ISA: A TRAFFIC JAM INFORMATION SYSTEM BASED ON THE IBM VIAVOICE TELEPHONY TOOLKIT

Carsten Günther[*], *Stefan W. Hamerich*[+], *Siegfried Kunzmann*[*], *and Thomas Roß*[*]

[*] European Speech Research – IBM Deutschland GmbH – Heidelberg – Germany

[+] Natural Language Systems Division – Computer Science Department – University of Hamburg – Germany

ABSTRACT

In this paper we introduce a development platform for telephone based dialogue systems – the IBM ViaVoice Telephony Toolkit. We describe ISA – IBM Traffic Jam Application (germ. *IBM Stau Applikation*) which is a voice operated real time information system for German. The system's domain consists of German motorway inquiries. It is built with the IBM ViaVoice Telephony Toolkit and was designed to make information stored on internet pages available via telephone. The system combines robust speech recognition with a grammar based dialogue module and a web interface to enable the user requesting the information in a natural way. The provided information is read to the user and could even be used for basic routing services. The system comes with a multimodal interface, which enables the user to get the requested information by voice or as an SMS message.

1. INTRODUCTION

The IBM ViaVoice Telephony Run Time and Tools (VVT) provides programmers with tools to develop phone recognition applications. The ViaVoice Telephony Run Time provides the speech recognition platform over which phone recognition applications run in several languages like German, English, and French. The ViaVoice Telephony Tools provide a set of utilities for developing and testing phone recognition applications. The Tools provide extensions for the used scripting languages to make application development and customization easier. Also included is a set of C++ classes and member functions that enable the programmer to develop object-oriented solutions. Additionally, there are several utilities that can be used e.g. to compile and test grammars and to create and check baseforms to support the unlimited vocabulary capability. Furthermore, a GUI is included to develop and administer the whole application itself. Finally a text-to-speech engine is delivered to generate the spoken output.

Based on this toolkit the traffic jam application (ISA) was build. ISA allows the access of information via telephone in a natural way. The backend uses extracted data taken from a web page and consists of traffic information on German motorways. The user can ask ISA for traffic jams on:

- a specified motorway
- a specified motorway in a specified state
- all motorways in a specified state
- all motorways between two specified cities.

This paper is organized as follows. Section 2 describes the system architecture. Section 3 gives information about the used speech recognition engine. Section 4 provides details of the related backend while section 5 reports on the answer generation used for this system and gives a sample dialog. Finally we conclude our major findings and outline the future work that will be done.

2. BASIC ARCHITECTURE OF ISA

2.1 VVT and ISA architecture

ISA represents a telephone based dialogue system that has been designed with the IBM ViaVoice Telephony Toolkit. The main architectural components of the ViaVoice Telephony Toolkit are:

- telephony interface
- process manager
- IVR script
- speech recognition engine
- TTS component

The telephony interface of the toolkit handles the basic telephony functions like accepting and disconnecting calls, detection of hang ups, recording, and playing back audio and DTMF tone detection and is working on the public telephone network.

The process manager is responsible for the startup of system resources in a configurable fashion. It is capable of running on separate machines and managing just the resources for it. In addition, it sequences the startup of the service. The process manager controls channel processes, which are carrying out the call flow.
The Interactive Voice Response (IVR) script is managing the call flow. This script is implemented in the Tcl scripting language.
The speech recognizer of the Telephony Toolkit consists of IBM's ViaVoice speech recognition engine and additional data files trained for telephony channels.
The Text-To-Speech (TTS) engine is provided by ViaVoice Outloud and is as well controlled by the IVR script.

For further details of the IBM ViaVoice Telephony Toolkit refer to [1].

Building a VVT-based application like ISA means the development of a special IVR script for handling the dialog flow, vocabulary and grammar enhancements and the construction of a backend interface to access requested data.

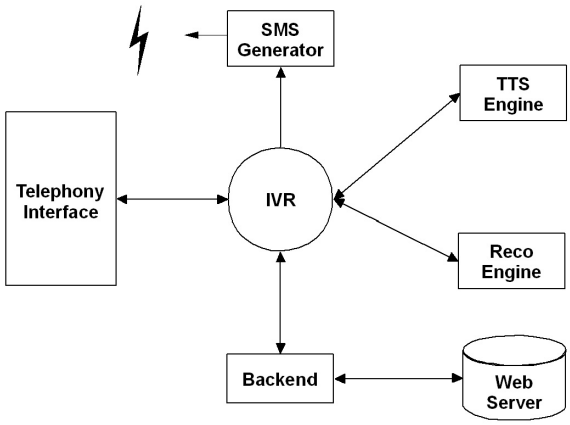

Figure 1: System architecture

2.2 Multimodal output enhancements

Telephony systems are used as natural human-machine interfaces to information systems or enterprise data. Usually the information flow in telephony applications is based on spoken input and spoken output. The speaker utters his request and is prompted by a system response. For longer or complicated messages, a spoken response soon becomes very cumbersome. One approach to overcome this shortage is the usage of textual output onto mobile devices. This could be done by enhancing such a voice operated system with a message generation module. This module extends the speech based communication flow in mobile phone communication situations by sending text messages. It is using the Short Message Service (SMS) of mobile phone networks.

