TRAINING A ROBUST COMMAND RECOGNIZER WITH THE TECNOVOZ DATABASE

José Lopes[1,3]*, Cláudio Neves*[1]*, Arlindo Veiga*[1]*, Alexandre Maciel*[1]*,*
Carla Lopes[1]*, Luís Sá*[1,2]*, Fernando Perdigão*[1,2]

[1] Instituto de Telecomunicações – Pólo de Coimbra, 3030-290 Coimbra, Portugal
[2] Dep. Eng. Electrotécnica e de Computadores, FCTUC, 3030-290 Coimbra, Portugal
[3] L2F – Spoken Language Systems Lab INESC-ID, Rua Alves Redol, 9, 1000-029 Lisboa, Portugal

ABSTRACT

This paper describes the development of a command-based robust speech recognition system for the Portuguese language. Due to an efficient noise reduction algorithm the system can be operated in adverse noise environments such as in vehicles or factories. The acquisition of a Portuguese database in the scope on the Tecnovoz project is addressed in this paper. The paper also describes a new noise-robust front-end and some experiments regarding the best acoustic model to use for a command-based speech recognizer. Results with whole-word, monophone and triphone models are presented and discussed.

1. INTRODUCTION

Tecnovoz [1] is a shared-cost project funded by the Portuguese government which aims to create a body of knowledge on voice technologies, particularly to the Portuguese language. This will materialize in a series of products for the market. The authors were responsible in the framework of the project for the development of a speech independent connected word recognizer which operates under noise adverse conditions, such as factories and vehicles. Therefore it has to incorporate advanced noise reduction techniques. Finally, the recognizer has to be computationally efficient in order to operate on small footprint embedded hardware platforms.

The speech database was collected in the scope of the Tecnovoz project. It has been designed regarding typical application demands, in terms of vocabulary and acoustic environments. The acoustic models are based on Hidden Markov Models (HMMs).

In order to deal with noise adverse conditions, a noise reduction front-end was designed, based on the Advanced Front-End (AFE) ETSI Standard [2]. Some modifications were made from the standard to enhance the performance and speed of the speech recognizer.

In order to improve the robustness of the speech recognizer, several experiments with different acoustic models were carried out using either whole-word HMM models or smaller unit HMM models, such as monophones and triphones.

The paper is organized as follows. In section 2 the database is described. Section 3 describes the front-end implementation. Section 4 refers to the approaches to the acoustic modelling. Finally, in section 5, results obtained with different acoustic models and front-end configurations, are presented.

2. SPEECH DATABASE

Three acoustical environments were considered during the database acquisition, namely: clean (TVFL), vehicle (TVV) and factory (TVF) environments. The collected speech database includes about 250 commands and several phonetically rich sentences. About 30 minutes of spoken content were recorded by each of the 368 speakers, which turn into about 184 hours of speech content and a total of 232,000 files. Table 1 and Table 2 show the distribution of the database according to gender and file types, respectively.

Gender	**TVFL**	**TVF**	**TVV**
Female	103	20	9
Male	197	16	23

Table 1: *Gender distribution.*

Content	**TVFL**	**TVF**	**TVV**
Words	141,992	30,090	19,648
Sentences	40,458	–	384

Table 2*: Speech file distribution.*

3. FEATURE EXTRACTION

The feature extraction system is based on the AFE standard, which incorporates a two-stage Wiener filtering system. In this standard, the Wiener filter is estimated in the linear frequency domain and is implemented by a time domain convolution. Li et al, [3], proposed a new algorithm where both filter estimation and operation are carried out in the Mel frequency domain. In our implementation some changes were made to Li et al approach in order to improve the front-end efficiency, as depicted in Figure 1 [4].

