V-Mail: Voice Email Communication over the Phone

Sheetal S. Sethi  
K S R Anjaneyulu

HP Labs India  
Hewlett-Packard India Software Operation Pvt. Ltd.  
Sterling Square  
No. 3, Madras Bank Road  
Bangalore - 560 001

sethisheetal@yahoo.co.in  
anji@india.hp.com

ABSTRACT

The main objective of this paper is to give an insight into the V-Mail project and how it tries to use speech recognition for providing a voice e-mail service to users. We first describe the motivation behind the project and the audience targeted to use the system. We then outline the applications for which the system could be put to use, and its architecture and design. We finally discuss the speech recognition issues, the current and the future plans for the project.

1. INTRODUCTION

One of the most widespread uses of computers is the purposes of communication. Electronic mail remains one of the most significant communication tools. However, in a developing nation like India, where more than half the population resides in rural areas, the use of e-mail is minimal. It is time to consider the needs of people who are relatively disconnected from their world by illiteracy, distance, or lack of access to inexpensive communications like e-mail. Also in a country like ours, which has close to 38 million landlines\(^1\) and 8 million cellular phones, compared to only 7 million people with access to the internet, it becomes imperative to provide services like e-mail over the phone in order to allow a larger number of people to benefit from the use of it. This application aims at doing just that. It would allow users to send voice messages to anyone who has an e-mail address or is registered with the service and has a V-Mail ID, and allow the user to receive and manage mails, which could be recorded messages or textual messages, received in response. A PC based prototype is currently being developed, however this application would be integrated with telephony in the future.

1.1. Purpose

It may be useful to start with an application scenario to demonstrate what we have in mind. This is outlined below:

The user dials up a number, logs in using his personal identification number (PIN), enters

\(^1\)The number of people who have ‘access’ to a phone is much larger given that there is a reasonably large percentage of the population that has access to a phone in a PCO booth for instance. Figures on landlines were obtained from the following source: [http://www.pcuf.fi/~tmo/telecom.htm](http://www.pcuf.fi/~tmo/telecom.htm)
a number indicating the action he wishes to carry out (say ‘send mail’), carries out that action and then logs out.

So essentially it is just like a voice mail service that runs in many organizations but the difference is that it would allow users to communicate with others who may or may not have an e-mail address.

The application aims to use voice as an input and output modality for allowing users to send e-mail and receive e-mail using voice. The users would need to record their message, and the recorded message’s file will be attached to the e-mail and sent to the recipient. The recipient can listen to the voice message using a ‘sound player’ on one’s computer system. The voice messages can be sent in the language of choice since the user just needs to record the message.

The main aims of the project are to:

- Test the usability of speech recognition for this application
- Define the type of interface that would be most appropriate for this application. The need for the simplest user interface that would be readily comprehensible and easy to master and extend is kept in mind.
- Prototype an application that can eventually be used over the phone without a computer.

The goal is to build an application that would allow a large number of users to benefit from e-mail as a means of communication.

1.2. Audience

The main target audience may be broadly described as those who fall into one or more of the following categories.

Those

- who find the keyboard the most intimidating feature of the computer.
- who would like to send messages in languages other than English
- who would like to send voice messages instead of text messages.
- who are physically challenged and therefore cannot use the keyboard.
- who cannot write.

However, the use of this application is not restricted to the audience mentioned above. It will prove useful to even those who have been using e-mail for communication but who for certain requirements may prefer using voice.

1.3. Advantages of Voice Email

Voice e-mail has some advantages over textual e-mail. Some of these advantages are:

- Use of voice e-mail allows messages to be multilingual in nature, which makes more sense in the Indian context. Traditional use of e-mail has the disadvantage that it is mainly used for English and uses mainly text messages.
- Talking is easier and faster than typing.
- Voice eliminates the need to use a keyboard thus making e-mail accessible to a larger part of the population
- Voice has a personal touch. Voice gives us the ability to convey humour, laughter, and other expressions better than emoticons².

² Emoticons: ASCII glyphs designed to show an emotional state in plain text messages. E.g. ☺ (happy) ☹ (sad) ;-) (winking)
This makes e-mail more "real," more personal, and even friendlier.

- Voice eliminates the need for editing, the need for spelling checks, and the need for grammar or punctuation checking.

