Extractive Summarization of Voicemail using Lexical and Prosodic Feature Subset Selection

Konstantinos Koumpis, Steve Renals, Mahesan Niranjan

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

Abstract

This paper presents a novel data-driven approach to summarizing spoken audio transcripts utilizing lexical and prosodic features. The former are obtained from a speech recognizer and the latter are extracted automatically from speech waveforms. We employ a feature subset selection algorithm, based on ROC curves, which examines different combinations of features at different target operating conditions. The approach is evaluated on the IBM Voicemail corpus, demonstrating that it is possible and desirable to avoid complete commitment to a single best classifier or feature set.

1. Introduction

There is a growing interest in mobile communications systems that allow users to use their voices to do more than speaking to other people; examples include accessing information services and interaction with booking services. An important issue related to the development of integrated voice/data communications is that of speech summarization: given a spoken passage produce a short, textual précis of its content. We are particularly interested in a system that transmits text summaries of a user's incoming voicemail messages, using the GSM Short Message Service (SMS) [1], reducing the need for users to listen to all of their messages.

Voicemail summarization differs from standard text summarization or abstracting, since it does not assume perfect transcriptions and is concerned with summarizing brief spoken messages (average duration about 40s) into terse (140 character) SMS summaries. Given this level of compression, "document flow" is not important in the summary, compared with the need to transmit the principal content words in the message. This approach assumes that an appropriate summary of a voicemail message may be constructed as a subset of the original message, and that each word may be considered independently.

It has been demonstrated in [2] that a combination of acoustic confidence measures with simple information retrieval techniques can be used to accept/reject words and phrases for inclusion in summaries. This basic architecture – in which the speech is transformed to text and summarized using text processing techniques – was also adopted in [3] and [4]. In the latter paper, we performed voicemail summarization by representing each message as a vector of weighted terms, with the weights derived from collection frequency, named entity (NE) lists and acoustic confidence measures.

Speech is a very rich communication medium and recently there have been efforts to find ways of incorporating information such as prosody in order to extent the capabilities of spoken dialogue and audio browsing/retrieval systems. Humans use prosody to disambiguate similar words, to group words into meaningful phrases, and to mark the importance of words or phrases. Spontaneous speech and read speech differ in regard to prosodic structure, with the former having shorter prosodic units. The acoustic correlates of prosody are among the cues least affected by noise, so it is likely that human listeners use prosody as a redundant cue to help them correctly recognize speech in noisy environments.

Tasks that have attracted research interest include identification of speech acts [5], sentence and topic segmentation [6, 7] and NE extraction [8]. The above approaches combine hidden Markov models (HMM), statistical language models, and prosody-based decision trees. In this paper, we use the Parcel feature subset selection algorithm [9] to evaluate which of the several and often correlated lexical and prosodic features are potentially optimal as classifier inputs for voicemail summarization. Parcel minimizes the management of classifier performance data, facilitates the comparison of a large number of classifiers, and allows clear visual comparisons and sensitivity analysis.

The rest of the paper is structured as follows: in section 2 we describe the experimental data. The prosodic and lexical features are presented in section 3. In section 4 we introduce the Parcel feature subset selection algorithm and its properties for comparing and visualizing classifier performance. A description of the evaluation metric and the summarization results are given in section 5, while the paper is concluded in section 6.

2. Experimental Data

Voicemail speech presents a challenging problem, since it is characterized by a variety of speaking rates, accents, tasks and acoustic conditions. Additionally, phenomena such as disfluencies, restarts, repetitions and broken words are common. In contrast to natural dialogue, voicemail speech is a "one-way" communication: speakers do not receive any direct feedback when they leave messages, resulting in many questions and instructions which are not present in conversational or dictated speech. The telephone channel also poses problems of low bandwidth and signal to noise ratio, since there are no restrictions on the location or type of phone used to leave a voicemail message.

The construction of a supervised classification system for the summarization task, requires a data set of labelled examples with which to train and test the system. The experiments reported in this paper have used as training set manually annotated data corresponding to the first 200 messages in the Voicemail Corpus Part I (distributed by LDC). The annotation of the data was not based only on the extraction of NE but it was rather a selection of any words that are necessary to understand the message content, without having to listen to it. 32% of the words in the corpus were marked as target words.

