THE CONTRIBUTION OF AUTOMATIC SPEECH RECOGNITION FOR KEYWORDS TO ASSIST IN THE INTEGRATED ORGANISATION OF DIGITAL MESSAGES

Gordon Hunter	Kingston University, London, KT1 2EE, U.K.
James Denholm-Price	Kingston University, London, KT1 2EE, U.K.
John Yardley	JPY Ltd., Surbiton, Surrey, KT6 4TW, U.K.
Thomas Michel	JPY Ltd., Surbiton, Surrey, KT6 4TW, U.K.
David Fox	JPY Ltd., Surbiton, Surrey, KT6 4TW, U.K.

1 INTRODUCTION

Nowadays, people may have to deal with electronically-transmitted messages of a variety of types: e-mails (potentially with attached files), SMS text messages and (digitised) telephone messages. Many e-mail systems allow users to browse through their e-mail archives, and search for specific e-mail messages, by subject, by sender or recipient(s), or by "keywords" in the content of the message, or possibly even in an attached file. However, this is not normally possible with telephone messages, even ones which have been recorded digitally. Even for text-based messages and files, most searching systems rely on finding an exact match between the entered search term or phrase and identical instances of the same text within the documents being searched. This method is not at all robust to small human errors (such as spelling or typing mistakes), or only part of the search phrase occurring in the documents.

In this paper, we describe initial experiments undertaken in an approach to address this situation, based on the assumption that the essence of the theme and meaning of a document can be summarized (and searched) in terms of the "keywords" – less common words which are directly related in meaning to the content and context of the document. Our approach also applies Automatic Speech Recognition (ASR) to recordings of telephone calls in an attempt to recover sufficient keywords from the recording to enable the content and theme of the telephone message to be identified. We hypothesise that, even if the recording is not of particularly good quality, or if the transcription obtained from the Automatic Speech Recognition system contains a relatively high proportion of incorrect words, enough correct relevant keywords will be obtained to make useful inferences about the topic of the conversation. This should enable integration of recorded telephone messages with e-mails and other text-based electronic messages, with a searchable database. In turn, this should lead to greater efficiency and save users time and effort searching for important information regarding their work and other aspects of their lives. A novelty of this system is that, whereas ASR is normally used to input information into a computer-based system, in this application we are using ASR to get information back from the system. Furthermore, because there is generally not the need to perform ASR in real-time in this situation, the choice of a suitable ASR system is much wider than for a "Dictaphone" or "real time input to a system" type of application.

2 THE *THREADS* SYSTEM

2.1 The Need for Intelligent Methods for Finding Messages

Readers will no doubt be familiar with the situation where one struggles to find an important document or message when it is needed in a hurry. Although the user can remember roughly what the document or message was about, he/she may not be able to recall where it was saved, what it was called (or exactly what its subject line was), or indeed, who sent it or who were its recipients. Conventional methods of searching mailboxes or files require exact matches to at least part of a search term – so if the search term used is completely incorrect, the search will clearly not return the desired result. Such schemes would be very difficult to apply to collections of audio recordings, for example of telephone messages, unless some kind of text summary of the recording were also provided. One alternative approach, for text documents at least, is to search by content, but this still relies on finding an exact match somewhere in the target document(s) of at least part of the search term. Another, more flexible and powerful methodology is to compare how similar each document or message in the set being searched through is to a document containing a set of words expressing what the user believes are the most important ideas, summarizing the topic or theme of the desired target document(s). Similarly, we can obtain an indication of whether two documents have essentially the same theme or topic by comparing how similar they are. Usually this similarity (or level of difference) is measured in terms of statistics regarding the words each of the two documents contain, including words they share in common and words appearing in one but not the other. Some such metrics are based on relatively simple word counts, others make use of these counts to compute information theoretic quantities such as entropy, cross entropy, mutual information and perplexity [1]. The "target" need not necessarily be a single document or message – it could equally well be a set of documents or messages all relating to the same topic or theme, for example all about the same project. However, some words impart far more information than others. A high proportion of words in both text and speech are "function words", present to serve a grammatical function, but which carry very information about the theme of the document or utterance. Examples of such function words are "is", "has", "and", "but", "he" and "so". Most of these are very common across a wide variety of documents and conversations, regardless of their topic. In contrast, words which give a large amount of information about the topic of the text or utterance – so-called "content words" – tend to be much less common generally. Examples of this include "technology", "contract", "acoustics" and "reverberation". Knowledge of which of these "content words", hereafter referred to as keywords, are in a given document, collection of documents or conversation, and how frequently, should be of great value in judging what the theme of the text or utterance is. Several authors [1, 2, 3] have noted that the distribution of words and their frequencies will vary greatly according to the type of document, speech or conversation under consideration, and also its topic.

