Summarizing multiple spoken documents: finding evidence from untranscribed audio

Xiaodan Zhu, Gerald Penn and Frank Rudzicz
University of Toronto
10 King's College Rd.,
Toronto, M5S 3G4, ON, Canada
{xzhu,gpenn,frank}@cs.toronto.edu

Abstract

This paper presents a model for summarizing multiple untranscribed spoken documents. Without assuming the availability of transcripts, the model modifies a recently proposed unsupervised algorithm to detect re-occurring acoustic patterns in speech and uses them to estimate similarities between utterances, which are in turn used to identify salient utterances and remove redundancies. This model is of interest due to its independence from spoken language transcription, an error-prone and resource-intensive process, its ability to integrate multiple sources of information on the same topic, and its novel use of acoustic patterns that extends previous work on low-level prosodic feature detection. We compare the performance of this model with that achieved using manual and automatic transcripts, and find that this new approach is roughly equivalent to having access to ASR transcripts with word error rates in the 33–37% range without actually having to do the ASR, plus it better handles utterances with out-of-vocabulary words.

1 Introduction

Summarizing spoken documents has been extensively studied over the past several years (Penn and Zhu, 2008; Maskey and Hirschberg, 2005; Murray et al., 2005; Christensen et al., 2004; Zechner, 2001). Conventionally called *speech summarization*, although *speech* connotes more than spoken documents themselves, it is motivated by the demand for better ways to navigate spoken content and the natural difficulty in doing so — speech is inherently more linear or sequential than text in its traditional delivery.

Previous research on speech summarization has addressed several important problems in this field (see Section 2.1). All of this work, however, has focused on single-document summarization and the integration of fairly simplistic acoustic features, inspired by work in descriptive linguistics. The issues of navigating speech content are magnified when dealing with larger collections — multiple spoken documents on the same topic. For example, when one is browsing news broadcasts covering the same events or call-centre recordings related to the same type of customer questions, content redundancy is a prominent issue. Multi-document summarization on written documents has been studied for more than a decade (see Section 2.2). Unfortunately, no such effort has been made on audio documents yet.

An obvious way to summarize multiple spoken documents is to adopt the transcribe-and-summarize approach, in which automatic speech recognition (ASR) is first employed to acquire written transcripts. Speech summarization is accordingly reduced to a text summarization task conducted on error-prone transcripts.

Such an approach, however, encounters several problems. First, assuming the availability of ASR is not always valid for many languages other than English that one may want to summarize. Even when it is, transcription quality is often an issue— training ASR models requires collecting and annotating corpora on specific languages, dialects, or even different domains. Although recognition errors do not significantly impair extractive summarizers (Christensen et al., 2004; Zhu and Penn, 2006), error-laden transcripts are not necessarily browseable if recognition errors are higher than certain thresholds (Munteanu et al., 2006). In such situations, audio summaries are an alternative when salient content can be identified directly from untranscribed audio. Third, the underlying paradigm of most ASR models aims to solve a

Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, pages 549–557,
Suntec, Singapore, 2-7 August 2009.

classification problem, in which speech is segmented and classified into pre-existing categories (words). Words not in the predefined dictionary are certain to be misrecognized without exception. This out-of-vocabulary (OOV) problem is unavoidable in the regular ASR framework, although it is more likely to happen on salient words such as named entities or domain-specific terms.

Our approach uses acoustic evidence from the untranscribed audio stream. Consider text summarization first: many well-known models such as MMR (Carbonell and Goldstein, 1998) and MEAD (Radev et al., 2004) rely on the reoccurrence statistics of words. That is, if we switch any word w_1 with another word w_2 across an entire corpus, the ranking of extracts (often sentences) will be unaffected, because no word-specific knowledge is involved. These models have achieved state-of-the-art performance in transcript-based speech summarization (Zechner, 2001; Penn and Zhu, 2008). For spoken documents, such reoccurrence statistics are available directly from the speech signal. In recent years, a variant of dynamic time warping (DTW) has been proposed to find reoccurring patterns in the speech signal (Park and Glass, 2008). This method has been successfully applied to tasks such as word detection (Park and Glass, 2006) and topic boundary detection (Malioutov et al., 2007).

