Computational Modelling for Coherence in Spoken Discourse

Undergraduate Thesis

Submitted in partial fulfillment of the requirements of
BITS F421T Thesis

By

Rajaswa Patil
ID No. 2017A3TS0334G

Under the supervision of:

Dr. Rajiv Ratn Shah
&
Dr. Ashish Chittora

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE PILANI, GOA CAMPUS
November 2021

Declaration of Authorship

I, Rajaswa Patil, declare that this Undergraduate Thesis titled, 'Computational Modelling for Coherence in Spoken Discourse' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: _____

Date: _____

Certificate

This is to certify that the thesis entitled, "*Computational Modelling for Coherence in Spoken Discourse*" and submitted by Rajaswa Patil ID No. 2017A3TS0334G in partial fulfillment of the requirements of BITS F421T Thesis embodies the work done by him under my supervision.

Supervisor	*Co-Supervisor*
Dr. Rajiv Ratn Shah	Dr. Ashish Chittora
Assistant Professor,	Assistant Professor,
IIIT Delhi	BITS Pilani, K. K. Birla Goa Campus
Date:	Date:

"Thinking and spoken discourse are the same thing, except that what we call thinking is, precisely, the inward dialogue carried on by the mind with itself without spoken sound."

Plato

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE PILANI, GOA CAMPUS

Abstract

Bachelor of Engineering (Hons.) Electrical & Electronics Engineering

Computational Modelling for Coherence in Spoken Discourse

by Rajaswa Patil

This thesis examines the role of audio modality in the coherence of spoken discourse. While there has been significant progress towards modelling coherence in written discourse, the work in modelling spoken discourse coherence has been quite limited. Unlike the coherence in text, coherence in spoken discourse is also dependent on the prosodic and acoustic patterns in the speech audio. The goal of this thesis is to provide evidence for the same by performing computational modelling for coherence in spoken discourse. The method followed includes modelling coherence in spoken discourse with audio-based coherence models and performing experiments with four coherence-related tasks with spoken discourses: Speaker Change Detection, Artificial Speech Evaluation, Discourse Topic Change Detection, and Speech Response Proficiency Scoring. In our experiments, we evaluate machine-generated speech against the speech delivered by expert human speakers. We also compare the spoken discourses generated by human language learners of varying language proficiency levels. The results show that incorporating the audio modality along with the text benefits the coherence models in performing downstream coherence related tasks with spoken discourses.

Acknowledgements

I would like to thank my supervisors Dr. Rajiv Ratn Shah and Dr. Ashish Chittora for their constant support and guidance throughout the process of carrying out this work. I would also like to thank Mr. Yaman Kumar (Doctoral Researcher, IIIT-Delhi) for his guidance, and SLTI Inc. for providing the data and feedback for this work. Lastly, I would like to thank my peers, family, and friends, who believed in me and provided a constant support and motivation during the process of carrying out this work.

Contents

List of Figures

List of Tables

Abbreviations

Coh-T	**Coh**erence **T**ext
Coh-A	**Coh**erence **A**udio
Coh-AT	**Coh**erence **A**udio + **T**ext
SCD	**S**peaker **C**hange **D**etection
ASE	**A**rtificial **S**peech **E**valuation
TCD	**T**opic **C**hange **D**etection
SRS	**S**peech **R**esponse **S**coring

Dedicated to the reader.

Chapter 1

Introduction

1.1 Motivation

Most of the previous work dealing with coherence modelling has been limited to evaluating the semantic organization of the discourse content with text-based coherence systems [23], and there has been limited work on modelling spoken discourse coherence as a task [69]. Unlike the coherence in text, coherence in spoken discourse is also dependent on the speech perception, which, in turn dependent on the prosodic and acoustic patterns in speech audio. The few studies which have tried to work on spoken discourse coherence have done so by transcribing speech and then applying text-coherence modelling methods on it [71]. Modelling coherence of a spoken discourse with its text-transcriptions is an inherently lossy and challenging task [59]. On the one hand, crucial cues of speech such as pauses, tonal variations, speed changes, stress, rhythm and intensity are lost while transcribing it, and on the other, the transcription is itself a cumbersome process with involved logistics and errors from automatic speech recognition (ASR) systems [11]. Various studies in linguistics have highlighted the importance of prosody in providing a structure to the spoken discourse [46, 6]. By incorporating the audio modality along with the text-transcription of the speech, we can better model the coherence in spoken discourse.

1.2 Goal

The goal of this thesis is to inspect the role of speech audio in the coherence of spoken discourse. This thesis does not attempt to introduce another model of coherence dealing with the semantic organization of the content. Instead, this thesis tries to inspect the role of speech perception in discourse coherence. Further, the thesis tries to lay down guidelines for any derivative studies that shall be conducted for modelling coherence in spoken discourse.

1.3 Approach

Since this is a preliminary work in modelling spoken discourse coherence, the main focus is on designing a variety of coherence-related tasks with spoken discourses. The tasks are designed in such a way, which allows us to inspect the wide role that the speech audio plays in discourse coherence individually as well as along with its text content. We do not focus on developing new models of discourse coherence, and instead borrow existing suitable state-of-the-art models from previous work in text-based coherence modelling. We use these models with text-only, audio-only, and text with audio settings. This allows us to provide independent insights for semantic organization and perception aspects of discourse coherence.

1.4 Chapter Summary

The goal of this thesis is to inspect the role of speech audio in the coherence of spoken discourse. This is a preliminary work in modelling spoken discourse coherence, where the main focus is on designing a variety of coherence-related tasks with spoken discourses, instead of proposing new models of discourse coherence. This allows us to provide independent insights for semantic organization and perception aspects of discourse coherence.

Chapter 2

Background

2.1 Chapter Overview

While the surface-level linguistic definitions of discourse coherence have been previously explored widely by evaluating the semantic organization of discourse content, the perception and participation aspects of discourse coherence have not yet been studied from a computational perspective. This chapter reviews previous work in theories and models of discourse coherence, which serves as a vital starting point for the thesis.

Section 2.2 discusses various theories of discourse coherence from linguistic and non-linguistic perspectives. Section 2.3 provides references to previous methods of modelling discourse coherence, which is divided into subsections for organizational convenience. Section 2.3.1 discusses preliminary studies of discourse coherence modelling, Section 2.3.2 discusses deep-learning based neural-network models of discourse coherence, Section 2.3.3 briefly describes various applications of modelling discourse coherence and Section 2.3.4 provides background in the form of early studies done towards understanding spoken discourse coherence. Finally, Section **??** summarizes the background and describes the motivation for possible research directions.

2.2 Theories of Discourse Coherence

Discourse is defined as a coherent group of written sentences or spoken utterances obtained from communication between a writer and reader, or a speaker and listener [24, 72]. Hence, coherence is the most fundamental property of any discourse, whether it be written, or spoken. Broadly, the existing theories of discourse coherence can be classified into two perspectives: *discourse-as-product* and *discourse-as-process* [72].

The theories of coherence with the *discourse-as-product* perspective define its coherence with surface-level linguistic definitions pertaining to the semantic organization of the discourse content, and the linguistic devices used to connect the ideas in a discourse [72]. The earliest such definition was presented by *Halliday and Hasan (1976)*, where they defined coherence as the existence of cohesion and register in the discourse content [19]. Around the same time, *van Dijk (1977)* defined coherence to be a semantic property of the discourse at linear and global levels [68]. Where, linear coherence refers to coherence relations between a sequence of consecutive sentences, while the global coherence characterizes larger spans in the discourse. *Mann and Thompson (1987)* introduced the Rhetorical Structure Theory (RST) [62], which models coherence as a hierarchical structure of functional chunks connected together with rhetorical relations. For a coherent text, the smaller chunks at the lower levels of the RST structure shall form a united structure [72]. Further, *Danes (1974)* and *Fries (1983)* defined coherence as the degree of the connectivity of themes in the sentences of a discourse [5, 14], and *Widdowson (1978)* defined coherence as the pragmatic relationship between the illocutionary acts used in the discourse [74].

