ARCHIVAL RESEARCH · ADITYA MOHAN

Voice Browsing Before Modern Smartphones

Research era: 2000–2002 (research and publication period) · Originally published on Blogger: 2012-10-31 · Archive recovered October 2026

> **2026 editorial context (not part of the original post).** Full technical text about constrained mobile devices, proxy-based voice search, accessibility, WAP, and diagrams.
>
> 2026 context: a documented antecedent of conversational information access on resource-constrained devices; avoid claiming authorship of later commercial assistants.

Original Blogger address: http://adiinsf.blogspot.com/p/voice-browsing-for-mobile-devices.html (live availability not established). Old report links below are carried forward but not guaranteed to work.

Original 2012 archive text

Some pdfs for publications

2000 ECUMN Voice Request/Response for Mobile
2000 Voice Browsing - Nortel Networks
2002 Voice Browsing - NCC
2002 Voice Browsing - NCC Presentation

Summary

We came up with a new technique to enable voice based information retrieval from the Internet on mobile devices, for example cell phones. In 2000, smart phones were not main stream and cell phones had limited processing and bandwidth capabilities. We believe that this was one of the first few researches in 2000 that focused on enabling cell phones with a system capable of acting on the basis of freely spoken voice and also allowing the response to be spoken instead of displayed. Such a system can assist people with visual disabilities to get access to information on the Internet via their mobile devices / cell phones that has limited capabilities. Mr. Mohan introduced the concept of voice templates that is generated through collective knowledge from many peers and aggregated to form a voice template library that is used for

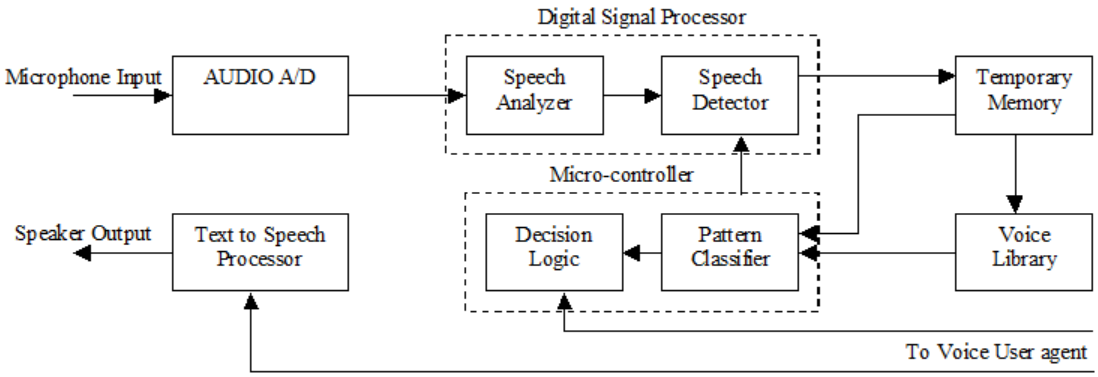
enabling the voice request and response.

Another important contribution of this research was that the algorithms and systems used can be used in countries where people are poor and can't afford high end smart phones – for example Apple iPhone with Siri.

IEEE International Conference on Trends in Communications, (EUROCON'2001 | Agenda) we proposed a system that makes it possible to use voice as a medium for communication between a mobile device and the wireless network for information retrieval from the Internet. A proxy agent approach is followed for wireless Internet access. The user is able to make his request by speaking simple sentences through the mobile device's microphone. The voice response, obtained from the proxy server, is played back using the existing device's loudspeakers. The paper tries to cover all the important components of the network that are responsible for generation of this response

Formants

Resonance Frequency that pass most acoustic energy
F1, F2, F3
F1 influenced by tongue body height
F2 influenced by tongue body frontness / backness

