A Project Progress Report on

Speech Summarization

Undergone at

National Institute of Technology Karnataka

Under the guidance of

Dr. Shashidhar G. Koolagudi

Submitted By

Yogendra Kumar

(Roll no. 172CS026)

in partial fulfillment of the requirements for the award of the degree of

Master of Technology

In

Computer Science and Engineering

Department of Computer Science and Engineering

National Institute of Technology Karnataka

November, 2018

Contents

List of Figures

1 Introduction

As the numbers of speech documents available on the Internet is very big, it becomes increasingly important to enable users to find relevant, significant and interesting parts of the documents automatically.The summarized document helps people to get important information in less time rather than going through the whole document.Summarizing recorded lectures, meetings, voicemail/voiceall, and broadcast news can help people get important things in less time.The summarization is the one way to fulfill the above requirement.There are two ways to do summarization of the speech documents, One is abstractive and another is extractive summarization. Abstractive summary is generated in the own words . And the extractive summary is a extraction of important segments from total segments.

The Speech Summarization process is mentioned in Figure1. The work of each steps is explained in the following points.

- First the speech document will be taken then preprocessing will be performed because some part of the speech might not important for the summarization.
- After that segmentation will be performed on the speech document. Each segment will be the sentence of the speech file.
- The feature extraction will be performed on the segments. Using extracted features of the speech document the summary will be generated.
- Finally the validation of the summary will be performed.

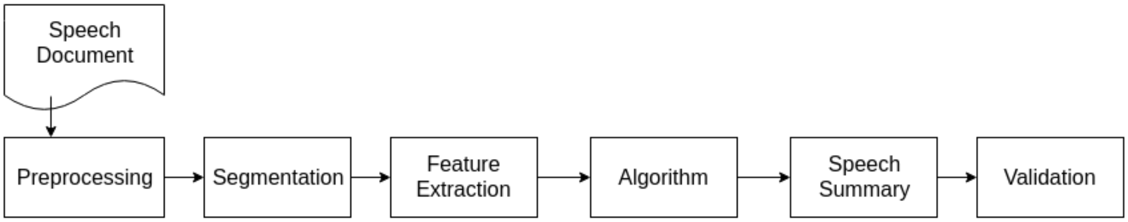


Figure 1: Speech Summarization Process

Following are the application of speech summarization.

- If someone missed the meeting and the meeting was too long and he/she do not want to listen to the whole meeting recording. So here speech summarization come into the picture. He/She will get the important things in less time.
- If someone coming back from a hard day at work and start the television to find a half hour show that has summarized all Broadcast News shows of his/her interest in all the available news channels. Here the speech summarization can be used to fulfill the requirement.
- There are many courses available on the internet. Each having many lectures. And if someone wants to get all the important information of the course in less time. Here speech summarization comes into the picture. He/She will get the important things in less time.
- Similarly, if someone have many voicemail/voicecall then he/she can get the important message from each voicemail/voicecall using speech summarization.

The aim of the project is to generate the speech form summary for the particular speech document without using any type of the text/transcript.The vision of the project is based on the two benefits.The first benefit of the project is that the summary can be generated for the speech document of any type of language because the generated summary does not use the linguistic feature of the speech document. And second benefit is that if the technique of text summarization used then the summary may loose the information related to the emotions present in the speech because the speech is less formal than text.So the idea is to extract and analyze the efficient features in the speech documents and generate automatic speech summarization using various combination of selected speech features.

2 Literature Survey

From Text to Speech Summarization

- Kathleen McKeown, Julia Hirschberg, Michel Galley and Sameer Maskey are the researchers of this paper[3].
- Here in this paper, the researchers presented that the speech summarization can be performed using the techniques of text summarization.
- But they also mentioned that spoken language is less formal than written language. So sentence extraction cannot be applied directly to speech summarization.
- Sometimes the speaker expresses words with some emotion and energy. So converting speech to text might be destroyed the emotion of the speaker that may useful for summarization.