On the other hand, the theories of coherence with the *discourse-as-process* perspective consider discourse to be *"a dynamic process of interaction between communicator and audience, during which language serves as a medium" (Wang and Guo, 2014)*. Here, coherence deals with the perception and participation aspects of the discourse rather than the discourse content itself [4, 72, 33]. *Brown and Yule (1983)* defined coherence with a psychological perspective, where the backward knowledge of the participants of a discourse played a vital role in the interpretation of its coherence [49]. *Hu Zhuanglin (1994)* defined discourse coherence from a situational and cultural context [77]. Further, the dynamic nature of discourse was studied under the Speech Act Theory [1] and the Conversational Implicature, where coherence was achieved in a linguistically incoherent phenomenon [72]. *Givon (1995)* defined discourse coherence to be a mental phenomenon where it was stated that: *"Coherence is not an internal property of a written or spoken text, (but) a property of what emerges during speech production and comprehension—the mentally represented text, and in particular the mental processes that partake in constructing that mental representation"* [15]. While these aspects of coherence might not always be that significant with written discourses, they are quite significant with spoken discourses, where speech's perception and delivery play a very important role.

2.3 Related Work

2.3.1 Coherence Modelling

Early work in modelling discourse coherence focused on extracting features based on the Centering Theory [16] and the entity transitions in the text [29, 9]. *Barzilay and Lapata (2008)* introduced

the entity grid representation of discourse [2], which was based on discourse entities and their grammatical role transitions. The entity grid model was further improved for coherence-related tasks by *Elsner and Charniak (2011)* [10], *Feng and Hirst (2012)* [12], and *Louis and Nenkova (2012)* [37]. Parallely, many works [53, 35, 13] performed coherence-related tasks based on discourse relations in the text, parsed with theories like the Rhetorical Structure Theory (RST) [62] and the Lexicalized Tree Adjoining Grammar for discourse (D-LTAG) [73] with the Penn Discourse Treebank (PDTB) [54] styled annotations. Notably, the features based on the RST-encodings were found to be useful for modelling coherence in spoken discourse [71] and more efficient than the PDTB-encodings for modelling text coherence as well [13]. Further, *Guinaudeau and Strube (2013)* [18], and *Mesgar and Strube (2015)* [40] proposed graph representation-based approaches to model coherence in text.

2.3.2 Neural Coherence Models

Following the advances in deep neural network architectures and distributed semantic representations, there has been much progress towards developing neural models of discourse coherence which provide significant performance gains over the traditional feature-based models. The entity grid representation of discourse got extended with neural architectures. *Tien Nguyen and Joty (2017)* [63] proposed the neural entity grid model, which performed convolutions over the entity grid representations. Further, *Joty et al. (2018)* [22] lexicalized the neural entity grid model by attaching the entities to their respective grammatical roles in the entity grid embeddings.

Neural coherence models can be broadly classified into two categories: *generative coherence models* and *discriminative coherence models*. On the one hand, generative coherence models deal with modelling the conditional probabilities of a sentence being coherent with a given set of preceding sentences [32, 36]. On the other hand, discriminative coherence models are trained to classify coherent and incoherent texts. It has been previously shown that modelling local coherence with discriminative models can be beneficial for capturing both the local and the global contexts of coherence with good approximation [45, 75]. Similarly, capturing relations and similarities between sentences at a local level with neural models can be helpful with coherence-related tasks [31, 39]. Recent work in coherence modelling has focused on building models in open-domain [32] and cross-domain [75] settings. More recently *Lai and Tetreault (2018)* [27] built coherence models and datasets for real-world texts. Some recent work has also focused on building benchmarks for applying coherence models in the qualitative evaluation of text-based natural language generation systems [44].

2.3.3 Applications of Coherence Modelling

Discourse coherence can be used as an auxiliary metric to evaluate the quality of a given discourse. Previously, discourse coherence has been used to evaluate written discourses for tasks like essay scoring and readability assessment [42, 3, 39]. Coherence based metrics and objectives have also been used to evaluate and improve text-based artificial natural language generation systems [29, 50, 25] for tasks like text-summarization [2, 48], machine-translation [56], language modelling [21, 28] and conversation thread reconstruction [22]. Therefore, modelling discourse coherence has become an essential task in computational linguistics with a variety of downstream applications.

2.3.4 Coherence in Spoken Discourse

Coherence deals with the *perception* of the discourse rather than the discourse content itself [4, 72, 33]. While the perception of a written discourse is only affected by the semantic organization of its lexical content, the perception of a spoken discourse is additionally dependent on its prosodic and acoustic features

2.3.4.1 Prosody and Coherence

[20]. Previous work in linguistics has highlighted the role of prosody in defining the structure for spoken discourse. *Nakajima and Allen (1993)* performed experiments with cooperative dialogues and demonstrated the role of prosodic information in defining the topic structure of a given spoken discourse [46]. Further, *Degand and Simon (2009)* introduced prosodic segmentation to define basic discourse units in speech [6].

Various previous studies have used prosodic attributes to perform coherence-related tasks with spoken discourse. *Nakajima and Allen (1993)* analysed the role of intonation and pause durations in modelling semantic relationships between discourse utterances at topic boundaries [46]. Further, *Tür et al. (2001)* used duration and pitch based features to perform the task of topic segmentation, which is closely related to both the local and global coherence of spoken discourse [65]. *Stifelman (1995)* used pitch patterns to perform emphasis detection with automated discourse segmentation [58]. This was further used to summarize and skim through spoken discourses, a task which is highly relevant to the comprehensibility and the perception of spoken discourse.

2.3.4.2 Automatic Speech Scoring

Apart from the previously mentioned tasks, automated speech scoring is an another important application of modelling discourse structure and coherence. Explicitly annotated coherence

based measures [69, 70] and features extracted from discourse structures in text-transcriptions of spoken discourses [71] help in improving the performance of the automated speech scoring systems significantly. Unlike the work done with essay scoring as an auxiliary evaluation task for coherence modelling, work in speech scoring has been limited to include discourse coherence related features from the text-transcriptions of spoken discourse along with other features relevant to speech scoring.

2.4 Chapter Summary and Research Directions

2.4.1 Chapter Summary

Past work has defined coherence from a *discourse-as-product* and *discourse-as-process* perspectives. While a lot of work has been done in coherence modelling for the former, the later has not been explored computationally before. Modelling discourse coherence has a lot of downstream applications in computational linguistics. Where, modelling spoken discourse coherence has a significant application in speech proficiency scoring.

2.4.2 Research Directions

The review of previous works indicates that the *perception* aspect of discourse coherence is as important as the semantic organization of discourse content, and the coherence in spoken discourse is highly dependent on the speech perception. Hence, the following research directions can be explored in modelling spoken discourse coherence:

- Developing coherence models with the audio modality.
- Building spoken discourse coherence related tasks and their corresponding datasets.
- Inspecting the role of spoken discourse coherence in speech response scoring with respect to the language proficiency.