Motivated by the work above, this paper explores the approach to summarizing multiple spoken documents directly over an untranscribed audio stream. Such a model is of interest because of its independence from ASR. It is directly applicable to audio recordings in languages or domains when ASR is not possible or transcription quality is low. In principle, this approach is free from the OOV problem inherent to ASR. The premise of this approach, however, is to reliably find reoccuring acoustic patterns in audio, which is challenging because of noise and pronunciation variance existing in the speech signal, as well as the difficulty of finding alignments with proper lengths corresponding to words well. Therefore, our primary goal in this paper is to empirically determine the extent to which acoustic information alone can effectively replace conventional speech recognition with or without simple prosodic feature detection within the multi-document speech summarization task. As shown below, a modification of the Park-Glass approach amounts to the efficacy of a 33-37% WER ASR engine in the domain of multiple spoken document summarization, and also has better treatment of OOV items. Park-Glass similarity scores by themselves can attribute a high score to distorted paths that, in our context, ultimately leads to too many false-alarm alignments, even after applying the distortion threshold. We introduce additional distortion penalty and subpath length constraints on their scoring to discourage this possibility.

2 Related work

2.1 Speech summarization

Although *abstractive summarization* is more desirable, the state-of-the-art research on speech summarization has been less ambitious, focusing primarily on *extractive summarization*, which presents the most important $N\%$ of words, phrases, utterances, or speaker turns of a spoken document. The presentation can be in transcripts (Zechner, 2001), edited speech data (Furui et al., 2003), or a combination of these (He et al., 2000). Audio data amenable to summarization include meeting recordings (Murray et al., 2005), telephone conversations (Zhu and Penn, 2006; Zechner, 2001), news broadcasts (Maskey and Hirschberg, 2005; Christensen et al., 2004), presentations (He et al., 2000; Zhang et al., 2007; Penn and Zhu, 2008), etc.

Although extractive summarization is not as ideal as abstractive summarization, it outperforms several comparable alternatives. Tucker and Whittaker (2008) have shown that extractive summarization is generally preferable to *time compression*, which speeds up the playback of audio documents with either fixed or variable rates. He et al. (2000) have shown that either playing back important audio-video segments or just highlighting the corresponding transcripts is significantly better than providing users with full transcripts, electronic slides, or both for browsing presentation recordings.

Given the limitations associated with ASR, it is no surprise that previous work (He et al., 1999; Maskey and Hirschberg, 2005; Murray et al., 2005; Zhu and Penn, 2006) has studied features available in audio. The focus, however, is primarily limited to prosody. The assumption is that prosodic effects such as stress can indicate salient information. Since a direct modeling of complicated compound prosodic effects like stress is dif-

ficult, they have used basic features of prosody instead, such as pitch, energy, duration, and pauses. The usefulness of prosody was found to be very limited by itself, if the effect of utterance length is not considered (Penn and Zhu, 2008). In multiple-spoken-document summarization, it is unlikely that prosody will be more useful in predicating salience than in single document summarization. Furthermore, prosody is also unlikely to be applicable to detecting or handling redundancy, which is prominent in the multiple-document setting.

All of the work above has been conducted on single-document summarization. In this paper we are interested in summarizing multiple spoken documents by using reoccurrence statistics of acoustic patterns.

2.2 Multiple-document summarization

Multi-document summarization on written text has been studied for over a decade. Compared with the single-document task, it needs to remove more content, cope with prominent redundancy, and organize content from different sources properly. This field has been pioneered by early work such as the SUMMONS architecture (Mckeown and Radev, 1995; Radev and McKeown, 1998). Several well-known models have been proposed, i.e., MMR (Carbonell and Goldstein, 1998), multiGen (Barzilay et al., 1999), and MEAD (Radev et al., 2004). Multi-document summarization has received intensive study at DUC. [1] Unfortunately, no such efforts have been extended to summarize multiple spoken documents yet.

Abstractive approaches have been studied since the beginning. A famous effort in this direction is the information fusion approach proposed in Barzilay et al. (1999). However, for error-prone transcripts of spoken documents, an abstractive method still seems to be too ambitious for the time being. As in single-spoken-document summarization, this paper focuses on the extractive approach.

