The Role of Prosody in a Voicemail Summarization System

The Role of Prosody in a Voicemail Summarization System

Konstantinos Koumpis and Steve Renals

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

Abstract

When a speaker leaves a voicemail message there are prosodic cues that emphasize the important points in the message, in addition to lexical content. In this paper we compare and visualize the relative contribution of these two types of features within a voicemail summarization system. We describe the system's ability to generate summaries of two test sets, having trained and validated using 700 messages from the IBM Voicemail corpus. Results measuring the quality of summary artifacts show that combined lexical and prosodic features are at least as robust as combined lexical features alone across all operating conditions.

1. Introduction

Speech is a very rich communication medium and recently there have been efforts to find ways of incorporating prosodic cues in order to extend the capabilities of spoken dialogue and audio browsing/retrieval systems. An important aspect of this approach is the combination of prosodic, acoustic and language information to achieve results that are more robust than those of single sources. Humans use prosody to disambiguate similar words, to group words into meaningful phrases, and to mark the importance of words or phrases. 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 [10]. Spontaneous and read speech differ in regard to prosodic structure, with the former having shorter prosodic units. A corpus-based analysis of prosodic correlates for spontaneous and read speech can be found in [4].

Tasks that have attracted research interest include identification of speech acts [20], sentence and topic segmentation [5, 18] and named entity (NE) extraction [2]. These approaches have combined hidden Markov models (HMMs), statistical language models, and prosody-based decision trees. In this paper, we are concerned with speech summarization, in particular the generation of short text summaries of a user's incoming voicemail messages. This is a potentially important component of integrated voice/data communication, and we have applied such a facility in a Short Message Service (SMS) based system [7].

Voicemail summarization differs from conventional text summarization or abstracting, since it does not assume a perfect transcription and is concerned with summarizing brief spoken messages (average duration about 40s) into terse summaries (140 characters in the case of SMS transmission). Given this level of compression, "document flow" is less important compared with the need to transmit the principal content words in the message. We have assumed 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.

Previously, we applied the Parcel feature subset selection algorithm [17] to evaluate which of the several and often correlated lexical and prosodic features are potentially optimal as classifier inputs for voicemail summarization [9]. In the present paper we extend this approach by using extra features and classifiers. We utilize a larger amount of training data as well as a validation set to compare and visualize the relative contribution of lexical and prosodic features in this task. A simple post-processing algorithm is presented to retain in the summaries information beyond the word level. Finally we evaluate the summarization performance with and without the effects of the speech recognizer and discuss the limitations of the evaluation metric.

The rest of the paper is structured as follows: in section 2 we describe the experimental data and the setup of the speech recognizer. The prosodic and lexical features are presented in section 3. In section 4 we describe 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.

2.1. Training, Validation and Test Data

The construction of a supervised classification system for a summarization task, requires a data set of labeled examples with which to train and test the system. The experiments reported in this paper have used manually annotated data corresponding to the first 500 and last 200 messages in the IBM Voicemail Corpus[1] as a training set. For testing and evaluation purposes, we use the development test set of this corpus comprising 42 messages (test42, 2K words) and a second test set containing 50 messages (test50, 4K words) provided by IBM who performed the original data collection [12]. The messages in the test50 set are on average twice as long as those in test42.

2.2. Annotation of Principal Content Words

The annotation of principal content words in the transcriptions was in part based on the extraction of NEs, along with the selection of additional words necessary for the understanding of the message. Labeling of important words in a message is not an easy task and possibly not very robust. On average 25% of the 70K words in the training, validation and test sets were

[1] http://www.ldc.upenn.edu/Catalog/LDC98S77.html

marked as target words. The annotation of an example message (vm103317) follows. Target words are shown in boldface.