The aim of this work is to investigate how easily and reliably a document or message can be classified – in our case, allocated correctly to a particular project – based on the statistics of keywords contained in the message, in comparison to other messages in the same, and in other, projects. A previous paper [4] described our initial experiments using a document similarity metric based on lexical content to classify text (e-mail messages) by their content. In this paper, we focus on the successful identification of keywords in telephone calls using Automatic Speech Recognition (ASR), with a view to this being applied in an analogous way for the classification of telephone calls by theme or project to which they are related.

2.2 The Development of *Threads*

The *Threads* project [5] evolved from the needs of a small company to share information whose representation had, over a period of 20 years, changed from physical to digital [6]. As such, the storage medium moved from shared physical access - filing cabinets - to private digital access - personal e-mail accounts. While the digital form was easier to store and search, it often ended up locked in private user files. This applied primarily to company e-mail, but also extended to documents (e.g. invoices, quotations, etc.) that were not confidential and routinely needed to be shared amongst employees. As e-mail became the de-facto currency of communication, where most documents were shared as email attachments, sharing the e-mail became the key to re-establishing the collaboration ethos [7].

Various procedures (e.g. using shared IMAP folders) were put into place to allow the sharing of e-mail, but none was really successful. Apart from the fact that these required strict discipline on the handling of e-mail, they often fell down when the user had several e-mail accounts, or was working from remote locations. Special purpose software - such as bespoke e-mail clients - were universally shunned, and rightly so [8]. However, gains made by digitisation were often lost due to unnecessary possessiveness of individuals. A culture of non-sharing had evolved and become the norm.

Threads [5] is a web application which was designed to obviate the need for staff to treat their mail specially. It was necessary that staff continue to use their preferred e-mail client and not change their working practices. While this was largely a technical exercise, it demanded a staff culture shift back to the days where the default case was to share information. After some initial reluctance, staff soon realised that, with the appropriate controls in place, sharing could be liberating and few would return to the possessive "old days". Staff quickly understood the benefits of such openness because, with *Threads*, they could find things they previously could not and, just as importantly, they were able to devolve responsibility when it was useful. When digital telephony was introduced within the company [9], including it within the same *Threads* message handling framework was immediately successful: once telephone calls could be located, retrieved and dealt with in the same manner as text-based messages - and vice-versa - the true power of message sharing was realised. That said, it became clear that it was important to unify digital messages such that, whatever the medium, they were presented to the user in a clear and natural way. This need for abstraction was a significant challenge.

Unification and association of messages in any meaningful way requires context. The useful information content of each message is not always obvious or available but, once the message has some context, then there is often sufficient information available to decide if it is worth reading or to what other messages or themes it relates. For *Threads*, the context is a combination of the project or job to which the message relates, and the people communicating through it.

