Evaluation of Extractive Voicemail Summarization

Konstantinos Koumpis and Steve Renals

Department of Computer Science, University of Sheffield
Regent Court, 211 Portobello St., Sheffield S1 4DP, UK
{k.koumpis,s.renals}@dcs.shef.ac.uk

Abstract

This paper is about the evaluation of a system that generates short text summaries of voicemail messages, suitable for transmission as text messages. Our approach to summarization is based on a speech-recognized transcript of the voicemail message, from which a set of summary words is extracted. The system uses a classifier to identify the summary words, with each word being identified by a vector of lexical and prosodic features. The features are selected using Parcel, an ROC-based algorithm. Our evaluations of the system, using a slot error rate metric, have compared manual and automatic summarization, and manual and automatic recognition (using two different recognizers). We also report on two subjective evaluations using mean opinion score of summaries, and a set of comprehension tests. The main results from these experiments were that the perceived difference in quality of summarization was affected more by errors resulting from automatic transcription, than by the automatic summarization process.

1. Introduction

Automatic text summarization has been an active research field for over 40 years (see Mani [1] for an introduction). There has been less work in speech summarization: Zechner's work on summarizing spoken dialogues [2] and the work on broadcast news summarization by Valenza et al [3] and Hori and Furui [4] are the most notable contributions.

In this paper we address the automatic generation of short text summaries of voicemail messages, that could be transmitted as Short Message Service (SMS) text messages. Voicemail summarization has several features that differentiate it from conventional text summarization:

1. Typical voicemail messages are short: the average duration of a voicemail message is 40s in the work reported here;
2. The summaries are extremely terse, designed to fit into a 140 character text message; hence coherence and document flow (style) are less important than the transmission of the main content of the message;
3. Since the voicemail message is transcribed by an automatic speech recognition (ASR) system, a significant word error rate (WER) must be assumed.

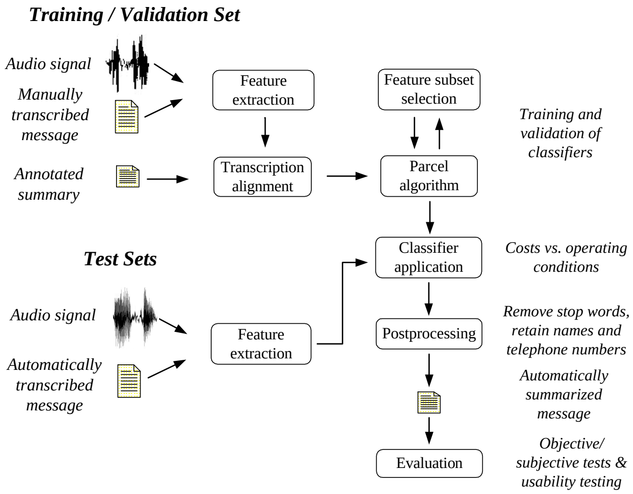

Training / Validation Set
Audio signal
Manually transcribed message
Annotated summary
Feature extraction
Transcription alignment
Feature subset selection
Parcel algorithm
Training and validation of classifiers
Test Sets
Audio signal
Automatically transcribed message
Feature extraction
Classifier application
Costs vs. operating conditions
Postprocessing
Remove stop words, retain names and telephone numbers
Automatically summarized message
Evaluation
Objective/ subjective tests & usability testing

Figure 1: An overview of the word-extractive approach that allows systematic comparisons and combination of patterns present in spoken audio for constructing summaries.

We have adopted a *word-extractive* approach to voicemail summarization. That is, we define a summary as a set of content words extracted from the original message transcription. Given this definition, we can frame the problem of voicemail summarization as a word-level discrimination task: for each word, decide whether or not it should be included in the summary.

Each word in the transcribed message is represented as a vector of lexical and prosodic features. We have trained classifiers on these feature vectors to discriminate between "summary words" and non-summary words. We used an ROC-based feature selection algorithm, known as Parcel [5], to determine the feature subsets. We have previously described this approach [6, 7], but we provide a brief overview here. In this paper we report on a number of evaluations of the voicemail summarization system, using human transcriptions of the voicemail messages and two different speech recognizer transcriptions.

K. Koumpis is currently with Domain Dynamics Ltd.

Using, slot error rate (SER), we compare the performance of the summarization system using the three different transcriptions. We have also performed some subjective tests. In the first of these tests we compared mean opinion scores (MOS) of users for human and automatic summaries from human and automatic transcriptions. A second set of tests assessed the summary quality in terms of comprehension, in which summaries were compared with the message audio.

