Prominence based scoring of speech segments for automatic speech-to-speech summarization

Sree Harsha Yella[1], Vasudeva Varma[1], Kishore Prahallad[1 2]

[1]International Institute of Information and Technology, Hyderabad, India.
[2]Language Technologies Institute, Carnegie Mellon University, USA.
sreeharshay@research.iiit.ac.in, {kishore,vv}@iiit.ac.in

Abstract

In this paper we propose prominence based features for ranking speech segments for automatic speech summmarization. Standard speech summarization systems depend on ASR transcripts or/and gold standard human reference summaries which limits application of such systems. The proposed method uses prominence values of syllables in a speech segment to rank the segment for summarization. The proposed method does not depend on ASR transcripts or gold standard human summaries. Evaluation results showed that summaries generated by the proposed method can generate summaries as good as the summaries generated using tf*idf scores and supervised system trained on gold standard summaries. Experiments are carried out on two types of speech corpora one containing read style news speech and the other spontaneous telephone conversations.

1. Introduction

Speech summarization systems produce extractive summaries where important segments from input speech signal are identified, ranked and concatenated without any alterations to form a summary. One of the crucial steps in extractive summarization is determining the importance of segments and ranking the segments for inclusion into a summary. Initial approaches to speech summarization obtained ASR output of speech files and applied methods based on tf*idf, maximum marginal relvance (MMR), latent semantic analysis (LSA) to rank the segments for summarization. Methods were proposed to reduce the effect of disfluencies present in speech and ASR errors to improve the quality of summaries [1, 2, 3, 4]. Recent methods have used acoustic features in combination with lexical and structural features derived from ASR transcripts of speech signals to perform summarization. In this type of approaches a supervised system is trained with the help of gold standard human reference summaries to classify a segment as belonging to summary or not. [5] scores the sentences based on prosodic features and lexical features. [6] combines lexical and acoustic features to train a supervised system to classify an segment as belonging to summary or not. [7] attempts to summarize speech without lexical features, using only acoustic features in a HMM frame work.

All the above mentioned methods depend on the availability of human/ASR transcribed speech, or gold standard human reference summaries for training. However, ASR systems may not be available for all languages, and it takes considerable amount of resources and effort in building an ASR system for a new language. Also, constructing gold standard human reference summaries is a tedious job and they are not easily available for all speech files. In the current paper, we propose a method to rank speech segments based on prominence features. Such a method does not require an ASR system or a gold standard human summary. The proposed method uses prominence values of syllables in a speech segment to rank the speech segment for summarization.

Section 2 describes the data set used for experiments in the current work, section 3 shows the significance of prominence for summarization by making use of hand labelled prominence markings, section 4 explains the proposed method for summarization based on automatic scoring of speech segments using prominence features and its evaluation, and section 5 presents our conclusions.

2. Data-set

The studies described in the current work are carried out on two different speech corpora.

1) One corpus is a subset of Boston university radio news corpus (BU-RNC) which contains read style news speech. The data subset used in current work contains 40 news stories on different topics spoken by a female speaker (f2b). The corpus consists of orthographic text transcript corresponding to each speech segment.

2) The second copus used is a subset of switchboard data corpus released by ICSI which contains spontaneous telephone conversations. The data subset we used consists 40 conversations on the issue of credit cards. It contains speakers from both genders (38 female and 42 male) coming from wide range of dialectal patterns of American English. The corpus contains corresponding orthographic text transcript and speaker turn information.

2.1. Construction of human reference summaries

The text transcripts of the speech files are presented to 4 human annotators along with corresponding audio files for constructing a summary. Each annotator was instructed to generate a summary for 30% compression ratio. They were instructed to pick meaningful phrases or sentences present in original story without altering them. The number of reference summaries and speech files is decided following the standard evaluation setup for text summarization at document understanding conference (DUC)[1] which uses 40 topics and 4 human reference summaries.