> HI **BLAINE KAREN GATES** JUST WANT TO LET YOU KNOW I HAD TO **MOVE** THE **BIWEEKLY** WITH ASH- WITH **ASHOUK** AND **DRAGUTIN FROM JUNE THIRTIETH MONDAY TO TUESDAY JULY FIRST ELEVEN THIRTY** TO **TWELVE** THE SAME TIME BUT THE NEXT DAY UH I- IT WILL **NOT** HAPPEN ON THE **THIRTIETH** OF **JUNE** WE'RE GOING TO PUT IT **JULY FIRST** THANKS BYE BYE

2.3. Message Transcription

We have constructed a 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 front-ends [8]. The baseline Word Error Rate (WER) for test42 was 46.5% while for test50 it was 48.2%.

Merging confusable words with other words with which they co-occur frequently and modeling coarticulated pronunciation at the boundaries has proven useful for this task [16]. We augmented both the vocabulary and the language model with 32 manually designed compound words specific to voicemail, reducing the WER to 44.4% and 46.6% for test42 and test50, respectively. Before computing the WER all compound words in the reference and decoded transcriptions are replaced with the corresponding sequence of words. We also split all words referring to acronyms to individual words of letters (e.g., C. E. O. instead of CEO). Hence part of the information would be retained in the case that some letters are misrecognized or not included in the summary, giving the user a chance to recognize a familiar acronym even in the presence of some errors.

In an attempt to extend the language model and the vocabulary while keeping the confusability as low as possible, we trained a trigram language model with the available voicemail data and measured the perplexity of every sentence in the Broadcast News and Switchboard corpora [1]. Approximately 1.3K sentences of both corpora that scored the lowest perplexity and contained at least ten words were added to the training data. We then trained a language model with the augmented texts and tested on the Voicemail test set. In all experiments the singleton bigrams and trigrams were excluded and a Witten Bell discount strategy was used. As shown in Table 1, the extended language model and the 2K extra entries in the vocabulary that reduced the OOV rate by 35%, leading to a WER of 41.1% for test42 and 43.8% for test50, respectively. Since the WER is not uniform, but rather bursty across and within messages, it is possible to perform useful summarization.

System Configuration	*test42*	*test50*
Baseline	46.5%	48.2%
Compound words	44.4%	46.6%
Low perplexity sentences	41.1%	43.8%

Table 1: *Improvements in transcription accuracy after augmenting both the language model and vocabulary with task specific compound words and Broadcast News and Switchboard sentences that scored low perplexity with respect to the Voicemail training set language model.*

3. Computation of Features

The system's architecture is shown in Figure 1. Lexical information is obtained from the speech recognizer while prosodic features may be extracted from audio data using signal processing algorithms or the recognizer's acoustic model. Alignment with the transcription enables the identification of features that correspond to each word in the recognizer's output. This information corresponds to the segments for which lexical and prosodic information has to be computed in order to score word hypotheses.

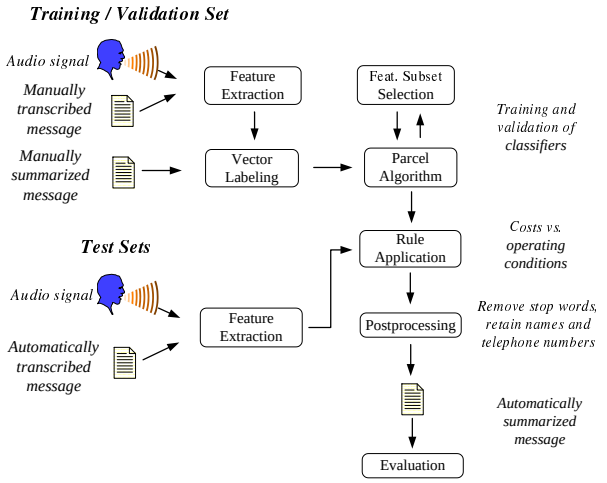

Figure 1: *System's architecture at a glance. Text summaries of spoken messages are constructed using a synchronized combination of prosodic and lexical features.*

The lexical and prosodic features we calculated are listed in Table 2. Features related to pauses and NE matching were treated as binary. The rest were normalized to zero mean and unit variance over the training set.