2. Voicemail speech data

Voicemail speech is a “one way” communication over a telephone channel, and is characterized by varying degrees of spontaneity. Phenomena such as disfluencies, false starts and repetitions are relatively common, although there is a greater degree of preparation compared with spontaneous dialogue. Additionally, since the speaker is receiving no immediate feedback, the language model is quite different to Switchboard, for example, and there are an increased number of questions and instructions.

We have used the IBM Voicemail Corpus-Part I [8], distributed by the Linguistic Data Consortium (LDC). This corpus contains 1801 messages (14.6 hours, averaging about 90 words per message). We have two test sets: the 42 message development test set distributed with the corpus (referred to as test42) and a second 50 message test set provided by IBM (test50). The messages in test42 are rather short, averaging about 50 words per message, whereas the messages in test50 are closer 100 words per message. The messages in to the training set average of 90 words per message. The messages in the training set were characterised as 27% business-related, 25% personal, 17% work-related, 13% technical and 18% in other categories.

As described in [7], we trained a hybrid MLP/HMM recognizer [9] on the voicemail data. The system used two MLPs, one trained using perceptual linear prediction acoustic features, the other using modulation filtered spectrogram features. The log posterior probabilities estimated by the two networks were averaged to produce an overall log posterior probability estimate. During speech recognition training, we reserved the last 200 messages of the corpus as a validation set, resulting in a 1601 message training set.

A bootstrap trigram language model was estimated using the training transcriptions. Those sentences from the Hub-4 Broadcast News and Switchboard language model training corpora which had a low perplexity with respect to the bootstrap language model were appended to the training data. The language model was then reestimated. The pronunciation dictionary contained around 10 000 words derived from the training data, with pronunciations obtained from the Abbot/SPRACH broadcast news system [9], plus 1 000 new words with pronunciations mainly constructed following the rules used to construct the broadcast news dictionary. Following [10], we used 32 manually designed compound words.

	Train	Validn	Test42	Test50
Messages	800	200	42	50
Transcribed words	66 049	17 676	1 914	4 223
Total content words	20 555	5 302	561	820
Proper names	2 451	666	111	170
Phone numbers	3 007	577	120	190
Dates and times	1 862	518	46	81
Other	13 235	3 541	284	379
Compression Rate	31%	30%	29%	19%

Table 1: Voicemail content word annotation.

The OOV rates were 1.6% on test42 and 2.0% on test50. The average test set WERs were 41% on test42 and 44% on test50. The WER is not uniform, but is bursty within and across messages. We denote these transcriptions SR_1. Additionally, we obtained a second set of transcriptions (denoted SR_2) produced by the more complex HTK system developed for Switchboard and adapted to Voicemail [11]. The WER for SR_2 was 31% for both test sets.

3. Annotation of summary words

Word-extractive summarization uses classifiers that detect content words. We marked the summary words in 1000 messages of the Voicemail corpus. The first 800 messages were used as a summarization training set, and the last 200 used as the validation set. The transcriptions supplied with the Voicemail corpus include marking of named entities, and we built on this using the following scheme:

1. Pre-annotated NEs were marked as targets, unless unmarked by later rules;
2. The first occurrences of the names of the speaker and recipient were always marked as targets; later repetitions were unmarked unless they resolved ambiguities;
3. Any words that explicitly determined the reason for calling including important dates/times and action items were marked;
4. Words in a stopword list with 54 entries were unmarked;
5. All annotation was performed using the human transcription only (no audio).

As shown in Table 1 the summarization training set contained about 66 000 words, of which about 31% were marked as summary words (the compression rate). To assess the level of inter-annotator agreement we compared the performance of 16 human annotators asked to create word-extractive summaries for five messages, at a compression rate of 20–30%. 14 out of 16 of the annotators produced their summaries by progressively eliminating irrelevant words (rather than selecting content words), and in nearly all cases the annotators tended to a compression rate of 29–30%. Inter-annotator agreement may be measured by the κ statistic:

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

where P_o is the proportion of times the annotators agree, and P_e is the expected chance agreement. In this case κ averaged 0.48, indicating a relatively good level of agreement.

4. Feature selection

Each word was represented by a set of lexical and/or prosodic features. Feature estimation and selection is described in more detail in [7]; we present a brief overview here.