Chapter 3

Datasets

3.1 Chapter Overview

Data plays a vital role in modelling discourse coherence. While a stimulus covering a wide range of coherent and incoherent discourses is required from the training data, building task-specific datasets from coherent discourses is equally important in order to evaluate and apply the coherence models. Developing datasets for coherence modelling in spoken discourses is quite challenging in this regard.

Section 3.2 discusses the prerequisites for a dataset to be ideal for spoken discourse coherence related tasks. Section 3.3 describes the datasets used in this thesis, and Section 3.4 describes the use of text-to-speech systems in generating artificial data for our experiments.

3.2 Requirements for the Datasets

A training dataset for modelling discourse coherence should possess a certain amount of structure, equally covering (or holding a superficial ability to cover) coherent and incoherent discourses. It is relatively easier to procure text datasets of such varying levels of structure from sources like Wikipedia, social networks, and other online medium. There is a significant lack of such structured discourse-rich datasets for speech. Therefore, choosing the correct datasets for training coherence models of spoken discourse is quite challenging. At the same time, building datasets for coherence related tasks with spoken discourses should possess both the content and perception aspects of the discourse.

3.3 Description of the Datasets

3.3.1 IBM Debater Dataset

We perform all our experiments with the debate speech samples from the IBM Debater dataset[1] [30]. The dataset consists of recordings of debate speeches delivered by nine expert debaters, with debate speeches on 200 distinct Wikipedia topics (such as social media, nuclear weapons, gambling, *etc.*) as the debate motions. Each motion topic is contested with two debate recordings from distinct experts, resulting in a total number of 400 unique speech samples. This makes the dataset rich with a variety of coherent speeches spread across a variety of open-domain topics, providing a high quality training signal to our coherence models.

3.3.2 L2 Simulated Oral Proficiency Interview (SOPI) Dataset

We use an another dataset of speech responses from non-native English language learners for one of the evaluation tasks, the details for which are given in Section 5.1.4. The statistics for all the datasets are mentioned in Table 5.1.

3.4 Synthesized Speech Datasets

We generate new datasets[2] for various evaluation tasks (explained in Section 5.1.4) using the responses from the IBM Debater dataset. For this, we use the text-transcriptions from the debate speech recordings to synthesize artificial speech responses with a standard text-to-speech (TTS) system based on the Microsoft Speech API (SAPI5) [41]. We use two distinct TTS voices across all our experiments: **S1** (US-male voice) and **S2** (US-female voice).

3.5 Chapter Summary

In this thesis, we use the IBM Debater Datasets to train our coherence models and perform our experiments. We also use a dataset of non-native English speakers to evaluate our models. Further, we generate new task-specific data with text-to-speech systems using the Microsoft Speech API for a variety of evaluation tasks.

[1] `https://www.research.ibm.com/haifa/dept/vst/debating_data.shtml`
[2] The datasets are available here

Chapter 4

Modelling

4.1 Chapter Overview

There has been no previous work done towards modelling coherence in spoken discourses with the audio modality. This chapter discusses the guidelines for developing coherence models of spoken discourse and describes our modelling methodology.

Section 4.2 discusses the general guidelines for developing coherence models of spoken discourse. Section 4.3 describes the text-based models used in this thesis. Section 4.4 describes the audio-based models used in this thesis, and Section 4.5 summarizes the chapter.

4.2 Coherence Models for Spoken Discourse

A coherence model for spoken discourse should be able to capture both the prosodic (pitch, intonation and stress) and the acoustic features (fundamental frequency, intensity and duration) of an audio sample. It is relatively easier to procure text datasets of varying levels of structure from sources like Wikipedia, social networks, and other online medium. Given the lack of such structured discourse-rich datasets for speech, the model should generalize beyond closed-domain settings [32, 75] and perform well on more open and cross-domain settings with limited training data [21, 44]. This becomes more important with spoken discourse as it has been shown that the audio modality is more vulnerable to change in data domain and background as compared to text [76]. Given the above mentioned challenges and pitfalls related to modelling coherence in spoken discourse, an ideal audio-based coherence model should:

1. learn a discourse coherence signal with a limited number of training samples

2. generalize across speech samples which vary in terms of the acoustic and prosodic features of the speech audio (Ex: accent, gender, rhythm, age, speed and intonation) and differences in the background of data like recording quality, sampling frequency, background noise *etc.*

3. and, similar to the text domain, generalize across speech samples which vary in terms of the spoken discourse's topic and content.

4.3 Text-based Discrimination Model

4.3.1 Distributed Word Representations

Recent development in deep-learning based neural network architectures has resulted in a progress towards distributed semantics. Distributed word representations obtained from such neural-network models are quite useful for a variety of downstream tasks in computational linguistics. Such distributed word representations are also more commonly know as word embeddings. In this work, we use pre-trained *Global Vectors for Word Representations* (GloVe) embeddings [52] to encode the text from the sentence into its corresponding text embedding.

4.3.2 Local Discrimination Model

The local discrimination algorithm proposed by *Xu et al. (2019)* is designed to maximize the local coherence scores of adjacent pair of sentences and minimize the local coherence scores for the non-adjacent pair of sentences in a given discourse [75]. Unlike the older discrimination models which suffer with class-imbalance between the coherent and incoherent permutations of written discourses, this approach captures local coherence with an effective negative sampling of the incoherent non-adjacent sentences. The model takes in a pair of sentence representations as an input, which are further passed through a multi-layered perceptron with a single hidden layer (Figure 4.1) to obtain a local coherence score. Experiments done by *Xu et al. (2019)* [75] show that the global aspects of coherence can be approximated by using the local coherence scores from their models with techniques like score-averaging across the discourse. Further, the local discrimination model learns to generalize in open-domain as well as cross-domain settings (as shown by their sentence-ordering and paragraph-reconstruction experiments with domain-separated Wikipedia articles), and is agnostic to the modality and background of the input data. Hence, for all our experiments, we use the local discrimination model proposed by *Xu et al. (2019)* [75].

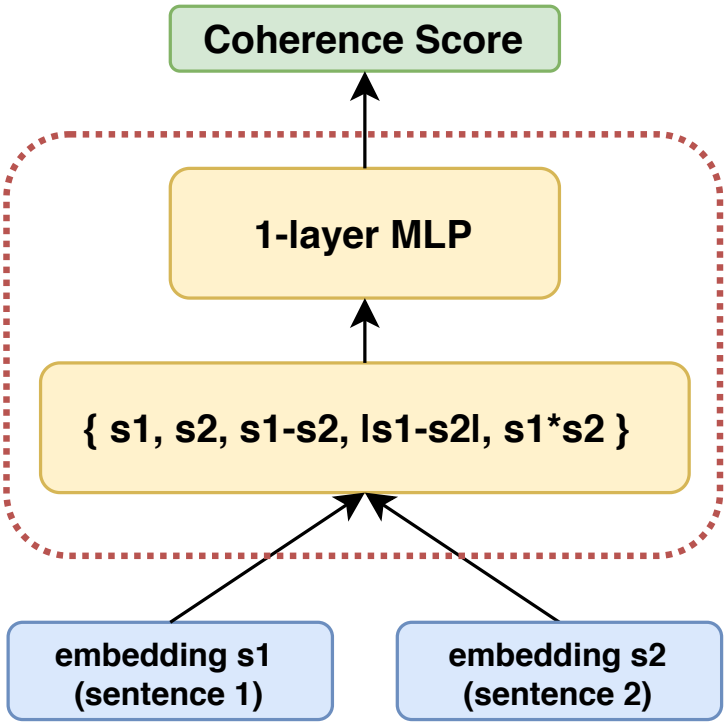

FIGURE 4.1: The local discrimination coherence model.

