Word Importance Modeling to Enhance Captions Generated by Automatic Speech Recognition for Deaf and Hard of Hearing Users

Sushant Kafle
sxk5664@rit.edu

Follow this and additional works at: https://scholarworks.rit.edu/theses

Recommended Citation
Word Importance Modeling to Enhance Captions Generated by Automatic Speech Recognition for Deaf and Hard of Hearing Users

by

Sushant Kafle

A dissertation submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in Computing and Information Sciences

B. Thomas Golisano College of Computing and Information Sciences

Rochester Institute of Technology

November, 2019
Word Importance Modeling to Enhance Captions Generated by Automatic Speech Recognition for Deaf and Hard of Hearing Users

by

Sushant Kafle

Committee Approval:
We, the undersigned committee members, certify that we have advised and/or supervised the candidate on the work described in the dissertation. We further certify that we have reviewed the dissertation manuscript and approve it in partial fulfillment of the requirements of the degree of Doctor of Philosophy in Computing and Information Sciences.

Dr. Matt Huenerfauth, Dissertation Advisor Date

Dr. Cecilia Ovesdotter Alm, Dissertation Committee Member Date

Dr. Vicki Hanson, Dissertation Committee Member Date

Dr. Emily Prud'hommeaux, Dissertation Committee Member Date

Dr. Jai Kang, Dissertation Chair Date

Certified by:

Dr. Pengcheng Shi, Director, Computing and Information Sciences Date
Word Importance Modeling to Enhance Captions Generated by Automatic Speech Recognition for Deaf and Hard of Hearing Users

by

Sushant Kafle

Submitted to the
B. Thomas Golisano College of Computing and Information Sciences
Ph.D. Program in Computing and Information Sciences
in partial fulfillment of the requirements for the
Doctor of Philosophy Degree
at the Rochester Institute of Technology

Abstract

People who are deaf or hard-of-hearing (DHH) benefit from sign-language interpreting or live-captioning (with a human transcriptionist), to access spoken information. However, such services are not legally required, affordable, nor available in many settings, e.g., impromptu small-group meetings in the workplace or online video content that has not been professionally captioned. As Automatic Speech Recognition (ASR) systems improve in accuracy and speed, it is natural to investigate the use of these systems to assist DHH users in a variety of tasks. But, ASR systems are still not perfect, especially in realistic conversational settings, leading to the issue of trust and acceptance of these systems from the DHH community. To overcome these challenges, our work focuses on: (1) building metrics for accurately evaluating the quality of
automatic captioning systems, and (2) designing interventions for improving the usability of captions for DHH users.

The first part of this dissertation describes our research on methods for identifying words that are important for understanding the meaning of a conversational turn within transcripts of spoken dialogue. Such knowledge about the relative importance of words in spoken messages can be used in evaluating ASR systems (in part 2 of this dissertation) or creating new applications for DHH users of captioned video (in part 3 of this dissertation). We found that models which consider both the acoustic properties of spoken words as well as text-based features (e.g., pre-trained word embeddings) are more effective at predicting the semantic importance of a word than models that utilize only one of these types of features.

The second part of this dissertation describes studies to understand DHH users' perception of the quality of ASR-generated captions; the goal of this work was to validate the design of automatic metrics for evaluating captions in real-time applications for these users. Such a metric could facilitate comparison of various ASR systems, for determining the suitability of specific ASR systems for supporting communication for DHH users. We designed experimental studies to elicit feedback on the quality of captions from DHH users, and we developed and evaluated automatic metrics for predicting the usability of automatically generated captions for these users. We found that metrics that consider the importance of each word in a text are more effective at predicting the usability of imperfect text captions than the traditional Word Error Rate (WER) metric.

The final part of this dissertation describes research on importance-based
highlighting of words in captions, as a way to enhance the usability of captions for DHH users. Similar to highlighting in static texts (e.g., textbooks or electronic documents), highlighting in captions involves changing the appearance of some texts in caption to enable readers to attend to the most important bits of information quickly. Despite the known benefits of highlighting in static texts, research on the usefulness of highlighting in captions for DHH users is largely unexplored. For this reason, we conducted experimental studies with DHH participants to understand the benefits of importance-based highlighting in captions, and their preference on different design configurations for highlighting in captions. We found that DHH users subjectively preferred highlighting in captions, and they reported higher readability and understandability scores and lower task-load scores when viewing videos with captions containing highlighting compared to the videos without highlighting. Further, in partial contrast to recommendations in prior research on highlighting in static texts (which had not been based on experimental studies with DHH users), we found that DHH participants preferred boldface, word-level, non-repeating highlighting in captions.
Acknowledgments

I would like to express my sincere gratitude to my advisor Dr. Matt Huenerfauth for his continuous support during my Ph.D. studies and research. It was a great pleasure to work under his advisement and to learn from him. His guidance has helped me not only in the time of research and writing of this thesis but also at times when I doubted myself and my intuitions. I could not have imagined having a better advisor and a mentor.