3.1. Lexical Features

For each word in the training, validation and test sets we calculated scores corresponding to acoustic confidence, collection frequency and NE matching. A description of these features follows:

Acoustic confidence quantifies 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 [21].

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 matching prioritizes words that may be classified as proper names, or as certain other classes such as organization names, dates, times and monetary expressions. In the current configuration all NE classes derived from Broadcast News and Voicemail data are treated equally.

The word lists from all messages were also stemmed using Porter's suffix stripping algorithm [13] and two variations of collection frequency and NE matching features were derived out of them. The algorithm reduces all words with the same root to the same stemmed form (e.g., computer, computation, computing, compute) on a purely lexical basis.

3.2. Prosodic Features

The prosodic features can be broadly grouped as referring to pitch, energy, word duration and pauses. A description of these features follows:

Duration features were extracted from the acoustic model and were normalized by corpus on the phoneme level and by the syllable rate of the particular message. For the rate of speech (ROS) estimation we used ICSI's `enrate` tool [11] which is based on the computation of the first spectral moment of the low frequency energy waveforms corresponding to a chosen time series segment.

Lexical Features
AC: acoustic confidence
CF_1: collection frequency of actual words
CF_2: collection frequency of stemmed words
NE_1: NE matching of actual words*
NE_2: NE matching of stemmed words*
Prosodic Features
DUR_1: duration normalized by corpus
DUR_2: duration normalized by message ROS
PP: preceding pause*
FP: succeeding pause*
E: mean RMS energy normalized by message
DP: delta of pitch normalized by message
PA: average pitch amplitude normalized by message
PR: pitch range
PON: pitch onset
POF: pitch offset

Table 2: *Lexical and Prosodic features calculated for each word in the voicemail training, validation and test sets. The features marked with an asterisk (*) are represented by binary variables (words possessing a property versus those not possessing it).*

Pauses refer to non speech regions exceeding 30 ms preceding and succeeding the word. Currently we do not consider filled pauses which might be informative about important words given that they tend to point to speakers' lexical search problems.

Pitch features are calculated every 16 ms using the `pda` function of Edinburgh Speech Tools [15] with default settings that implements a super resolution pitch determination algorithm. The output values were smoothed using a window ranging three frames preceding and following each word. The mean, range and slope of the pitch regression line over the word, the pitch onset (the first non zero value in segment) and the pitch offset (the last non zero value in segment) were calculated.

Energy features (mean of RMS energy) were calculated every 16 ms using the `energy` function of Edinburgh Speech Tools [15] with default settings.

In case there were not enough pitch or energy samples in the examined window to calculate an adequate feature value (e.g., for short words such as articles), each missing value was replaced by the minimum of the corresponding variable over those words for which a value was available.

4. Feature Subset Selection

Apparently many tens of lexical and prosodic features can be identified and calculated. It is desirable to select a subset of such features and to discard the remainder. This can be useful if there are features which carry little useful information for the particular task, or if there are very strong correlations between sets of inputs so that the same information is repeated in several features. Furthermore, one might wish to reduce the dimensionality simply in order to make the classification calculations quicker, to save storage space or to permit rapid feature extraction.

The feature selection problem can be viewed as a search problem. The search process starts with either an empty set or a full set. The two simplest optimization methods are forward selection (keep adding the best feature) and backward elimination (keep removing the worst feature). An optimal subset is always relative to a certain evaluation function (i.e., an optimal subset chosen using one evaluation function may not be the same as that which uses another evaluation function). Typically, an evaluation function tries to measure the discriminating ability of a feature or a subset to distinguish the different class labels.

Feature selection methods may be classified as filters or wrappers. The direct approach (the wrapper method) retrains and re-evaluates a given model for many different feature sets. An approximation (the filter method), which is independent of the inductive algorithm, instead optimizes simple criteria which tend to improve performance [6]. Decision trees that are widely used to incorporate prosodic features in spoken language systems can be classed as wrappers if the constructed tree is used for classification, or as filters if the tree is used to select features that will subsequently be used for another algorithm.