1.4. Related Applications

There have been other attempts at similar types of applications. However, our unique angle is to try to create a service that can be used over the phone and use speech recognition for the e-mail addresses. Some attempts are listed below:

- Rediffmail provides the facility “Send Voice” of recording one’s voice messages and sending it as an attachment with the e-mail and thus allowing users to send e-mails in Indian languages. This service has been widely accepted by people, and if it’s popularity is anything to go by, then we have good reasons to believe that our application will be a success [3]. This service however needs to be used from a computer.
- Navin Communications, Inc provides “TeliVoice” [4]. This allows e-mail to be sent from one’s phone but uses manual operators to take the messages and send them.

2. POSSIBLE APPLICATIONS

Once this application is integrated with telephony it could play a major role in enhancing communication. Some of the possible applications are:

- allow easy communication with family and friends. The only other alternative for people in the rural areas to stay in touch with their

loved ones is through STD/ISD. Such an application would provide a cheaper means of communication.

- allow easier access to information to people in rural India. This could be achieved by subscriptions to news items or other information that can be delivered through e-mail and then be read out to the user over the telephone. For instance, market prices of products relevant to farmers.

3. ARCHITECTURE/DESIGN

3.1. Application Functionalities

The application can handle the following functionalities:

- Send mail: The sender starts the application, logs in, reads out the e-mail address, records his message and sends the e-mail.
- Receive mail: The receiver starts the application, logs in, asks for new e-mails, listens to his e-mails in the order of their arrival.
- Change PIN: The user starts the application, logs in, enters the old PIN, a new PIN and then reconfirms the new PIN and leaves the application.

3.2. Application Flow

The following flowchart shows the flow of the “Send Mail” functionality in the application. The Receive Mail and Change

---

3 Assuming they have access to e-mail or they are V-mail users

4 STD/ISD, which stand for long distance and international dialing, however do not allow asynchronous communication that may be required for instance if the recipient is either not present or does not have a dedicated phone to use.

5 In the telephony version, dialing a number would start the application
PIN functionalities would have similar flows.

![Flow Diagram](image)

**Figure 2.** Flow Diagram for Send Mail functionality

### 3.3. Technologies being used

The application is being developed using:

- Visual Studio VC++ Version 6.0
- Microsoft Speech API (SAPI 5.0) for Speech Recognition
- Microsoft Text To Speech (TTS)
- Messaging API (MAPI)

The Speech API has been used to interface our application to the Microsoft Speech Engine. The TTS is used to play back messages that are received if they are text messages\(^6\) else an audio player is used (if they are voice files). MAPI is used to integrate the application created to the backend mail server to send out and receive messages.

### 3.4. User Interface Design Issues

Convenience and usability have been the core issues while designing the interface. In addition we have kept in mind that finally the application would need to run without a visual interface, since the user would be using a phone in place of a PC.

For the current PC based prototype, a suitable and simple Graphical User Interface has been designed for a user to:

- record, hear and send messages
- receive and playback messages
- delete messages
- change his/her Personal Identification Number (PIN).

Proper prompts have been incorporated to make the application hands-free\(^7\) and eyes-free. The user would be prompted to respond in case there is no response from him for a long time.

It also allows users without requiring a good knowledge of the English language to send e-mails in their own language. The user would however need to be able to read out the e-mail address in English. The prompts provided by the application can be in any language based on the need.

---

\(^6\) Currently we support English messages, however messages in Indian languages can be supported with appropriate TTS engines.

\(^7\) The application would require minimal interaction with the keypad/telephone dialpad.
4. SPEECH RECOGNITION ISSUES

A speaker dependent engine has been chosen for the application, where the user would have to provide training to the engine. The choice was made considering the fact that the user will use the V-Mail service on a continuous basis. Also given that the service requires user authentication, it is possible to identify the appropriate speaker-model. Speaker dependent systems also achieve better recognition performance than speaker independent systems.

Speaker dependent engines have the disadvantage of requiring the user to undergo an enrollment/training process, which is not the case with speaker independent engines. Also the requirements of model loading for every user affect the speed of an application. However, such individualized training is likely to have higher levels of accuracy in recognition.

If the word models of speaker dependent systems are tailored to suit the needs of the application, then we could achieve a high level of accuracy even in environments where there is some noise. Further, the enrollment process enhances the overall accuracy by acclimatizing users to the recognizer and the application. Accuracy is also improved when the enrollment process is done in the environment in which the application will be used.

The application allows the user to give a word in place of the individual alphabet, in case the engine does not recognize the individual alphabet. For instance the user could say “apple” for “a”; “ball” for “b” and so on. The words chosen for the purpose are phonetically different to allow better recognition of these words.

