DESIGN AND IMPLEMENTATION OF AUTO ATTENDANT SYSTEM FOR THE T.U.C. CAMPUS USING SPEECH RECOGNITION

K. Koumpis, V. Digalakis,
Technical University of Crete
731-00 Chania, GREECE
email: {koumpis,vas}@telecom.tuc.gr

H. Murveit
Nuance Communications
Menlo Park, California, USA
email: hy@nuancecom.com

ABSTRACT

We present an auto-attendant system, which is based on a statistical speech recognizer and has been developed for the Technical University of Crete (TUC) Campus. Auto-attendants allow remote callers to reach a person or department by simply speaking an appropriate name. This is the first speech recognition system in Greece operating in continuous speech and speaker-independent modes, and we describe our approaches for solving several special phenomena specific to the Greek language. The high recognition accuracy of the engine supports several hundreds of names. Evaluation on our database yielded more than 97.5% name retrieval for a dictionary of 350 names of persons and services.

1. INTRODUCTION

The field of automatic speech recognition has attracted a lot of interest during the last decade, due to its demonstrable increase in productivity by greatly assisting human operators or by replacing the human element altogether [7]. In this paper we deal with the implementation of an auto-attendant system, which allows callers to reach a person by simply speaking the name of a person or department. Speech recognition of names over the telephone is a difficult task, because each telephone channel has its own and unique signal to noise ratio (SNR) and frequency response. The complexity of problems, which are faced in this work, need fast and accurate techniques, which provide robustness for both speaker and acoustic variability (variation in speech rate, context and dialect, background noise, room acoustics, unknown channels). Speech transmitted over telephone lines can be non-linearly distorted and corrupted by transient interference.

Hidden Markov models (HMMs) are used to represent speech in our system. Hidden Markov modeling is a powerful technique capable of robust and succinct modeling of speech [9]. We used phonemes as our basic acoustic units, which were represented by Gaussian-mixture continuous HMMs. We compared systems with different degrees of tying among the Gaussian mixtures, maintaining a small total number of Gaussians because of the limited amount of training data that we had in our disposal.

With their efficient maximum-likelihood training and recognition algorithms, HMMs can be successfully applied today to constrained tasks in real-world applications. An auto-attendant is such an example, where the system functions as an automated receptionist, asking users to answer specially designed questions and waiting for their responses. The utterances are recognized in real time using Nuance Communications' recognition engine [6], and the recognition result is processed to execute the conference between the two parties.

The organization of the paper is as follows. In Section 2 we describe the system's components and the data collection. The development of language and acoustic models is presented in Sections 3 and 4. The paper is concluded in Section 5.

2. DATA COLLECTION

To train our HMM-based speech recognizer, we implemented an over-the-telephone data-collection system using Dialogic hardware resident in a 200 MHz Pentium Pro PC with 128 Mbytes of RAM operating under Solaris 2.5.1 (see Figure 1). The utterances were spoken by respondents into telephone handsets and recorded directly through an analog connection to the usual switched telephone network in 8-bit mu-law digital form.

The corpus consists of subjects reading various prompts organized in sections, as in the Macrophone corpus [8]. The sections are adapted to our application, and include a set of newspaper sentences, names of members and departments of the university, single-word answers, spelled words, digit sequences, and natural numbers.

The data collection process began by distributing the material to prospective callers in the form of unique prompting sheets. The data collection had two phases. During the first phase, in which we tried to capture speaker variability, 180 prompting sheets of 47 prompts were distributed, and 120 calls were received. Since the total amount of training data was small, we continued the data collection with a second phase, in which we collected a larger amount of task-specific data per speaker from a smaller number of speakers, using a 136-prompt sheet of task-specific sentences. During the second phase, the subjects were instructed not to read the prompts, but to simulate real usage of the system and ask for the persons or departments using natural language. The data from some of the callers of the second phase were reserved for system testing. On sheets for training we had added 20 newspaper sentences. This phase resulted in 12 callers for training and 10 calls for testing (equal number of males and females). The details of the training and testing data collected during the two phases are summarized in Table 1.