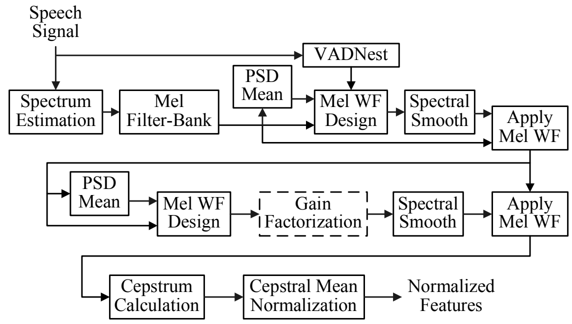

Figure 1: *Block diagram of the feature extraction system.*

It can be seen that the speech signal is processed by a two-stage Wiener filter as in the ETSI standard. The estimated signal spectrum is applied to a Mel filter-bank and the frames are then classified as "noise only" or "speech with noise" by the VADNest block. The Wiener filter design depends on this classification in order to estimate the noise spectrum. The "Spectral Smooth" block presents some modifications: the operations involved in the smoothing of the Wiener filter coefficients were reduced to a single matrix multiplication [4]. Apart from the gain factorization block, that were not found valuable for the final system performance, the second Wiener filter stage is similar to the one proposed in [3].

The de-noised frames are then converted to cepstral coefficients by a discrete cosine transform (DCT) and their means are normalized by a real-time algorithm, resulting on a feature vector with 39 components, comprising 12 cepstral coefficients plus log energy and their first and second time derivatives. The feature extraction algorithm is described in detail in [4].

4. MODEL TRAINING

Acoustic models were built using the Tecnovoz speech database and only files corresponding to command utterances with Signal-to-Noise Ratio above 15 dB were considered. There were a total of 137,860 files (120,459 from TVFL, 8,760 from TVF and 8,641 from TVV). From these files 75 % were picked for training, 20 % for test and 5 % for development. From the first trained models a recognition test was performed on the training database. The results allow us to detect transcription errors, and consequently, some annotation files had their marks re-adjusted, others were deleted and wrong labels were changed according to the word effectively pronounced. From these procedures the total number of files was reduced to 137,237 (119.975 from TVFL, 8,633 from TVF and 8,629 from TVV).

The model training was carried out using the HTK toolkit [5]. During the training three approaches were explored for the acoustic models: word-level, context-free phones and context-dependent triphones. The word-level approach tries to create HMM models for the whole-word, whereas context-free monophone models split the words into the corresponding monophone transcription to provide data for monophone training. Finally, triphone training tries to profit from left and right contexts of each phone, which naturally influence the acoustic realization of each phone, to create a new model. The advantages and disadvantages of each method will be discussed in the next sub-sections.

4.1. Word-level training

For word-level training, each of the 254 words is represented by an HMM with left-to-right topology. The number of states of the HMM depends on the word length in terms of phones. For example for the command "stop", the transcription is /s t O p/ (in SAMPA), which results in a 12-state HMM for this word, using 3 states for each phone.

The models "ruido" and "sil" are used to model noise and silence, respectively. They are represented with 3-state HMM's with left-to-right topology with an extra transition from the first to the last emitting states and vice-versa.

The model initialization was done with the HTK tools HInit and HRest. Afterwards, the training was carried out with the embedded re-estimation HTK tool HERest. Word-level models were trained with mixture increment, up to 10 Gaussian mixtures for each state.

4.2. Monophone training

The first step consisted in defining the phone set for the Portuguese language. A list of 40 phones was taken, including models for silence and pause. All phone HMMs have 3 states with a left-to-right topology and were initialized with the "flat start" method [5]. Multiple pronunciations were considered for some words, which permitted to realign the training data after 5 iterations of embedded re-estimation. The number of mixture components was then incremented up to 16 Gaussians, as depicted in Figure 2.

4.3. Triphone training

Triphones depends on the two adjacent phones, which gives considerable robustness to variations in pronunciations in specific contexts [6].

Since there is no annotated speech data at phone level, monophone models were used (initialized with the flat start procedure) to develop the intra-word triphone models. As triphones are phones with context, it was used a straightforward procedure to convert from one notation to another (e.g.: "dez" ("ten") → /d E S/ → /d+E d-E+S E-S/).