4.1. Variable Costs and ROC Analysis

In many applications, such as speech summarization, the cost of different types of errors is not known at the time of designing the system. Additionally the costs may change over time. Finally, some costs cannot be specified quantitatively: in speech summarization such costs include coherence degradation, readability deterioration and topical under-representation. 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 [14].

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-positive (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. As one curve can dominate in some interval of thresholds and the other dominates in other intervals, an end user can pick a point on the curve, that represents an operating classifier with the most desirable true- and false-positive rates. A graphical way of finding the optimal performance of a given classifier for specified costs is illustrated in [3].

4.2. Single Feature Comparison

The ROC curves for the best performing lexical and prosodic features that offer maximum discrimination between words are shown in Figure 2. Among the lexical features, collection frequency is the one with the highest correlation with the target words followed by NE matching. We observed a minor improvement in separability offered by collection frequency when it was calculated over stemmed words (CF_2) compared to CF_1. However, the NE matching feature when estimated over stemmed words (NE_2) proved to be worse than NE_1.

Considering the prosodic features, the one with the highest correlation between the important words proved to be durational, followed by energy (E). Normalization of duration by ROS (DUR_2) offered almost identical class separability as the one obtained by DUR_1. Pitch information did not offer significant discrimination and this is in accordance with the results presented in [19] where it was shown that pitch relevant features of the syllabic nuclei play a much less important role in prosodic stress than duration and energy. Pitch range (PR) and pitch amplitude (PA) proved to be the most useful pitch related features. Both pitch onset (PON) and offset (POF) features offered similar and rather low separability in our task. There is also a very weak correlation of important words and pauses in this task, perhaps due to the spontaneous nature of voicemail speech. We found that important words tend to precede pauses instead of succeeding them.

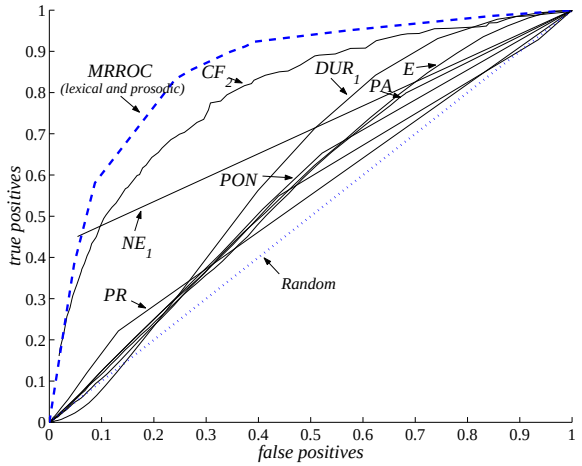

Figure 2: *The ROC curves produced using single features with respect to the validation set. For simplicity only the best (potentially optimal) types of features are shown with collection frequency, NE scoring, duration, energy, pitch onset, pitch amplitute and pitch range offering maximum discrimination.*

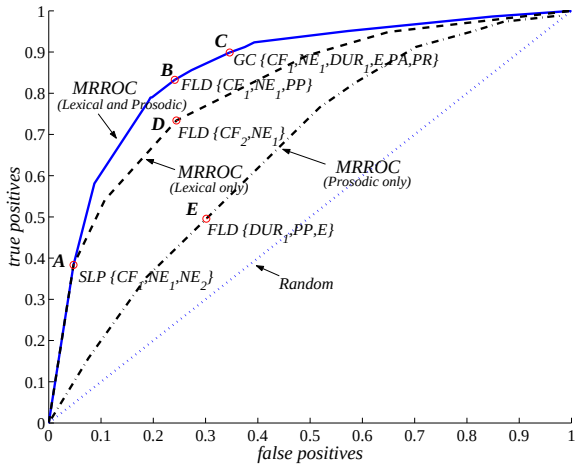

Figure 3: *The MRROC curves produced by Parcel on the validation set using lexical only, prosodic only and combination of lexical and prosodic features. Lexical features as classifier inputs clearly dominate prosodic features in all intervals of thresholds. The combination of lexical and prosodic features gives superior performance than any single constituent classification system.*