4.4 Audio-based Discrimination Models

4.4.1 Speech Representation Learning

Learning latent representations from a time-frequency signal of audio data is very important for a variety of tasks in speech processing. Recent developments in deep-learning based Convolutional Neural Network architectures have enabled researchers to develop state-of-the-art speech representation learning models. These models take in the raw audio waveforms and spectrograms as their inputs, and output a latent vector representation.

The audio waveforms are usually passed through filters, filter-banks, and a series of time-frequency transforms. The resultant data is then used by the models on a power-scale or a log-scale. The models are trained to reconstruct segments of audio signal given certain parts of the signal as the context. The latent representations learnt by these models in the process, are rich in prosodic, lexical and acoustic information from the speech audio, and can be used for a variety of downstream applications.

4.4.2 Multimodal Local Discrimination Model

To incorporate the audio modality into the coherence model, we encode an audio based sentence embedding, similar to the sentence embedding obtained from the text modality. We use a

pre-trained audio language model: wav2vec [55] to encode the audio segment of a sentence into its corresponding audio embedding as shown in Figure 4.2. Wav2vec is pre-trained with an unsupervised objective of next time-step prediction task for audio segments. This objective function aligns significantly with that of the local discrimination coherence model, providing rich audio representations for our task. Similarly, we use pre-trained GloVe embeddings [52] to encode the text from the sentence into its corresponding text embedding.

In order to inspect the role of the audio modality in modelling coherence for spoken discourses, we experiment with three different learning settings (Coh-T, Coh-A and Coh-AT) for the local discrimination coherence model (Figure 4.2). Similar to all the previous work, with the first setting, we just use the text modality as the input to the coherence model (**Coh-T**). In the second setting, we use only the audio modality as the input to the coherence model, to establish an audio-only control setting for our experiments (**Coh-A**). Finally, in the third setting we obtain a multimodal input representation by fusing the text and audio modalities together (**Coh-AT**). In order to get a minimal trainable aggregated fusion of the two modalities, we pass the audio and the text embeddings through a bi-linear layer as shown in Figure 4.2. Following *Xu et al. (2019)*'s approach, we aggregate the coherence scores from both the forward model *(sentence-1, sentence-2)* and the backward model *(sentence-2, sentence-1)* [75].

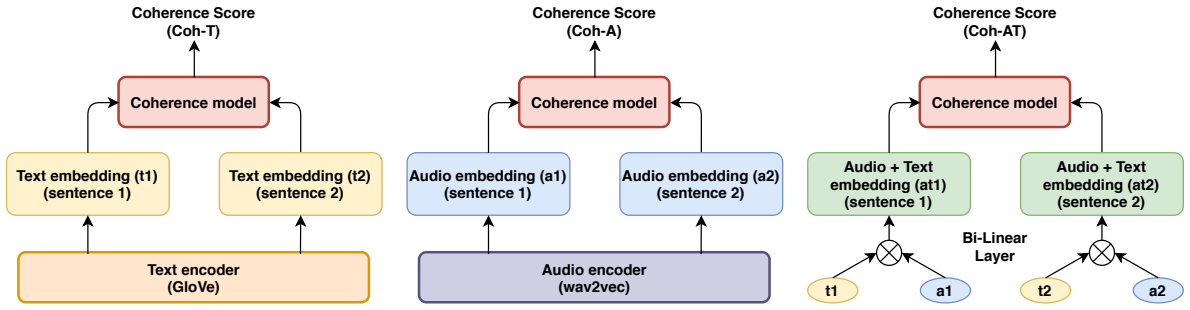

FIGURE 4.2: Model architecture across the three input settings: Coh-T, Coh-A and Coh-AT.

4.5 Chapter Summary

This chapter described the details about the methodology followed in this work to develop coherence models of spoken discourses.

Chapter 5

Experiments

5.1 Evaluation Tasks

Empirical results from previous work in coherence modelling for text has shown that the traditional synthetic tasks like sentence ordering do not effectively capture the models' ability to perform downstream discourse coherence related tasks with real-world data [27, 44]. An ideal coherence model should perform well for both spoken discourses delivered by humans and machine-generated speech. Keeping this in mind, we design four tasks for evaluating coherence in spoken discourse: *Speaker Change Detection* (**SCD**), *Artificial Speech Evaluation* (**ASE**), *Discourse Topic Change Detection* (**TCD**) and *Speech Response Scoring* (**SRS**). While the first three tasks focus on artificially generated speech and speech delivered by expert human speakers, the fourth task focuses on comparing the spoken discourses delivered by non-native language learners of varying proficiency levels.

5.1.1 Speaker Change Detection

Modelling coherence in a conversational setting is a very important task with various downstream applications [22, 67]. To obtain a structured conversational text-transcript from a given speech audio, we first need to perform speaker diarization. Hence, speaker diarization is an important aspect of modelling conversational spoken discourse [43]. Acoustic cues play an important role in speaker segmentation of conversational speech [51]. We construct a new dataset for this task using the speech samples from the IBM Debater dataset [30]. We sample a ten-sentences long segment from the middle of every response. The first five sentences from the sampled segment are synthesized with a TTS voice (S1) and the next five sentences are synthesized with another TTS voice (S2). Consequently, the overall synthesized speech response consists of a speaker change at the end of the fifth sentence, while maintaining a continuation in the discourse topic.

This speaker change in the response can be detected with an audio-based coherence model, where the event of speaker change can be depicted with the least inter-sentence local coherence score. Further in a separate experiment, we reverse the order of speakers, such that the first five sentences are synthesized with TTS voice S2 and the next five sentences are synthesized with TTS voice S1.

5.1.2 Artificial Speech Evaluation

Following the work done in evaluating the quality of text-based natural language generation systems with coherence-based measures [29, 50, 25], we propose evaluating the TTS systems with coherence models of spoken discourse. TTS systems often face issues while naturalizing the synthesized speech to make it more human-like and intelligible across longer contexts [61]. Modelling discourse relations and coherence can benefit a TTS system in delivering expressive and intelligible speech [7]. In this task, we evaluate and compare a TTS speech sample which lacks a certain amount of prosodic variation in terms of intonation, stress and speaking rate, to the speech from a expert human speaker. The underlying hypothesis for this particular task is that the lack of prosodic variation in the TTS sample makes it relatively incoherent as compared to the human speech, even though the delivered lexical content is same across both the samples. An ideal audio-based coherence model should assign a lesser coherence score to the TTS response as compared to its corresponding expert human speech response.

5.1.3 Discourse Topic Change Detection

Following the work done by *Tür et al. (2001)* [66] in automatic topic segmentation with prosodic cues, we propose a coherence-based topic change detection task for spoken discourse. For this, we construct a new dataset using the responses from the IBM Debater dataset [30]. These samples majorly evaluate the extent to which a model captures the prosodic features at topic boundaries [46]. For constructing a sample, we select a five-sentences long segment from the middle of every response, so that the sampled segment represents a developed topic rather than introductory definitions or concluding statements. Subsequently, we combine it with a similarly sampled segment from a different motion topic. This results in a new ten-sentences long speech response which covers a particular topic in its first five sentences and a different topic in the next five sentences. We use a text-to-speech system to synthesize the speech audio for this newly generated response. The underlying hypothesis for this particular task is that the local inter-sentence coherence score should be the lowest for the fifth and the sixth sentence, depicting a change in discourse topic for the given speech response.

