Speech Summarization Methods Using Speaker Tracking and Prominence Based Ranking

Thesis submitted in partial fulfillment

of the requirements for the degree of

Master of Science

in

Computer Science and Engineering

Sree Harsha Yella

200402047

sreharshay@research.iiit.ac.in

Search and Information Extraction Lab

International Institute of Information Technology

Hyderabad - 500 032, INDIA

August 2010

International Institute of Information Technology

Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled " Speech summarization methods using speaker tracking and prominence based ranking" by Sree Harsha Yella , has been carried out under my supervision and is not submitted elsewhere for a degree.

Adviser: Dr. Kishore Prahallad	Adviser: Dr. Vasudeva Varma
Date	Date

To my parents and teachers

Acknowledgments

First, I would like to thank my advisor Dr. Kishore Prahallad for his guidance and motivation. I will always be indebted to him for teaching me the basics of research. I am also grateful to him and Speech lab for the support during my thesis work.

I would also like to thank Dr. Vasudeva Varma, for providing an excellent research environment at SIEL, IIIT Hyderabad. I would also like to thank Prof. Yegnanarayana for his stimulating lectures on speech signal processing which prompted me to pursue a research career in this field.

I want to thank all the people in Speech lab for their company during the past years. I thank my colleagues and friends Harish, Sudheer, Avinash, Bapineedu, Venkatesh, Raghavendra, Srinivas and Raju for spending their valuble time to help me in various issues during my masters thesis. I would also like to thank senior members of Speech lab, Dhanunjaya and Guru bhai for clarifying various doubts during my early days at speech lab. I take this opportunity to thank my batch mates of 2K4 batch at IIIT who have made my stay at IIIT memorable.

Last but most important mention is for the endless love and support given to me by my family, my mother Usha Rani and my father Dileep Kumar. Without their continuous support and encouragement I would not have completed this work.

Abstract

Automatic speech summarization is the task of generating a concise summary of a speech signal using a digital computer. The existing speech summarization systems rely on automatic speech recognition (ASR) transcripts and gold standard human summaries to generate summaries of speech signals. The limitations with these approaches are, ASR errors make summaries less usable by humans, also ASR systems are not available for all languages, especially for less resource languages and it takes considerable resources and effort in building one. Gold standard human summaries are not available for all speech signals and building them is tedious and time consuming task. In this work, we propose two techniques for summarization:

1) Exploiting anchor speaker role in broadcast news (BN) show to construct summaries,

2) A generalized ranking of speech segments based on prominence values of syllables in them.

By analyzing manual summaries of news shows, it was found that anchor speaker segments are mostly picked in manual summaries. Therefore it is desirable for automatic summaries to exhibit this characteristic. We proposed two techniques to perform anchor speaker tracking, based on auto associative neural network model and Bayesian information criterion method. Audio summaries are generated for desired summary length by concatenating anchor speaker segments based on their positional features. These summaries are evaluated with the help of ROUGE, an automatic text summarization evaluation pack-

age by transcribing the audio summary into text. ROUGE-N metric measures the N-gram overlap between human reference summaries and the automatic summary. The f-measure scores of the proposed system for ROUGE-1 and ROUGE-2 metrics are 0.561 and 0.392 respectively. These scores showed that the system is capable of generating summaries that are as good as supervised speech summarization system trained using gold standard human summaries which achieved 0.553 and 0.382 for ROUGE-1 and ROUGE-2 metrics respectively. Also, we performed a task based evaluation where, humans were asked to listen to the summary and answer questions regarding the contents of a news show. The percentage of questions answered by the humans was 71 % for the proposed system which is better than 60.2 % of the supervised speech summarization system. The coherence of the summaries was also evaluated by asking the users to rate the summaries on a scale of 1-5 where 1 corresponds to very bad and 5 corresponds to very good. The mean opinion scores (MOS) of these ratings for the proposed method and the supervised speech summarization system are 4.05 and 3.2 respectively. The task based evaluation of these summaries by humans showed that, they prefer the summaries generated by the proposed techniques over the summaries generated by standard speech summarization methods.

In other part of the work, a technique to rank segments in a speech signal using prosodic features that indicate importance is proposed. When humans convey message through speech, they attract listeners' attention to information bearing parts of speech through variations in pitch, amplitude, duration and stress. Speakers make some words prominent and reduce other words. The proposed method computes syllable level prominence values as a function of syllable nucleus duration, sub-band energy (300-2200 Hz), and pitch variation and these values are used to obtain a segment level score, which is used for ranking the segment for summarization. It is shown that this type of scoring captures the prosodic information relevant to summarization in an unsupervised framework. We have also proposed a method to combine lexical and positional features with the prominence based scoring

when text transcripts of speech signals are available. The proposed prominence based scoring captures complimentary information to lexical features derived from text transcripts of speech signals. The combination of these features perform better than the individual features. The proposed method was evaluated on two types of speech data; read style news speech and spontaneous telephone conversations. The proposed system based on prominence scoring achieved ROUGE-1 and ROUGE-2 f-measure scores of 0.508, 0.341 on read style news speech and 0.666, 0.464 on spontaneous conversations respectively. In read style speech the basic unit of extraction was obtained based on pause based segmentation which does not give semantically meaningful segments, where as in spontaneous telephone conversations we have considered speaker turns which are semantically meaningful units as basic unit of extraction.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The amount of multimedia data available has increased rapidly in recent years due to increase in number of information sources and availability of cheap and efficient storage means. Speech data forms a major part of this multimedia data. Speech files belong to different genres such as broadcast news shows, telephone conversations, dialogues, meeting recordings, voice mail and messages, public addressings etc. Users do not have time and patience to go through each document fully. Therefore, in this era of information explosion there is need for systems that can distill this huge amount of data with less complexity and in less time. Automatic summarization systems are a type of such systems, that help in providing most relevant and important data to the user by condensing large amount of data.

1.1 Summarization

The goal of automatic summarization is to take an information source, extract content from it, and present the most important content to the user in a condensed form and in a manner sensitive to the user's or application's needs. Sparck Jones [36] defines summary as a condensed derivative of source, i.e. reduction of content through either selection or generalization on what is important in the source. In general, the functions of a summary include

- Announcement: announce existence of the original document
- Screening: determine the relevance of the original document
- Substitution: replace the original document
- Retrospection: point to the original document

Depending on the length and requirement of the summary some of these can be included while discarding others. Summaries are influenced by a broad range of factors. Sparck Jones [37] broadly classified these factors into three types:

- **Input factors**: source form, subject type and unit

  Input factors characterize the properties of the input to summarization system. Source form subsumes structure of the document, scale of document, medium and genre. Based on the presumed knowledge of the reader, the subject type can be broadly classified as ordinary, specialized and restricted. Unit distinguishes whether single unit or multiple units need to be summarized.

- **Purpose factors**: situation, audience and use

  Situation refers to the context with in which the summary is intended to be used, while audience can be further be categorized into targeted or general.

- **Output factors**: material, format, style, expression and brevity

  Material characterizes the relation of the summary to the source text, whether it covers all main concepts of the summary or only some concepts of the summary. Format and expression refers to the representation of the summary while style refers to the function (indicative, informative etc.) of the summary. Brevity tells the condensation ratio of the text.

1.1.1 Basic Notions of Summarization

A number of basic notions of summarization depend on the type of relationship between the summary and its input. A fundamental distinction in summaries is between **extracts** and **abstracts**. An extract is a summary consisting entirely of material copied from the input. Thus, a typical extract at a condensation rate of 25 % will take some 25 % of the material in the document. The basic units for extraction can be words, phrases, sentences or paragraphs. The choice of extraction unit is also determined by the condensation rate. An abstract is a summary at least some of whose material is not present in the input. Typically an abstract contains some degree of paraphrase of input content. In general, abstracts offer the possibility of higher degrees of condensation: a short abstract may offer more information than a longer extract.

Another way to look at summaries is in terms of **indicative** and **informative** summaries. An indicative summary provides a reference function for selecting documents for more in depth reading. Thus, an indicative abstract is aimed at helping the user to decide whether to read the information source or not. An informative summary covers all salient information in the source at some level of detail. The distinction between indicative and informative summaries can be extended to a three way distinction, between indicative, informative and **critical evaluative** summaries. A critical summary evaluates the subject matter of the source, expressing the abstractor's views on the quality of work of the author.

User focused (or **topic focused** or **query focused**) summaries are tailored for the requirements of a particular user or group of users. This means that the summary takes into account some representation of users' interests, which can range from full blown user models to profiles recording subject area terms or even a specific query containing terms that are deemed to express users' information need. **Generic summaries** are aimed at a particular usually broad readership community. Traditionally, generic summaries written by

authors or professional abstractors served as surrogates for full text. These summaries can be indicative or informative in nature.

Summaries may be of a single input document, or of multiple documents, as in the case of multi-document summarization (MDS). In MDS, the summarizer identifies what is common across the documents, or different in a particular one.

1.2 Speech Summarization

The aim of speech summarization is to generate a concise summary of a given speech signal. Speech is the most natural way of communication among human beings and it encodes various aspects of communication. A Speech signal contains linguistic, para-linguistic and extra-linguistic information. Linguistic information indicates the direct meaning of the spoken utterance. Para-linguistic information indicates speaker's current affective such as tone of voice and emotion. Extra-linguistic information indicates speaker specific information such as physiological features of vocal tract system, pitch range, cultural and social background. A speech summarization system must aim at modelling and capturing all these sources of information in a speech signal in order to summarize it effectively.

Speech summaries can be produced in the form of text or audio. Summaries in the form of text contain errors due to automatic speech recognition (ASR) and also they do not carry para-linguistic and extra-linguistic information conveyed by a speech signal. But these summaries have an advantage that they can easily be indexed and stored for further retrieval and also information extraction and retrieval techniques can easily be applied on them to serve users' information need. Summaries in audio form can be generated in two ways; by synthesizing the output text summary into speech and the by concatenating important parts of original speech signal. The state of art speech synthesizers can produce speech that is intelligible but are still far off in synthesizing speech with natural variations. Therefore,

speech summaries in the form of audio are generally extractive summaries where, important segments in the speech signal are identified, ranked and concatenated without any alterations to form a summary. Abstractive summaries are relatively harder for a machine to generate as they require additional knowledge resources such as ontologies to provide a degree of generalization, or linguistic knowledge to construct sentences.

1.2.1 Issues in Speech Summarization

Text documents have word, sentence and paragraph boundaries defined which makes it easier to choose the desired processing unit reliably. Speech, however, is one long stream of audio signal with none of these boundaries. Such lack of segmentation makes it difficult to process speech in meaningful semantic units. This problem is typically addressed by employing speech segmentation algorithms.

In order to process speech documents we need to convert speech signals into a sequence of words that is meaningful to users. Automatic Speech Recognition (ASR) engines that convert speech to text have limited accuracy, even though they have improved in recent years. Poor accuracy effects speech summarizers because word errors degrade the overall performance of a system that assumes well-formed sentences as input.

Another problem faced by speech summarization systems is disfluency. Even though humans write well-formed grammatical sentences, when they speak they repeat or repair phrases, insert filled pauses such as uh, and oh. Text-trained natural language processing (NLP) tools such as parsers and taggers suffer with reduced accuracy on speech documents because of such disfluency. Not having adequate NLP tools that work well with speech, added to other problems of processing speech, makes summarization of speech more challenging than text summarization.

Even though such problems make speech summarization harder, there is extra information available in speech that does not exist in text documents. Speech has acoustic information that may help in identifying topic shifts or acoustically significant segments. Spoken documents such as news broadcasts tend to have multiple speakers who play different roles in the broadcasts. Identifying these roles may provide cues to the structure of a broadcast, and can be exploited to deduce the significance of segments for extractive summarization. Also, a speakers emphasis of particular segments of speech may indicate the significance he or she attaches to that segment.

1.3 Problem Statement

Automatic speech summarization systems depend on ASR transcripts and gold standard human summaries to produce automatic summaries. This thesis focusses on speech summarization methods that do not depend on ASR transcripts and gold standard human summaries. In this thesis we aim to summarize two types of speech data; broadcast news speech and spontaneous conversations. To summarize broadcast news we aim to use speaker roles to extract segments relevant to summary. To summarize any given speech signal such as spontaneous conversations which have no explicit structure, we aim to rank speech segments using acoustic features that indicate important content in the speech signal. A detailed analysis of these techniques is to be performed by comparing them with baseline text summarization system and state of the art speech summarization system.

1.4 Outline of Speech summarization Approaches

Speech summarization systems can be broadly classified into two categories. One type of systems take ASR output of speech signals and apply automatic text summarization

approaches on it to obtain summary of speech signals and the other type of systems train a classifier using various features such as acoustic, lexical, structural and discourse features that can be derived from speech signal and corresponding text transcript. But these type of systems require gold standard human summaries to train the classifier. The block diagram of these two types of systems are shown in Fig. 1.1 and 1.2 respectively.

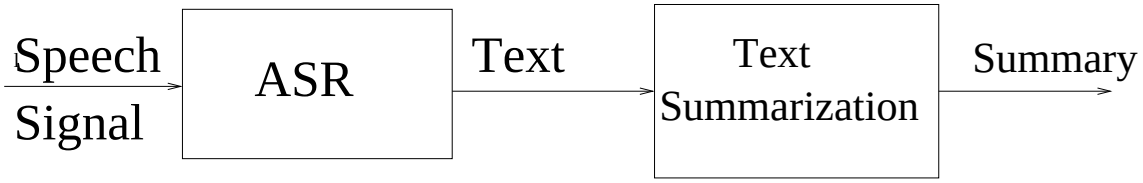

Figure 1.1 *Block diagram of speech summarization system based on text summarization approaches.*

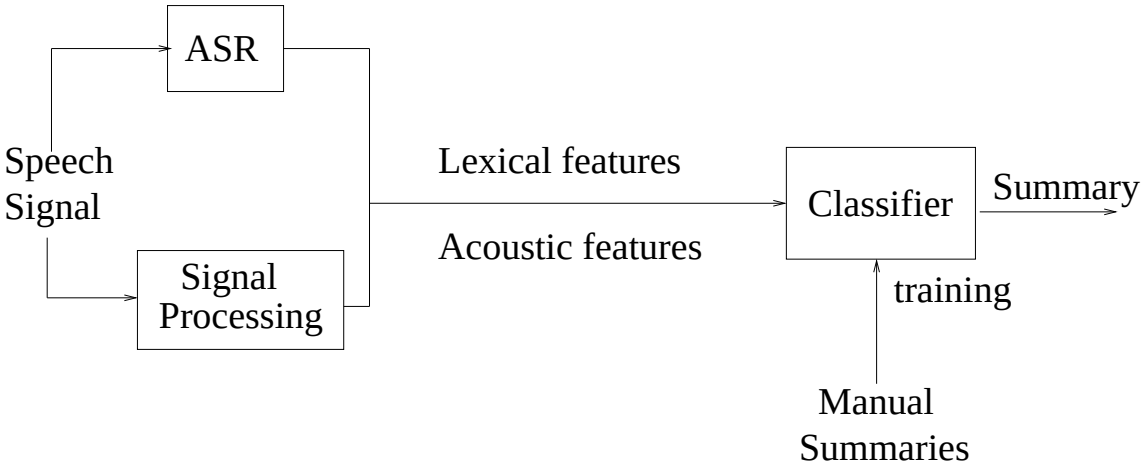

Figure 1.2 *Block diagram of supervised speech summarization system.*

1.4.1 Outline of Proposed Approaches

In this thesis, we propose two methods; exploiting the role of anchor speakers to summarize broadcast news shows and to rank speech segments based on prominence based features to summarize a given speech signal.

1.4.2 Broadcast News Summarization by Anchor Speaker Tracking

The block diagram of the proposed summarization system is shown in Fig. 1.3. We

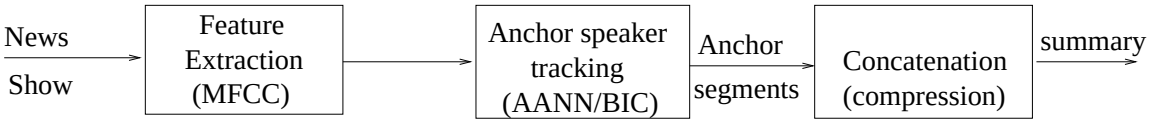

Figure 1.3 *Block diagram of the proposed broadcast news summarization system using anchor speaker tracking.*

analyzed human summaries of broad cast news shows, and found that most of the segments picked in human summaries contain anchor speaker segments. Also agreement between human annotators is more in anchor speaker segments. It was also observed that anchor speaker segments in the beginning of a news story were picked in almost all human summaries. This shows that human annotators prefer anchor speaker segments to other speakers in a news show and anchor speaker segments at the beginning of a news story are very relevant to the summary. We propose two techniques to perform anchor speaker tracking. The first technique is based on auto associative neural network model [101] which performs speaker tracking of a modelled speaker. In the training phase, the model is trained on speech of anchor speaker and that model is used for tracking his/her speech in the news show in testing phase. In the second technique, the broadcast news show is first segmented into homogeneous regions containing speech of a single speaker using Bayesian information criterion (BIC) [6] method. After obtaining single speaker segments, agglomerative clustering of these segments is done using BIC as a distance measure. The cluster containing highest number of speaker turns is hypothesized as the cluster containing anchor speaker segments. This technique does not require initial training data to perform speaker tracking and also it can easily be extended to multiple anchor speakers. After obtaining anchor speaker segments, isolated segments are filtered out and continuous regions containing his/her speech are considered as beginning of a news story in the news show.

The required summary length is divided among the anchor speaker regions (approximately equal to number of news stories in the show) and segments from the beginning of each anchor speaker region are concatenated in the order of their occurrence in the news show to generate the summary. These summaries are evaluated with the help of ROUGE, an automatic text summarization evaluation package by transcribing the audio summary into text. ROUGE-N metric measures the N-gram overlap between human reference summaries and the automatic summary. The f-measure scores of the proposed system for ROUGE-1 and ROUGE-2 metrics are 0.561 and 0.392 respectively. These scores showed that the system is capable of generating summaries that are as good as supervised speech summarization system trained using gold standard human summaries which achieved 0.553 and 0.382 for ROUGE-1 and ROUGE-2 metrics respectively. Also, we performed a task based evaluation where, humans were asked to listen to the summary and answer questions regarding the contents of a news show. The percentage of questions answered by the humans was 71 % for the proposed system which is better than 60.2 % of the supervised speech summarization system. The coherence of the summaries was also evaluated by asking the users to rate the summaries on a scale of 1-5 where 1 corresponds to very bad and 5 corresponds to very good. The mean opinion scores (MOS) of these ratings for the proposed method and the supervised speech summarization system are 4.05 and 3.2 respectively.