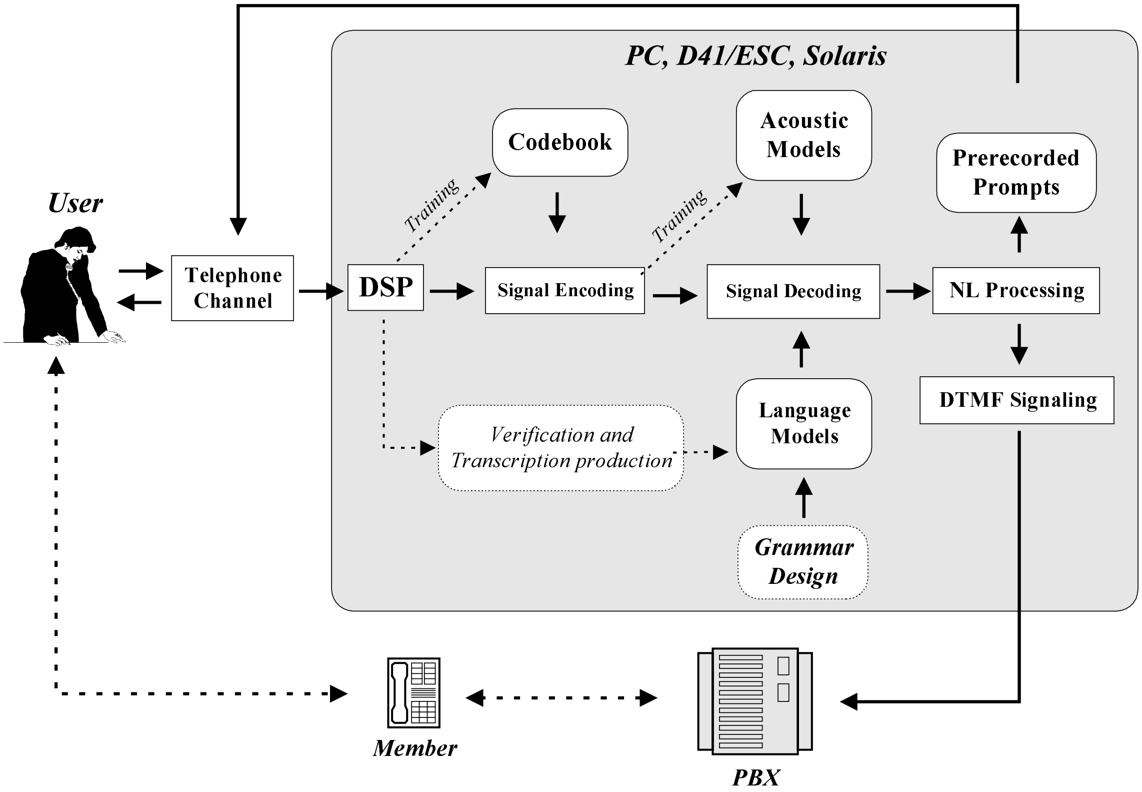

Figure 1: PC-based auto-attendant system using Nuance's speech recognition engine and a Dialogic D41/ESC card.

	1^{st} Phase Training	2^{nd} Phase Training	2^{nd} Phase Testing
Number of Callers	120	12	10
Newspaper sent.	589	239	-
Names of members	1353	1193	1342
Related words	974	-	-
Answers of one word	1042	-	-
Spelled words	455	-	-
Numbers	689	11	-
Total	*5102*	*1443*	*1342*

Table 1: Greek auto-attendant corpus

3. GRAMMAR DESIGN

Our goal was the implementation of a system without restricting the way someone could ask for a particular person, department or university office. Our experience has shown that in Greek there is a much larger variation than in English in the natural language constructions that can be used in this type of task.