We used raw prosodic features containing information about pitch, energy, duration and pauses. Word duration was normalized both across the corpus and by an estimate of message rate-of-speech. The pause features were binary indicating the presence of a preceding or following pause. Energy, average pitch and delta pitch were all normalized within the message. We used three other pitch features: the range, onset and offset. Two principal lexical features, obtained from the transcript, were used: the collection frequency (inverse document frequency) of words (including a variation based on Porter stemmed words) and a binary feature indicating whether the word (or its stem) was present in a list of named entities. The latter list was derived from a combination of the Voicemail corpus pre-annotated named entities along with entities extracted from the Broadcast News corpus.

We employed a feature selection approach in which we aimed to use the data to guide us to an optimal subset of features. Instead of specifying a single classifier and feature set – optimized for a particular precision/recall tradeoff – we maintain a set of classifiers/feature sets, optimizing for all possible precision/recall tradeoffs. We achieve this by considering the ROC curves of the trained classifiers (with respect to the validation set) and forming the convex hull of those ROC curves. Provost and Fawcett [12] proved that switching between the classifiers corresponding to the vertices enables the attainment of any operating point on the convex hull. The combination of these classifiers is referred to as the maximum realisable ROC (MRROC) classifier and is equal to, or better than, all existing classifiers. Scott, Niranjan and Prager [5] observed that this approach could be used for feature selection, resulting in an algorithm they called Parcel.

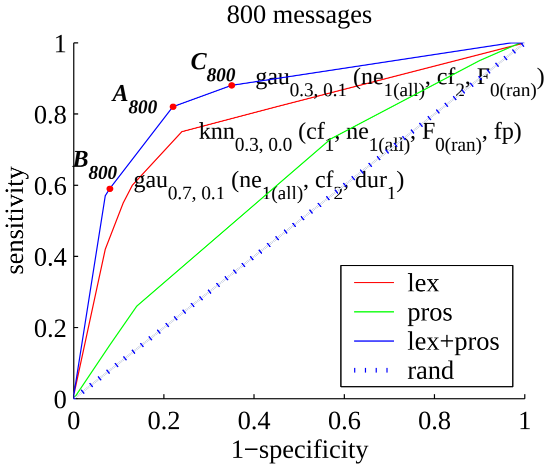


Figure 2: The MRROC curves produced by Parcel on the validation set selecting from lexical only, prosodic only, and lexical and prosodic features. Classifier A (k-nearest neighbour using collection frequency, named entity match, pitch range and following pause) performs best at moderate precision/recall tradeoff; B performs best at high precision; and C performs best at high recall.

Parcel selects those feature subsets and classifiers that extend the MRROC. It does not select a single feature subset (or classifier), rather it selects as many subsets that are required to maximize performance at all operating points. We evaluated features using a sequential forward selection method. Parcel starts by estimating classifiers using single features only, and forms the MRROC. Those classifiers that are vertices of the MRROC are retained. If there are n total features, and a retained classifier uses a subset of k features, then $n - k$ new classifiers are generated, by adding each of the unused features to the feature set. The new classifiers are trained, the MRROC is updated and the process continues. The algorithm terminates when retraining any of the retained classifiers with an additional feature does not extend the MRROC.

Figure 2 shows the MRROC on the validation set selecting from lexical only, prosodic only and all features. Training was performed on the hand transcriptions. Although lexical features alone clearly dominate prosodic features alone, there is a clear benefit to augmenting the lexical features with prosodic features such as pitch range and pause information. We note that named entity match and collection frequency were the most important single features. Given a desired operating point in ROC space, Parcel enables us to choose a classifier that is optimal (with respect to the validation set) for that point.

5. Experiments

The design of the automatic voicemail summarization system for mobile messaging requires trade-offs between

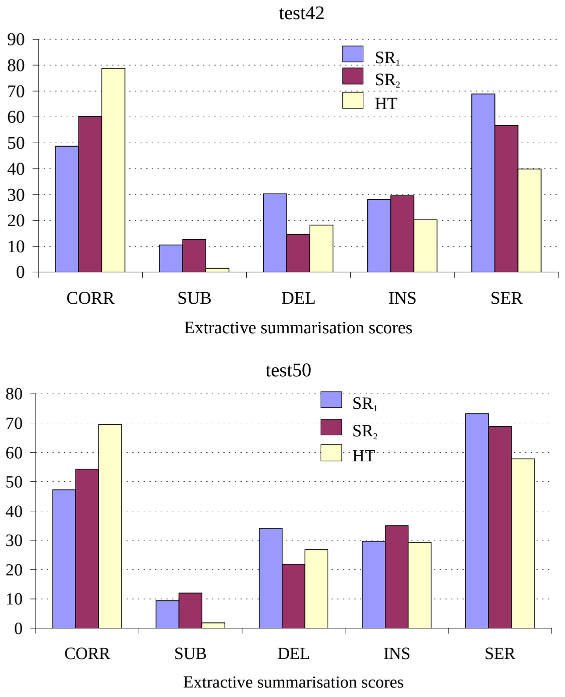


Figure 3: Slot error rate of extractive summarization on test42 and test50 for summaries based on human transcription (HT), SR$_1$ and SR$_2$. CORR refers to correct word transcription and classification.