1.4.3 Prominence based Ranking of Speech Segments

The block diagram of the proposed summarization system is shown in Fig. 1.4. In order to summarize any given speech file such as spontaneous conversations, we propose a ranking method that ranks the segments based on acoustic features indicating importance of the segment. The speech segments are ranked with the help of prominence values of the syllables present in them. Prominence is defined as perceptual salience of a language

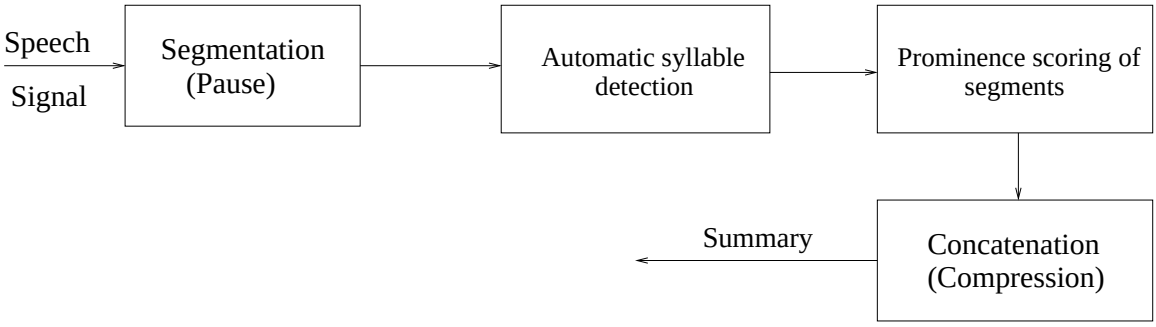

Figure 1.4 *Block diagram of the proposed speech summarization system using prominence based ranking.*

unit. It was shown by previous studies that prominent words occur while introducing new concepts and it is widely accepted that content words (nouns, adjectives, adverbs) are made prominent than function words (conjunctions, inter-junctions). The prominence value of a syllable was computed as a function of syllable nucleus duration, sub-band energy, and pitch variation. We experimented with four scoring functions to obtain an acoustic score for each segment from prominence values of syllables in the segment. The speech segments are ranked using these acoustic scores and top ranking segments are concatenated in chronological order of occurrence in the speech file to form a summary. The experiments were carried out on both read style news speech and spontaneous telephone conversations. These summaries are evaluated in two ways; one using ROUGE evaluation package and the other, task based evaluation by humans. The proposed system achieved a ROUGE-1 and ROUGE-2 scores of 0.508, 0.341 on read style news speech and 0.666, 0.464 on spontaneous conversations respectively. In read style speech the basic unit of extraction was obtained based on pause based segmentation which does not give semantically meaningful segments, where as in spontaneous telephone conversations we have considered speaker turns which are semantically meaningful units as basic unit of extraction. The proposed system performed better than baseline text summarization system based on tf*idf scores of ASR transcripts of speech files and supervised speech summarization system trained us-

ing gold standard human reference summaries. The summaries generated by the proposed method achieved higher recall scores as target summary length increased without significant fall in precision scores. This shows that the system is capable of generating summaries of different length without degradation in the quality of the summaries. .

1.5 Thesis Organization

Chapter 2 presents a review of summarization techniques used earlier in the fields of text summarization and speech summarization. Chapter 3 presents the proposed method to summarize broadcast news shows using anchor speaker tracking techniques. Chapter 4 presents a method to rank speech segments for summarization using prominence based features. The proposed method is compared with a baseline text summarization system and a standard speech summarization system. Chapter 5 presents summary and conclusions of the thesis with future directions.

Chapter 2

Overview of Previous Work in Summarization

2.1 Review of Text Summarization Methods

Luhn's work [50] focussed on recognizing keywords in text is among the earliest works on automatic text summarization. Lush showed that the words with highest resolving power are words with medium or moderately high frequency in a given document. A decade later, Edmundson [13] began to look beyond keywords for the summarization of scientific articles. He focused on four features: cue phrases, keywords, title words, and location. Cue phrases are phrases that are very likely to signal an important sentence, and could include phrases such as 'significantly', 'in conclusion' or 'impossible' in the scientific articles domain. There are also Stigma phrases that may signal 'negative relevance': specifically, these might be hedging or belittling expressions. The Title feature, weights each sentence according to how many times its constituent words occur in section or article titles. The Location feature weights sentences more highly if they occur under a section heading or occur very early or late in the article. Edmundsons summarization system works by scoring and extracting sentences based on a linear combination of these four features. The weights associated with these features are manually tuned depending on the corpus. Similar features are used today in machine learning frameworks.

The ADAM system of the 1970s (Rush et al., [82]; Mathis, [58]; Pollock and Zamora, [72]) relies on cue phrases, but its goal is to maximize coherence by analyzing whether a candidate sentence contained anaphoric references [14]. In the case that a candidate does contain anaphoric references, the system tries to either extract the preceding sentences as well or to rewrite the candidate sentence so that it could stand alone. If neither of these are possible, the candidate is not chosen.

In the late 1970s and early 1980s, Paice [69] investigated the idea of using self indicating phrases to detect informative sentences from journal papers. These phrases explicitly signal that a sentence is relevant to the document as a whole, e.g. 'This report concerns...'. Contemporary work by Janos [34] divided documents into meta-text and the text proper. Janos found that while most meta-text could be discarded in the summarization process, certain thematical meta-text sentences were able to form a semantic nucleus for the summary as a whole. The summarization work of Paice is also similar to the ADAM summarization system in its treatment of exophoric sentences. The primary difference is that Paice evaluated both anaphoric and cataphoric references.

In the 1980s, several summarization methods that were inspired by findings in psychology and cognitive science (DeJong, [12]; Fum [18]; Jacobs and Rau [32]) were proposed. These methods use human processing and understanding of text as a model for automatic summarization. The source is interpreted and inferences are made based on prior knowledge. For an automatic summarization method, a schemata is created relating to the domain of the data being summarized. The major difference between these methods and the earlier summarization methods described above is that the input is interpreted and represented more deeply than before. For example, the FRUMP system [12] uses sketchy scripts to model events in the real world for the purpose of summarizing news articles. For example, a sketchy script relating to earthquakes contains entries, such as the magnitude on the Richter scale, the location of the epicenter, the number of deaths and the amount of dam-

age inflicted. When a particular sketchy script is activated, these pieces of information are sought in the source data. These approaches are limited by being very domain specific and requiring prior knowledge about the data being summarized. Further information on such approaches can be found in [15].

In late 1980s, summarization research underwent a major resurgence primarily due to the explosion of data available from sources such as the web. Due to the volume and variety of data to be summarized, the summarization techniques were more often extractive than abstractive. Extractive summaries are more domain independent, require little or no prior knowledge, and can process a large amount of data efficiently. Therefore, the methods for summarization tended to move away from the schema based, cognition inspired approaches of the 1980s. Much of the work of this period revisited the seminal work of Edmundson [13] and his investigation of cue phrases, keywords, title words, and location features. The newer work incorporated these same features into machine learning frameworks where classifiers are trained on human gold standard extracts (Kupiec [44]; Teufel Moens, [94]), rather than manually tuning the weights of these features as in the work of Edmundson. For the tasks of summarizing engineering papers [44] and computational linguistics papers [94], the most useful features were found to be cue phrases and locational features.

Other researchers investigated the use of rhetorical relations for the purpose of text summarization, particularly in the framework of Rhetorical Structure Theory (RST) (Mann Thompson, [52]). A hypothesis of RST is that a given document can be represented as a single binary branching rhetorical tree comprised of nuclei satellite pairs, where a particular rhetorical relation exists between each nuclei satellite pair. By pruning such a rhetorical tree, a summary of the entire text can be generated [68, 53, 54].

Contemporary work utilized linguistics resources such as WordNet, a database of lexical semantics, in order to derive relations between terms or phrases in a document. In work by Barzilay and Elhadad [2] lexical chains were detected according to the relatedness of

document terms, and sentences corresponding to the strongest chains were extracted. The SUMMARIST system [28] utilizes WordNet for concept detection in the summarization of news articles.

Also in the late 1990s, interest in multi document summarization was growing. Creating a single summary of multiple documents presented, and still presents, an interesting challenge. Given a set of relevant documents to the query, the summarizer must not extract the same information from multiple sources and identify unique information present in each document. Carbonell and Goldstein [5] introduced the Maximal Marginal Relevance (MMR) algorithm, which scores a candidate sentence according to how relevant it is to a query (or how generally relevant, for a generic summary) and how similar it is to sentences that have already been extracted. The latter score is used to penalize the former, thereby reducing redundancy in the resultant summary. MMR remains popular both as a stand alone algorithm in its own right as well as a feature score in more complex summarization methods [105]. Work by Radev [77, 76] addressed single and multi document summarization by a centroid method. A centroid is a pseudo document consisting of important terms and their associated term weight scores, representing the source document(s) as a whole. The authors address the redundancy problem by the idea of cross sentence information subsumption, whereby sentences that are too similar to other sentences are penalized, similar to the MMR method.

The work of Maybury [59] extended summarization work from merely processing and summarizing text to summarizing multi modal event data. In the domain of battle simulation, the researchers took as input battle events such as missile fire, refuelling, radar sweeps and movement, and generated summaries based on the frequencies of such events and relations between such events. Not only are the inputs multi-modal events, but the output can be a combination of textual and graphical summaries in order to give a quick perception and comprehension of the battle scene. The researchers also took into account that such

summaries should be tailored to the user: for example, an intelligence officer might care more about enemy size and position whereas a logistician will care about refuelling and supplies.

Since 2001, the Document Understanding Conference has encouraged research in the area of multi document, query dependent summarization. For the text summarization community, this annual conference provides the benchmark tasks for comparing and evaluating state of the art summarization systems. While the data used has primarily been news wire data, DUC has recently added tracks relating to the summarization of web-log opinions. Though a wide variety of systems have been entered in DUC, one finding is that the most competitive systems have extensive query expansion modules [33, 71, 96]. In fact, query expansion forms the core of many of the systems [29].

Automatic text summarization is closely intertwined with automatic text retrieval, and this connection can especially be seen in query dependent summarization, wherein a query and a document or set of documents must be represented in such a way that similarity between the query and a candidate document or sub-document can be gauged. A major difference between the tasks of text retrieval and query-dependent summarization is that text retrieval in its basic form concerns the determination of whether or not a document is relevant to a query, whereas summarization goes a step further and condenses the relevant documents. The basic formulation of the text retrieval task is that there is an archive of documents, a user who generates a query, and a process of retrieving the documents in the archive that satisfy the querys information need [79]. An efficient way of representing queries and documents is by a vector space representation where words are associated with term weights, with an example weighting scheme being tf.idf [38, 79, 83], where a word has a high score if it occurs often in the candidate document but rarely across the set of documents.

The vector space representation is useful because if both the query and candidate document are represented as vectors, similarity can be easily gauged using the cosine of the two vectors. Alternatively, probabilistic information retrieval systems [55, 79] estimate the probability of relevance for a document D, $P(R|D)$. This is arrived at using Bayes theorem, with probability $P(D|R)$ equal to the product of the individual term probabilities in the simplest formulation [86]

$$P(D|R) = \prod P(t_i|R) \times \prod (1 - P(t_j|R)) \tag{2.1}$$

where t_i is a term common to the query and the document and term t_j is a term present in the query but missing from the document. Since realistically the relevance information is not known, there are numerous methods for estimating the probability of a term given the relevance information, and Croft and Harper [11] illustrate an estimation method that is closely approximated by inverse document frequency [38].

Automated information retrieval as a field took root in the 1940s with the germinal work of Bush [4], and it was Luhn [50], mentioned above, who put forth the idea that words could act as indices for documents in a collection. Probabilistic information retrieval was developed in the early 1960s [55], and further refined in the 1970s and 80s [38, 11]. Since the early 1990s, the Text Retrieval Conference (TREC) [21] has encouraged the development of effective retrieval methods for large corpora [86].

For Further overview of text summarization research and directions, see [51, 15, 37].

2.2 Text to Speech Summarization

McKeown [60] provided an overview of text summarization approaches and discussed how text-based methods might be extended to speech data. The authors described the challenges in summarizing differing speech genres such as broadcast news and meeting

speech and which features are useful in each of those domains. Their summarization work involved components of speaker segmentation, topic segmentation, detection of agreement/disagreement, and prosodic modelling, among others. For meetings in particular, their research involved finding the prosodic and lexical correlates of topic shifts, and they investigated known useful features of monologue speech such as pauses and cue phrases and concluded that these are informative for segmenting multi-party dialogue speech as well.

Christensen [7] investigated how well text summarization techniques for news-wire data could be extended to broadcast news summarization. In analyzing feature subsets, they found that positional features were more useful for text summarization than for broadcast news summarization and that positional features alone provided very good results for text. In contrast, no single feature set in their speech summarization experiments was as dominant, and all of the features involving position, length, term-weights and named entities made significant contributions to classification. They also found that increased word-error rate (WER) only caused slight degradation according to their automatic metrics, but that human judges rated the error filled summaries much more severely.

In the following sections we first provide an overview of early research on speech summarization, then describe speech summarization research from four particular domains: newscasts, meetings, lectures, and voice-mail.

2.3 Review of Speech Summarization Methods

In the early 1990s, simultaneous with the development of improved automatic speech recognition, researchers became increasingly interested in the task of automatically summarizing speech data. Here we describe several early summarization projects from a variety of speech domains.

Rohlicek [80] created brief summaries, or gists, of conversations in the air traffic control domain. The basic summarization goals were to identify flight numbers and classify the type of flight, e.g. takeoff or landing. Such a system required components of speaker segmentation, speech recognition, natural language parsing and topic classification. The authors reported that the system achieved 98% precision of flight classification with 68% recall.

One of the early projects on speech summarization was VERBMOBIL [78], a speech-to-speech translation system for the domain of travel planning. The system is capable of translating between English, Japanese and German. Though the focus of the project was on speech-to-speech translation, an abstractive summarization facility was added that exploited the information present in the translation modules knowledge sources. A user can therefore be provided with a summary of the dialogue, so that they can confirm the main points of the dialogue were translated correctly, for example. The fact that VERBMOBIL is able to incorporate abstractive summarization is due to the fact that the speech is limited to a very narrow domain of travel planning and hotel reservation; normally it would be very difficult to create such structured abstracts in unrestricted domains.

Simultaneously work was being carried out on the MIMI dialogue summarizer [39], which was used for the summarization of spontaneous conversations in Japanese. Like VERBMOBIL, these dialogues were in a limited domain; in this case, negotiations for booking meetings rooms. The system creates a running transcript of the transactions so far, by recognizing domain specific patterns and merging redundant information.

2.3.1 Summarization of Newscasts

One of the domains of speech summarization that has received the most attention and has perhaps the longest history is the domain of broadcast news summarization. Summa-

rizing broadcast news is an interesting task, as the data consists of both spontaneous and read segments and so represents a middle ground between text and spontaneous speech summarization. In [24], a user interface tool is provided for browsing and information retrieval of spoken audio in this case, using TREC-7 SDR data [97]. The browser adds audio paragraphs, or paratones, to the speech transcript, using intonational information. This is a good example of how structure can be added to unstructured speech data in order make it more readable as well as more amenable to subsequent analysis incorporating structural features. Their browser also highlights keywords in the transcript based on acoustic and lexical information.

Another example of adding structure to speech data is in the work of [1]. The authors focus on classifying speaker roles in radio broadcasts, automatically discerning between anchors, journalists and program guests using lexical and durational cues. This speaker role identification can be valuable for quickly indexing a large amount of broadcast data and especially for finding the transitions between stories.

In [95], summarization of the American Broadcast News corpus was carried out by weighting terms according to an acoustic confidence measure and a term-weighting metric from information retrieval called inverse document frequency. The units of extraction are n-grams, utterances and keywords, which in the case of n-grams and utterances are scored according to the normalized sums of their constituent words. When a user desires a low word-error rate (WER) above all else, a weighting parameter can be changed to favor the acoustic confidence score over the lexical score. One of the most interesting results of this work is that the WER of summaries portions are typically much lower than the overall WER of the source data, a finding that has since been attested in other work [62]. [95] also provide a simple but intuitive interface for browsing the recognizer output.

In work by Hori and Furui [25] on Japanese broadcast news summarization, each sentence has a subset of its words extracted based on each words topic score a measure of

its significance and a concatenation likelihood, the likelihood of the word being concatenated to the previously extracted segment. Using this method, they reported that 86% of the important words in the test set are extracted.

[42] used a series of multi-layer perceptrons to summarize news casts, by removing ASR errors according to recognizer confidence scores and then selecting units at increasing levels of granularity, based on term weighting and Named Entity features. They found that their summarizer performed very well according to a question answering evaluation and ROUGE analysis, but slightly less well on subjective fluency criteria.

More recently in the broadcast news domain, Maskey and Hirschberg [56] found that the best summarization results in this domain utilized prosodic, lexical and structural features, but that prosodic features alone resulted in good quality summarization. The prosodic features they investigated were broadly features of pitch, energy, speaking rate and sentence duration. Work by [67] explored using only prosodic features for speech-to-speech summarization of Japanese newscasts, finding that such summaries rated comparably with a system relying on speech recognition output.

[8] have developed a system for skimming broadcast news transcripts, consisting of three steps of automatic speech recognition, story and utterance segmentation, and determination of the most informative utterances, which are then highlighted in the transcript. Salience is determined by features of position, length, tf.idf score and cosine similarity of utterance and story term vectors. They evaluated their system both intrinsically with recall, precision and f-score, and extrinsically by a question-answering task. Two relevant findings are that ASR did not seriously affect the determination of salience, but that errors in story segmentation had a detrimental impact on downstream processes.