3. Significance of prominence for speech summarization

3.1. Prominence

Prominence is defined as perceptual salience of a language unit [8]. It is the property by which linguistic units are perceived as standing out from their environment [9]. Prominence is also de-

[1]http://www-nlpir.nist.gov/projects/duc/guidelines.html

scribed in terms of distribution of accents. F2B corpus contains hand labelled pitch accent markings by experienced human labelers. These pitch accent markings are treated as prominence markers. In this section we describe experiments done using these manual prominence markings.

3.2. Content and Function words

Previous studies reported [10, 11, 12] have shown that content words are made prominent than function words in continous speech. In order to validate these observations on current data set, we have analyzed the nature of words that are marked as prominent by human labellers. Out of total 9090 words in the corpus 2852 words were marked as prominent. Out of 2852 words that are marked as prominent, 2614 (91.6%) are content words and 238 (8.3%) are function words. The content and function words classification was based on POS tags given in the corpus. This observation shows that prominence can be used to distinguish content and function words.

3.3. Acoustic measure for prominence

The hand labelled prominence markings in the corpus provide information about whether a syllable is prominent or not, but does not assign any prominence score to it. In order to obtain prominence scores for syllables in an segment, we followed the method described in [13]. This method computes a prominence score for a syllable based on acoustic features like syllable duration, filtered energy(300-2200 Hz) and pitch variation. A brief description of this method is presented below, further details can be obtained from [13].

Prominence value (p_i) of a syllable (i) in a speech segment is given by $p_i = max(F1_i, F2_i)$ where,
$F1_i = dur^i \times en^i_{300-2200}$. Here dur^i is the syllable duration and $en^i_{300-2200}$ is the energy in frequency band $300-2200Hz$.
$F2_i = en^i_{ov} \times (A^i_{event} \times D^i_{event} \times R^i_{event})$, where
en^i_{ov} is the overall syllable energy, A^i_{event} , D^i_{event} are amplitude and duration of an intonational event respectively and R^i_{event} is a normalizing factor. The intonational events considered here are those events that contain a rise followed by a fall in the pitch profile. These type of intonation events were shown to correlate well with human prominence judgements [14].

Figure 1 shows the distribution of prominence values for prominent and not prominent syllables in the corpus.

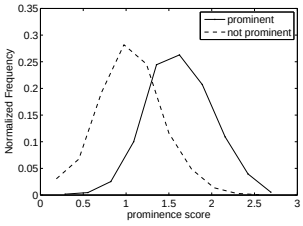

Figure 1: *Distributions of prominence values for prominent and non prominent syllables.*

It can be observed from figure 1 that prominence values for syllables marked as prominent are higher than values of non prominent syllables. Therefore, the computed value for a syllable can be treated as a measure of its prominence.

3.4. Usefulness of prominence for speech summarization

In order to investigate usefulness of prominence for summarization, we use hand labelled prominence markings for automatic summarization. The prominence scores of syllables that are hand-labelled as prominent are obtained by the method described in Section 3.3. Acoustic score of a speech segment used for it's ranking is obtained by taking the mean of prominence values of syllables that are hand marked as prominent. Speech segments are ranked in decreasing order of acoustic scores and top ranking segments are concatenated in chronological order of their occurrence in the news show until desired summary length is reached.

The distribution of acoustic scores for speech segments belonging to summary class and non summary class is shown in Figure 2. It can be observed from the Figure 2 that segments belonging to summary class tend to have high prominence score than segments not in summary. This shows that prominence based scoring of speech segments helps in automatic summarization. In order to formally evaluate the usefulness of prominence for summarization, we compare the summaries generated by prominence based scoring with summaries generated by tf*idf based scoring of manual transcripts (section 3.5) and summaries generated by a supervised system trained on gold standard human reference summaries (section 3.6).

