An Advanced Integrated Architecture for Wireless Voicemail Data Retrieval

Konstantinos Koumpis[†] Charalampos Ladas[‡] Steve Renals[†]

[†]Dept. of Computer Science
University of Sheffield
Sheffield, S1 4DP, UK
{k.koumpis,s.renals}@dcs.shef.ac.uk

[‡]Dept. of Electronic and Electrical Eng.
University of Sheffield
Sheffield, S1 3JD, UK
c.ladas@shef.ac.uk

Abstract

This paper describes an alternative architecture for voicemail data retrieval on the move. It is comprised of three distinct components: a speech recognizer, a text summarizer and a WAP Push Service initiator, enabling mobile users to receive text summaries of their voicemail in real-time without an explicit request. Our approach overcomes the cost and usability limitations of the conventional voicemail retrieval paradigm which requires a connection establishment in order to listen to spoken messages. We report performance results on all different components of the system that has been trained and tested on a database containing 1843 North American English messages as well as on the duration of the corresponding data path. The proposed architecture can be further customized to meet the requirements of a complete voicemail value-added service.

Keywords: voicemail data retrieval, automatic speech recognition, text summarization, Wireless Application Protocol, Short Message Service.

1. Introduction

There is a growing interest in mobile communication systems that allow users to use their voices to do more than simply talk to other users. Speech recognition can provide access to several types of content based applications and services through a number of portable solutions, including mobile phones and personal digital assistants. Voicemail is a common part of any office solution today. Sophisticated personal voicemail systems can be built using low cost computers and modern voiceband modems. These systems can provide many, if not all, of the features of large centralized voicemail systems found in large organizations. The personal voicemail system can record and store voice messages digitally while the user is away or simply unavailable and can be reviewed when the user returns. Additionally, the user can call in on a touch tone phone and review stored messages.

Although, several advances in voicemail retrieval scheme have been proposed (e.g. [4], [11]), the limitations of the old paradigm still exist: users of voicemail systems on the receipt of a notification have to call their answering machine and download/listen to their actual/compressed messages. In order to overcome these limitations, the following interdependent issues have to be resolved:

- how to access the vast volume of background knowledge that is needed to interpret a random simple spoken message
- how to make it instantly and securely available to the recipient

In an attempt to provide answers to the above, we integrated our system that transcribes and summarizes voicemail messages [7] with a Wireless Application Protocol (WAP) Push Service initiator designed for this application. The motivation lies on the fact that such systems can reduce delays to important decision-making as they are suitable for capturing and distributing information quickly, no matter one's location and without human intervention. This makes voicemail retrieval on the move much more efficient and cost effective by proactively delivering text voicemail summaries to mobile terminals. The notion of these services is that the content is delivered to the mobile terminal directly from a source without an explicit user request. This is also beyond the conventional request/response model found in most WAP applications (e.g. [9]) and the WWW, where a user enters a URL (the request) which is sent to a server, and the server answers by sending a web page (the response) to the user. Additionally, the proposed system offers uninterrupted information flow in noisy places (crowded streets, train stations, airports) or in so called 'mobile phone free' environments (conference/meeting rooms), better message management (visual listing and indexing of messages) and lower cost of receiving calls while roaming abroad.

The rest of the paper is organized as follows: in section 2 a brief description of the WAP and its applicability to the voicemail retrieval task is given, while in section 3 we describe the voicemail data and the experimental setup of the speech recognizer. Details about the restrictions in message length as well as the need to use summarization are outlined in section 4. The text analysis for the summarization purposes is presented in section 5. Sections 6 and 7 describe the data conversion required by WAP as well as the testbed used to estimate submission delays, while the paper is concluded in section 8.

2. Voicemail retrieval using WAP

WAP is an application environment and set of communication protocols for wireless terminals designed to overcome resource limitation issues such as low processing capabilities, high bit error rate, limited memory and operation over a variety of bearer networks. It comprises a five-layer model, consisting of transport, security, transaction, session and application layers [16], giving greater flexibility to the already existing Internet model. WAP does not force applications to use the entire protocol stack. If an application, for example, does not require security but reliable data transmission, it can use a service of the transport layer without using the security layer.

