ISCA Archive
http://www.isca-speech.org/archive

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

INCORPORATING MULTIPLE-HMM ACOUSTIC MODELING IN A MODULAR LARGE VOCABULARY SPEECH RECOGNITION SYSTEM IN TELEPHONE ENVIRONMENT

A. Gallardo-Antolín, J. Ferreiros, J. Macías-Guarasa, R. de Córdoba and J. M. Pardo*

Grupo de Tecnología del Habla. Departamento Ingeniería Electrónica. Universidad Politécnica de Madrid.
E.T.S.I. de Telecomunicación, Ciudad Universitaria s/n, 28040, Madrid, Spain
e-mail: gallardo@tsc.uc3m.es, (jfl, macias, cordoba, pardo)@die.upm.es

ABSTRACT

The use of multiple acoustic models has reported great improvements when facing speaker independent difficult tasks. In this paper, we are applying this strategy to a flexible, large vocabulary, speaker-independent, isolated-word hypothesis generation system in a telephone environment with vocabularies up to 10000 words. The new problem addressed here is how to efficiently integrate the multiple model scheme in the system, as due to its bottom-up approach (phonetic string generation followed by a lexical access process), multiple possibilities arise (apart from the alternatives in the training stage), and its not clear what combination would achieve the best results. In the paper, full details on every alternative are shown, along with results showing actual improvements in the system.

1. INTRODUCTION

When facing the design and implementation of real-word public information services using the telephone network and working in real time, important aspects arise, as opposed to the conditions found in laboratory environments. At ICSLP'96 [1] and ICSLP'98 [2] we presented a large-vocabulary, speaker-independent, isolated word preselection system in a telephone environment with different improvements in order to take into account low computational demands and reasonable recognition rates. Techniques for handling non-speech sounds or for using variable-length preselection lists have shown its usefulness in this context.

Other common sources of recognition errors in speaker-independent real-world applications are pronunciation variations between different speakers (even between different utterances pronounced by the same speaker). The use of multiple acoustic modes per phonetic unit can help to reduce these variations and hence to increase the overall system performance.

Specifically, the approach of using gender specific model sets has been widely used and great improvements have been reported. In this paper, we apply this strategy to the preselection module of the ASR system mentioned above. The novelty here resides in the fact that due to the highly modular architecture of the system, incorporating multiple acoustic models can be done in different ways for each sub-module. We will describe these different approaches in both training and recognition stages, showing remarkable improvements.

2. SYSTEM OVERVIEW

The general system architecture is based on the hypothesis-verification paradigm, so that the output of a rough analysis module, with low computational demands, is fed to a detailed matching module. In this paper, we focus on the hypothesis generation stage, which has been designed according to a bottom-up approach [1] and consists of three main modules, as shown in Figure 1:

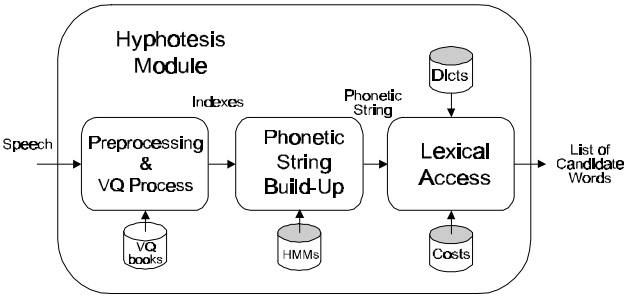

Figure 1: Hypothesis Module Architecture.

Acoustic Processing (AP): The input speech signal is preprocessed obtaining a vector of parameters composed of 8 MFCCs, 8 delta-MFCCs, log-energy and its first derivate. They are quantized for discrete HMMs (DHMMs) or soft quantized if semi-continuous (SCHMMs) are used, with 2 codebooks and 256 centroids each.

Phonetic String Build-Up (PSBU): the resulting indexes are passed to the phonetic string build-up module with generates a string of alphabet units using a frame-synchronous one-pass algorithm. Each set of gender-dependent models (or single models) consists of 23 allophone-like context-independent HMMs. In order to reduce the effect of endpointing inaccuracies 2 additional silence units are also introduced.

Lexical Access (LA): The phonetic string is matched against the whole dictionary, using a dynamic programming algorithm

* Ascensión Gallardo-Antolín is currently at Departamento de Tecnologías de las Comunicaciones, Universidad Carlos III de Madrid, E.P.S., C/ Butarque, 15, 28911-Leganés (Madrid), Spain.

and alignment costs for substitution, insertion and deletion errors [4].

3. DATABASES AND DICTIONARIES

In our experiments, we have used part of the VESTEL database [5], a speaker-independent speech corpus collected over commercial telephone lines, composed of digits, numbers, commands, city names, etc.

