Automatic summarization of voicemail messages using lexical and prosodic features

Konstantinos Koumpis* Steve Renals†

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

Abstract

This paper presents trainable methods for extracting principal content words from voicemail messages. The short text summaries generated are suitable for mobile messaging applications. The system uses a set of classifiers to identify the summary words, with each word being identified by a vector of lexical and prosodic features. We use an ROC-based algorithm, Parcel, to select input features (and classifiers). We have performed a series of objective and subjective evaluations using unseen data from two different speech recognition systems, as well as human transcriptions of voicemail speech.

keywords: voicemail, automatic speech recognition, automatic summarization, pattern classification, prosody, feature subset selection, receiver operating characteristic, short message service, wireless application protocol, evaluation, usability testing.

1 Introduction

The increased proliferation of audio content has recently motivated several projects in the field of extracting and accessing information from audio archives. Some notable successes have been spoken document retrieval (SDR) and named entity (NE) extraction. A number of SDR systems, operating on an archive of broadcast news, were evaluated as part of the Text REtrieval Conference (TREC) from 1997-2000, giving the important result that retrieval performance on ASR output was similar to that obtained using human-generated reference transcripts, with little or no dependence on transcription errors (Garofolo et al., 2001). This is not the case for all tasks which involve accessing information in spoken audio: it has been observed that the accuracy of NE identification is strongly correlated with the number of transcription errors (Kubala et al., 1998; Gotoh and Renals, 2000; Palmer et al., 2000).

This paper is about the generation of short text summaries of voicemail messages. Automatic summarization may be defined as the distillation of the most important information from a source, producing an abridged version, given a particular user and task (Mani and Maybury, 1999). The majority of research in this area has been concerned with the summarization of written text, reviewed by Mani (2001). The growth of information and communication systems that deal with audio and visual media has stimulated the need to expand

*Now at Domain Dynamics, Reading RG1 1LX, UK
†Now at Centre for Speech Technology Research, University of Edinburgh, Edinburgh EH8 9LW, UK

summarization systems from text to multimedia. For example, the existence of automatic speech summarizers would enable many practical applications such as the construction of automatically annotated audio archives, integrated mixed media communication systems and innovative multimodal interfaces.

A complete speech summarization system demands both spoken language understanding and language generation, and is well beyond the current state-of-the-art. However, it is possible to use simpler techniques to produce summaries that are of some use. The earliest reported work in speech summarization concerned the generation of crude summaries based on acoustic emphasis (Chen and Withgott, 1992) and the classification of parts of dialogue (Rohlicek et al., 1992). More recently, with the advent of large vocabulary speaker-independent continuous ASR, speech summarization research has focused on the application of text-based methods to ASR output (Valenza et al., 1999; Hori and Furui, 2000; Zechner, 2001). At the same time, researchers have begun to combine prosodic, acoustic and language information in an attempt to achieve results that are more robust than those of single sources. Application domains include identification of speech acts (Warnke et al., 1997), sentence and topic segmentation (Hirschberg and Nakatani, 1998; Shriberg et al., 2000) and NE identification (Hakkani-Tür et al., 1999).

Voicemail involves a conversational interaction between a human and a machine, with no feedback from the machine. Voicemail systems can record and store voice messages digitally while the user is away or simply unavailable and can be reviewed upon the user's return. Alternatively, the user can call in on a touch tone phone and review stored messages. Voicemail messages are typically short, conveying the reason for the call, the information that the caller requires from the voicemail recipient and a return telephone number.

The slow, sequential nature of speech makes it hard to find important information quickly. Although, several advances in voicemail retrieval schemes related to pause removal for faster playback and efficient audio coding have been proposed (Kato, 1994; Paksoy et al., 1997), the limitations of the old paradigm remain. Users of voicemail systems on the receipt of a notification have to call their voicemail system and download/listen to their actual/compressed messages. The ScanMail system (Hirschberg et al., 2001) allows users to browse and search the full message transcription of their voicemail messages by content through a graphical user interface. However, voicemail users are likely to want to receive their messages on handheld devices – especially for messages taken by voicemail systems other than the one provided by the network operator, e.g. home or corporate voicemail system. In general, there is a lot of time sensitive content in voicemail, but which the user cannot access either because it is not known when it becomes available (i.e. lack of notification mechanism), or because the notification refers only to changes in status (e.g. arrival of new messages) and not to actual content.

We have proposed an efficient voicemail retrieval scheme (Koumpis et al., 2001a) which 'pushes' text summaries of incoming messages to the handheld device directly from a server without an explicit user request. Figure 1 compares the two approaches for accessing voicemail content. In our architecture the spoken messages collected by the voicemail system are forwarded to the content server where they are automatically transcribed and summarized. There is no restriction on the location of the voicemail system, so access to answering services other than the one provided by the network operator is possible. The message initiator contacts the gateway over the Internet and delivers the messages. The gateway examines the message and performs the required encoding and transformation. The messages are then transmitted hop-by-hop in the mobile network to the mobile client. The message initiator is then notified by the gateway about the final outcome of the operation.

Automatically produced text summaries from voicemail messages may serve multiple goals such as the rapid digest of content, and the indexing of messages with the intention of retrieving the original recordings when more information is needed. Voicemail summarization has several features that differentiate it from conventional text summarization.

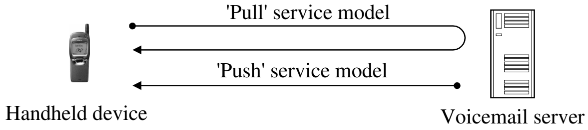

Figure 1: 'Pull' and 'push' service models for accessing voicemail. The 'pull' model employs a conventional request/response approach similar to that of the web – a user enters a URL (the request) which is sent to a server, and the server answers by sending a web page (the response) – while in the 'push' model content is delivered to the handheld device without a specific user request.

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, in this case designed to fit into a 140 character text message and therefore coherence and document flow (style) are less important than content.
3. Only one speaker speaks at a time and due to the relatively short message length, segmentation is unnecessary (in contrast to spoken dialogues or broadcast news).
4. Since the voicemail messages are transcribed by an automatic speech recognition (ASR) system, a significant word error rate (WER) must be assumed.

