Digit Recognition Using the SPEECHDAT Corpus

Frederico Rodrigues and Isabel Trancoso

INESC, Instituto de Engenharia de Sistemas e Computadores - IST, Instituto Superior Técnico
INESC, Rua Alves Redol, Nº 9, 1000 Lisboa

Abstract

With the remarkable evolution of telecommunications as we reach the end of this century, it becomes clear that speech recognition via the telephone network will play an increasingly important role, mainly due to the widespread use of both cellular and non-cellular telephones. For many applications of speech recognition over the telephone, digit recognition is fundamental.

This paper describes a set of digit recognition experiments with the SPEECHDAT corpus for European Portuguese. We present techniques and results obtained with isolated and connected digits with both known and unknown length grammars. Error rates of 0.6% and 1,9% were achieved, respectively, for isolated digit and connected digit strings.

I. INTRODUCTION

During the past decade, research in automatic speech recognition has produced increasingly better results for a wide range of tasks. On the other hand, and at same time, the evolution of telecommunications and the widespread use of telephones are urging telephony service providers to apply speech technology [1].

For many applications of speech recognition over the telephone, such as credit card and account number validation, catalogue ordering, and many interactive voice response systems, digit recognition is fundamental. This kind of application demands a high level of accuracy in order to be of any use. However, speaker independent recognition of telephone speech is more difficult than clean speech recognition because besides speaker variability problems, we also have to deal with a potentially very large variability in channels and microphones, with many different kinds of channel and environmental noise.

This paper describes a set of digit recognition experiments with the SPEECHDAT corpus for European Portuguese. A baseline recognition system is described. We also present an extension to this baseline system that consists of an improved filler model used to absorb extraneous speech events that surround the uttered digits.

In the following section we describe the corpus used and in section III we present the model topology, training strategy and experimental results. Conclusions and future work are discussed in section IV.

II. TELEPHONE CORPUS

We have used a subset of the SPEECHDAT multilingual speech database [2], which has been collected in two phases (1000 and 4000 speakers for the first and second phases, respectively). The subset used consists of 3788 isolated digits (subset I1, without the feminine words for “um” (one) and “dois” (two), respectively, “uma” and “duas”) and 3641 connected digit strings (subset B1 only present in the second phase). All files were orthographically transcribed and extraneous speech events were marked, although distinctly in each phase. In the second phase, four broad noise categories were defined, namely, stationary noises (e.g. channel noise, voice babble, public place background noise, street noise), intermittent noises (e.g. music, background speech, baby crying, phone ringing, door slam), speaker noises (e.g. lip smack, cough, grunt, throat clear, tongue click) and filled pauses (e.g. uh, um, er, ah, mm). In the first phase these categories were subdivided. Hence, a more detailed annotation was achieved. Two special characters (~, *) in the beginning or end of words were used to indicate that the word was truncated or mispronounced. All the corresponding files were rejected.

A selection procedure defined by the project partners was used so that age, region and gender distribution is approximately equal in both train and test sets. A test set with 500 speakers was obtained and fixed for all items of the second phase corpus, to allow the comparison of results between different research teams. We extended this procedure to select 200 extra test speakers from the first phase. Finally, to reach an overall 20% test ratio we selected 300 additional speakers from the second phase to be used strictly as unseen test data or for cross-validation purposes.

Table I
Corpus size for each set

	Train Set	Test Set	Develop. Set
I1	2954	768	-
B1	2905	491	277

[1] Email: {fspr, Isabel.Trancoso@inesc.pt}
This work is part of Frederico Rodrigues’s Master thesis,
“Reconhecimento de Dígitos Ligados e Números Naturais”,
sponsored by a FCT scholarship (GGP XXI / BM / 3782 / 96)

Table I shows the corpus size for each set. As can be seen, for instance, in the subset B1 only 491 files were used for testing because this subset only appeared in the second phase of the corpus and because 9 files were marked as truncated or mispronounced.

III. EXPERIMENTS AND RESULTS

A. Feature Extraction and Model topology

The front-end stage of the speech recognition system used MFCC's (Mel-Frequency Cepstral Coefficients) with 30 parameter vectors per frame, specifically, 14 cepstral coefficients, 14 delta-cepstral coefficients, energy and delta-energy. The speech signal was band-limited between 200 and 3800 Hz and we used a Hamming window of 25 ms each 10 ms. An effective, yet simple procedure, CMN (Cepstral Mean Normalisation), provided some channel and speaker normalisation.

Each digit was modelled with left-to-right continuous density HMM's (Hidden Markov Model) with no skips between the states. State distribution for each whole-word model was determined based on the average length of each of 15 hand labelled examples of each digit. It ranges from 3 to 8 states for digits, and exactly 3 to all fillers and silence model (Table II). These also have 3 states and skips from the first to last state, and vice versa. A graphic representation is given in figure 1.