Although not a technical point, it is worth noting that, in our case, the integration of telephone calls into *Threads* was never motivated by legal compliance or disclosure issues. Indeed, early on, we took the decision that calls would only ever be shared internally. What prompted this initially was to provide the knowledge that certain phone calls had occurred within the "thread" of a project - without regard to the actual content of these calls. With JPY's conversion to IP telephony [9], and the availability of call recordings in digital format, it seemed an obvious next step to include these within our *Threads* system. Once they were freely accessible to all staff, users realised they no longer needed to rely on scribbled notes or their recollection of a call. With *Threads*, they could find a call just as they would an e-mail, listen to it as many times as they wished, or pass the "link" to relevant colleagues. A schematic diagram illustrating how Threads stores and processes both text e-mail messages and telephone calls is shown in Figure 1, with screenshots of the "general" and "telephone call" Threads user interfaces in Figures 2 and 3 respectively.

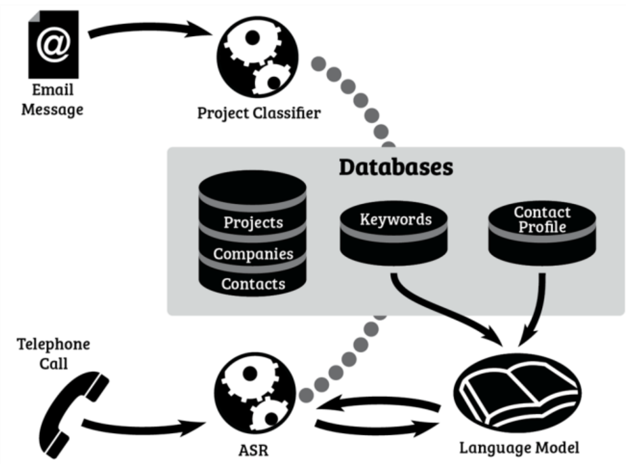


Figure 1: Schematic representation of how messages (both e-mails and telephone calls) are processed and stored by *Threads*.

2.3 Previous Work on Text (E-Mail) Classification Using Keywords

In a previous paper [4], we described a set of experiments in which keywords within text documents (in our case, e-mail messages) were identified, and the documents subsequently clustered in such a way that documents which contained similar sets of keywords would be put into the same cluster. This methodology was applied to two sets of e-mail messages – one from the company JPY, the other from the publicly available Enron dataset [10, 11] – for which it was known that each message belonged to a particular "project" or "folder", and it was believed that messages within the same project or folder should share a common theme, and hence should probably contain similar keywords. The results of that study indicated that summarizing text messages according to the keywords they contained could, in conjunction with the use of an appropriate text similarity metric, could lead to a useful method of classifying e-mails by theme. As noted above, the aim of this present paper is to investigate whether a similar approach could be suitable for classifying recordings of spoken messages, such as telephone calls.