None of the existing GSM bearers for WAP – the Short Message Service (SMS), Unstructured Supplementary Services Data (USSD) and Circuit Switched Data (CSD) – are optimized for this protocol. As a matter of fact, push services already exist in mobile phone networks, using the well-know SMS and the Cell-Broadcast mechanism in GSM networks. Many people already use SMS regularly, for information on key news items, or notification that a new message has arrived in their voicemail box. However, there is a very important element missing in the above: the service is *not* interactive. WAP provides the required interactive behavior through specifications such as the Wireless Telephony Application (WTA), as outlined in section 6.

In this work we have selected SMS instead of CSD as an adequate bearer for WAP. SMS has several unique features that can be summarized as message storage if the recipient is not available, confirmation of delivery to the sender and simultaneous transmission with voice, data and fax services. SMS is available on GSM, CDMA, North American IS-136 and Japanese PVC among other networks. Although the underlying protocols may vary [12] all of them define a mechanism for sending and receiving short text messages. In this work we have focused on SMS within the GSM, in which the maximum length of each Short Messages (SM) is 140 octets, sufficient coding for 160 7-bit ASCII characters. This implies that for a typical voicemail message only a couple of SMs may in fact have to be sent.

On the contrary, CSD is inappropriate as a bearer network for Push Services because a connection establishment requires the mobile client acting as a remote access server, a feature not supported by mobile terminals in order to keep the client's processing load low. But even if mobile terminals had this feature, CSD lacks immediacy in the sense that it requires a certain amount of time to establish the connection and reserves the voice channel prohibiting voice calls to be placed at the same time.

The architecture of the proposed system is shown in Figure 1 and encompasses three distinct phases of processing:

1. transcription of voicemail messages
2. construction of transcription summaries by selecting important terms according to their weights
3. formation of summaries and delivery via the WAP Push Service

The spoken messages collected by the voicemail system are forwarded to the Content Server where they are automatically transcribed and summarized. There is clearly no restriction on where the voicemail system is located and will most likely not to be located anywhere geographically close to the Content Server, allowing access to answering services other than the one provided by the network operator. The Push Initiator contacts the Push Proxy gateway over the Internet and delivers the messages. The Push Proxy gateway examines the message and performs the required encoding and transformation of the WAP domain. The messages are then transmitted hop-by-hop in the mobile network to the mobile client. The Push Initiator is then notified by the gateway about the final outcome of the Push operation.

3. Transcription of voicemail messages

The system presented in this paper is trained using the 14.6 hours of speech contained in the Voicemail Corpus Part I that is distributed by LDC[1] and we refer to this set as Vmail15. This corpus was collected from volunteers at various IBM sites in the United States, comprising 1801 messages in the training set and 42 messages in the development test set. In our implementation the first 1601 messages are used as training set and the remaining 200 as validation set. The vocabulary contains 10K entries and the out-of-vocabulary (OOV) rate for the test set was 7.3%.

Voicemail represents telephone bandwidth spontaneous speech and is characterized by a variety of speaking rates, accents, acoustic conditions and a variety of tasks [10]. Additionally, the natural unit seems to be the phrase rather than the sentence, and phenomena such as disfluencies, restarts, repetitions and broken words are common. Another feature

[1] Linguistic Data Consortium, http://www.ldc.upenn.edu

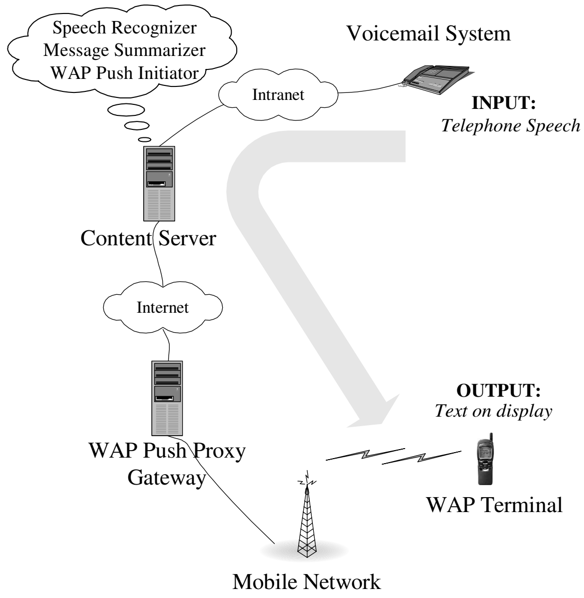

Figure 1. The architecture of the voicemail text summaries delivery system using the WAP Push framework.

of this corpus is that speakers do not get any direct feedback when they leave messages. This leads to many questions and instructions, which are absent from read or conversational speech. As the caller leaving a voicemail message could be calling from any location on any type of phone (wired, wireless, cellular) telephone channel data also poses problems of low bandwidth and low signal-to-noise ratio (SNR), while some degradation is due to the file compression method used by voicemail systems.