2.3.2 Summarization of Meetings

In the domain of meetings, [98] implemented a modified version of MMR applied to speech transcripts, presenting the user with the n best sentences in a meeting browser interface. The browser contained several information streams for efficient meeting access, such as topic tracking, speaker activity, audio/video recordings and automatically generated summaries. However, the authors did not research any speech specific information for summarization; this work was purely text summarization applied to speech transcripts.

Zechner [103] investigated summarizing several genres of speech, including spontaneous meeting speech. Though relevance detection in his work relied largely on tf.idf scores, Zechner also explored cross speaker information linking and question/answer detection, so that utterances could be extracted not only according to high tf.idf scores, but also if they were linked to other informative utterances. This work also focused on detecting disfluencies such as filled pauses, false starts and repairs in order to increase summary readability and informativeness.

On the ICSI corpus, Galley [20] used skip-chain conditional random fields to model pragmatic dependencies such as question-answer between paired meeting utterances, and used a combination of lexical, prosodic, structural and discourse features to rank utterances by importance. The types of features used were classified as lexical features, information retrieval features, acoustic features, structural and durational features and discourse features. Galley found that while the most useful single feature class was lexical features, a combination of acoustic, durational and structural features exhibited comparable performance according to Pyramid evaluation.

Simpson and Gotoh [85], also working with the ICSI meeting corpus, investigated speaker-independent prosodic features for meeting summarization. A problem of working with features relying on absolute measurements of pitch and energy is that these features

vary greatly depending on the speaker and the meeting conditions, and thus require normalization. The authors therefore investigated the usefulness of speaker-independent features such as pauses, pitch and energy changes across pauses, and pitch and energy changes across units. They found that pause durations and pitch changes across units were the most consistent features across multiple speakers and multiple meetings.

[49] reported the results of a pilot study on the the effect of disfluencies on automatic speech summarization, using the ICSI corpus. They found that the manual removal of disfluencies did not improve summarization performance according to the ROUGE metric. Zhu and Penn [105] showed how disfluencies can be exploited for summarization purposes and found that non-lexicalized filled pauses were particularly effective for summarizing SWITCHBOARD speech.

[62, 64] compared text summarization approaches with feature based approaches incorporating prosodic features, with human judges favoring the feature based approaches. In subsequent work [65], they began to look at additional speech specific characteristics such as speaker and discourse features. One significant finding of these papers was that the ROUGE evaluation metric did not correlate well with human judgements on the ICSI test data.

2.3.3 Summarization of Lectures

[26] developed an integrated speech summarization approach, based on finite state transducers, in which the recognition and summarization components are composed into a single finite state transducer, reporting results on a lecture summarization task. Summarization accuracy results (word accuracy between an automatic summary and the most similar string from the referent summary word network) were reported, with scores in the range of 25-40 for a 50% summarization ratio and 35-56 for the 70% summarization ratio. Also in

the lectures domain, [17] attempted to label cue phrases and use cue phrase features in order to supplement lexical and prosodic features in extractive summarization. They reported that the use of cue phrases for summarization improved the summaries according to both f-scores and ROUGE scores. [104] compared feature types for summarization across domains, concentrating on lecture speech and broadcast news speech in Mandarin. They found that acoustic and structural features are more important for broadcast news than for the lecture task, and that the quality of broadcast news summaries is less dependent on ASR performance.

2.3.4 Summarization of Voicemail

The SCANMail system [23] was developed to allow a user to navigate their voicemail messages in a graphical user interface. The system incorporated information retrieval and information extraction components, allowing a user to query the voicemail messages, and automatically extracting relevant information such as phone numbers. [31] and Jansche and Abney [35] also described techniques for extracting phone numbers from voicemails. Koumpis and Renals [43] investigated prosodic features for summarizing voicemail messages in order to send voicemail summaries to mobile devices. They reported that while the optimal feature subset for classification was the lexical subset, an advantage could be had by augmenting those lexical features with prosodic features, especially pitch range and pause information.

2.3.5 Summary

The speech summarization approaches have explored various lexical, prosodic, structural and discourse features for summarization. It was also shown that prosodic information alone can be useful in generating summaries that are as good as summaries based on

other features. One common feature of the speech summarization systems using prosodic features is that they require gold standard human reference summaries to train a supervised classifier that classifies a given utterance as belonging to summary or not. In the current work, we propose a method to rank the speech segments based on prominence values of syllables in the segments to capture the prosodic information relevant to summarization in an unsupervised framework. Also, most of the speech summarization systems provide summaries in the form of text, which requires an ASR system which introduces errors. It was reported in many studies that though automatic metrics do not show a great degradation due to ASR errors, human evaluators penalize summaries with ASR errors more severely. Also summaries in the form of text do not provide extra information in the form of para-linguistic and extra-linguistic information which summaries in the form of speech do. In this current thesis, we aim to generate summaries in audio form and in the case of broadcast news shows we explore the importance of anchor speaker segments for summarization and produce summaries that are acceptable and useful to humans.

Chapter 3

Broadcast News Summarization Using Anchor Speaker Tracking

3.1 Introduction

Broadcast news (BN) is one of the most common media through which people obtain news besides newswire. BN contains one or more speakers presenting, discussing or analyzing current events that are deemed important. A team of producers, screenwriters, audio and video editors, reporters and anchors are involved in the production of BN and they generally follow a standard format of news reporting. Most BN shows contain a sequence of reports on significant current events followed by some commercials, weather, sports and entertainment news. This standard formatting of BN can be useful for automatic processing of BN.

In this thesis, we propose an approach to summarize BN using anchor speaker tracking. We analyze human reference summaries of BN and find that anchor speaker segments are important as they are picked in most of human reference summaries. We propose two methods to perform anchor speaker tracking; 1) based on auto-association neural network (AANN) model [101] and 2) based on Bayesian information criterion (BIC) technique [6]. Once the segments of anchor-speaker's speech are extracted, a summary is obtained for de-

sired compression ratio by using positional features of these segments. The summaries are provided in audio format as it prevents errors due to automatic speech recognition (ASR) and preserves characteristics of natural speech. The idea lies in exploiting the characteristics of broadcast news, where a specific structure is followed to deliver the news content. We make use of the fact that in broadcast news, there is a pattern of anchor-speaker and on-field reporter taking turns to cover each story.

Broadcast news show follows a certain structure depending on the genre of the show. Most of the broadcast news have an anchor speaker who starts the show by reading the headlines and then presents each story where reporters and others speakers may be involved. Our approach assumes following structure of a broadcast news show.

- Anchor (Headlines): .....
- Anchor (Story 1): Its not often when an US president quotes lines...
- Reporters and other speakers: ....
- Anchor (Story 2): Its several days now since opposition leader....
- Reporters and other speakers: .....

Our aim is to find the segments in the news show, that when concatenated together form a meaningful and coherent audio summary that is acceptable and useful for humans. The summaries generated by current techniques will be indicative or informative, extractive summaries.

3.2 Data Set

3.2.1 BBC News Corpus

All the news shows used in the experiments belong to globalnews podcast of BBC podcasts[1] available on-line. The show provides a daily update of global news and features different anchor speakers. We have used a total of 20 news shows each around 30 min of duration. Each show was sampled at $16\ kHz$ and contains a single anchor speaker and multiple other speakers. There are a total of eight anchor speakers in 20 shows, of which three are male and five are female speakers.

3.2.2 Human Reference Summaries

The text transcripts of the speech files along with their corresponding audio are presented to 4 human annotators for constructing a summary. All the annotators are graduate students with a good background of English. The annotators were instructed to generate a summary of five minutes in length. They were instructed to pick meaningful phrases or sentences present in original story without altering them. Their aim was to generate a generic extractive speech summary that is coherent and meaningful. These human summaries are used to study how humans perform summarization of broadcast news and also for evaluating the automatic summaries. The standard evaluation setup for text summarization at document understanding conference (DUC)[1] uses 4 human reference summaries. The number of human reference summaries used in this work was fixed following DUC framework.

[1]http://www.bbc.co.uk/podcasts/series/globalnews/
[1]http://www-nlpir.nist.gov/projects/duc/guidelines.html

3.3 Analysis of Human Reference Summaries

The way in which human abstractors perform summarization may help us a great deal in building automatic summarization systems [51]. Professional abstractors do not focus on understanding a document for summarizing it, instead they make use of the properties of structure of the document such as title, position of a sentence in the paragraph (beginning and ending) and also cue phrases to find important parts in the document. Once they have found the parts of the document that describe the content of the document, they construct simple sentences on the contents of these segments to present it as an abstract. Hence, to summarize any document it is important to first find informative sections in the document.

In order to study how humans perform summarization of broadcast news, we have asked four graduate students with good English knowledge to summarize each news show in the data set. They were instructed to generate a five minute generic summary for each show. These audio summaries are transcribed into text manually for analysis purpose. Given these multiple human reference summaries for a news show, it would be interesting to observe the measure of overlap between them and also type of segments present in the overlap. This would help us to identify the features in the input that humans use and agree on, to pick segments in summary. If such features can be identified, it would help in design of automatic summarization systems.

As anchor speaker performs an important task of delivering news and running the show, we investigate his/her contribution to human reference summaries. Tab. 3.1 shows the % of anchor speaker sentences (An) in human summaries, % of sentences picked in all human summaries which indicates overlap (Ov) among human summaries, % of anchor speaker sentences in the overlap (An_Ov) and % of initial sentences (first two) in each news story (In) that are picked in human summaries.

Table 3.1 *Statistics of human summaries averaged over 20 news shows.*

type	An	Ov	An_Ov	In
%	74%	63%	92%	89%

Tab. 3.1 shows that human annotators give importance to anchor speaker utterances while summarizing and they also have a good agreement on this (92 % of the segments in the overlap belong to anchor speaker segments). The bias of human annotators towards anchor speaker segments may be due to their preciseness and salience which are essential for an audio summary. Also the picking of 89% of initial sentences in a story (In) shows the importance of anchor speaker utterances in the starting of story.

3.4 Anchor Speaker Tracking

In this section we present two techniques for anchor speaker tracking and features used for speaker tracking.

3.4.1 Feature Extraction from Speech Signal

To perform speaker tracking, speaker-specific features are extracted from the speech signal. Typically these features represent the short-time spectral information such as mel-frequency cepstral coefficients (MFCCs) which describe the vocal tract properties of an individual broadly [74]. In our study, 13 MFCC features were extracted for each speech frame, with a frame length of 10 ms and frame shift of 5 ms. These features are used to train an auto-associative neural network (AANN) model in the first method and as data points to compute the parametric models of two windows between whom a speaker change is hypothesized based on dissimilarity measure computed using BIC.

3.4.2 Anchor Speaker Tracking Using AANN Models

Artificial Neural Network (ANN) models consist of interconnected processing nodes, where each node represents the model of an artificial neuron, and the interconnection between two nodes has a weight associated with it. ANN models with different topologies perform different pattern recognition tasks [100]. For example, a feed-forward neural network can be designed to perform the task of pattern mapping, whereas a feedback network could be designed for the task of pattern association. A special case of feed-forward network is auto-associative neural network (AANN) models which perform identical mapping of input space. It has been shown such networks effectively captures speaker characteristics and could be used for speaker recognition and tracking [101].

The structure of AANN model is similar to the one followed in [101]. The network structure that was used in our experiments consists of 5 layers: $13\ L\ 39\ N\ 4\ N\ 39\ N\ 13\ L$, where the numbers indicate the number of nodes in the corresponding layer. L represents linear output function and N represents tangential output function. The AANN network layout is shown in Fig. 3.1.

The above structure was estimated over a few trials with different number of units in each layer. 13 MFCC features extracted for each frame are given as input to the network with the same feature vector as desired output. The weights in the network are modified by standard backpropagation learning law [100]. The weights of the network are adjusted for 200 cycles of presentation of data, where each cycle involves presentation of all training data once.

The proposed speaker tracking method follows an iterative technique to identify the segments of speech belonging to anchor speaker and the speaker model is refined in each iteration. An AANN model is trained with initial 30 s of speech of the show which contains anchor speaker's speech mostly. This is a reasonable assumption to make as in most cases,

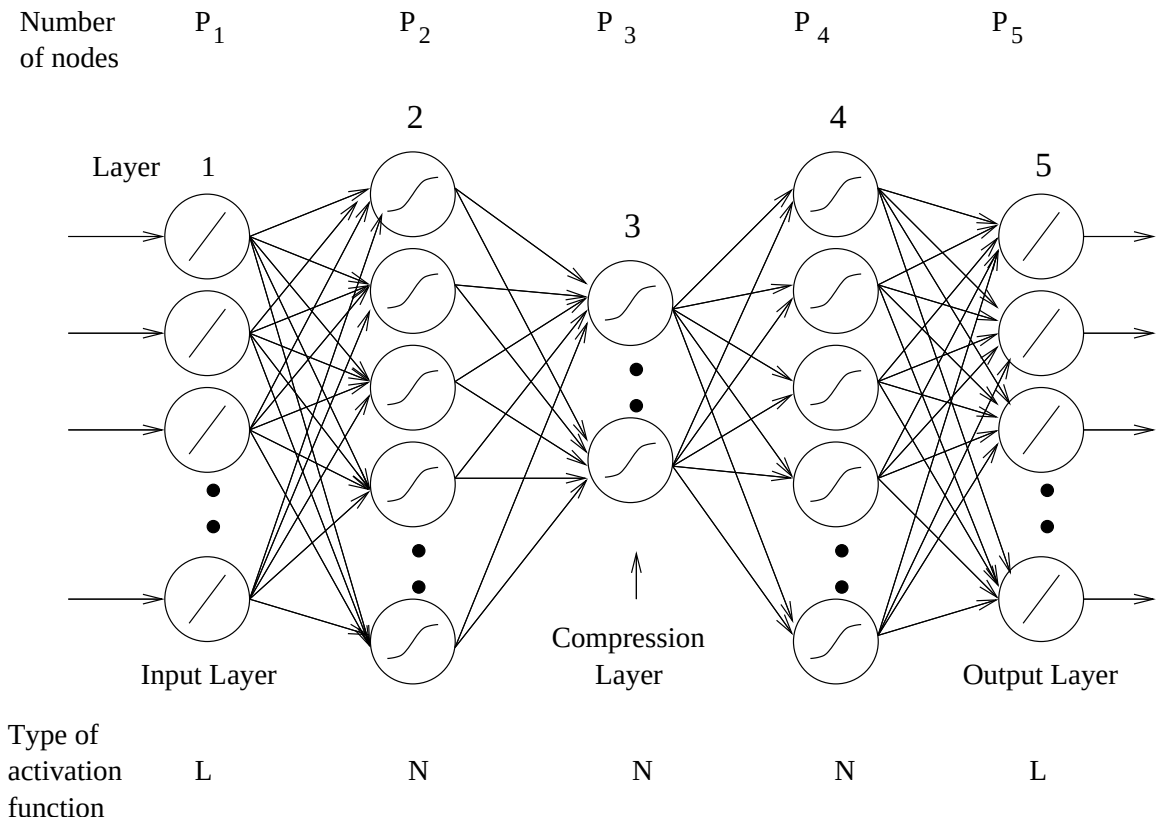

Figure 3.1 *Five layer Auto Associative Neural Network.*

anchor speaker starts the show by greeting the audience and reading the headlines. 13 MFCC features of each frame (generated by a frame length 10 ms and shift of 5 ms) of the show are given as input to the model. The mean squared error ($e[n]$) between the actual output and desired output is calculated. When MFCC features are given as input to AANN model, error as a function of time is not uniform in time. So, we used a confidence measure similar to the one proposed in [101] defined as,

$$c[n] = exp(-e[n]) \tag{3.1}$$

where, $e[n]$ is the mean squared error for the nth frame. $c[n]$ is the confidence score for the nth frame.

The confidence score will be high for the regions belonging to the speaker on whom the model is trained. These confidence scores are smoothed by a moving average window of

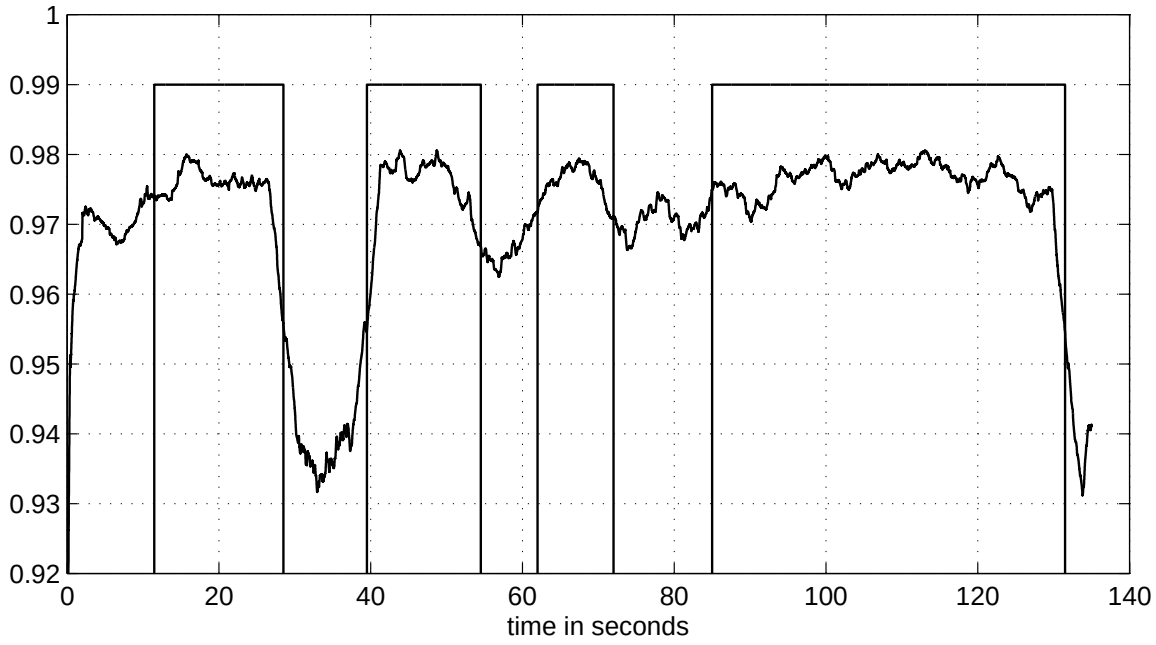

Figure 3.2 *Smoothed confidence contour with a moving average window of* 2 *s with anchor speaker regions marked.*

length 2 s. The valleys in the smoothed confidence contour belong to speech of speakers other than modelled speaker. The smoothed confidence contour is shown in Fig. 3.2.