5.1.4 Speech Response Scoring

Coherence scores follow a monotonic relationship with the holistic language proficiency grades. Previous work in text-based coherence modelling has used essay scoring as an auxiliary evaluation task [42, 39]. Conversely, modelling coherence has also proved to be beneficial in essay-scoring benchmarks [60, 34]. In a like manner, coherence-based features extracted from the text-transcriptions of spoken discourses have proved to be useful in scoring speech responses from language learners and non-native English speakers [69, 70]. Following this, we test our coherence models on a dataset of spoken discourses delivered by non-native English language learners from Philippines [17]. The dataset comprises of speech responses recorded in a test environment where the candidates are asked to respond to six distinct prompts. They are subsequently double scored by expert annotators using a holistic language proficiency level on a 6-point CEFR scale council2001common. We construct pairs of speech responses, such that every pair contains speech responses from two different speakers, for the same prompt. Given such a pair of speech responses, we hypothesize that the response graded with higher holistic proficiency level, should be assigned higher coherence scores by a coherence model.

Spoken discourse from non-native speakers is usually less structured as compared to a discourse delivered by a native expert speaker [76], making it more challenging to model discourse coherence for non-native speakers. The text-transcripts of the speech responses in the dataset are not structured with proper punctuation, which is needed to obtain sentence-level segments of the speech response. So, we punctuate the text-transcripts from the dataset with a punctuator model [64]. Given the background and the pre-processing involved, this dataset is more noisy and challenging as compared to the IBM Debater dataset. Moreover, while the discourses from the IBM Debater dataset are more argumentative and informative in nature, the responses in this dataset are more descriptive and narrative in nature. Hence, both the datasets vastly differ in terms of discourse modes [57, 8].

Task	**Number of speech responses**		
	Train	**Validation**	**Test**
SCD*	-	-	398
ASE*	197	78	120
TCD*	-	-	786
SRS	463	181	234

TABLE 5.1: Statistics: Dataset Samples Distribution

5.2 Experimental Setup

Audio Processing: Across all our experiments, we use a speaking-rate of 150 words per minute to synthesize speech responses with the TTS systems, a value which is recognized as the average speaking rate for a native US-English speaking adult [47]. Unlike structured texts, there are no explicit cues to perform a sentence-level segmentation in speech. We use a pre-trained Montreal Forced Aligner (MFA[1]) model [38] and the punctuation from the structured text-transcriptions to get the sentence-level alignments for the speech audio files. Further, all the speech responses are resampled to a 16kHz mono-channel audio file as required by the pre-trained wav2vec model.

Coherence Modelling: Following *Xu et al. (2019)*'s [75] training protocol, we sample the incoherent pair of sentences within the same speech response. This avoids the model pitfalls related to the topic and speech based features with the local discrimination setting . Further, to make the model generalize well across various domains and sources of the data, we do not fine-tune the pre-trained audio and text encoders on the training data. We train the model to optimize the local coherence scoring based margin loss objective as shown in Equation 5.1, where f^+ and f^- are coherence scores for the adjacent (coherent) and non-adjacent (incoherent) pairs of sentences, respectively. We use 50% of the topics in the dataset to train our model, while the rest 20% and 30% of the topics are used for validation and testing purposes, respectively. The model parameters are optimized with Adam optimizer [26] with a learning rate of 0.001. We validate the models with an early stop callback on the validation loss, with a patience of two epochs. Given, the cross-domain adaptation abilities of the local discrimination model, we borrow the hyperparameter settings from *Xu et al. (2019)* [75] and do not perform any extensive hyperparameter tuning during our experiments.

$$L(f^+, f^-) = \max(0, 5 - f^+ + f^-) \tag{5.1}$$

$$k_{change} = \min_{k \in [1, N-1]} \{{f^+}_k\} \tag{5.2}$$

$$coherence\ score = \frac{1}{N-1} \sum_{k=1}^{N-1} {f^+}_k \tag{5.3}$$

[1] `https://montreal-forced-aligner.readthedocs.io/`

While the SCD task and TCD task are evaluated at a local level with inter-sentence coherence scores (Equation 5.2), the SRS task is evaluated with response-level coherence scores as shown in Equation 5.3, where N is the number of sentences in the speech response.

Chapter 6

Discussion

6.1 Results

Task SCD: For a ten-sentences long response, task SCD results in a nine-way classification setting. The top-k (*k=1,2,3*) accuracy scores for the SCD task are shown in Table 6.1. As expected, the Coh-T model fails to capture the speaker change boundaries (with the accuracy scores being almost equal to that of random guessing) due to the lack of access to the acoustic information from the speech audio. The Coh-A model shows impressive accuracy for this task, consistent across both the orders of speaker-change. The model predicts almost all the speaker change boundaries for *k=3*. The Coh-AT model does not match up in performance against the Coh-A model, suggesting a difference in audio-based learning between the two input settings.

Model	Change	k=1	k=2	k=3
Coh-T	-	0.0944	0.2041	0.3138
Coh-A	**S1 → S2**	0.9770	0.9898	**0.9949**
	S2 → S1	0.9796	0.9974	**1.0000**
Coh-AT	**S1 → S2**	0.7398	0.8954	0.9311
	S2 → S1	0.5026	0.6913	0.7730

TABLE 6.1: Top-k accuracy scores for the SCD task.

Negative sampling within the same speech response restricts the model to look at audio segments from different speakers under the audio-based settings. Consequently, the model is only exposed to small prosodic and acoustic variations from the same speaker during training. This shows that the local discrimination model captures even large acoustic changes in a given conversational spoken discourse, by modelling local coherence with cues from small acoustic and prosodic variations in monologue speech.

Speaker		Coh-T	Coh-A	Coh-AT
Human Expert	f^+	0.75	0.97	0.84
	f^-	-1.14	-1.76	-2.52
	% diff	-252%	-281.40%	**-400%**
TTS voice (S1)	f^+	0.75	1.76	1.08
	f^-	-1.14	1.62	-0.92
	% diff	-252%	-7.90%	-185.20%
TTS voice (S2)	f^+	0.75	2.73	0.77
	f^-	-1.14	2.55	-1.39
	% diff	-252%	-6.60%	-280.50%

TABLE 6.2: Mean coherence scores of positive (adjacent) and negative (non-adjacent) pairs of sentences from the speech samples in the test set along with the relative difference (% diff) between them. A higher % diff value is indicative of better coherence models and more coherent spoken discourses.

Task ASE: Given the difference in the audio data backgrounds for expert human speakers and TTS systems, we compare their coherence by monitoring the relative difference between the mean coherence scores of the coherent (adjacent) and incoherent (non-adjacent) pairs of sentences sampled from the speech responses generated by them (Table 6.2). In accordance with the training objective function, the incoherent sentences are scored lesser than the coherent sentences (negative relative difference) across all the speaker and model settings. A higher relative difference between the coherence scores of coherent and incoherent pairs of sentences not only indicates the coherence model's ability to effectively model coherence (horizontal traversal across Table 6.2), but it also indicates the speaker's ability to produce more coherent discourses (vertical traversal across Table 6.2). While the Coh-T model gives a relative difference of -252% on the samples from human experts, the audio-based Coh-A and Coh-AT models give much higher relative differences of -281.40% and -400%, respectively. This shows that incorporating the audio modality highly benefits a coherence model to capture the difference between coherent and incoherent samples. Further, while the Coh-T model is independent of any changes in the speech audio (same mean coherence scores of 0.75 and -1.14 for all the speakers), comparing the relative differences across the speakers, we observe that the TTS voices S1 and S2 show significantly lower relative differences across both the audio-based settings (-7.90% and -6.60% for Coh-A and, -185.20% and -280.50% for Coh-AT, respectively). Hence, under the ASE task, we find that the speech synthesized by TTS systems is relatively incoherent and more difficult to perceive as compared to human-generated speech.