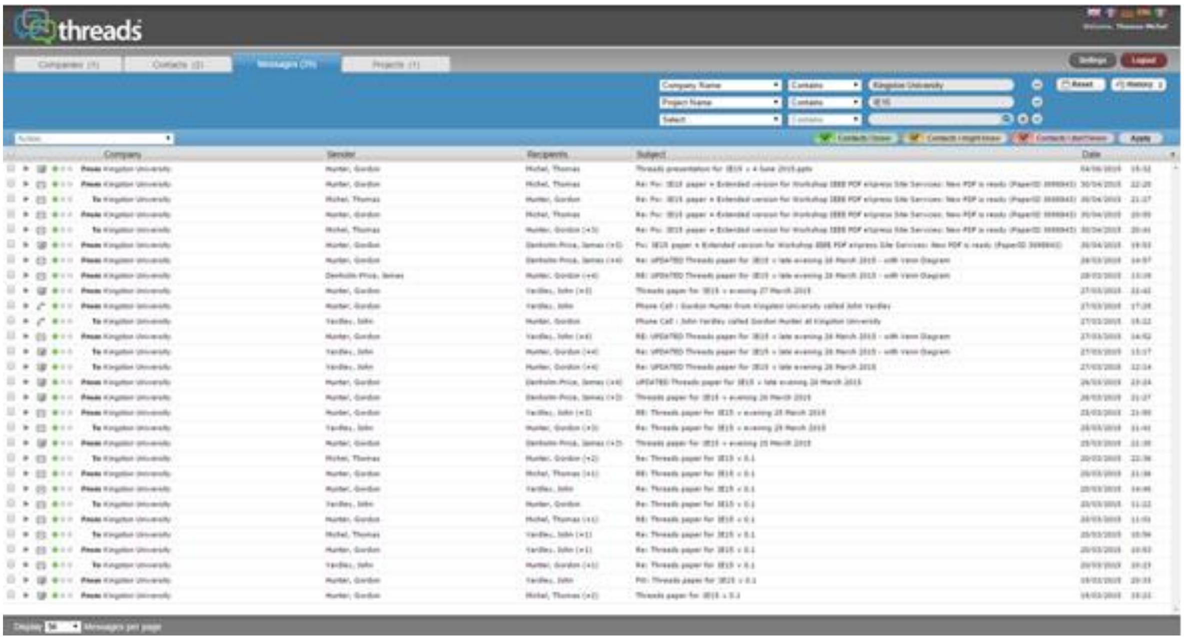


Figure 2 : *Threads* general User Interface for all types of messages.

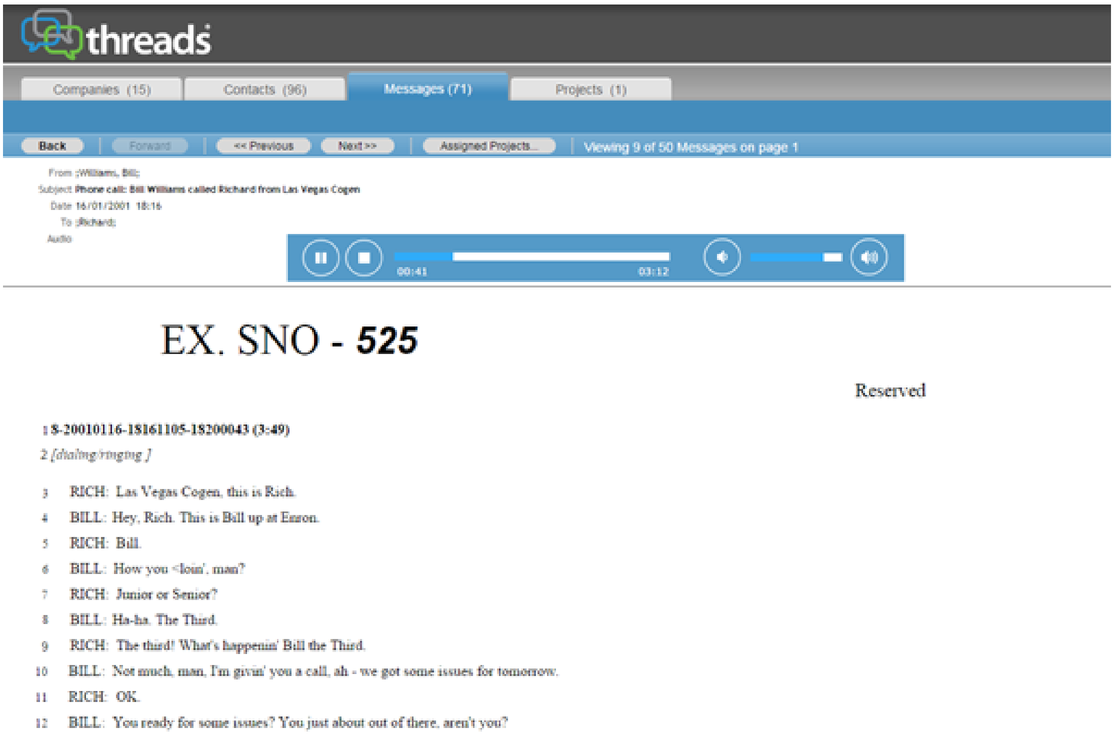


Figure 3 : *Threads* User Interface for telephone calls, showing both the audio recording and the transcript (if available).

