Intonational Phrases for Speech Summarization

Sameer R. Maskey†, Andrew Rosenberg, Julia Hirschberg

IBM T.J Watson Research Center, Yorktown Heights, NY, 10598†
Department of Computer Science, Columbia University, New York, NY, 10027
smaskey@us.ibm.com, {amaxwell, julia}@cs.columbia.edu

Abstract

Extractive speech summarization approaches select relevant segments of spoken documents and concatenate them to generate a summary. The extraction unit chosen, whether a sentence, syntactic constituent, or other segment, has a significant impact on the overall quality and fluency of the summary. Even though sentences tend to be the choice of most the extractive speech summarizers, in this paper, we present the results of an empirical study indicating that intonational phrases are better units of extraction for summarization. Our study compared four types of input segmentation: sentences, two pause-based segmentation, and intonational phrases (IP). We found that IPs are the best candidates for extractive summarization, improving over the second highest-performing approach, sentence-based summarization, by 8.2% F-measure.

Index Terms: Speech Summarization, Intonational Phrases, Segmentation

1. Introduction

Extractive speech summarization algorithms (e. g. , [3, 13, 16]) operate by selecting segments from the source spoken documents and concatenating them to generate a summary. Recently, text summarization approaches have been transitioning away from extractive summarization towards generative summaries, where the source documents are paraphrased to construct the final summary. While this leads to more concise and accurate summaries of text material there are significant barriers towards applying these techniques to speech documents. Resynthesis of a speech for inclusion in a summary is likely to result in the loss of information bearing qualities of the original speech, such as voice quality and intonational variation. Additionally, it is difficult to paraphrase transcribed spoken data to construct a generative summary. As Automatic Speech Recognition (ASR) transcripts of speech documents are noisy, word hypothesis and boundary errors in these transcripts make them difficult to manipulate using text-based tools and subsequently difficult to paraphrase. Moreover, speech is often disfluent; repairs, restarts and filled pauses add more complexity when attempting to generate a fluent summary from transcribed spontaneous speech. For these reasons, it is likely that speech summarization will continue to rely upon segment extraction for some time to come. Because of this foreseeable reliance on extractive summarization, the choice of extracted unit of speech will likely remain an important decision in the process of building a speech summarizer.

Generally, the speech segments extracted for summarization should be semantically meaningful and coherent stretches of speech. Segmentations currently used or proposed for extractive summarization include words, phrases, sentences, or speaker turns [3]. The choice of segmentation unit greatly influences the length and quality of the resulting summary. If speaker turns are extracted, the shortest summary will be a single turn, which may contain many sentences, not all of which may be important. If there are only ten turns in a document, a compression ratio of less than 10% is impossible — a significant limitation for many summarization tasks. We have the most control over the length of the summary if we extract individual words. However, by sacrificing higher-level structural, semantic and syntactic information from the source document, this approach is likely to be limited to a set of key words. Sentences are extracted by many of the current summarization systems and may be a better choice of segmentation for extraction; They are shorter than turns, affording finer control over the length of a summary and are semantically and syntactically meaningful units. However, automatic prediction of sentence boundaries is errorful in speech, and longer sentences may include modifiers, phrases and clauses which are not essential for the summary. Syntactic phrase extraction is a promising alternative to sentence extraction, but identifying phrases in speech transcripts using current Natural Language Processing (NLP) tools is errorful, for reasons described above. It is, however, possible to identify intonational phrase units using acoustic and prosodic information from the speech source. In this paper, we show that extracting automatically predicted intonational phrase boundaries produces the best summaries when compared to the extraction of sentences and segments based on 250ms and 500ms pauses.

In Section 2 we discuss related work. We describe our corpus in Section 3. In Section 4 we describe how we built our automatic segmentation modules and in Section 5, and we present our conclusions in Section 6.