This smoothed confidence contour is divided into non overlapping segments of 5 s each and mean confidence score is calculated for each segment. Length of the segment is chosen as 5 seconds as average length of speaker turn in a news show is around 5 seconds. Mean confidence score of a segment is compared against a threshold to classify it as belonging to anchor speaker or not. The threshold is calculated automatically as mean value of the smoothed confidence contour in the region belonging to initial 30 s (training) speech. All the segments that have mean confidence score greater than or equal to the threshold are identified as anchor speaker's speech. The MFCC features of these identified segments are used as training data for the next iteration. The above process is repeated until the model converges. The threshold ensures that only the segments that have a high likelihood of belonging to the modeled speaker are identified.

3.4.2.1 Evaluation

The speaker tracking efficiencies for each iteration are calculated in terms of precision and recall. A segment is considered as belonging to the anchor speaker if it contains more than half of his speech. The speaker tracking performance for each iteration is shown in Table 3.2.

Table 3.2 *Speaker tracking performance.*

iter-no	Recall	Precision
1	0.220	0.968
2	0.458	0.962
3	0.541	0.967
4	0.652	0.970
5	0.784	0.942

We can observe from Table 3.2 that the recall increases for each iteration while the precision values are fairly constant. The process is stopped when the model converges and no new anchor speaker segments are identified. The identified anchor speaker regions are used to construct summaries. One limitation of this method is, initial training data used for model adaption is not available always. The problem becomes more prominent when there are more than one speakers to be tracked; in the case of BN shows with multiple anchor speakers. In order to overcome this limitation, we used BIC based method to perform speaker tracking.

3.4.3 Anchor Speaker Tracking Using BIC

Speaker tracking using BIC method is performed in two stages. In the first stage, the BN show is divided into homogeneous regions containing speech from a single speaker, by detecting speaker change points. In the second stage, agglomerative clustering of these segments is performed using BIC as distance measure. As, anchor speaker has more speech

instances spread across the show, the cluster containing more speaker turns is hypothesized as the cluster belonging to anchor speaker.

3.4.3.1 Speaker Change Detection

The speaker change detection is performed by the dissimilarity measurement between two adjacent windows based on the comparison of their parametric models. The comparison is performed using Bayesian Information Criterion (BIC) [6]. Bayesian Information Criterion (BIC) is a maximum likelihood criterion penalized by the model complexity (number of model parameters). If X is a sequence of data and M is a parametric model with m parameters, and likelihood $L(X, M)$ is maximized, the BIC for model M is defined as

$$BIC(M) = logL(X, M) - \lambda \frac{m}{2} logN_x \tag{3.2}$$

where N_x is the number of points in the data sequence.

The first term represents the extent of match between model and the data. The second term denotes the model complexity. The value of λ is data dependent (theoretical value of λ is 1). The BIC allows us to select a model that best fits the data with less complexity. For speaker change detection, two hypothesis are tested. Consider two windows X and Y adjacent to each other. The first hypothesis (H_1) is that there is no speaker change between X and Y and the second hypothesis H_2 states that a speaker change occurs between the two windows. In H_1 a single multi-dimensional Gaussian distribution is assumed to model the data in the two windows better. In H_2 two multi-dimensional Gaussian distributions one for each window are assumed to model the data better. Let N_x, N_y be the number of data points in X and Y windows respectively and Z be the combined sequence of X and Y windows (N_z = N_x + N_y).

The ΔBIC value between the two hypothesis H_1 and H_2 is given by

$$\Delta BIC(H_1, H_2) = \frac{N_z}{2} log|\Sigma_z| - \frac{N_x}{2} log|\Sigma_x| - \frac{N_y}{2} log|\Sigma_y| + \frac{\lambda}{2}(p + \frac{p(p+1)}{2}) log N_z$$

where λ is a tuning factor which is data dependent and p denotes dimensionality of feature vector (in present case 13). A positive ΔBIC value indicates that a speaker change occurs between two windows. The windows are slid along time axis to detect speaker changes. A speaker change point is hypothesized at time instant i such that

$$max_i \Delta BIC(i) > 0. \tag{3.3}$$

The performance of the above technique on the current data set is reported in terms of false alarm rate (FAR) and missed detection rate (MDR) in Tab. 3.3.

Table 3.3 *Performance of ΔBIC on current data set*

error type	FAR	MDR
%	9.8%	11%

The BIC technique works better for long speaker turns as there is sufficient data to compute the dissimilarity measure reliably. The window size used in our experiments for computation of BIC was five seconds as speaker turns in news data are typically long. The graph of ΔBIC values with actual speaker change points marked is shown in Fig. 3.3.

It can be observed from Fig. 3.3 that the speaker change points coincide with the peaks in smoothed ΔBIC graph. These peaks are considered as speaker change points.

3.4.3.2 Clustering Anchor Speaker Segments

Homogeneous segments containing speech from a single speaker are obtained by taking segments between two speaker change points. To find the segments of anchor speaker, the segments are clustered by using the ΔBIC values as the distance measure. Initially each

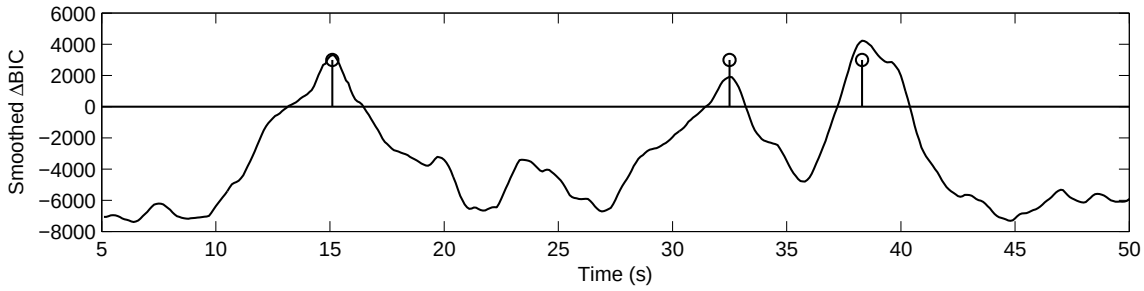

Figure 3.3 *ΔBIC based smoothed distance graph with actual speaker change points marked.*

individual segment is treated as a cluster and the ΔBIC is calculated for each segment with all other segments. The segments that have ΔBIC values less than or equal to zero are assigned to the cluster of corresponding segment. Ideally, the cluster containing highest number of segments will be anchor speaker's, as anchor speaker has more turns in a show. But it was observed that there are a few missed anchor speaker segments in this cluster. To reduce these, a global similarity matrix is constructed, by the intuition that segments of same speaker will have similar clusters. Similarity (s_{AB}) between two clusters A and B is given by

$$s_{AB} = i_{AB} - d_{AB}. \tag{3.4}$$

Here i_{AB} denotes the number of segments in the intersection of two clusters A and B.

$$i_{AB} = n(A \cap B). \tag{3.5}$$

And d_{AB} denotes symmetric difference between two clusters A and B. d_{AB} is given by

$$d_{AB} = n(A \Delta B), \tag{3.6}$$

All the clusters that have similarity score (s_{AB}) greater than a threshold (empirically decided as 1) with the cluster containing highest number of segments are treated as clusters of anchor speaker. All these clusters are merged into one cluster and this cluster represents the segments of anchor speaker. The technique can be easily extended to multiple anchor

speakers by taking top n clusters which are most dissimilar to each other according to the global similarity matrix. n is equal to the number of anchor speakers. The assumption here is that anchor speakers have more turns in the show than other speakers. The performance of anchor speaker tracking is reported in Tab. 3.4.

Table 3.4 *Performance of anchor speaker tracking*

error type	FAR	MDR
%	14%	3%

3.5 Summary Construction

Each anchor speaker segment can be treated as start of a news story in the show. But there are also instances where anchor speaker interacts with the other speakers within a story. Such segments are typically small and filtered out by removing anchor speaker segments less than 5 seconds in duration.

3.5.0.3 Concatenation with Compression

After removing short segments, we obtain final anchor speaker regions that need to be concatenated to form a summary. The compression ratio (cr) is defined as the ratio of desired summary length to the total length of a document. The required summary length (Sl) is obtained from the given compression ratio (cr) as

$$Sl = cr \times Tl, \tag{3.7}$$

where Tl is the total length of the show in seconds. The number of stories is approximately equal to the number of final anchor speaker regions (N). Duration (D) of each news story in a summary is obtained as

$$D = Sl/N. \tag{3.8}$$

Initial D seconds of speech from each anchor speaker region are taken as candidates for concatenation. This type of selection makes sure that all news stories are covered in the summary. If anchor speaker's speech in a particular news story is less than D seconds then the boundary is adjusted accordingly to the end point of his speech. The boundaries of these candidate regions are not meaningful, either acoustically or linguistically, and they may be abrupt. To make them smooth the boundaries of these regions are extended to the nearest 250 ms pause in the signal. The final candidates are concatenated to form a meaningful audio summary.

3.6 Evaluation

The evaluation is done on 20 news shows of globalnews podcast of BBC news, details of which are presented in Sec. 3.2.1. Two types of evaluations are carried out, one using traditional text summary evaluation system ROUGE and the other using human evaluation for audio summaries. ROUGE based evaluation provides an objective measure of quality of the summaries where as human evaluation was done to evaluate the usefulness of the audio summaries for humans. The summaries generated by proposed techniques are compared with summaries generated by a text summarization system , and a supervised state of the art speech summarization system similar to the systems proposed in the literature.

3.6.1 Text Summarization System

The manual transcripts of speech files corresponding to each BN show are given as input to the text summarizer to generate a summary. The text summarizer is built using MEAD [75] which uses positional features and tf.idf scores for ranking sentences in a document. The top ranking sentences are picked into the summary until desired summary length is reached. The summaries are generated for a compression ratio of 30 %.

3.6.2 Supervised Speech Summarization System

An artificial neural network classifier is trained on gold standard human labelled summaries which contains segments from all four human summaries. The classifier is 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 energy (I), ΔI, F_0, ΔF_0 over each segment and duration of the segment. The F_0 and I contours are normalized using z-score normalization. The corpus is 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 a compression ration of 30 %.

3.6.3 ROUGE based Evaluation

Recall oriented understudy for gisting evaluation (ROUGE) [48] which is commonly used for evaluating text summaries, measures overlap units between automatic and manual summaries. ROUGE-N computes the n-gram overlap between the summaries where N indicates the size of n-grams. We report ROUGE-1, ROUGE-2 and ROUGE-SU4 scores. ROUGE-SU4 indicates the skip bi-gram score within a window length of four. The ROUGE scores of the current system are compared against a baseline text summarization system built using MEAD and supervised speech summarization system trained on gold standard human reference summaries. Audio summaries generated by the system are transcribed manually into text for evaluation purpose. In order to evaluate the summarization capability of the proposed techniques for different summary lengths, summaries are generated for different compression ratios (5, 10, 15, 20, 25 and 30). The size of human reference summaries was not altered for evaluating automatic summaries of different compression

ratios. The ROUGE scores of audio summaries for different compression ratios (5, 10, 15, 20, 25 and 30) are presented in Fig. 3.4 and Fig. 3.5.

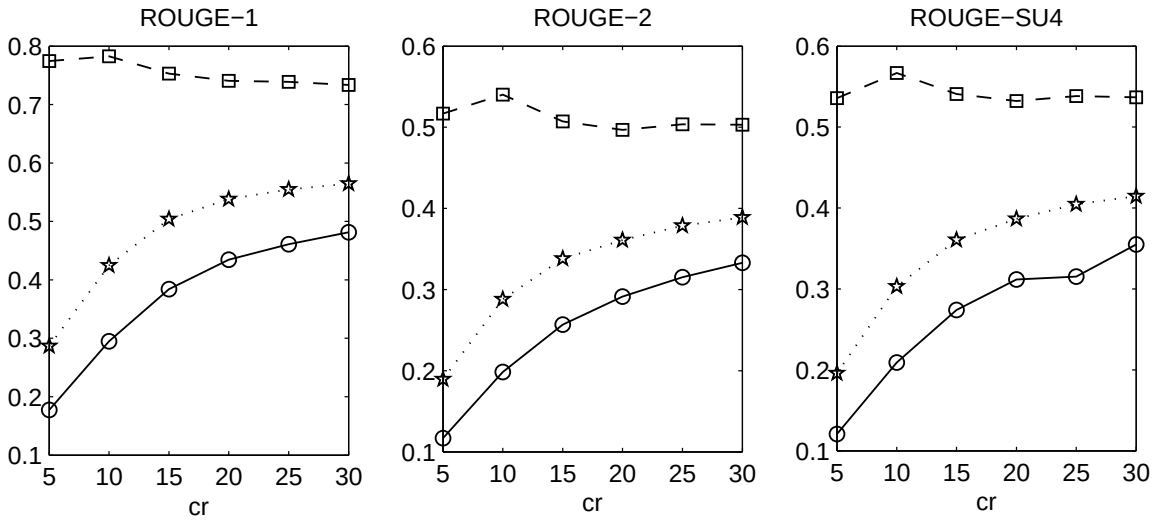

Figure 3.4 *Plots showing recall (solid line), precision (dashed line) and F-measure (dotted line) values for various compression ratios (cr) of audio summary generated using BIC based speaker tracking.*

It can be observed from Fig. 3.4 and Fig. 3.5 that recall values of the summaries increase with increase in compression ratio as expected. The precision values are fairly constant for all compression ratios which shows that the new segments that are being added to the summary due to increase in desired summary length are relevant to summary. Precision values are important for an extractive summary, because if the number of extracts is increased, the recall values might increase but the percentage of segments relevant to summary might drop.

The ROUGE scores for summaries generated using proposed speaker tracking techniques, text summarizer built using MEAD and supervised speech summarizer trained on gold standard human summaries for 30 % compression ratio are presented in Tab. 3.5

It can be observed from Tab. 3.5 that the proposed speaker tracking techniques produce summaries as good as MEAD based text summarizer and supervised system.

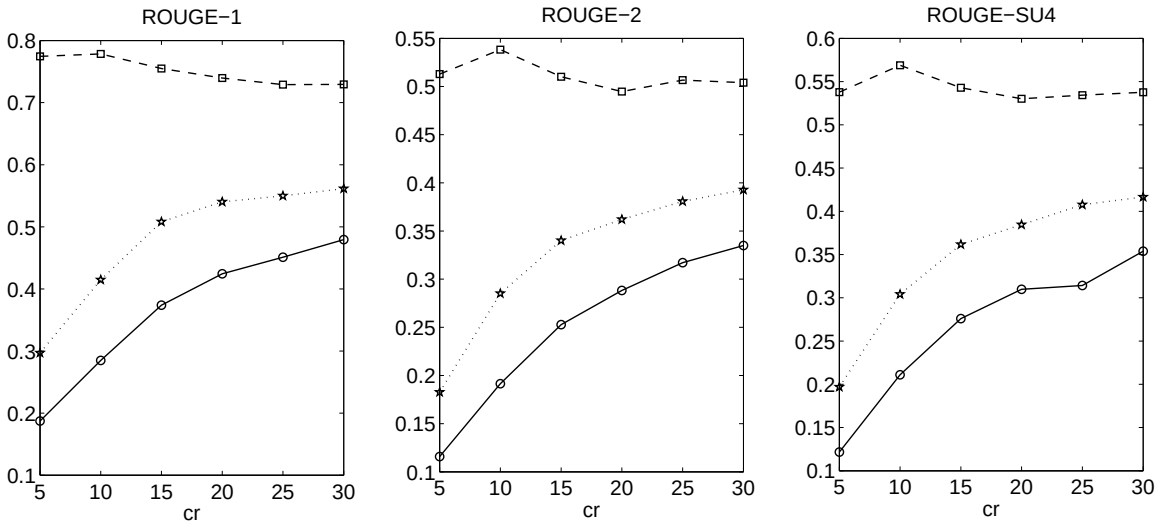

Figure 3.5 *Plots showing recall (solid line), precision (dashed line) and F-measure (dotted line) values for various compression ratios (cr) of audio summary generated using AANN based speaker tracking.*

Table 3.5 *F-measure values and 95% confidence intervals for ROUGE-1 (R-1), ROUGE-2 (R-2), ROUGE-SU4 (R-SU4) metrics for speaker tracking based summaries, summaries generated by supervised system and MEAD summarizer.*

system	R-1	R-2	R-SU4
AANN	0.561[0.54 0.58]	0.392[0.36 0.40]	0.416[0.39 0.43]
BIC	0.564[0.54 0.58]	0.388[0.36 0.40]	0.414[0.39 0.43]
supervised	0.553[0.52 0.57]	0.382[0.36 0.40]	0.402[0.38 0.42]
MEAD	0.572[0.55 0.59]	0.394[0.37 0.41]	0.421[0.40 0.44]

3.6.4 Human Evaluation

3.6.4.1 Question & Answer based Evaluation

In human evaluation, 5 human subjects were 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. As the aim of our summarizer is to generate indicative summaries, which announce the contents of a document, the questionnaire consisted of simple questions based on facts of a news story. The questions are of type what, when, who, where etc. All the subjects were asked

to answer the questionare before listening to summaries to factor out their prior knowledge on the news stories. The subjects were not restricted from listening to a summary multiple times. The percentage of the questions answered correctly after factoring out their prior knowledge for each compression ratio is presented in Tab. 3.6.

Table 3.6 *Percentage of questions answered correctly for different compression ratios (cr)*

cr	5	10	15	20	25
AANN	43.6 %	54.3 %	61.1 %	66.0 %	70.8 %
BIC	42.4 %	55.6 %	62.0 %	65.5 %	71.0 %
Supervised	36.2 %	41.6 %	47.3 %	53.4 %	60.2 %

The results of Q&A based evaluation in Tab. 3.6 show that humans are able to understand the audio summaries produced by anchor speaker tracking easily and were able to get more information from them than summaries generated by a supervised system.