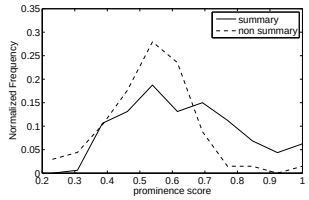

Figure 2: *Distributions of acoustic scores for segments belonging to summary and non summary classes.*

3.5. Summaries based on tf*idf scores

The tf*idf scores are computed from manual transcripts provided along with the corpus. The tf*idf based score of a segment is computed as similarity measure between the segment and the whole document. Sentences are ranked in decreasing order of their similarity scores. The similarity between a segment and the document is computed by the dot product between correponding vectors with terms as dimensions and tf*idf scores of the terms as magnitudes of corresponding dimensions.

3.6. Supervised system using acoustic features

An artificial neural network classifier was trained on gold standard human labelled summaries which contained segments from all four human summaries. The classifier was trained with class labels -1 for class 'non summary' and 1 for class 'summary'. The features on which the classifier is trained consist of minimum, maximum, mean, standard deviation of RMS intensity (I), ΔI, F_0, ΔF_0 over each segment. The F_0 and I contours are normalized using z-score normalization. The corpus was divided randomly into two non overlapping halfs. Classifier was trained on one half and tested on the other. While testing the classifier outputs a score between -1 and 1 for a given speech segment. This score is used for ranking the speech segments to generate audio summaries for desired length. Summaries are generated for 30% compression ratio.

3.7. Evaluation

The evaluation of summaries generated by the three techniques explained in sections 3.4, 3.5, 3.6 was done by estimating how close they are with human reference summaries. The summaries are evaluated using standard text summarization evaluation system ROUGE[15]. ROUGE measures n-gram overlap between human reference summaries and automatic summaries. Four human reference summaries are provided as model ref-

erence summaries for each news story. We report ROUGE-1, ROUGE-2 and ROUGE-SU4 scores for these summaries in Table 1. ROUGE-N measures N-gram overlap between human reference summaries and automatic summary. ROUGE-SU4 measures the skip Bi-gram overlap within a window of four. Audio summaries are transcribed into text by picking corresponding text segments from the manual transcripts provided with the corpus.

Table 1: *F-measure values and 95% confidence intervals for ROUGE-1 (R-1), ROUGE-2 (R-2), ROUGE-SU4 (R-SU4) metrics for prominence based summaries and summaries based on acoustic features.*

system	R-1	R-2	R-SU4
prominence	0.5157 [0.49 0.53]	0.3510 [0.33 0.37]	0.3451 [0.32 0.36]
supervised	0.4786 [0.45 0.49]	0.3403 [0.32 0.36]	0.3373 [0.31 0.35]
tf*idf	0.5141 [0.49 0.53]	0.3371 [0.31 0.35]	0.3443[0.32 0.36]

From Table 1 it can be seen that prominence based features generate summaries as good as summaries generated by supervised system trained on standard acoustic features and summaries based on tf*idf scores of manual transcripts. The advantage of prominence based summaries is, they do not depend on ASR output or gold standard human labelled summary for training. In this experiment, we have made use of prominence markings provided by human experts. This was done primarily to demonstrate that explicit modelling of prominence helps in ranking speech segments for automatic summarization in an unsupervised framework. In the next section we propose a speech-to-speech summarization method where syllable boundaries of a speech segment are automatically computed and the segment is ranked using prominence scores of syllables in the segment.

4. Speech summarization using automatic prominence scoring

The speech files given as input are first segmented by extracting speech segments based on pause duration. A segment boundary is assumed whenever a pause greater than 250 ms is encountered. In order to rank the speech segments automatically by their acoustic score, we need syllable boundaries. To obtain syllable boundaries of a speech segment automatically, we followed the method used in [13]. The errors in syllable segmentation on the present data set is reported in terms of missed detection rate (MDR) and false alarm rate (FAR). The MDR and FAR values on the current data set are 12.3% and 9.4% respectively. Prominence value of each syllable in the segment is computed as described in section 3.3. To obtain acoustic score of a segment from prominence values of syllables present in it, four types of scoring functions are experimented. First function calculates mean prominence score (mp) of a segment by taking mean of prominence values of syllables in it.

$$mp = \frac{\sum_{i=1}^{N} p_i}{N}, \quad (1)$$

where p_i is prominence value of i^{th} syllable and N is total number of syllables in a segment. Second function scores a segment by maximum prominence value (Mp) of syllables in it.