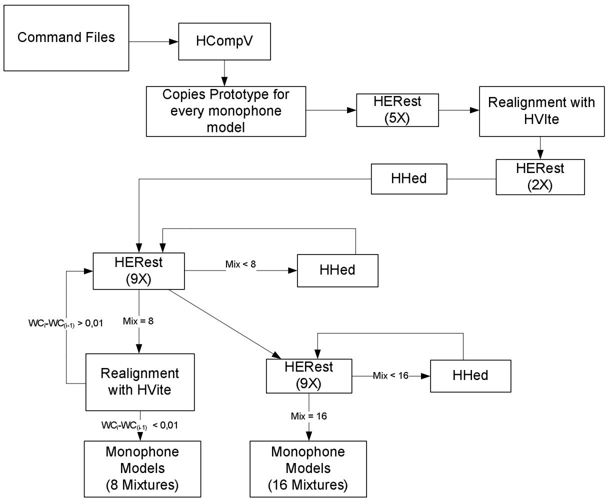
Command Files
HCompV
Copies Prototype for every monophone model
HERest (5X)
Realignment with HVIte
HERest (2X)
HHed
HERest (9X)
Mix < 8
HHed
WC_i-$WC_{(i-1)}$ > 0,01
Mix = 8
Realignment with HVite
HERest (9X)
Mix < 16
HHed
WC_i-$WC_{(i-1)}$ < 0,01
Mix = 16
Monophone Models (8 Mixtures)
Monophone Models (16 Mixtures)

Figure 2: *Monophone training procedure.*

The resulting number of triphone models is 872 for the command vocabulary. This results in much less training material for each triphone compared with monophones. To overcome this problem and taking into consideration that there are many similar triphones in the model list, some models can be tied in order to reduce the total number of physical models. For this purpose two methods were considered: data-driven clustering (DDC) and tree-based clustering (TBC). The data-driven clustering uses a similarity measure between HMM states, while tree-based clustering builds a binary decision tree. This tree attempts to find those contexts which make the largest difference to the acoustics and which should therefore distinguish clusters. The latter method has the advantage of accommodating the construction of systems which have used unseen triphones [5]. Different likelihood thresholds in TBC and distance thresholds in DDC were taken into account as sources of variability in training. These thresholds have a strong influence on the number of physical models that need to be trained, and consequently in the total number of Gaussians, which is a major concern as recognizer will be working over low performance hardware. The number of Gaussians was incremented up to 16, as depicted in Figure 3.

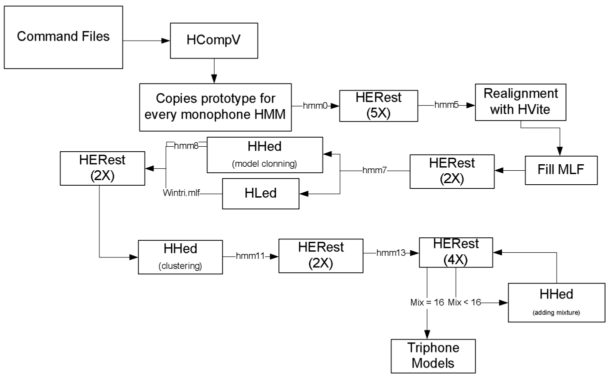

Figure 3: *Triphone training procedure.*

5. RESULTS

In this section, the results obtained with each acoustic modeling approach are presented. Tests were carried out using the HTK decoder tool HVite.

To perform the experiments, a task grammar must be defined in order to provide information about the sequence of events that can be found in the test utterances. The used grammar consisted in taking all the command words in parallel, with an optional silence before and after a command, as shown in Figure 4.

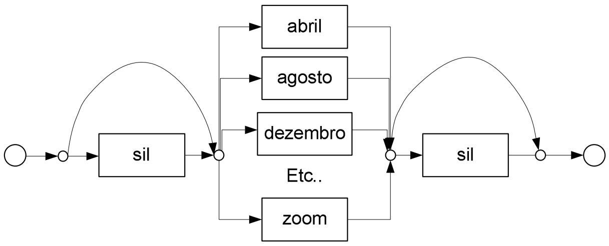

Figure 4: *Task grammar.*

Table 3 shows the achieved recognition rates of both the original and proposed front-ends in terms of the *Word Correctness* rate. The improvements made at the front-end level resulted in a system' performance improvement of about 2% (absolute points). This result suggests that the ETSI's AFE may be biased towards the database used for the evaluation of the algorithm (the Aurora 2 database).

Front-End	Word Correctness
Original ETSI AFE	94.88 %
Efficient Front-end (E-AFE)	96.88 %

Table 3: *Comparison between AFE and E-AFE.*

Three different versions of the whole-word models were tested. The first one corresponds to the models created with the first alignment of the training database (V1). The models' second version resulted from the new label files with re-adjusted marks (V2). The third one was created with a modification in the front-end, that consists in removing the gain factorization block, indicated in Figure 1 (V3). Results for 8 Gaussian mixtures are presented in Table 4. As expected, the consecutive modifications made on training procedure, label files and front-end improved the whole-word models.