3.6.4.2 Coherence Evaluation

In order to evaluate coherence of the audio summaries, subjective evaluation by is performed by 10 subjects. The subjects are asked to evaluate the summaries based on coherence, ease of understanding and appropriateness as a summary. They are provided with text transcript of the news show before they listen to the summaries, so that they get an idea of the contents of the show. They are asked to rate the summaries at five levels: 1-very bad, 2-bad, 3-normal, 4-good, 5-very good. The mean opinion scores (MOS) of these ratings for summaries of 20 news shows are presented in Tab. 3.7

Table 3.7 *MOS of summaries generated by various methods.*

method	AANN	BIC	Supervised
MOS	4.0	4.05	3.2

From Tab. 3.7 it can be inferred that human beings prefer summaries generated by the proposed techniques than summaries generated by standard speech summarization systems based on a supervised classifier.

3.7 Summary

In this chapter, we have demonstrated an automatic speech-to-speech summarization system for BN shows. The proposed approach does not require any transcripts or reference summaries, and summaries are generated in speech such that the naturalness in the original signal is preserved. We have demonstrated the importance of anchor speaker segments for constructing an extractive audio summary of a news show by analyzing human summaries of broadcast news. This property of human summaries was incorporated into an automatic summarization system by performing anchor speaker tracking and constructing audio summaries based on the positional features of the identified anchor speaker segments. The proposed system generates summaries for different compression ratios without degrading the quality of the summaries. Good recall and precision scores indicate that it is possible to build extractive speech summarization systems with performance comparable to text summarization systems provided they have some inherent structure that can be identified.

Chapter 4

Prominence Based Ranking of Speech Segments

4.1 Introduction

Speech summarization systems produce extractive summaries where important segments from the 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 important segments and ranking the segments for inclusion in a summary. Initial approaches to speech summarization used automatic speech recognition (ASR) output of speech files and applied methods based on tf*idf (term frequency, inverse document frequency), maximum marginal relevance (MMR), and 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 [102, 63, 19, 42]. Recently acoustic features were used 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. [56] combines lexical and acoustic features to train a supervised system to classify a segment as belonging to summary or not. [57] 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. However, ASR systems may not be available for all languages, and it involves 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 this work, we propose a method to rank speech segments based on prominence features. The proposed method does not require an ASR system or a gold standard human summary as it uses prominence values of syllables in a speech segment to rank the speech segment for summarization. When humans convey message through speech, they attract listeners' attention to information bearing parts of speech through variations in pitch, amplitude, duration and stress [10]. Speakers make some words prominent and reduce other words. It is widely accepted that in English, content words (nouns, verbs) are stressed or made prominent than functional words (articles, conjunctions, inter-junctions) [70]. It was also shown that prominent words occur to introduce new concepts [27]. A study on prominent words and their importance showed that words that are not prominent had low value of information retrieval (IR) index, while words that are mostly prominent had higher IR index [84]. Traditional text summarization systems rank the sentences based on tf*idf scores of their constituent words. This type of ranking gives high scores to sentences with more content words. As content words are shown to be stressed or prominent in speech, we investigate whether prominence based ranking of speech segments could help in automatic summarization.

To perform this investigation, a method for scoring syllables based on prominence is required. In the scope of this work, we wish to use existing methods for estimating prominence value of a syllable, and focus on using prominence for speech summarization. The current work differs from previous works on speech summarization in the following ways.

- Main distinction is, the features used in current study are computed with respect to a syllable rather than at a segment level.

- Though previous works on speech summarization have used basic prosodic features such as F_0, duration and intensity, there is a difference in the way these features are used to model prominence. For example, instead of computing mean of raw F_0 values, the shape of F_0 contour is modelled to detect intonatinal events which indicate prominence and only these events are considered in further computations.

- Also, the intensity values are not just root mean square values of signal amplitudes but the intensity values of the signal that is band-pass filtered between 300-2200 Hz. The intensity values in this band show a greater discrimination between prominent and non prominent syllables.

This way of modelling prominence provides a way of scoring speech segments based on prominence values of syllables which in turn provides a way of performing speech summarization in an unsupervised frame work. The main aim of the current study is to evaluate the usefulness of this scoring for speech summarization. Also, when an ASR system is available, we propose a method to combine lexical features derived from ASR transcripts with prominence based scoring.

4.2 Prominence

The definition of prominence in literature is perceptually motivated. Prominence is defined as perceptual salience of a language unit [88]. It is the property by which linguistic units are perceived as standing out from their environment [92]. Prominence is associated with suprasegmental characteristics of speech primarily duration, frequency and amplitude. In order to objectively study prominence, it needs to be quantified. Several approaches have

been made in literature to quantify prominence at different levels. Portele and Heuft [73] defined prominence on a scale from 0 to 30 at word level. Terken [93] defines prominence on a scale of 0 to 10. It should be noted that it is challenging for human annotators to label prominence at these levels. Streefkerk [88] has used binary markings of prominence 0 or 1 at word level. Prominence is also described in terms of distribution of accents. The tones and break indices (ToBI) [3] is a standard followed widely for annotating accents and prosodic phrase boundaries in continuous speech. The ToBI annotation standard was developed to address the issue of representing prosodic events in spoken language in an unambiguous fashion. It uses four interrelated tiers of annotation to represent prosodic events in spoken utterances. The tone tier marks the presence of pitch accents and prosodic phrase boundaries. A pitch accent can be broadly thought of as a prominence mark. Two basic types of accents high (H) and low (L) are defined, depending on the value of F_0 with respect to its vicinity. Other complex accents such as low-high (L+H*) and high-low (H+L*) are also marked based on shape of F_0 contour in the immediate vicinity of the accent. An example of these annotations are shown in the Fig. 4.1.

4.2.1 Acoustic Correlates of Prominence

Prominence cannot be attributed to a single production mechanism such as, vibration of vocal folds to fundamental frequency (F_0). Prominence can be achieved by varying any of the acoustic properties such as intensity, duration, pitch or by a combined effect of any of them [47]. Numerous studies have been conducted in literature to study acoustic correlates of prominence. There is a rough agreement in the literature that syllable duration, pitch pattern and intensity(sub-band energy) have close correlation with prominence.

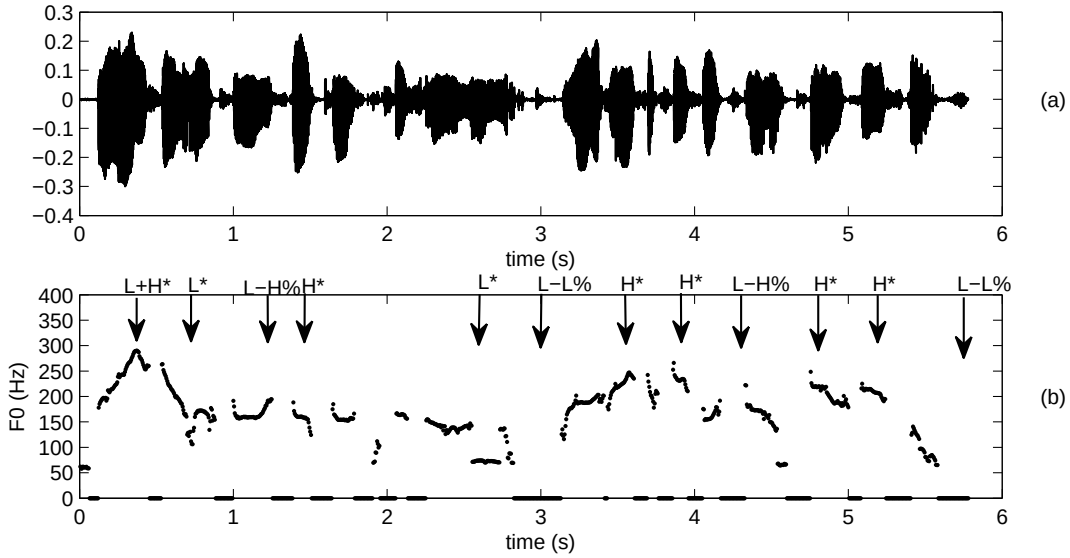

Figure 4.1 *An example of ToBI based prosodic annotations for a given speech signal: (a) Speech signal (b) F_0 contour of speech signal shown in (a) with manual ToBI based prosodic event markings.*

4.2.1.1 Syllable Duration

It was shown by Sluijter and Van Heuven [87] that speakers tend to stretch the constituent syllable durations when they try to emphasize a specific word. Usually vowel or semi vowel which constitutes a syllable nucleus, is stretched more than consonant parts of the syllable. Tamburini and Caini [89] have shown that syllable nucleus duration is as effective as syllable duration in discriminating prominent and non prominent syllables. This is a helpful observation as syllable nucleus boundaries can be identified with greater accuracy than syllable boundaries.

4.2.1.2 Pitch Pattern

Pitch patterns have been shown to correlate strongly with human prominence judgements by Streefkerk [88]. Many studies have tried to explore valid pitch patterns that indicate prominence. In [93] the distance between F_0 maxima and the corresponding virtual

baseline at any instant has been proposed as a valid indication of pitch accent. Streefkerk [88] used pitch median and pitch range as a measure of accent. Sluijter and Van Heuven [87] used pitch variation. Tamburini [89] applied the sum of rise and fall amplitudes measured from Taylor's [91] tilt parameters as a measure of pitch accent. Knight [40] showed that pitch plateau is related to prominence perception.

4.2.1.3 Spectral Intensity

Spectral intensity was also used widely as a feature to indicate prominence. Sluijter and Van Heuven [87] showed that energy in the 300-2200 Hz band has maximum correlation with prominence. Beyond the straightforward measure of such sub-band energy, there has been research in measuring various transforms of spectral intensity. There has been a notion of loudness [16], an approximation to steady state perceptual loudness, with various measures for it such as through power spectral density [41].

4.2.2 Acoustic Measure of Prominence

In order to obtain prominence values of syllables in a speech segment, we followed the method described in [89]. This method computes prominence value of a syllable based on acoustic features like syllable nucleus duration, sub-band energy(300-2200 Hz) and pitch variation. A brief description of this method is presented below.

4.2.2.1 Estimation of Syllable Nucleus Duration

To reliably identify the syllable nuclei in a segment and measure their duration to obtain the acoustic parameter needed for subsequent computations, we applied a modified version of the convex-hull algorithm [61] to the segment energy profile. The energy profile was computed after band-pass filtering (300-900 Hz) the speech samples, as suggested in [30], to filter out energy information not belonging to vowel phones, which form the syllable

nucleus. The duration parameter is then normalized by dividing with the maximum of durations of the syllable nuclei in the segment. This is a standard technique for rate-of-speech normalization, described, in [66].

4.2.2.2 Estimation of Sub-band Energy

In order to compute sub-band energy the speech signal was passed through a band pass (300-2200 Hz) FIR filter. The sub-band energy of each syllable nucleus is computed as RMS energy of the filtered speech signal within the syllable nucleus. The RMS energy is computed as

$$E_j^{RMS} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} a_{ji}^2} \tag{4.1}$$

where, N is the number of samples per frame and j is the frame index. The frame width used to compute the RMS energy is 10 ms with a frame shift of 5 ms.

4.2.2.3 Modelling of Pitch Patterns

Taylor proposed a model to capture intonation events in continuous speech by representing pitch contour in form of rise/fall/connection (RFC) segments. He defined a set of parameters capable of uniquely describing events in pitch contour (pitch accent shapes and boundary tones). This set, called TILT, consists of five parameters defined as:

$$tilt_{amp} = \frac{|A_{rise}| - |A_{fall}|}{|A_{rise} + A_{fall}|} \tag{4.2}$$

$$tilt_{dur} = \frac{|D_{rise}| - |D_{fall}|}{|D_{rise}| + |D_{fall}|} \tag{4.3}$$

$$tilt = \frac{|A_{rise}| - |A_{fall}|}{2 \times (|A_{rise}| + |A_{fall}|)} + \frac{|D_{rise}| - |D_{fall}|}{2 \times (|D_{rise} + |D_{fall}|)} \tag{4.4}$$

$$A_{event} = |A_{rise}| + |A_{fall}| \tag{4.5}$$

$$D_{event} = |D_{rise}| + |D_{fall}| \tag{4.6}$$

where $A_{rise}, A_{fall}, D_{rise}, D_{fall}$ are respectively the amplitude and duration of the rise and fall segments of the intonation event.

In order to extract these parameters the F_0 contour is first converted into an intermediate RFC model. To do that the contour is segmented into frames 25 ms long; next, the data in each frame is linearly interpolated using a least median squares method to obtain robust regression and deletion of outliers [81]; then every frame interpolating line is classified as rise, fall or connection, depending on its gradient, as suggested in [22] and [90]. After that, subsequent frames with the same classification are successively merged into one interval and the duration and amplitudes of the rise and fall sections are measured to finally derive the TILT parameter set. An example RFC representation of pitch contour is shown in Fig. 4.2.

As described by Taylor [91], an intonational event that can be considered as a good candidate for pitch accent exhibits a rise followed by a fall in the pitch profile. The pitch variation inside each syllable nucleus is measured from the amplitudes and durations of such intonational event within the syllable nucleus. To measure pitch variation, the event amplitude, which is one of the TILT parameters, can be considered as a proper measure, being the sum of the absolute amplitude of the rise and fall sections of an intonational event. A further refinement can be obtained by multiplying the event amplitude (A_{event}) by its duration (D_{event}). This measure of pitch variation of each syllable nucleus in a segment is multiplied by a normalizing factor (R_{event}) which is computed as event amplitude divided by maximum pitch value in the segment.

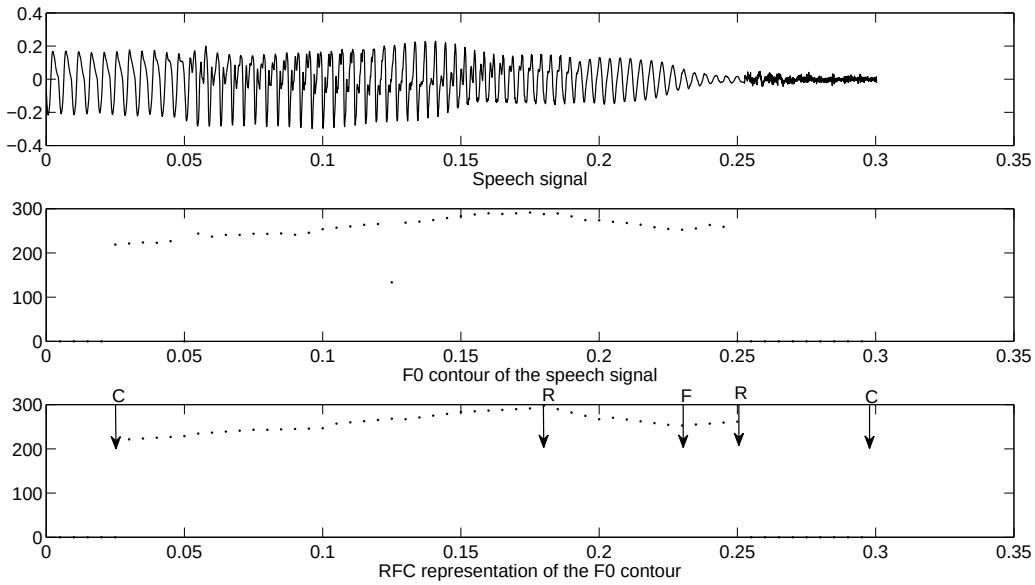

Figure 4.2 *RFC representation of F_0 contour.*

4.2.2.4 Prominence Value of a Syllable

Prominence value (p_i) of a syllable (i) in a speech segment is given by

$$p_i = max(\gamma_i, \eta_i), \tag{4.7}$$

where γ_i is computed as

$$\gamma_i = dur^i \times en^i_{300-2200} \tag{4.8}$$

here, dur^i is the syllable nucleus duration and $en^i_{300-2200}$ is the energy in frequency band $300 - 2200Hz$.

η_i is computed as

$$\eta_i = en^i_{ov} \times (A^i_{event} \times D^i_{event} \times R^i_{event}), \tag{4.9}$$

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.

4.3 Data-Set

The studies described in the current work are carried out on two different speech corpora. One corpus is a subset of Boston university radio news corpus (BU-RNC) which contains read style speech. This corpus was used for analysis and testing of the proposed prominence based speech summarization as it contains human prominence markings. The second corpus used is a subset of switchboard data corpus released by ICSI which contains spontaneous telephone conversations. This corpus was used to verify the performance of the proposed technique on spontaneous speech.

4.3.1 Boston University Radio News Corpus

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. It contains ToBI-style [3] prosodic annotations for part of data which include hand labelled prominence markings by experienced labelers. It also contains word and phone level time alignments and POS tags corresponding to each token in the orthographic transcription. The prosodic annotations, phone level alignments and POS tags are not used in current experiments. The orthographic text transcripts for segments are used in creating human reference summaries for evaluation purpose.

4.3.2 ICSI Switchboard Corpus

Switchboard audio corpus contains spontaneous discussions between two individuals over a telephone on a specific topic such as automobiles, sports, politics, credit cards for

several minutes. 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. Manual prominence markings are not available.

4.3.3 Human Reference Summaries

The text transcripts of the speech files are presented to 4 human annotators along with corresponding audio for constructing a summary. The annotators were 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 standard evaluation setup for text summarization at document understanding conference (DUC) [1] uses 40 topics and 4 human reference summaries. The number of speech files from each corpus and human reference summaries used in this work was fixed following DUC framework.

4.4 Significance of Prominence for Summarization

4.4.1 Experiments using Hand-labelled Prominence Markings

In this section experiments carried out using hand labelled prominence markings present in f2b corpus are reported which give motivation for exploring automatic prominence based scoring of speech segments for summarization.

4.4.1.1 Content and Function Words

Previous studies [99, 84] have shown that content words are made prominent than function words in continuous speech. In order to validate these observations, we analyzed the

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

nature of words that are marked as prominent by human labelers in f2b corpus. Out of 9090 words in the corpus 2852 words are marked as prominent, of which 2614 (91.6%) are content words and 238 (8.3%) are function words. This observation shows that prominence can be used to distinguish content and function words. The content and function words classification was based on POS tags given in the corpus. The words carrying NNP (noun), VBN (verb), JJR (adjective), RBS (adverb) tags are treated as content words while the rest are treated as function words.