A number of techniques have been proposed to extract key pieces of information from voicemail messages. Huang et al. (2001) discuss three approaches to extract the identity and phone number of the caller: 200 hand-crafted rules; grammatical inference of subsequential transducers; and log-linear classifiers using a set of 10 000 bigram and trigram features. Jansche and Abney (2002) proposed a phone number extractor based on a two-phase procedure that employed a hand-crafted component derived from empirical data distributions, followed by a decision tree. These techniques rely explicitly on lexical information and the best performing methods are based on hand-crafted rules.

In this paper we present an approach to voicemail summarization based on the extraction of content words from the message transcription. Each word is characterized by a set of lexical and prosodic features, and we have trained classifiers on these feature vectors to discriminate "summary words" from non-summary words. The set of features that we use for the classification was obtained using Parcel (Scott et al., 1998), an ROC-based feature selection methodology. We have carried out a number of experiments using a corpus of Voicemail speech, collected and transcribed by IBM (Padmanabhan et al., 1998), in which the behaviour of our summarization approaches, using speech recognizers with varying error rates, was evaluated using both objective error measurements (with respect to a human generated reference) and subjective user tests.

2 Summarization as a classification problem

We have adopted a *word-extractive* approach to voicemail summarization (Koumpis et al., 2001b), in which a summary is defined as a set of content words extracted from the original message transcription. Given a spoken message $\mathcal{S}$, the word-extractive summarization can be framed as the mapping of each transcribed word into a predefined summary class. This classification problem is hard since there can be a large degree of within-class variability, relative to the between-class variability. Increasing the dimensionality of the feature space can enhance the training set discrimination but at a cost to generalization performance. If a "gold standard" reference is available, with summary class labels for each word, then this approach can be evaluated using standard metrics based on the true positive and true negative rates, also known as sensitivity and specificity:

$$\text{sensitivity} = \frac{\text{TP}}{\text{TP}+\text{FN}} = \text{true positive rate} \tag{1}$$

$$\text{specificity} = \frac{\text{TN}}{\text{TN}+\text{FP}} = \text{true negative rate} \tag{2}$$

$$\tag{3}$$

where TP is the number of true positives (when a word is correctly classified as belonging to a class), TN is the number of true negatives, and FP and FN are the numbers of false positives and false negatives, respectively. A receiver operating characteristic (ROC) curve gives a compound representation of sensitivity and specificity, by plotting sensitivity against [1-specificity] (Zweig and Campbell, 1993; Provost and Fawcett, 2001). For a binary classifier, the sensitivity and specificity are typically controlled by an acceptance threshold: for a strict threshold the sensitivity will be low while the specificity very high. If the threshold is lowered, specificity will fall while sensitivity will rise. In this way we can compare classifiers at particular operating points.

For a given task, two classifiers may be compared using their ROC curves. One classifier dominates another classifier if it has a higher sensitivity at all specificities; in other circumstances one classifier may be more sensitive at some specificities and the other may be more sensitive at others (i.e., the curves cross). To obtain maximal sensitivity at all specificities, Provost and Fawcett (2001) showed that a set of component classifiers could be combined to give a composite classifier whose ROC curve is defined by the convex hull of the component classifier ROC curves. This convex hull is referred to as the maximum realizable ROC (MRROC) curve. Any operating point on the MRROC curve can be achieved by switching between the classifiers corresponding to the vertices of the convex hull.

3 The Voicemail corpus

We have used the IBM Voicemail Corpus-Part I (Padmanabhan et al., 1998), distributed by the Linguistic Data Consortium (LDC). This corpus contains 1801 messages (14.6 hours, averaging about 90 words per message). We used 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 to the training set average of 90 words per message. The messages in this corpus may be categorized as 27% business-related, 25% personal, 17% work-related, 13% technical and 18% in other categories.

We built a hybrid multi-layer perceptron (MLP) / hidden Markov model (HMM) speech recognizer for the voicemail task (Koumpis and Renals, 2000, 2001). The essence of the hybrid approach is to train neural network classifiers to estimate the posterior probability of context independent phone classes, then to use these probabilities (converted into likelihoods by dividing with the priors) as inputs to a HMM decoder (Morgan and Bourlard, 1995). 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 development set, resulting in a 1601 message training set. An initial trigram language model was estimated using the training transcriptions. This training set was augmented with those sentences from the Hub-4 Broadcast News and Switchboard language model training corpora which had a low perplexity with respect to the initial language model, and the language model reestimated. We used a pronunciation dictionary containing around 10 000 words derived from the training data, with pronunciations obtained from the SPRACH broadcast news system (Robinson et al., 2002), plus 1 000 new

	Training	Development	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.

words with pronunciations mainly constructed following the rules used to construct the broadcast news dictionary. The OOV rates were 1.6% on test42 and 2.0% on test50. Additionally we used 32 manually designed compound words (Saon and Padmanabhan, 2001). The average test set WERs were 41.1% on test42 and 43.8% on test50. We denote these transcriptions SR-SPRACH. Additionally, we obtained a second set of transcriptions (denoted SR-HTK) using the more complex HTK Switchboard system, adapted to the Voicemail corpus (Cordoba et al., 2002). The WER for SR-HTK was 31% for both test sets.

We annotated summary words in 1 000 messages of the Voicemail corpus. The first 800 messages were used as a summarization training set, and the last 200 used as the development set. The transcriptions supplied with the Voicemail corpus include marking of NEs, 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;

All annotation was performed using the human transcription only (no audio).

As shown in Table 1 the compression rate in our training, development and testing material was in the range of 19% to 31%. 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:

$$\kappa = \frac{\mathrm{P_o} - \mathrm{P_e}}{1 - \mathrm{P_e}} \tag{4}$$

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

4 Lexical and prosodic features

The architecture of the voicemail summarization system is shown in Figure 2. Lexical information is obtained from the ASR transcriptions, while prosodic features are extracted from audio data using signal processing algorithms or (in the case of pause and durational features) may be extracted by the speech recognizer. Each word in the transcription is represented by a set of lexical and prosodic features (listed in Table 2).