3 EXPERIMENTS USING AUTOMATIC SPEECH RECOGNITION

In order to investigate whether the same approaches could be applied to telephone messages, with the benefit of transcriptions provided by an Automatic Speech Recognition (ASR) system, we performed experiments on keyword recognition from audio recordings of telephone calls from two different sources, for which manually made transcriptions were available. Although it is acknowledged that automatically produced transcriptions are very likely to contain errors, the purpose of this study is to investigate whether knowledge of the keywords contained in a message proves sufficient to classify the message reliably. Thus, the role of the ASR system here is not necessarily to supply a reliable verbatim transcript, but only to identify the majority of the keywords correctly. Two experiments were performed for each dataset – one to study the proportions of keywords correctly identified using ASR without using any tailoring of the vocabulary, the other identified a set of relevant keywords from a set of e-mail messages from the same company, and used these as a restricted vocabulary for the ASR. The ASR system used throughout was Nuance's *Dragon Dictate for Macintosh*, version 4 [12]. The "correct" keywords for each telephone call were assumed to be those found in the manually produced transcript of that call.

3.1 Datasets

Our methodologies have been tested on two datasets of recorded 'phone calls for which manual transcriptions (assumed to be reliable) were available. The first of these was a collection of 10 recordings of telephone calls either made by, or made to, people in the company JPY Ltd. The second was a set of 10 telephone calls made by or two employees of the former US energy company Enron, which are now available in the public domain [10, 11].

3.1.1 JPY Dataset

JPY Ltd., the authors of *Threads*, is a small software development, distribution and consultancy company, based in a South-Western suburb of London. JPY has around 10 employees, who need to communicate with each other, their suppliers, customers and collaborators by e-mail and telephone on a regular basis. The nature of the company means that several of the employees, and possibly the directors, will be involved in the same project, and hence benefit from the ability to access and search all e-mails and recordings of telephone calls relating to a particular project in which they are involved. The JPY dataset used here for testing our method consisted of a total of 10 recordings of telephone calls and manually-created transcriptions of each, made by the JPY employee involved in the call. Those 10 calls each contained between 450 and 1025 spoken words in total.

3.1.2 Enron Dataset

The ENRON Corporation of the U.S.A. was, until it went bankrupt in 2001, one of the World's largest energy providers and employed around 20 000 people. However, the bankruptcy triggered an investigation, which led to a major fraud trial, and the U.S. legal authorities required scrutiny of all still extant ENRON e-mails and recorded telephone conversations [13]. This resulted in a large corpus of some 250 000 e-mails, relating to 350 projects, involving around 28 000 contacts from 1 500 companies, plus 76 recordings (with transcriptions) of telephone calls being available in the public domain [10, 11]. For the purposes of this experiment, in order to use a sample of comparable size to the JPY dataset, we have taken a sample of 10 of these ENRON telephone calls, each containing between 440 and 1250 spoken words in total. It is not known who made the transcriptions of these calls (presumably carried out for purposes associated with the legal proceedings), but it was assumed that the messages were all transcribed accurately.

Since it offers a substantial resource of text from real business e-mails, and is in the public domain, the Enron dataset has attracted a large amount of attention from researchers, and has been the subject of several published research projects, e.g. [10, 14, 15].