4.4.1.2 Correlation between Prominence Values (p_i) and Prominence Markings

In order to verify the relevance of the features such as syllable nucleus duration, sub-band energy and pitch variation to prominence, we have computed prominence values of the syllables using the above features as explained in Sec. 4.2.2.4 . As the f2b corpus contains hand labelled prominence markings, we have plotted the distributions of each acoustic feature and the computed prominence values for prominent and non prominent syllables. For computation of these features and prominence values, syllable nucleus boundaries which are given in the corpus are used. The Fig. 4.3 shows distributions of various acoustic features such as sub-band energy, syllable nucleus duration and pitch variation for prominent and non prominent syllables.

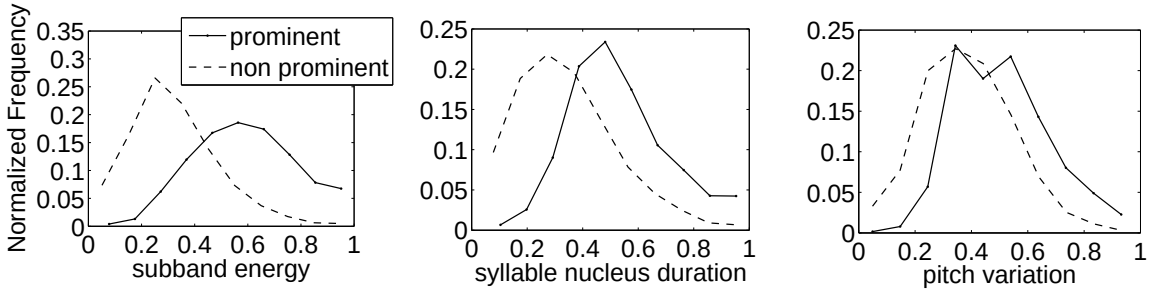

Figure 4.3 *Distributions of various acoustic features for prominent and non prominent syllables.*

The Fig. 4.4 shows the distribution of prominence values (p_i) computed as explained in Sec. 4.2.2.4 for prominent and not prominent syllables in the corpus.

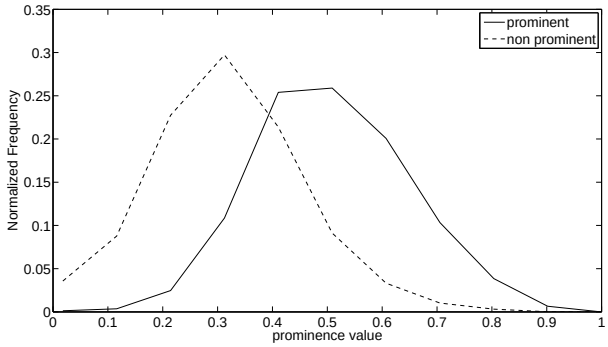

Figure 4.4 *Distributions of prominence values (p_i) for prominent and non prominent syllables.*

It can be observed from Fig.4.4 that prominence values for syllables marked as prominent are higher than values of non prominent syllables. Therefore, the computed prominence value for a syllable (p_i) can be treated as a measure of its prominence.

4.4.1.3 Computation of Segment Level Acoustic Score (α) based on Prominence Values of Syllables

To rank speech segments for automatic summarization, a segment level score is required. The acoustic score of a segment (α) is computed from prominence values of syllables marked as prominent in the segment. The prominence values (p_i) of syllables that are hand-labelled as prominent are obtained by the method described in Sec. 4.2.2.4. Acoustic score of a speech segment (α) based on prominence is computed as the mean of prominence values of syllables that are hand marked as prominent. The acoustic scores of segments (α) in a speech document are normalized by dividing them with the maximum value of α in the speech document. The distribution of acoustic scores (α) for speech segments belonging to summary class and non summary class is shown in Fig. 4.5.

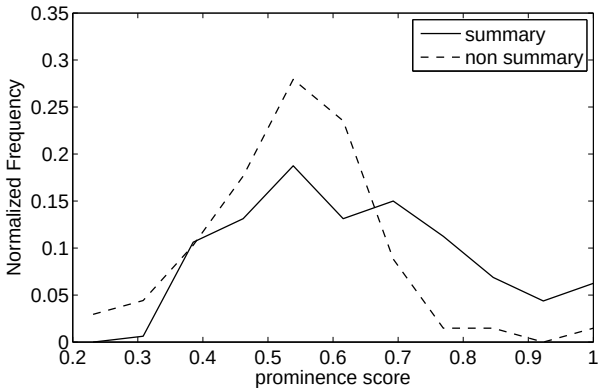

Figure 4.5 *Distributions of α scores for segments belonging to summary and non summary classes.*

It can be observed from the Fig. 4.5 that segments belonging to summary class tend to have high α score than segments not in summary. This shows that prominence based scoring of speech segments helps in automatic summarization.

4.4.1.4 Speech Summarization using Segment Level Acoustic Scores (α)

Speech segments are ranked in decreasing order of α and top ranking segments are concatenated in chronological order of their occurrence in the news show until desired summary length is achieved. In order to formally evaluate the usefulness of prominence for summarization, we compare the summaries generated by prominence based acoustic scores with summaries generated by tf*idf based scoring of manual transcripts (Sec. 4.4.2.1) and summaries generated by a supervised system trained on gold standard human reference summaries (Sec. 4.4.2.2).

4.4.2 Evaluation

The evaluation of summaries was done by estimating how close they are with human reference summaries. Audio summaries are transcribed into text by picking corresponding text segments from the manual transcripts provided with the corpus. The summaries

are evaluated using standard text summarization evaluation system ROUGE[48]. ROUGE measures n-gram overlap between human reference summaries and automatic summaries. Four human reference summaries are provided as model reference summaries for each news story. ROUGE-1, ROUGE-2 and ROUGE-SU4 scores for these summaries are reported in Tab. 4.1. 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.

4.4.2.1 Comparison with 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. Segments are ranked in decreasing order of their similarity scores. The similarity between a segment and the document is computed by the dot product between corresponding vectors with terms as dimensions and tf*idf scores of the terms as magnitudes of corresponding dimensions.

4.4.2.2 Comparison with Supervised System Trained Using Gold Standard Human Summaries

An artificial neural network classifier is trained on gold standard human labelled summaries which contains segments from all four human summaries. The classifier is 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 energy (I), ΔI, F_0, ΔF_0 over each segment and duration of the segment. The F_0 and I contours are normalized using z-score normalization. The corpus is 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.

Table 4.1 *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 generated by supervised system and tf*idf scores.*

system	R-1	R-2	R-SU4
prominence	0.515[0.49 0.53]	0.351[0.33 0.37]	0.345[0.32 0.36]
supervised	0.478[0.45 0.49]	0.340[0.32 0.36]	0.337[0.31 0.35]
tf*idf	0.514[0.49 0.53]	0.337[0.31 0.35]	0.344[0.32 0.36]

From Tab. 4.1 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 values (p_i) of syllables in the segment.

4.5 Speech Summarization Using Automatic Prominence Scoring

4.5.1 Proposed Approach

The block diagram of proposed summarization system is shown in Fig 4.6. The speech file is first segmented by extracting speech segments based on pause duration. A segment boundary is assumed whenever a pause greater than 250 ms is encountered. Syllable nu-

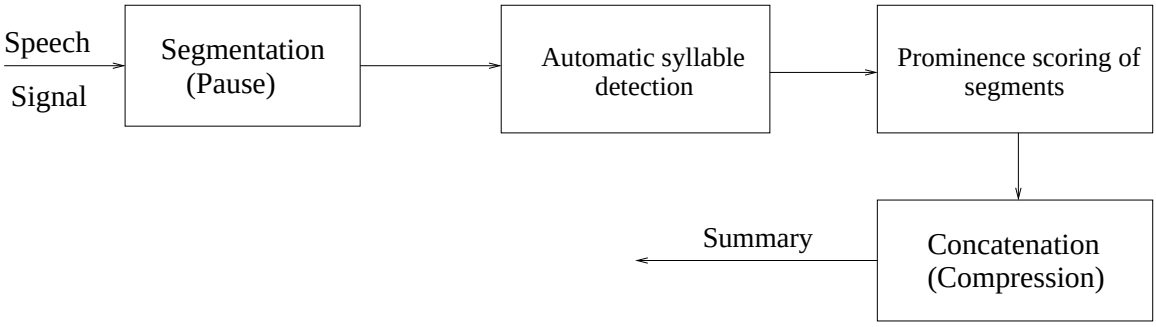

Figure 4.6 *Block diagram of the summarization system.*

cleus boundaries are identified using the method explained in Sec. 4.2.2.1. 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 Sec.4.2.2.4. 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 present in it.

$$mp = \frac{\sum_{i=1}^{N} p_i}{N}, \tag{4.10}$$

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

$$mdp = \frac{\sum_{i=1}^{N-1} |p_{i+1} - p_i|}{N-1}, \tag{4.11}$$

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 achieved.

The distributions of the four scoring functions for summary and non summary class phrases are shown in Fig. 4.7.

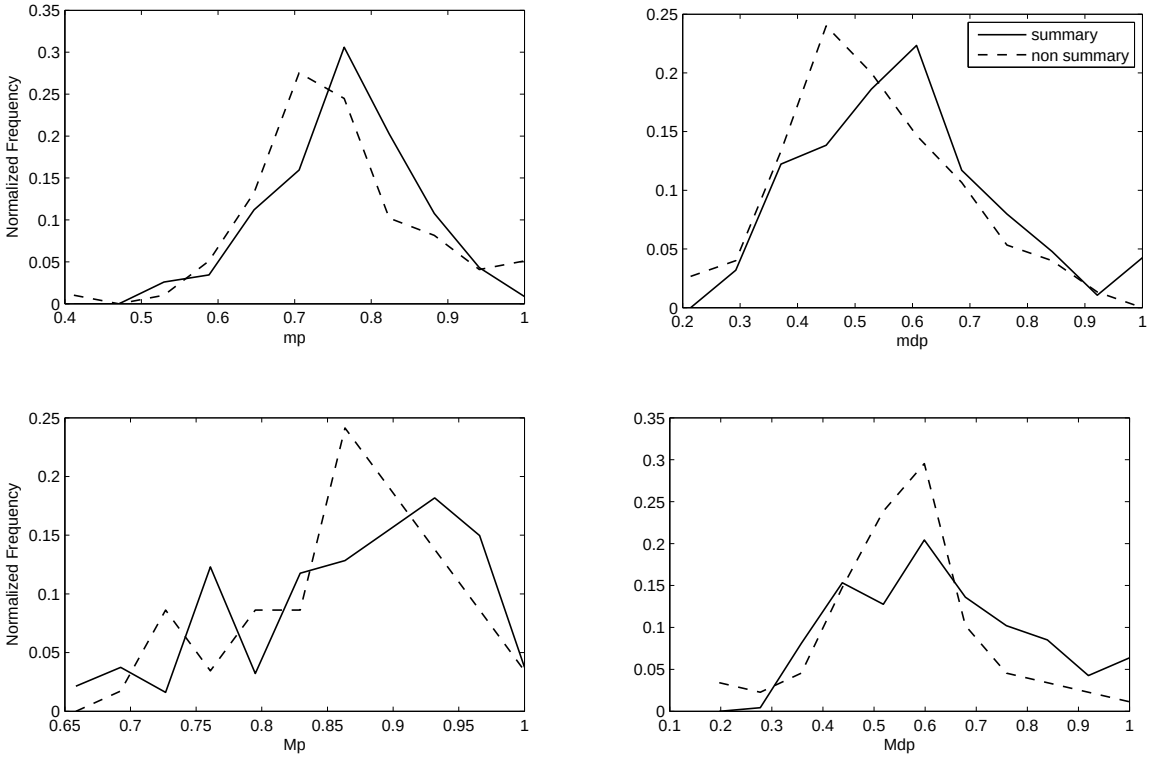

Figure 4.7 *Distributions of segment level acoustic scores obtained from four different scoring function (mp,mdp,Mp,Mdp) for segments belonging to summary and non summary classes.*

4.5.2 Evaluation

4.5.2.1 Results on f2b Corpus

The evaluation of summaries generated by automatic prominence detection was done in two ways, one based on text summarization evaluation package ROUGE [48] and the other

based on task based evaluation by humans. ROUGE based evaluation gives an objective measure for the quality of the summary, while task based evaluation was done to evaluate the quality of the audio summaries.

All the summaries are generated for 30% cr (same as model summaries). 4 human summaries are provided as model reference summaries for each story. ROUGE scores for different prominence scoring functions are reported in Tab. 4.2. It can be observed that mdp performs better than other scoring functions. The summaries generated by automatic prominence scoring using mdp (Tab. 4.2) have less ROUGE scores than summaries generated by manual prominence markings and tf*idf based scores (Tab. 4.1). But the difference is not statistically significant as the 95% confidence intervals of these systems overlap significantly. 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 Fig. 4.8. It can be observed from Fig. 4.8 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 4.2 *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	0.496[0.47 0.51]	0.316[0.29 0.33]	0.322[0.29 0.34]
Mp	0.474[0.45 0.49]	0.297[0.27 0.31]	0.305[0.28 0.32]
mdp	0.508[0.48 0.52]	0.341[0.32 0.36]	0.343[0.32 0.36]
Mdp	0.489[0.46 0.50]	0.323[0.30 0.34]	0.328[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

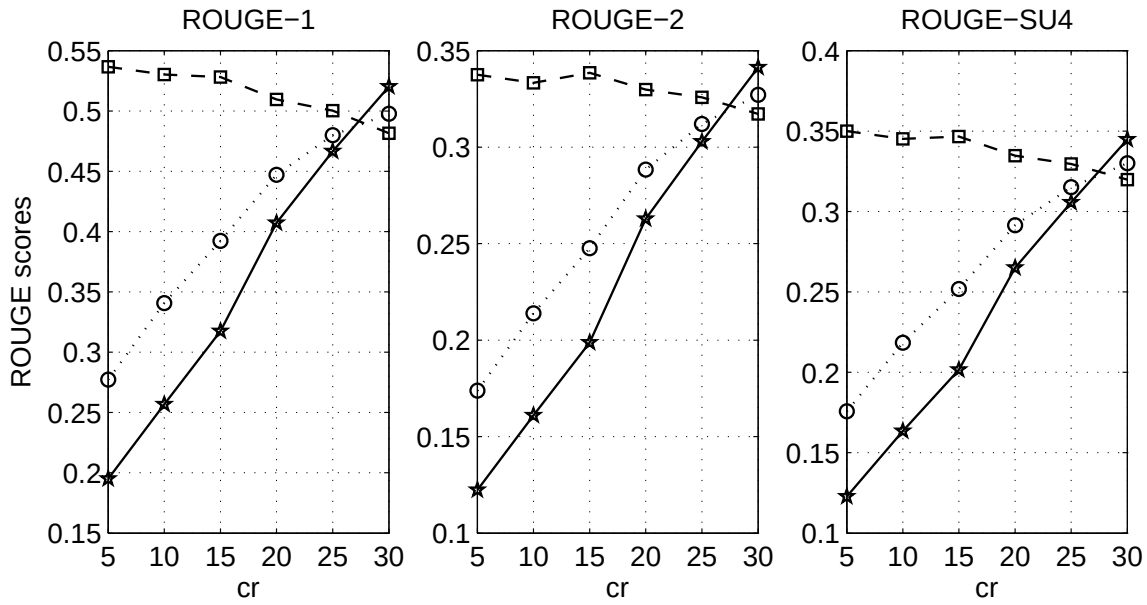

Figure 4.8 *Figure showing recall (solid line), precision (dashed line) and f-measure (dotted line) values of different ROUGE metrics for different compression ratios (5, 10, 15, 20, 25, 30) of audio summaries generated by mdp scoring function.*

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 Tab. 4.3.

Table 4.3 *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 (Tab. 4.3) show that humans 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 (Fig. 4.8).

4.5.2.2 Results on Switchboard Corpus

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 Sec. 4.5.1. Top scoring speaker turns are concatenated until 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 performance of the proposed method along with tf*idf based scores and supervised system trained on switchboard data is reported in terms of ROUGE scores in Tab. 4.4.

Table 4.4 *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	0.666[0.64 0.68]	0.464[0.41 0.49]	0.491[0.45 0.52]
tf*idf	0.653[0.62 0.68]	0.461[0.40 0.49]	0.490[0.42 0.53]
supervised	0.628[0.59 0.65]	0.456[0.40 0.48]	0.474[0.40 0.52]

It can be observed from Tab. 4.4 that the prominence based ranking (mdp) performs as good as tf*idf scores and supervised system even on spontaneous speech data. The ROUGE scores for switchboard corpus are higher than ROUGE scores for f2b corpus. This might be because of the choice of the extraction unit. In the case of switchboard data speaker turns which are linguistically and semantically meaningful segments are considered, where as in the case of f2b corpus segments were based on pause.

4.6 Lexical and Positional Features

4.6.1 Lexical Features

Previous works on speech summarization have successfully used lexical features derived from ASR/manual transcripts of speech files. Though ASR is not available for many languages in world especially for less resource languages, it would not be wise to exclude it for languages like English where great amount of effort has been invested in building resources and techniques for ASR [45]. In the present study we use an open source Sphinx speech recognition tool available online [2] [46] to obtain ASR transcripts of speech segments. The recognition system uses open source acoustic models trained using speech from hub4 data which contains 140 hours of English broadcast news data collected between 1996 and 1997 and language model built using Gigaword corpus (1200M words) which contains news wire text. The accuracy of recognition on present data set (f2b) is 67%.

