An Interactive Directory Assistance Service for Spanish with Large-Vocabulary Recognition

R. Córdoba, R. San-Segundo, J.M. Montero, J. Colás, J. Ferreiros, J. Macías-Guarasa, J.M. Pardo

Grupo de Tecnología del Habla. Departamento de Ingeniería Electrónica. Universidad Politécnica de Madrid
E.T.S.I. Telecomunicación. Ciudad Universitaria s/n, 28040 Madrid, Spain
cordoba@die.upm.es
http://www-gth.die.upm.es

ABSTRACT

In the EU funded IDAS project (LE4-8315), demonstrators providing an automated interactive telephone-based directory assistance service have been developed by ten partners from Germany, Greece, Spain and Switzerland [6]. In this paper we will focus in the Spanish demonstrator. In particular, we will describe the following aspects:

The general architecture of the system, paying special attention to the speech recognition module. We will present new alternatives for the estimation of continuous HMMs and the agglomerative clustering of context-dependent units.

The most common problems encountered in the development of this kind of systems and their operation in a real environment.

Impressions, opinions and scores from real-world users of the system.

Keywords: large vocabulary recognition, telephone-based, directory assistance service, dialog.

1. INTRODUCTION

In the IDAS project, we address the challenging problem of automating the provision of directory assistance services to the public over the telephone network. The technical challenge that has to be tackled makes high demands on each of the speech processing components:

- A speech recognizer for Large Vocabulary over the telephone.
- A speech production system able to speak out any imaginable phone directory entry.
- A dialogue component that can interpret user inputs and ask the right questions in order to guide the users quickly to the desired information.

Directory assistance services are very interesting for telephone companies, because they save operator time and the information that has to be provided is very reduced (the desired telephone number). This aspect reduces user rejection, specially if the service is not free.

It is important to achieve user satisfaction from the beginning, if a system is operating in real life. However, speech technology is still far from being perfect. One of the project's main focuses was therefore to provide the user with a high success rate, independently of the current state of technology. To this end, the system design incorporates an operator fallback component.

The main difficulty in a system like this is the noise and the reduced signal to noise ratio that is common in a telephone line. The second difficulty is the high degree of confusability that arises when you consider 10,000 surnames in Spanish, because we need their exact transcription. As a perfect recognition system is not feasible, we have to introduce new alternatives in the dialog: user confirmation and spelling. We have developed a spelling module that is very robust and helps to disambiguate a great number of entries.

In this paper, we will describe a series of improvements that have been applied to a large vocabulary isolated-word recognition system using continuous models. We will cover improvements in the techniques for continuous HMMs and agglomerative clustering.

2. DESCRIPTION OF THE SYSTEM

The demonstrator presented in the paper has the following characteristics:

- Representative database with some 1 million registers.
- 4 different vocabularies: cities, first-names, surnames and company names. All of them, using 10,000 words. We have obtained them from the results of project Onomastica (except for the company names).
- Provides telephone numbers for private users and companies.
- System-driven dialogue optimized to increase the transaction success.
- With operator fallback if the recognition fails.
- We have recorded the most common system prompts to improve the general acceptance of the system voice answer.
- All the messages used for confirmation were generated using our Spanish text-to-speech system [1]. Special attention has been dedicated to names pronunciation.

When the recognition module fails, and before falling back to the operator, the user is asked to spell the misrecognized data, as an intermediate step. If the spelling fails too, the operator receives a dialog box where all information about the call is present: recognition results, unrecognized entries and different icons that play what the user has said in each step of the dialog. This way, the operator is able to complete the missing entries, allowing the system to make the query to the database with the correct information. Everything is solved in transparently, without any intervention from the user, and there is a warranty that all calls can be handled.

This demonstrator has been implemented using a TADE (Telephone Application Development Environment) system, developed completely in our research group. This system has a high level language for designing telephone applications. This environment provides tools for the whole application life: designing, compilation and execution. A telephone application specified with our system consists in several sections: variables initialization, error handling, subroutines, and application, where we specify the high level instructions.