We have constructed a baseline speech recognizer for the Voicemail task using a hybrid HMM/multi-layer perceptron (MLP) framework along with a combination of perceptual linear prediction and modulation-filtered spectrogram frontends [4]. For testing and evaluation purposes, we use the manually annotated development test set of this corpus comprising 42 messages (test42) and a second test set containing 50 messages (test50) provided by IBM who performed the original data collection [10]. The messages in test50 set are on average twice as long as those in test42. The Word Error Rate (WER) for test42 was 46.5% while for test50 it was 48.2%. These WER figures are not uniform, but they are bursty, both across and within messages and therefore it is possible to perform useful summarization.

3. Feature Description

Lexical information is obtained from the speech recognizer. Prosodic features may be extracted from audio data using signal processing algorithms or the recognizer's acoustic model. Alignment with the corresponding transcription enables the identification of features that correspond to each word in the recognizer's output. The features we used are listed in Table 1.

3.1. Lexical Information

For each word in the training and test sets we calculated scores corresponding to acoustic confidence, collection frequency and NE matching as in [4]. Confidence measures quantify how well a model matches some spoken utterance, where the values are comparable across utterances. A discriminating confidence measure was obtained using a duration normalized sum of log phone posterior probability estimates. Collection Frequency is based on the fact that words which occur only in a few messages are often more likely to be relevant to the topic of that message than ones that occur in many. NE lists derived from Broadcast News data were also employed in order to prioritize words that may be classified as proper names, or as certain other classes such as organization names, dates, times and monetary expressions. In the present work all the NE classes are treated equally.

3.2. Prosodic Information

The prosodic features can be broadly grouped as referring to pitch (mean, range and slope of F0 regression line over the word), energy (mean of RMS energy), duration of the word and pauses (non speech regions exceeding 30 ms preceding and following the word). Duration and pause features were extracted from the acoustic model while pitch and energy features were calculated every 16 ms using `pda` and `energy` functions of the Edinburgh Speech Tools [11] with default settings. The former implements a super resolution pitch determination algorithm and the output values were smoothed using a window ranging three frames before and after the word, and normalized within a message.

4. Feature Subset Selection

Many tens of lexical and prosodic features may be identified. It is desirable to select a subset of these features, thus reducing the effects of the curse of dimensionality and the inclusion of redundant or irrelevant features. In feature selection approaches, features which seem irrelevant for modeling are removed. This is a combinatorial optimization problem. The direct approach (the "wrapper" method) retrains and re-evaluates a given model for many different feature sets. An approximation (the "filter" method) instead optimizes simple criteria which tend to improve performance [12]. The two simplest optimization methods are forward selection (keep adding the best feature) and backward elimination (keep removing the worst feature).

In many applications such as speech summarization, the cost of different types of errors is not known at the time of designing the system. One also can find applications where the costs change over time. Further, some costs cannot be specified quantitatively including coherence degradation, readability deterioration and topical under-representation in speech summarization. Thus, we resort to specifying the classifier in the form of an adjustable threshold and a receiver operating characteristic (ROC) curve obtained by setting the threshold to various possible values [13].

Lexical Features
f1: acoustic confidence
f2: collection frequency
f3: NE matching*
Prosodic Features
f4: duration normalized by corpus
f5: precedent pause*
f6: following pause*
f7: mean RMS energy normalized by message
f8: slope of pitch linear regression normalized by message
f9: average pitch amplitude normalized by message
f10: pitch range

Table 1: *Lexical and Prosodic features calculated for each word the voicemail training and test sets. The features marked with an asterisk (*) are represented by binary variables instead of continuous.*

ROC curves quantify the accuracy of classification systems without regard to the probability distributions of training and test set pattern vectors or decision bias. This measurement system uses a forced classification method for binary outcomes. Two rates can be calculated for any series of classifications: the true-possitive (sensitivity) and the false-positive (1-specificity) rates. A true-positive has occurred when a important word is correctly included in the summary, and a false-positive when a non-important word is incorrectly included in the summary. By varying the level of the threshold, different degrees of true-positive and false-positive rates can be achieved. Given the ROC curve, an end user can pick a point on the curve, that represents an operating classifier with the most desirable true- and false-positive rates.