Among the extractive models, MMR (Carbonell and Goldstein, 1998) and MEAD (Radev et al., 2004), are possibly the most widely known. Both of them are linear models that balance salience and redundancy. Although in principle, these models allow for any estimates of salience and redundancy, they themselves calculate these scores with word reoccurrence statistics, e.g., `tf.idf`, and yield state-of-the-art performance. MMR iteratively selects sentences that are similar to the entire documents, but dissimilar to the previously selected sentences to avoid redundancy. Its details will be revisited below. MEAD uses a redundancy removal mechanism similar to MMR, but to decide the salience of a sentence to the whole topic, MEAD uses not only its similarity score but also sentence position, e.g., the first sentence of each new story is considered important. Our work adopts the general framework of MMR and MEAD to study the effectiveness of the acoustic pattern evidence found in untranscribed audio.

[1] http://duc.nist.gov/

3 An acoustics-based approach

The acoustics-based summarization technique proposed in this paper consists of three consecutive components. First, we detect acoustic patterns that recur between pairs of utterances in a set of documents that discuss a common topic. The assumption here is that lemmata, words, or phrases that are shared between utterances are more likely to be acoustically similar. The next step is to compute a relatedness score between each pair of utterances, given the matching patterns found in the first step. This yields a symmetric relatedness matrix for the entire document set. Finally, the relatedness matrix is incorporated into a general summarization model, where it is used for utterance selection.

3.1 Finding common acoustic patterns

Our goal is to identify subsequences within acoustic sequences that appear highly similar to regions within other sequences, where each sequence consists of a progression of overlapping 20ms vectors (*frames*). In order to find those shared patterns, we apply a modification of the segmental dynamic time warping (SDTW) algorithm to pairs of audio sequences. This method is similar to standard DTW, except that it computes multiple constrained alignments, each within predetermined bands of the similarity matrix (Park and Glass, 2008).[2] SDTW has been successfully applied to problems such as topic boundary detection (Malioutov et al., 2007) and word detection (Park and Glass, 2006). An example application of SDTW is shown in Figure 1, which shows the results of two utterances from the TDT-4 English dataset:

[2] Park and Glass (2008) used Euclidean distance. We used cosine distance instead, which was found to be better on our held-out dataset.

I: the explosion in aden harbor *killed seventeen* u.s. *sailors* and injured other thirty nine last month.

II: *seventeen sailors* were *killed*.

These two utterances share three words: *killed*, *seventeen*, and *sailors*, though in different orders. The upper panel of Figure 1 shows a matrix of frame-level similarity scores between these two utterances where lighter grey represents higher similarity. The lower panel shows the four most similar shared subpaths, three of which correspond to the common words, as determined by the approach detailed below.

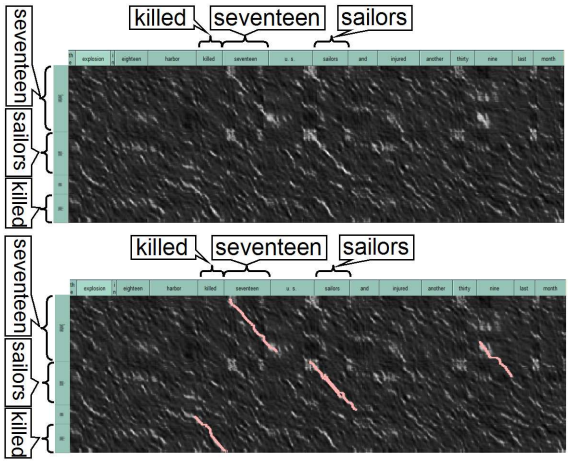

Figure 1: Using segmental dynamic time warping to find matching acoustic patterns between two utterances.

Calculating MFCC

The first step of SDTW is to represent each utterance as sequences of Mel-frequency cepstral coefficient (MFCC) vectors, a commonly used representation of the spectral characteristics of speech acoustics. First, conventional short-time Fourier transforms are applied to overlapping 20ms Hamming windows of the speech amplitude signal. The resulting spectral energy is then weighted by filters on the Mel-scale and converted to 39-dimensional feature vectors, each consisting of 12 MFCCs, one normalized log-energy term, as well as the first and second derivatives of these 13 components over time. The MFCC features used in the acoustics-based approach are the same as those used below in the ASR systems.