Table II
Number of states for each model

# States	Models
3	"um", fillers, silence
6	"cinco", "zero", "nove", "quatro", "oito", "três"
7	"sete"
8	"dois", "seis"

In order to introduce gender dependent models, which are now a standard in speaker independent speech recognition, we duplicated each model and labelled it accordingly.

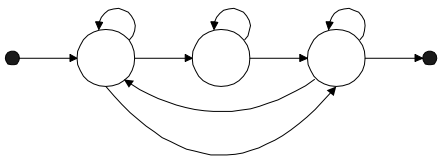

Fig. 1 Filler and silence model architecture

B. Baseline System

This section describes the baseline system where we had, along with the ten digits, 9 filler models plus the silence model. Each filler model represented an extraneous speech event close to the classification defined in the first phase of the SPEECHDAT corpus.

The HMM parameters of the isolated digit models were initialised with the global mean and variance of the training set for that particular digit. As for the filler and silence models, because they could appear in any file, the whole training set was used for this operation.

We started by choosing from the training set all the isolated digit utterances with no noise marks (about 1400 utterances) and did a series of embedded Baum-Welch re-estimations. At each 2 or 3 steps of this iterative process the performance was evaluated against the training set, with Viterbi decoding over a known length grammar where only one digit could occur. When no significant improvements were observed we decided to increment the number of gaussian mixtures per state from 1 to 2, and from 2 to 3 in a subsequent iteration. Viterbi recognition was performed now with a known length grammar where the digit could be preceded or followed by any number of filler models. A 99.9 accuracy score was obtained on the training set for these files.

The next step was to introduce all files from the first phase, which also had noise marks and do another series of re-estimations. This provided initial training of our filler models. Mixtures were incremented up to 3 for each filler model. The remaining files were from phase two corpus and hence had only 4 kinds of noise marks. We had two ways of mapping these to the 9 models we had. One was at the dictionary level, providing multiple pronunciations for each class. The other was to perform recognition over the training files (preserving the digits, which are always correctly annotated) and use the output as the new transcription. We chose the second because the first (simple alignment) only substitutes each model for the pronunciation it finds more suitable. Therefore the recognition alternative could and did lead us to a more complete transcription. A new series o re-estimations was performed and filler models mixtures incremented from 3 to 4, reaching a 99.6 accuracy for the training set.

As for connected digit strings, we started training using our best isolated digit models as bootstrap models. We did a series of re-estimations and then incremented the number of mixture components, only for digit models, from 3 to 5. The best results are shown in table III.

C. Extension to the Baseline System

The extension to the baseline system consisted of a new way of modelling the filler models.

It is widely accepted [3], [4], that connected digit accuracy degrades significantly in noisy environments or when the string length is unknown. This is mainly due to the fact that there is almost no grammatical structure, i.e., each digit may appear in any order surrounded or not by extraneous speech events. Also, it is very hard to explicitly mark every one of these events. In the SPEECHDAT corpus annotation process, annotators were instructed to mark some events such as stationary or intermittent noises only the first time

they appeared or at the beginning of the transcriptions. What we did in the baseline system to overcome this problem was to replicate this mark (for stationary noises only) over the transcription between each digit. However, this was far from an optimum solution because in the connected digits task we can have very small or no pauses between the digits and thus lead us to improper training of the digit models, which are our main concern.

In this extension we proceeded like in the baseline system until we trained the 9 noise models plus the silence model. However, this time, these models did not have a backward transition.

We then used these models to build a unique filler model that consists of all these models in parallel. The next figure shows a block diagram of this filler model, where each box corresponds to a noise model as the one depicted in figure 1, but without the backward transition. The transition from the end to the starting position is only a graphic simplification since what really happens is that the last state of each sub-model has a transition to the initial state of each of the others, and of course to the exit state.