Summarizing Speech Without Text Using Hidden Markov Models

- Sameer Maskey, Julia Hirschberg are the researchers of the this paper [2].
- Here in this paper, the researchers presented a method for summarizing speech documents without using any type of transcript/text in a Hidden Markov Model framework.
- They build the model in which each state tells that this sentence will be included in the summary or not.
- They have extracted 12 acoustic features for training. It includes speaking rate (the ratio of voiced/total frames); F0 minimum, maximum, and mean; F0 range and slope; minimum, maximum, and mean RMS energy (minDB, maxDB, meanDB); RMS slope (slopeDB); sentence duration (timeLen = endtime - starttime).
- They trained and tested the model using the TDT-2 corpus.

- Then they compared the summary with the human-generated summary for performance evaluation.

Automatic Broadcast News Speech Summarization

- Sameer Raj Maskey is the author[1].
- Here in this paper, the researcher presented a system for speech summarization.
- In which the lexical features, acoustic/prosodic features, structural features, discourse features of the speech used for the automatic speech summarization system.
- He also generated the summary by using only acoustic properties only. These all feature are intensity, pitch, speaking rate and duration. These all extracted using Praat Software. He also showed that summarization using speech to text can be improved by using acoustic information.
- They have created two corpora.The first one is the Columbia BN Speech Summarization Corpus I (CBNSCI) and the second one is the Columbia BN Speech Summarization Corpus II (CBNSCII). Both corpora consist of CNN Headline news shows collected over the period of 1998 to 2003 with annotations for summarization and IE experiments.
- In this paper additional problems in speech summarization also addressed like word error, dis-fluency errors and boundary errors and how to overcome them also proposed.

Speech Under Stress: Analysis, Modeling and Recognition

- John H. L. Hansen, Sanjay Patil are the authors[4].
- They are identifying that what could be perceived as stress and how it affects the speech production system.
- In that they worked for development of algorithms to estimate classify or distinguish different stress conditions.
- They used some features like stuttering, repetition, tongue-slip, pauses between utterances, speed of word production, energy, intensity, pitch (fundamental frequency), formant locations / structure (vocal tract), glottal structure (spectral slope), duration.
- Then they conclude with revealing what might be in store for understanding stress, and the development of techniques to overcome the effects of stress for speech recognition and human-computer interactive systems.

Intensity Feature for Speech Stress Detection

- Laszlo Czap, Judit Maria Pinter are the authors[5].
- They compare the amplitude of the actual vowel to that of its average to show the stressed or unstressed nature of the syllable.
- They got average energies of vowels are obtained from a speech recognizer trained with voices of hundreds of speaker.
- They also mentioned that in creation of the stress three main factors individually or collectively play role. These factors are fundamental frequency, higher articulation intensity and duration of the vowel of stressed syllable.

Spectral Analysis of Stressed Speech for Speech Recognition

- Sumitra Shukla is the author[6].
- The objective of this thesis is to analyze the stress information in the spectral features of stressed speech.
- They explained about glottal parameters and how these parameters will be useful in identifying stress information.

3 Proposed Solution

The objective of the project are

- Extraction and analysis of the features efficient in speech summarization.
- Automatic speech summarization using various combination of selected speech features.

So solution to the problem is to extract efficient features and use these features to select the important sentences. On the basis of important sentences the extracted summary will generated. Features extraction on words are good compare to features extraction on the sentences. So sentence segmentation and word segmentation have to be done on the each preprocessed file. First the features extraction will done on the words then important words will be selected then after that corresponding sentences will be used to generate the extracted summary. For validating the extracted summary, the opinion score will be used.

The progress work is divided into four parts.

- Manual Work
- Feature Extraction
- Important Word Selection, Summary Generation and Validation.
- Result

3.1 Manual Work

Three task has been performed manually before applying the feature extraction. First preprocessing has been performed for removing unnecessary parts of the speech document. Removal of the unnecessary parts is necessary from the audio because these parts does not help summarization. Second is the sentence segmentation and third is the word segmentation. The audio files have been segmented into sentences and words before processing to feature extraction. After that feature extraction will be done for each words of the file. Then important words will be selected on the basis of the sum metric.