2. Related Work

In recent years there has been a growing interest in speech summarization. Zechner [13] proposed a system to produce a summary of spontaneous speech using a Maximal Marginal Relevance technique. Hori [3]'s extractive summarization used a word-based approach, selecting a set of words to produce a given summarization ratio, where words are selected using ASR word confidence scores, linguistic scores, word significance scores, and word concatenation scores based on a Dependency Grammar. Kolluru, et al. [15] extracted phrases by using a multi-stage filtering process in which perceptrons were employed at different stages of summarization to remove words with low confidence and to find significant segments. Zhu [16] extracted sentences of spontaneous speech using a number of different feature sets. Maskey and Hirschberg [7, 8] proposed a sentence extraction system that identifies significant segments using acoustic, lexical, discourse and structural features in a ma-

chine learning framework. Each of these systems extracts some type of segment — words, phrases, or sentences — although the approaches vary in the length of the segment as well as in their extraction technique. However, none of this previous work has examined the impact of the extraction unit chosen on the performance of their summarization system.

3. Corpus

The corpus we used for our experiments is a subset of the TDT4 corpus [11]. TDT4 consists of newswire and BN in three languages: English, Arabic and Mandarin. Our subset of TDT4 consists of 12 CNN "Headline News" broadcasts. These broadcasts were manually segmented into 419 BN Stories. One human labeler was asked to generate a manual summary with a length of less than 30% of the original story. The labeler were also asked to use words and phrases directly from the story in the summary whenever possible. These annotator-generated summaries were based on manual transcripts provided with TDT4. This resulted in training material comprising 419 human summaries of manually-segmented BN stories.

ASR transcripts for these stories were provided by SRI as a part of the DARPA GALE task [20]. These ASR transcripts contain automatically hypothesized words and confidence scores. Additionally, our system had access to automatically generated story boundaries [14] and automatic speaker segmentations (diarization) [23] for the 12 shows. This module identified 96 CNN stories. All of our automatic summarization experiments were run on the automatically annotated and segmented stories, using only automatically generated words, word boundaries, and confidence scores.

We automatically aligned the manual summaries with the ASR transcripts to obtain the summary labels (i.e. should this word or phrase be included in the summary or not). We used an alignment procedure based on minimum edit distance with a higher insertion and deletion cost and a lower matching cost. Hence, the aligner found the optimal match between the words of the manual summary with the ASR transcript words. The forced alignment of summary and ASR transcripts provided us with summary labels for each word in the ASR transcripts. We used these word level summary labels to generate summary labels for each candidate segment described below. For example, to create the training data for sentence extraction, we counted the percentage of words in a given sentence that appeared in the human summary as correctly included in our automatically generated summary. If more than 50% of a segment was aligned to a manual summary, it was labeled for inclusion during the training of our summarizer, otherwise it was labeled for exclusion.

4. Speech Segmentation

In order to determine which segmentation type is best for extractive summarization of spoken documents, we must first produce candidate segmentations at different levels of granularity. We describe these segmentations and the techniques used to generate them next.

4.1. Pause-Based Segmentation

To generate pause-based segments, we calculate the pause duration between each pair of ASR-hypothesized words. We insert a segmentation boundary at every pause that exceeds a manually determined threshold. For these experiments, we construct two input segmentations — one using a 250ms threshold, another with a threshold of 500ms. Obviously, the set of boundaries selected with a 250ms threshold is a superset of those selected with a 500ms threshold. We hesitate to use a threshold below 250ms due to the potential confusion of stop gaps with phrasal boundaries [22].

Table 1: *Segmentation Statistics*

Segmentation	number per story	avg. length
250ms Pause	43.2	11.47 wds
500ms Pause	19.1	25.97
Sentence	26.9	18.46
Intonational Phrase	71.2	6.96

4.2. Automatic Sentence Segmentation

We use an automatic sentence boundary detector from ICSI [19] to produce sentence segmentation together with confidence scores for each hypothesized boundary. This system was trained on human transcriptions of BN and combines both a language model and a prosodic model. On automatically recognized speech, it operates with an error rate of 57.23%.