I would also like to thank my wonderful thesis committee members: Drs. Cecilia Alm, Vicki Hanson, and Emily Prud’hommeaux, for their encouragement and insightful comments. I will forever be indebted to their stimulating discussions and hard questions that have helped shape this research tremendously.

I thank my colleagues and research assistants in the Center for Accessibility and Inclusion Research (CAIR) lab for their collaboration and support. Special thanks to Larwan Berke, a fellow researcher at the CAIR lab, for invaluable discussions and collaborations without which this research would not have been possible.

Last but not least, I would like to thank my family, especially my parents Dev Raj Kafle and Kabita Kafle, for their endless love and support, and my girlfriend Swapnil Sneham for believing in me, always.
Contents

List of Figures .................................................. xvii
List of Tables ................................................... xxv

1 Introduction ................................................. 1
  1.1 Motivating Challenges ................................. 3
  1.2 Research Questions Investigated in this Dissertation .... 6
  1.3 Overview of The Chapters ............................ 8

2 Background on Automatic Speech Recognition Technology 10
  2.1 Conventional Speech Recognition Architecture ....... 10
    2.1.1 Acoustic Models ................................... 11
    2.1.2 Language Models .................................. 12
    2.1.3 Decoding ......................................... 12
  2.2 Recent Advancements: End-to-End ASR ................. 13
  2.3 Other Terminology ..................................... 15
    2.3.1 Confidence Scores ............................... 15
    2.3.2 Word Error Rate ................................. 16
## 6 Supervised Models of Word Importance

6.1 Text-based Model of Word Importance .......................... 50
   6.1.1 Model Architecture .................................... 51
   6.1.2 Experimental Setup .................................... 53
   6.1.3 Experiment 1: Performance of the Models ............... 53
   6.1.4 Experiment 2: Comparison with Human Annotators .... 54
   6.1.5 Limitations of this Research ............................. 55

6.2 Speech-based Importance Model ................................. 56
   6.2.1 Model Architecture .................................... 57
   6.2.2 Acoustic-Prosodic Feature Representation ............... 60
   6.2.3 Experimental Setup .................................... 63
   6.2.4 Experiment 1: Comparison of the Projection Layers ..... 64
   6.2.5 Experiment 2: Ablation Study on Speech Features ..... 65
   6.2.6 Experiment 3: Comparison with the Text-based Models 66
   6.2.7 Limitations of this Research ............................. 68

6.3 Text- and Speech-based Importance Model ..................... 68
   6.3.1 Prior Work on Joint Modeling of Speech and Text ...... 69
   6.3.2 Lexical-Prosodic Feature Representation ............... 71
   6.3.3 Experimental Setup .................................... 74
   6.3.4 Experiment 1: Error Analysis of Unimodal Models ..... 78
   6.3.5 Experiment 2: Comparison of Fusion Strategies ....... 79

6.4 Conclusions .................................................. 81

Epilogue for Part I ................................................. 83
CONTENTS

PART II: AUTOMATIC CAPTION QUALITY EVALUATION 86

PROLOGUE TO PART II 87

7 Prior Approaches to ASR Evaluation 90

7.1 Limitations of the Word Error Rate Metric .......................... 90
7.2 Other Methods of ASR Evaluation ........................................ 91
7.3 Metric of ASR Quality for DHH users ................................. 94

8 Collection of Understandability Scores from DHH users for
Text with Errors 97

8.1 Understanding the Effect of Recognition Errors ...................... 97
8.2 User Study (QUESTION-ANSWER STUDY) ............................ 98
  8.2.1 ASR Error Category ............................................ 98
  8.2.2 Study Resources ................................................ 100
  8.2.3 Recruitment and Participants .................................... 103
  8.2.4 Study Procedure ................................................. 103
8.3 Summary of the data .................................................... 104