In order to estimate a posteriori context-independent phone class probabilities we use a standard feed-forward multi-layer perceptron (MLP) [1]. Such approach is able to model long-term acoustic context without strong assumptions on the distribution of the observations. A nine frame window centered on the frame of interest was used as an input to our three-layer MLP networks that have a single sigmoidal hidden layer of 2,000 units and an output of 54 phoneme classes. The outputs were generated by a softmax function from the weighted hidden units.

The underlying statistical model was an extremely simple hidden Markov model (HMM). For each of the 54 phonetic classes, we had a HMM consisting of strictly left-to-right model with multiple states tied to a single distribution; multiple repeated states were used to establish a minimum duration of each phone. Transition probabilities were set to 0.5. The emission probabilities of the HMM were scaled likelihoods estimated by dividing the network outputs by the priors of each class.

System configuration	***WER%***
(1) BN Acoustics, bigram	67.0
(2) VMail15 Acoustics, bigram	56.1
(3) Combination of (1) and (2), bigram	55.2
(4) Embedded training of (3), bigram	54.2
(5) PLP alone, trigram	51.2
(6) MSG alone, trigram	51.8
(7) PLP+MSG, trigram	46.8

Table 1. Recognition performance of the hybrid connectionist system on the VMail15 task.

The speech signal is segmented into 32 ms frames with a 16 ms frame step. For the feature extraction we created 12th order Perceptual Linear Prediction (PLP) [2] files plus log energy (13 features in total). Feature vectors for a given utterance were normalized according to the mean and variance of each utterance in the training data. In order to produce the labels for our initial system we passed Voicemail data through a network trained on BN speech (bandlimited to 8 KHz) with a word error rate (WER) of 36% on that task. That system gave a 67.0% WER for VMail15 as shown in Table 1. We then trained the MLP network with the acoustics of Vmail15 and tested it using a bigram language model to get a WER of 56.1%. By combining probability streams framewise in the log domain the WER came down to 55.2%. The next step was to perform an embedded training of that system and the WER then was 54.2%. After replacing the bigram language model with a trigram trained on the Vmail15 data the WER was reduced to 51.2%.

We also employed the Modulation-Filtered Spectrogram (MSG) feature extraction technique, which is a 28 elements wide robust representation in adverse acoustic conditions and is based on two signal processing strategies modeled after human speech perception [5]. gain control. Although MSG features were not as good as PLP features (a degradation of 0.6% absolute was observed), the WER was significantly improved by combining the PLP and MSG systems. MSG features offered a significant benefit for messages that were degraded in some manner, and were therefore used as the basis of our subsequent work further details of which can be found in [7]. From these experiments it has been demonstrated that competitive recognition performance can be achieved using fewer parameters than those used in mixture of Gaussian systems (e.g. [10]).

4. Message length restrictions

SMS as a bearer introduces an essential limitation on the amount of characters that each message can deliver. The 140 octets payload per SM is reduced further when considering the headers required by the WAP protocols. Wireless Datagram Protocol (WDP) serves as a transport protocol and offers segmentation and reassembly when it is used over SMS [20]. Each SM is then required to have an 11 octet long WDP header containing information on how to reconstruct the messages at the receiver end. The Wireless Session Protocol (WSP) headers and the Push Over-the-Air (OTA) headers are necessary for effective message delivery. These headers are compacted according to their contents as specified in [19] and [21], respectively.