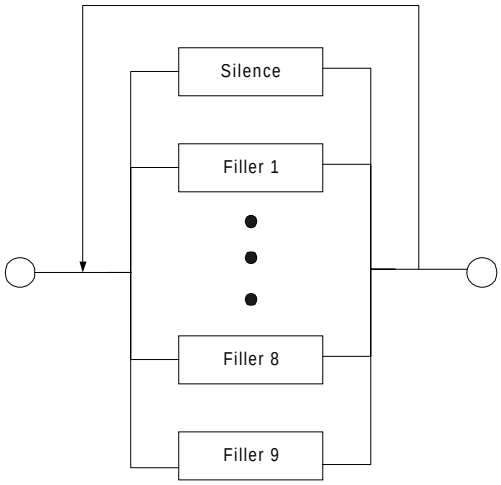

Fig. 2 New filler model architecture

The procedure to construct this filler model was as follows. We build a model with 32 states (3 states for each noise model and for the silence model) plus two non-emitting states. The first (non-emitting) state connected to the first state of every sub-model (states 2, 5, 8,…, 29) with the same transition probability (0.1). The first and intermediate states of each block remained exactly as they were. To determine how to distribute the transition probability of the last state of each block (states 4, 7, 10, …, 31), we performed a count on the training set to see how many times a noise mark appeared followed by another and how many times it appeared followed by a digit. Around 60 % of the times it was just before a digit, so the model exit transition probability was weighted with this value and we added 10 backward transitions, each with a tenths of the remaining 40 % former exit transition probability.

The filler was trained over the first phase utterances and then used in the alignment of the remaining isolated digit utterances. In this process, mixtures were incremented from 1 to 3, and from 3 to 5 only for the digit models.

The connected digit strings were trained from the best isolated digit models and in a first iteration substituting all noise marks by a filler mark in the transcriptions. Due to the described architecture all consecutive occurrences of filler marks were discarded since the model is supposed to be able to jump from block exiting only if and when a digit appears. However, this procedure has the problem described in the last section in respect to incomplete transcriptions. What we did to overcome this problem was to create a dictionary where each digit could be modelled by the digit model or by the digit model followed by a filler model. We performed an alignment based on this dictionary to let the decoder choose, according to the speech data and models we had, what was the best “pronunciation”.

As usual, several re-estimation/recognition steps were performed and evaluated over the development test. The best results are shown in the next section and correspond to digit models and filler model with 5 and 3 gaussian mixture components, respectively.

D. Results

A summary of the results is shown in the next table, where correctness is the percentage of correctly recognised models (taking into account deletions and substitutions) and accuracy stands for the percentage of correctly recognised models minus insertions. BL stands for baseline system and EXT stands for extension to the baseline system.

Table III
Recognition Results

	% Correctness		Accuracy	
	BL	EXT	BL	EXT
Isolated Digits	99.0	**99.4**	99.0	**99.4**
Connected Digits Known-length grammar	97.8	**98.1**	97.6	**98.0**
Connected Digits Unknown-length grammar	97.2	**97.9**	95.1	**96.1**

We can see that a 40 % decrease in the error rate was obtained for the isolated digit task with the extension to the baseline system. Considerable improvements were also obtained for the connected digit task.

To our knowledge, the best results reported so far for a Portuguese isolated digit task are the ones described in [5], for the TELEFALA telephone speech database with a recognition score of 98 %, achieved with a significantly smaller training set.

IV. CONCLUSIONS AND FUTURE WORK

Although the results for connected digit recognition were about 98.0%, we believe this score can be improved in the future and match state of the art recognition scores for other languages, [3], [6].

Two main reasons lead us to draw this conclusion. First, this task presents many variations of each digit, especially at digit boundaries, which are difficult to model with simple word models. Second, explicitly modelling fillers is also a difficult task, which we think we can still improve on as we optimise our filler model. Another point to take into account is that the subset used for connected digit recognition is not yet representative enough, in the sense that, even though each digit can appear in any order, most of the strings do not include any digit repetitions.

Future work will focus on dealing with these problems in order to improve connected digit recognition scores. In particular, we expect to work on context dependency, bootstrapping context dependent models from our context independent ones. This will certainly improve connected digit recognition because it will provide models to the digits boundaries in different contexts.

Also planned for the near future is the recognition of natural numbers, either of money amounts, of strictly natural numbers, or of hybrid natural number and connected digit sentences.

V. REFERENCES

[1] D. Johnston et al., “Current and Experimental Applications of Speech Technology for Telecom Services in Europe”, Speech Communication 23, pp. 5-16, 1997.

[2] SPEECHDAT – EU-project

[3] M. Rahim, C.H. Lee, B.H. Juang, “Robust Utterance Verification for Connected Digits Recognition”, Proc. EUROSPEECH’95, Madrid, Spain, 1995.

[4] R. J. Perdue, “The Way We Were: Speech Technology, Platforms and Applications in the ‘Old’ AT&T”, Speech Communication 23, pp. 31-39, 1997.

[5] F. Perdigão, “Modelos do Sistema Auditivo Periférico no Reconhecimento Automático da Fala”, PhD. Thesis, Coimbra, 1998.

[6] F.J. Caminero, L. Hernandez-Gomez, C. De La Torre, , C. Martin del Alamo, “Improving Utterance Verification Using Hierarchical Confidence Measures in Continuous Natural Numbers Recognition”, Proc. ICASSP’97, Munich, Germany, 1997.