As in (Park and Glass, 2008), an additional whitening step is taken to normalize the variances on each of these 39 dimensions. The similarities between frames are then estimated using cosine distance. All similarity scores are then normalized to the range of $[0, 1]$, which yields similarity matrices exemplified in the upper panel of Figure 1.

Finding optimal paths

For each similarity matrix obtained above, local alignments of matching patterns need to be found, as shown in the lower panel of Figure 1. A single global DTW alignment is not adequate, since words or phrases held in common between utterances may occur in any order. For example, in Figure 1 *killed* occurs before all other shared words in one document and after all of these in the other, so a single alignment path that monotonically seeks the lower right-hand corner of the similarity matrix could not possibly match all common words. Instead, multiple DTWs are applied, each starting from different points on the left or top edges of the similarity matrix, and ending at different points on the bottom or right edges, respectively. The width of this diagonal band is proportional to the estimated number of words per sequence.

Given an M-by-N matrix of frame-level similarity scores, the top-left corner is considered the origin, and the bottom-right corner represents an alignment of the last frames in each sequence. For each of the multiple starting points $p_0 = (x_0, y_0)$ where either $x_0 = 0$ or $y_0 = 0$, but not necessarily both, we apply DTW to find paths $P = p_0, p_1, ..., p_K$ that maximize $\sum_{0 \leq i \leq K} sim(p_i)$, where $sim(p_i)$ is the cosine similarity score of point $p_i = (x_i, y_i)$ in the matrix. Each point on the path, p_i, is subject to the constraint $|x_i - y_i| < T$, where T limits the distortion of the path, as we determine experimentally. The ending points are $p_K = (x_K, y_K)$ with either $x_K = N$ or $y_K = M$. For considerations of efficiency, the multiple DTW processes do not start from every point on the left or top edges. Instead, they skip every T such starting points, which still guarantees that there will be no blind-spot in the matrices that are inaccessible to all DTW search paths.

Finding optimal subpaths

After the multiple DTW paths are calculated, the optimal subpath on each is then detected in order to find the local alignments where the similarity is maximal, which is where we expect actual matched phrases to occur. For a given path $P = p_0, p_2, ..., p_K$, the optimal subpath is defined to be a continuous subpath, $P^* = p_m, p_{m+1}..., p_n$

that maximizes $\frac{\sum_{m \le i \le n} sim(p_i)}{n-m+1}$, $0 \le n \le m \le k$, and $m - n + 1 \ge L$. That is, the subpath is at least as long as L and has the maximal average similarity. L is used to avoid short alignments that correspond to subword segments or short function words. The value of L is determined on a development set.

The version of SDTW employed by (Malioutov et al., 2007) and Park and Glass (2008) employed an algorithm of complexity $O(Klog(L))$ from (Lin et al., 2002) to find subpaths. Lin et al. (2002) have also proven that the length of the optimal subpath is between L and $2L - 1$, inclusively. Therefore, our version uses a very simple algorithm—just search and find the maximum of average similarities among all possible subpaths with lengths between L and $2L - 1$. Although the theoretical upper bound for this algorithm is $O(KL)$, in practice we have found no significant increase in computation time compared with the $O(Klog(L))$ algorithm—L is actually a constant for both Park and Glass (2008) and us, it is much smaller than K, and the $O(Klog(L))$ algorithm has (constant) overhead of calculating right-skew partitions.

In our implementation, since most of the time is spent on calculating the average similarity scores on candidate subpaths, all average scores are therefore pre-calculated incrementally and saved. We have also parallelized the computation of similarities by topics over several computer clusters. A detailed comparison of different parallelization techniques has been conducted by Gajjar et al. (2008). In addition, comparing time efficiency between the acoustics-based approach and ASR-based summarizers is interesting but not straightforward since a great deal of comparable programming optimization needs to be additionally considered in the present approach.

3.2 Estimating utterance-level similarity

In the previous stage, we calculated frame-level similarities between utterance pairs and used these to find potential matching patterns between the utterances. With this information, we estimate utterance-level similarities by estimating the numbers of true subpath alignments between two utterances, which are in turn determined by combining the following features associated with subpaths:

Similarity of subpath

We compute similarity features on each subpath. We have obtained the average similarity score of each subpath as discussed in Section 3.1. Based on this, we calculate relative similarity scores, which are computed by dividing the original similarity of a given subpath by the average similarity of its surrounding background. The motivation for capturing the relative similarity is to punish subpaths that cannot distinguish themselves from their background, e.g., those found in a block of high-similarity regions caused by certain acoustic noise.