For example, two phenomena that were faced are the accusative case and clitic doubling. The accusative case is very common in asking names, while nominative is also allowable but more impersonal in conversations. For some speakers of Greek, the clitic doubling is obligatory, at least for indirect objects, while for others it is optional, with an emphatic meaning.

Greek: *“Τον Κουμπή* ***(τον)*** *Κωνσταντίνο”*
English: *“Koumpis Konstantinos”*

We also had to distinguish members according to their specialty, since in a campus environment there is a variety of occupations. A professor, for example, could be named both with his title “καθηγητής” (Professor), as well as with the title “*κύριος*” (sir), while someone else only with “*κύριος*”.

For example, for an assistant professor with first name “Βασίλης” (Vassilis) and last name “Διγαλάκης” (Digalakis), some allowable expressions are (the accusative case for this name is “Βασίλη Διγαλάκη”):

“*Θα ήθελα τον Βασίλη Διγαλάκη”*

“*Θα ήθελα τον Βασίλη το Διγαλάκη”*

“*Δώστε μου το Διγαλάκη Βασίλη”*

“*Δώστε μου το Διγαλάκη το Βασίλη”*

“Με το Διγαλάκη Βασίλειο”

“*Ναι, το Διγαλάκη το Βασίλη”*

“*Μπορώ να μιλήσω με τον κύριο Διγαλάκη Βασίλη;”*

“*Συνδέστεμε με τον Διγαλάκη το Βασίλη”*

“*Με το Βασίλη Διγαλάκη παρακαλώ”*

“*Θα ήθελα τον επίκουρο καθηγητή Βασίλη Διγαλάκη”*

“Μπορώ να μιλήσω με το Βασίλη Διγαλάκη”

“Μου δίνετε τον καθηγητή Διγαλάκη Βασίλη;”

“Τον επίκουρο καθηγητή Διγαλάκη Βασίλη παρακαλώ”

“Τον καθηγητή Διγαλάκη Βασίλη, θα ήθελα παρακαλώ”

“Με συνδέετε με τον κύριο Διγαλάκη Βασίλη;”

“Είναι εύκολο να μιλήσω με τον κύριο Διγαλάκη Βασίλη;”

We designed a grammar using Nuance Developer’s Toolkit™ [6]. The grammar was written in the toolkit’s grammar specification language that implements finite state grammars, and covered a large number of natural language expressions. Grammar constraints, like the ones implied by the case used (accusative or nominative) or the gender, were implemented using different paths in the grammar. Members of the university were clustered according to their occupation (e.g. professors vs. others), and their names appeared in different subgrammars in order to implement consistency constraints between the possible ways of asking a person and the person’s real occupation. Understanding was achieved using the toolkit’s natural language capabilities. Interpretations for spoken commands are achieved by filling the values of certain slots based on the particular grammar path that was followed in the recognizer string.

4. SYSTEM DEVELOPMENT

In this Section we present the experiments which we did in order to optimize the accuracy of the recognition engine by testing several acoustic models. By using sophisticated modeling techniques to exploit all available training data, our system connects to the correct user 97.5% of time.

Experiments were carried out using Nuance Communication’s batchrec™ [6], which is based on SRI’s DECIPHER™ system [4] and the full grammar described in Section 3. The system’s front-end was configured to output 8 cepstral coefficients, cepstral energy and their first and second derivatives. The cepstral features are computed with a fast Fourier transform (FFT) filterbank. We investigated both phonetically tied mixture (PTM), where the same Gaussians are shared among all allophones of the same phone, and genonic HMM models [4], where the same Gaussians are shared among automatically derived clusters of HMM states. The initial context-independent boot models were constructed by mapping similar sounding context-independent PTM English models to the phones contained in the Greek phone set. We used two measures to evaluate the different systems, the conventional word-error rate (WER) as well as the natural-language error rate (NL error rate, NLER). WER is based on the transcriptions of the test sentences, whereas NLER is the percentage of the times the system connects the caller to an incorrect extension. In all the comparisons between different acoustic models that follow, the test set consisted of sentences which might not be fully covered by the grammar, but consisted of valid commands, i.e. the caller tried to connect to a person or department that was in the dialer database.