Third function assigns mean value of absolute difference between prominence values of consecutive syllables (mdp) in a segment as its score. The use of difference between prominence values serves to normalize data against variation between speakers, but preserves variations produced by prosody.

$$mdp = \frac{\sum_{i=1}^{N-1} |p_{i+1} - p_i|}{N-1}, \quad (2)$$

Fourth function assigns maximum of absolute difference (Mdp) between prominence values of consecutive syllables in a segment as its score. Segments are ranked in decreasing order of their acoustic score and top ranking segments are concatenated in chronological order of their occurrence in the news story until the desired summary length is reached. The desired length is obtained from given compression ratio which is defined as ratio of summary length to document length.

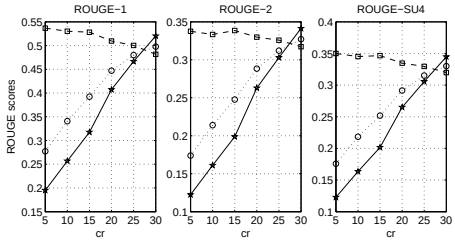

Figure 3: *Figure showing recall (solid line), precision (dashed line) and f-measure (dotted line) values of different ROUGE metrics for different compression ratios (5, 10, 15, 20, 25, 30) of audio summaries generated by mdp scoring function.*

4.1. Evaluation

Evaluation of the summaries generated by automatic prominence detection was done in two ways, one using ROUGE [15] and the other based on task based evaluation by humans. Task based evaluation was done to evaluate the quality of the audio summaries.

All the summaries are generated for a compression ratio of 30% (same as model summaries). 4 human summaries are provided as model reference summaries for each story. ROUGE scores for different prominence scoring function are reported in table 2. It can be observed that mdp performs better than other scoring functions. In order to evaluate the summarization capability of the proposed technique for different compression ratios, ROUGE scores for summaries of different compression ratios (5, 10, 15, 20, 25, 30) with mdp as scoring function are reported in figure 3. It can be observed from figure 3 that precision values do not drop much with increase in compression ratio. This shows that system is capable of generating summaries of different lengths without compromising on the quality of summaries.

Table 2: *F-measure values and 95% confidence intervals of ROUGE-1 (R-1), ROUGE-2 (R-2), ROUGE-SU4 (R-SU4) metrics for various scoring functions.*

system	R-1	R-2	R-SU4
mp	0.4964 [0.47 0.51]	0.3165 [0.29 0.33]	0.3225 [0.29 0.34]
Mp	0.4740 [0.45 0.49]	0.2978 [0.27 0.31]	0.3050 [0.28 0.32]
mdp	0.5085 [0.48 0.52]	0.3413 [0.32 0.36]	0.3431 [0.32 0.36]
Mdp	0.4893 [0.46 0.50]	0.3237 [0.30 0.34]	0.3285 [0.30 0.34]

In task based evaluation, five human subjects are asked to listen to a summary of a given compression rate and answer a questionnaire given to them. All the subjects are in the age group of 20-23 and are graduate students who can understand and speak English. The questionnaire consisted of simple questions based on facts of the news story. The questions are of type what, when, who, where etc. The subjects were given strict instructions not to use their prior knowledge on the news stories in answering the questions. They answered the questions based on the information present in the summary. The subjects were not restricted from listening to a summary multiple times. The percentage of the questions answered correctly for each compression ratio is presented in table 3.