Lexical Features
ac: acoustic confidence
cf_1: collection frequency
cf_2: collection frequency (stem)
$ne_{1(all)}$: all NEs match*
$ne_{2(all)}$: all NEs match (stem)*
$ne_{1(nam)}$: proper names match*
$ne_{2(nam)}$: proper names match (stem)*
$ne_{1(tel)}$: telephone numbers match*
$ne_{2(tel)}$: telephone numbers match (stem)*
$ne_{1(d/t)}$: dates and times match*
$ne_{2(d/t)}$: dates and times match (stem)*
$ne_{1(oth)}$: other NEs match*
$ne_{2(oth)}$: other NEs match (stem)*
pos: word position in message
Prosodic Features
dur_1: duration norm. over corpus
dur_2: duration norm. over message ROS
pp: preceding pause*
fp: succeeding pause*
e: mean RMS energy norm. over message
ΔF_0: delta of F_0 norm. over message
F_0: average F_0 norm. over message
$F_{0(ran)}$: F_0 range
$F_{0(on)}$: F_0 onset
$F_{0(off)}$: F_0 offset

Table 2: Lexical and prosodic features calculated for each word in the voicemail training, development and test sets for the summarization tasks. The features marked with an asterisk (*) are represented by binary variables.

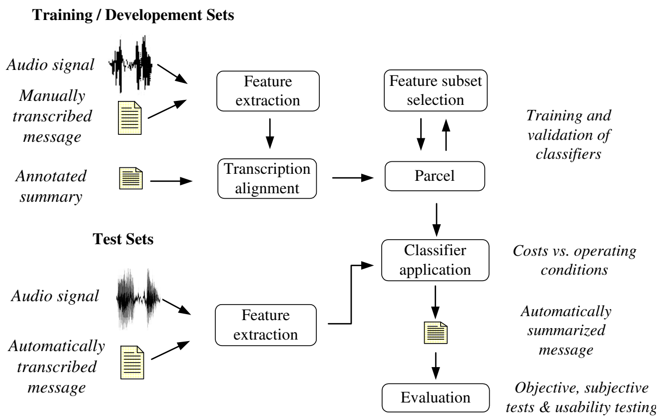

Figure 2: An overview of the word-extractive summarization approach based on systematic comparisons and combination of patterns present in spoken audio.

4.1 Lexical features

For each word in the training, development and test sets we calculated scores corresponding to collection frequency, NE matching, word positioning and acoustic confidence.

4.1.1 Collection frequency

Collection frequency (Robertson and Sparck Jones, 1997) is an information retrieval measure which models the fact that words which occur only in a few messages are likely to be more informative than words which occur often in the entire corpus. For a term $\mathrm{w_i}$ the collection frequency is defined as:

$$\mathrm{CFW_{w_i}} = \log \frac{\mathrm{N}}{\mathrm{n_{w_i}}} \tag{5}$$

where N is the number of messages in the training data and $\mathrm{n_{w_i}}$ is the number of messages that word $\mathrm{w_i}$ occurs in.

4.1.2 Named entity matching

Often the most important pieces of information in a message are the named entities (NEs): people, places, organizations, numbers and dates. Identification of NEs in voicemail is less straightforward than for text. Rather than train or adapt a statistical NE identifier (Gotoh and Renals, 2000) for voicemail, we used matches with an NE list of 3 400 entries, 2 800 of which were derived from the Hub-4 BN corpus (Stevenson and Gaizauskas, 2000), the remainder derived from the Voicemail training data transcriptions.

4.1.3 Word positioning

It is well known (Edmundson, 1969) that the location of terms and sentences within a document can be a good indicator of their relevance to its content. We thus derived a related feature by associating each word in the voicemail transcriptions with a position index which was normalized across messages.

4.1.4 Acoustic confidence

Ideally, we would like to extract only those words that were recognized correctly. Acoustic confidence measures, which may be extracted directly from the acoustic model for MLP/HMM speech recognizers (Williams and Renals, 1999), quantify how well a recognized word matches the acoustic data, given the model.

4.2 Prosodic features

Prosodic features concern the way in which sounds are acoustically realized and can disambiguate a text transcription (e.g. question or statement) or add new information (e.g. the speaker's emotional state). The main focus of existing computational theories of prosody is on stress and intonation, primarily as reflections of the lexical, syntactic and information structures. One such theory developed by Pierrehumbert and colleagues (Pierrehumbert, 1980; Beckman, 1986) has three main distinguishing features. First, it assumes that phrasal intonation is comprised of a string of tones generated by a finite-state automaton. In general, this will consist of an optional boundary tone, a series of pitch accents, a phrase accent, and an optional final boundary tone. The second feature of the theory is the decomposition of the text to be associated with the tune into some metrical representation, indicating stressed and unstressed syllables. The third feature of the theory is the system of rules for associating tune with text. Thus, given some metrical representation of the text and intonational string of tones, there is a mechanism which associates the two. Ladd (1996) made another distinction for intonation, between the contour interaction theories, which treat pitch accents on words as local differences of a global contour for the phrase, and the tonal sequence approaches, which treat phrasal tune as compositional from a sequence of elements associated with the word. Computational theories of prosody however have not yet progressed to a point where interesting generalizations can be made for an engineering approach to voicemail summarization. Hence, we decided to use raw prosodic features without addressing any formal theory of prosody in our modelling.

The manual annotation of prosody can be a very complex task, requiring a great deal of time and training. Most linguistic prosody research still relies heavily on the hand-labelling of speech, augmented by semi-automated computer analysis tools, since this is by far the most accurate way to obtain precise estimates of prosodic features. However, a machine learning approach to automatic speech summarization requires large quantities of data for training purposes, for which prosody can not be expertly transcribed. Using signal processing algorithms or the output of the speech recognizer we automatically extracted and computed the correlates of basic prosodic features associated with each transcribed word. These features can be broadly grouped as referring to pitch, energy, word duration and pauses. Various versions for some features were used and a more detailed description of them follows.