We initially performed some experiments to determine the best phone set. In these experiments we used a PTM system with a 100 Gaussians per phone, and the training procedure consisted of two context-independent and two context-dependent iterations of the forward-backward algorithm [1]. The context-dependent models were triphones, with backoff to the corresponding biphone models when a triphone was not seen a sufficient number of times in the training. The best phone set consisted of 40 phones, presented in Table 2. We found it helpful to distinguish between stressed and unstressed vowels (indicated by the symbol ‘+’), and this is particularly convenient since in Greek stress helps to disambiguate the meaning of certain homophone words. Distinction between long and short vowels is not very important in modern Greek [2], except for the vowel “ι” (phones “ι” and “y”, which correspond to the English phones “ih” and “iy”). In addition, we used allophones for certain consonants that are used when these consonants appear before the vowel “ι”, and these are denoted by the symbol 2 (e.g. “χ2”).

-, α, α+, β, γ, γ2, δ, ε, ε+, ζ, ι, ι+, y, θ, κ, κ2, λ, λ2, μ, ν, ν2, ξ, ο, ο+, π, ρ, σ, τ, φ, χ, χ2, ψ, ου, ου+, γκ, γκ2, μπ, ντ, τζ, τσ

Table 2: Greek phone set used in our system.

In a second set of experiments, we evaluated the effect on performance that the amount and origin of the training data had. Specifically, we compared three PTM systems trained with data from the first phase only (PTM_04 - 5120 general training sentences from 120 speakers), with data from the second phase only (PTM_05 - 1443 task-specific sentences from 12 speakers) and with the combined data from both phases (PTM_01). The results are presented in Figure 2, and we see that the system trained on many speakers is significantly better than the one trained on 12 speakers, since it has over three times more training data. However, simply adding the small amount of task-specific data of the second phase halves the NLER, reducing it to 3.3%. The exact word and NL error rates of the various systems evaluated are summarized in Table 3.

Experiment	Word %ERROR	NL %ERROR
PTM_01	20,26	3,28
PTM_04	24,61	6,41
PTM_05	31,02	10,13
PTM_06	20,08	3,20
PTM_07	19,77	3,58
PTM_08	21,32	3,50
PTM_09	21,63	4,55
GEN_01	18,97	2,46
GEN_02	18,28	2,98
GEN_03	17,72	2,91

Table 3: Word- and natural-language error rates of various acoustic models.

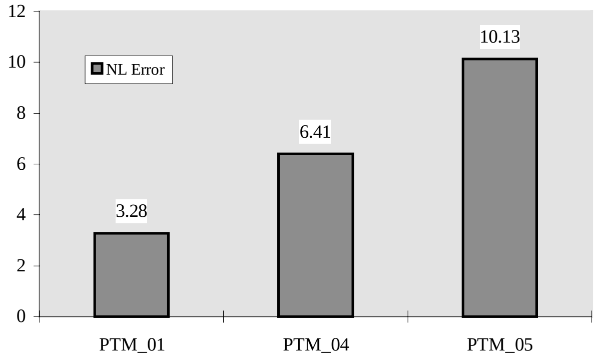

Figure 2: Comparison of models using different amounts and types of training data.

In the next phase, we used a different booting procedure. We did not initialize the models of Greek phones by copying similar U.S. English models, but initialized the Gaussians from alignments using the expectation-maximization (EM) algorithm [3]. We evaluated four different PTM systems, PTM_06, PTM_07, PTM_08 and PTM_09 with 100, 64, 50 and 32 Gaussians per phoneme, respectively. We can see from Table 3 that the performance of the system PTM_06 is similar to that of PTM_01, which was initialized from English phones, whereas the systems with smaller numbers of Gaussians per phoneme are significantly worse.