Our experience with Microsoft’s speech engine has been satisfactory so far. It allows a fair deal of flexibility. The recognition has been found to be accurate to a great extent. The engine claims to achieve 95 percent accuracy after sufficient amount of user training. We undertook approximately 5 minutes of application specific training, and with that training along with the provision of using words for characters, as mentioned in the previous paragraph, we were able to get almost 85% of the characters recognized accurately on the very first attempt. Close to 15% of the characters were recognized accurately on the second attempt, while on some few occasions another attempt was required for accurate recognition. We would subsequently need to check accuracy with telephone quality speech.

We also need to consider the trade off between higher levels of accuracy and faster speed of recognition.

Finally, speech recognition is highly sensitive to the cooperation of the system users, unlike many other modes of input. So it becomes important to consider user-acceptance factors like perceived benefit of the system, fear of the computer/technology, etc. to enhance acceptability of the system.

5. CURRENT PLANS

The application in its current stage of development allows the user to send voice mail. The user needs to dictate the e-mail address character-by-character, record the message, and send the mail. A simple user interface has been chosen for the application.

The User Interface designed allows the user to change the last character of the e-mail address spoken and identified incorrectly by the speech engine or completely reset it for reentry. Each character spoken out is spoken back to the user for his confirmation. For the purpose of repeating the character the Microsoft TTS service has been used and the character is spoken using the default voice object. In the future, we may also use

---

8 The list of characters includes alphabets, digits, and few special characters like ‘@’, ‘_’, ‘.’
prerecorded confirmation messages where required to improve intelligibility.

The UI design is such that it restricts the use of the keyboard to the set of keys common with the telephone. This has been done keeping in mind the long-term plan of interfacing the application with telephony.

The user can at any point of time abort a process by pressing ‘*’. Pressing ‘0’ on the numeric pad invokes the help for the respective service the user is using.

5.1. Limitations

The application has certain limitations in terms of the services that it offers:

- The application does not allow for attachments to be sent. The user is only allowed to record his/her message, which goes as an attachment to the e-mail. Since the medium is voice the types of attachments, which would be useful, would be other voice messages. For instance the user may want to forward a voice message to someone else adding his/her own comments to it.
- The application does not allow the user to sort his mails based on fields like subject, date or from fields. This issue needs to be explored. Ideally there should be a mechanism to ‘manage’ mails just like in typical mail clients. However, the functionality need not be exactly analogous.
- There is a limit on the duration of the recorded message that can be sent. The user would be prompted when the limit is reached.

5.2. Issues to be handled

The following issues will be addressed in terms of the current prototype that is being created:

- Carry out usability studies. We hope to test the effectiveness of the application with several audiences. What needs to be estimated is the percentage of users who would accept the service and those who would be able to use it effectively. The dropout\(^9\) percentage should also be low in order to call the project a success. Using these studies, we would need to obtain an understanding of the minimum sophistication level of the users who will be able to operate the system.
- Provide for creation of aliases for e-mail addresses that have previously been given, whereby the user could simply speak out the name instead of speaking out the e-mail address character by character.
- Improve speed of e-mail address recognition. Speed of a recognizer’s response becomes significant when recognition must be done in real-time.
- One will need to look at the key factors that result in high error rates. Some factors are:
  - Quality of the microphone.
  - User training: This has relevance in case of a speaker dependent speech engine. We will investigate whether more task specific training would help improve accuracy.

---

\(^9\) Dropouts are those who start an action and do not complete it.
6. FUTURE PLANS

The following is proposed as a follow-up to this project:

- Converting the application to a service whereby more than one user can interact with it at the same time.
- Providing a telephony interface. We have to look at how this service can be provided over the telephone, wherein the caller would just need to hook up to a voice service platform, log in and send his/her mails as recorded messages and listen to the received mails.
- One of the key issues that would need to be considered when interfacing with telephony would be noise elimination. All handset models contribute their own distortion and additive noise. Various noise reduction techniques would have to be applied for the service.
- The application should allow multiple recorded files to be sent to the recipient (instead of only one at present).
- Allowing e-mail addresses to be fed offline might be useful to allow those who have a pre-existing address book to be able to use that without having to dictate e-mail addresses to start with.
- Allowing multiple recipients for the same message.
- Audio compression issues may need to be looked at. However, sending compressed audio files would require the recipient of the mail to allow decompression of the same before listening.

7. CONCLUSION

This application has been designed keeping in mind the needs of a large number of users who have no access to the Internet or who may not be comfortable using it. Electronic Mail services have been a success, mostly with the web-savvy people. Now what needs to be seen is the acceptability and usability of a service like that proposed here for those who may want to access it over the phone.

8. REFERENCES

[2] MSDN Online documentation, Microsoft