Distortion score

Warped subpaths are less likely to correspond to valid matching patterns than straighter ones. In addition to removing very distorted subpaths by applying a distortion threshold as in (Park and Glass, 2008), we also quantitatively measured the remaining ones. We fit each of them with least-square linear regression and estimate the residue scores. As discussed above, each point on a subpath satisfies $|x_i - y_i| < T$, so the residue cannot be bigger than T. We used this to normalize the distortion scores to the range of [0,1].

Subpath length

Given two subpaths with nearly identical average similarity scores, we suggest that the longer of the two is more likely to refer to content of interest that is shared between two speech utterances, e.g., named entities. Longer subpaths may in this sense therefore be more useful in identifying similarities and redundancies within a speech summarization system. As discussed above, since the length of a subpath $len(P')$ has been proven to fall between L and $2L - 1$, i.e., $L \le len(P') \le 2L - 1$, given a parameter L, we normalize the path length to $(len(P') - L)/L$, corresponding to the range [0,1).

The similarity scores of subpaths can vary widely over different spoken documents. We do not use the raw similarity score of a subpath, but rather its rank. For example, given an utterance pair, the top-1 subpath is more likely to be a true alignment than the rest, even if its distortion score may be higher. The similarity ranks are combined with distortion scores and subpath lengths simply as follows. We divide subpaths into the top 1, 3, 5, and 10 by their raw similarity scores. For subpaths in each group, we check whether their distortion scores are below and lengths are above

some thresholds. If they are, in any group, then the corresponding subpaths are selected as "true" alignments for the purposes of building utterance-level similarity matrix. The numbers of true alignments are used to measure the similarity between two utterances. We therefore have 8 threshold parameters to estimate, and subpaths with similarity scores outside the top 10 are ignored. The rank groups are checked one after another in a decision list. Powell's algorithm (Press et al., 2007) is used to find the optimal parameters that directly minimize summarization errors made by the acoustics-based model relative to utterances selected from manual transcripts.

3.3 Extractive summarization

Once the similarity matrix between sentences in a topic is acquired, we can conduct extractive summarization by using the matrix to estimate both similarity and redundancy. As discussed above, we take the general framework of MMR and MEAD, i.e., a linear model combining salience and redundancy. In practice, we used MMR in our experiments, since the original MEAD considers also sentence positions [3] , which can always been added later as in (Penn and Zhu, 2008).

To facilitate our discussion below, we briefly revisit MMR here. MMR (Carbonell and Goldstein, 1998) iteratively augments the summary with utterances that are most similar to the document set under consideration, but most dissimilar to the previously selected utterances in that summary, as shown in the equation below. Here, the sim_1 term represents the similarity between a sentence and the document set it belongs to. The assumption is that a sentence having a higher sim_1 would better represent the content of the documents. The sim_2 term represents the similarity between a candidate sentence and sentences already in the summary. It is used to control redundancy. For the transcript-based systems, the sim_1 and sim_2 scores in this paper are measured by the number of words shared between a sentence and a sentence/document set mentioned above, weighted by the `idf` scores of these words, which is similar to the calculation of sentence *centroid values* by Radev et al. (2004).

[3]The usefulness of position varies significantly in different genres (Penn and Zhu, 2008). Even in the news domain, the style of broadcast news differs from written news, for example, the first sentence often serves to attract audiences (Christensen et al., 2004) and is hence less important as in written news. Without consideration of position, MEAD is more similar to MMR.

Note that the acoustics-based approach estimates this by using the method discussed above in Section 3.2.

$$Nextsent = \underset{t_{nr,j}}{argmax}(\lambda\, sim_1(doc, t_{nr,j}) - (1-\lambda) max_{t_{r,k}} sim_2(t_{nr,j}, t_{r,k}))$$

4 Experimental setup

We use the TDT-4 dataset for our evaluation, which consists of annotated news broadcasts grouped into common topics. Since our aim in this paper is to study the achievable performance of the audio-based model, we grouped together news stories by their news anchors for each topic. Then we selected the largest 20 groups for our experiments. Each of these contained between 5 and 20 articles.