4.3. Intonational Phrase Segmentation

To produce training material for a ToBI-based [18, 24] intonational phrase (IP) model, we asked an expert ToBI labeler to manually annotate one TDT4 show, an ABC "World News Tonight" broadcast (20010131_1830_1900_ABC_WNT) for (binary) pitch accent presence on each word and for level 4 phrase boundaries. The annotator annotated the ASR transcript of this show, marking a hypothesized word as ending an intonational phrase if the phrase ended after or within the ASR-hypothesized word boundaries. After omitting regions of ASR error, silence and music, the training material included approximately 20 minutes of annotated speech and 3326 hypothesized words.

Using the J48 java implementation in weka [12] of Quinlan's C4.5 algorithm [17] to train a model on this single annotated show, we classify each ASR word in our summarization material as either preceding an intonational boundary or not. This decision tree is trained using feature vector containing only acoustic information: pitch, duration and intensity features. These features are extracted from raw and speaker normalized pitch and intensity tracks calculated using Praat [1]. The features that were the most indicative of the presence of a phrse boundary were: a long pause following the word, a descending change of energy over the final quarter of the word, lower minimum energy relative to the two preceding words, and decreased standard deviation of pitch. On the training material, based on ten-fold cross-validation experiments, intonational phrase boundaries are predicted with 89.1% accuracy, and an f-measure of 66.5% (precision: 68.3%, recall: 64.7%).

5. Experiments and Results

We next built summarizers that extract segments based on these four segmentation units: pauses of 250ms and 500ms, sentences, and IPs. We extracted features for the corpus for each segment type and assigned summary labels to each unit, using the forced alignments with manual summaries described in Section 3. We constructed each of our four summarizers as a binary Bayesian Network classifier [21] where the summarizer's task

is to determine whether a given segment should be included in the summary or not.

5.1. Feature Extraction

[8] have shown that speech can be summarized using just the acoustic information and for our segment comparison experiments we would like to exclude the effects of word errors, hence we extracted acoustic and structural features and built summarizers based on these features only.

We extracted aggregated acoustic features over the candidate segments. We extracted the minimum, maximum, standard deviation, mean of f0, Δ f0, RMS intensity (I) and ΔI over each segment. We also extracted the z-score ($\frac{value-mean}{std.dev.}$) of the maximum and minimum within the segment over these four acoustic information streams. Using the pitch (f0) tracks only, we extracted three pitch reset features. These were calculated by taking the difference between the average of the last 10, 5 or 1 pitch points (calculated using a 10ms frame) in the current segment, and the average of the first 10, 5 or 1 pitch points in the following segment. We speaker normalized the f0 and intensity tracks using z-score normalization, and calculated speaker normalized versions of the previously described features. We also included in the feature vector the average word length within the segment. The raw pitch tracks were extracted using Praat's [1] 'To Pitch (ac)...' function, intensity tracks using 'To Intensity...'. These features totaled 87, most based on duration, energy, and pitch.

Using hypothesized story boundaries [14] and speaker turn boundaries provided by ICSI [23], we extracted a set of structural features for each segmentation. For each segment, we identified its length, absolute and relative start time, relative position in the current speaker turn and relative position in the story. Additionally, based on the speaker identifications produced by the diarization module, we calculated the relative position of the current segment based on all of the material spoken by its speaker.

5.2. Results

The results of 10-fold stratified cross validation experiments are shown in Table 2. We find that the best summarization (using F-measure) is obtained from the summarizer which uses IPs as its input segmentation unit. The summaries produced represent a significant improvement of 8.2% in F-measure over sentence-based summaries. Note that, on average, there are about 2.75 intonational phrases for every sentence. This enables the summarizer to extract smaller segments for inclusion in summaries. However, the superior performance of the IP-based summarizer improvement is not simply due to its ability to extract smaller segments. When we compare the IP-based summarizer to the summarizer trained on 250ms pause-based segments, we see a considerable difference in F-measure. Note that, while this pause-based segmentation does operate on more units than the sentence-based summarizer (there are 1.6 250ms-pause-based segments for each sentence), the sentence-based results are slightly better — by almost 5%. Thus, merely extracting shorter segments does not necessarily improve summarization performance. Using more linguistically meaningful units, even when they are somewhat errorful, provides the best summarizer performance.