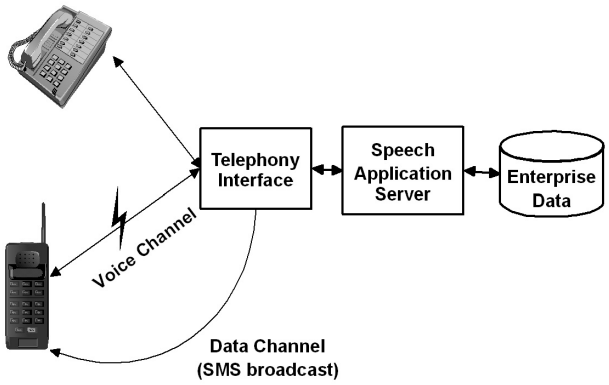

Figure 2: Mulitmodal system architecture

Figure 2 shows the overall scenario. A call arrives through the telephony interface. The interface detects whether the call is coming from a mobile or a landline phone. Sending of SMS messages is only supported for mobile phone devices. The telephony interface can be connected to the Public Network directly, via a PBX, or an IP connection (VoIP). A spoken request from the calling person is put through to the Speech Application Server. The speech application server interprets the request, retrieves the data from a backend enterprise database or directly from the web, generates an answer and puts it back as generated speech through the voice channel. Additionally, if the telephony interface has detected that the call is coming from a mobile phone, the speech application server (answer generator, TTS engine) will generate a question whether or not the caller wants additional paged output (SMS message) of his request. If yes, the speech application server generates a formatted message and broadcasts it back to the caller's device through the data channel. The message has to meet the SMS limitations with respect to message length and screen size. Long responses (more than 160 characters) will be spliced into multiple message chunks. Caller identification is done automatically by the telephony interface (CLIP - Calling Line Identification Presentation) or can be supported by the Speech Application Server via digit recognition capabilities.

3. SPEECH RECOGNITION

As described above the speech recognition engine used here is the IBM ViaVoice decoder. In this section, a short summary of basic aspects of the training as well as the recognition is given.

The training of the 8kHz system is a bootstrap procedure based on an initial acoustic model built from downsampled 22kHz data only. In a first step, 13 Mel-Frequency-Cepstrum coefficients and their first and second order derivatives are computed every 10 milliseconds. With the initial (bootstrap) system, the training data is then viterbi aligned against its transcription. Based on this alignment, a decision tree is constructed for subphonetic HMMs by querying phone context. The data corresponding to a subphone (leaf) is clustered and consequently modeled by a mixture of Gaussians with diagonal covariance matrices.

Training data consists of both real telephony data (landline, cordless, cellular phone) and bandsampled 22kHz office correspondence data. The latter consists to a much higher degree of rich context sentences (read speech). In a first step, an acoustic model was built from downsampled data only. To improve the coverage of channel characteristics of the target domain, approximately the same amount of real telephony data was added to the training set. The telephony data set comprises utterances from various domains: digit strings, numbers, time, date, spellings, a limited amount of office correspondence and spontaneous utterances. Using additional downsampled data is mainly motivated by increasing the amount of rich context utterances in the training set. This has shown better results than training with real telephony data only. In particular, this is true for high quality ISDN data present in the test set.

The clustering process attempts to model each leaf with approximately the same number of mixture components. Each HMM state's output probability is modeled by approximately the same number of Gaussians. The models created with this procedure are recomputed by running two or four iterations of the forward-backward training algorithm, see e.g. [2].

The speech recognition engine used here is the IBM ViaVoice decoder and is described in more detail e.g. in [3]. Each speech frame, taken at 10 millisecond intervalls, is labeled and passed to the acoustic fast match which uses continuous density HMMs. For all words with a high fast match score (the fast match list) a language model (or grammar) score is computed based on the sequences of words decoded so far. This reduces the number of words for which in the next processing step the computationally more expensive so-called detailed match has to be computed. An heuristic search algorithm determines at each step which paths to expand. The best path covering all input data is chosen as decoder output.