We compare our acoustics-only approach against transcripts produced automatically from two ASR systems. The first set of transcripts was obtained directly from the TDT-4 database. These transcripts contain a word error rate of 12.6%, which is comparable to the best accuracies obtained in the literature on this data set. We also run a custom ASR system designed to produce transcripts at various degrees of accuracy in order to simulate the type of performance one might expect given languages with sparser training corpora. These custom acoustic models consist of context-dependent tri-phone units trained on HUB-4 broadcast news data by sequential Viterbi forced alignment. During each round of forced alignment, the maximum likelihood linear regression (MLLR) transform is used on gender-dependent models to improve the alignment quality. Language models are also trained on HUB-4 data.

Our aim in this paper is to study the achievable performance of the audio-based model. Instead of evaluating the result against human generated summaries, we directly compare the performance against the summaries obtained by using manual transcripts, which we take as an upper bound to the audio-based system's performance. This obviously does not preclude using the audio-based system together with other features such as utterance position, length, speaker's roles, and most others used in the literature (Penn and Zhu, 2008). Here, we do not want our results to be affected by them with the hope of observing the difference accurately. As such, we quantify success based on ROUGE (Lin, 2004) scores. Our goal is to evalu-

ate whether the relatedness of spoken documents can reasonably be gleaned solely from the surface acoustic information.

5 Experimental results

We aim to empirically determine the extent to which acoustic information alone can effectively replace conventional speech recognition within the multi-document speech summarization task. Since ASR performance can vary greatly as we discussed above, we compare our system against automatic transcripts having word error rates of 12.6%, 20.9%, 29.2%, and 35.5% on the same speech source. We changed our language models by restricting the training data so as to obtain the worst WER and then interpolated the corresponding transcripts with the TDT-4 original automatic transcripts to obtain the rest. Figure 2 shows ROUGE scores for our acoustics-only system, as depicted by horizontal lines, as well as those for the extractive summaries given automatic transcripts having different WERs, as depicted by points. Dotted lines represent the 95% confidence intervals of the transcript-based models. Figure 2 reveals that, typically, as the WERs of automatic transcripts increase to around 33%-37%, the difference between the transcript-based and the acoustics-based models is no longer significant. These observations are consistent across summaries with different fixed lengths, namely 10%, 20%, and 30% of the lengths of the source documents for the top, middle, and bottom rows of Figure 2, respectively. The consistency of this trend is shown across both ROUGE-2 and ROUGE-SU4, which are the official measures used in the DUC evaluation. We also varied the MMR parameter λ within a typical range of 0.4–1, which yielded the same observation.

Since the acoustics-based approach can be applied to any data domain and to any language in principle, this would be of special interest when those situations yield relatively high WER with conventional ASR. Figure 2 also shows the ROUGE scores achievable by selecting utterances uniformly at random for extractive summarization, which are significantly lower than all other presented methods and corroborate the usefulness of acoustic information.

Although our acoustics-based method performs similarly to automatic transcripts with 33-37% WER, the errors observed are not the same, which

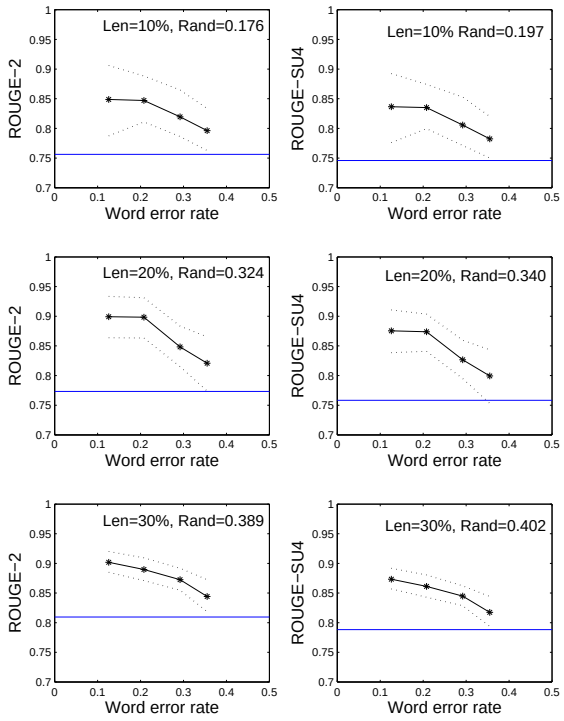

Figure 2: ROUGE scores and 95% confidence intervals for the MMR-based extractive summaries produced from our acoustics-only approach (horizontal lines), and from ASR-generated transcripts having varying WER (points). The top, middle, and bottom rows of subfigures correspond to summaries whose lengths are fixed at 10%, 20%, and 30% the sizes of the source text, respectively. λ in MMR takes 1, 0.7, and 0.4 in these rows, respectively.