the target summary length and the retaining of essential content words. The way message transcriptions are processed to construct summaries can affect everything from a user's perception of the service to the allocation and management of the mobile network's resources. Summaries are inherently hard to evaluate because their quality depends both on the intended use and on a number of other factors, such as how readable an individual finds a summary or what information an individual thinks should be included in it. The following experiments were conducted using unseen test data and the questions we are looking to answer are the effects of WER and the effects of automatic summarization.

5.1. Objective evaluation

The slot error rate (SER) treats substitution errors (correct classification, wrong transcription), insertion errors (false positives) and deletion errors (false negatives) equally. Of the classifiers forming the MRROC in Figure 2, classifier A, which has the shortest Euclidean distance from the perfect classifier, is most appropriate if the aim is to minimize SER. Figure 3 shows these errors for summarization using classifier A applied to human, SR$_1$ and SR$_2$ transcriptions for test42 and test50. Summaries based on SR$_1$ result in a higher SER than SR$_2$ summaries, and it can be seen that the additional errors for SR$_1$ are deletions, which may arise due to more summary words being misrecognized. Summaries based on the human transcription have the lowest SER, as expected. Although the deletion rate is higher for the human transcription compared with SR$_2$, recognition errors also give rise to summarization substitutions.

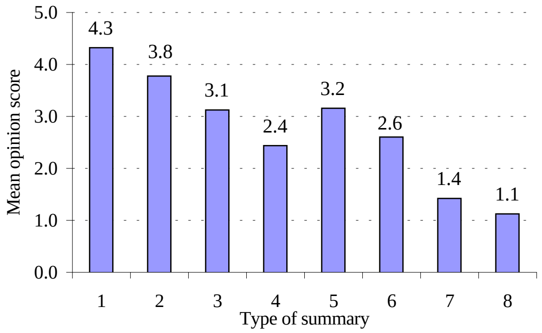


Figure 4: Average MOS on 8 summaries for 5 messages from test42, judged by 10 subjects.

For HT, 80% and 72% correct content and classification was achieved on test42 and test50, respectively. For the SR$_1$ transcriptions, 49% and 47% correct classification was achieved on test42 and test50, respectively. At the same time, for the SR$_2$ transcription scores were consistently higher, 60% and 55% correct content and classification on test42 and test50, respectively. Deletion errors were 26% and 33% for SR$_1$ while for SR$_2$ these were lower at 15% and 22%. SER scores for test50 follow the same patterns with those for test42 while being slightly poorer primarily due to a higher deletions rate as a result of the relatively long duration of the messages contained in the test50.

5.2. Subjective and usability evaluation

The quality of a service is not related to a single measure, but is rather to a combination of several factors, including learnability, effectiveness and user satisfaction. Quality is a property that must be assessed by having representative users interact with each application built. Usability testing ensures that application designs are on target and allow users to accomplish their tasks with ease and efficiency. Poor usability of voicemail summarization applications has a direct cost. Every user who cannot determine the key content from a summary has to retrieve the original audio recording.

We have conducted some subjective and usability tests on the system in a controlled environment. These tests compared manual and automatic summaries presented in random order from human (1, 2), SR$_1$ (3, 4) and SR$_2$ (5, 6) transcriptions, along with the first 30% (7) and a random set (8) of the words of the human transcription. The MOS determined by 10 human subjects for 5 messages summarized in these 8 ways are shown

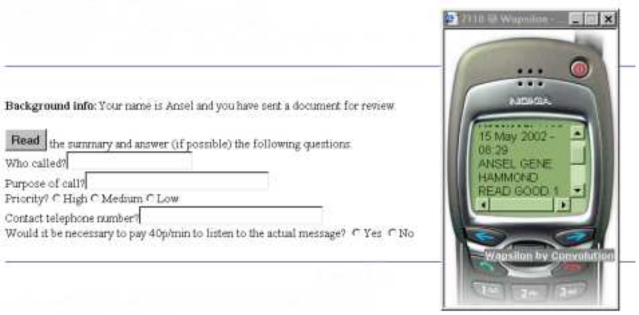
Question	HT	SR_1
caller name	94%	57%
reason for calling	78%	78%
priority	63%	58%
contact number	82%	80%
retrieve audio	*30%*	*53%*