4.2.1 Durational features

The durations of the recognized words and phones may be extracted from the speech recognizer output (assuming that Viterbi decoding is used), and normalized within a message. Phone durations were expressed relative to the expected duration, normalizing to zero mean and unit variance. Word durations were normalized in a similar way, with expected durations computed as a sum of the expected durations of constituent phones (using the pronunciation dictionary). We also extracted rate-of-speech (ROS) information using the enrate tool (Morgan et al., 1997) which calculates the syllable rate based on the computation of the first spectral moment of the low frequency energy waveforms corresponding to a chosen time series segment.

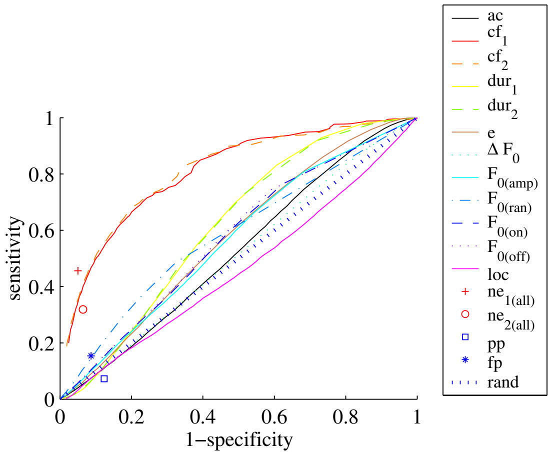

Figure 3: The ROC curves produced by linear classifiers with respect to the development set for voicemail summarization, using the features listed in Table 2 (excluding those referring to class specific NE matching).

4.2.2 Pause features

Typically, pauses reflect the speaker's uncertainty in formulating utterances marking a conflict between speech planning and speech production. ASR systems in general treat silence as an additional subword unit and recognize it in the same way as other phone models. Therefore, from a practical perspective pauses may be seen as the duration of the silence models, which are easily extracted from the recognizer output. Due to the spontaneous nature of speech in Voicemail corpus we decided not to use raw pause durations themselves. Instead we defined binary features for preceding and succeeding pause, which took non-zero values if non-speech regions preceding or succeeding a word exceeded a duration of 30 ms[1]. Although we did not explicitly consider filled pauses, these might be informative about important words (Maclay and Osgood, 1959; Shriberg, 2001).

4.2.3 F_0 features

The fundamental frequency (F_0), was computed using the pda function of the Edinburgh Speech Tools (Taylor et al., 1999). This function implements a super resolution pitch determination algorithm proposed by Medan et al. (1991). To correct for estimation errors, we smoothed the output values using a 5–frame median filter.

We used a number of features derived from the estimate of F_0: the mean, range and slope of the F_0 regression line over a window ranging three frames preceding and following each word; the F_0 onset (the first non zero value in segment); and the F_0 offset (the last non zero value in segment). In case there were not enough F_0 samples in the examined window to calculate an adequate feature value (e.g. for short words such as articles), each missing value was set to the minimum available value from the words in the window's vicinity.

4.2.4 Energy features

Energy features were calculated using the energy function of Edinburgh Speech Tools (Taylor et al., 1999). This function calculates the RMS energy for each frame of the waveform.

[1]The selection of 30 ms as a threshold to identify pauses within a message is somewhat arbitrary and was derived by studying a subset of forced alignments of the training data.

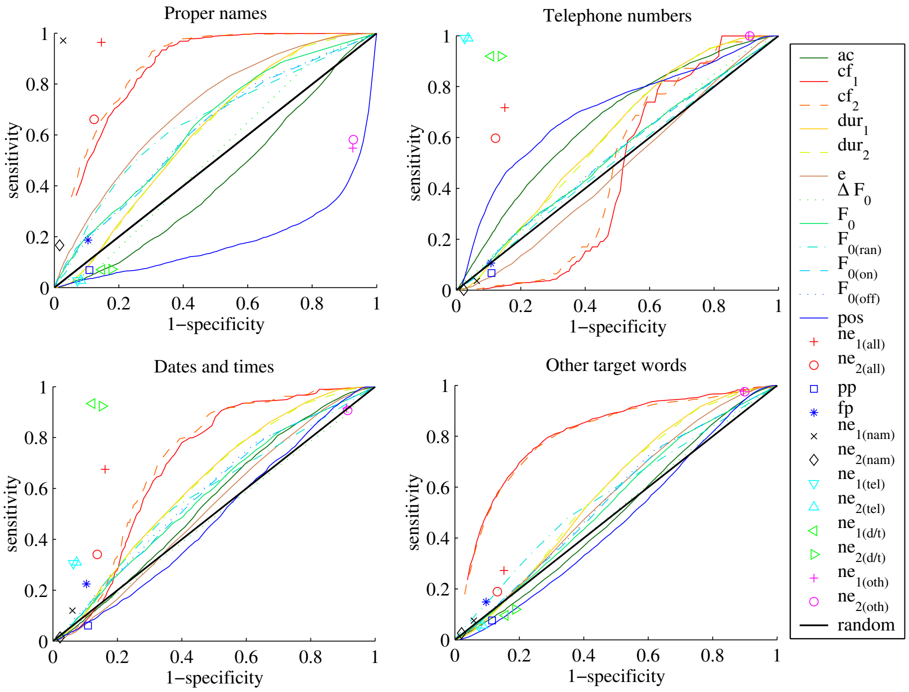

Figure 4: The ROC curves produced by linear classifiers with respect to the development set for the four target classes using the individual features listed in Table 2.

5 Feature selection

Each word in a transcribed voicemail message was represented by a vector of lexical and prosodic features, as described above. Some of these features provide more information for the task at hand than others, and some features may be redundant given other features. In this section we assess the informativeness of these features for the voicemail summarization task first by considering single feature classifiers, then developing optimal feature subsets using an ROC-based algorithm, Parcel.

5.1 Performance of individual features

We investigated the informativeness of each of the lexical and prosodic features listed in Table 2 for the voicemail summarization task by training linear classifiers on each feature in turn.

5.1.1 Single summary class

In Figure 3 we show the ROC curves given by linear classifiers each trained on a single feature, testing on a development set. The best features for extracting summary words were lexical: collection frequency and NE matching. Of the prosodic features, the most important were durational, followed by energy. Features based on F_0 information did not offer significant discrimination, when used alone.