we attribute to fundamental differences between these two methods. Table 1 presents the number of different utterances correctly selected by the acoustics-based and ASR-based methods across three categories, namely those sentences that are correctly selected by both methods, those appearing only in the acoustics-based summaries, and those appearing only in the ASR-based summaries. These are shown for summaries having different proportional lengths relative to the source documents and at different WERs. Again, *correctness* here means that the utterance is also selected when using a manual transcript, since that is our defined topline.

A manual analysis of the corpus shows that utterances correctly included in summaries by

	Summ. length	Both	ASR only	Aco.-only
WER=12.6%	10%	85	37	8
	20%	185	62	12
	30%	297	87	20
WER=20.9%	10%	83	36	10
	20%	178	65	19
	30%	293	79	24
WER=29.2%	10%	77	34	16
	20%	172	58	25
	30%	286	64	31
WER=35.5%	10%	75	33	18
	20%	164	54	33
	30%	272	67	45

Table 1: Utterances correctly selected by both the ASR-based models and acoustics-based approach, or by either of them, under different WERs (12.6%, 20.9%, 29.2%, and 35.5%) and summary lengths (10%, 20%, and 30% utterances of the original documents)

the acoustics-based method often contain out-of-vocabulary errors in the corresponding ASR transcripts. For example, given the news topic of the bombing of the U.S. destroyer ship *Cole* in Yemen, the ASR-based method always mistook the word *Cole*, which was not in the vocabulary, for *cold, khol, and called*. Although named entities and domain-specific terms are often highly relevant to the documents in which they are referenced, these types of words are often not included in ASR vocabularies, due to their relative global rarity. Importantly, an unsupervised acoustics-based approach such as ours does not suffer from this fundamental discord. At the very least, these findings suggest that ASR-based summarization systems augmented with our type of approach might be more robust against out-of-vocabulary errors. It is, however, very encouraging that an acoustics-based approach can perform to within a typical WER range within non-broadcast-news domains, although those domains can likewise be more challenging for the acoustics-based approach. Further experimentation is necessary. It is also of scientific interest to be able to quantify this WER as an acoustics-only baseline for further research on ASR-based spoken document summarizers.

6 Conclusions and future work

In text summarization, statistics based on word counts have traditionally served as the foundation of state-of-the-art models. In this paper, the similarity of utterances is estimated directly from recurring acoustic patterns in untranscribed audio sequences. These relatedness scores are then integrated into a maximum marginal relevance linear model to estimate the salience and redundancy of those utterance for extractive summarization. Our empirical results show that the summarization performance given acoustic information alone is statistically indistinguishable from that of modern ASR on broadcast news in cases where the WER of the latter approaches 33%-37%. This is an encouraging result in cases where summarization is required, but ASR is not available or speech recognition performance is degraded. Additional analysis suggests that the acoustics-based approach is useful in overcoming situations where out-of-vocabulary error may be more prevalent, and we suggest that a hybrid approach of traditional ASR with acoustics-based pattern matching may be the most desirable future direction of research.

One limitation of the current analysis is that summaries are extracted only for collections of spoken documents from among similar speakers. Namely, none of the topics under analysis consists of a mix of male and female speakers. We are currently investigating supervised methods to learn joint probabilistic models relating the acoustics of groups of speakers in order to normalize acoustic similarity matrices (Toda et al., 2001). We suggest that if a stochastic transfer function between male and female voices can be estimated, then the somewhat disparate acoustics of these groups of speakers may be more easily compared.

References

R. Barzilay, K. McKeown, and M. Elhadad. 1999. Information fusion in the context of multi-document summarization. In *Proc. of the 37th Association for Computational Linguistics*, pages 550–557.

