among all pairs of media: again, some cases are easy (e.g. E-mail to FAX), and some are daunting (speech to E-mail).

We are cautiously optimistic that a multimedia message retrieval research project focused on 'spotting of preregistered words or names,' and perhaps allowing the selection of 'important messages,' as defined above, is a reasonable candidate for near-term research. Some research is certainly required, since the content and quality of the messages are unconstrained. But the strategy of constraining the query narrowly in this way, together with exploiting the time that messages wait in mailboxes to improve the accuracy of pre-indexing, gives hope that it is technically feasible.

References

[1] P. Schnofr, "Integrating Video into an Application Framework," *ACM Multimedia*,93, pp. 411–417, Aug. 1993.

[2] A. Pizano and T.Y. Hou, "Integrated Multimedia Messaging Concepts and Applications," in *Workshop on Multimedia Applications, 24th ACM CSC Conference*, Feb 1996.

[3] F. R. Chen, L. D. Wilcox and D. S. Bloomberg, "Detecting and locating partially specified keywords in scanned images using Hidden Markov Models", *Proc. 2nd ICDAR, Tsukuba City, Japan,* Oct 1993, pp. 133-138.

[4] Richard C. Rose, "Discriminant wordspotting techniques for rejecting non-vocabulary utterances in unconstrained speech", *Proc. ICASSP 92, Vol.2, San Francisco,* March 1992, pp. 105-108.

the word (spoken by the user). The response is to select the chosen messages.

It may be possible to implement all 9 combinations of message-type and query-type. Much of the computer processing can be performed off-line, immediately after the message is received (using the full list of prere- gistered words), long before the user asks for them. Thus the user suffers only a small delay, while the query is being understood. And, the computing power available to accomplish word-spotting is much less constrained than if it had to be accomplished while a user was waiting. So, the user enjoys two quality advantages: faster response and higher reliability.

B. Important Messages

E.g. "Anything important?"

This is an extension of the case above: instead of searching for one preregistered word, in this case all of the words will be searched for. All messages in which any of these words are spotted will be selected.

This will be, perhaps, only slightly harder than A., if the number of preregistered words is small (say, a few dozen). This seems a reasonable assumption for most users. However, we expect that some users may gradually add words (there may be little incentive to delete any) until hundreds accumulate (at which time it is tedious to delete any).

C. Word Spotting

E.g. "Anything about 'media'?"

This differs from the above cases in that the word is an ordinary English word whose spelling, speech attributes, and pronunciation can be reliably looked up in a table (not an arbitrary proper name). No 'preregistration' of the word is required, either by spelling or by voice recording. The response is to select the right messages.

The overall difficulty may be much greater that the above cases. The number of possible words is much greater. This query type may be confusing: users may have a hard time remembering, or even understanding, that non-English words (e.g. proper names) are not supported.

D. Name Spotting

E.g. "Anything from 'Juang'?"

This differs from C. above in that the word is an arbitrary proper name, whose spelling, speech attributes, and pronunciation may not be known in advance. (And, it is not 'preregistered'.) The response is to select the right messages.

Conclusions

Queries of the types discussed above do not exhaust the technical challenges of multimedia messaging. Users may want messages to be summarized: in general, this is immensely hard, but some special cases may be easy: e.g. E-mail or FAX messages with 'Subject:' lines. And, users may wish to be able to get messages in any medium which is handy, so requiring conversion

than others. An example: the task of searching for an arbitrary proper name (say, 'Kung'), expressed as ASCII text, is almost trivial in E-mail messages, is often but not always feasible in FAXes [3], and is a long-term research challenge in speech signals [4].

If we stress the principle of uniformity over all the media, it might be interesting to resist the temptation to allow far richer queries in one medium than in another, where it is easy to do so: as, in the above example, we would allow complex queries in E-mail but not in speech. Instead, we might try to constrain the queries so that all media can be served by the same queries (if not uniformly well, perhaps). A significant benefit of such a policy could be that impatient users won't have to remember special cases.

To push the example one step further, we may find that it is easier to 'spot' a proper name in a speech signal if it has been 'registered' earlier in some way, say, as both a spoken utterance and as text (perhaps 'spelled out' verbally). If so, then we require names to be registered before they can be queried, in email and FAXes as well. Such a policy may seem peculiarly rigid, but it is interesting to note that it might permit the names to be queried verbally (that is, the query is spoken) and then applied successfully to FAXes and email, which would otherwise have been technically difficult or impossible!

This is just one example of interesting and perhaps unexpected interactions among the media, due to the varied capabilities that the state-of-the-art allows, which we may encounter and exploit in such a 'uniformly multi-media' research project.

We won't attempt at this time to specify all the technical components of such a system, or all of the research and engineering tasks that must be coordinated. However, each of the following research fields can make a significant contribution: textual information retrieval (topic identification); speech recognition (word-spotting, phrase-spotting, speech-directed queries); speech synthesis (email and FAX reading, prompting); speech coding (coding features for searching); document image analysis (FAX OCR readers, language identification, word-spotting); intelligent agents (adapting queries to various media; 'standing queries'); natural language understanding (summaries & verbal rendering of text); dialogue systems (understanding informally posed questions); indexing and searching (feature extraction, query languages, search algorithms); computer/telephony integration.

Technical Challenges

We list a few potentially challenging and useful research projects in the indexing and retrieval of multimedia messages. For lack of space, we do not go into detail about their comparative difficulty.

A. Pre-registered Word/Name Spotting

E.g. "Anything about 'Reibman'?"