The ROC curves for the best performing lexical and prosodic features that offer maximum discrimination between words are shown in Figure 1. Among the lexical features, collection frequency is the one with the highest correlation with the target words followed by NE matching. When a speaker leaves a voicemail message there will be prosodic cues that emphasize the important points in the message, beyond the simple lexical content. Considering the prosodic features, the one with the highest correlation between the important words proved to be duration, followed by energy. Pitch information did not offer significant discrimination and this is in accordance with the results presented in [14] where it was shown that pitch relevant features of the syllabic nuclei play a much less important role in the prosodic stress than duration and energy. There is also a very weak correlation of important words and pauses in this task, perhaps due to the spontaneous nature of voicemail speech. An indication that important words tend to precede a pause instead of following a pause remains to be further examined.

4.1. Parcel

In [13], Provost and Fawcett suggested that classifiers may be combined by random switching to achieve any operating point on the convex hull of their ROC curves. Such a combination is

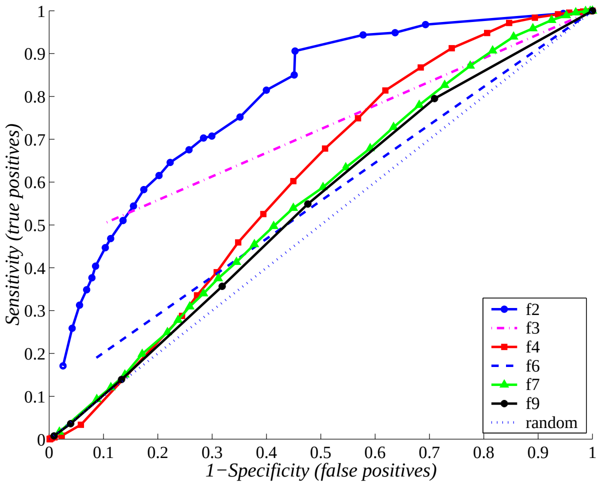

Figure 1: *The ROC curves produced using a single feature with respect to the training set. Only the six best (potentially optimal) features are shown with collection frequency, NE scoring, duration and RMS energy offering maximum discrimination.*

Classifier Type
C1: k-nearest neighbours, $k = 3$
C2: Gaussian
C3: perceptron
C4: MLP comprised 10 hidden units

Table 2: *Classifiers used within the Parcel framework.*

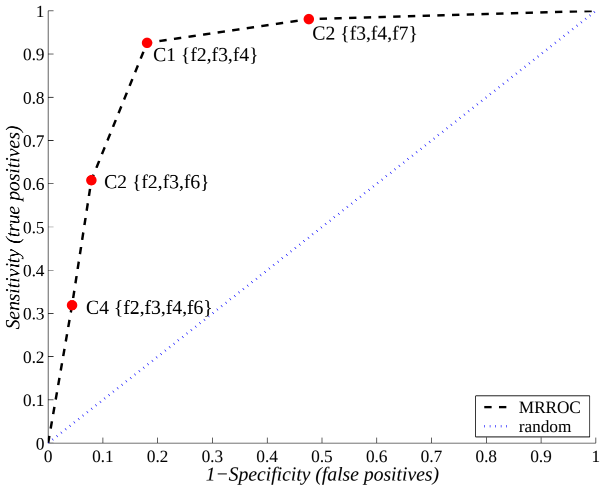

Figure 2: *The ROC produced by Parcel on the training set. Each vertex label indicates the classifier and feature subset used to produce that vertex. Parcel instead of selecting a single best feature subset, selects as many as different subsets are necessary to produce satisfactory performance across* all *target operating conditions.*

referred to as the Maximum Realizable ROC (MRROC) classifier. Subsequentally, Scott, Niranjan and Prager [9] derived the Parcel algorithm that sequentially selects features to maximize the MRROC. It is the objective of Parcel to produce a MRROC that has the largest possible area underneath it, i.e. to maximize the Wilcoxon statistic associated with the classification system defined by the MRROC. This is achieved by searching for, and retaining, those classifiers that extend the convex hull defined by the MRROC. We have applied the Parcel algorithm to lexical and prosodic feature subset selection in the task reported here. Four simple classifiers[1] were implemented for this task (Table 2): a Gaussian classifier; k-nearest neighbours ($k = 3$); a single layer network; and an MLP with 10 hidden units.