Table 3: *Percentage of questions answered correctly for different compression ratios (CR)*

cr	5	10	15	20	25
correct(%)	32.4%	41.5%	45.6%	51.3%	56.8%

The results of task based evaluation (table 3) show that hu-

mans are able to understand the audio summaries and are able to get some useful information from these audio summaries. The number of questions answered correctly increased with the compression ratio which agrees with the ROUGE based evaluation (figure 3).

4.2. Correlation between tf*idf based summaries and prominence based summaries

Figure 4 shows scatter plot between tf*idf scores and prominence score (mdp) for phrases picked in prominence (mdp) based summaries(a) and tf*idf based summaries (b) for two news stories 1 and 2. In Figure 4 it can be observed from 1(a) and 2(a)

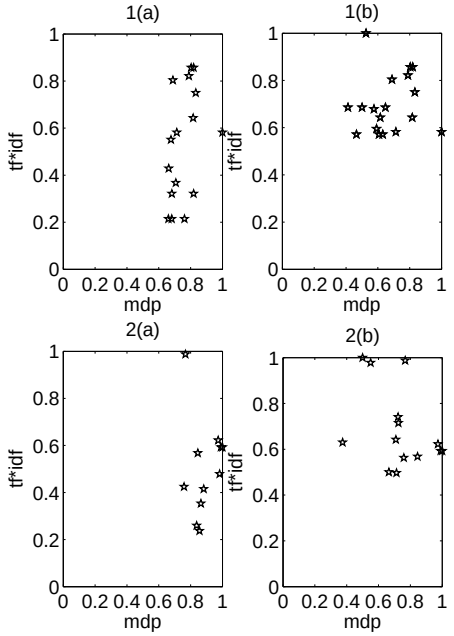

Figure 4: *Scatter plots between tf*idf scores and prominence score (mdp) for summaries of two news stories 1 and 2. (a) shows the scatter plot of scores for phrases picked in summaries based on prominence (mdp) scores. (b) shows the scatter plot of scores for phrases picked in summaries based on tf*idf scores.*

that some phrases picked in prominence based summaries have low tf*idf scores, where as it can be observed from 1(b) and 2(b) (tf*idf based summaries) that phrases having high tf*idf scores also have high prominence (mdp) scores. This shows that prominence based ranking provides some complementary information to tf*idf based ranking. In order to capture this complementary information, segments are ranked by a combined score computed from prominence score and tf*idf score of segments. The scores obtained from prominence scoring and tf*idf scoring for a document are normalized between 0 and 1 and a combined score is obtained by adding these two scores. The ROUGE scores for these summaries are reported in Table.

Table 4: *F-measure values and 95% confidence intervals of ROUGE-1 (R-1), ROUGE-2 (R-2), ROUGE-SU4 (R-SU4) metrics for summaries generated by combined score.*

system	R-1	R-2	R-SU4
mdp + tf*idf	0.5204 [0.50 0.54]	0.3503 [0.33 0.37]	0.3562 [0.33 0.37]

4.3. Experiments on switchboard data

The evaluation of the proposed method is also done on switchboard data which contains spontaneous telephone dialogues. A conversation is segmented at speaker turns that are provided with the corpus. These speaker turns are treated as basic units while performing extractive summarization. Each speaker turn is assigned an acoustic score as described in section 4. Top scoring speaker turns are concatenated untill desired summary length is reached. Evaluation of these summaries was carried out using ROUGE package. Similar to the results obtained on f2b corpus mdp scoring function performed better than other scoring functions. The perfomance of the proposed method along with tf*idf based scores and supervised system on switchboard data is reported in terms of ROUGE scores in table 5.