In all the experiments reported herein the compacted header sizes of both the session and Push protocols were 26 octets long. The 44 octets long WML headers and tags are encoded using the WAP Binary XML (WBXML) [23] which performs a lossless compression algorithm. In particular, rather than apply HTTP's zip/deflate compression options in WSP, WBXML tokenizes normal XML by preparsing it into a tree structure, extracting common text stings, and transmitting it according to a compression state machine at both ends. A description of the above procedure follows:

1. embed summary into a Wireless Markup Language (WML)[2] template
2. compact page into WBXML format
3. add Push OTA header
4. add Push WSP header
5. fragment message into 129 octet chunks
6. add WDP headers to each message fragment
7. add SMS header to each message

The total amount of overhead produced for each transmitted message is 81 octets. If only one SM was to be sent per voicemail message, the summary would have to be up to 59 characters. This was considered insufficient and we decided to send two SMs for each voicemail summary instead. Therefore, in the current implementation the available length to fit the actual message information is:

$$\begin{aligned} L_{Summary} &\leq 2 \times (L_{SMS} - L_{WDP}) - [(L_{WSP} + L_{Push\ OTA}) + L_{WML}] \\ &\leq 2(140-11) - (26+44) = 188 \end{aligned} \tag{1}$$

[2]WML is a tag-based browsing language that inherits technology from HDML and HTML supporting screen management (text, images), data input (text, selection lists) and navigation support through hyperlinks. The definition of WML has only 37 markup tags, as opposed to 90+ of HTML due to the limited capabilities of current mobile terminals.

where L represents the length of each data field. This corresponds to a level of compression – defined as the ratio of summary length to source length – around 30%–70%, as the transcribed messages are typically 300–500 characters long.

5. Summarization of voicemail messages

An effective summarizer distills the most important information from a source to produce an abridged version of the original message for a particular user. In theory, we need to generate message summaries that use only information that is not in error. Since we can never guarantee that we can perfectly identify the words in decoded audio that are not in error, we cannot depend on sentence-level parsing because even a single incorrect word can completely garble syntactic structure. Instead, we need to choose information in another way, one that does not depend on syntactic or semantic analysis. It has been demonstrated in [15] that a combination of confidence measures with simple information retrieval (IR) and IE techniques can be used to accept/reject words and/or phrases for inclusion in summaries.

In our work the task of message summarization maybe cast as a linear combination of statistical and prior knowledge information, where the decision to accept or to reject words or phrases in a summary is based upon speech recognition confidence measures, term collection frequency and named entity (NE) lists. The idea behind term weighting is selectivity: what makes a term a good one is whether it can express correct information content from the original message. Each message is represented as a vector of weighted terms. The computation of the weights reflects empirical observations of the data. We did not remove stop words (e.g. prepositions and conjunctions) via a stop word list as they normally contain important information. The algorithm shown in Figure 2 removes the words with the lowest score till the limitations of message length as described by (1) are met. A description of the three main weight factors accompanied with a lossless summarization technique follows:

Confidence measures quantify how well a model matches some spoken utterance, where the values must be comparable across utterances. It has been demonstrated in [24] that hybrid connectionist models are well suited to producing computationally efficient acoustic confidence measures. The most discriminating measures are achieved by normalizing all phones in a word by the duration of the entire hypothesized word, rather than normalizing each phone constituent by its own duration. For a phone q_k which a hypothesized start time n_s end time n_e given some acoustic data x^n the

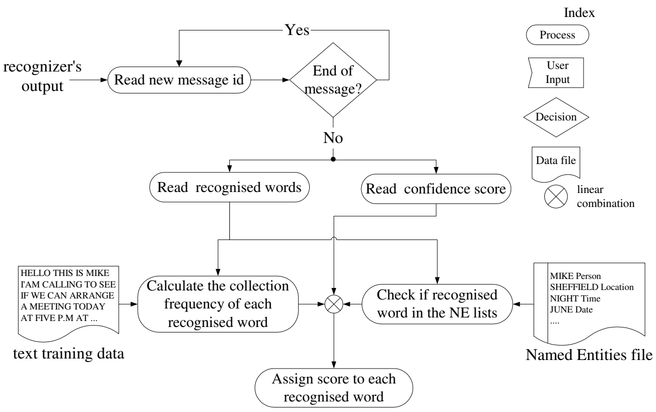

Figure 2. A schematic outline of the mechanism that assigns scores to each recognized word reflecting empirical observations regarding a combination of statistics and prior knowledge.