In the final set of experiments, we compared the PTM system to systems with smaller degrees of tying. We clustered the PTM_01 system to genonic systems with 100 (GEN_01), 193 (GEN_02) and 340 (GEN_03) genones (Gaussian codebooks) with 32 Gaussians per genone, following the procedure described in [4]. The genonic systems outperformed the PTM system significantly, reducing the word error rate by 12.5% and the NL error rate by 25%.

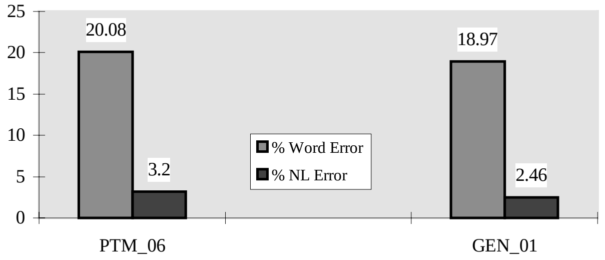

Figure 3: Comparison between the best PTM and Genonic Model.

The best system achieved a NL error rate of 2.46%, which means that the caller was connected to the correct person or department 97.5% of the time. It is worth noticing that the systems PTM_01 and GEN_01 had the same total number of Gaussians, but with different degrees of tying: PTM_01 had 30 codebooks with 100 Gaussians each, whereas GEN_01 had 100 codebooks with 32 Gaussians each. These two systems are compared in Figure 3.

5. CONCLUSIONS

In this work, by using sophisticated modeling techniques to exploit all available training data, we managed to reach a higher level of accuracy for an auto-attendant application. Our experiments showed that our system connects to the correct user 97.5% of time. This auto attendant is now in public use at the university. Our system is the first speech recognition system in Greece operating in speaker-independent and continuous-speech modes, and we addressed issues relating to the phonetic alphabet and the grammatical constraints of the Greek language. In addition, we compared continuous HMM systems with similar numbers of Gaussian distributions, and we found that those with a smaller degree of tying among different HMM states were significantly better.

This work can be extended in several aspects in the future. For example, by working in auto-attendant applications with a much larger number of PBX subscribers [5]. Such applications are expected to be a critical component of further implementation of computer-based secretarial assistants and database access and management [7].

REFERENCES

[1] Baum "An Inequality and Associated Maximizitation Technique in Statistical Estimation for Probabilistic Functions of Markov Processes" *Inequalities, Vol. 3, pp. 1-8,1972.*

[2] Bernard Comrie "*The World's Major Languages*" Oxford University Press, *1990.*

[3] A. P. Dempster, N. M. Laird, and D.B. Rubin "Maximum Likelihood from Incomplete Data via the EM" *J. Roy. Stat. Soc., Vol. 39, no. 1, pp. 1-38, 1977.*

[4] V. Digalakis, P. Monaco and H. Murveit "Genones: Generalized Mixture Tying in Continuous Hidden Markov Model-Based Speech Recognizers", *IEEE Transactions on Speech and Audio Processing, July 1996.*

[5] J. C. Junqua, "SmarTSpell: A Multipass Recognition Systemfor Name Retrieval over the Telephone", *IEEE Transactions on Speech and Audio Processing, Vol. 5, No. 2, March 1997.*

[6] Nuance Speech Recognition System, Version 5, Developer's Manual. *Nuance Communications, 1996.*

[7] L.R. Rabiner, "The Impact of voice processing on modern telecommunications" in *Speech Communication, 17, pp. 217-226, May 1995.*

[8] K. Taussig and J. Bernstein, "Macrophone: An American English Telephone Speech Corpus" *Proceedings DARPA Workshop, SRI International, 1995.*

[9] S. J. Young, "A Review of Large Vocabulary Continuous Speech Recognition" *IEEE Signal Processing Magazine pp. 45-57, September 1996.*