The training set consists of 5820 utterances pronounced by 3011 different speakers (46.74 % of training data belongs to male speakers and 53.26 % to female ones). The test set is composed by 1434 utterances from 1351 speakers. In both sets, no noticeable non-speech sounds are presented.

Experiments have been performed using three dictionaries, which have been built with 2000, 5000 and 10000 words, extracted from the application domain, if available, or added from the ONOMASTICA project results. In all of them, graphemes had never been seen in the training data (vocabulary-independent task).

4. TRAINING MULTIPLE-HMM ACOUSTIC MODELING

For training multiple-HMM acoustic models, a partitioning of the training data is required; each part groups utterances with similar acoustic characteristics. Here, this division has been performed based on gender-dependency due to the clear acoustic differences between utterances pronounced by male and female speakers.

We have developed two different methods for training gender-dependent models (in both cases, the models were trained using the Viterbi algorithm):

- **Independent training**: In this case, we have trained each set of models, using only the part of the database assigned to them. The main disadvantage of this method is the a priori assignment of the training data to a particular set. This situation is not very suitable if the acoustic characteristics of the utterance fit better into the other set. In addition, if the training data is not balanced enough between both sets, a poor modeling may result for the one containing less examples.
- **Joint training**: This approach tries to overcome the problems mentioned before. In this case, all the material for both sets of models is used, and a weighting function controls the influence of each utterance in the modeling of each set [3]. The weights for both sets are calculated as follows:

$$w_B = \left(\frac{P_B}{P_A + P_B} \right)^{\alpha} \qquad w_A = 1 - w_B$$

where, P_A and P_B are the likelihoods for the utterance with the set of models A and B, respectively, α is an adjustment factor, and ω_A and ω_A are the weights to be applied to the reestimation formulae in the training stage.

The adjustment factor α allows to assign more training data to a particular set. For example, for α = 1.25, the training of models belonging to set B will be emphasized.

5. INCORPORATING MULTIPLE-HMMs IN THE RECOGNITION STAGE

5.1. Multiple-HMMs in the PSBU module

In the PSBU module, two approaches are tested:

- **"Combined-sets"**: In this first approach, phonetic strings are composed by concatenating models coming from any set. No modification is required on the one-pass algorithm in the PSBU implementation; the only difference is that the number of acoustic models is duplicated.
- **"Single-set"**: In this case, strings are forced to be generated by only one set of models (the one that produces the best score). In the experiments performed, we observed that only 11% of the words uttered by male speakers produce better scores when comparing to the opposite set models (female speakers). In the same way, 5.5% of the test data belonging to female's group fitted better to male models. This fact corroborates that gender-dependent models make an adequate discrimination between both sets. In fact, in many cases, phonetic strings generated by using the opposite models were composed of a non-sense concatenation of allophones.

Regarding to the increase of computational load, both methods duplicate the number of numeric operations when compared to the single-modeling approach.

5.2. Multiple-HMMs in the LA module

In the LA module, two strategies are also used:

- **Shared-costs**: In this approach, all the allophones have the same behaviour in the lexical access stage even if they have been generated from different sets of models, so that both sets share the same confusion matrix. This is an optimal approach when the data for training substitution, deletion and insertion costs is limited.
- **Set-dependent-costs**: In this second strategy, costs are gender-dependent as in the acoustic modeling, so it is necessary to train one squared confusion matrix (for "single-set" PSBU strategy) per set or a single rectangular confusion matrix (for "combined-sets").

The implementation of these techniques is not very expensive due to the low computational load of the LA sub-module itself. In summary, by combining these different strategies for PSBU and LA modules, we have four different alternatives for introducing Multiple-HMM modeling in the recognition stage, as is shown in Figure 2. Note that the combination of "single-set" PSBU with a "set-dependent-costs" LA strategy has not been implemented in this work, as the use of rectangular confusion matrices complicates remarkably the LA module, and, no significant improvement was expected taking into account preliminary experimentation.