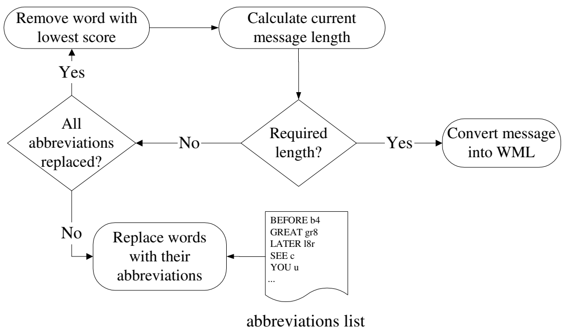

Figure 3. A schematic outline of the mechanism that removes the words with the lowest score till the message length restrictions are met. A list of common abbreviations is also used to reduce the length without losing information.

confidence measure is:

$$CMW_{npost}(q_k) = \frac{1}{n_e - n_s} \sum_{n=n_s}^{n_e} \log\left(p(q_k|x^n)\right) \qquad (2)$$

These confidence measures are computed directly by CHRONOS decoder [13] where the language model is used to constrain the search for the optimal state sequence but is not used in the computation of the confidence estimates.

Collection Frequency is inspired from IR and is based on the fact that terms that occur only in a few messages are often more likely to be relevant to the topic of that message than ones that occur in many. For a term t_i the collection frequency is defined as:

$$CFW_{t_i} = log \frac{N}{n_{t_i}} \qquad (3)$$

where N is the number of messages in the training data and n_{t_i} is the number of messages term t_i occurs in. The CFW_{t_i} weights are then normalized to the number of messages.

Named entity lists are used in order to prioritize words that may be classified as proper names, or as certain other classes such as organization names, dates, times and monetary expressions. This is less straightforward than identifying NE in written text, since speech recognition output is missing features that may be exploited by "hard-wired" grammar rules or by attachment to vocabulary items, such as punctuation, capitalization and numeric characters. Our NE lists constitute of 3.4K entries, 2.8K of which derived from BN data [14] and the remaining were classified manually from the Vmail15 transcriptions. This allowed us to retain in the summaries certain types of the terms containing important information, i.e. series of digits comprising telephone numbers.

Shorter synonyms and digits is another way to reduce the length of the message without losing information and is based on the fact that users of SMS are familiar with text messaging abbreviations[3]. As an example the phrase "SPEAK TO YOU LATER" can be replaced with "spk 2 u l8r". In our system we have been using a set of approximately 40 such abbreviations that offer a reduction of 10% in the message length. We also replaced full words describing numbers with the respective digits, i.e. the word "FOURTEEN" was replaced by "14" in the summaries offering a further reduction in length and increasing message readability.

Summaries are inherently hard to evaluate because the quality of a summary depends both on the use for which it is intended and on a number of other human factors, such as how readable an individual finds a summary or what information an individual thinks should be included in a summary. The recently introduced Slot Error Rate (SER) [8], which is analogous to the WER was used to evaluate how the summaries reflect the information content of the original messages. SER is equal to the sum of the three types of errors – substitutions S, deletions D and Insertions I – di-

[3] These abbreviations are written with small letters to show the difference from the human transcription and the decoder's output.

vided by the total number of slots in the reference (where C is the number of the correctly recognized words):

$$SER_{msg} = \frac{S+D+I}{C+S+D} \quad (4)$$

We define a slot to be any term or group of terms containing key and essential information for the message recipient. The baseline SER for the 42 messages test set was 40.3%. For a 40% compression level the SER of the summaries was 51.7%, while for a 20% compression (corresponding to two SMs or one WML card per voicemail message) the SER raised to only 55.9%. This indicates that the SER is mainly due to the transcription errors and the 7.3% OOV rate, rather than the summarization approach. These results are in consistence with the output of voicemail summaries evaluation survey reported in [7]. From these initial experiments is shown that the statistical model in combination with prior knowledge sources is an effective and robust approach to message summarization. We are currently investigating whether prosodic features such as pitch, energy, duration and pauses can add useful information for the summarization process.