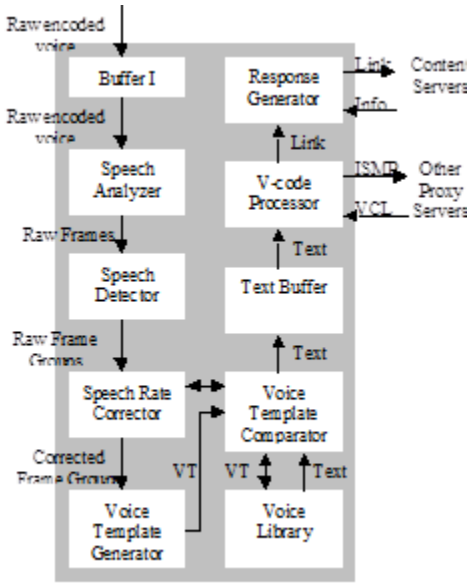


Architecture: Mobile Device

Major Processing Blocks

1. Speech Analyzer - Signal Processing Front End
2. Speech Detector
3. Voice Template Generation

4. Voice Analyzer
5. Voice Library Organization



Proxy Server Architecture

Voice User Agent

Forwards request either to the Proxy Server(in form of V-code) or to Web micro Browser
Forwards the request to Telephony related User Agent, if the voice request relates to Telephony related command

V-code To Link Mapping

Contains links and its corresponding V-code
Resides in both Proxy and Mobile Device
Link can be either a XML Link or a telephony related link
www.nortelnetworks.com
<telephony>://gsm/ch;1 for Call Hold
Unidirectional Graph Structure

Voice User Agent

Forwards request either to the Proxy Server(in form of V-code) or to Web micro Browser
Forwards the request to Telephony related User Agent, if the voice request relates to Telephony related command

Voice Response

- Text Summarizer
- Mapping in the Gateway is formed by the combined request from all the clients, using that Gateway
- Tags in the XML (or its equivalent) page to indicate, what has to be said on device

Here are some pieces from my paper on Voice Enabled Request & Response I presented at National Conference on Communication (January 2000) in IIT-Bombay, India.

