# INTRODUCTION TO THE IST-HLT PROJECT SPEECH-DRIVEN MULTIMODAL AUTOMATIC DIRECTORY ASSISTANCE (SMADA)

*Fréderic Béchet[4], Elisabeth den Os[1], Lou Boves[2], Jürgen Sienel[3]*

[1]KPN Research, Department of Multimedia, Leidschendam, The Netherlands
[2]KUN, Department of Language and Speech, Nijmegen, The Netherlands
[3]Alcatel SEL AG, Stuttgart, Germany
[4]Université d'Avignon et des Pays de Vaucluse, LIA, Avignon, France

## ABSTRACT

This paper introduces the IST-HLT project SMADA. This project started in January 2000 and it will run for three years. We present information on the topics to be addressed in the project; in addition some preliminary results are given. Information is given on functional specifications and technical and Human Factors evaluations of Automatic Directory Assistance systems in four European countries, on the technology issues taken into account (e.g., automatic derivation of pronunciation variants and grammars, combination of multiple decoders, confidence measures), on multimodal access to Directory Assistance and in relation to this on Distributed Speech Recognition.

## 1. INTRODUCTION

Directory Assistance (DA) is the service that provides information on telephone numbers, once a name and address are provided by the customer. For many European countries the annual number of calls to this service is about 10 times the number of subscribers. DA belongs to the Concession Services, that the traditional Telecom Operators are obliged to offer in one way or another, for a limited price. Thus, Operators are forced to make the service as efficient as possible. Automatic speech recognition may help. In January 2000 four Telecom Operators (France Télécom, Telecom Italia, Royal Dutch PTT, and Swisscom), a telecom equipment manufacturer (Alcatel), and three universities (Nijmegen University, University of Avignon, Politecnico di Torino) joined forces in the 5th Framework project SMADA to develop the advanced ASR technology needed for the automation of a large proportion of the DA service. Increasingly, calls come from mobile environments. In addition to aggravating the task for ASR this also creates a demand for multimodal interfaces to mobile information services like DA. The SMADA project will also deal with these issues.

In the next sections we will describe the research issues addressed by SMADA. In addition, some preliminary results will be given.

## 2. FUNCTIONAL SPECIFICATIONS AND EVALUATIONS OF THE DA SYSTEMS

It appears that there are substantial differences between the functional specifications of the DA service in different countries. In this section we review the functional specifications, and we describe ways in which automated DA services can be evaluated.

### 2.1 Functional specifications of the Systems

Country specific properties of human operated and automatic DA services are caused by cultural differences between the countries as well as by differences in the infrastructures of the operational services. For example, the French and Swiss systems will implement spelling, whereas the Italian and Dutch systems will not. In Italy spelling is highly unusual, probably because of the regular relation between orthography and pronunciation. In French homophone heterographs abound; thus, spelling is quite usual. Dutch (and Swiss German) have an intermediate position with respect to regularity of orthography and pronunciation. In the Netherlands customers behave unpredictably when asked to spell. Provided that spelled letters can be recognized with a reasonably low error rate, the letter sequence can be combined with the recognition result of the full names. Several different combinations are possible, depending on the search strategy and the confidence measures attached to the representations of the names. Some systems (like the system under development by France Télécom) only prompt for spelling if the result of the recognition of the full name is too ambiguous.

The databases of the DA services in Italy, France, and Switzerland contain surnames and first names of private listings. This means that both items can be combined for searching the database and/or for reducing the number of items to be recognized in a sub-lexicon (e.g. once a first name has been recognized correctly, the number of corresponding surnames to be recognized can be reduced). In the Netherlands, however, few first names are listed in the DA database; most of the time only initials are present. Thus, first names cannot be used for the search. Trying to recognize initials will not help. Firstly, because the recognition of initials is a very difficult ASR task. Moreover, initials may not always be known to customers, because they often do not correspond with the first letter of the first name (if this first name is known in the first place).

All automatic systems in SMADA will use a structured dialog, eliciting isolated words as much as possible. After the introduction to the system (and sometimes after a question related to the type of listings (private or business) the customer is looking for), the first prompts asks to speak the city name. There are large differences in the number of cities that must be recog-

nized. The Italian system that is already partly operational contains a lexicon of 9,325 city names. A large number of hamlets (about 90,000) is not covered; if one of these is mentioned an OOV word occurs, that must be rejected. The French demonstrator system aims for the recognition of about 36,000 French town names; spelling will be used to support the recognition. For the Dutch system a lexicon of about 2,400 city names is sufficient to cover half of the named localities in the Netherlands (the remaining 2000 hamlets need not to be modeled, because they cannot be used to search the directory). The Swiss system that only covers the German speaking part of the country, contains a lexicon of about 2,600 city names.