We have implemented functions for:

- Phone-line management, call redirection and call progress analysis
- Database queries
- E-mail sending
- Speech tools for recording, playing, synthesis and recognition
- Automatic generation of recognition dictionaries from lists of words or expressions
- Multimedia elements management for video, images, and speech
- Generation of detailed logs needed for the evaluation of the system

This system is working now in several customer care centers with success, for applications different from the directory assistance service.

3. SPEECH RECOGNITION MODULE

3.1 Database

We have used the SpeechDat database, the isolated speech part, with 1,000 speakers who utter the following items: application words, isolated digits, cities, companies, names, surnames, and spelled names. We have divided this database into two parts: 9,069 words for training and 3,840 for recognition.

3.2 System architecture

As real-time is a must in the system, we decided to use a hypothesis-verification approach to reduce the number of candidates that have to be considered with detailed modeling in our Large Vocabulary recognition system.

- In the hypothesis step, we use a fast preselection module [4]: with context-independent semi-continuous HMMs (CI-SC), we obtain a sequence of recognized phones which is followed by a lexical access module. This module passes N-best candidates to the verification step.
- In the verification step, using context-dependent semi-continuous HMMs (CD-SC) or continuous HMMs (CD-C), and whole word recognition we obtain the recognized word.

In this paper, we will concentrate in the verification step. We had previously used CD-SC with success [2][3], so we concentrated our efforts in the CD-C, which we had used before for another task in [3].

3.3 Context-dependent continuous models

We have followed an agglomerative clustering approach at the state level, as in [3]. These are the new techniques considered for this process:

3.3.1 Best "distance between states"

We had previously used the distance from **¡Error! No se encuentra el origen de la referencia.**, but weighed by the number of vectors available in training for each state to favor the clustering of states with little training data:

$$D(i,j)=\sqrt{\left(\frac{n_i * n_j}{n_i + n_j}\right)\left(\frac{1}{codes}\right)\sum_{s=1}^{codes}\left[\left(\frac{1}{param}\right)* \sum_{K=1}^{param}\left(\frac{(\mu_{isk}-\mu_{jsk})^2}{\sigma_{isk}*\sigma_{jsk}}\right)\right]}^{1/2}$$

where n_i, n_j are the number of vectors for state i and j.

In the best experiment we got a 9.56% error rate with a vocabulary of 1,000 words.

We have switched to a new distance, which is based in the minimum loss of information produced by the clustering, weighed by the number of vectors assigned to the states in training.

$$D'(i,j)=(n_i+n_j)*\sum_{k=1}^{d}\ln(\sigma_k)-n_i*\sum_{k=1}^{d}\ln(\sigma i_k)-n_j*\sum_{k=1}^{d}\ln(\sigma j_k)$$

where σ_k is the standard deviation of the distribution obtained after the clustering, σi_k, σj_k are the standard deviations of the original distributions *i* and *j*, and *d* is the dimension of the vector of parameters.

This distance has produced better results: 8.32% error rate with a similar number of total mixtures (13% improvement over D).

3.3.2 Best strategy for the clustering

We compared two alternatives:

A) *Clustering with single-mixture models.* After the clustering, we reestimate the models increasing the number of mixtures in each state (similar to context-independent (CI)).

B) *Clustering with multi-mixture models*. We followed the following iterative approach:
   1) Reduce the number of units: find the closest states and merge them.
   2) Increase the number of mixtures (similar to CI).
   3) Reestimate the models. If number of units > objective, go to step 1. If not, stop.

The results obtained with this approach show a great adaptation to the training data (18% improvement with the same number of parameters), but offers little improvement for the recognition set (7%).

3.3.3 Optimum number of mixtures in each state

We compared three alternatives:

A) Same number of mixtures for all states.

B) Number proportional to the number of vectors assigned to the state in the training.

C) An innovative approach: we split the mixtures that provide the largest reduction in entropy. So we adapt the process of increasing the mixtures to the training data.

The results obtained with options b and c are the best. Again, there is no significant difference between them for the recognition set, but for the training set the improvement of option c over b is 27.7%, which shows its effectiveness.