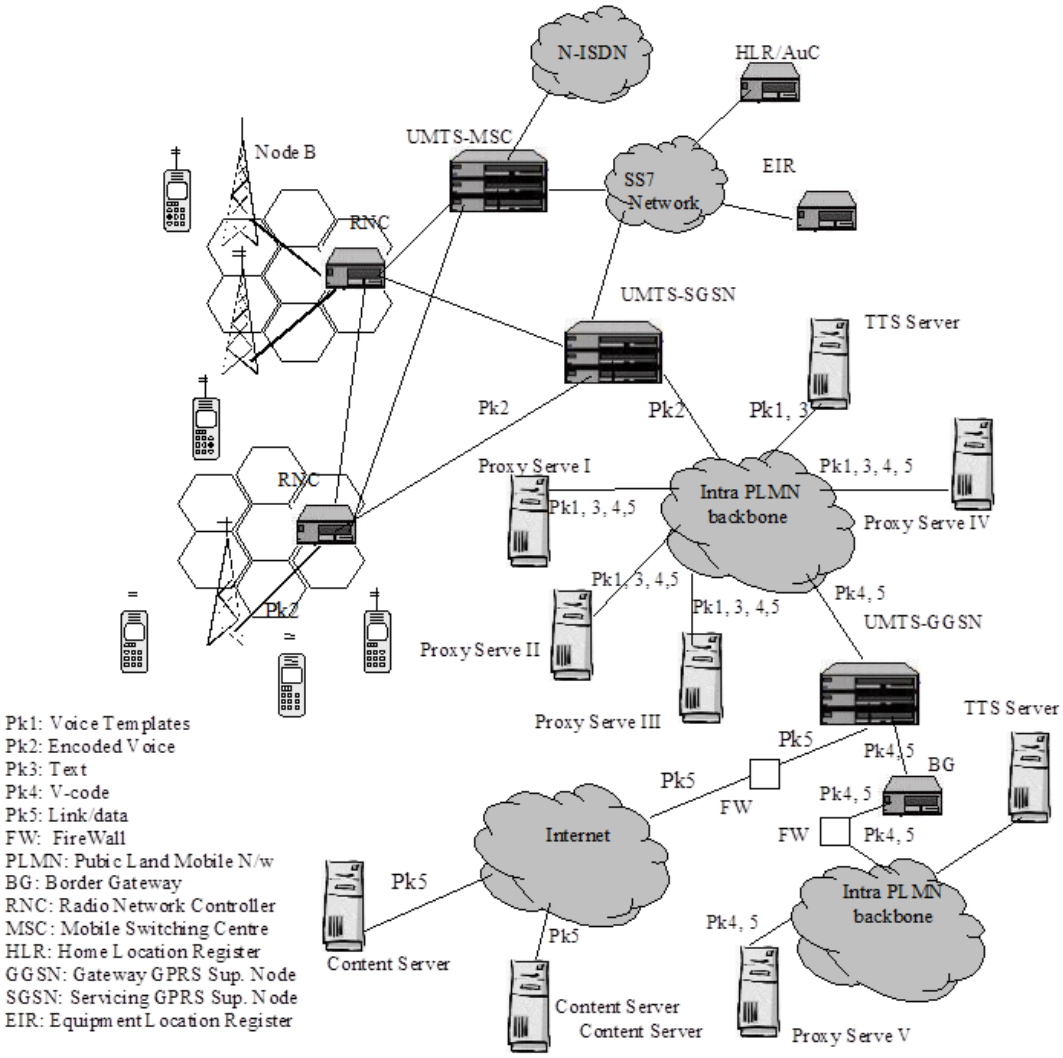
Abstract

The paper proposes a system that makes it possible to use voice as a medium for communication between a mobile device and the wireless network for retrieval of information from the Internet. A Proxy agent approach is being followed for wireless Internet access. The user is able to speak his request using simple sentences, through the mobile device's microphone. The voice response, obtained from the Proxy Server, is played back on the device's speakers.

Our Value Proposition

What distinguishes our system with the others is that service is quite different from the trivial Voice Enabled services available today. In the usual Voice Enabled services, a circuit switched connection is first established between the subscriber mobile equipment (which can be a cellular phone) and the servicing network. The user is now given some options in the form of voice that he can accept or use. The user can choose any of these options by either entering his request number through the keyboard of the phone via tone dials or can speak up the number himself. In both the cases the user may be expected to proceed to the next set of instruction if needed. This system is less preferable for some service scenarios like a businessman wanting to check the stock quote while driving, as he has to apply some sort of logic and attention for making the voice request. The existing Interactive Voice Response (IVR) system lacks the intelligence and hence each IVR system has its scope limited to narrow knowledge domain. A system used for railway ticket reservation is of little use for a service like weather updates over your phones.

System Overview



Epitome of the morphosis of the net and the mobile, a device that listens and speaks to you! Be it getting the current stock prices and news updates for the business executive, checking email for the net-savvy or hotel reservations, movie bookings, street directions, searching for information or making phone calls. Being interactive with the input from the user, it is able to learn day-to-day tasks (opening a web page, calling the doctor on the phone!). A user can command his mobile phone to answer his calls or dial a number by saying, “Andy please call Akshay” (Andy being the name of your phone). With a simple voice request “get my mails from yahoo and auto-answer all any new e-mails”, the system can open your email account and auto answer emails.

Figure above gives the network topology for an UMTS network with the additional components (TTS Server and the Proxy Servers) in place. The Proxy Servers and the TTS server form the part of the Core network of UMTS. A Proxy Server is located with each SGSN (Serving GPRS Support Node). Depending upon the region to which a particular wireless subscriber belongs, he is assigned a Home Proxy Server, which is responsible for receiving the subscriber’s voice request. It also coordinates the Voice response generation and finally sending the response back to the mobile device. Different Proxy Servers can communicate with each other via the IP backbone through the SGSN. To send a voice request, the MT initiates a connection with the radio switch, namely the RNC[1] and transfers the user’s voice request. It’s now the RNS’s job to forward the request to the Home Proxy Server via the SGSN. The Home Proxy Server is responsible for