Another difference between the systems concerns the focus on private or business listings. Although it is clear that the large majority of the requests are for business listings, some systems will focus on private listings. Automating private listings is easier, especially when first and surnames are listed in the directory. CSELT reported that in a field trial in which 9454 requests for private listings were handled, 30% of the calls could be automated completely [1]. The difficulty of automating business listings is related to the fact that these listings can be expressed by a great number of different formulations that are difficult to predicted from the listing in the directory. Therefore, a lot of (hand) work is needed with very large speech corpora to model these different formulations and pronunciations. One of the goals of SMADA is to automate this process.

## 2.2. Evaluation

In the SMADA project, three evaluations of the DA systems are planned: one at the end of 2000, one in the middle of the project and one at the end of the project (December 2002). The last two evaluations include human factors and technology, the first one is mainly focused on technology.

*Human Factors evaluation*

Today customers expect a high quality DA service. Automatic versions of the service must attain a sufficient level of user satisfaction. The best speech recognition performance will be reached when customers stick to predictable expressions as much as possible. To accomplish this, much attention must be paid to the formulation of prompts. In the Italian system, many different prompt formulations have been tried. The best prompts turned out to be those that give explicit examples of what customers should say. In the Dutch demonstrator system, short direct prompts were used. Also these prompts resulted in relatively large proportion of predictable utterances for city names (85%). The remaining 15% include silence, utterances that are too long and OOV utterances. However, a marketing study involving a large number of customers showed that the prompts were considered as rather unfriendly and impolite.

An automatic service cannot be expected to handle all calls completely. Thus, operator fall back must be provided. If an automatic system has been able to acquire part of the information that is needed to find a telephone number, the operator should not be required to start the dialog from scratch. However, if the call cannot be completed by the automatic system, part of the information may be incorrectly understood. This introduces the need for human factor research related to the needs of the operators.

*Technology evaluation*

The first year of the project is meant to evaluate the state of the art of the technology, and to set the base line. The advanced technology that will be developed in the project will be incorporated in new versions of the systems. This should lead to measurable improvements, in terms of the proportion of calls that are fully automated and in a reduction of the word error rate that can be measured off-line.

First experiments by CSELT have shown that substantial improvements can be obtained for city name recognition by adding a small number of application specific acoustic models. Also a richer set of acoustic features helped to improve the performance. Some experiments were carried out on a spontaneous speech database of 8775 tokens of a city name recorded in the Italian Directory Assistance service prototype. The use of the vocabulary dependent sub-words units for the most common city names significantly improves the performance; further improvement is obtained by the addition of whole word models for the most frequent city names. Table 1 summarizes the results obtained with Continuous Density HMMs.

| Models | Errors | Error rate % |
|---|---|---|
| Vocabulary independent | 1536 | 17.8% |
| + 40 vocabulary dependent sub words | 1122 | 12.8% |
| + 20 whole word models | 1000 | 11.4% |

**Table 1:** Improved Italian city name recognition with additional models (8775 tokens)

# 3. TECHNOLOGY ISSUES

It is well known that human speech recognition performance is still almost an order of magnitude better than ASR. Moreover, recognition of names, often without any linguistic or non-linguistic context is among the most difficult speech recognition tasks for humans. Thus, to deal successfully with an application like automatic Directory Assistance substantial improvements of the state-of-the-art of the basic ASR technology are needed. In this paper we discuss some of the issues in ASR that need to be addressed in SMADA. More detailed information about the state-of-the-art and the research agenda of SMADA in this respect is given in [2].

## 3.1 Semi-automatic learning

In DA there are two problems that will require automatic learning, viz. the derivation of the way in which customers express business listings; if the words and expressions in the lexicon and grammar are known, there is still much to be learned about the likely pronunciation variants. Fortunately, field tests of operational automatic DA systems, as well as Wizard of Oz experiments in preparation of such tests, have provided very large databases of recordings of expressions that were actually used by customers.

In SMADA several partners are developing (semi-)automatic methods that will allow to exploit these recordings to derive expressions and pronunciation variants. The (semi-)automatic derivation of pronunciation variants is especially urgent in the case of foreign names, especially the English business names that abound in non-English speaking countries [3].

However, taking into account all possible pronunciation variants of proper names gives rise to a very large number of highly confusable word models. Furthermore, a given speaker will introduce only certain types of variants of the canonical representations. Thus, a search based on a network with all possible pronunciation forms may lead to an increase in word error rate, because the full set will comprise models which are inconsistent with the behavior of a given speaker.