Human transcribed message
HI MARY IT'S MARY ANN SHAW I JUST HAVE A QUICK QUESTION FOR YOU THE DEFENSIVE DRIVING COURSE THAT IS TOMORROW AND THURSDAY CAN YOU LET ME KNOW IF THAT'S IN THE HAWTHORNE ONE OR HAWTHORNE TWO JUST WANTED TO MAKE SURE I'M NOT SURE THE WAY THEY YOU KNOW SET UP THEIR ROOMS SO IF YOU COULD GIVE ME A CALL I'M ON TIELINE EIGHT TWO SIX SIXTEEN OH TWO BYE
Automatically transcribed message
HI GARY IT'S MARY INITIAL I SHOULD HAVE A QUICK QUESTION FOR YOU DID DEFENSIVE DRIVING COURSE IT IS TOMORROW AND THURSDAY YOU LET ME KNOW THANKS IN THE HAWTHORNE WONDERFUL POINT TWO JUST WANTED TO MAKE SURE MUCH SURE THE WAY YOU KNOW SET UP THERE SO IF YOU COULD GIVE ME A CALL ON TIELINE EIGHT TWO SIX SIXTEEN OH TWO BYE
Automatically summarized message
MARY I A QUICK QUESTION 4 u DRIVING COURSE IT IS 2moro AND THURSDAY u LET ME KNOW IN 2 WANTED 2 MAKE SURE SURE THE WAY u KNOW SET THERE SO u GIVE ME A CALL ON TIELINE 8 2 6 16 0 2 BYE
Summarized message in WML format

```
<?xml version="1.0"?>
<!DOCTYPE wml PUBLIC "-//WAPFORUM//DTD WML 1.1//EN"
"http://www.wapforum.org/DTD/wml_1.1.xml">
<wml><card id="MainCard" title="Voicemail Notification">
<p> <b>Caller:</b> 07979123456 <br/>
<b>Date and Time:</b> Fri 14 Apr 2000 - 14:56</p>
<p align="center"> <b>Message</b><br/>
MARY I A QUICK QUESTION 4 u DRIVING COURSE IT IS 2moro AND
THURSDAY u LET ME KNOW IN 2 WANTED 2 MAKE SURE SURE THE
WAY u KNOW SET THERE SO u GIVE ME A CALL ON TIELINE 8 2 6 16
0 2 BYE
<a href="wtai://wp/mc;01141234567,4557">Listen to message</a>
</p></card></wml>
```

Table 2. Human transcription, automatic transcription, summary and WML template of the vm1dev26 spoken message.

6. Delivery of summaries via WAP

The Push Service features all the underlying services that a WAP connection may have i.e. secure connection, connection oriented or connectionless sessions. Our experimental setup uses connectionless sessions based on the simplest submission service, which comprises of the datagram layer, the session layer, the Push OTA layer and finally the application layer that contains the actual summary.

After the voicemail summaries have been produced, they are fitted into a WML template designed particularly for this service. The template allows specific piece of information such as caller's number or name and time of call to be displayed in a more formalized way. The WML page is then forwarded to a mechanism that encapsulates it within the appropriate protocol headers. The page is then sent to a WAP Push Proxy gateway, which handles the compaction of the page as well as the communication with the SMS gateway. For the delivery of the push content to the WAP Push Proxy gateway the Push Access Protocol (PAP) is used. PAP runs over TCP/IP and it carries information to the WAP Push Proxy gateway when the messages should be sent, through which network bearer, and a number of other options that can be found in the PAP specification [17].

The Nokia WAP Toolkit[4] 1.3 beta was used in order to compile the WML code. Its emulator offers a real-time depiction of the content on a WAP enabled handset and figure 4 shows the summary of vm1dev26 message on such a display. In case a text summary suffers from coherence degradation, readability deterioration and topical under-representation, a link that allows a direct connection to the voicemail system is offered to the user in order to listen to the particular message. This functionality is specified by the WTA [22] and save users from typing numbers and passwords while accessing telephony services.

Figure 4. Summary retrieval of the vm1dev26 on the display of a WAP compatible phone. An optional connection to the voicemail system in order to listen to the particular message is provided by the WTA.

7. Message submission experiments

The testbed used to estimate the text summaries submission delays is described in Figure 5. It consists of two PCs,

[4] Nokia Wireless Data Forum, http://www.forum.nokia.com

each connected to a V.DOT. The latter are self-contained GSM phones with voice, data, fax and SMS capabilities controlled by standard and extended AT commands from the RS232 serial port. Expressway 2.0 of Dialogue Communications was used on both computers as SMS gateway software. The WBXML encoding as well as the software to embed the summary text to the WML template was especially written for the purposes of this application. The headers of the protocols have been hardcoded because at the time of writing there were no WAP phones on the market supporting push capabilities, as this feature was specified in the WAP 1.2 specifications [18].

