce the search effort or to improve the efficiency of the algorithms used. Our system, being a preselection module, offers different alternatives to achieve the same goal. The most obvious one consists of using specific techniques during the search process. For example, implementing the lexical access in a tree structure, or using well known beam search techniques. In our baseline system, the preselection module offers a list of candidate words to the verification stage, and the length of this list is fixed. We try to make this length variable, depending on any available system parameter.

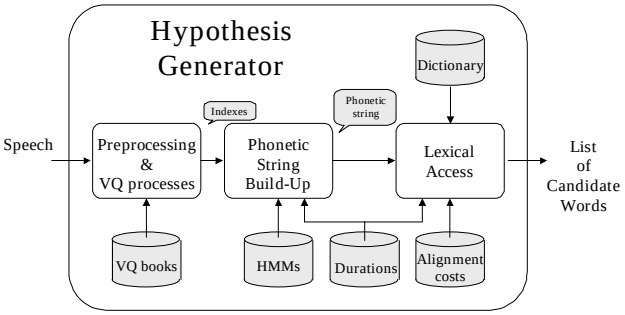

Figure 1: Hypothesis Module Architecture

2. SYSTEM OVERVIEW

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 have implemented a hypothesis subsystem, to be run before the integrated module, to allow increasing dictionary size (without loosing too much recognition accuracy); or increasing the number of recognizers that fit into one hardware board.

The main preselection modules are: Preprocessing (P), Phonetic-string build-up (PSBU) and Lexical Access (LA), and its modular structure is shown in **Figure 1**. The hypothesis module divides the recognition process in two: the first one generates a phonetic string, which is taken by a lexical access module to give a list of candidate words to the verification stage. A detailed description can be found in [6].

3. EXPERIMENTAL SETUP

For our experiments 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.).

The training data is divided in two sets: 5820 utterances, with no noticeable non-speech sounds in the speech signal (from now on,

we will refer to it as the "CLEAN" training set; and 1375 utterances, with non-speech sounds present and manually labeled to allow training specific models (from now on, the "NOISY" training set).

The test material is also divided in two sets: the first one, containing 1434 utterances, in which no noticeable non-speech sounds are presented (from now on, the "CLEAN" test set); and the second one composed of 313 utterances, in which non-speech sounds are known to be present (from now, on the "NOISY" test set). None of the words in the testing material lists have been previously seen in the training set

4. BASELINE SYSTEM AND RESULTS

The inclusion rate of the preselection module actually limits the performance of the overall system. We wanted to achieve 2% error rate for the tasks under study, using 1200, 2000, 5000 and 10000 words dictionaries.

In this paper, we will refer to the 10000 words case, in which we decided that a preselection list of length less than 10% of the dictionary size would be reasonable (for example., a preselection list composed of 900 words (9% of the dictionary size)).

In **Figure 2** we show that we achieved this requirement, when recognizing the CLEAN test set, using 23 automatically clustered phoneme-like unit SCHMM, plus 2 models for silence (referred as "*normal modeling*" from now on).

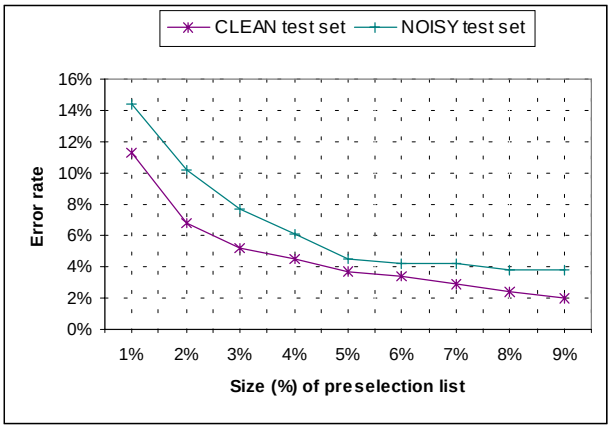

Figure 2: Baseline experiment for the CLEAN and NOISY test sets with no special non-speech sound modeling.

5. HANDLING NON-SPEECH SOUNDS

Presence of non-speech sounds in telephone speech is not negligible at all. Almost 20% of the utterances recorded in VESTEL show these effects. Clicks due to the switching technology, tongue clicks and respiration and lips noises from the speakers are common enough to be taken into account. Although overall degradation is not dramatic, not facing them means certain users will never be correctly recognized, what is clearly undesirable in a real world system.