Another important problem arises from the fact that it is difficult to distinguish between non-canonical form of a word and the possibility that certain sounds intended by the speaker are confused by the recognizer. Yet, humans often perceive different pronunciations of a word as similar, if not equal to the canonical form for that word.

To alleviate these problems we propose to use multiple decoders in parallel. The best word hypothesis is the one for which the global score from the K decoders is maximum:

$$\hat{W} = \arg\max_{W} P(W|A) = \arg\max_{W} \sum_{k=1}^{K} P(Wz_k|A) = \arg\max_{W} \sum_{k=1}^{K} P(AWz_k)$$

Each decoder corresponds to a particular unit model, for which it is possible to write:

$$P(Az_k W) = P(A|z_k W)P(z_k)P(W|z_k) \approx P(A|z_k)P(z_k)P(W|z_k)$$

Where

$$P(A|z_k)P(z_k)$$

is given by the k-th decoder.

Each decoder is trained in such a way that it gives high scores to pronunciations belonging to a set of *word patterns*. These patterns are obtained by an automatic learning procedure based on Semantic Classification Trees [4]; they represent word clusters such that in each cluster word pronunciations belonging to the cluster receive on the average higher score than words belonging to other clusters. Preliminary experimental results were obtained in terms of perplexity reduction by using multiple decoders.

59K French last names with 90K corresponding phonetic transcriptions were made available by France Télécom R&D. 72K phonetic strings were used for training, and the remaining 18K were used for testing. Ten patterns were obtained, yielding an average perplexity reduction of 23.3%.

18K English surnames with 24K phonetic transcriptions that are freely available on the ISIP Web site [5]. Using 19.3K phonetic strings for training and 4.7K for testing ten patterns were obtained, resulting in an average perplexity reduction of 18.7%.

A small corpus of 2369 phonetically transcribed Dutch city names was used for both training and testing (because of the small number of names). Five patterns were obtained with an average perplexity reduction of 13%.

From these results it is evident that the idea of using multiple decoders in parallel deserves further exploration.

## 3.2 Confidence measures

An estimate of the confidence with which a user utterance has been recognised is essential for user friendly human-machine dialogs. The simplest use of confidence measures (CM) is for the Dialog Manager to decide whether the recognised hypothesis should be accepted or rejected. Confidence measures are also useful for unsupervised adaptation of acoustic models, pronunciation variants and grammars.

Several different approaches have been developed to decide whether the best match between a lexicon entry and a speech signal is good enough.

A popular approach is closely related to hypothesis testing, which leads to computing the ratio between the likelihood of the recognition result and the likelihood of that the signal represented anything but the best matching word(s) [6]. Formalisms have been developed to weight the contributions of the phoneme segments assuming that they do not all convey the same discriminative information. CMs based on the comparison of competing solutions (e.g., the first and second best decoding) behave differently depending on the vocabulary. If the vocabulary contains similar words, they will obviously lead to similar acoustic scores, and this will lead to high false rejection on the corresponding words. Such similarity should thus be integrated in the confidence measure computation

An interesting approach is the use of non-spectral features (especially prosodic features, but also segmental features like voicing and energy) as additional information to estimate the confidence of the recogniser output [7].

CMs are also needed to determine the next move in a dialog. If the CM is below a certain threshold, it is probably wise to ask explicitly for confirmation. With very high CM values it may very well be profitable to skip confirmation altogether, or to use some form of implicit confirmation if additional information items must be elicited.

# 4. MULTIMODAL ACCESS TO DA

In the past directory information was available via two sources, the printed directory and the operator. Today, electronic access to directory data is rapidly gaining in popularity. At the same time we see an extremely fast growth of the number of mobile phones, that are often used in situations where no printed directory is available. This leaves the operator based (or its automated equivalent) and Internet as the only alternatives. Internet access is possible by means of speech only, e.g. using emerging standards like VoiceXML. However, there is a growing demand for multimodal user interfaces, that allow callers to select the mode (speech, text/graphics, point and click) that suites their requirements best. This creates a demand for multimodal access

to information services in general and to DA in particular.