When we applied the same technique to context-independent models, where the training data available is more adequate to the number of parameters estimated, option C was better than option B both for the test set (3.5%) and the training set (8%).

We are now preparing a larger database to check the improvements of both techniques.

3.4 Results with SpeechDat database

We applied our best systems to the final vocabulary used in the IDAS system (10,000 words) and we obtained the following WER for each system:

- CD-C: 23.1% (first candidate) and 14.8% (including two candidates).
- CD-SC: 25.4%.
- CI-C: 34.7%.

Another aspect that should be remarked is that results could be better as there are some mistakes in the database that have not been excluded from the test.

3.5 Spelling module

This module is specially difficult in Spanish, because people is not used to spell and they make a lot of mistakes. We followed these steps [5]:

1) Use a Continuous HMM "word" model for each letter and for noise. The output in this stage is a sequence of phonemes.
2) Align the sequence of letters with all the names in the dictionary. The output is a set of the best 50 candidates in the dictionary.
3) Repeat step 1 without noise and considering only the best candidates of step 2.

The results are really good (a WER of 3.9% in first candidate with a 1,000 words vocabulary) and the module recovers from many of the mistakes made in the recognition stage.

4. FIELD TESTS

The audience of our field test consisted of all kinds of users, most of them not used to speech processing systems. We announced our system in some distribution lists, at the University and through e-mailing.

We prepared a questionnaire for the users with the following sections:

- General comments about the system. The focus was made on the organization of the test, not in the system itself, and the way to fill the form.
- A form with 10 company entries and 10 private entries. The user had to take down for every item if it was recognized in first place, second place, after the spelling or not recognized.
- A satisfaction questionnaire. We will see the questions in next section.

The system first asks for the city. Then, for private entries, we ask for the surname and, at last, for the name. For company entries, we just ask for its name.

In each query, the user is asked to confirm the result. If he/she rejects it, a second candidate is proposed. If it is rejected again, then he/she is asked to spell the name. If the result from the spelling is rejected, then there is a fallback to the operator, but only with the item that has been rejected.

In Table 1, we can see the recognition rate obtained for each dictionary:

- In first place: with the first candidate.
- In second place: with the first or second candidate.
- In the spelling module. This module is only used when the previous recognition fails, so the noise is usually high and the results decrease
- Global rate: all modules are considered.

All dictionaries used in these tests were 1,000 words in size, except for surnames, where 10,000 words were used.

	City	Company	Surname	Name
1st place	60.56%	64.55%	30.13%	51.17%
2nd place	68.10%	69.63%	38.24%	56.19%
Spelling	45.92%	38.60%	36.34%	37.69%
Global	82.75%	81.36%	60.68%	72.90%

Table 1. Recognition rate for 1,420 queries

Names is a special case because: we did not offer a second candidate to speed up the dialog; and if the recognition of the city and surname was correct, we restricted the recognition to an average of 10 names and obtained a 100% recognition rate. So, in Table 1, what we present is the result for names when the recognition of city and/or surname was incorrect, and that is the reason of the low results: considers only calls with previous mistakes, that usually have a lot of noise.

The global rate for query success without operator intervention is **58.80%**. The average duration of the dialog has been 84.2 seconds for private entries and 62.4 for companies, including system answers, confirmations and spelling, which is particularly slow.

The error rate for the companies is lower. The reason is the longer average duration for company names (14.3 phonemes per word) which allows a better discrimination. The worst recognition rate corresponds to first names and surnames, as they have the shorter average duration (7.2 and 6.7 phonemes, respectively).

Another important factor is the confusability between words belonging to the same vocabulary. With a dynamic programming algorithm we computed the phonetic distance between every word and the closest one in the vocabulary. We obtained the following average distances: surnames 1.3, first names 2.3, cities 3.3 and companies 6.8. As we expected, the worst error rates correspond to the vocabularies with smaller average phonetic distance.

4.1 Problems encountered

These are some problems that affect the system:

- Some times, there is a problem with the end-point detector, and recognition ends before the user has finished speaking. It is always related to very noisy telephone lines.
- The dialog is often affected by poor perception of names as synthesized in confirmation questions.