4.3. Parcel Algorithm

Classifiers may be combined by random switching to achieve any operating point on the convex hull of their ROC curves [14]. Such a combination is referred to as the Maximum Realizable ROC (MRROC) classifier. Scott, Niranjan and Prager [17] derived the Parcel algorithm that sequentially selects features and classifiers to maximize the MRROC. This implies that different trade-offs in the ROC curve require different optimal feature sets and classifiers. 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 features and classifiers that extend the convex hull defined by the MRROC. The Parcel algorithm seeks not to select a single best feature subset, but rather to select as many as different subsets as are necessary to produce satisfactory performance across all costs.

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. One of the most powerful uses of this algorithm is that the points on the convex hull (realisable classifiers) can be found as combinations of classifiers from the vertices. If the constituent classification rules have produced probabilities, then these can be averaged. For a weighted combination, weights need to be specified.

4.4. Classifiers and Search Method

Although theoretically it is possible to obtain a single optimal subset, in practice it has been shown that the subset chosen will be highly dependent upon the classifier used [6]. The Parcel algorithm requires neither a fixed classifier, nor a single classifier. Five simple classifiers were implemented for this task (Table 3): a k-nearest neighbours ($k = 3$); a Gaussian classifier; a single layer network; an MLP with 20 hidden units and a Fisher linear discriminant. For a comprehensive description of these classifiers see [3].

Sequential forward selection was adopted for searching, in which the best single feature is found and taken as the first feature in the subset. Then all the remaining features are examined to identify that one which, when combined with the first, yields greatest between class-separability. This is repeated, at each step adding that feature which, when combined with those already chosen, leads to the best results. Note that a set of n features chosen in this manner may not be the best set of n. Some potential subsets might not have been examined at all by this procedure.

Classifier Type
KNN: k-nearest neighbours, $k = 3$
GC: Gaussian classifier
SLN: single layer network
MLP: MLP comprised 20 hidden units
FLD: Fisher linear discriminant

Table 3: *Classifiers used within the Parcel framework.*

4.5. MRROC produced by Parcel

Figure 3 depicts the MRROC produced by Parcel on the validation set using lexical only, prosodic only and combination of lexical and prosodic features. Each vertex was produced by using a particular classifier and feature subset. Five different feature subsets at different target operating conditions are shown in detail (A,B,..,E) and will be used to report summarization results on the two test sets in the next section.

Lexical features as classifier inputs clearly dominate prosodic features in all intervals of thresholds. The combination of lexical and prosodic features gives superior performance compared with any single constituent classification system. The lowest false positives rate is achieved using lexical features alone (lower left of ROC curve). Prosodic features extend the system's classification capabilities for true positives of 0.4 and above. In contrast, prosodic features alone tend to produce relatively high false positive and true positive rates (upper right of ROC curve). We also found that every single feature subset corresponding to the MRROC of lexical and prosodic features contains NE matching, while almost all the remaining subsets contain collection frequency features.

test42

			Prosodic and Lexical						***Lexical only***		***Prosodic only***	
System	*Baseline*		*A*		*B*		*C*		*D*		*E*	
	SR	*HT*	*SR*	*HT*	*SR*	*HT*	*SR*	*HT*	*SR*	*HT*	*SR*	*HT*
CORR(%)	23.2	31.9	45.2	69.1	56.3	79.5	48.3	73.5	53.4	77.6	47.1	65.7
SUB(%)	-	-	9.4	0.4	8.9	1.7	10.0	1.0	10.9	1.4	10.4	4.0
DEL(%)	-	-	35.9	30.1	23.9	15.1	31.6	24.6	25.0	17.6	30.0	24.3
INS(%)	-	-	9.5	5.9	15.8	12.7	12.6	8.4	17.8	14.5	18.1	21.4
SER(%)	-	-	54.8	36.4	48.6	29.6	54.2	34.0	53.7	33.5	58.5	49.7