The word to be spotted must have been 'registered' in advance by the user, by typing in (or, speaking) the correct spelling, and recording the sound of

- what's it about?
'Summarize the long messages.'

- what language is it in?
'Put all German messages in mailbox G.'

- 'standing' queries
'Call me when the FAX from the IRS arrives.'

- ... (others)

Similarly, the response to the various queries can come in different media as indicated in Figure 1.

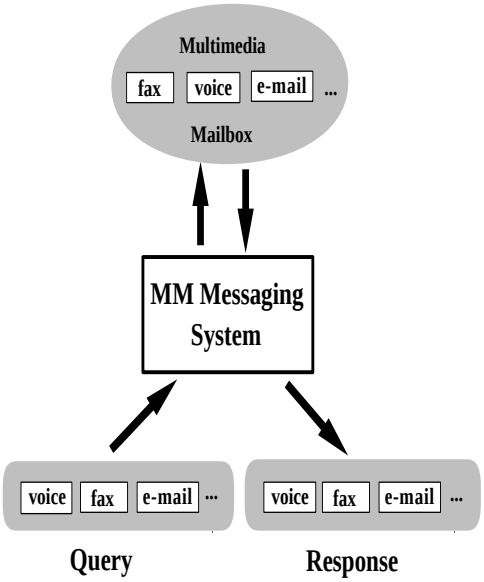

Figure 1: The telecommunications messaging environment.

A Research Strategy

Such a system has the potential to be 'multi-media' in an unusually strong sense:

- any query can be applied to messages in ALL three media;
- the queries themselves can be posed in ANY of the media – text via keyboard (KBD) or E-mail, voice control, or via FAX; and
- replies to a query can be expressed in ANY of the media: in text on the screen, via email, via recorded or synthesized speech, and of course by FAX.

Some of these combinations make more sense than others from a user's point of view. And, of course some are technically much harder to achieve

munications messaging applications [2]. Both will probably require roughly – sometimes exactly – the same kinds of technical advances for success. The scale of both applications can be very large, in the aggregate, even though each individual message stream (or associated MM 'mailbox') may contain only a small number of messages.

However, we expect that they will lead to different styles of research. For example, queries among messages may often require near-real-time replies, and so 'on-line' query processing may matter more than 'off-line' methods. At the same time, messages may linger in mailboxes for a long time, on average, before being queried: this can create opportunities for off-line indexing in which the time available for indexing is comparatively large and, for individual messages, potentially unbounded.

Also, the relative importance of the various media may differ in the two cases: for example, video images are found in large data bases today, whereas it may be some years before they occur routinely in telecommunications messaging [1](though this has already begun).

Further, the typical user may be materially different in each case. The user of a large, static digital library may be patient and may eventually become expert, because he/she has no other way to get the data; thus we may expect a demand for complex queries and willingness to use them. By contrast, a person querying message streams is likely to be more impatient – in fact, his/her impatience with a backlog of messages, cluttering the mailbox, may be the greatest incentive to use the retrieval system – so we can expect that the queries will almost always be simpler: ideally, in nearly natural language, idiomatic, and easy to recall.

This paper discusses some aspects of next-generation multimedia indexing and retrieval systems applied to message streams.

The Telecommunications Messaging Environment

The MM data to be queried is messages passing among business office workers via enterprise-wide telephony and data networks. The media of interest here as illustrated in Figure 1 include voice, email, and FAXes, initially: at some later date, binary-file 'attachments' (including electronic business forms), static and video images, electronic ink and other media might be added. We assume, for the sake of this discussion, that mechanisms for the transmission and storage of these messages exist, allowing interaction via telephone, computer terminal, or FAX machine.

We allow queries of various kinds, which are applied to all the waiting messages, in whatever medium – for example:

- who are they from?

'Any word from Chen?'

- what people/companies/universities/topics are mentioned?

'Is there a submission from BellCore?'

NEXT-GENERATION MULTIMEDIA MESSAGING

Henry S. Baird
Jianying Hu
Ramanujan S. Kashi

Bell Laboratories
Lucent Technologies
700 Mountain Avenue, Room 2C-322
Murray Hill, NJ

{hsb|jianhu|ramanuja}@bell-labs.com

Abstract - This paper discusses some technical aspects of multimedia messaging systems. By 'messages,' we mean non–real–time communications containing various media (text, speech, FAXed document images, video, electronic ink, etc), passing between users over enterprise-wide telephony and data networks, and collecting in 'multimedia mailboxes.' Messaging, like 'digital libraries,' is an application domain for indexing, retrieval, searching, etc — but it offers some significantly different challenges and opportunities. We discuss some technical implications of the expected contents of messages, waiting time in mailboxes, and the particular needs of messaging users.

INTRODUCTION

The motivation for research into 'multi-media (MM) indexing & retrieval' is to make collections of data in various media (text, speech, FAXed document images, general images, video, electronic ink etc) more readily accessible to users. The services that result should ideally allow users to 'query' the MM data for various purposes, *e.g.* search, select, retrieve, summarize, route, alert, etc. This querying can itself be a multimedia operation in several different ways: the queries, the data that is queried, and the reply to the query ALL may be expressed in various media; and thus there may arise a need for 'translation' among media.

We find it helpful to distinguish two ways in which the MM data to be queried can be presented:

(1) data bases: concentrated, and more or less static, within, e.g., a digital library; and

(2) message streams: distributed, and more or less dynamic, within, e.g., the office telephony and LAN networks of an enterprise.

The first of these leads to research which is similar in spirit to today's academic "digital library" initiatives. The second arises in modern telecom-