5.1.2 Separate summary classes

In Figure 4 we consider each of the summary classes (names, numbers, dates/times, other) separately and show an ROC curve for each feature and each summary class.

Proper names were identified very accurately by matching to named entity lists. In particular, matching with the unstemmed proper name list resulted in a very high true positive rate with low false positive rate. The unstemmed general NE list also performed well, with stemmed variants being rather less accurate. Collection frequency also offered good discrimination with the stemmed variant performing slightly but consistently better than the unstemmed variant (cf_1). Word position (pos) had strong negative correlation with this summary class, indicating that proper names are mostly positioned at the beginning of voicemail transcriptions where the position features have low values. Regarding the prosodic features, mean RMS energy, features based on F_0 and duration (in descending order) gave useful discrimination. A weak correlation with following pauses (fp) was also observed.

Telephone numbers were also identified accurately by specific named entity lists. The date/time specific named entity lists also matched well for this class (both name lists contain digits). Word position (pos) offered a good discrimination as telephone numbers typically appear towards the end of a message. Collection frequency had an interesting correlation with this class. For words with low collection frequency the correlation was strongly negative, while the correlation was slightly positive for words with a collection frequency above the average. It is also notable that the telephone numbers class had the highest acoustic confidence among all summary classes. Of the prosodic features only the durational ones proved to be correlated with telephone numbers. The rest of prosodic features did not offer any useful discrimination.

The remaining two classes (dates/times and other) were less accurately identified by name matching. For dates and times, the specific named entity list was a good predictor, as were the collection frequency features. The prosodic features were not particularly good predictors for this class, with the best being following pause and the durational features. For the other class, matching to named entity lists was not useful, with the most informative features being the collection frequencies. Among the prosodic features the most useful were the word durations, energy and the F_0 range.

5.2 Selection of multiple features

We used a feature selection approach, in which the data was used to guide us to an optimal feature subset. Instead of demanding a single classifier and feature set (which would be optimized for a particular operating point in ROC space) we adopted an approach that maintained a set of classifiers and feature sets, enabling optimal performance at all points in ROC space. This approach, referred to as Parcel (Scott et al., 1998), builds on the notion of the MRROC curve formed as the convex hull of component ROC curves (section 2).

Parcel is an iterative algorithm that selects those classifiers and feature sets that can extend the MRROC. It does not select a single feature subset (or classifier), but selects as many feature subset/classifier combinations required to maximize performance at all operating points. The operation of Parcel for feature selection is illustrated in Figure 5. In this example, the objective is to find a MRROC for a problem with a data set described by the features $\{a\}$,$\{b\}$ and $\{c\}$. Sequential forward selection (SFS) is used in our implementation to search the feature space but any combinatorial search algorithm could be used instead. SFS starts with an empty set of features and at each iteration adds to the current subset the feature from those remaining that best satisfies the evaluation criterion.

Phase A: estimate single feature classifiers and generate the ROC curves for each candidate feature. For continuous output classifiers vary a threshold over the output range to plot the ROC curve. The $MRROC_{(old)}$ is the diagonal.

Phase B: form the convex hull of the ROC curves and retain those classifiers that correspond to the vertices of the convex hull. If $MRROC_{(new)}$ differs[2] from $MRROC_{(old)}$,

[2] Each new classifier/feature either extends the existing convex hull or does not. The degree of difference

the algorithm proceeds. Set $\mathrm{MRROC}_{(\mathrm{old})}$ equal to $\mathrm{MRROC}_{(\mathrm{new})}$. In the example of Figure 5, as classifiers produce a continuous output to which different thresholds have been applied to predict class membership, the convex hull, $\mathrm{MRROC}_{(\mathrm{new})}$, has five vertices.[3] Two use feature subset $\{\mathrm{b}\}$, and three use $\{\mathrm{a}\}$.

Phase C: for each retained classifier c in the vertices of $\mathrm{MRROC}_{(\mathrm{old})}$ if there are N total features and c has $\mathrm{n_c}$ features, then form $\mathrm{N}-\mathrm{n_c}$ new classifiers, each with $\mathrm{n_c}+1$ features, formed by adding each remaining feature to the input feature set. Generate ROC curves for the new classifiers and recompute the convex hull.

Phase D: retain those classifiers that are used to form the vertices of the convex hull (In Figure 5 two use feature subset $\{\mathrm{a, c}\}$, the others using $\{\mathrm{b}\}$, $\{\mathrm{a, b}\}$ and $\{\mathrm{b, c}\}$). If the new convex hull does extend the old convex hull go to Phase C. Otherwise, terminate and return the set of classifiers that are the vertices of the convex hull.

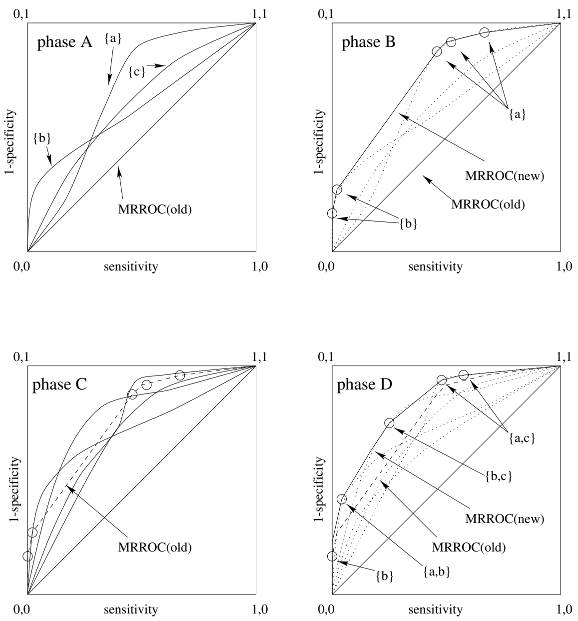

Figure 5: The operation of the Parcel algorithm in searching for the feature subsets that produce the MRROC, after Scott et al. (1998). Only those systems that their operating points lie on the MRROC are saved, as the rest can never be optimal. Clear visual comparisons and sensitivity analysis can be performed at each step of the algorithm's operation.