Version	Word Correctness
V1	95.92 %
V2	96.55 %
V3	96.76 %

Table 4: *Whole-word models results.*

As referred to in section 4.2, the label files were automatically aligned several times, in order to improve robustness. Table 5 shows the word correctness for monophone models, with 8 Gaussian mixtures, for several realignment iterations. Besides the low rates obtained with the monophone models, an improvement of 6 % was observed by realigning the training data 3 times.

Number of Realignments	Word Correctness
0	83.46 %
1	87.57 %
2	87.54 %
3	89.28 %

Table 5: *Monophone models results.*

To evaluate triphone model performance, experiments were carried out with no clustering and with both clustering methods. Results obtained with 8 Gaussian mixture models are presented in Table 6.

Clustering Method	Threshold	Word Correctness
TBC	7500.0	96.06 %
TBC	1000.0	97.03 %
TBC	300.0	97.06 %
DDC	0.3	96.81 %
No clustering	–	97.03 %

Table 6: *Triphone models results.*

Results indicate that the lower the thresholds in TBC, the better are the results. This is due to the number of physical models resulting from the cluster which is higher when the likelihood threshold is lower. With about the same number of physical models, the DDC clustered models presents a slightly lower score. Nevertheless, the clustering method seems to be useless, since with no clustering a very similar recognition rate is achieved.

In order to compare the performance of the three acoustic model types, the best score from each approach is presented in the same table as well as the number of Gaussians that the ensemble of models have. According to Table 7, the triphone models have the best performance, comparing to whole-word or monophone models. As the recognizer should work on low performant hardware, a trade-off between computational load (dependent on the number of Gaussians), and the recognition rate should be made. The triphone models not only have less computational load when compared to whole-word models, as achieve higher recognition rate. As a result, most commands in the vocabulary are represented by triphone models in our recognition engine. Only smaller commands, where whole-word models seem to be more accurate, use this kind of models.

Acoustic Model	Word Correctness	Total Number of Gaussians
Whole-word	96.76 %	37,344
Monophone	89.28 %	952
Triphone	97.03 %	16,204

Table 7: *Comparison of acoustic models.*

6. CONCLUSIONS AND DISCUSSION

In this paper some modifications are proposed to the ETSI's AFE regarding noise robustness of a command-based speech recognition system for the Portuguese language. The new proposal outperformed the ETSI standard in about 2%.

Three different acoustic models (whole-word, monophone and triphone models) were also tested and compared. Results show that triphone models achieved the best performance.

Another interesting conclusion is that new word models can be easily built using the monophone models. The user just need to add the sequence of phones that compose a new command in order to be accepted by the recognizer engine. With triphones it is not that simple, because only a small set of triphones are available. An algorithm that associates to an unseen triphone the better one that is already on the initial triphone list is currently being developed. This tying takes into account acoustic and phonetic similarities between triphones. With this association the recognizer will be prepared to recognize any command.

7. REFERENCES

[1] Tecnovoz website (2007), http://www.tecnovoz.pt/web/home_english.asp.

[2] ETSI ES 202 050 v1.1.3, "Speech Processing Transmission and Quality Aspects (STQ); Distributed Speech Recognition; Advanced Front-End Feature Extraction Algorithm; Compression Algorithms", Technical Report ETSI ES 202 050, November 2003.

[3] J.-Y. Li, B. Liu, R.-H. Wang, and L.-R. Dai, "A complexity reduction of ETSI standard advanced Front-End for DSR", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol. 1, pp. 61-64, Montreal, Canada, May 2004.

[4] C. Neves, A. Veiga, L. Sá, and F. Perdigão, "Efficient Noise-Robust Speech Recognition Front-End Based on the ETSI Standard", IEEE 9th International Conference on Signal Processing (ICSP), Beijing, China, October 2008.

[5] S. Young, G. Everman, et al, "The HTK Book (For Version 3.4)", University of Cambridge, England, 2006.

[6] S. Abate, and W. Menzel, "Automatic Speech Recognition for an Under Resourced Languaged – Amharic", Proc. Interspeech, pp. 1541-1544, Antwerp, Belgium, August 2007.