test50

			Prosodic and Lexical						***Lexical only***		***Prosodic only***	
System	*Baseline*		*A*		*B*		*C*		*D*		*E*	
	SR	*HT*	*SR*	*HT*	*SR*	*HT*	*SR*	*HT*	*SR*	*HT*	*SR*	*HT*
CORR(%)	22.9	38.8	34.0	49.2	47.1	65.7	31.7	49.4	43.0	60.9	36.8	50.8
SUB(%)	-	-	7.0	0.2	10.4	4.0	7.6	0.2	7.5	1.3	7.5	2.7
DEL(%)	-	-	52.1	50.4	30.0	24.3	53.1	50.1	40.0	34.4	46.1	41.8
INS(%)	-	-	6.4	2.2	20.1	23.4	7.9	2.2	14.9	12.6	16.9	15.0
SER(%)	-	-	65.5	52.8	60.5	51.7	68.6	52.5	62.4	48.3	70.5	59.5

Table 4: *Extractive summarization scores on the two test sets. CORR refers to correct content and correct extent while SUB denotes wrong content and correct extent. DEL refers to 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. Results are given for both the speech recognition output and the human transcription.*

5. Summarization Performance

Evaluating automatic summarization is hard, not least because there is no such thing as the best, or 'canonical' summary – especially when the summary is constructed as an extract. As we are dealing with imperfect transcriptions, the summarizer should not act passively on the transcript it is given.

5.1. Error Analysis: Strengths and Limitations

A weighted Slot Error Rate (SER) metric was chosen as an appropriate evaluation metric for this task. As a 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. The error for each word in the target summary 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 condition 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.

Despite the fact that the above metric allows summary evaluation in terms of accuracy and completeness by determining whether key content has been correctly transcribed and captured, a drawback can exist whenever content words are repeated several times during a message and fewer instances of them have been identified as target words by the human annotator. This can lead to both deletion and insertion errors even if the correct words appear in both the target and hypothesized summaries.

5.2. Summary Post-processing

In another situation, a hypothesized summary may contain a subset of the true summary, but missing one or two words which can distract the meaning (e.g., a missing 'not' in a statement) or diminish the usability of the extracted information (e.g., a missing digit in a telephone number). In order to overcome the latter problem we have implemented a simple algorithm to post-process the output of the summarizer so as to retain information context beyond the word level.

At the first stage of post-processing all stop words are removed from the summarizer's output. Stop words lack significance to the determination of the content of a message at the rather general level at which message summarization works e.g., 'a', 'an', 'the'. Our stop word list contains 35 entries. Less than 2% of the hypothesized words were stopped. This is a clear indication that our models are well trained to exclude from the summaries frequently occuring and insignificant words.

Subsequently, we search the summarizer's output for proper names and check whether the word preceding and following them is also a proper name. In case the word in the vicinity of a hypothesized proper name is also a proper name and has not been identified as an important word by the summarizer, we include it in the summary given that its acoustic confidence score is above a certain threshold. We repeat the above procedure for acronyms and digits with the search taking place in a wider window so as to retain in the summaries complete acronyms and telephone numbers.

5.3. Results and Discussion

After having performed the feature subset selection and chosen the operating points for our trained classifiers and automatically post-processed the output, 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 results are given in Table 4, where we evaluate the summarization performance with and without the effects of the other component technology, speech recognition.

The columns entitled Baseline show the ratio of words with correct content and correct extent over the number of target words in a message when as summary we consider the first 25% of the words contained in the speech recognizer's output and the human transcription. Any stop words were excluded from these transcriptions prior to generating the summaries. Calculation of other types of errors is not possible from these partial alignments corresponding to the beginning of each message. The comparison of these scores with the ones obtained by the actual systems shows a clear superiority of the latter. For the automatic

transcription 56% and 47% correct content and extent classification was achieved on test42 and test50, respectively while the baseline systems got only 23%. The significant difference between scores when the human transcription is assumed shows what a bottleneck speech recognition can be. Deletions which could be considered as the most crucial type of error count for about 24% and 30% in system B that has the best trade-off between true and false positives using two lexical and one prosodic features as inputs to a Fisher linear discriminant. SER scores for test50 are substantially poorer than those for test42 primarily due to a higher deletions rate as a result of the relatively long duration of the messages contained in the test50.