Table 2: Average percentage of correct answers in message comprehension.

in Figure 4. We found that subjects tend to agree more on which summaries are of low rather than high quality and the overall κ statistic was in the range 0.26 to 0.41. The scores indicate that the automatic summaries (2, 4, 6) are considered to be better than selecting the first *n* words (7) or random selection (8), but are inferior to the corresponding human-generated summaries (1, 3, 5). Moving from human to automatic summaries reduces the MOS by about 0.6, whereas moving from a human transcription to a speech recognizer with 30–40% WER reduces the MOS by over 1 point.

A second set of tests aimed to assess the summary quality in terms of comprehension. Subjects answered questions about message content ("caller name?", "reason for calling?", "message priority?", "contact number?") based on the audio and the text summaries. We used a WAP phone emulator to simulate transmitted summaries, and the audiovisual interface is shown in Figure 5. The tests were carried out by 16 subjects who were presented with the summaries and audio of 15 voicemail messages. The summaries used the human and SR_1 transcriptions, and the results are shown in Table 2. Human transcription was considerably more reliable in determining caller identity (94% vs. 57%), but there was less difference in determining the contact phone number (82% vs. 80%). The users were able to determine the reason of calling with equal accuracy (78%) for both types of transcriptions. The above results indicate that summaries produced using automatic transcriptions are particularly beneficial for tasks such as determining the reason for calling, priority of messages and contact numbers. It seems that users were able to associate the words included in summaries to make global judgements about the message content. This evaluation also showed that the users were much more likely to request the message audio, when presented with summaries generated from the speech recognized message, compared with summaries generated from human transcriptions (53% vs. 30%).

Message priority could be determined relatively accurately from the summaries: classifying priority as high/medium/low, the priority obtained from the summary agreed with that obtained from the audio 58% of the time for SR_1 and 63% of the time for human transcriptions. The cases where the subjects completely misjudged the message priority from the text summaries were 2% (judged as high while from the summary they thought it was low) and 5% (judged as low while from the summary they thought it was high). The above results suggest that transcription errors affect mainly the identity of the caller while they lead to 23% more retrievals of audio recordings as users were not confident that the information they read in a summary corresponded to the full and correct content of voicemail messages.



Background info: Your name is Ansel and you have sent a document for review

Read the summary and answer (if possible) the following questions

Who called?

Purpose of call?

Priority? ○ High ○ Medium ○ Low

Contact telephone number?

Would it be necessary to pay 40p/min to listen to the actual message? ○ Yes ○ No

15 May 2002 -
08:29
ANSEL GENE
HAMMOND
READ GOOD 1

Figure 5: Audiovisual interface used for summarization assessment allowing users to access the original audio and the text summaries.

Figure 6 summarizes the time taken by users to answer the comprehension questions about the voicemail messages, comparing summaries based on human and SR_1 transcriptions, and the original audio. Although not directly comparable (since each message was used in one form only), the average comprehension time for speech recognition summaries was about 30% greater than for the human transcription case. These times are about 1.5 times longer than performing the same task using the audio. Note that these figures include the time required to type the answers in the appropriate template fields (Figure 5). This favours the audio retrieval scenario, where users can listen to the recording while typing their answers. At the same time, while retrieving the text summaries they had to browse the mobile display to find the appropriate bit of information prior to typing it. In practice, retrieving the audio would also involve connection overheads, such as typing a PIN. Despite the fact that in the above experiment the digestion of text summaries was not found to be as rapid as that achieved by listening to the audio, the advantages of summarization e.g. indexing and uninterrupted information flow in noisy places need to be considered.

Finally, 13 out of the 16 subjects (81%) that took part in this evaluation would likely use such a service regularly to access their voicemail messages while away from office or home. This suggests that even average quality automatic summaries might be preferable given the elaborate nature of accessing spoken audio.