9 Metric for ASR Evaluation for Captioning Applications 105

9.1 Automatic-Caption Evaluation Framework ............................... 105
  9.1.1 Word Importance Sub-score .................................... 107
  9.1.2 Semantic Distance Sub-score .................................... 107
  9.1.3 The Weighting Variable ......................................... 108
9.2 Research Methodology and Hypotheses ................................ 109
  9.2.1 Four Phases of this Research .................................... 109
9.3 Phase 1: Designing and Evaluating the ACE Metric ....... 112
  9.3.1 Computing the Word Importance Sub-score .......... 113
  9.3.2 Computing the Semantic Distance Sub-score ........ 114
  9.3.3 From Individual-Error Impact Scores to an Overall Sentence Error Score .............. 115
  9.3.4 Designing Stimuli for Metric Evaluation (PREFERENCE-2017 Study) ............... 118
  9.3.5 Experimental Study Setup and Procedure ............. 122
  9.3.6 Results and Discussion ............................ 124
  9.3.7 Summary and Discussion of Limitations of ACE .... 126
9.4 Phase 2: Improving the ACE Metric to Create ACE2 ....... 127
  9.4.1 Improving the Word Importance Sub-score .......... 128
  9.4.2 Alternatives for Combining Individual Error Scores into a Sentence Score .......... 132
9.5 Phase 3: Comparison with Prior Metrics ................. 139
  9.5.1 Human Perceived Accuracy (HPA) .................... 140
  9.5.2 Information Retrieval Based Evaluation Metrics .. 140
  9.5.3 Word Information Lost (WIL) ......................... 141
  9.5.4 Weighted Word Error Rate (WWER) ................. 141
  9.5.5 Weighted Keyword Error Rate (WKER) and Keyword Error Rate (KER) .......... 141
9.6 Phase 4: User-Based Evaluation of ACE and ACE2 (PREFERENCE-2018 Study) ............. 142
  9.6.1 Designing Stimuli .................................. 143
CONTENTS

9.6.2 User Study Setup ............................................. 145
9.6.3 Results and Discussion ...................................... 146
9.7 Conclusions .......................................................... 148

Epilogue for Part II ................................................... 152

Part III: Enhancements to Improve Caption Usability ............. 155

Prologue to Part III ....................................................... 156

10 Prior Work on Caption Accessibility .............................. 159
  10.1 Caption Accessibility Challenges .............................. 159
  10.2 Improving Caption Accessibility .............................. 162
  10.3 Importance-based Highlighting in Text ....................... 163
    10.3.1 Style Guidelines for Highlighting ....................... 164
    10.3.2 Visual Markup of Text in Captions ...................... 167

11 Evaluating the Benefits of Highlighting in Captions ........... 168
  11.1 Background and Introduction ................................. 168
    11.1.1 Research Questions Investigated in this Chapter ...... 171
  11.2 Formative Studies: Method and Results ........................ 172
    11.2.1 Highlighting Configurations for Formative Studies ...... 173
    11.2.2 Stimuli Preparation for Formative Studies .............. 174
    11.2.3 Recruitment and Participants for Formative Studies ... 175
    11.2.4 Questionnaires for Smaller Studies ..................... 175
    11.2.5 Round-1 Results: Comparing Markup-Styles ............. 178
11.2.6 Round-2 Results: Comparing Highlight Percentage . . . 178
11.2.7 Round-1 and Round-2 Results: Interest in Highlighting . 181
11.2.8 Discussion of Results from Round-1 and Round-2 . . . . 182
11.3 Larger Study: Method and Results . . . . . . . . . . . . . . . . 183
11.3.1 Preparation of the Stimuli Video . . . . . . . . . . . . . 183
11.3.2 Study Setup and Questionnaires . . . . . . . . . . . . . 185
11.3.3 Recruitment and Participants . . . . . . . . . . . . . . . 188
11.3.4 Results . . . . . . . . . . . . . . . . . . . . . . . . . . . 188
11.4 Discussion and Conclusion . . . . . . . . . . . . . . . . . . . . 193
11.5 Limitations of this Research and the Need for an Additional Study196

12 Evaluating the Designs for Highlighting Captions 197
12.1 Background and Introduction . . . . . . . . . . . . . . . . . . 197
12.1.1 Harmful Effects of Inappropriate Highlighting . . . . . 199
12.1.2 Research Questions Investigated in this Chapter . . . . 200
12.2 Methodology . . . . . . . . . . . . . . . . . . . . . . . . . . . . 202
12.2.1 Four Phases in the Study . . . . . . . . . . . . . . . . . 202
12.2.2 Details of Video Stimuli Creation for Each Condition . . 207
12.2.3 Questions Asked in the Study . . . . . . . . . . . . . . . . 208
12.2.4 Recruitment and Participants . . . . . . . . . . . . . . . 210
12.3 Results . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 211
12.3.1 Text Decoration Style for Highlighting . . . . . . . . . . 211
12.3.2 Granularity for Highlighting . . . . . . . . . . . . . . . . 214
12.3.3 Handling Key Term Repetition . . . . . . . . . . . . . . . 216
CONTENTS

12.3.4 Interest in Highlighting Applications ......................... 217
12.4 Discussion of the Results ........................................ 218
12.5 Conclusions ....................................................... 221

Epilogue for Part III .................................................. 222

13 Limitations and Future Work ...................................... 226
13.1 Word Importance Modeling ....................................... 226
13.1.1 Modeling Importance at a Larger Semantic Units ........... 227
13.1.2 Unsupervised (and Semi-supervised) Models of Word
Importance ................................................................. 227
13.2 Automatic Caption Quality Evaluation .......................... 228
13.3 Highlighting in Captions to Improve Caption Usability ...... 230
13.4 Using Word-Importance Models during the Training or Decoding
of ASR Systems .......................................................... 232
13.4.1 N-best list Re-scoring Technique ............................... 234
13.4.2 Improved Optimization Strategy (End-to-End Models) .... 235