Task TCD: Similar to the SCD task, the TCD task comes up with a nine-way classification setting. The top-k (*k=1,2,3*) accuracy scores for the TCD task are shown in Table 6.3. The Coh-A model does not perform well on the TCD task individually, with the accuracy scores

Model	Speaker	k=1	k=2	k=3
Coh-T	-	**0.2513**	0.3980	0.5434
Coh-A	**S1**	0.1148	0.2156	0.3444
	S2	0.1135	0.2742	0.3801
Coh-AT	**S1**	0.2385	**0.4031**	0.5383
	S2	0.2449	**0.4056**	**0.5663**

TABLE 6.3: Top-k accuracy scores for the TCD task.

being almost equal to that of random guessing. While the Coh-T model slightly outperforms the Coh-AT model for *k=1*, the Coh-AT model shows slight improvements over the Coh-T model for *k=2* with both the TTS voices S1 and S2. Further, for *k=3*, the Coh-AT model shows significant improvement over the Coh-T model for TTS voice S2. Even though topic segmentation is predominantly a text-based task, the slight improvements shown by Coh-AT model over the text-only settings can be explained by the presence of cues related to the prosodic patterns observed at topic boundaries 10.5555/898272.

	Coh-T	Coh-A	Coh-AT
Train	0.477	0.5594	0.4216
Valid	0.5278	0.6944	0.5417
Test	0.4641	**0.7046**	0.5569

TABLE 6.4: Accuracy scores for the SRS task with non-native speech dataset.

Task SRS: For this task, we monitor the accuracy scores for the binary classification setting based on the holistic language proficiency grades, using response-level coherence scores as the proficiency measure (Table 6.4). While the Coh-A model performs significantly well with an accuracy score of 0.70 on the test set, the Coh-T and Coh-AT models fail to converge. Given the lack of structure in non-native speech and the noise in the text-transcriptions of the speech-responses, text-based settings do not capture the complex holistic grades efficiently. On the other hand, the audio-based setting seems to be resistant to this lack of structure and noise in transcriptions and it effectively captures the holistic language proficiency grades.

Chapter 7

Conclusion

7.1 Summary

The near-perfect performance of the Coh-A model in predicting the speaker changes suggests that modelling coherence with the audio modality can turn out to be quite beneficial for a variety of discourse related tasks in conversational speech such as speech act detection, conversation disentanglement, etc. To further the efforts made in naturalizing the speech synthesized with text-to-speech systems, one can come up with better coherence-based objectives to train the TTS systems. Building up on the topic change detection task, coherence models for spoken discourses can be also evaluated on related downstream applications like topic-segmentation in spoken lectures, podcasts, political spoken discourses, etc. Further, the significantly higher performance of the Coh-A model on the SRS task shows that modelling coherence with audio modality can highly compensate for the lack of structure and errors in text-transcriptions of the speech. This can be quite useful while modelling coherence with data from non-native speakers, language learners or while using error-prone text-transcriptions from automatic speech recognition (ASR) systems.

In this thesis, we performed experiments with four coherence-related tasks for spoken discourse. In our experiments, we compare the speech synthesized with text-to-speech systems against the expert human speakers. We also evaluate coherence in spoken discourses delivered by non-native language learners of varying language proficiency levels. Our experiments show that incorporating the audio-modality betters the coherence-modelling for spoken discourses significantly.

7.2 Future Work

While this is a preliminary work done towards modelling coherence in spoken discourse, there are many future directions that one can follow building up on it. Building real-world datasets for coherence-related tasks is a significant challenge that one can take on. Further, designing more interpretable tasks for spoken discourse coherence is a good future research direction as well. Wit developments in tasks and datasets, we can then further push for developments in modelling spoken discourse coherence, with applications in various downstream tasks in speech processing and computational linguistics.

Bibliography

[1] John L Austin. "How to do things with words: Lecture I". In: *How to do things with words: JL Austin* (1962), pp. 1–11.

[2] Regina Barzilay and Mirella Lapata. "Modeling Local Coherence: An Entity-Based Approach". In: *Computational Linguistics* 34.1 (2008), pp. 1–34. DOI: 10.1162/coli.2008.34.1.1. URL: https://www.aclweb.org/anthology/J08-1001.

[3] Jill Burstein, Joel Tetreault, and Slava Andreyev. "Using Entity-Based Features to Model Coherence in Student Essays". In: *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Los Angeles, California: Association for Computational Linguistics, June 2010, pp. 681–684. URL: https://www.aclweb.org/anthology/N10-1099.

[4] Henri Cohen, Josée Douaire, and Mayada Elsabbagh. "The role of prosody in discourse processing". In: *Brain and Cognition* 46.1 (2001), pp. 73 –82. ISSN: 0278-2626. DOI: https://doi.org/10.1016/S0278-2626(01)80038-5. URL: http://www.sciencedirect.com/science/article/pii/S0278262601800385.

[5] Frantisek Danes. *Papers on functional sentence perspective*. Berlin, Boston: De Gruyter Mouton, 10 Mar. 2015. ISBN: 978-3-11-167652-4. DOI: https://doi.org/10.1515/9783111676524. URL: https://www.degruyter.com/view/title/59162.

[6] Liesbeth Degand and Anne Catherine Simon. "On identifying basic discourse units in speech: theoretical and empirical issues". In: *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics* 4 (2009).

[7] Rodolfo Delmonte and Rocco Tripodi. "Semantics and Discourse Processing for Expressive TTS". In: *Proceedings of the First Workshop on Linking Computational Models of Lexical, Sentential and Discourse-level Semantics*. Lisbon, Portugal: Association for Computational Linguistics, Sept. 2015, pp. 32–43. DOI: 10.18653/v1/W15-2704. URL: https://www.aclweb.org/anthology/W15-2704.

[8] Swapnil Dhanwal et al. "An Annotated Dataset of Discourse Modes in Hindi Stories". English. In: *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, May 2020, pp. 1191–1196. ISBN: 979-10-95546-34-4. URL: https://www.aclweb.org/anthology/2020.lrec-1.149.

[9] Micha Elsner, Joseph Austerweil, and Eugene Charniak. "A Unified Local and Global Model for Discourse Coherence". In: *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*. Rochester, New York: Association for Computational Linguistics, Apr. 2007, pp. 436–443. URL: https://www.aclweb.org/anthology/N07-1055.

[10] Micha Elsner and Eugene Charniak. "Extending the Entity Grid with Entity-Specific Features". In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, Oregon, USA: Association for Computational Linguistics, June 2011, pp. 125–129. URL: https://www.aclweb.org/anthology/P11-2022.

[11] Rahhal Errattahi, Asmaa El Hannani, and Hassan Ouahmane. "Automatic Speech Recognition Errors Detection and Correction: A Review". In: *Procedia Computer Science* 128 (2018). 1st International Conference on Natural Language and Speech Processing, pp. 32–37. ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2018.03.005. URL: http://www.sciencedirect.com/science/article/pii/S1877050918302187.