We have developed a simple strategy to face them, in an attempt to achieve better results while keeping computational cost under control. In the following figures, we give error rates versus the size of the preselection list needed to get those rates. The preselection list size is given as a "percentage" of words calculated over the whole dictionary size (i.e. for a 10000 words task, a 5% in the figures would mean we used a preselection list composed of 500 words). To compare the results obtained, we calculate the relative error rate reduction obtained, averaged between the first 9 steps in the preselection results (that is, from 1% to 9%). In this way, we smooth the different behavior of the error rate reductions in the range of interest for our system.First of all, we measured the error rate using the same set of SCHMMs than in the baseline experiment and applied to the NOISY test set. In **Figure 2** and **Table 1** (in which we summarize the results for both subsets in the test database and the overall performance), we can see that for the range of interest (between 1% and 9%), error rate increases, in average, in 39% when comparing with the CLEAN test set.

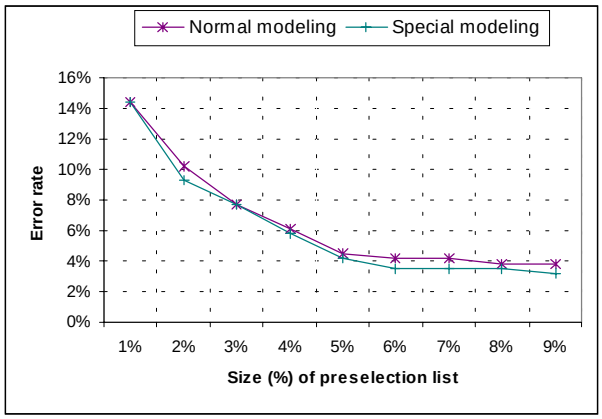

Figure 3: NOISY test set. With and without specific non-speech sound modeling.

Then, we used the hand labeled non-speech sounds in the NOISY training set, to obtain a single "non-speech sound" SCHMM, with the same topology than the others. In order not to increase computational load, we decided to keep the same number of models considered in the PSBU module, so we used only one silence model. Summarizing, we kept the 23 phoneme-like unit SCHMMs, one model for silence, and one for non-speech sounds (referred as "*special modeling*" from now on); and run the same experiment. As in the baseline system, both silence and non-speech sound models are discarded prior to forward the phonetic string to the LA module. We tested how this reduction in number of silence models affected performance in the CLEAN test set, and we found a slight increase in average error rate of 0.2%, showing the method does not significantly degrade the results obtained in the baseline experiment.

In **Figure 3**, we compare the recognition performance on the NOISY test set, using specific non-speech sound modeling (normal modeling) and not using it (special modeling). Error rate reduction, on the average, is almost 7% and shows a consistent improvement behavior in the range of interest.

An additional point of comparison was made measuring the difference in performance between test sets with and without non-speech sounds. When using the specific model to handle non-speech sounds, error rate increase is almost 30%, compared with the previous 39% figure. Obviously, we still have worse

performance, but the differences are lower. As a final result, **Figure 4** shows the overall error rate reduction using the whole test database. Average error rate actually decreases compared with the baseline experiment (in which the CLEAN test set was used) in around 1.4%, a slight overall improvement, due to the lower proportion of utterances with non-speech sounds in the database, and the slight degradation that the method produces on the CLEAN test database.

TEST SET	Normal modeling	Special modeling
CLEAN	4.7%	4.71%
NOISY	6.54%	6.11%
CLEAN+NOISY	5.03%	4.96%

Table 1: Average error rates for the test database sets.

Similar comparative rates have been obtained in experiments using 1200, 2000 and 5000 words dictionaries, even in other tasks, showing that the behavior of the approach used is consistent and that it is important of introducing additional mechanisms to face non-speech sounds in speech signals.

6. DYNAMIC PRESELECTION LISTS

In the baseline system described above, the preselection module generated a list of candidate words of fixed length, to be given to the verification stage. Our idea is making this length variable, depending on any available system parameter (number of frames, phonetic string length, PSBU probability estimation with different normalizations, lexical access cost, etc.). We were thinking in, for example, determining whether the word length (number of frames) was somehow related to recognition confidence, taking into account that, usually, longer words were possibly better recognized and viceversa. So, for longer words, a shorter preselection list could be used, and computational demands could also decrease.