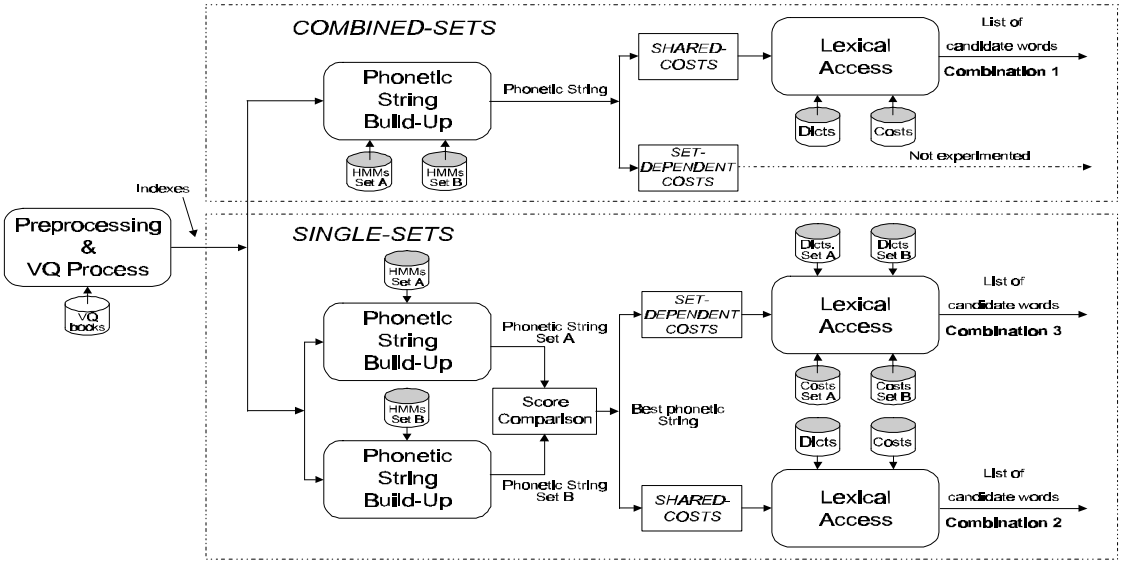

Figure 2: Strategies for incorporating Multiple-HMMs in the recognition stage.

6. EXPERIMENTAL RESULTS

The performance of the preselection module has been measured in terms of the inclusion rate, i.e. the percentage of words calculated over whole dictionary that are necessary to be included in a preselection list in order to achieve a certain recognition rate. For example, for a 10000 words task, a 10 % in the figures would mean we used a preselection list of 1000 words. In the case of the 2000 words dictionary, we estimated a preselection list 4% of dictionary size would be reasonable. For the 5000 and 10000 words dictionaries, a preselection list 10% of dictionary size is a good choice for achieving an adequate overall performance.

6.1. Independent-training vs. joint-training

The aim of this set of experiments was to determine the best procedure for training multiple acoustic models. In this case, we trained discrete models (DHMM). Combination 2 (see Figure 2) was the strategy used in the recognition stage.

Figure 3 shows the results obtained with single-DHMM and multiple-DHMM trained according to the two possibilities mentioned in Section 4: independent-training and joint-training with three different values for the adjustment factor.

The introduction of multiple-DHMMs improved considerably the system performance for almost all the alternatives. Regarding to the comparison between independent and joint training, the last technique does not increase significantly the recognition rate (even, recognition rate decreases when using α = 1.25).

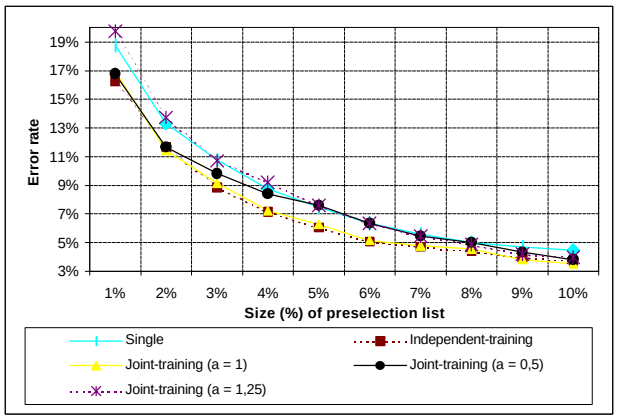

Figure 3: Comparison of error rates for different training strategies: single-DHMM, independent-training multiple-DHMM and joint-training multiple-DHMM (2000 words dict).

There are two possible explanations for this phenomenon: on one hand, the automatic classification performed by the joint-training procedure is very similar to the one obtained by simply dividing the training data into male and female speakers. In fact, only 2 % of male utterances fits better to female models, and conversely, only 4 % of female utterances is better recognized using the opposite modeling. On the other hand, the training data is reasonably well-balanced, so for values of α different from 1.0, no improvement is achieved. Therefore, we adopt the independent strategy for training multiple-HMMs in the remaining of experiments.

6.2. Alternatives for PSBU and LA modules

Figure 4 shows the performance of the preselection module when the different combinations of PSBU and LA strategies listed in Figure 2 are applied. From the results, we can extract two conclusions:

- The use of combined sets in PSBU does not produce better results due to the increase of insertion errors (see experiment labeled as “Combination 1” vs. “Combination 2”).
- Using one confusion matrix (“shared costs”) performs slightly better than using two confusion matrices. The reason is the lack of data when training two matrices of costs (see experiment labeled as “Combination 2” vs. “Combination 3”).