Using important words sentence selection will be done.

3.2 Feature Extraction

The details of the four extracted features is given below :-

3.2.1 Average Pitch

There is considerable evidence that topic shifts may be marked by changes in pitch. Mean pitch can be used for normalizing pitch differences between speakers of different genders. While a maximum and minimum pitch and their difference can provide important clues on the effective emphasis a person may be putting while he/she is speaking.

3.2.2 Average Intensity

Extracting the intensity (amplitude) of the speech signal with the motivation that speakers may speak louder if they want to emphasize significant segments of speech.

3.2.3 Average Slope Ratio and Average Duration Ratio

These parameters are opening slope, closing slope, opening duration, top duration, closing duration, and closed duration. These parameters control glottal air flow. The slopes of glottal opening and closing provide control over the amount of acoustic energy produced and the duration's of glottal closure decide the length of pitch period. For Stressed word these two ratios vary significantly . So using these parameters will help in identifying the important word.

3.3 Important Word Selection, Summary Generation and Validation

The figure no.2 shows the process after feature extraction. Below is the description about the whole process after feature extraction.

- Column Standardization is performed to make every column scale free because every column have different unit to measure.

- Then generated the sum metric which is sum of all the features.

- On the basis of highest sum the important sentences will be marked.

- On the basis of important words corresponding sentences will be appended and appended sentences will become summary.