The key factor to evaluate different methods in our case is calculating the *average effort*, defined as the average preselection list length required to ensure that the error rate in this stage is under 2%. If this average effort is lower than the fixed preselection list length, the method is acceptable, and it would be better as this average effort decreases.

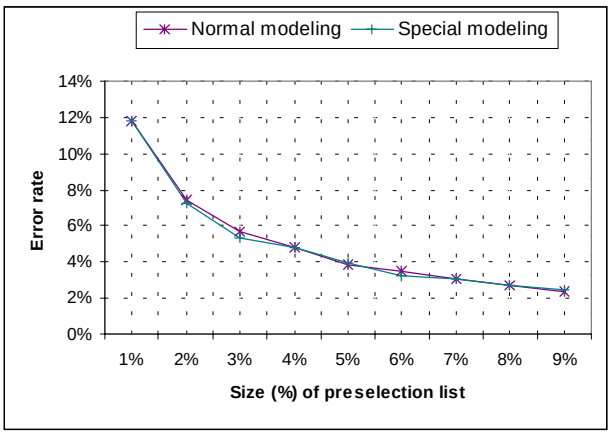

Figure 4: Whole test database results. With and without non-speech sound modeling

We applied parametric and non-parametric approaches, using data-driven techniques to estimate the relationship between the available parameters and preselection list length.

In the non-parametric case, our work was oriented towards building a pruning histogram, relating any given parameter (or parameter range) to a certain preselection list length. We start with a fixed preselection list length for every parameter value (or range): the one needed to ensure 100% inclusion rate. Then, the idea is iteratively search for the utterances in which the system performed worst (that it, the ones for which the preselection list must be longer), and discard them if possible (basically if this discarding does not affect more than one word) That means modifying the pruning histogram, using a lower value for the preselection list length for this utterance parameter value (or range). At the end of every iteration, the updated histogram is used to test the inclusion rate achieved and if conditions are met (error rate is below 2%), the process starts over again.

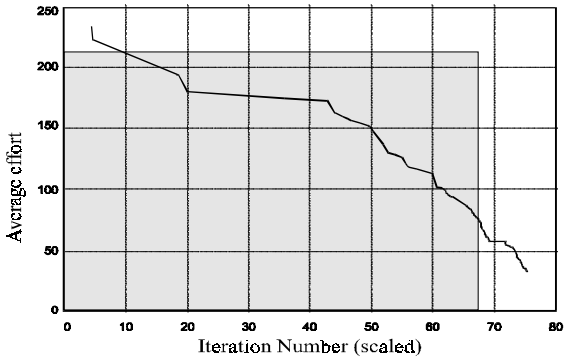

Figure 5: Average effort evolution using the number of frames as the control parameter in the non-parametric approach

In **Figure 5**, we show the *average effort* evolution as a function of iteration number, in an experiment in which the fixed list length to ensure 2% error rate was 216 candidates, using the number of frames as the control variable. The grey area shows the region for which the average effort was below 216 and the error rate measured was under 2%. In the optimal case, it would mean that we could get an average effort of around 75, that is, a reduction of up to 65%. This approach seems to be promising but has two main drawbacks. First of all, the pruning histograms are heavily dependent on the training data, so that they should be further smoothed to content with unknown data, leading to an average effort that will always be above the optimal value mentioned (75). The second refers to the granularity used to discretize the continuous range of certain parameters (for example, the log-likelihood computed in the PSBU module): as we make the intervals smaller, we obtain histograms too close to the training data. If we choose bigger intervals, more files will be affected by the pruning threshold imposed by the worst of them, so that the reduction in average effort is lower.

In the parametric case, we impose a fixed analytic relationship between the parameter and the list length. Our first attempt in this direction has been using a linear function of the control parameter. For example, when using the number of frames as the control variable, the longer the word, the smaller the list, and viceversa. In **Figure 6**, we show for every file, the pairs (number of frames, position in which the file was recognized). So, we only

need to estimate the line equation leading to the required results. In our example:

$$\text{preselection list length} = -\frac{C0}{P0} \cdot NumFrames + C0$$

where P0 and C0 are values to calculate (graphically, they are the intersecting values of the line equation in the two axis of **Figure 6**). In the upper dark area of **Figure 7** we show the values of pairs (P0, C0), for which we obtain average effort below 216 candidates, and error rate below 2%. In the lower dark area of the same Figure, we have plotted the actual average effort. From these two sets of graphical data, we can see (points marked in **Figure 7**) that for, approximately, P0=125 we get a minimum value of average effort, which is roughly 160, what means a reduction of almost 26%. This would correspond to a value of C0=410, given the valid area shape in the upper graphic, the value of P0 and the estimation formula used, searching for C0's minimum value. This optimum line equation (P0=125, C0=410) is shown in **Figure 6**.