ASR transcript corresponding to each segment is obtained from the speech recognition system. In order to compute importance of an segment based on its lexical features, we use tf*idf based scores of the respective segments. The tf*idf based scores are computed using the method explained in Sec. 4.4.2.1. The tf*idf based scores are computed for all the segments in a news story. In order to measure the degradation of automatic summaries due to ASR errors, we have also generated tf*idf scores for manual transcripts of corresponding segments. The tf*idf scores derived from ASR transcripts are referred as $asr\_tf * idf$ and those obtained from manual transcripts are referred as $tf * idf$. Speech segments are ranked in decreasing order of these scores and top ranking segments are concatenated to generate a summary of desired length. ROUGE scores for the summaries generated using tf*idf based scores ($tf * idf$, $asr\_tf * idf$) are presented in Tab. 4.5.

[2] http://cmusphinx.sourceforge.net/wordpress/download/

Table 4.5 *F-measure values of ROUGE-1 (R-1), ROUGE-2 (R-2), ROUGE-SU4 (R-SU4) metrics for summaries based on lexical features ($tf * idf$, $asr\_tf * idf$).*

system	R-1	R-2	R-SU4
$tf * idf$	0.514[0.49 0.53]	0.337[0.31 0.35]	0.344[0.32 0.36]
$asr\_tf * idf$	0.470[0.45 0.49]	0.264[0.24 0.28]	0.276[0.25 0.29]

4.6.2 Correlation Between tf*idf based Summaries and Prominence based Summaries

In order to investigate the nature of the segments picked by the prominence based scoring we plotted a scatter plot between tf*idf based scores and acoustic scores (mdp) of speech segments. Fig. 4.9 shows scatter plot between tf*idf scores and acoustic scores (mdp) for segments picked in prominence (mdp) based summaries (a) and tf*idf based summaries (b) for two news stories 1 and 2. In Fig. 4.9 it can be observed from 1(a)

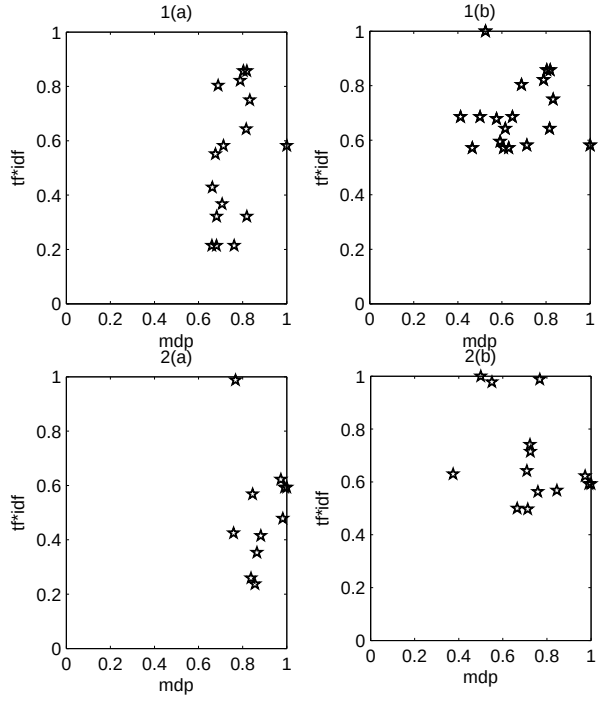

Figure 4.9 *Scatter plots between tf*idf scores and prominence score (mdp) for summaries of two news stories 1 and 2. (a) shows the scatter plot of scores for phrases picked in summaries based on prominence (mdp) scores. (b) shows the scatter plot of scores for phrases picked in summaries based on tf*idf scores.*

and 2(a) that some phrases picked in prominence based (mdp) 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. Summaries are generated for 30% cr. The ROUGE scores for these summaries are reported in Tab. 4.6.

Table 4.6 *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	0.520[0.50 0.54]	0.350[0.33 0.37]	0.356[0.33 0.37]

The ROUGE scores in Tab. 4.6 show that summaries generated by combined scores based on prominence and tf*idf scores are better than summaries generated by the individual systems (tf*idf in Tab. 4.1, mdp in Tab. 4.2).

4.6.3 Positional Information

The scoring of segments by acoustic scores and lexical scores is aimed at capturing acoustical evidence and lexical evidence for importance but, it is well known that in single document news article summarization, positional features of sentences also play a major role [44, 9]. It is widely accepted that initial sentences of a news article serve as good candidates for extractive summarization. Automatic summaries based on positional features are generated by picking segments from the beginning of a news story until the desired summary length is reached. ROUGE scores for summaries generated for 30% compression ratio using positional features are presented in Tab. 4.7.

Table 4.7 *F-measure values of ROUGE-1 (R-1), ROUGE-2 (R-2), ROUGE-SU4 (R-SU4) metrics for summaries based on positional features (lead).*

system	R-1	R-2	R-SU4
lead(pos)	0.512[0.48 0.52]	0.349[0.32 0.36]	0.352[0.33 0.36]

The ROUGE scores for lead summary are similar to those of summaries generated by tf*idf scores and acoustic score (mdp) as the data set used in current experiments belongs to news genre, and we are aiming at single document summarization.

4.6.4 Unsupervised System Using Prominence, Lexical and Positional Features.

In this section we propose a method to combine prominence, lexical and positional features to rank a speech segment for summarization in an unsupervised framework. Initial segments of the summary are extracted based on the positional features. The segments present in initial 5% of the speech document are included in the summary as they provide relevant background to the summary. The scores obtained from prominence based scoring (mdp) and lexical features ($tf * idf$, $asr\_tf * idf$) are normalized between 0 to 1 and a combined score is obtained by adding the normalized scores. Speech segments are ranked in decreasing order of their combined 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 summaries are generated for 30% compression ratio. ROUGE scores for summaries generated by combination of lexical ($tf * idf$, $asr\_tf * idf$), prominence (mdp) and positional (pos) features are reported in Tab. 4.8.

It can be observed that ROUGE scores of summaries generated using tf*idf scores ($tf * idf$) (Tab. 4.5) and positional features (lead) (Tab. 4.7) are slightly greater than the scores of summaries generated using prominence based acoustic scores (mdp) (Tab. 4.2). It can be observed that 95% confidence intervals for these systems overlap significantly and the

Table 4.8 *F-measure values of ROUGE-1 (R-1), ROUGE-2 (R-2), ROUGE-SU4 (R-SU4) metrics for summaries generated by unsupervised system using prominence score (mdp), lexical (mmr, $asrmmr$) and positional features (pos).*

system	R-1	R-2	R-SU4
mdp +pos	0.531[0.51 0.55]	0.368[0.34 0.38]	0.371[0.35 0.39]
mdp + $tf * idf$	0.520[0.50 0.54]	0.350[0.33 0.37]	0.356[0.33 0.37]
mdp + $tf * idf$ + pos	0.547[0.53 0.56]	0.390[0.38 0.40]	0.391[0.38 0.40]
mdp + $asr\_tf * idf$ + pos	0.496[0.47 0.51]	0.310[0.29 0.33]	0.319[0.29 0.33]

difference in the ROUGE scores is not statistically significant. It can observed that ROUGE scores of summaries generated by combination of lexical ($tf*idf$), positional features (pos) and prominence based acoustic scores (Tab. 4.8) are better than summaries generated by individual features.

4.6.5 Supervised System Using Prominence, Lexical and Positional Features.

We have also built a supervised system using prominence based acoustic scores (Sec. 4.5.1), tf*idf scores (Sec. 4.4.2.1), and positional features to evaluate its performance in comparison with supervised system trained on standard acoustic features such as F_0, duration and intensity (Sec. 4.4.2.2) along with tf*idf scores and positional features. The supervised system is an artificial neural network classifier similar to the one described in Sec. 4.4.2.2. The system was trained using gold standard human labelled summaries. The data set of 40 stories is randomly divided into two non overlapping halfs of which one is used for training and the other for testing. The feature vector on which the classifier was trained consists of prominence based acoustic scores (Sec. 4.5.1) such as mean prominence values of syllables (mp), maximum of prominence values of syllables (Mp), mean of difference between prominence values of consecutive syllables (mdp) and maximum of differences between prominence values of consecutive syllables (Mdp) as described in Sec.4.5.1

, tf*idf scores ($tf * idf$) obtained for manual transcripts of each segment as described in Sec. 4.4.2.1 and positional features. Positional feature of an segment was assigned three values based on the occurrence of the segment in a news story. All the segments in the initial 5% of the news story were assigned a feature value 1, all segments in the final 5% of a news story are assigned a value -1 and the rest are assigned 0. The tf*idf and prominence based acoustic scores are z-score normalized to bring them to zero mean and unit variance. Similarly another classifier was trained using acoustic features described in Sec. 4.4.2.2 along with tf*idf scores ($tf * idf$) and positional features of speech segments as explained above. Summaries are generated for 30% compression ratio.

Evaluation was done using ROUGE evaluation system. The text corresponding to audio summaries is obtained from transcripts provided with the corpus. We report ROUGE scores for two supervised systems, one trained on prominence based acoustic scores along with lexical ($tf * idf$) and positional features ($prom$ + $tf * idf$ + pos) and the other trained with standard acoustic features along with lexical ($tf * idf$) and positional features (A + $tf * idf$ + pos) in Tab. 4.9.

Table 4.9 *F-measure values of ROUGE-1 (R-1), ROUGE-2 (R-2), ROUGE-SU4 (R-SU4) metrics for summaries generated by supervised systems.*

system	R-1	R-2	R-SU4
$prom$ + $tf * idf$ + pos	0.583[0.57 0.59]	0.401[0.38 0.42]	0.410[0.39 0.43]
A + $tf * idf$ + pos	0.556[0.53 0.7]	0.381[0.36 0.40]	0.382[0.36 0.40]

It can be observed from Tab. 4.9 that supervised system trained using prominence features along with lexical ($tf * idf$) and positional features performs better than supervised system trained using standard acoustic features along with lexical ($tf * idf$) and positional features. The difference between scores of these two systems is statistically significant. This shows that when sufficient number of gold standard human reference summaries are

available to train an supervised system, using prominence based acoustic scores as features helps in generating better summaries than standard acoustic features.

4.7 Summary

In this chapter an automatic speech summarization system based on prominence was proposed. The proposed technique does not require ASR/manual transcripts or human reference summaries for training. Significance of prominence for speech summarization was shown by ranking speech segments with the help of hand labelled prominence markings. It was shown that prominence based ranking of speech segments captures prosodic information relevant to summarization. An automatic method to score the segments using prominence values of syllables within them was proposed. Evaluation results showed that the proposed technique generates summaries that are as good as summaries generated by text summarizer based on tf*idf scores and summaries generated by as supervised system trained on standard acoustic features. It was also shown that the proposed prominence based scoring captures complimentary information to td*idf based scoring and their combined scores generated summaries that are better than the summaries generated by individual features. It was also shown that supervised system trained on prominence based features generated better summaries than supervised system trained on standard acoustic features. The summaries are produced in form of speech such that characteristics of original signal are preserved. Summaries for desired compression ratios can be generated without loss in quality of summaries.

Chapter 5

Summary and Conclusions

5.1 Summary of the Work

In this thesis, techniques which do not depend on ASR transcripts and gold standard human summaries to summarize speech signals are explored. These methods explore the structural and prosodic features relevant for summarization. The experiments are carried out on two genres of speech documents namely, broadcast news shows where a specific structure is followed to deliver news and spontaneous telephone dialogues.

Broadcast news shows are a special type of speech documents which contain explicit structure and well defined speaker roles to deliver news. Human summaries of broadcast news shows are analyzed, which showed that anchor speaker segments are preferred to other speakers' in a summary. It is also observed that anchor speaker segments in the beginning of a news story are highly relevant to summary. This property is exploited to develop a method for summarizing broadcast news by performing anchor speaker tracking.

Two methods are proposed to perform anchor speaker tracking; based on auto-associative neural network (AANN) models and Bayesian information criterion (BIC). The features used for speaker tracking purpose are MFCC features which describe the spectral characteristics of vocal track. The former method based on AANN model requires a training phase where the model is trained on MFCC features extracted from the speech signal of the

speaker that needs to be tracked. It was observed that for effective tracking of the speaker the model requires a training speech of about 60 s. The amount of training data may vary from speaker to speaker depending on the variability within the speaker. To overcome this problem and to reduce the initial training speech required for speaker tracking, an iterative technique to train AANN model is proposed in the current thesis. As we are interested in tracking anchor speaker, the initial part of news show (20 s) where the anchor speaker delivers headlines is used for training AANN model in first iteration. The model is retrained after each iteration by adding features extracted from the newly tracked speaker segments in the iteration. The model converges after a few iterations. The final anchor speaker regions are obtained using this model. Though, this model effectively tracks anchor speaker in a news show, it still requires an initial training data and also multiple models need to be trained for the news shows with multiple anchor speakers. To overcome this problem, a method for tracking anchor speaker was proposed based on BIC. This method does not require initial training data and it can be easily extended to multiple anchor speakers. This method detects speaker change points by computing Δ BIC value between two windows. Speaker change points across the news show are obtained by shifting the windows along the time axis. Anchor speaker segments are obtained by performing agglomerative clustering of the segments between speaker change points using Δ BIC as distance measure. After obtaining the anchor speaker regions in a news show, segments from beginning of each anchor speaker region are concatenated until the desired summary length is reached. While concatenating the anchor speaker segments for summarization, it was observed that boundaries of these segments are abrupt and therefore, the resulting audio summaries are not coherent. To overcome this we have extended the boundaries to nearest 250 ms duration pause in the signal. As the speech signals belong to news style speech these pauses largely coincided with sentence and phrase boundaries. It was observed that this improved the quality of the audio summaries. The evaluation results showed that these summaries

are rated highly by humans showing that they are coherent and carry relevant information about news story. Audio summaries generated by anchor speaker tracking can be converted into text with less errors due to ASR as it is mostly read style speech and contains very few disfluencies when compared to other speakers in news show.

In this thesis, a technique to score speech segments based on prominence values of syllables present in them is proposed to capture prosodic information relevant to summarization. When humans convey message through speech they attract listener's attention towards information bearing parts of the signal by variations in pitch, intensity and duration. Speakers make some words prominent than others. It was shown in previous studies that content words (nouns,verbs and adjectives) are made prominent than function words and prominent words occur while introducing new concepts. Therefore, modelling prominence might help in capturing important content in the speech signals. Prominence value of a syllable is computed as a function of syllable nucleus duration, sub-band energy (300-2200 Hz) and pitch variation. In order to obtain syllable nucleus without text transcript in an unsupervised way, we applied modified convex hull algorithm on the filtered (300-900 Hz) energy envelope of speech signals. We experimented with four different scoring functions to obtain a segment level score from prominence values of syllables in the segment. Out of these mdp which computes the mean of the absolute difference between prominence values of consecutive syllables was found to be effective. This types of scoring is robust to speaker variations as the use of difference operator captures the variation in the prominence values rather than their absolute values which might be dependent on the speaker. ROUGE based evaluation showed that summaries generated by this method are as good as summaries generated by supervised system trained on gold standard human summaries and baseline text summarization system based on tf*idf scores. We have shown that this type of scoring captures complementary information to tf*idf scores of the text transcripts of the speech segments and a combination of these two features produced summaries of

better quality than individual systems. In case of spontaneous telephone dialogues where ASR transcripts are highly error prone, the proposed method produced summaries better than text summarization system taking ASR transcripts as input.

5.2 Conclusions from the Work

Experiments conducted in the current work showed that anchor speaker segments at the beginning of a news story in a news show are good candidates for generating extractive audio summaries of news shows. Picking these segments into audio summaries increased coherence of the summaries which are subsequently rated high by humans. We have also shown that prominence values of syllables in a speech segment can be used as a measure of importance attached by the speaker to the spoken segment. By ranking speech segments based on prominence values of syllables present it them, summaries that are as good as summaries produced by baseline text summarization system using tf*idf scores and supervised speech summarization system trained on standard acoustic features can be produced. It was also shown that proposed prominence based ranking carries complimentary information to tf*idf based ranking and increases the performance of the speech summarization based on text when incorporated in it.

5.3 Contributions of the Work

The important contributions of the thesis are exploiting speaker roles in broadcast news and prominence based features for summarization of speech signals. The major contribution of the thesis is in developing methods for the following:

- **Exploiting anchor speaker role to summarize broadcast news shows**. We analyzed the human summaries of broadcast news shows and found that anchor speaker's

speech is most relevant to summaries and segments in the beginning of a news story inside a news show are important for the summary.

- **Speaker tracking using auto-associative neural network model**. We proposed an iterative speaker tracking algorithm to train an auto associative neural network which can be trained using limited initial training data to track anchor speaker in a news show.

- **Anchor speaker tracking using ΔBIC as distance measure**. We proposed a method to cluster anchor speaker segments using Δ BIC as distance measure which does not need any training data a-priori.

- **Significance of prominence for speech summarization**. We showed that scoring speech utterances using prominence values of hand marked prominent syllables helps in generating speech summaries that are as good as tf*idf based scoring and supervised system trained using standard acoustic features with help of human labelled summaries.

- **Automatic scoring of speech segments using prominence values of syllables.** We proposed a method to score speech segments from the prominence values of syllables which captures the variations in prominence values and is robust to inter speaker variations. This type of scoring was shown to capture complimentary information to tf*idf based scoring. We proposed a method to combine prominence based scores and tf*idf scores. The resulting combined scores generated summaries that are better than the summaries generated by the individual features.

5.4 Limitations and Scope for Future Work

The work presented in this thesis can be extended and improved in certain aspects. Possible improvements and future research directions are given below.

- The prominence detection method described can be used only for languages similar to English and cannot be applied to tonal languages like Thai, Chinese, etc.,. In tonal languages the variations in pitch have a linguistic function where they discriminate between meaning of two words with same orthography. Therefore, features indicating prominence in tonal languages differ from other languages and there is a need to incorporate these features that indicate prominence in tonal languages in a speech summarization system based on prominence for these languages.

- The syllable segmentation method used in the current study is an unsupervised and threshold based technique. Therefore, it is sensitive to speaking rate variations and does not give reliable boundaries when the variation is high.

- The proposed prominence based ranking method makes an assumption that speaker emphasizes important content in his/her speech which may not be true in all cases.

- Determining inherent structure in speech documents by detecting topic shifts and discourse structure will be helpful for summarization. As observed in the case of broadcast news, where anchor speaker tracking helped in detecting topic shifts, acoustic features that indicate topic shifts need to be investigated in other genres of speech documents where there is no explicit structure.