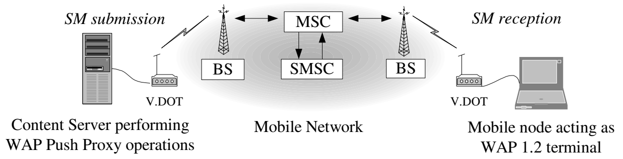

Figure 5. Testbed architecture overview.

In order to calculate the delay of the voicemail summary delivery, two types of measurements have been performed; the time needed for the V.DOT device to send a SM and the time needed for the SM to be received by the mobile client. SMs can be transmitted over the mobile phone's air interface using the signaling channels so there is no delay for call setup. SMs are stored by an entity called Short Message Service Center (SMSC) and sent to the recipient when the subscriber connects to the network. The average delay time for a successful delivery of a pair of SMs is given by the following:

$$D_{service} = 2 \times T_{setup} + T_{submission} \quad (5)$$

where T represents the corresponding time. This is because our testbed puts the SMs in a queue, while the operators SMSC supports multiple submission sessions. Figure 6 shows the average delays after delivering SMs corresponding to the summaries of 25 voicemail messages (50 SMs in total) during four time ranges in a single day over BTCellnet's network. The submission mechanism over this network gives a stable performance by delivering a message summary WAP page after approximately 10 sec. Further performance analysis and results on SMS submission utilizing different protocols over multiple networks can be found in [6] and [3]. Although the implementation of SMS in GSM networks is very much standardized and is defined in the GSM specifications, each operator provides a proprietary interface. By using WAP Push instead, the only information that is different from users under one operator to another is the location of the Push Proxy gateway while the rest of the protocol (PAP) would be the same making much easier the development of portable applications across different networks. In order to estimate the full data path delay

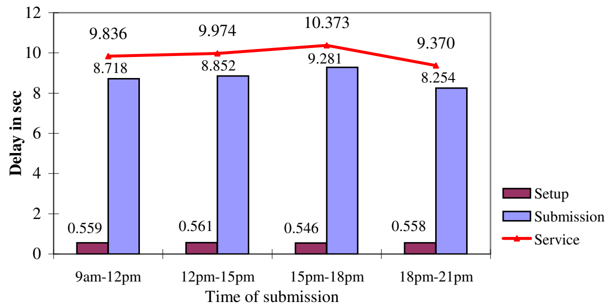

Figure 6. Average setup and submission delays corresponding to the delivery 25 voicemail messages (50 SMs in total) within each time range.

we have to add the recognition (including feature extraction), summarization and message formating times. Text summary encoding into the WAP domain takes a fraction of second and can be regarded as insignificant. The summarization algorithm presented in Figures 2 and 3 is implemented in C programming language and its execution time depends mainly on the length of the NE lists. For the NE lists described herein the summary of a transcribed message is produced within 3-4 sec. The latter delay can be reduced by using more efficient indexing techniques such as hash tables. On a PII PC with 512MB of RAM, fast and memory efficient CHRONOS decoder [13] gives a 1.3× real-time recognition. Consequently, a typical spoken message contained in the Voicemail I corpus with duration of 31 sec can be delivered on a WAP compatible terminal as a text summary within approximately 54 sec. The above figures are approximate and provided only to indicate that significant delay savings are possible. Further investigation into feature extraction methods, NE lists length, configuration of the mobile network and traffic load are likely to have a significant effect.

8. Conclusion

This paper has presented a real-time end-to-end solution that enables mobile users to receive text summaries of their voicemail without an explicit request by taking advantage of the WAP Push framework. By transcribing and extracting information from messages content can be delivered quickly, without human intervention and no matter the location of the user. The above architecture can be further customized to meet the requirements of a complete voicemail value-added service, offering significant benefits to both the network operators and their subscribers. Among the most interesting future extensions of such a service would be the construction of summaries according to prespecified user profiles. Message filtering can also be employed in order to deliver only the summaries of preselected message types i.e. personal, professional or urgent.