J. G. Carbonell and J. Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval*, pages 335–336.

H. Christensen, B. Kolluru, Y. Gotoh, and S. Renals. 2004. From text summarisation to style-specific

summarisation for broadcast news. In *Proceedings of the 26th European Conference on Information Retrieval (ECIR-2004)*, pages 223–237.

S. Furui, T. Kikuichi, Y. Shinnaka, and C. Hori. 2003. Speech-to-speech and speech to text summarization. In *First International workshop on Language Understanding and Agents for Real World Interaction*.

M. Gajjar, R. Govindarajan, and T. V. Sreenivas. 2008. Online unsupervised pattern discovery in speech using parallelization. In *Proc. Interspeech*, pages 2458–2461.

L. He, E. Sanocki, A. Gupta, and J. Grudin. 1999. Auto-summarization of audio-video presentations. In *Proceedings of the seventh ACM international conference on Multimedia*, pages 489–498.

L. He, E. Sanocki, A. Gupta, and J. Grudin. 2000. Comparing presentation summaries: Slides vs. reading vs. listening. In *Proceedings of ACM CHI*, pages 177–184.

Y. Lin, T. Jiang, and Chao. K. 2002. Efficient algorithms for locating the length-constrained heaviest segments with applications to biomolecular sequence analysis. *J. Computer and System Science*, 63(3):570–586.

C. Lin. 2004. Rouge: a package for automatic evaluation of summaries. In *Proceedings of the 42st Annual Meeting of the Association for Computational Linguistics (ACL), Text Summarization Branches Out Workshop*, pages 74–81.

I Malioutov, A. Park, B. Barzilay, and J. Glass. 2007. Making sense of sound: Unsupervised topic segmentation over acoustic input. In *Proc. ACL*, pages 504–511.

S. Maskey and J. Hirschberg. 2005. Comparing lexial, acoustic/prosodic, discourse and structural features for speech summarization. In *Proceedings of the 9th European Conference on Speech Communication and Technology (Eurospeech)*, pages 621–624.

K. Mckeown and D.R. Radev. 1995. Generating summaries of multiple news articles. In *Proc. of SIGIR*, pages 72–82.

C. Munteanu, R. Baecker, G Penn, E. Toms, and E. James. 2006. Effect of speech recognition accuracy rates on the usefulness and usability of webcast archives. In *Proceedings of SIGCHI*, pages 493–502.

G. Murray, S. Renals, and J. Carletta. 2005. Extractive summarization of meeting recordings. In *Proceedings of the 9th European Conference on Speech Communication and Technology (Eurospeech)*, pages 593–596.

A. Park and J. Glass. 2006. Unsupervised word acquisition from speech using pattern discovery. *Proc. ICASSP*, pages 409–412.

A. Park and J. Glass. 2008. Unsupervised pattern discovery in speech. *IEEE Trans. ASLP*, 16(1):186–197.

G. Penn and X. Zhu. 2008. A critical reassessment of evaluation baselines for speech summarization. In *Proc. of the 46th Association for Computational Linguistics*, pages 407–478.

W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery. 2007. Numerical recipes: The art of science computing.

D. Radev and K. McKeown. 1998. Generating natural language summaries from multiple on-line sources. In *Computational Linguistics*, pages 469–500.

D. Radev, H. Jing, M. Stys, and D. Tam. 2004. Centroid-based summarization of multiple documents. *Information Processing and Management*, 40:919–938.

T. Toda, H. Saruwatari, and K. Shikano. 2001. Voice conversion algorithm based on gaussian mixture model with dynamic frequency warping of straight spectrum. In *Proc. ICASPP*, pages 841–844.

S. Tucker and S. Whittaker. 2008. Temporal compression of speech: an evaluation. *IEEE Transactions on Audio, Speech and Language Processing*, pages 790–796.

K. Zechner. 2001. *Automatic Summarization of Spoken Dialogues in Unrestricted Domains*. Ph.D. thesis, Carnegie Mellon University.

J. Zhang, H. Chan, P. Fung, and L Cao. 2007. Comparative study on speech summarization of broadcast news and lecture speech. In *Proc. of Interspeech*, pages 2781–2784.

X. Zhu and G. Penn. 2006. Summarization of spontaneous conversations. In *Proceedings of the 9th International Conference on Spoken Language Processing*, pages 1531–1534.