14 Summary and Contributions ........................................ 237
14.1 Summary of the Contribution of This Research .................. 238
14.2 Final Comments ...................................................... 243

Bibliography .............................................................. 245

Appendices ................................................................. 279

A Publications ............................................................ 280
List of Figures

1.1 A deaf student collaborating with two other hearing students using automatic speech recognition technology installed in their mobile devices during our exploratory study of the usefulness of such a service. .......................... 3

1.2 Research focus of this thesis. .......................... 5

4.1 Figure showing how the language model is used to make inference about the predictability of a word given its context in an example sentence. Reader can refer to Section 4.2 for mathematical detail on how this score is computed. .......................... 32

4.2 Diagram of neural word predictability model demonstrating how the context of a word \( w(i) \) is captured using bi-directional recurrent units. .......................... 37
5.1 Visualization of importance scores assigned to words in a sentence by a human annotator on our project, with the height and font-size of words indicating their importance score (and redundant color coding: green for high-importance words with score above 0.6, blue for words with score between 0.3 and 0.6, and gray otherwise).

6.1 General unfolded network structure of our model, adapted from Lample et al. [90]. The bottom layer represents word-embedding inputs, passed to bi-directional LSTM layers above. Each LSTM takes as input the hidden state from the previous time step and word embeddings from the current step, and outputs a new hidden state. $C_i$ concatenates hidden representations from LSTMs ($L_i$ and $R_i$) to represent the word at time $i$ in its context.

6.2 Confusion matrices for each model for classification into 6 classes: $c_1 = [0, 0.1)$, $c_2 = [0.1, 0.3)$, and so forth.

6.3 Example of conversational transcribed text, right where you move from, that is difficult to disambiguate without prosody. The intended sentence structure was: Right! Where you move from?
LIST OF FIGURES

6.4 Architecture for feature representation of spoken words using time series speech data. For each spoken word \( w \) identified by a word-level timestamp, a fixed-length interval window \( \tau \) slides through to get \( n = \frac{\text{time}(w)}{\tau} \) sub-word interval segments. Using an RNN network, a word-level feature \( s \), represented by a fixed-length vector, is extracted using the features from a variable-length sub-word sequence. 58

9.1 Graphical illustration of research activities presented in this article. 110
9.2 Visual illustration of n-gram-based word-importance scoring, based on the predictability of words in the context of a sentence, with higher bars indicating less predictable words. 113
9.3 Visual Illustration of word2vec based semantic distance scoring of different alignment pairs (reference word → hypothesis word) in an example sentence. The height of the black bar indicates the semantic distance between the words. 115
9.4 Preparation of a fake meeting transcript. 119
9.5 Time based alignment of reference (R) and hypothesized (H) text. The grouping with red dotted arrowhead lines indicates individualized errors aligned with corresponding reference text based on word level timestamps. 121
LIST OF FIGURES

9.6 Screenshot from the study, with side-by-side comparison of caption-text automatically generated by ASR. Each pair of texts (left and right) have identical WER scores, but one text in each pair was preferred by our ACE metric. 123

9.7 Average usability rating variation among participants. 124

9.8 Analysis of the ACE metric with participant’s usability rating. 125

9.9 Visual illustration of Neural-based word-importance scoring, based on the predictability of words in the context of a sentence, with higher bars indicating less predictable words. 129

9.10 Visual illustration of TF-IDF based word-importance scoring, based on the predictability of words in the context of a sentence, with higher bars indicating less predictable words. 130

9.11 An example of error impact scoring in a sentence, with “c” indicating a correct word was recognized, and “i” indicating an incorrect word. 133

9.12 Various position-based weighting functions we considered. 135

9.13 This figure corresponds to the example sentence in Fig. 9.11, and it displays a plot of impact scores for each error (black bars) and the region of impact due to error (overlaid grey region) represented using the error-spread model. 136

9.14 Screenshot from the study to measure usability of caption-text automatically generated by ASR. 146

9.15 Analysis of the ACE metric with participant’s usability rating. 147
LIST OF FIGURES

xxi

11.1 A typical arrangement of elements in an online educational video:
with instructor, slides, and captions [62] . . . . . . . . . . . . . 170
11.2 Round-1 Formative Study: Comparison of different visual markupstyles for highlighting in captions on easy to follow question. . . 176
11.3 Round-1 Formative Study: Comparison of different visual markupstyles for highlighting in captions on distracting question. . . . 177
11.4 Round-1 Formative Study: Comparison of different visual markupstyles for highlighting in captions on easy to read question. . . . 177
11.5 Round-2 Formative Study: Comparison of the percentage of
words highlighted in captions on easy to follow question. . . . . 179
11.6 Round-2 Formative Study: Comparison of the percentage of
words highlighted in captions on distracting question. . . . . . . 179
11.7 Round-2 Formative Study: Comparison of the percentage of
words highlighted in captions on easy to read question. . . . . . 180
11.8 Percentage distribution of participants’ responses on the ease of
following the content of the video and the caption. . . . . . . . 189
11.9 Percentage distribution of participants’ responses on the readability of the caption.