Using Parcel, it is possible to use multiple classification algorithms and to carry out the search for suitable classifiers to form the MRROC by not only varying the feature subset, but also the classification algorithm. We used five classifiers within this framework: k-nearest neighbour (knn, k=5); Gaussian classifier (gau); single layer network (sln); multi-layer perceptron (mlp); and Fisher linear discriminant (fld).

The training performance of the Parcel algorithm is shown in Figure 6, which graphs the MRROC curves of the development set for each of the classifiers (left), and selecting from lexical only, prosodic only and all features (right). The classifiers in this case were trained on the human transcriptions. The k-nearest neighbours classifier gave very good

is implementation dependent. In our experiments we required a 5% minimum difference for the algorithm to proceed.

[3]The convex hull of a set of points is the smallest convex hull that contains the points.

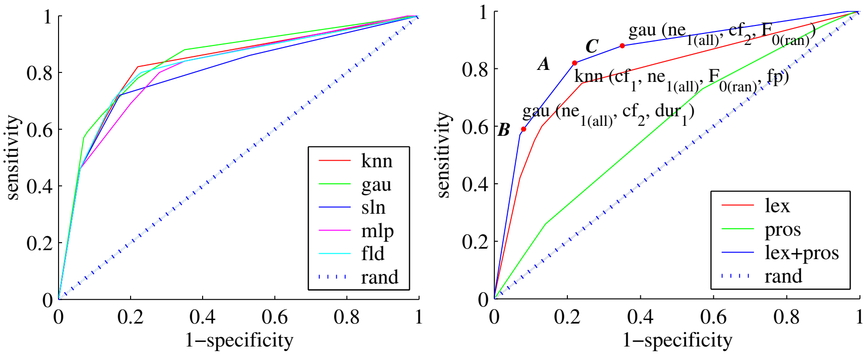

Figure 6: The MRROC curves produced by Parcel on the development set, using the features listed in Table 2 (excluding those referring to class specific NE matching). The left graph compares the role of the five classifiers employed while the right one depicts the MR-ROC produced by all classifiers from lexical only, prosodic only, and lexical and prosodic features. Classifier A is optimal at moderate precision/recall tradeoff; B is optimal at high precision; and C is optimal at high recall.

trade-off between TP and FP for all four sizes of available training data. The Gaussian classifier produces relatively high number of both TP and FP covering a wide range of operating points. Finally, the results from the single layer network were relatively poor.

Although selecting from lexical features alone dominates selecting from prosodic features alone at all operating points, it can be seen that 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 matching 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 development set) for that point.

6 Evaluation

The design of the automatic voicemail summarization system for mobile messaging requires trade-offs between 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 speech recognition WER and of automatic summarization. Speech recognition WER was varied using human transcriptions (denoted SR-Human with 0% WER) and the speech recognition transcriptions described in section 3: SR-SPRACH (41–44% WER) and SR-HTK (31% WER). The effect of automatic summarization was obtained by comparing the automatic system described above with manual summarization, and baseline automatic approaches (random selection of words, and first 30% of the message).

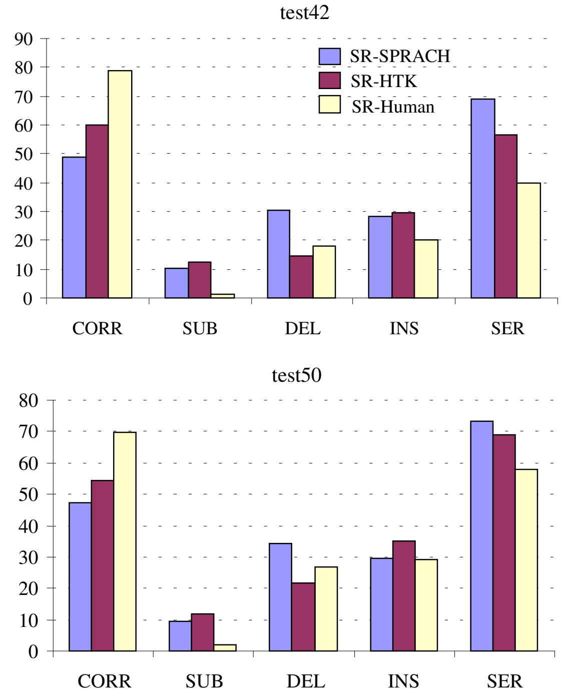

Figure 7: Extractive summarization scores on test42 and test50 for SR-SPRACH, SR-HTK and SR-Human input, respectively.

6.1 Objective evaluation

We have used the slot error rate (SER)(Makhoul et al., 1999) to compare an automatically generated summary against a human generated gold standard. The SER is analogous to the WER, and treats substitution errors (correct classification, wrong transcription), insertion errors (false positives) and deletion errors (false negatives) equally. Of the classifiers forming the MRROC in the right of Figure 6, classifier A (using named entity match, collection frequency, F_0 range and following pause features) was used, since it has the shortest Euclidean distance from the perfect classifier, and is most appropriate if the aim is to minimize SER. Figure 7 shows these errors for summarization using classifier A applied to human (SR-Human), SR-SPRACH and SR-HTK transcriptions for test42 and test50. Increasing speech recognition WER results in an increased SER. The highest WER system, based on SR-SPRACH, has a significantly higher deletion rate compared with SR-Human and SR-HTK which may arise due to more summary words being misrecognized. Recognition errors also give rise to substitutions in the summaries (compared with the gold standard) and this can be seen by comparing the low level of substitutions for the system based on human transcriptions, compared with the systems based on SR-SPRACH and SR-HTK.

For SR-Human, 80% and 72% correct content and classification was achieved on test42 and test50, respectively. For the SR-SPRACH transcriptions, 49% and 47% correct classification was achieved on test42 and test50, respectively. At the same time, for the SR-HTK 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-SPRACH while for SR-HTK 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 short gold standard summaries of the messages contained in the test50.