The post-processing algortihm decreases insertions due to the use of the stop word list. It also tends to decrease deletions and increase correct extent due to the way proper names, acronyms and digits are handled. Finally, substitution score remains unchanged as post-processing does not perform any word replacements. Note that substitution errors for configurations that make use of the human transcriptions are non zero as one would expect. This error is introduced by the method we use to examine word extent based on the 50% requirement of time-aligned reference and hypothesis transcriptions.

Our ultimate goal is to select a dimension-reduced vector of prosodic and lexical features that is adequate for classifying the words contained in spontaneous spoken messages according to their importance with the best trade-off between true- and false positives as defined by an end-user. Work in progress includes evaluation of additional word-level prosodic features as well as extension of the reliability of methods used to extract them. There are still questions as to which is the most effective way to normalize individual raw prosodic features by word, by message/speaker and by corpus. Finally, the shortcomings of the objective evaluation based on the SER measure points to the need for a novel summary evaluation framework that will incorporate factors that are used by subjective methods.

6. Conclusion

The design and performance of a voicemail summarization system under development are presented. The system integrates in a transparent way multiple sources of knowledge encoded as lexical and prosodic features at the word level to generate terse summaries. We have described our message summarization approach and discussed the challenges of automatic speech transcription, extraction of verbal and non-verbal cues and classifier/feature subset selection process that characterize the key content words in spoken messages. We believe that a number of additional cues and structural information can be extracted automatically, allowing the construction of concise text voicemail summaries.

7. Acknowledgements

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

8. References

[1] Fetter, P., Kaltenmeier, A., Kuhn, T., Regel-Brietzmann, P., “Improved Modeling of OOV Words in Spontaneous Speech”, Proc. ICASSP, Vol. 1, pp. 534-537, Atlanta, USA, 1998.

[2] 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.

[3] Hand, D., J., “Construction and Assessment of Classification Rules”, John Wiley and Sons, 1997.

[4] Hirschberg, J., “A Corpus-based Approach to the Study of Speaking Style”, pp. 335-350 in Horne, M., (ed.), Prosody: Theory and Experiment, Kluwer Academic Publishers, The Netherlands, 2000.

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

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

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

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

[9] Koumpis, K., Renals, S., Niranjan, M., “Extractive Summarization of Voicemail using Lexical and Prosodic Feature Subset Selection”, Proc. Eurospeech, Vol. 5, pp. 2377-2380, Aalborg, Denmark, 2001.

[10] Lindfield, K., C., Wingfield, A., Goodglass, H., “The Role of Prosody in the Mental Lexicon”, Brain and Language, Vol. 68, No. 1-2, pp. 312-317, 1999.

[11] Morgan, N., Fosler, E., Mirghafori, N., “Speech Recognition Using On-Line Estimation of Speaking Rate”, Proc. Eurospeech, Vol. 4, pp. 2079-2082, Rhodes, Greece, 1997.

[12] 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.

[13] Porter, M., F., “An Algorithm for Suffix Stripping”, Program, Vol. 14, No. 3, pp. 130-137, 1980.

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

[15] 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

[16] Saon, G., Padmanabhan, M., “Data-Driven Approach to Designing Compound Words for Continuous Speech Recognition”, IEEE Trans. on Speech and Audio Processing, Vol. 9, No. 4, pp. 327-332, 2001.

[17] 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

[18] 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.

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

[20] 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.

[21] Williams, G., Renals, S., “Confidence Measures from Local Posterior Probability Estimates”, Computer Speech and Language, Vol. 13, No. 4, pp. 395-411, 1999.