. . . . . . . . . . . . . . . . . . . . . . . 189

11.10Percentage distribution of participants’ responses on being able
to identify the important words and concepts. . . . . . . . . . . 190
11.11Percentage distribution of participants’ responses on the understandability of the content of the content of the video and the
captions. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 191


11.12 Percentage distribution of participants’ responses on the mental demand when reading and understanding the captions in the video. ............................................. 191
11.13 Percentage distribution of participants’ responses on the temporal demand of reading and understanding the captions in the video. ............................................. 192
11.14 Percentage distribution of participants’ responses on the difficulty of reading and understanding the captions in the video. ........... 193

12.1 Samples of video stimuli in Phase 1 of the study with videos containing different text decoration styles for highlighting. ... 204
12.2 Samples of video stimuli in Phase 2 of the study with videos containing different levels of granularity for highlighting. .... 205
12.3 Samples of video stimuli in Phase 2 of the study with videos containing different strategies for handling repeated keyword when highlighting. .................................................. 206
12.4 Participants’ responses to Phase 1 of the study, comparing decoration styles (italics, boldface and underline) for question *I was able to identify important words and concepts*, with significant differences marked with asterisks. ...................... 212
12.5 Participants’ responses to Phase 1 of the study, comparing decoration styles (italics, boldface and underline) for question *I found the captions distracting*, with significant differences marked with asterisks. ...................... 213
12.6 Participants’ responses to Phase 1 of the study, comparing decoration styles (italics, boldface and underline) for question *It was easy to read the caption*, with significant differences marked with asterisks. ........................................... 213

12.7 Participants’ responses to Phase 2 of the study, which compared the granularity of highlighting (at the sentence-level or the word-level), for question *I was able to identify importance words and concepts*, with significant differences marked with asterisks. ........ 215

12.8 Participants’ responses to Phase 2 of the study, which compared the granularity of highlighting (at the sentence-level or the word-level), for question *I found the captions distracting*, with significant differences marked with asterisks. ............. 215

12.9 Participants’ responses to Phase 3 of the study for whether repeated keywords should be highlighted only once (once) or every important occurrence (always), for question *I found the captions distracting*, with significant differences marked with asterisks. ................................. 216

12.10 Responses in Phase 4, on whether participants thought that highlighting important words in captions would be useful for six different applications, with significant pairwise differences marked with asterisks. ................................. 217

B.1 IRB Decision Form for “Creating the Next Generation of Live-Captioning Technologies”. ................................. 285
B.2 IRB Decision Form for “Identifying the Best Methods for Displaying Word-Confidence in Automatically Generated Captions for Deaf and Hard-of-Hearing Users”. . . . . . . . . . . . . . . . . . . . . . . . . . . . 286
List of Tables

5.1 Guidance for the annotators to promote consistency and uniformity in the use of numerical scores. ................................. 46

6.1 Model performance in terms of RMS deviation and macro-averaged $F_1$ score, with best results in **bold** font. ................... 54

6.2 Performance of the speech-based models on the test data under different projection layers. Best performing scores highlighted in **bold**. ................................................................. 64

6.3 Speech feature ablation study. The minus sign indicates the feature group removed from the model during training. Markers (* and †) indicate the biggest and the second-biggest change in model performance for each metric, respectively. .................... 65

6.4 Comparison of our speech-based model with a prior text-based model, under different word error rate conditions. .............. 67
6.5 Comparative performance of lexical and prosodic unimodal models. RMS column represents the overall RMS score, whereas RMS (oov words only) represents the RMS deviation of the prediction on oov words only. Bold font shows the best scores. 78

6.6 Comparison of different models combining lexical and prosodic cues. Per column, the top two results are marked with ⋆ & † symbols, respectively. Our proposed model demonstrates lower RMS error both overall as well as for OOVs specifically. 79

6.7 Comparison of models on ordinal-range classes, and Kendall-tau (τ-b) rank-prediction correlation. The top two results per column are marked with ⋆ & † symbols. Our proposed model performs better for high and low importance words. 80

9.1 Comparison of error-impact prediction models (based on three different word-importance models) for predicting the comprehensibility of error-containing texts for DHH users. Note: The ngram_LM model corresponds to our original ACE metric from Phase 1. 131