Table 5: *F-measure values and 95% confidence intervals of ROUGE-1 (R-1), ROUGE-2 (R-2), ROUGE-SU4 (R-SU4) metrics for prominence based summaries, tf*idf based summaries and supervised system on switchboard data.*

system	R-1	R-2	R-SU4
mdp	0.6660 [0.64 0.68]	0.4640 [0.41 0.49]	0.4914 [0.45 0.52]
tf*idf	0.6534 [0.62 0.68]	0.4616 [0.37 0.54]	0.4918 [0.42 0.56]
supervised	0.6280 [0.59 0.65]	0.4568 [0.40 0.48]	0.4749 [0.40 0.52]

5. Conclusions

We proposed an automatic speech summarization system based on prominence. The proposed technique does not require ASR/manual transcripts or human reference summaries for training. Evaluation results showed that the proposed technique generates summaries that are as good as summaries generated by text summarizer based on tf*idf and summaries generated by as supervised system trained on standard acoustic features. The summaries for desired length are produced without loss in the quality. The output summaries are presented in form of speech by preserving characteristics of the input speech signal.

6. References

[1] K. Zechner, "Automatic generation of concise summaries of spoken dialogues in unrestricted domains," *R and D in IR*, pp. 199–207, 2001.

[2] G. Murray, S. Renals, and J. Carletta, "Extractive summarization of meeting recordings," in *Proceedings of the 9th European Conference on Speech Communication and Technology*, 2005, pp. 593–596.

[3] S. Furui, T. Kikuchi, Y. Shinnaka, and C. Hori, "Speech-to-text and speech-to-speech summarization of spontaneous speech," *Speech and Audio Processing, IEEE Transactions on*, vol. 12, no. 4, pp. 401–408, July 2004.

[4] B. Kolluru, H. Christensen, and Y. Gotoh, "Multi-stage compaction approach to broadcast news summarisation," in *Proceedings of Eurospeech*, 2005, pp. 69–72.

[5] A. Inoue, T. Mikami, and Y. Yamashita, "Improvement of speech summarization using prosodic information," in *Proc. Speech Prosody*, Japan, 2004.

[6] S. Maskey and J. Hirschberg, "Comparing lexical, acoustic/prosodic, structural and discourse features for speech summarization," in *ICSLP*, 2005.

[7] S. Maskey and J.Hirschberg, "Summarizing speech without text using hidden markov models," in *NAACL-HLT*, 2006.

[8] B. M. Streefkerk, L. C. W. Pols, and L. F. M. T. Bosch, "Acoustical features as predictors for prominence in read aloud dutch sentences used in ann's," in *Proceedings of the European Conference on Speech Processing and Technology*, 1999, pp. 551–554.

[9] J. Terken, "Fundamental frequency and perceived prominence of accented syllables," *The Journal of the Acoustical Society of America*, vol. 89, no. 4, pp. 1768–1776, 1991. [Online]. Available: http://link.aip.org/link/?JAS/89/1768/1

[10] S. Pan and K. R. Mckeown, "Word informativeness and automatic pitch accent modeling," in *In Proceedings of EMNLP/VLC99*, 1999, pp. 148–157.

[11] G. B. M. Horne, P. Hansson, and J. Frid, "Prosodic correlates of information structure in swedish human-human dialogues," in *In Proceedings Eurospeech*, 1999, pp. 29–32.

[12] R. Silipo and F. Crestani, “Prosodic stress and topic detection in spoken sentences,” in *SPIRE ’00: Proceedings of the Seventh International Symposium on String Processing Information Retrieval (SPIRE’00)*. Washington, DC, USA: IEEE Computer Society, 2000, p. 243.

[13] F. Tamburini and C. Caini, “An automatic system for detecting prosodic prominence in american english continuous speech,” *International Journal of Speech Technology*, vol. 8, no. 1, pp. 33–44, January 2005. [Online]. Available: http://dx.doi.org/10.1007/s10772-005-4760-z

[14] P. Taylor, “Analysis and synthesis of intonation using the tilt model,” *Journal of the Acoustical Society of America*, vol. 107, pp. 1697–1714, 2000.

[15] C. Y. Lin, “Rouge: a package for automatic evaluation of summaries,” in *ACL Text summarization Workshop*, 2004.