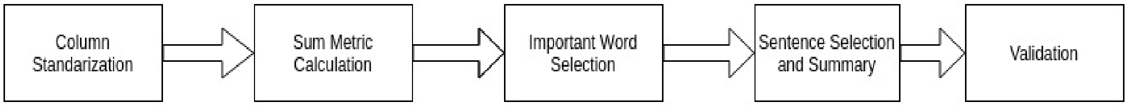


Figure 2: Important Word Selection, Summary Generation and Validation Process

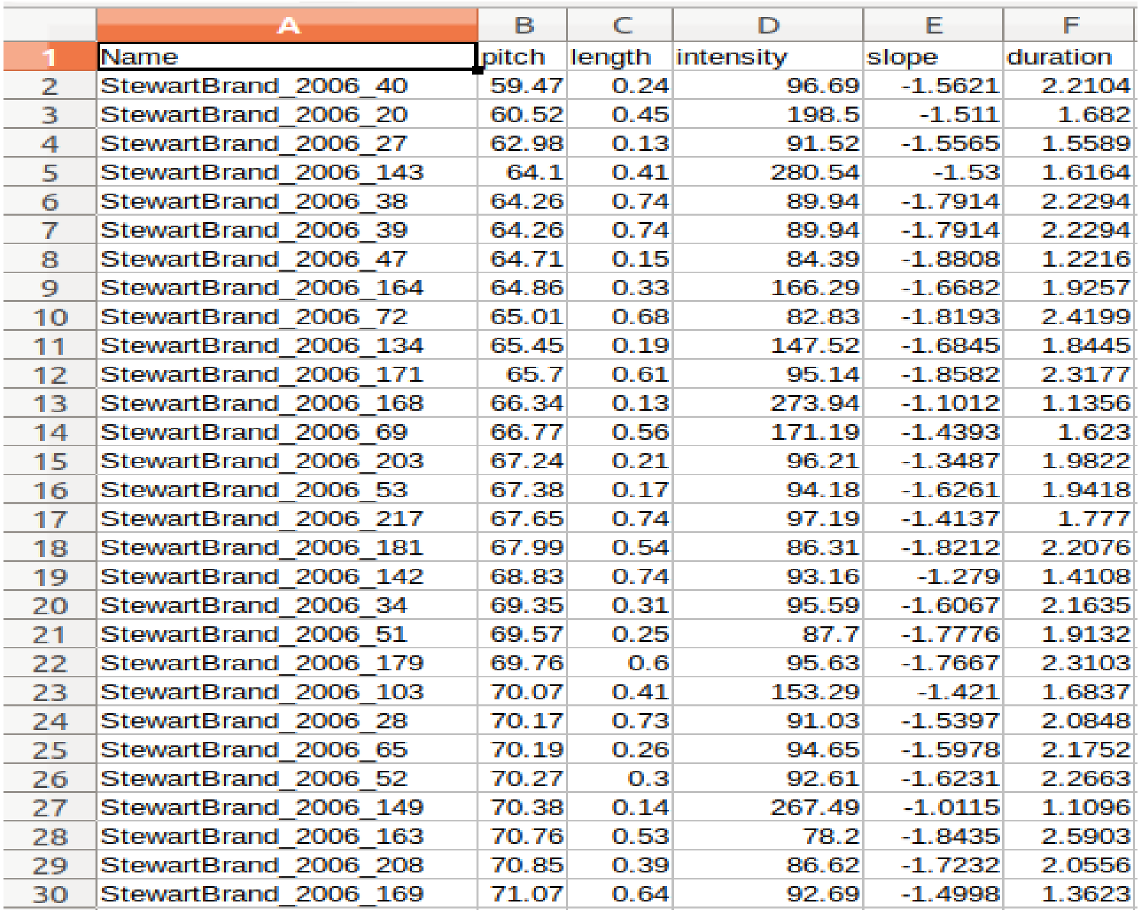
3.4 Results

3.4.1 Dataset

The TED-LIUM is the dataset used. TED is a non-profit organization for spreading ideas usually in 18 minutes or short. In Ted talks, topics are from science or business or global issues. It is in English language TED talks with transcriptions. It contains about 118 hours of speech and the size is about 20 GB.

3.4.2 Extracted Features For Each Words

The experiment has been performed on one of the ted-talk named **Stewart Brand.** First preprocessing has been performed the audio file. Then segmented the audio file into sentences and words. Then feature extraction have been done for each words of the audio file. The screen-shot of the extracted feature for each words is shown below in the figure no 3.