9.2 Comparison of different methods for calculating a sentence score, based on individual error scores contained within the text; each model below utilizes the Neural_LM-based error impact model (for calculating individual error scores) discussed in Section 9.4.1.138

10.1 Design aspects in prior text highlighting research, e.g. granularity (full sentences or individual words highlighted) 165
LIST OF TABLES

11.1 Results of preliminary studies used in final study. . . . . . . . . 183
11.2 List of questions used in the final study . . . . . . . . . . . . . . . 187
12.1 Sources of video content for stimuli in each phase. . . . . . . . . . 203
Chapter 1

Introduction

People who are Deaf and Hard-of-Hearing (DHH) make use of a wide variety of communication technologies to access spoken information, including services like captioning (e.g., offline captioning for pre-recorded television programming or real-time captioning services in classrooms, meetings, and live events) or sign language interpreting. In particular, captioning technology produces a digital textual output, which can be easily processed, transmitted, or stored as a transcript. Such captions are useful in various scenarios, such as classrooms or meetings, where these captions may be viewed in real-time or transcripts can be reviewed later.

While trained service providers, either in the form of professional captioning or sign language interpreting, are most often used for making real-time aural information accessible to DHH individuals, these services are not legally required, affordable, nor available in many settings, for e.g, impromptu communication such as small-group meetings or extremely brief conversational interactions.
As discussed in Chapter 7 and 10, prior work has established that insufficient communication access can be detrimental to DHH individuals’ professional or academic success. Therefore, technology for automatically transcribing real-time spoken message to textual output is beneficial for these users.

Advances in Automatic Speech Recognition (ASR) technology, which uses machine-learning techniques to automatically transcribe speech to text, holds exciting promise for providing such a service. However, current state-of-the-art ASR technology is still imperfect, especially in realistic conversational settings: In the complex audio environment of multiparty meetings, ASR systems have been shown to produce low-quality output, and these errors can be harmful for DHH users when their success in the workplace or educational settings depends on full and accurate communication. Prior research on fully automated real-time captioning using ASR in settings such as classrooms [79] or in simulated live meetings [8] has revealed that DHH users are interested in the promise of ASR supporting their conversations, but when users actually try such systems, they are very concerned about low accuracy.

With ASR systems growing in popularity, there is a risk that a cost-savings motivation could encourage automatic captioning to be deployed before the output of such technology is of acceptable quality and accuracy. Surveys of DHH users have revealed their fears that current services (e.g. ASL interpreting) could be replaced by lower quality automated systems [137]. Therefore, there is an ethical imperative on researchers to evaluate and enhance the usability of such systems for these users before their deployment.
1.1 Motivating Challenges

With the advent of cloud-enabled services, ASR systems today are cheap, scalable and highly available, which makes them promising for real-time captioning applications for DHH users. Today, we can easily envision such a system being installed on mobile phones or tablets and being used on-demand for transcribing spoken messages to digital texts.

Figure 1.1: A deaf student collaborating with two other hearing students using automatic speech recognition technology installed in their mobile devices during our exploratory study of the usefulness of such a service.

Fig. 1.1 shows how ASR system installed on mobile devices could be used to enable participation of DHH users in mainstream meetings with their hearing peers.

Despite the recent leaps in the accuracy of ASR systems, the performance
of these systems are generally not on par with humans, who currently provide most caption text for DHH users. Hence, these systems need to be properly designed and evaluated in order for them to be trusted and accepted by DHH users for real-time captioning applications. However, despite the enormous potential of ASR-based captioning, research into these issues are still largely unexplored.

With this motivation, this dissertation addresses some of the challenges in evaluating and improving the usability of ASR technology for supporting communication between DHH users with their hearing peers. This research began by exploring methods for identifying which words in spoken messages were most important for understanding its meaning, and this word-importance model is used as a building block for research activities in later phases of this dissertation. Identifying semantic importance of words in a spoken message allows us to accurately investigate the understandability of automatically generated captions, thus informing our research into the issues of usability of these automatic systems for captioning applications.

Specifically, we investigate two main challenges discussed below (also illustrated by two rectangles in Fig. 1.2):

- **Automatic Caption Quality Evaluation Challenges**: Commonly used metrics for evaluating ASR system performance are very simplistic, i.e. based on simply counting the number of errors without considering whether the errors occur on important words. Prior research (not with DHH users nor in a captioning context) had found that these metrics were not well correlated with performance of humans of tasks that depend
upon the ASR output. Thus, there was a need for research to determine whether simplistic metrics correlated with the judgments of DHH users about the quality of captions based on ASR, and understand if there is a potential need for better metrics of ASR performance that correlates better with actual DHH users’ perception of caption quality. (This is addressed in Part II of this thesis.)