processing the voice request and generating the response with the help of other proxy servers. This is described in details in the Proxy Server Architecture section. The communication between the different Proxy Servers happens through the IP backbone. The TTS server responsible for speech synthesis (if needed), is connected to the GGSN (Gateway GPRS Support Node). The TTS Server – GGSN interface is same as that of the Proxy – SGSN interface i.e. the Gn. As shown in the figure 1. The system is capable of learning the various voice requests initiated by the user and their corresponding action. This gives the system its intelligence. This information form the part of the V-code/Link mapping and the Voice Library as explained in detail in later sections of this paper. Hence each time the user issues a voice request, the system updates its database. So the next time another user gives a *similar**[2]*** request, the response will be generated immediately without any more user intervention. In theory, the V-code-Link mapping is a database formed by combined efforts by all subscribers that are using that Server as their Home Proxy. The response is generated by the Proxy Server from the data received from the various content servers on the Internet. For conversion of text to speech stream, voice templates (VT – see [1] for more details) are collected from the various proxy servers and TTS servers (if required). The VTs are collected and buffered by the Home Proxy, before being sent to the RNC.

Packetized Voice

One of the challenges in the design of this system is to transfer the voice request signal and data stream to the Proxy Server. More over there need to be exchange of Voice templates between the Proxy Servers, which makes it inappropriate to use the circuit switched network components. Another reason for using the packet service of UMTS instead of the circuit switch domain is the fact that by using the latter will restrict the number of users that can access the network. Also since a voice request is normally not more that few seconds, setting up a dedicated link will waste bandwidth. Hence it was decided to use the Packet service of the UMTS User plane for voice data transfer. The heart of the system, the Proxy Server is connected to the SGSN via the Gn interface with AAL2 and GTP[3]/UDP/IP at the transport level. AAL2 is the ATM Adaptation Layer specified in ITU-T Recommendation I.363.2 (1997). AAL2 provides the needed requirement of voice compression, silence detection / suppression, and idle channel removal. Although ATM networks face the problem of end-to-end delay (mainly due to packetization, transformation of the analog signal to digital, and buffering), the delay incurred is tolerable. As the more of operation in our system does not require full interactivity, hence Voice over ATM for transfer of Voice Request and Response between the MT and the network comes out to be excellent alternative.

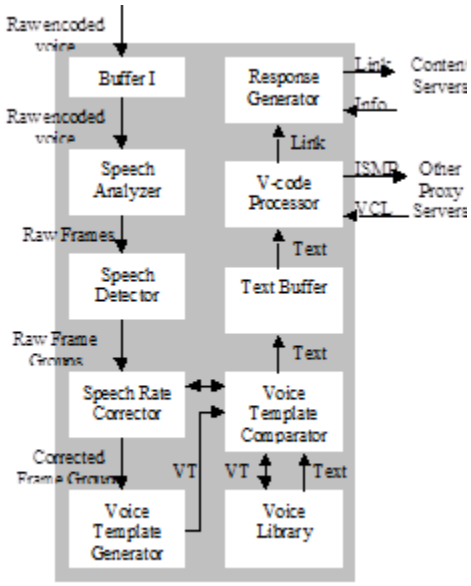
Voice Encoding and Decoding

Voice Compression and decompression is another factor that affects the performance of the system tremendously. The user request at the mobile device has to be first compressed (Encoded) before it can be sent to the network nodes. Again at the Proxy Server, this needs to be decompressed (Decoded) back to the original voice signal. Since these processes are computationally expensive and takes finite time, the encoding decoding technique has to be carefully selected. More this same method is used for generation of the VTs. Hence if compression is too high, then this will result in the Voice Template comparison to be even more processor and memory intensive. Although at the Proxy Server, processing power is not a bottleneck, but when we see the subscriber-base[4] that we are aiming, even the fastest processor can be rendered useless. It was decided to use a speech-specific analysis-by-synthesis model for speech conversion, which parameterizes speech and later reconstructs it using the source system model. The ACELP (Algebraic Code Excited Linear Predictive Coder) speech compression algorithm was used for this purpose. Typical voice codec capable of coding network quality speech at low bit rates were too high on complexity to provide a viable solution.