9. Acknowledgments

This project was sponsored in part by Dialogue Communications, a Sheffield-based company that develops and implements mobile data and Internet messaging solutions. We also acknowledge discussions with S. Cvetkovic.

References

[1] H. Bourlard and N. Morgan. *Connectionist speech recognition: A hybrid approach*. Kluwer Academic Publishers, Boston, USA, 1994.

[2] H. Hermansky. Perceptual linear predictive (PLP) analysis of speech. *Journal of the Acoustical Society of America*, 87:1738–1752, 1990.

[3] H. Jiang. Reliability, costs and delay performance of sending short message service in wireless systems. In *Proc. of the ICUPC*, pages 1073–1077, Florence, Italy, 1998.

[4] Y. Kato. Voice message summary for voice services. In *Proc. of the ISSIPNN*, pages 622–625, Hong-Kong, 1994.

[5] B. Kingsbury, N. Morgan, and S. Greenberg. Robust speech recognition using the modulation spectrogram. *Speech Communication*, 25:117–132, 1998.

[6] K. Koumpis, S. Cvetkovic, and G. Peersman. Performance evaluation of SMS-based email and voicemail notification architecture. In *Proc. of the 5th Workshop on Emerging Technologies in Telecommunications*, pages 282–286, Bayona, Spain, 1999.

[7] K. Koumpis and S. Renals. Transcription and summarization of voicemail speech. In *Proc. of the ICSLP*, volume 2, pages 688–691, Beijing, China, 2000.

[8] J. Makhoul, F. Kubala, R. Schwartz, and R. Weischedel. Performance measures for information extraction. In *Proc. of the DARPA Broadcast News Workshop*, pages 249–252, Virginia, USA, 1999.

[9] M. Metter and R. Colomb. Wap enabling existing HTML applications. In *Proc. of the 1st Australasian User Interface Conference*, pages 49–57, Canberra, Australia, 2000.

[10] M. Padmanabhan, E. Eide, G. Ramabhardan, G. Ramaswany, and L. Bahl. Speech recognition performance on a voicemail transcription task. In *Proc. of the ICASSP*, volume 2, pages 913–916, Seattle, USA, 1998.

[11] E. Paksoy, A. McCree, V. Viswananthan, and J. Linn. A variable-rate CELP coder for fast remote voicemail retrieval using a notebook computer. In *Proc. of the IEEE Workshop on Multimedia Signal Processing*, pages 119–124, Princeton, USA, 1997.

[12] G. Peersman, S. Cvetkovic, P. Griffiths, and H. Spear. The global system for mobile communications short message service. *IEEE Personal Communications Mag.*, 7:6–14, 2000.

[13] T. Robinson, J. Christie, and G. Cook. Time-first search for speech recognition. *Submitted to Speech Communication*, 2000.

[14] M. Stevenson and R. Gaizauskas. Using corpus-derived named lists for named entity recognition. In *Proc. of the ANLP*, pages 290–295, Seattle, USA, 2000.

[15] R. Valenza, T. Robinson, M. Hickey, and R. Tucker. Summarization of spoken audio through information extraction. In *Proc. of the ESCA Workshop on Accessing Information in Spoken Audio*, pages 111–116, Cambridge, UK, 1999.

[16] WAP Forum, http://www.wapforum.org. *Wireless Application Protocol Architecture*, Ver. 30-Apr. 1998.

[17] WAP Forum, http://www.wapforum.org. *Push Access Protocol*, Ver. 08-Nov. 1999.

[18] WAP Forum, http://www.wapforum.org. *Push Architectural Overview*, Ver. 08-Nov. 1999.

[19] WAP Forum, http://www.wapforum.org. *Push OTA Protocol*, Ver. 08-Nov. 1999.

[20] WAP Forum, http://www.wapforum.org. *Wireless Datagram Protocol*, Ver. 05-Nov. 1999.

[21] WAP Forum, http://www.wapforum.org. *Wireless Session Protocol*, Ver. 05-Nov. 1999.

[22] WAP Forum, http://www.wapforum.org. *Wireless Telephony Application*, Ver. 08-Nov. 1999.

[23] WAP Forum. *WAP Binary XML content format*, Ver. 4-Nov. 1999.

[24] G. Williams and S. Renals. Confidence measures from local posterior probability estimates. *Computer Speech and Language*, 13:395–411, 1999.