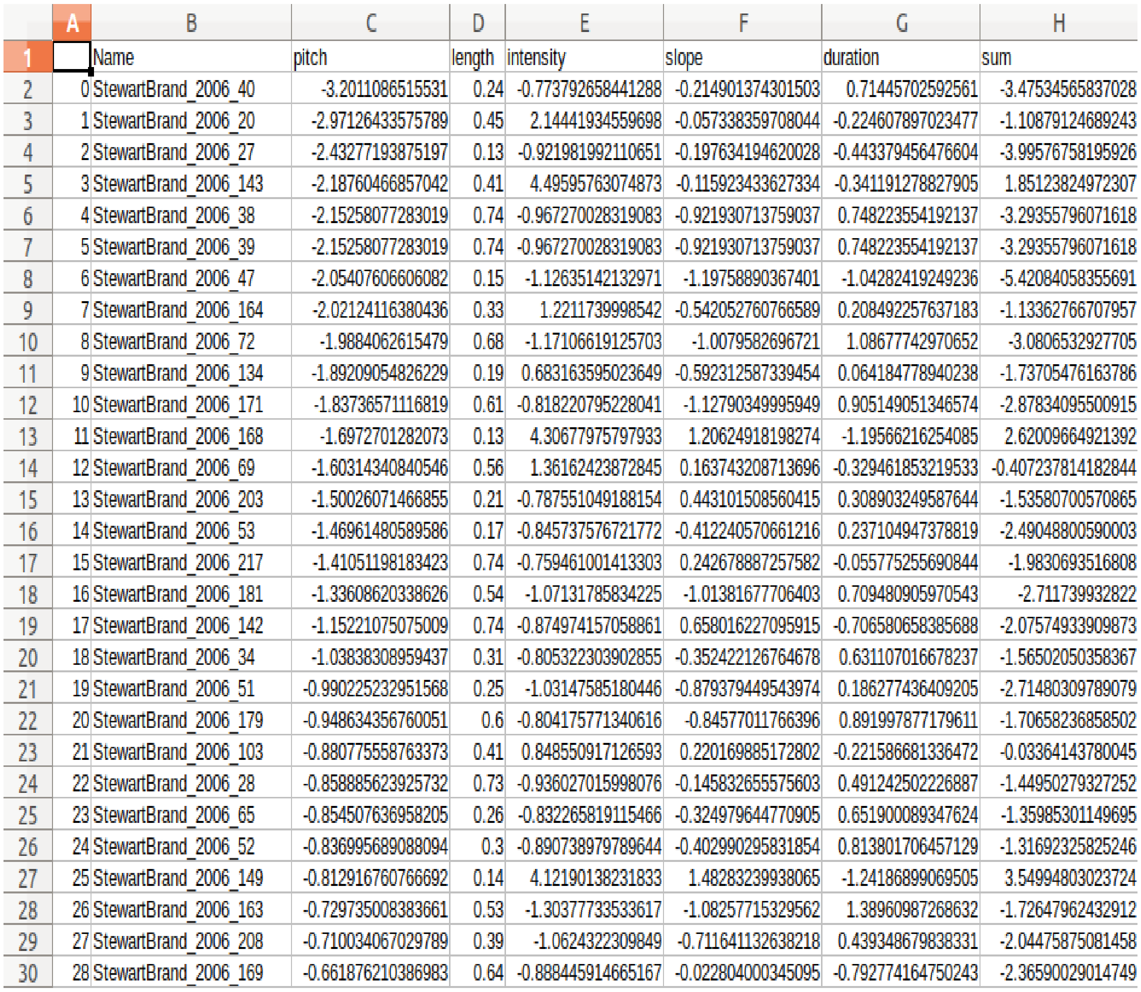


	A	B	C	D	E	F
1	Name	pitch	length	intensity	slope	duration
2	StewartBrand_2006_40	59.47	0.24	96.69	-1.5621	2.2104
3	StewartBrand_2006_20	60.52	0.45	198.5	-1.511	1.682
4	StewartBrand_2006_27	62.98	0.13	91.52	-1.5565	1.5589
5	StewartBrand_2006_143	64.1	0.41	280.54	-1.53	1.6164
6	StewartBrand_2006_38	64.26	0.74	89.94	-1.7914	2.2294
7	StewartBrand_2006_39	64.26	0.74	89.94	-1.7914	2.2294
8	StewartBrand_2006_47	64.71	0.15	84.39	-1.8808	1.2216
9	StewartBrand_2006_164	64.86	0.33	166.29	-1.6682	1.9257
10	StewartBrand_2006_72	65.01	0.68	82.83	-1.8193	2.4199
11	StewartBrand_2006_134	65.45	0.19	147.52	-1.6845	1.8445
12	StewartBrand_2006_171	65.7	0.61	95.14	-1.8582	2.3177
13	StewartBrand_2006_168	66.34	0.13	273.94	-1.1012	1.1356
14	StewartBrand_2006_69	66.77	0.56	171.19	-1.4393	1.623
15	StewartBrand_2006_203	67.24	0.21	96.21	-1.3487	1.9822
16	StewartBrand_2006_53	67.38	0.17	94.18	-1.6261	1.9418
17	StewartBrand_2006_217	67.65	0.74	97.19	-1.4137	1.777
18	StewartBrand_2006_181	67.99	0.54	86.31	-1.8212	2.2076
19	StewartBrand_2006_142	68.83	0.74	93.16	-1.279	1.4108
20	StewartBrand_2006_34	69.35	0.31	95.59	-1.6067	2.1635
21	StewartBrand_2006_51	69.57	0.25	87.7	-1.7776	1.9132
22	StewartBrand_2006_179	69.76	0.6	95.63	-1.7667	2.3103
23	StewartBrand_2006_103	70.07	0.41	153.29	-1.421	1.6837
24	StewartBrand_2006_28	70.17	0.73	91.03	-1.5397	2.0848
25	StewartBrand_2006_65	70.19	0.26	94.65	-1.5978	2.1752
26	StewartBrand_2006_52	70.27	0.3	92.61	-1.6231	2.2663
27	StewartBrand_2006_149	70.38	0.14	267.49	-1.0115	1.1096
28	StewartBrand_2006_163	70.76	0.53	78.2	-1.8435	2.5903
29	StewartBrand_2006_208	70.85	0.39	86.62	-1.7232	2.0556
30	StewartBrand_2006_169	71.07	0.64	92.69	-1.4998	1.3623