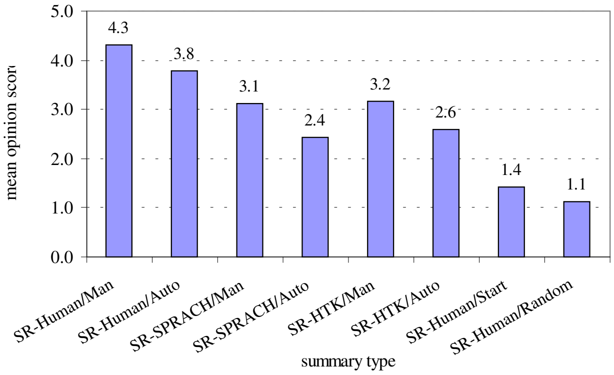

Figure 8: Average MOS on 8 summaries for 5 messages from test42, judged by 10 subjects.

Question	SR-Human	SR-SPRACH
caller name	94%	57%
reason for calling	78%	78%
priority	63%	58%
contact number	82%	80%
retrieve audio	*30%*	*53%*

Table 3: Average percentage of correct answers in message comprehension.

6.2 Subjective and usability evaluation

The quality of a service cannot be represented by a single measure, but it is rather a combination of several factors, including learnability, effectiveness and user satisfaction. Such factors 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. Each time a user cannot determine the key content from a summary, they have 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 SR-Human, SR-SPRACH and SR-HTK transcriptions, along with the first 30% and a random (but sequentially ordered) set of the words in the human transcription. The mean opinion score (MOS) determined by 10 human subjects for 5 messages summarized in these 8 ways are shown in Figure 8. We found that subjects tended 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 are considered to be better than selecting the first 30% of words or random selection, but are inferior to the corresponding human-generated summaries. 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 9. 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-SPRACH transcriptions, and the results are shown in Table 3. Human transcription was considerably more reliable in determining caller identity (94% vs. 57%), but there

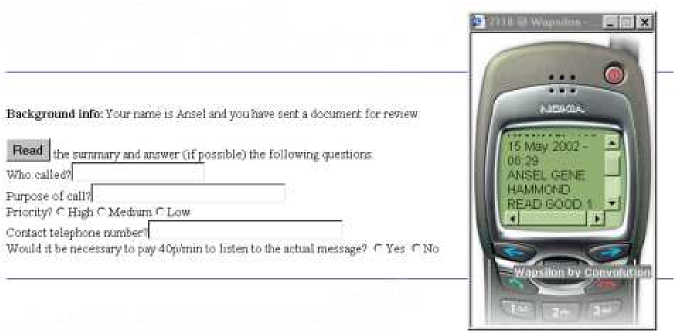

Figure 9: Audiovisual interface used for summarization assessment allowing users to access the original audio and the text summaries.

was less difference in determining the contact phone number (82% vs. 80%). The users were able to determine the reason for calling with equal accuracy (78%) for both types of transcriptions. The above results indicate that summaries produced using automatic transcriptions are particularly useful 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. The above supports the hypothesis that even a few relevant words extracted from a transcription can lead to good message perception and allow potential action to be taken. 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-SPRACH 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.

Figure 10 summarizes the time taken by users to answer the comprehension questions about the voicemail messages, comparing summaries based on human and SR-SPRACH 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 9). 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%) who took part in this evaluation would likely use such a service regularly to access their voicemail messages while away from office or

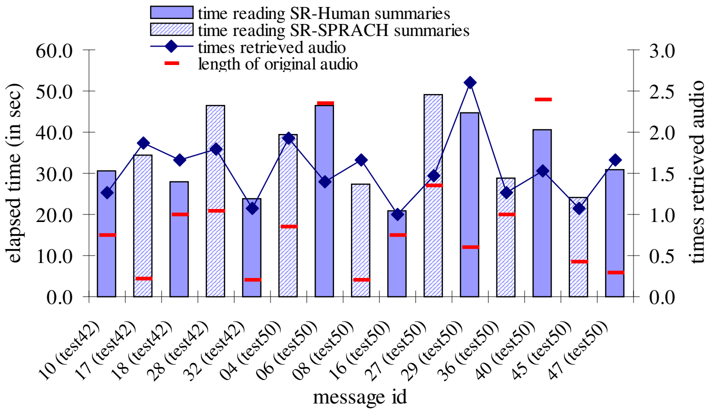

Figure 10: Message comprehension times comparing accessing the original audio to summaries produced from human and SR-SPRACH transcripts.

home. This suggests that even average quality automatic summaries might be preferable given the elaborate nature of accessing spoken audio.

6.3 Discussion

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 user satisfaction will change when other trade-offs are made (Walker et al., 1998). This evaluation framework was extended in Koumpis (2002) with the aim to determine which metrics maximize summary quality and minimize delivery costs within this 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 Walker et al. (1998), 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.

Although treating transcribed words independently proved to work relatively well and allowed us to study the correlation between word classes and a variety of features, it is expected that if modelling is extended beyond the word level classification can be based on the expectations from syntax, semantics and pragmatics and lead to better text coherence. HMMs are a well developed probabilistic tool for modelling sequences of observations, although the amount of annotated data requirements will need to be addressed.

It remains to be seen whether a similar approach can be used to combine acoustic and lexical features to rank messages by accuracy. This would have applications in filtering in order to deliver only the summaries of preselected message types i.e., personal, or professional.

7 Conclusion

In this paper we have presented a framework for voicemail summarization, based on the extraction of words from speech recognition transcriptions. The word extraction process operated by training classifiers to identify words as summary words or not, with each word represented by a vector of lexical an prosodic features. The features used in the summarizer were selected using Parcel, a method based on ROC curves, which returned a collection of feature sets and classifiers which together were optimal at all points in ROC space. Although lexical features (named entity list matching and collection frequency) were most informative, we found that a significant improvement could be observed by augmenting with some prosodic features.

We evaluated the resultant voicemail summarization system through comparison with human-generated gold standard summaries (using slot error rate) and through subjective user testing. We assessed the effect of transcription word error rate, comparing the performance of automatic summarization approaches with respect to transcriptions produced by hand and produced by recognizers with average word error rates of 31% and 42%. The summarization slot error rate was dependent on the word error rate, but the difference between the two speech recognition systems was small; however, the human transcribed system was significantly better. 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 for providing us with the HTK transcriptions of the voicemail test sets and Mark Stevenson for providing the BN corpus derived NE lists. We also acknowledge discussions with Mahesan Niranjan and Rob Gaizauskas. This work was supported by EPSRC ROPA award GR/R23954.