During coding, the system parameterization uses an initial open loop stage followed by an analysis-by-synthesis model for determination of the various speech parameters. The open-loop analysis consists of the linear prediction thus modeling the digital filter used for the synthesis of speech. The closed loop analysis-by-synthesis stage consists of calculation of the excitation source, pitch and the gain parameters by minimizing the error between synthesized and the input speech. In the decoding or the synthesis stage, these parameters are decoded and used in conjunction with a reconstruction mechanism to form speech. Also in has an algorithmic delay period of 20msec (coding and decoding) making it suitable for real-time implementation. It is suitable for applications such as voice over frame relay/ATM, teleconferencing or visual telephony, wireless telephony and voice logging thus ideal for the system under discussion. It can be implemented easily on the commonly used DSPs (Digital Signal Processor) available commercially.

The V-code Protocol

V-code represents the information about the voice request (after being transformed into appropriate text). All irrelevant information after being removed from the input text request forms a V-code sequence [3]. This is done by applying a set of rules to the input text. (Refer [3] for details) The V-code Protocol is used to exchange information across the three major components of the system, namely the mobile device, the proxy server, and the TTS server. The protocol basically runs on top of the UMTS stack at the Application layer level. The protocol defines the communication between Mobile Device and Home Proxy Server, Home Proxy Server and other Proxy Servers, TTS Sever and Proxy Server. The V-code Protocol also defines a communication mechanism for negotiating a set of parameters between the mobile device and the proxy server. These parameters pertain to the system functioning like enable/disable voice request/voice response. A user may not always want to use the ‘Voice Request and Response’ feature.



Proxy Servers

The system is based on a Proxy server model for accessing the Internet. The Voice request from the user is processed by any one of the Proxy Servers on the network (See Figure 1.for network topology). Depending upon the location to which a MT is registered with, it is assigned a Home Proxy Server. All voice requests from that MT are handled by its Home Proxy Server and are responsible for accepting the user's voice input and generating the voice response with the help of other proxy servers and the TTS server. It generates Voice Templates from the text input during voice response. When the user wishes to establish a voice session with one of the proxy servers, it first performs QOS and Parameter Negotiation.

For communication between a home proxy server and other proxy servers, the two important packets exchanged are the ISMP (Inter Server Mapping Packet) and the VCLP (V-code to Link packet). ISMP contains the v-code sequence that is finally obtained from a voice request. Each proxy server has a V-code/Link Mapping composed of v-codes and their related links [3]. A search for the v-code sequence in the mapping returns a link. The home proxy sends the ISMP to other proxy servers, which responds by sending the VCLP. A VCLP contains the returned link and the score (measure of match between the v-code sequence and the mapping). A home proxy can send the same ISMP to more than one proxy server. This ensures higher accuracy, as more options are available to the user.

ISMP (Inter Server Mapping Packets) are used by the Home Proxy to obtain a VT from a Proxy Server or the TTS Server. The response in the form of a VT is encapsulated in a Voice Template packet format before being sent back to the Home Proxy. This is useful when the Home Proxy needs to generate the Voice Response and the VTs some VTs are not present in its Voice Library. A Home Proxy can also ask a Proxy for help in transforming the VTs from the Voice Request to an appropriate text. This is done by sending a VTP packet to other Proxies, which respond by sending a V-code.

The Proxy Server is connected to the network via the SGSN with Gn Interface between the two. At the Proxy Server side the bearer is IP/AAL2/ATM. The V-code protocol is running over the connection oriented TCP and communicates pear to pear with Mobile Terminal (MT).

System Architecture

The encoded voice request received by the server is buffered to account for network delays and the decoded speech then goes for voice recognition. Voice recognition is performed using a linear predictive analysis approach followed by a vector-quantizer based two level decision logic that is capable of handling differences in the rates of speaking as well. Main functional blocks of the system are described below.