Figure 3: Extracted Features For Each Word

3.4.3 Result After Column Standardization

After features extraction column standardization has been done on the same collected extracted features. The screen-shot for particular file is shown below.



	A	B	C	D	E	F	G	H
1		Name	pitch	length	intensity	slope	duration	sum
2	0	StewartBrand_2006_40	-3.2011086515531	0.24	-0.773792658441288	-0.214901374301503	0.71445702592561	-3.47534565837028
3	1	StewartBrand_2006_20	-2.97126433575789	0.45	2.14441934559698	-0.057338359708044	-0.224607897023477	-1.10879124689243
4	2	StewartBrand_2006_27	-2.43277193875197	0.13	-0.921981992110651	-0.197634194620028	-0.443379456476604	-3.99576758195926
5	3	StewartBrand_2006_143	-2.18760466857042	0.41	4.49595763074873	-0.115923433627334	-0.341191278827905	1.85123824972307
6	4	StewartBrand_2006_38	-2.15258077283019	0.74	-0.967270028319083	-0.921930713759037	0.748223554192137	-3.29355796071618
7	5	StewartBrand_2006_39	-2.15258077283019	0.74	-0.967270028319083	-0.921930713759037	0.748223554192137	-3.29355796071618
8	6	StewartBrand_2006_47	-2.05407606606082	0.15	-1.12635142132971	-1.19758890367401	-1.04282419249236	-5.42084058355691
9	7	StewartBrand_2006_164	-2.02124116380436	0.33	1.2211739998542	-0.542052760766589	0.208492257637183	-1.13362766707957
10	8	StewartBrand_2006_72	-1.9884062615479	0.68	-1.17106619125703	-1.0079582696721	1.08677742970652	-3.0806532927705
11	9	StewartBrand_2006_134	-1.89209054826229	0.19	0.683163595023649	-0.592312587339454	0.064184778940238	-1.73705476163786
12	10	StewartBrand_2006_171	-1.83736571116819	0.61	-0.818220795228041	-1.12790349995949	0.905149051346574	-2.87834095500915
13	11	StewartBrand_2006_168	-1.6972701282073	0.13	4.30677975797933	1.20624918198274	-1.19566216254085	2.62009664921392
14	12	StewartBrand_2006_69	-1.60314340840546	0.56	1.36162423872845	0.163743208713696	-0.329461853219533	-0.407237814182844
15	13	StewartBrand_2006_203	-1.50026071466855	0.21	-0.787551049188154	0.443101508560415	0.308903249587644	-1.53580700570865
16	14	StewartBrand_2006_53	-1.46961480589586	0.17	-0.845737576721772	-0.412240570661216	0.237104947378819	-2.49048800590003
17	15	StewartBrand_2006_217	-1.41051198183423	0.74	-0.759461001413303	0.242678887257582	-0.055775255690844	-1.9830693516808
18	16	StewartBrand_2006_181	-1.33608620338626	0.54	-1.07131785834225	-1.01381677706403	0.709480905970543	-2.711739932822
19	17	StewartBrand_2006_142	-1.15221075075009	0.74	-0.874974157058861	0.658016227095915	-0.706580658385688	-2.07574933909873
20	18	StewartBrand_2006_34	-1.03838308959437	0.31	-0.805322303902855	-0.352422126764678	0.631107016678237	-1.56502050358367
21	19	StewartBrand_2006_51	-0.990225232951568	0.25	-1.03147585180446	-0.879379449543974	0.186277436409205	-2.71480309789079
22	20	StewartBrand_2006_179	-0.948634356760051	0.6	-0.804175771340616	-0.84577011766396	0.891997877179611	-1.70658236858502
23	21	StewartBrand_2006_103	-0.880775558763373	0.41	0.848550917126593	0.220169885172802	-0.221586681336472	-0.03364143780045
24	22	StewartBrand_2006_28	-0.858885623925732	0.73	-0.936027015998076	-0.145832655575603	0.491242502226887	-1.44950279327252
25	23	StewartBrand_2006_65	-0.854507636958205	0.26	-0.832265819115466	-0.324979644770905	0.651900089347624	-1.35985301149695
26	24	StewartBrand_2006_52	-0.836995689088094	0.3	-0.890738979789644	-0.402990295831854	0.813801706457129	-1.31692325825246
27	25	StewartBrand_2006_149	-0.812916760766692	0.14	4.12190138231833	1.48283239938065	-1.24186899069505	3.54994803023724
28	26	StewartBrand_2006_163	-0.729735008383661	0.53	-1.30377733533617	-1.08257715329562	1.38960987268632	-1.72647962432912
29	27	StewartBrand_2006_208	-0.710034067029789	0.39	-1.0624322309849	-0.711641132638218	0.439348679838331	-2.04475875081458
30	28	StewartBrand_2006_169	-0.661876210386983	0.64	-0.888445914665167	-0.022804000345095	-0.792774164750243	-2.36590029014749