3.2 Experiments

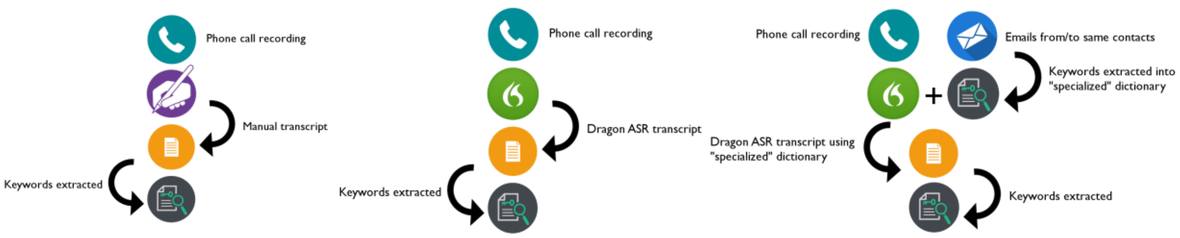
For each dataset, "keywords" were identified as those words present which were neither in the 5000 most common words found in the British National Corpus (a large dataset of modern British English, compiled during the 1990s) [16], nor in the most common 1000 words found in the dataset of current interest (JPY or Enron, as appropriate), but which contained at least 4 characters. A "word stemming" algorithm was also applied, so that simple variants of the same word were considered to be identical. For example, the words "technical" and "technically" were considered to be the same. Between 1.2% and 6.4% of the total words in any one call were keywords.

Within each dataset, the keywords present in the manual transcription of each telephone call were found (Figure 4(a)). This set is considered as the "reference standard" of the correct set of keywords for that telephone message. For each such message, at least one of the speakers is known, and there are some e-mail messages either sent by, or to, or both, that individual present in the corresponding text dataset described in [4]. The keywords found in the e-mails to/from that individual were also noted separately. We shall refer to these as the corresponding text keywords.

For each telephone call, an automatically-produced transcription was created using the Dragon [12] ASR system without using any specialized language model or user profile (Figure 4(b)). Typically, the accuracy of this automatic transcription was not very good (less than 25%). The "keywords", according to the definition given above, within this transcription were found, and this set compared with those found in the manual transcription of the same call.

An attempt to improve the keyword recognition rate was made by amending the vocabulary by giving the corresponding text keywords (found in e-mails to/from the known speaker in the telephone call) "prioritized word" status. The recording was then processed by Dragon again, producing a revised automated transcription based on this amended vocabulary, and the keywords found in this new transcription again compared with those in the manual transcript (Figure 4(c)).

Counts were made of the "True Positive", "False Positive", "True Negative" and "False Negative" keywords obtained in each case. The results are presented in the next section.



(a) (b) (c)

Figure 4 : Schematic Representations of the Experimental Procedures on Telephone Calls (a) Manual Transcription (used later as "reference standard"), (b) Automated Transcription made by Dragon ASR system with default "general purpose" vocabulary, (c) Dragon vocabulary amended by noting "keywords" from related e-mails, then telephone call transcribed automatically using this modified "specialized" vocabulary.

3.3 Results

The results of our keyword recognition experiments, described above on the two datasets are given in Table 1 (for the JPY data) and Table 2 (for the Enron data) below. In both tables, TP denotes "True Positives" (i.e. correctly identified keywords), FP "False Positives" (i.e. keywords "identified" in the call which should not really be there) and FN "False Negatives" (i.e. keywords which were present in the call, but were not identified). The number of "True Negatives" (TN – keywords in the dictionary correctly not included in the transcript) can be inferred from the other values given.

3.3.1 JPY Data

JPY Telephone Call Number	1	2	3	4	5	6	7	8	9	10
Keyword count in manual transcript	11	12	20	11	27	14	12	27	29	19
Keywords correctly recognised in transcript (b) : TP(b)	3	5	8	2	5	5	2	3	3	2
Keywords correctly recognised in transcript (c) : TP(c)	6	5	12	4	10	7	6	11	10	5
Improvement in TP	**3**	**0**	**4**	**2**	**5**	**2**	**4**	**8**	**7**	**3**
Keywords incorrectly included in transcript (b) : FP(b)	28	17	36	31	57	17	23	30	38	30
Keywords incorrectly included in transcript (c) : FP(c)	10	9	33	11	37	11	7	16	25	9
Improvement (decrease) in FP	**18**	**8**	**3**	**20**	**20**	**6**	**16**	**14**	**13**	**21**
True keywords missed in transcript (b) : FN(b)	8	7	12	9	22	9	10	24	26	17
True keywords missed in transcript (c) : FN(c)	5	7	8	7	17	7	6	16	19	14
Improvement (decrease) in FN	**3**	**0**	**4**	**2**	**5**	**2**	**4**	**8**	**7**	**3**