Sequential Forward Selection (SFS) was adopted for searching, in which the best single feature is found and taken as the first feature in the subset. Next, each of the other features are evaluated with the first one to find the best two features (retaining the first). This is repeated until the desired number of features have been chosen. One of the most powerful uses of this technique is that the points on the convex hull (realisable classifiers) may be found as combinations of classifiers from the vertices.

Figure 2 depicts the MRROC produced by Parcel on the training set. Four different combinations of features at different target operating conditions are shown. This implies that different trade-offs in the ROC curve require different optimal feature sets. For instance, the feature set {f2,f3,f4,f6} should be used as an input to the MLP in order to obtain the lowest false-positive rate. The Gaussian classifier with feature sets {f2,f3,f6} has almost the same false-positive but significantly better sensitivity. The classifier with the best trade-off between true- and false-positives is a k-nearest neighbours with feature set {f2,f3,f4}. Finally the highest sensitivity and false-positive rate is obtained by using a Gaussian classifier with feature set {f3,f4,f7}.

5. Summarization Performance

Evaluating summaries is not trivial, at least because there is no such thing as the best, or ‘canonical’ summary – especially when the summary is constructed as an extract.

5.1. Error Analysis

A summarizer should not act passively on the transcript it is given by a speech recognizer and therefore a weighted Slot Error Rate (SER) metric was used. As voicemail task involves both transcription and summarization, there are two possible types of error: *content* where an important word has been located but the recognizer has failed to transcribe it correctly and *extent*, where a non-important word has been hypothesized. SER is set to zero if content and extent are all correct, otherwise a 0.5 penalty is added for every content (substitution error) or extent mismatch (insertion error). A word hypothesis w_{hyp} may only be marked as correct extent if an identical word w_{ref} exists in the time-aligned reference transcription such that greater than 50% of the interval spanded by w_{hyp} overlaps with that of w_{ref} and vice versa. The last conditions makes it possible to identify deletion errors. Although the above metric does not forgive recognition errors, it penalizes them partially and therefore it is a good diagnostic while developing a summarization system.

5.2. Summarization Results

After having performed the feature subset selection and chosen the operating points for our trained classifiers we evaluated the summarization performance on the two held-out sets by aligning the content words flagged by the summarizer with those annotated in a human-generated reference transcription. The re-

[1] Although theoretically it is possible to obtain a single optimal subtest, in practice it has been shown that the subset chosen will be highly dependent upon the classifier used [12].

Classifier	C4	C2	C1	C2
Feature set	{f2,f3,f4,f6}	{f2,f3,f6}	{f2,f3,f4}	{f3,f4,f7}
test42				
CORR(%)	41.0	44.3	44.5	40.5
SUB(%)	12.2	14.5	15.3	11.8
DEL(%)	34.6	26.7	24.9	35.8
INS(%)	9.8	10.4	10.4	8.9
SER(%)	56.6	51.6	50.6	56.5
test50				
CORR(%)	38.2	44.3	46.5	40.5
SUB(%)	16.0	14.3	15.1	12.6
DEL(%)	29.8	27.1	23.3	34.3
INS(%)	18.5	18.0	18.3	21.0
SER(%)	64.3	59.4	56.7	67.9

Table 3: *Extractive summarization scores on the two test sets. CORR indicates correct content and correct extent while SUB denotes wrong content and correct extent. DEL indicates words in the reference that failed to be identified by the summarizer as important and INS denotes non-important words that have been included in the summary. SER is equal to the sum of the three types of errors – SUB, DEL and INS.*

sults are given in Table 3, where it is shown that 45% correct content and extend classification was performed with regard to the annotated data across both test sets. Deletions which could be considered as the most crucial type of error count for about 25%. While the SER scores for test50 are substantially poorer by approximately 10% than those for test42, primarily due to a high insertions rate, this is explained by the long duration of the messages contained in the test50. It should be noted that these results are not directly comparable with those we reported for test42 in [4], as in that case summaries were constructed by removing low score words (some having been replaced by their abbreviations) according to some criteria and the SER was calculated over the remainder of the transcription.