Figure 4: Extracted Features For Each Word with Sum Metric

3.4.4 Opinion Score

After getting the important words on the basis of the sum metric, corresponding sentences has been selected. After appending the sentences summary has been generated. After getting the summary for the audio document, summary has been given to few people and asked to rate this summary on 0-10 scale.Every person has been given different opinion score. So the average opinion score of 21 people is 7.07 on 0-10 scale.

4 Plan of Work

Till now the summary is generated for the audio file using four acoustic features and opinion score is collected of summary for validation.

The next plan of the work is mentioned in the following points

- Implement the paper mentioned in reference no 2.
- Get the result for same set of input.
- Compare the results of the paper and proposed solution.

In the paper, they presented a method to summarizing the speech documents. For summarizing the speech documents they used Hidden Markov Model framework. On the basis of variable they decided that the sentence will be included in the summary or not. They tested and trained the model using TDT-2 Corpus. This dataset includes 216 stories of 20 CNN shows. The stories comprising 10 hours of audio data. In modeling , they used 12 acoustic features. The features are speaking rate ; minimum, maximum, mean, range, slope(minimum, maximum) of F0; minimum, maximum, mean, slope of RMS energy; duration of the sentences. They used the Praat software to extract the acoustic features. After generating the summary they compared with human generated summary and calculated precision, recall, and F-Measure for validation to the summary.

References

[1] Sameer Raj Maskey. "Automatic Broadcast News Speech Summarization" Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Graduate School of Arts and Sciences COLUMBIA UNIVERSITY, 2008.

[2] Sameer Maskey, Julia Hirschberg "Summarizing Speech Without Text Using Hidden Markov Models". Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL, pages 89–92,New York, June 2006. .

[3] Kathleen McKeown, Julia Hirschberg, Michel Galley and Sameer Maskey "FROM TEXT TO SPEECH SUMMARIZATION". Conference Paper in Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on , April 2005

[4] John H.L. Hansen and Sanjay Patil "Speech Under Stress: Analysis, Modeling and Recognition". Part of the Lecture Notes in Computer Science book series (LNCS, volume 4343)

[5] Laszlo Czap, Judit Maria Pinter "Intensity Feature for Speech Stress Detection". IEEE Conference paper in 16th International Carpathian Control Conference (ICCC) 2015.

[6] Sumitra Shukla "Spectral Analysis of Stressed Speech for Speech Recognition", Thesis submitted in the Department Of Electronics And Electrical Engineering IIT Guhawati - 781039, India, 2013