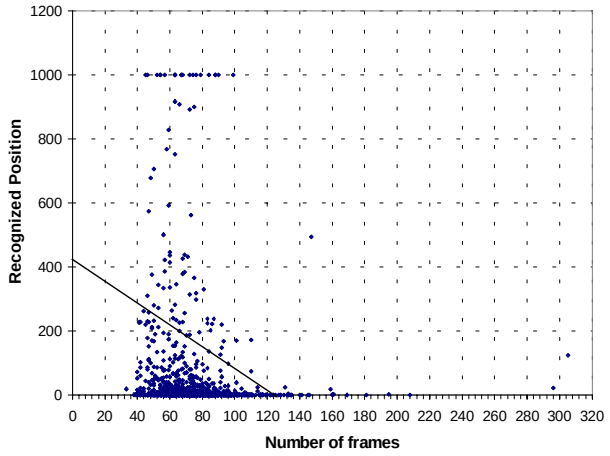

Figure 6: Plot of pairs (number of frames, position in which the file was recognized). Line shows the optimum in the parametric case. Recognized position has been clipped to a value of 1000

The main drawback in this case is the simplicity of the preselection list length estimation function. We are confident in this being a more promising approach, specially when combined with the non parametric one, but we need to derive more complete functions, including more parameters and better parameter estimation methods. Unfortunately, none of the approaches applied has given a definitive solution yet, and lot of work is to be done to achieve an implementable strategy.

7. CONCLUSIONS AND FUTURE WORK

Presence of non-speech sounds in telephone speech is fairly common (around 20% of the cases) and important enough to be taken into account when designing and testing real-world systems. We have applied a simple strategy to face this problem, achieving reasonable results. The overall performance using the whole test set is slightly better than using the CLEAN subset (1.4% decrease in error rate), and the results have been considerably improved when comparing the degradation between the two test sets. In any case, we are more concerned with the idea of being able to correctly recognize some of those speakers systematically presenting non-speech sounds in their utterances, even if their proportion is low. We are working on more complex non-speech sound modeling to improve performance.

We have also presented some preliminary ideas towards achieving dynamic preselection lists, using both parametric and non parametric techniques. None of them have been intensively tested nor have been refined to allow a final implementation, but we consider this is a good starting point in the topic. In this area, we are currently working in combining the parametric and non parametric approaches, applying regression analysis methods to make the parameter estimation process and studying the application of techniques based in neural networks as a novel approach to preselection list length estimation.

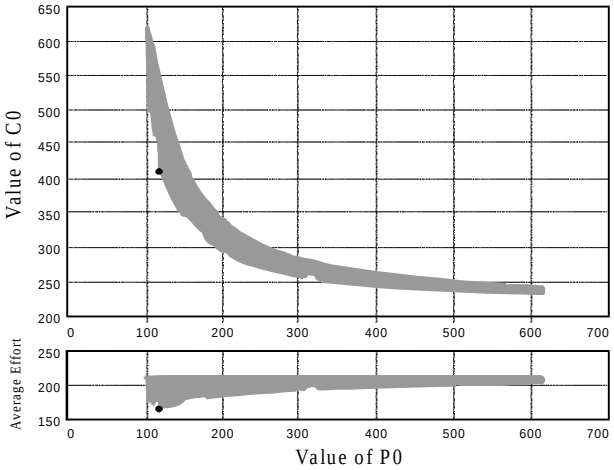

Figure 7: Valid pairs of P0, C0 and average effort using a linear function of the number of frames in the parametric approach

8. 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. Fissore, L., Laface, P., Micca, G. and Pieraccini, R. *"Lexical Access to Large Vocabularies for Speech Recognition"*. IEEE Trans. ASSP Vol. 37, n. 8. 1197-1213. 1989

4. 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.

5. 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

6. Macías-Guarasa, J., Gallardo, A., Ferreiros, J., Pardo, J.M. and Villarrubia, L. "Initial Evaluation of a Preselection Module for a Flexible Large Vocabulary Speech Recognition System in Telephone Environment". ICSLP'96, 1343-1346.