Table 1 : Number of Keywords found in the manual transcript of each JPY telephone call studied, together with the corresponding counts for the automated transcriptions produced using the Dragon ASR system, using the default “general purpose” vocabulary (b), and the amended vocabulary (c).

3.3.2 Enron Data

Enron Telephone Call Number	1	2	3	4	5	6	7	8	9	10
Keyword count in manual transcript	19	8	24	19	18	25	18	15	18	15
Keywords correctly recognised in transcript (b) : TP(b)	2	2	2	1	1	2	0	1	0	1
Keywords correctly recognised in transcript (c) : TP(c)	5	4	7	6	4	7	2	5	7	5
Improvement (increase) in TP	**3**	**2**	**5**	**5**	**3**	**5**	**2**	**4**	**7**	**4**
Keywords incorrectly included in transcript (b) : FP(b)	15	12	14	11	22	23	29	20	13	29
Keywords incorrectly included in transcript (c) : FP(c)	8	9	11	8	15	11	22	10	8	20
Improvement (decrease) in FP	**7**	**3**	**3**	**3**	**7**	**12**	**7**	**10**	**5**	**9**
True keywords missed in transcript (b) : FN(b)	17	6	22	18	17	23	18	14	18	14
True keywords missed in transcript (c) : FN(c)	14	4	17	13	14	18	16	10	11	10
Improvement (decrease) in FN	**3**	**2**	**5**	**5**	**3**	**5**	**2**	**4**	**7**	**4**

Table 2 : Number of Keywords found in the manual transcript of each Enron telephone call studied, together with the corresponding counts for the automated transcriptions produced using the Dragon ASR system, using the default “general purpose” vocabulary (b), and the amended vocabulary (c). The terminology used is as for Table 1.

4 DISCUSSION

From Tables 1 and 2, it can be observed that, in general, the recognition rates for keywords using the Dragon ASR system applied to these datasets is rather poor, with the TP rates being quite low, and the FP and FN rates both being high. (This is particularly true for the Enron dataset.) This could be due to a number of factors – the quality of the audio recordings not being particularly high, and the limited bandwidth of telephone speech. Furthermore, although Dragon can be used in a speaker-independent way (as is the case here), its performance generally improves considerably as it becomes trained to a particular speaker’s voice and common vocabulary. A further complication for the Enron dataset was that most participants in the calls were speakers of North American English, whereas the version of Dragon used here was optimised for spoken Standard Southern British English. However, in both cases, the results improved considerably (TP increased, FP and FN both decreased) when the vocabulary used was adapted, taking into account the content of e-mail messages involving one of the participants from the telephone call under consideration, suggesting that, as might be expected, the vocabulary used varies from person to person, and knowledge of who the speaker is may provide useful information regarding the context and likely content of a telephone call.

5 CONCLUSIONS AND FUTURE WORK

The experiments described in this paper have shown that use of automated transcriptions using an ASR system may be a useful method for finding relevant keywords in recordings of telephone calls, which could in turn be used for classification of the message. However, the successful recognition rates for keywords, without using a more specialised language model, are poor. Modification of the language model by prioritising keywords obtained from related e-mail messages (e.g. sent by or to one of the speakers) greatly improves the correct recognition success rates for keywords, which should in turn improve the accuracy of automated classification of telephone messages.

It is planned to extend the scope of these experiments through the use of individual speaker voice profiles and language models for known callers (or calls from known numbers, from which a caller's identity may be inferred). For situations where a speaker is an employee of the host company (in this case, JPY) or is known to be from a small set of people who are regular callers, closed set Speaker Recognition, Identification or Classification (e.g. in the latter case, to identify speakers of a particular variety of English, such as North American or Scottish English) could also be used to identify appropriate acoustic and/or language models for that speaker.