With the scenario given in the introduction (all german motorways, all states and all major cities included) the vocabulary size amounts to 500 words. Initial tuning on a small test set (16 speakers and 320 utterances) resulted in a word error rate of 15.3 % and a success rate of 91.3 %.

4. BACKEND

The backend has to provide the information requested by the caller. This could be done using an enterprise database or by internet data servers. In ISA we implemented an information access to web pages containing traffic information in Germany. The content of these pages is very dynamic and may change every 10 minutes. The data is provided sorted for federal states or for motorways.

To overcome usual bandwidth limitations we implemented a cache. The data was passed to an HTML parser and then to a Java script which stored the data in certain text files. Additionally it generated a new indexing which covers the numbers of the motorways and the federal states to allow more dedicated user requests. The Java script also prepares the written data for spoken output like text normalization for abbreviations.

5. ANSWER GENERATION

There are two possibilities to present the requested information to the user in ISA. The first is spoken output which could be generated by the text-to-speech component, also a playback of recorded utterances is possible. The second way is the multimodal answering via SMS or other services.

5.1 Text-To-Speech Processing

The IVR script constructs a text message based on the cached content. General control flow messages are as well possible. To generate the spoken response, the IBM ViaVoice Outloud module is used. This is a formant-based synthesizer and allows dynamic modifications of prosodic features like phone duration etc.

5.2 SMS

From the system response an SMS message is generated. The text message is reduced to a template based on the most important information. If a message exceeds 160 characters it is spliced into two or more messages. To broadcast the message different approaches can be used. The message can be sent via an IP connection to the SMS gateway of a cellular network provider or via a connected modem to a service hotline number. Another possiblity is the use of a directly connected cellular phone for sending SMS messages.

5.3 Example Dialog

```
letzte Änderung: 15:33 Uhr,
es gibt eine Meldung.
[A7] Flensburg in Richtung Hannover
zwischen HH-Schnelsen-Nord und Elbtunnel
11 kilometer Stau.

last change: 3:33 PM,
there is one report.
[A7] Flensburg in direction Hannover
between HH-Schnelsen-Nord and the Elbtunnel
11 kilometers traffic jam.
```

Table 1: German traffic jam report used for ISA with English translation

Table 1 gives a German example of a traffic jam report text and the English translation of this report. In table 2 a German example dialog with ISA is given.

The caller has also the possibility to ask for help, if he does not know what he can say to ISA. As well the system provides the chance to repeat the last read message in case the user did not understand everything clearly.

```
S: IBM Stauinformationssystem –
   IBM Traffic Jam Information System

C: Gibt es Staus in Hamburg? –
   Are there any traffic jams in Hamburg?

S: letzte Änderung: 15:33 Uhr,
   es gibt eine Meldung.
   [A7] ...
   Worüber möchten sie als nächstes
   Staumeldungen hören? –
   last change: 3:33 PM
   there is one report.
   [A7] ...
   Which reports do you want to hear next?

C: Ich möchte Informationen über Staus in
   Baden-Württemberg auf der A5 bitte. –
   I want information over traffic jams in Baden-
   Wurttemberg on the A5 please.

S: Hier sind derzeit keine Staus gemeldet.–
        No traffic jams are reported here.
```

Table 2: Spoken dialog (German original and English translation, C = caller, S = system)

```
[A7] Flensburg -> Hannover zw. HH-Schnelsen-Nord
und Elbtunnel 11 km Stau
```

Table 3: SMS message

Table 3 shows the SMS message which will be generated in the example dialog. Figure 3 shows how this message will appear at the caller's mobile phone screen.

6. CONCLUSION AND FUTURE WORK

In this paper we have described ISA, which is a running prototype of a telephony application built with the IBM ViaVoice Telephony Toolkit. This application allows spoken access to data stored in the web or on any database server. In this application the voice operated communication flow is enhanced by a message broadcasting module (SMS generator) which generates textual output on cellular phones.

Future extensions of ISA will target towards the integration of WAP (Wireless Application Protocol) services. As soon as transmission rates in mobile phone data channels are high enough, the communication can be completely handled via data channel. Additionally a multilingual version of ISA is under construction to provide even foreign drivers on German motorways with helpful information.

Figure 3: Appearance of an SMS message at a mobile phone's screen

7. REFERENCES

[1] IBM Corp. *IBM ViaVoice Telephony Tools – programmer's reference*. August 1999.

[2] F. Jelinek. *Statistical Methods for Speech Recognition*. MIT Press, Cambridge MA, 1997.

[3] P. Gopalakrishnan, D. Nahamoo, L. Bahl, P. de Souza, and M. Picheny. *Context-dependent vector quantization for continuous speech recognition*. Proc. of the IEEE Int. Conference on Acoustics, Speech, and Signalprocessing, Minneapolis, 1993.