However, all these alternatives outperform the single-HMMs system. In summary, the “single-set” technique in PSBU module and “shared-costs” technique in LA module is the combination that achieves better results.

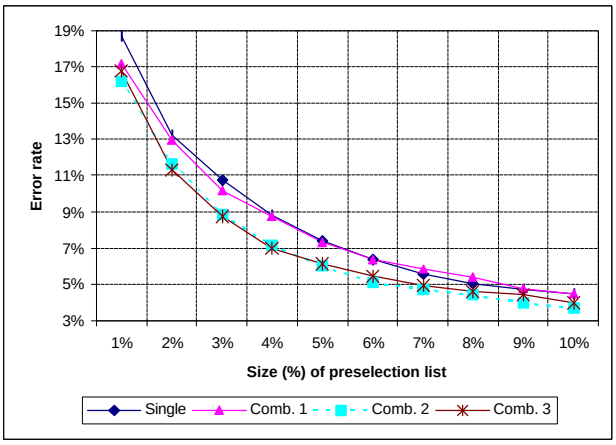

Figure 4: Comparison of error rates for different alternatives of introducing multiple-DHMM in the recognition stage of the preselection module (2000 words dictionary).

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

Although the previous experiments have been carried out using DHMM modeling, a better acoustic modeling is needed in order to achieve reasonable results in a real-world system. So, we decided to use multiple SCHMM modeling. We experimented with the choice that achieves the least error-rate for multiple-DHMM, i.e. independent-training, “single-set” in PSBU and “shared-costs” in LA module.

The comparison between single-SCHMM and multiple-SCHMM is shown in Figure 5. We carried out three set of experiments using 2000, 5000 and 1000 words dictionaries. As it can be observed, multiple-SCHMM performs significantly better than single-SCHMM. In fact, a relative error reduction around 25% is achieved with a dictionary of 10000 words.

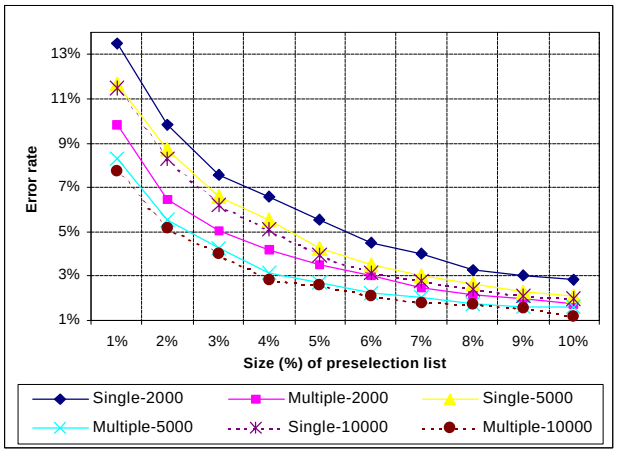

Figure 5: Comparison of error rates for different dictionary sizes. Single-SCHMM vs. Multiple-SCHMM (independent-training + combination 2)

7. CONCLUSIONS

The use of multiple acoustic models per phonetic unit allows increasing acoustic modeling robustness in difficult tasks.

Different combinations of the techniques explained above have been tested and the best results were obtained using independent-training, "single-set" technique in PSBU module, and only one confusion matrix in LA. Applying this strategy to multiple-SCHMM modeling and a dictionary of 10000 words, we have achieved a relative error reduction around 25% (compared to the single-SCHMM system).

8. REFERENCES

1. Macias-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. Vol .2, pp. 1343-1346. 1996.
2. Ferreiros, J., Macías-Guarasa, J., Gallardo-Antolín, A., Cordoba, R., Pardo, J. M. and Villarrubia, L. "Recent Work on a Preselection Module for a Flexible Large Vocabulary Speech Recognition System in Telephone Environment". ICSLP 98. Vol. 2, pp. 321-324. 1998.
3. Córdoba, R. “Sistemas de reconocimiento de habla continua y aislada: comparación y optimización de los sistemas de modelado y parametrización”. Phd. Thesis. ETSI Telecomunicacion. Universidad Politécnica de Madrid. 1995
4. Fissore, L., Laface, P., Micca, G. and Pieraccini, R. “Lexical Access to Large Vocabularies for Speech Recognition”. IEEE Trans. ASSP. Vol. 17, n. 8, pp. 1197-1213. 1989.
5. Tapias, D., Acero, A., Esteve, J., and Torrecilla, J.C. "The VESTEL Telephone Speech Database". ICSLP 94, pp. 1811-1814. 1994.