References

Beckman, M. (1986). *Stress and Non-Stress Accent*. Foris Publications, Dordrecht, Holland/Riverton.

Chen, F. and Withgott, M. (1992). The use of emphasis to automatically summarize a spoken discourse. In *Proc. IEEE ICASSP*, volume 1, pages 229–232, San Francisco, CA, USA.

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

Edmundson, H. P. (1969). New methods in automatic extracting. *Journal of the ACM*, 16(2):264–285.

Garofolo, J., Lard, J., and Voorhees, E. (2001). TREC-9 spoken document retrieval track: overview and results. In *Proc. 9th Text Retrieval Conference (TREC-9)*, Gaithersburg, MD, USA.

Gotoh, Y. and Renals, S. (2000). Information extraction from broadcast news. *Philosophical Transactions of the Royal Society of London, Series A*, 358:1295–1310.

Hakkani-Tür, D., Tür, G., Stolcke, A., and Shriberg, E. (1999). Combining words and prosody for information extraction from speech. In *Proc. Eurospeech*, pages 1991–1994, Budapest, Hungary.

Hirschberg, J., Bacchiani, M., Hindle, D., Isenhour, P., Rosenberg, A., Stark, L., Stead, L., Whittaker, S., and Zamchick, G. (2001). SCANMail: Browsing and searching speech data by content. In *Proc. Eurospeech*, Aalborg, Denmark.

Hirschberg, J. and Nakatani, C. (1998). Acoustic indicators of topic segmentation. In *Proc. ICSLP*, volume 4, pages 1255–1258, Sydney, Australia.

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

Huang, J., Zweig, G., and Padmanabhan, M. (2001). Information extraction from voicemail. In *39th Annual Meeting of the Association for Computational Linguistics*, Toulouse, France.

Jansche, M. and Abney, S. (2002). Information extraction from voicemail transcripts. In *Proc. Conference on Empirical Methods in NLP*, Philadelphia, PA, USA.

Kato, Y. (1994). Voice message summary for voice services. In *International Symposium on Speech, Image Processing and Neural Networks*, pages 622–625, Hong-Kong.

Koumpis, K. (2002). *Automatic Voicemail Summarisation for Mobile Messaging*. PhD thesis, University of Sheffield, UK.

Koumpis, K., Ladas, C., and Renals, S. (2001a). An advanced integrated architecture for wireless voicemail retrieval. In *Proc. 15th IEEE International Conference on Information Networking*, pages 403–410, Beppu, Japan.

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

Koumpis, K. and Renals, S. (2001). 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.

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

Kubala, F., Schwartz, R., Stone, R., and Weischedel, R. (1998). Named entity extraction from speech. In *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, VA, USA.

Ladd, D. R. (1996). *Intonational Phonology*. Cambridge University Press, Cambridge, UK.

Maclay, H. and Osgood, C. (1959). Hesitation phenomena in spontaneous english speech. *Word*, 1:19–44.

Makhoul, J., Kubala, F., Schwartz, R., and Weischedel, R. (1999). Performance measures for information extraction. In *Proc. of the DARPA Broadcast News Workshop*, pages 249–252, Herndon, VA, USA.

Mani, I. (2001). *Automatic Summarization*. John Benjamins Publishing, Amsterdam, The Netherlands.

Mani, I. and Maybury, M. (1999). *Advances in automatic text summarization*. MIT Press, Cambridge, MA, USA.

Medan, Y., Yair, E., and Chazan, D. (1991). Super resolution pitch determination of speech signal. *IEEE Trans. Acoustics, Speech and Signal Processing*, 39(1):40–48.

Morgan, N. and Bourlard, H. (1995). An introduction to hybrid HMM/connectionist continuous speech recognition. *IEEE Signal Processing Magazine*, pages 25–42.

Morgan, N., Fosler, E., and Mirghafori, N. (1997). Speech recognition using on-line estimation of speaking rate. In *Proc. Eurospeech*, volume 4, pages 2079–2082, Rhodes, Greece.

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

Paksoy, E., McCree, A., Viswananthan, V., and Linn, J. (1997). 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.

Palmer, D., Ostendorf, M., and Burger, J. D. (2000). Robust information extraction from automatically generated speech transcriptions. *Speech Communication*, 32(1–2):95–109.

Pierrehumbert, J. (1980). *The Phonology and Phonetics of English Intonation*. PhD thesis, MIT, Cambridge, MA, USA.

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

Robertson, S. E. and Sparck Jones, K. (1997). Simple proven approaches to text retrieval. Technical report, TR-356, Cambridge University Computer Laboratory, Cambridge, UK.

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

Rohlicek, J. R., Ayuso, D., Bates, M., Bobrow, R., Boulanger, A., Gish, H., Jeanrenaud, P., Meteer, M., and Siu, M. (1992). Gisting conversational speech. In *Proc. IEEE ICASSP*, volume 2, pages 113–117, San Francisco, CA, USA.

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

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

Shriberg, E. (2001). To “errrr” is human: Ecology and acoustics of speech disfluencies. *Journal of the International Phonetic Association*, 31(1):153–169.

Shriberg, E., Stolcke, A., Hakkani-Tür, D., and Tür, G. (2000). Prosody-based automatic segmentation of speech into sentences and topics. *Speech Communication*, 32(1–2):127–154.

Stevenson, M. and Gaizauskas, R. (2000). Using corpus-derived named lists for named entity recognition. In *Proc. of Applied NLP and the N. American Chapter of the ACL*, pages 290–295, Seattle, WA, USA.

Taylor, P., Caley, R., Black, A. W., and King, S. (1999). Edinburgh speech tools library. Technical report, ftp://ftp.cstr.ed.ac.uk, Edinburgh, UK.

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

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

Warnke, V., Kompe, R., Niemann, H., and Nöth, E. (1997). Integrated dialog act segmentation and classification using prosodic features and language models. In *Proc. Eurospeech*, volume 1, pages 207–210, Rhodes, Greece.

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

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

Zweig, M. H. and Campbell, G. (1993). Receiver-operative characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, 39:561–577.