- **User-Experience Challenges:** ASR-output text containing errors can be more difficult to understand, as compared to transcripts produced by humans. For instance, even if both are imperfect, prior work has found that the errors produced by human transcriptionists are less confusing than the errors produced by ASR [89]. Consequently, to enhance the user-experience of ASR systems as a captioning tool for DHH users, it was necessary to investigate how to enhance the usability of caption-text output, even in the presence of errors. Authors of textbooks have tra-
ditionally used highlighting as a method to draw readers’ attention to important segments of a text. Prior research has found that such highlighting enhances the reading experience, and in an educational context, highlighting has been found to enable faster browsing and recall of information by students. However, the use of importance-based highlighting in the captions of videos has been largely unexplored, and highlighting words in such text may require special consideration: Unlike books or documents, captions are dynamic (with the speed determined by the live speaker or the video playing), with shorter text segments, which are usually shown with only 1 or 2 lines at a time, with each appearing for 2 to 4 seconds [89]. Moreover, users are known to be sensitive to caption display parameters such as speed, font size, or decorations: Several researchers have measured the influence of such visual parameters of caption appearance on the readability of captions for DHH users [12, 89, 165]. (This is investigated in Part III of this thesis)

In the coming sections, we discuss how we use the information about the importance of words in a text to design solutions for tackling these challenges.

1.2 Research Questions Investigated in this Dissertation

In this work, we conduct research to understand the challenges of ASR-based captioning technologies for producing more usable captions for users who are Deaf or Hard of Hearing (DHH) and, provide methodological solutions to these
challenges, validated through studies with the users. More specifically, our work address the set of research questions listed below:

**RQ1:** How can we identify words in a spoken message that are important to its understandability for DHH readers? The task of predicting the importance of words in a spoken message for understanding serves an important purpose in this thesis: Through our preliminary studies, we identified that answering this research question might help us investigate the issues of usability of ASR-based captioning technologies such as evaluation of ASR system quality (addressed in RQ2) and usability enhancement of captioning through importance-based highlighting in captions (addressed in RQ3). As we will discuss in Section 3.1.3, existing methods for identifying important words (for the understandability of a text) have some inherent challenges when focusing on a more conversational style of texts. With this motivation, Part I of this research investigates this question in detail.

**RQ2:** Do our models of estimating the quality of ASR systems for generating captioning for DHH users accurately predict the quality of the output? Current methods of evaluation of ASR system quality, such as the Word Error Rate metric, have been shown to be inefficient in predicting actual human task performance using these systems in various applications (as discussed in Section 7.1). Therefore, there is a need for a way to measure the quality of output of an ASR system to determine whether it is accurate enough to be used to produce captions automatically for DHH users. We are
also interested in exploring other metrics that are more informed about the impact of various errors on the understandability of texts (discussed in Part I of our work), for evaluating the quality of automatic captions for DHH users. This will be discussed in Part II of this research.

RQ3: Are DHH users receptive to the premise of importance-based word highlighting in captions, and what are their highlighting preferences in captions? As discussed in section 10.1, viewing captions during a video can be challenging due to the need to split visual attention between the text and other sources of information in the video. For this reason, some form of emphasis of which words are essential for the meaning of a text might be useful to visually convey to users. Hence, Chapter 11 and 12 discusses importance-based highlighting in captions, especially in the context of educational lecture videos for DHH viewers. Specifically, in Chapter 11, we study the benefits of highlighting in captions for DHH individuals when viewing online lecture videos. Further, in Chapter 12, we investigate DHH users’ preferences on different design choices for highlighting in captions through experimental studies with these users. This will be discussed later in Part III of this research.

1.3 Overview of The Chapters

To provide readers with essential background knowledge, Chapter 2 quickly introduces Automatic Speech Recognition (ASR) technology, their architecture and other important concepts that might be useful for discussion later in this work.
In Part I, we begin by discussing the prior work on estimating importance of words in texts in Chapter 3, for various applications. In the subsequent chapters, we present our investigation into the task of word importance prediction in spoken dialogues. Specifically, Chapter 4 presents our initial method for estimating the importance of words in a text based on its predictability in the text. This work was inspired by previous eye-tracking research on the reading strategies of DHH readers. Next, Chapter 5 and 6 discusses other supervised models of word importance based on human-labelled data of word importance.

Part II of our work begins by exploring current practices in evaluating the quality of ASR systems for various applications, which is discussed in Chapter 7. Chapter 8 discusses our methods for understanding the effect of various recognition errors in the understandability of text for DHH readers. Later, Chapter 9 draws upon these results to design and evaluate various automatic metrics for measuring ASR performance in real-time captioning application for DHH users.