It is also intended to integrate these speech technology-based features, together with intelligent message classification, into the *Threads* system to make it an even more powerful tool for business organisation, message retrieval and processing.

6 REFERENCES

1. Hunter, G.J.A.. & Huckvale, M.A. (2005) "An Evaluation of Statistical Language Models of Spoken Dialogue Using the British National Corpus", Proceedings of 1st IET International Conference on Intelligent Environments (IE'05), University of Essex, U.K., July 2005
2. Rosenfeld, R. (2000) "Two Decades of Statistical Language Modelling : Where do we go from here?", Proceedings of the IEEE, Vol. 88(8), pp 1270 - 1278
3. Hunter, G. & Huckvale, M. (2006) "Cluster-Based Approaches to the Statistical Modelling of Dialogue Data in the British National Corpus", Proceedings of 2nd IET International Conference on Intelligent Environments (IE'06), Athens, Greece, July 2006
4. Hunter, G., Denholm-Price, J., Michel, T., Yardley, J. & Fox, D. (2015) "Progress Towards a Smarter Office via a Novel Intelligent System for Message Organisation by Unifying E-Mails & Phone Calls", Proceedings of 4th International Workshop on Smart Offices and Other Workplaces (SOOW'15), Prague, July 2015, IOS Press Series on Ambient Intelligence and Smart Environments, Vol. 19, pp 15-26.
5. JPY Ltd. (2015). *Threads* Website. http://www.threads.uk.com (Last accessed 15th September 2015). Patent Applications "Digital Messaging System", UK Patent Application No. 1408302.6, International (PCT) Patent Application No. PCT/GB2015/050580
6. Chowdhury. G. (2010) "Introduction to Modern Information Retrieval" (3rd ed.), Facet Publishing.
7. Whittaker, S. & Sidner, C. (1996) "E-mail Overload : Exploring Personal Information Management of E-mail", Proceedings of CHI '96, pp 276-283
8. Schuff, D., O. Turetke, D. Croson, F. (2007) "Managing E-mail Overload: Solutions and Future Challenges", IEEE Computer Society, vol.40, No.2, pp. 31-36
9. Yardley, John (2010) "Adopting an Open Source VoIP phone system" http://www.jpy.com/support/articles/adopting-open-source-voip
10. Klimt, B. & Yang, Y. (2004) "The Enron Corpus: A New Dataset for Email Classification Research", Proceedings of European Conference on Machine Learning (ECML), Pisa, Italy
11. JPY Ltd. (2015) *Threads* Enron Database http://www.threads.uk.com/threads-enron-database (Last accessed 15 September 2015)
12. Nuance Communication Inc. (2015) "Nuance Dragon for Mac" http://www.nuance.com/for-individuals/by-product/dragon-for-mac/software/index.htm (Accessed 15 September 2015)
13. Smith, R. & Emshwiller, J. (2011) "24 Days: How Two Wall Street Journal Reporters Uncovered the Lies that Destroyed Faith in Corporate America", Harper Collins, New York
14. Shetty, J. & Adibi, J. (2005) "Discovering Important Nodes through Graph Entropy :The Case of Enron Email Database", Proceedings of 3rd International workshop on Link Discovery (LinkKDD'05), pp 74-81, ACM, New York
15. Wang, M., He, Y. & Jiang, M. (2010) "Text Categorization of Enron Email Corpus Based on Information Bottleneck and Maximal Entropy", Proceedings of 10th IEEE International Conference on Signal Processing (ICSP2010), Beijing, China, pp 2472 - 2475
16. Burnard, L. (2007) "Reference Guide for the British National Corpus" http://www.natcorp.ox.ac.uk/ (Last accessed 16 September 2015)