Further investigation on which other prosodic features or variations of those already examined are strongly correlated with the important words in spoken messages is in progress. We also plan to annotate more messages from the Voicemail corpus so as to use a larger training set. The latter is crucial, because an increase in the number of features, without increasing the number of training examples, creates a sparsely populated input for the classifiers. It also remains to be seen whether prosodic information can be used to perform message filtering in order to deliver only the summaries of preselected message types i.e. personal, professional or urgent.

6. Concluding Remarks

These initial experiments substantiate the claim that prosodic information can be useful for summarizing spoken audio, particularly when the WER is high. Given the limited amount of training data utilized, the performance for the two held-out test sets shows that this is a promising approach. The configuration with the best trade-off between true- and false-positives proved to be a k-nearest neighbours classifier with collection frequency, NE matching and duration as input features. However, other combinations of classifiers and features should also be considered as the cost of different types of errors can not be easily specified. We strongly believe that significant improvements are possible, for example by incorporating more features and/or classifiers. Parcel is the natural choice for such a complex feature subset selection task.

7. Acknowledgements

We wish to thank Yoshi Gotoh for good discussions and Mukund Padmanabhan (IBM) for making available the second test set. This work is supported by an EPSRC ROPA (GR/R23954/01).

8. References

[1] Koumpis, K., Ladas, C., Renals, S., “An Advanced Integrated Architecture for Wireless Voicemail Data Retrieval”, Proc. ICOIN, pp. 403-410, Beppu, Japan, 2001.

[2] Valenza, R., Robinson, T., Hickey, M., Tucker, R., “Summarization of Spoken Audio through Information Extraction”, Proc. of the ESCA Workshop on Accessing Information in Spoken Audio, pp. 111-116, Cambridge, UK, 1999.

[3] Hori, C., Furui, S., “Improvements in Automatic Speech Summarization and Evaluation Methods”, Proc. ICSLP, Vol. 4, pp. 326-329, Beijing, China, 2000.

[4] Koumpis, K., Renals, S., “Transcription and Summarization of Voicemail Speech”, Proc. ICSLP, Vol. 2, pp. 688-691, Beijing, China, 2000.

[5] Warnke, V., Kompe, R., Niemann, H., Noeth, E., “Integrated Dialog Act Segmentation and Classification using Prosodic Features and Language Models”, Proc. Eurospeech, Vol. 1, pp. 207-210, Rhodes, Greece, 1997.

[6] Hirschberg, J., Nakatani, C., “Acoustic Indicators of Topic Segmentation”, Proc. ICSLP, Vol. 4, pp. 1255-1258, Sydney, Australia, 1998.

[7] Shriberg, E., Stolcke, A., Hakkani-Tür, D., Tür, G., “Prosody-Based Automatic Segmentation of Speech into Sentences and Topics”, Speech Communication, Vol. 32, No. 1-2, pp. 127-154, 2000.

[8] Hakkani-Tür, D., Tür, G., Stolcke, A., Shriberg, E. “Combining Words and Prosody for Information Extraction from Speech”, Proc. Eurospeech, pp. 1991-1994, Budapest, Hungary, 1999.

[9] Scott, M., Niranjan, M., Prager, R., “Parcel: Feature Subset Selection in Variable Cost Domains”, CUED TR-323, Cambridge, UK, 1998, available from ftp://svr-ftp.eng.cam.ac.uk/pub/reports

[10] Padmanabhan, M., Eide, E., Ramabhardan, G., Ramaswany G. Bahl, L., “Speech Recognition Performance on a Voicemail Transcription Task”, Proc. ICASSP, vol. 2, pp. 913-916, Seattle, USA, 1998.

[11] Taylor, P., Caley, R., Black, A. W., King, S., “The Edinburgh Speech Tools Library Version 1.2.0”, available from ftp://ftp.cstr.ed.ac.uk

[12] Kohavi, R., John G., “Wrappers for Feature Subset Selection”, Artificial Intelligence, Vol. 97, No. 1-2, pp. 273-324, 1997.

[13] Provost, F., Fawcett, T., “Robust Classification for Imprecise Environments”, Machine Learning, Vol. 42, No. 3, pp. 203-231, 2001.

[14] Silipo, R., Greenberg, S., “Automatic Transcription of Prosodic Stress for Spontaneous English Discourse”, Proc. ICPhS, San Francisco, USA, 1999.