The mobile devices of the future will continue to shrink in weight and size. That enhances their portability, but the small size also creates human factor problems: if there is no room for a keyboard, composing texts, even short ones, like the name and address information in a DA service, becomes cumbersome. It is assumed that ASR will play a pivotal role in this respect. However, little is known so far about the ergonomics of combining speech with other input and output modalities. Especially the fact that ASR is not error free might pose strict limits on the conditions under which speech will be the preferred input modality. In a similar vein, synthetic speech may not be the modality of choice to present names. SMADA intends to investigate these issues. Since there are no commercial multimodal terminals available (nor the transmission protocols to support truly multimodal interaction) much of the research work will use mock-ups of terminals and services.

## 4.1 Distributed Speech Recognition

One of the difficulties that must be overcome in the development of multimodal mobile terminals is the limited memory size and computing power that is available. Actually, the power that can be packed in a light weight handset is by no means enough to support advanced applications like automated DA. Distributed Speech Recognition (DSR) has been proposed to solve this problem. The basic idea behind DSR is "divide and conquer": do parameter extraction –that does not require much power-- in the handset, transmit the parameters as data, and do the search and all additional post-processing in a powerful server in the network. Actually, this approach kills several birds with one stone: ASR for calls coming from cellular networks suffers from substantial performance degradations caused by the different coding methods of the speech signal in the mobile network (full-rate GSM, half-rate, etc) and by radio transmission errors, which cannot be repaired at the receiver end.

DSR intends to solve these problems by defining standard acoustic feature that can be computed in the handsets and transmission protocols that minimize the impact of transmission errors (by enhancing error correction capabilities). In addition, the DSR approach can remove the bandwidth limitation of the telephony speech signal, since the speech can be processed at a higher sampling rate without increasing the data rate. ETSI has launched the *AURORA* project that intends to establish standards for feature extraction methods in terminals and corresponding transmission protocols. The first version the *AURORA* standard specified a cepstral feature representation, intended to form the baseline for an enhanced version that is more robust against acoustic background noise[8]. SMADA is working to develop a proposal for a noise reduction technique. Preliminary experiments have shown that word error rates for noisy speech can be reduced by over 30%.

DSR is another area where the lack of suitable transmission protocols has become evident. Therefore, the *AURORA* project has initiated contacts with the WWW community, to define and standardize protocols that are able to support services where data and speech can be mixed in a single session, and where the computational load for speech recognition can be shared between the terminal and network. This also creates a link with research on speech recognition in Voice over IP [9].

# 5. CONCLUDING REMARKS

The SMADA project started only six months ago. Already now the collaborative research has resulted in interesting suggestions for the solution of the multitude of problems that ASR is facing in advanced applications like DA. The project will be working at the border between basic research and commercial applications. This combination is essential to make progress in both directions. The availability of huge amounts of real-life data that become available through operational services will allow basic research that was not previously possible. At the same time the success of the operational services depends very much on the ability of the research to improve ASR performance.


### Acknowledgments

The SMADA project is partially funded by the European Commission, under the Action Line Human Language Technology in the 5th Framework IST Programme.


# 6. REFERENCES


1. Billi, R., Canavesio, F., and Rullent, C. "Automation of Telecom Italia directory Assitance Service: Field Trial Results," *IEEE 4th Workshop Interactive Voice Technology for Telecommunications Applications*: 11-16, 1998.

2. Boves, L., Jouvet,D., Sienel, J., de Mori,R., Béchet, F., Fissore, L. and Laface P. "ASR for automatic directory assistance: the SMADA project", *Proc. ASR2000*, Paris, Sept. 2000.

3. Bartkova, K. and Jouvet, D. "Language based phone model combination for ASR adaptation to foreign accent", *Proc. ICPHS'99*, San Francisco, USA, August 1-7, 1999, vol.3, pp. 1725-1728.

4. Kuhn R. and De Mori R., "The Application of Semantic Classification Trees to Natural Language Understanding," *IEEE Trans. On Pattern Analysis and Machine Intelligence*, Vol. 17, N. 5, pp. 449-460. May, 1995.

5. *http://www.isip.msstate.edu/projects/nbest_pronunciations/*

6. Lee, C.-H. "A Unified Statistical Hypothesis Testing Approach to Speaker Verification and Verbal Information Verification", *Proc. COST 250*, Rhodes, Greece, 1997, pp. 63-72.

7. Bartkova, K. and Jouvet, D. "Usefulness of phonetic parameters in a rejection procedure of an HMM based speech recognition system", *Proc. EUROSPEECH'97*, September 1997, Rhodes, Greece, pp. 267-270

8. ETSI Standard ETSI ES 201 108, European Telecommunications Standards Institute, Feb. 2000.

9. Millner, B. Mobile and IP access to network-based speech recognisers". *Proc. Workshop Voice Operated Telecom Services*, Ghent 2000, pp. 83-86.