When we have very big vocabularies that try to consider all possible cities and names, the effect is that there are many entries in the vocabulary that are totally unknown to the caller. So, when the synthesizer wants to confirm the recognition result, the user does not understand what the synthesizer says, as it has no meaning for him/her.

We can conclude that if the system is easy to use, the users can adapt themselves to its deficiencies and they obtain a best impression of the system behavior.

5. USER SATISFACTION

The total number of questionnaires we received was 58, 39 male and 19 female, with ages ranging from 14 to 51 years old. Each question should be assigned a qualification from 1 to 5, with the following meaning: 5- I fully agree, 4- I agree, 3- Rather yes, 2- Rather no, 1- Disagree.

Table 2 shows the results for each question of the questionnaire.

Question	**Average result**
You have experience in this kind of systems.	2.42
The system understands what you say.	3.00
System responses are clear and precise.	3.51
I understand what the system says.	3.25
You access quickly to the telephone number.	3.71
The system is easy to use.	4.32
The system is easy to learn.	4.42
The system helps me during the interaction. It explains and feedbacks correctly.	3.68
In case of error, the correction process was easy.	2.53
The system asks me in a logical sequence.	4.00
I prefer to dial to the system instead of looking for the number in the White pages.	3.02
In general, it is a good system.	3.32

Table 2. User satisfaction

The global results are really good. Most of the users (61%) do not have experience with automatic recognition systems and think (63%) that the system is able to understand what they say.

57.6% of the users think that the system responses are clear and precise. There is a significant peak of users (27.1%) that do not think so. Probably the culprit is the synthesis module used for the recognition confirmations, as we mentioned in section 4.1. May be we should have made a separate question in the questionnaire for the recorded speech and the synthesized speech.

69.5% of the users think that they can access quickly to the phone number, which is a great score and shows that the global dialog is quick. 85% of the users think that the system is easy to use and 88% that it is easy to learn, which again is positive.

6. CONCLUSIONS

We have successfully developed an Interactive Directory Assistance Service. The system is working with real users. The architecture using a recognition module and a spelling module is able to recover from many errors and improves the automation rate of the service.

The transaction rate is more than acceptable and we provide a user-friendly operator module to solve the recognition mistakes.

User satisfaction can be considered as outstanding.

New techniques have been developed for continuous HMMs and agglomerative clustering with outstanding results.

7. REFERENCES

[1] Pardo, J.M., Giménez de los Galanes, F.M., Vallejo, J.A., Berrojo, M.A., Montero, J.M., Enríquez, E., Romero, A. "Spanish text to speech: from prosody to acoustics". 15th International Congress on Acoustics, pp. 133-136. 1995.

[2] Córdoba, R., X. Menéndez-Pidal, J. Macías, A. Gallardo, J.M. Pardo. "Development and improvement of a real-time ASR system for isolated digits in Spanish over the telephone line". Eurospeech 95. Vol. II, pp. 1537-1540.

[3] Córdoba, R., J.M. Pardo. "Different strategies for distribution clustering using discrete, semicontinuous and continuous HMMs in CSR", ICSLP'96, pp. 1097-1100.

[4] Ferreiros, J., J. Macías-Guarasa, A. Gallardo, J. Colás, R. Cordoba, J.M. Pardo. and L. Villarrubia "Recent work on preselection module for flexible large vocabulary speech recognition system in telephone enviroment", ICSLP'98, pp. 1689-1692.

[5] San-Segundo, R., J. Colás, J. Ferreiros, J. Macías-Guarasa, J. M. Pardo. "Spanish recogniser of continuously spelled names over the telephone". ICLSP'00.

[6] Lehtinen, G., S. Safra, …, J.M. Pardo, R. Córdoba, R. San-Segundo, et al., "IDAS : Interactive Directory Assistance Service", VOTS-2000 Workshop, Belgium.

[7] S.J.Young, P.C. Woodland, "State clustering in hidden Markov model-based continuous speech recognition". Computer Speech and Language 1994 vol. 8, pp. 369-383.