Distributed speech Recognition Engine

[Image missing from Blogger export: original authoring-file reference.]

The speech recognition consists of a LPC speech analyzer and a two level decision logic for pattern comparison and classification. The speech analyzer parameterizes the acoustic events in the digitized reconstructed signal, in terms of a compact and efficient set of speech parameters using linear predictive analysis. The basic entities on which speech analysis is performed are the frames. A frame consists of 160 speech samples (sampling rate being 8-KHz). Thus speech is discretized into frames of 30-msec separated by intervals of 10-msec. This analysis provides a set of spectral vectors (test pattern) that after end point detection by a hybrid approach are compared using the voice template comparator. The set of spectral vectors corresponding to a word form a voice template.

The recognition system follows a distributed approach and consists up of all the Proxy Servers located in the same Intra – PLMN backbone. Each server has its own Voice Library and V-code – Link Mapping. A Voice Request from a user, under the coordination of the Home Proxy Server is acted upon more than one Proxy.

Some important components of the system are explained below briefly.

Voice Library Organization

A vector quantization (VQ) representation is used for coding the spectral vectors of a speech signal. Thus the voice library consists of codebooks (one for each word) where each spectral vector is represented by one of the code words in the codebook. Thus it sores the voice templates corresponding to a word.

Voice Template Comparator

It performs the function of a pattern classifier and contains the decision logic to find the best match. The voice templates stored in the temporary memory are compared with those in the voice library to find the best match. It consists of the preprocessor and the postprocessor. The preprocessor is a vector-quantizer based speech recognizer that performs frame-by-frame vector–quantization and provides the possible alternatives. Based on the number of alternatives given by it and a threshold, it is decided to take the output as the recognized word, further processed or ask the user to input it manually using the keypad. The post processor consists of a dynamic time warping processor (Speech Rate Corrector). The best matching alternative is taken as recognized speech. The recognized text is then encoded and sent to the v-code processor

for the generation of the V-code.

V-code Processor

The text received from the voice template processor is resolved to a web-link by the v-code processor. It transforms the text to a v-code [3] that is resolved to a link by searching a V-code to Link Mapping.
V-code/link mapping

The V-code/Link mapping contains the relation between the V-code sequences and the links. The database for the mapping is organized as a Unidirectional Graph Structure. A link can either be a normal hypertext link or a set of rules/sequence of actions to be preformed (e.g. fetching the emails and sending an 'out of station response' to all new emails).

Response Generator

The response generator sends the link to the relevant content servers, and collects the response. It can be in the form of XML, HTML pages or other forms of data. This information is fed to the text summarizer, if needed. This condensed information is fed back to the voice template processor, which converts it to a voice response and finally the response is sent to the user.

TTS Server

The TTS Server is responsible for generating the VTs from the v-code. It receives request from proxy servers via ISVPs. The v-code extracted from the packet is converted to normal text and fed to the TTS Engine, which generates the VTs by encoding them using the Speech Encoder. These templates are sent back to the initial proxy after encapsulating them in VTPs. The mobile device uses the VTs for only playing back the response (after they have been decoded by voice decoder) and are not added to the voice library of the mobile device.

For speech synthesis at the TTS Server, one can use a concatenative speech synthesis system with dynamic unit selection using diaphones as the units for synthesis. As there is only one TTS server, a can also use a high quality speech synthesis system.

NOTES

[1] Radio Network Controller
[2] The voice request “Get me to Nortel Networks site” is *similar* to “Hey go to Nortel Networks” or even “Nortel Networks please”.
[3] GPRS Tunneling Protocol: The protocol between GPRS Support Nodes (GSNs) in the UMTS/GPRS backbone network. It includes both the GTP signaling (GTP-C) and data transfer (GTP-U) procedures
[4]According to recent Nortel Networks' studies, there are more than 200 million people using the Internet today. By 2004, some studies are predicting that nearly 1 billion people will be online across the globe.