[12] Vanessa Wei Feng and Graeme Hirst. "Extending the Entity-based Coherence Model with Multiple Ranks". In: *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Avignon, France: Association for Computational Linguistics, Apr. 2012, pp. 315–324. URL: https://www.aclweb.org/anthology/E12-1032.

[13] Vanessa Wei Feng, Ziheng Lin, and Graeme Hirst. "The Impact of Deep Hierarchical Discourse Structures in the Evaluation of Text Coherence". In: *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Dublin, Ireland: Dublin City University and Association for Computational Linguistics, Aug. 2014, pp. 940–949. URL: https://www.aclweb.org/anthology/C14-1089.

[14] Peter H Fries. "On the status of theme in English: Arguments from discourse". In: *Micro and macro connexity of texts* 45 (1983).

[15] T. Givón. *Functionalism and Grammar*. John Benjamins, 1995. URL: https://www.jbe-platform.com/content/books/9789027273994.

[16] Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. "Centering: A Framework for Modeling the Local Coherence of Discourse". In: *Computational Linguistics* 21.2 (1995), pp. 203–225. URL: https://www.aclweb.org/anthology/J95-2003.

[17] Manraj Singh Grover et al. *Multi-modal Automated Speech Scoring using Attention Fusion*. 2020. arXiv: 2005.08182 [cs.CL].

[18] Camille Guinaudeau and Michael Strube. "Graph-based Local Coherence Modeling". In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Sofia, Bulgaria: Association for Computational Linguistics, Aug. 2013, pp. 93–103. URL: https://www.aclweb.org/anthology/P13-1010.

[19] Michael Alexander Kirkwood Halliday and Ruqaiya Hasan. *Cohesion in english*. Routledge, 2014.

[20] Julia Hirschberg and Christine H. Nakatani. "A Prosodic Analysis of Discourse Segments in Direction-Giving Monologues". In: *34th Annual Meeting of the Association for Computational Linguistics*. Santa Cruz, California, USA: Association for Computational Linguistics, June 1996, pp. 286–293. DOI: 10.3115/981863.981901. URL: https://www.aclweb.org/anthology/P96-1038.

[21] Dan Iter et al. "Pretraining with Contrastive Sentence Objectives Improves Discourse Performance of Language Models". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 4859–4870. DOI: 10.18653/v1/2020.acl-main.439. URL: https://www.aclweb.org/anthology/2020.acl-main.439.

[22] Shafiq Joty, Muhammad Tasnim Mohiuddin, and Dat Tien Nguyen. "Coherence Modeling of Asynchronous Conversations: A Neural Entity Grid Approach". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 558–568. DOI: 10.18653/v1/P18-1052. URL: https://www.aclweb.org/anthology/P18-1052.

[23] Shafiq Joty et al. "Discourse Analysis and Its Applications". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 12–17. DOI: 10.18653/v1/P19-4003. URL: https://www.aclweb.org/anthology/P19-4003.

[24] Daniel Jurafsky and James H. Martin. *Speech and Language Processing (2nd Edition)*. USA: Prentice-Hall, Inc., 2009. ISBN: 0131873210.

[25] Chloé Kiddon, Luke Zettlemoyer, and Yejin Choi. "Globally Coherent Text Generation with Neural Checklist Models". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 329–339. DOI: 10.18653/v1/D16-1032. URL: https://www.aclweb.org/anthology/D16-1032.

[26] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. 2015. URL: http://arxiv.org/abs/1412.6980.

[27] Alice Lai and Joel Tetreault. "Discourse Coherence in the Wild: A Dataset, Evaluation and Methods". In: *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 214–223. DOI: 10.18653/v1/W18-5023. URL: https://www.aclweb.org/anthology/W18-5023.

[28] Zhenzhong Lan et al. "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations". In: *International Conference on Learning Representations*. 2020. URL: https://openreview.net/forum?id=H1eA7AEtvS.

[29] Mirella Lapata and Regina Barzilay. "Automatic Evaluation of Text Coherence: Models and Representations". In: *Proceedings of the 19th International Joint Conference on Artificial Intelligence*. IJCAI'05. Edinburgh, Scotland: Morgan Kaufmann Publishers Inc., 2005, 1085–1090.

[30] Tamar Lavee et al. "Towards Effective Rebuttal: Listening Comprehension Using Corpus-Wide Claim Mining". In: *Proceedings of the 6th Workshop on Argument Mining*. 2019, pp. 58–66.

[31] Jiwei Li and Eduard Hovy. "A Model of Coherence Based on Distributed Sentence Representation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 2039–2048. DOI: 10.3115/v1/D14-1218. URL: https://www.aclweb.org/anthology/D14-1218.

[32] Jiwei Li and Dan Jurafsky. "Neural Net Models of Open-domain Discourse Coherence". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, Sept. 2017, pp. 198–209. DOI: 10.18653/v1/D17-1019. URL: https://www.aclweb.org/anthology/D17-1019.

[33] Junyi Li. "From Discourse Structure To Text Specificity: Studies Of Coherence Preferences". In: (2017).

[34] Xia Li et al. "Coherence-Based Automated Essay Scoring Using Self-attention". In: *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*. Springer, 2018, pp. 386–397.

[35] Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. "Automatically Evaluating Text Coherence Using Discourse Relations". In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, Oregon, USA: Association for Computational Linguistics, June 2011, pp. 997–1006. URL: https://www.aclweb.org/anthology/P11-1100.

[36] Lajanugen Logeswaran, Honglak Lee, and Dragomir R Radev. "Sentence Ordering and Coherence Modeling using Recurrent Neural Networks". In: *AAAI*. 2018.

[37] Annie Louis and Ani Nenkova. "A Coherence Model Based on Syntactic Patterns". In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Jeju Island, Korea: Association for Computational Linguistics, July 2012, pp. 1157–1168. URL: https://www.aclweb.org/anthology/D12-1106.

[38] Michael McAuliffe et al. "Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi." In: 2017.

[39] Mohsen Mesgar and Michael Strube. "A Neural Local Coherence Model for Text Quality Assessment". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 4328–4339. DOI: 10.18653/v1/D18-1464. URL: https://www.aclweb.org/anthology/D18-1464.

[40] Mohsen Mesgar and Michael Strube. "Graph-based Coherence Modeling For Assessing Readability". In: *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*. Denver, Colorado: Association for Computational Linguistics, June 2015, pp. 309–318. DOI: 10.18653/v1/S15-1036. URL: https://www.aclweb.org/anthology/S15-1036.

[41] Microsoft. *Microsoft Speech API (SAPI) 5.4*. 2009. URL: https://docs.microsoft.com/en-us/previous-versions/windows/desktop/ee125663(v=vs.85)).

[42] E. Miltsakaki and K. Kukich. "Evaluation of Text Coherence for Electronic Essay Scoring Systems". In: *Nat. Lang. Eng.* 10.1 (Mar. 2004), 25–55. ISSN: 1351-3249. DOI: 10.1017/S1351324903003206. URL: https://doi.org/10.1017/S1351324903003206.

[43] Mohammad Hossein Moattar and Mohammad M Homayounpour. "A review on speaker diarization systems and approaches". In: *Speech Communication* 54.10 (2012), pp. 1065–1103.

[44] Tasnim Mohiuddin et al. *CohEval: Benchmarking Coherence Models*. 2020. arXiv: 2004.14626 [cs.CL].

[45] Han Cheol Moon et al. "A Unified Neural Coherence Model". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 2262–2272. DOI: 10.18653/v1/D19-1231. URL: https://www.aclweb.org/anthology/D19-1231.