- In the current work basic units of extraction are obtained based on pause duration (for read style speech); detecting meaningful semantic units of extraction for summarization and developing algorithms to extract these units reliably from a given speech signal are necessary to improve quality and coherence of the summaries.

Related Publications

- Sree Harsha Yella, Vasudeva Varma and Kishore Prahallad, 'Significance of Anchor Speaker Segments for Generating Extractive Audio Summaries of Broadcast News', Accepted for publication at *IEEE workshop on Spoken Language Technologies 2010*, Berkeley, CA, USA.

- Sree Harsha Yella, Vasudeva Varma and Kishore Prahallad, 'Prominence based Scoring of Speech Segments for Automatic Speech-to-Speech Summarization'. accepted for publication at *Interspeech 2010*, Makuhari, Japan.

- Sree Harsha Yella, Kishore Prahallad and Vasudeva Varma, 'Summarization of Broadcast News using Speaker Tracking', in *Proc. International Conference on Natural Language Processing 2009, Hyderabad, India*.

Bibliography

[1] R. Barzilay, M. Collins, J. Hirschberg, and S. Whittaker. The rules behind roles: Identifying speaker role in radio broadcasts. In *AAAI*, pages 679–684, Austin, Texas, USA, 2000.

[2] R. Barzilay and M. Elhadad. Using lexical chains for text summarization. In *ACL-WS1997A*, 1997.

[3] M. Beckman, J. Hirschberg, and S. Shattuck-Hafnagel. The original tobi system and evolution of the tobi framework. *Prosodic Models and Transcription: Towards Prosodic Typology*, pages 9–54, 2004.

[4] V. Bush. As we may think.. *The Atlantic Monthly*, 1976(1):101–108.

[5] J. Carbonell and J. Goldstein. 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, New York, NY, USA, 1998. ACM.

[6] S. S. Chen and P. S. Gopalakrishnan. Speaker, environment and channel change detection and clustering via the bayesian information criterion. In *DARPA Speech Recognition Workshop*, pages 127–132, 1998.

[7] H. Christensen, Y. Gotoh, B. Kolluru, and S. Renals. Are extractive text summarisation techniques portable to broadcast news? In *IEEE Speech Recognition and Understanding Workshop*, pages 489–494, 2003.

[8] H. Christensen, Y. Gotoh, and S. Renals. A cascaded broadcast news highlighter. *Audio, Speech, and Language Processing, IEEE Transactions on*, 16(1):151 –161, jan. 2008.

[9] H. Christensen, B. Kolluru, Y. Gotoh, and S. Renals. From text summarisation to style-specific summarisation for broadcast news. In *ECIR*, Sunderland, UK, 2004.

[10] J. Clark and C. Yallop. *Introduction to phonology and phonetics*. Blackwell, Oxford, UK, 1990.

[11] W. Croft and D. Harper. Using probabilistic models of information retrieval without relevance information. *Journal of Documentation*, 35:285–295, 1979.

[12] G. F. DeJong. An overview of the FRUMP system. In W. G. Lehnert and M. H. Ringle, editors, *Strategies for Natural Language Processing*, pages 149–176. Lawrence Erlbaum Associates, Publishers, 1982.

[13] H. P. Edmundson. New methods in automatic extracting. *JACM*, 16(2):264–285, April 1969.

[14] B. Endres-Niggemeyer. A grounded theory approach to expert summarizing. In *AAAI-SS1998A*, pages 133–135, 1998.

[15] B. Endres-Niggemeyer. *Summarizing information*. Springer, Berlin, 1998.

[16] H. Fletcher and W. A. Munson. Loudness, its definition, measurement and calculation. *The Journal of the Acoustical Society of America*, 5(2):82–108, 1933.

[17] Y. Fujii, N. Kitaoka, and S. Nakagawa. Automatic extraction of cue phrases for important sentences in lecture speech and automatic lecture speech summarization. In *Interspeech*, page 28012804, Antwerp, Belgium, 2007.

[18] D. Fum, G. Guida, and C. Tasso. Forward and backward reasoning in automatic abstracting. In *Proceedings of the Ninth International Conference on Computational Linguistics (COLING '82)*, pages 83–88, Prague, 1982.

[19] 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*, 12(4):401–408, July 2004.

[20] M. Galley. A skip-chain conditional random field for ranking meeting utterances by importance. In *EMNLP*, pages 364–372, Sydney, Australia, 2006.

[21] D. Harman. Overview of the first text retrieval conference. In *TREC 1992*, pages 1–20, Gaithersburg, MD, USA, 1992.

[22] J. Hieronymus. Automatic sentential vowel stress labelling. In *Eurospeech*, pages 1226–1229, Paris, France, 1989.

[23] J. Hirschberg, M. Bacchiani, D. Hindle, P. Isenhour, A. Rosenberg, L. Stark, L. Stead, S. Whittaker, and G. Zamchick. Scanmail: Browsing and searching speech data by content. In *Interspeech*, page 12991302, Aalborg, Denmark, 2001.

[24] J. Hirschberg, S. Whittaker, D. Hindle, F. Pereira, and A. Singhal. Finding information in audio: A new paradigm for audio browsing and retrieval. In *ESCA ETRW Workshop*, pages 117–122, Cambridge, UK, 1999.

[25] C. Hori and S. Furui. Automatic speech summarization based on word significance and linguistic likelihood. In *ICASSP*, pages 1579–1582, Istanbul, Turkey, 2000.

[26] T. Hori, C. Hori, and Y. Minami. Speech summarization using weighted finite state transducers. In *Interspeech*, pages 2817–2820, Geneva, Switzerland, 2003.

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

[28] E. Hovy and C. Y. Lin. Automated text summarization in summarist. In I. Mani and M. T. Maybury, editors, *Advances in Automatic Text Summarization*, pages 81–94. MITP, 1999.

[29] E. Hovy, C. Y. Lin, and L. Zhou. Evaluating DUC 2005 using basic elements. In *DUC2005*, 2005.

[30] A. W. Howitt. *Automatic Syllable Detection for Vowel Landmarks*. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, USA, 2000.

[31] J. Huang, G. Zweig, and M. Padmanabhan. Information extraction from voicemail. In *ACL*, pages 290–297, Toulouse, France, 2001.

[32] P. S. Jacobs and L. F. Rau. SCISOR: Extracting information from on-line news. *CACM*, 33(11):88–97, 1990.

[33] J. Jagarlamudi, P. Pingali, and V. Varma. Capturing sentence prior for query-based multi-document summarization. In *RIAO*, 2007.

[34] J. Janos. Theory of functional sentence perspective and its application for the purposes of automatic extracting. *Information Processing Management*, 15(1):19–25, 1979.

[35] M. Jansche and S. Abney. Information extraction from voicemail transcripts. In *EMNLP*, pages 320–327, Philadelphia, USA, 2002.

[36] K. S. Jones. What might be in a summary. *Information Retrieval 93: Von der Modellierung zur Anwendung, 9–26*, 1993.

[37] K. S. Jones. Automatic summarizing: Factors and directions. In *Inderjeet Mani and Mark T. Maybury, editors, Advances in Automatic Text Summarization*, pages 1–13. MIT Press, 1999.

[38] S. Jones, K. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28:11–21.

[39] M. Kameyama and I. Arima. Coping with aboutness complexity in information extraction from spoken dialogues. In *ICSLP*, pages 87–90, Yokohama, Japan, 1994.

[40] R. A. Knight. The realisation of intonational plateaux: Effects of foot structure. *Cambridge Occasional Papers in Linguistics, L. Astruc and M. Richards, Eds. Cambridge, UK*, pages 157–164, 2004.

[41] G. Kochanski, E. Grabe, J. Coleman, and B. Rosner. Loudness predicts prominence: Fundamental frequency lends little. *The Journal of the Acoustical Society of America*, 118(2):1038–1054, 2005.

[42] B. Kolluru, H. Christensen, and Y. Gotoh. Multi-stage compaction approach to broadcast news summarisation. In *Eurospeech*, pages 69–72, Lisbon, Portugal, 2005.

[43] K. Koumpis and S. Renals. Automatic summarization of voicemail messages using lexical and prosodic features. *ACM Transactions on Speech and Language Processing*, 2:1–24, 2005.

[44] J. Kupiec, J. Pedersen, and F. Chen. A trainable document summarizer. In *SIGIR '95: Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 68–73, New York, NY, USA, 1995. ACM.

[45] K.-F. Lee. *Large-vocabulary speaker-independent continuous speech recognition: the sphinx system*. PhD thesis, Carnegie Mellon University, Pittsburgh, PA, USA, 1988. AAI8826533.

[46] K.-F. Lee, H.-W. Hon, and M.-Y. Hwang. Recent progress in the sphinx speech recognition system. In *HLT '89: Proceedings of the workshop on Speech and Natural Language*, pages 125–130, Morristown, NJ, USA, 1989. Association for Computational Linguistics.

[47] I. Lehiste, U. zu Koln., and L. Institute. *Suprasegmentals*. M.I.T. Press Cambridge, Mass.,, 1970.

[48] C. Y. Lin. Rouge: a package for automatic evaluation of summaries. In *ACL Text summarization Workshop*, 2004.

[49] Y. Liu, F. Liu, B. Li, and S. Xie. Do disfluencies affect meeting summarization: A pilot study on the impact of disfluencies. In *MLMI*, page poster, Brno,Czech Republic, 2007.

[50] H. P. Luhn. The automatic creation of literature abstracts. *IBM Journal of Research Development*, 2(2):159–165, 1958.

[51] I. Mani. *Automatic Summarization*. John Benjamins, 2001.

[52] W. Mann and S. Thompson. Rhetorical structure theory: Towards a functional theory of text organization. *Text*, 8(3):243–281, 1988.

[53] D. Marcu. Discourse trees are good indicators of importance in text. In I. Mani and M. T. Maybury, editors, *Advances in Automatic Text Summarization*, pages 123–136, Cambridge, MA, 1995. MITP.

[54] D. Marcu. *The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts*. PhD thesis, Department of Computer Science, University of Toronto, December 1997.

[55] E. Maron, M and L. Khuns, J. Probabilistic indexing and information retrieval. *Journal of the ACM*, 7(3):216–244.

[56] S. Maskey and J. Hirschberg. Comparing lexical, acoustic/prosodic, structural and discourse features for speech summarization. In *ICSLP*, pages 621–624, Lisbon, Portugal, 2005.

[57] S. Maskey and J. Hirschberg. Summarizing speech without text using hidden markov models. In *NAACL '06: Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers on XX*, pages 89–92, Morristown, NJ, USA, 2006. Association for Computational Linguistics.

[58] B. Mathis. *Techniques for the evaluation and improvement of computer produced abstracts*. Ohio State University Technical Report OSU-CISRC-TR-72-15, Ohio, USA.

[59] M. T. Maybury. Generating summaries from event data. *IPM*, 31(5):735–751, September 1995.

[60] K. McKeown, J. Hirschberg, M. Galley, and S. Maskey. From text to speech summarization. In *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on*, volume 5, pages v/997 – v1000 Vol. 5, 18-23 2005.

[61] P. Mermelstein. Automatic segmentation of speech into syllabic units. *The Journal of the Acoustical Society of America*, 58(4):880–883, 1975.

[62] G. Murray, S. Renals, and J. Carletta. Extractive summarization of meeting recordings. In *Interspeech*, Lisbon, Portugal, 2005.

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

[64] G. Murray, S. Renals, J. Carletta, and J. Moore. Evaluating automatic summaries of meeting recordings. In *ACL MTSE Workshop*, pages 33–40, AnnArbor, MI, USA, 2005.

[65] G. Murray, S. Renals, J. Moore, and J. Carletta. Incorporating speaker and discourse features into speech summarization. In *HLT-NAACL*, pages 367–374, New York City, USA, 2006.

[66] L. Neumeyer, H. Franco, M. Weintraub, and P. Price. Automatic text-independent pronunciation scoring of foreign language student speech. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, volume 3, pages 1457 –1460 vol.3, 3-6 1996.

[67] K. Ohtake, K. Yamamoto, Y. Toma, S. Sado, S. Masuyama, and S. Nakagawa. Newscast speech summarization via sentence shortening based on prosodic features. In *ISCA and IEEE Workshop on Spontaneous Speech Processing and Recognition*, pages 167–170, Tokyo, Japan, 2003.

[68] K. Ono, K. Sumita, and S. Miike. Abstract generation based on rhetorical structure extraction. In *COLING*, pages 344–348, Kyoto, Japan, 1994.

[69] C. D. Paice. The automatic generation of literary abstracts: An approach based on identification of self-indicating phrases. In O. R. Norman, S. E. Robertson, C. J. van Rijsbergen, and P. W. Williams, editors, *Information Retrieval Research*, London: Butterworth, 1980.

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

[71] P. Pingali, R. Katragadda, and V. Varma. Iiit hyderabad at duc 2007. In *working notes of DUC at the annual meeting of Document Understanding Conferences (DUC)*, 2007.

[72] J. J. Pollock and A. Zamora. Automatic abstracting research at chemical abstracts service. *Journal of Chemical Information and Computer Sciences*, 15(4), 1975.

[73] T. Portele and B. Heuft. Towards a prominence-based synthesis system. *Speech Commun.*, 21(1-2):61–72, 1997.

[74] L. Rabiner and B.-H. Juang. *Fundamentals of speech recognition*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1993.

[75] D. Radev, T. Allison, S. Blair-Goldensohn, J. Blitzer, A. Çelebi, S. Dimitrov, E. Drabek, A. Hakim, W. Lam, D. Liu, J. Otterbacher, H. Qi, H. Saggion, S. Teufel, M. Topper, A. Winkel, and Z. Zhang. MEAD - a platform for multidocument multilingual text summarization. In *LREC*, Lisbon, Portugal, May 2004.

[76] D. R. Radev, S. Blair-Goldensohn, and Z. Zhang. Experiments in single and multi-document: Summarization using MEAD. In *DUC2001*, 2001.

[77] D. R. Radev, H. Jing, and M. Budzikowska. Centroid-based summarization of multiple documents: Sentence extraction, utility-based evaluation, and user studies. In *NAACL-WS2000A*, 2000.

[78] N. Reithinger, M. Kipp, R. Engel, and J. Alexandersson. Summarizing multilingual spoken negotiation dialogues. In *ACL*, pages 310–317, Morristown, NJ, USA, 2000.

[79] V. Rijsbergen. *Information Retrieval*. Butterworth, London, U.K.

[80] R. Rohlicek, J. Gisting continuous speech. In *ICASSP*, pages 384–387, San Francisco, USA, 1992.

[81] P. Rousseeuw. *Robust regression and outlier detection*. Wiley, New York, 1987.

[82] J. E. Rush, A. Zamora, and R. Salvador. Automatic abstracting and indexing. ii, production of abstracts by application of contextual inference and syntactic coherence criteria. *JASIS*, 22(4):260–274, 1971.

[83] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24:513–523, 1988.

[84] 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)*, page 243, Washington, DC, USA, 2000. IEEE Computer Society.

[85] S. Simpson and Y. Gotoh. Towards speaker independent features for information extraction from meeting audio data. In *MLMI*, page poster, Edinburgh, UK, 2005.

[86] A. Singhal. Modern information retrieval: A brief overview. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 24(4), 35-43.

[87] A. M. Sluijter, A. M. C, S. Vincent, and V. J. van Heuven. Acoustic correlates of linguistic stress and accent in dutch and american english. In *Proceedings of the International Conference on Spoken Language Processing*, pages 630–633, 1996.

[88] 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*, pages 551–554, 1999.

[89] F. Tamburini and C. Caini. An automatic system for detecting prosodic prominence in american english continuous speech. *International Journal of Speech Technology*, 8(1):33–44, January 2005.

[90] P. Taylor.

[91] P. Taylor. Analysis and synthesis of intonation using the tilt model. *Journal of the Acoustical Society of America*, 107:1697–1714, 2000.

[92] J. Terken. Fundamental frequency and perceived prominence of accented syllables. *The Journal of the Acoustical Society of America*, 89(4):1768–1776, 1991.

[93] J. M. B. Terken. Variation of accent prominence within the phrase: Models and spontaneous speech data. *Computing Prosody for Spontaneous Speech, Y. Sagisaka, W. Campbell and N. Higuchi (Eds). Berlin, Germany*, pages 95–116, 1997.

[94] S. H. Teufel and M. Moens. Sentence extraction as a classification task. In *ACL-WS1997A*, 1997.

[95] R. Valenza, T. Robinson, M. Hickey, and R. Tucker. Summarization of spoken audio through information extraction. In *ESCA Workshop on Accessing Information in Spoken Audio*, pages 111–116, Cambridge, UK, 1999.

[96] V. Varma, P. Pingali, R. Katragadda, and Others. Iiit hyderabad at tac 2008. In *working notes of Text Analysis Conference (TAC) at the joint meeting of the annual conferences of TAC and TREC*, 2008.

[97] E. Voorhees and D. Harman. Overview of the seventh text retrieval conference (trec-7). In *TREC*, pages 1–24, 1999.

[98] A. Waibel, M. Bett, M. Finke, and R. Stiefelhagen. Meeting browser: Tracking and summarizing meetings. In *Broadcast News Transcription and Understanding Workshop*, pages 281–286, Lansdowne, VA, USA, 1998.

[99] D. Wang and S. Narayanan. An acoustic measure for word prominence in spontaneous speech. *Audio, Speech, and Language Processing, IEEE Transactions on*, 15(2):690–701, Feb. 2007.

[100] B. Yegnanarayana. *Artificial Neural Networks*. Prentice-Hall of India Pvt.Ltd, 2004.

[101] B. Yegnanarayana, K. S. Reddy, and S. P. Kishore. Source and system features for speaker recognition using aann models. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pages 409–412, 2001.

[102] K. Zechner. Automatic generation of concise summaries of spoken dialogues in unrestricted domains. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 199–207, New York, NY, USA, 2001. ACM.

[103] K. Zechner. Automatic summarization of open-domain multiparty dialogues in diverse genres. *Computational Linguistics*, 28(4):447–485, 2002.

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

[105] X. Zhu and G. Penn. Summarization of spontaneous conversations. In *Interspeech*, pages 1531–1534, Pittsburgh, USA, 2006.