Engineering-oriented metrics and user input can be correlated with system properties to identify what components of the system affect usability and to predict how

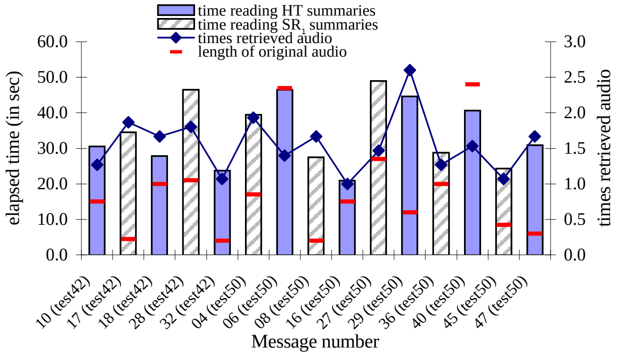


Figure 6: Message comprehension times comparing accessing the original audio to summaries produced from human and SR$_1$ transcripts.

user satisfaction will change when other trade-offs are made [13]. This evaluation framework was extended in [14] with the aim to determine which metrics maximize summary quality and minimize delivery costs within our automatic voicemail summarization system for mobile messaging. One disadvantage of this framework is the amount of data required from subjective evaluations. Instead of solving for weights on the success and cost measures using multivariate linear regression as in [13], one could use Parcel to calculate the role of each metric to the overall system performance. This is a straightforward and possibly much more robust process as the metrics are numerical values that can be used as inputs to simple classifiers that will be trained and validated using task completion as perceived by human subjects as an external criterion.

6. Conclusions

We have evaluated a system for the word-extractive summarization of voicemail, based on the selection of prosodic and lexical features. Speech summarization can be evaluated by many methods and at several levels. We assessed the effect of transcription word error rate, comparing the performance of automatic summarization approaches with respect to transcriptions produced by hand, and by recognizers with average word error rates of 31% and 42%. The summarization slot error rate is dependent on the word error rate, but the difference between the two speech recognition systems is small. We conducted a set of usability tests, using human subjects, based on mean opinion score of summaries, and on a set of comprehension tests. The main results from these experiments were that the automatic summaries were inferior to human summaries, but there was a greater perceived quality difference between summaries derived from hand- and automatically-transcribed messages, than between manual and automatic summarization.

Acknowledgements

We thank Mark Gales (Cambridge University) for providing us with the HTK transcriptions of the voicemail test sets. This work was supported by EPSRC ROPA award GR/R23954.

References

[1] I. Mani. *Automatic Summarization*. John Benjamins Publishing, Amsterdam, The Netherlands, 2001.

[2] K. Zechner. Automatic generation of concise summaries of spoken dialogues in restricted domains. In *Proc. ACM SIGIR*, pages 199–207, New Orleans, LA, USA, 2001.

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

[4] C. Hori and S. Furui. Improvements in automatic speech summarization and evaluation methods. In *Proc. ICSLP*, volume 4, pages 326–329, Beijing, China, 2000.

[5] M. Scott, M. Niranjan, and R. Prager. Parcel: Feature subset selection in variable cost domains. Technical report, CUED TR-323, ftp://svr-ftp.eng.cam.ac.uk/pub/reports, Cambridge, UK, 1998.

[6] K. Koumpis, S. Renals, and M. Niranjan. Extractive summarization of voicemail using lexical and prosodic feature subset selection. In *Proc. Eurospeech*, pages 2377–2380, Aalborg, Denmark, 2001.

[7] K. Koumpis and S. Renals. The role of prosody in a voicemail summarization system. In *Proc. ISCA Workshop on Prosody in Speech Recognition and Understanding*, pages 87–92, Red Bank, NJ, USA, 2001.

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

[9] A. J. Robinson, G. D. Cook, D. P. W. Ellis, E. Fosler-Lussier, S. J. Renals, and D. A. G. Williams. Connectionist speech recognition of broadcast news. *Speech Communication*, 37:27–45, 2002.

[10] G. Saon and M. Padmanabhan. Data-driven approach to designing compound words for continuous speech recognition. *IEEE Trans. Speech and Audio Processing*, 9(4):327–332, 2001.

[11] R. Cordoba, P. C. Woodland, and M. J. F. Gales. Improving cross task performance using MMI training. In *Proc. IEEE ICASSP*, volume 1, pages 85–88, Orlando, FL, USA, 2002.

[12] F. Provost and T. Fawcett. Robust classification for imprecise environments. *Machine Learning*, 42(3):203–231, 2001.

[13] M. A. Walker, D. J. Litman, C. A. Kamm, and A. Abella. Evaluating spoken dialogue agents with PARADISE: Two case studies. *Computer Speech and Language*, 12(3):317–347, 1998.

[14] K. Koumpis. *Automatic Voicemail Summarisation for Mobile Messaging*. PhD thesis, University of Sheffield, UK, 2002.