[46] Shin Nakajima and James F. Allen. *A Study on Prosody and Discourse Structure in Cooperative Dialogues*. Tech. rep. USA, 1993.

[47] NCVS. *National Center for Voice and Speech (NCVS)*. 2020. URL: http://ncvs.org/e-learning/tutorials/qualities.html.

[48] Ani Nenkova and Kathleen McKeown. *Automatic summarization*. Now Publishers Inc, 2011.

[49] Marion Owen. "G. Brown and G. Yule, Discourse analysis. Cambridge: Cambridge University Press, 1983. Pp. xii 288. - M. Stubbs, Discourse analysis. Oxford: Basil Blackwell, 1983. Pp. xiv 272." In: *Journal of Linguistics* 21.1 (1985), 241–245. DOI: 10.1017/S0022226700010161.

[50] Cesc C Park and Gunhee Kim. "Expressing an Image Stream with a Sequence of Natural Sentences". In: *Advances in Neural Information Processing Systems 28*. Ed. by C. Cortes et al. Curran Associates, Inc., 2015, pp. 73–81. URL: http://papers.nips.cc/paper/5776-expressing-an-image-stream-with-a-sequence-of-natural-sentences.pdf.

[51] Tae Jin Park and Panayiotis Georgiou. "Multimodal Speaker Segmentation and Diarization Using Lexical and Acoustic Cues via Sequence to Sequence Neural Networks". In: *Interspeech 2018* (2018). DOI: 10.21437/interspeech.2018-1364. URL: http://dx.doi.org/10.21437/Interspeech.2018-1364.

[52] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. "GloVe: Global Vectors for Word Representation". In: *Empirical Methods in Natural Language Processing (EMNLP)*. 2014, pp. 1532–1543. URL: http://www.aclweb.org/anthology/D14-1162.

[53] Emily Pitler and Ani Nenkova. "Revisiting Readability: A Unified Framework for Predicting Text Quality". In: *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. Honolulu, Hawaii: Association for Computational Linguistics, Oct. 2008, pp. 186–195. URL: https://www.aclweb.org/anthology/D08-1020.

[54] Rashmi Prasad et al. "The Penn Discourse TreeBank 2.0." In: *LREC*. Citeseer. 2008.

[55] Steffen Schneider et al. "wav2vec: Unsupervised Pre-Training for Speech Recognition". In: *Interspeech 2019* (2019). DOI: 10.21437/interspeech.2019-1873. URL: http://dx.doi.org/10.21437/interspeech.2019-1873.

[56] Karin Sim Smith, Wilker Aziz, and Lucia Specia. "The Trouble with Machine Translation Coherence". In: *Proceedings of the 19th Annual Conference of the European Association for Machine Translation*. 2016, pp. 178–189. URL: https://www.aclweb.org/anthology/W16-3407.

[57] Wei Song et al. "Discourse Mode Identification in Essays". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, July 2017, pp. 112–122. DOI: 10.18653/v1/P17-1011. URL: https://www.aclweb.org/anthology/P17-1011.

[58] Lisa J Stifelman. "A discourse analysis approach to structured speech". In: 1995.

[59] Heike Tappe and Frank Schilder. "Coherence in Spoken Discourse". In: *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 2*. Montreal, Quebec, Canada: Association for Computational Linguistics, Aug. 1998, pp. 1294–1298. DOI: 10.3115/980691.980780. URL: https://www.aclweb.org/anthology/P98-2211.

[60] Yi Tay et al. *SkipFlow: Incorporating Neural Coherence Features for End-to-End Automatic Text Scoring*. 2017. arXiv: 1711.04981 [cs.AI].

[61] Paul Taylor. *Text-to-speech synthesis*. Cambridge university press, 2009.

[62] Sandra A Thompson and William C Mann. "Rhetorical structure theory: A framework for the analysis of texts". In: *IPRA Papers in Pragmatics* 1.1 (1987), pp. 79–105.

[63] Dat Tien Nguyen and Shafiq Joty. "A Neural Local Coherence Model". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, July 2017, pp. 1320–1330. DOI: 10.18653/v1/P17-1121. URL: https://www.aclweb.org/anthology/P17-1121.

[64] Ottokar Tilk and Tanel Alumäe. "Bidirectional Recurrent Neural Network with Attention Mechanism for Punctuation Restoration". In: *Interspeech 2016*. 2016.

[65] Gökhan Tür et al. "Integrating prosodic and lexical cues for automatic topic segmentation". In: *Computational linguistics* 27.1 (2001), pp. 31–57.

[66] Gökhan Tür et al. "Integrating Prosodic and Lexical Cues for Automatic Topic Segmentation". In: *Computational Linguistics* 27.1 (2001), pp. 31–57. DOI: 10.1162/089120101300346796. eprint: https://doi.org/10.1162/089120101300346796. URL: https://doi.org/10.1162/089120101300346796.

[67] Svitlana Vakulenko et al. "Measuring Semantic Coherence of a Conversation". In: *The Semantic Web – ISWC 2018*. Ed. by Denny Vrandečić et al. Cham: Springer International Publishing, 2018, pp. 634–651. ISBN: 978-3-030-00671-6.

[68] Teun A Van Dijk. "The semantics and pragmatics of functional coherence in discourse". In: *Journal of Pragmatics* 4 (1980), pp. 233–252.

[69] Xinhao Wang, Keelan Evanini, and Klaus Zechner. "Coherence Modeling for the Automated Assessment of Spontaneous Spoken Responses". In: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Atlanta, Georgia: Association for Computational Linguistics, June 2013, pp. 814–819. URL: https://www.aclweb.org/anthology/N13-1101.

[70] Xinhao Wang et al. "Modeling Discourse Coherence for the Automated Scoring of Spontaneous Spoken Responses". In: *Proc. 7th ISCA Workshop on Speech and Language Technology in Education*. 2017, pp. 132–137. DOI: 10.21437/SLaTE.2017-23. URL: http://dx.doi.org/10.21437/SLaTE.2017-23.

[71] Xinhao Wang et al. "Using Rhetorical Structure Theory to Assess Discourse Coherence for Non-native Spontaneous Speech". In: *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*. Minneapolis, MN: Association for Computational Linguistics, June 2019, pp. 153–162. DOI: 10.18653/v1/W19-2719. URL: https://www.aclweb.org/anthology/W19-2719.

[72] Yuan Wang and Minghe Guo. "A short analysis of discourse coherence". In: *Journal of Language Teaching and Research* 5.2 (2014), p. 460.

[73] Bonnie Webber. "D-LTAG: extending lexicalized TAG to discourse". In: *Cognitive Science* 28.5 (2004). 2003 Rumelhart Prize Special Issue Honoring Aravind K. Joshi, pp. 751 –779. ISSN: 0364-0213. DOI: https://doi.org/10.1016/j.cogsci.2004.04.002. URL: http://www.sciencedirect.com/science/article/pii/S0364021304000655.

[74] Henry George Widdowson. *Teaching language as communication*. Oxford University Press, 1978.

[75] Peng Xu et al. "A Cross-Domain Transferable Neural Coherence Model". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 678–687. DOI: 10.18653/v1/P19-1067. URL: https://www.aclweb.org/anthology/P19-1067.

[76] Duanli Yan, André A Rupp, and Peter W Foltz. *Handbook of automated scoring: Theory into practice*. CRC Press, 2020.

[77] Hu Zhuanglin. "Discourse Cohesion and Coherence". In: *Shanghai Foreign Language Education Press* (1994).