Table 2: *Information Retrieval-based Summarization Results*

Segmentation	Precision	Recall	F-Measure
250ms Pause	0.333	0.622	0.432
500ms Pause	0.255	0.756	0.381
Sentence	0.362	0.540	0.434
Intonational Phrase	0.428	0.650	0.516

F-measure evaluation assesses exact matches of predicted summary sentences to a labeled summary sentences. This measure is generally considered too strict for summarization purposes, because a segment classified incorrectly as a part of a summary may be very close in semantic content to another sentence which *was* included in the gold standard summary. A commonly used used in summary evaluation is ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [5]. ROUGE measures overlap units between automatic and manual summaries. Units measured can be n-gram, word sequences or word pairs. N-grams overlaps are computed by ROUGE-N where N indicates the size of n-grams computed. In addition to ROUGE-1 and ROUGE-2, we compute ROUGE-L, a ROUGE variant that measures the longest common subsequence between the initial document and the summaries.

$$ROUGE-N = \frac{\sum_{S \in Ref.Sum} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in Ref.Sum} \sum_{gram_n \in S} Count(gram_n)}$$

Note, that due to potential error in automatic story segmentation, a different number of summarized stories may comprise the human and automatic summaries of each show. In fact, there are more than 4 times as many manually annotated stories as those derived from automatic segmentation. Therefore, in order to evaluate the performance of the summarizer against human summaries, we compare summaries of entire shows against each other, rather than a story-by-story comparison. Thus, the target summaries for each of the twelve training shows are constructed by concatenating human summaries of manually-transcribed and manually-segmented stories. The automatic summaries of each show are concatenated automatic summaries of each automatically-segmented story.

We can note the improvement in summarization is even more pronounced in ROUGE evaluation framework. IP based ROUGE scores are higher than those obtained by extracting 250ms pause-based segments, 13.5% using ROUGE-1, 8%, using ROUGE-2 and 14.4% using ROUGE-L . Moreover, IP-based summaries are dramatically better than sentence-based summaries. The improvements seen in summarization scores, F-measure, ROUGE-1, ROUGE-2 and ROUGE-L confirm that hypothesized IPs are an excellent unit for extractive summarization of broadcast news.

Table 3: *ROUGE-based Summarization Results*

Segmentation	ROUGE-1	ROUGE-2	ROUGE-L
250ms Pause	0.437	0.103	0.415
500ms Pause	0.440	0.128	0.412
Sentence	0.394	0.096	0.377
Intonational Phrase	0.572	0.183	0.559

6. Conclusion and Future Work

We have presented results of an empirical study attempting to determine which segmentation unit should be used in extractive summarization of spoken documents. Based upon a comparison of four types of input segmentation, sentences, two pause-based segmentations, and IP segmentation, and using summarizers trained on the same corpus and using the same features, we found that IPs are the best candidates for extractive summarization, improving over the second highest-performing approach, sentence-based summarization, by 8.2% F-measure and 17.8% on ROUGE-1, 8.7% on ROUGE-2 and 18.2% on ROUGE-L. We attribute the superior performance of the IP-based summarizer to the fact that IPs are shorter than sentences in our corpus but are more linguistically and semantically meaningful units than simple pause-based segments.

The summarization technique presented in this paper operates exclusively on automatically produced input – word, intonational phrase, sentence, and story boundaries. In the future, we plan to closely examine the impact that these noisy inputs as well as common speech disfluencies – repeats, repairs, filled pauses – have on summarizer performance.

Extractive speech summarization that produces speech output poses a number of unique problems, including but not limited to 1) how to smoothly concatenate extracted speech segments and 2) how to select and order extracted segments such that the concatenated speech is understandable. We intend to extend this work in order to produce spoken summaries of BN, and will need to address these issues.

7. Acknowledgements

This work was funded by the Defense Advanced Research Projects Agency (DARPA) under Contract No. NR0011-06-C-0023. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA).

8. References

[1] Boersma, P. “Praat, a system for doing phonetics by computer”, Glot International 5:9/10, 341-345. 2001.

[2] Hirschberg J. “Communication and Prosody: Functional Aspects of Prosody”, Speech Communication, Vol 36, pp 31-43, 2002.

[3] Hori, C., Furui, S., Malkin, R., Yu, H., Waibel, A. “Automatic Speech Summarization Applied to English Broadcast News Speech,” ICASSP 2002.

[4] Kupiec, J., Pedersen, J., Chen, F. “A trainable document summarizer”, SIGIR 1995.

[5] Lin, Chin-Yew “ROUGE: A Package for Automatic Evaluation of Summaries”, Proc. Workshop on Text Summarization, ACL 2004, Barcelona.

[6] Maskey, S., Hirschberg, J., “Automatic Summarization of BroadcastNews using Structural Features”, Eurospeech 2003.

[7] Maskey, S., Hirschberg, J., “Comparing Lexical, Acoustic/Prosodic, Structural and Discourse Features”, Proc. of ICSLP 2005, Lisbon, Portugal.

[8] Maskey, S., Hirschberg, J., “Summarizing Speech without Text Using Hidden Markov Models”, HLT-NAACL, 2006, New York

[9] Schiffman, B., Nenkova, A., McKeown, K. “Experiments in multidocument summarization”, HLT 2002.

[10] Shriberg, E., Stolcke, A., Tur, D.H., Tur, G. “Prosody-Based Automatic Segmentation of Speech into Sentences and Topics”, Speech Communication, Vo. 32. pp 127-154 2000.

[11] Language Data Consortium “TDT-2 Corpus”, Univ. of Pennsylvania.

[12] Witten, I.H., E. Frank, L. Trigg, M. Hall, G. Holmes and S. J. Cunningham, “Weka: Practical machine learning tools and techniques with Java implementations,” in H. Kasabov and K. Ko, eds., ICONIP/ANZIIS/ANNES’99 International Workshop, Dunedin, 1999.

[13] Zechner, K. “Automatic Generation of Concise Summaries of Spoken Dialogues in Unrestricted Domains”, R and D in IR, 199-207, 2001.

[14] Rosenberg, A., Hirschberg J., “Story Segmentation of Brodcast News in English, Mandarin and Arabic”, Proc. of HLT/NAACL 2006, New York

[15] Kolluru B., Gotoh Y., Christensen H., “Multistage Compaction approach to Broadcast News Summarization”, Proc of Interspeech, 2005, Lisbon, Portugal

[16] Zhu X., Penn G., “Roles of textual, acoustic and spoken-language features on spontaneous-conversation summarization”, Proc of HLT/NAACL, 2006

[17] Quinlan, J. R., *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, 1993.

[18] Pitrelli, J., Beckman, M., and Hirschberg, J., “Evaluation of Prosodic Transcription Labeleing Reliability in the ToBI Framework”. In Proc. ICSLP’94, pp 123–126, 1994, Yokohama.

[19] Liu, Y., Shriberg, E., Stolcke, A., Hillard, D., Ostendorf, M., Harper, M., ”Enriching Speech Recognition with Automatic Detection of Sentence Boundaries and Disfluencies”, IEEE Transactions on Audio, Speech, and Language Processing, V14(5), pp 1526-1540, September, 2006.

[20] Stolcke, A., et al., “Recent innovations in speech-to-text transcrip- tion at sri-icsi-uw.” IEEE Transactions on Audio, Speech & Language Processing, vol. 14, no. 5, pp. 17291744, 2006.

[21] Jensen, F. V. Bayesian Networks and Decision Graphs. Springer, 2001.

[22] Luce, P., and Charles-Luce, J., Contextual effects on vowel duration, closure duration, and the consonant/vowel ratio in speech production, Journal of the Acoustical Society of America, vol. 78, no. 19491957, 1985.

[23] Wooters, C., Fung, J., Peskin, B., and Anguera, X., Towards robust speaker segmentation: The icsi-sri fall 2004 di- arization system, in RT-04F Workshop, November 2004.

[24] Beckman, M., Hirschberg, J. and Shattuck-Hufnagel, S.. “The original ToBI system and the evolution of hte ToBI framework” Ch. 2, pp. 9–54 in *Prosodic Models and Transcription: Towards Prosodic Typology* S A. Jun ed., Oxford University Press, 2004