Lastly, in Part III, our work examines strategies to improve the usability of captioning systems by focusing on enhancing the user-experience surrounding the use of these systems. For this, we focus on importance-based highlighting in captions with a goal to improve the readability of the captions and reduce their reading times. Chapter 11 discusses our work on evaluating the benefits of highlighting key words in captions for DHH users, especially when viewing educational lecture-type videos. As a follow-up to this study, Chapter 12 studies DHH users' preference on the different design choices for highlighting in captions.
Chapter 2

Background on Automatic Speech Recognition Technology

The task of an Automatic Speech Recognition (ASR) system is to transcribe aural information to visual text. This chapter aims to provide a quick overview of the working of an ASR system, and some related terminology that might be a useful background information for this document.

2.1 Conventional Speech Recognition Architecture

One of the most popularly used models for speech recognition is a type of generative statistical model based on a source-channel architecture, where the source i.e., the sequence of words in speaker’s mind ($W$), is passed through a noisy communication channel (consisting of the speaker’s vocal apparatus) that produces the speech waveform ($X$), which we are interested to process.
with the help of our *speech processing engine*. The goal of this engine is to decode the speech waveform back to text ($\hat{W}$). A typical speech-recognition system consists of three main components: **Acoustic models** represents the knowledge about the speech, **Language models** represents knowledge about the language of the speech and **Decoder** makes use of these models to decode the speech to text, as:

$$\hat{W} = \arg\max_W P(W|X) = \arg\max_W P(W)P(X|W)$$

(2.1)

where the goal of the decoder is to search the optimal word sequence $\hat{W} = w_1, w_2, ..., w_n$ that has the maximum posterior probability $P(W|X)$. $P(W)$ and $P(X|W)$ represent the probabilities computed by the language modeling and the acoustic modeling components, respectively. The remainder of this section will provide brief discussion on each of these components, and how they are realized in practice.

### 2.1.1 Acoustic Models

Acoustic models are often central to speech recognition systems, responsible for representing the knowledge about the statistical properties in speech. More accurately, it represents the likelihood of the model generating the observed speech waveform ($X$) given the linguistic units. Traditionally, a Hidden Markov Model (HMM), which is a finite state machine, is used to make probabilistic inferences about their temporal structure. A HMM is often used along side a Gaussian mixture model (GMM) that is used to compute the observation prob-
abilities from the input feature vectors of speech. More recently, several Deep Neural Network (DNN) based acoustic models have been proposed, which include hybrid-HMMs that use deep neural network to approximate the likelihood probability $P(X|W)$ [66,108], to fully DNN (particularly Recurrent Neural Networks) based acoustic models which directly model sequential acoustic signals to generate posterior probabilities of the acoustic states [153,175].

### 2.1.2 Language Models

The task of language models in speech recognition is to compute the probabilistic parameter $P(W)$ in Eq. 2.1, which refers to the probability that a given string of words $(W = w_1, w_2, ..., w_n)$ belongs to a language.

A common way to represent a language model is through a $n$-gram model, which is based on estimates of word string probabilities from large collections of text. In order to make these estimates tractable, the probability of a word given the preceding sequence is approximated to the probability given the preceding one (bigram) or two (trigram) words or three (fourgram) and so on – thus, these models are commonly referred to as $n$-gram models. While $n$-gram based language models have been dominant in the past, recently Recurrent Neural Network (RNN) based language models have been popular [114,147].

### 2.1.3 Decoding

The final step in speech recognition, as shown in Eq. 2.1, is the decoding process which involves using the acoustic and language model components to best match the input speech features to a sequence of words. Since the acoustic
states in acoustic models (e.g., HMMs) are often represented by phones, an additional lexicon, called the pronunciation lexicon is utilized to maps the sequence of phones to words. Generally, a search algorithm such as Viterbi decoding is used which looks for a sequence of words $\hat{W} = w_1, w_2...w_n$ that has the best maximum posterior probability $P(W|X)$ for the given input speech waveform $(X = x_1, x_2,...,x_t)$.

2.2 Recent Advancements: End-to-End ASR

Recently, there has been much interest in end-to-end speech recognition which rather than relying on separate components and training for the acoustic and language models, learns all the components of the system together. This model is also referred to as a discriminative model (in contrast to the traditional generative model) which consists of a single end-to-end trained sequence-to-sequence model. This model can directly output words or graphemes from raw audio, which greatly simplifies the speech recognition pipeline.

The development of end-to-end ASR systems started with Connectionist Temporal Classification (CTC) proposed by Graves et al. [57] which provides a way to train an acoustic models without requiring frame-level alignments between the acoustics and the transcripts. Specifically, CTC introduces a special symbol $\text{blank}$ to encode a sequential input, and works by maximizing the total probability of the label sequence by marginalizing over all possible alignments identified by the encoded input. One of the early efforts towards end-to-end ASR used CTC with phoneme output targets, which achieved state-of-the-art
performance over conventional ASR baselines. Later, Graves and Jaitly [58] proposed a system with character-based CTC which directly outputs word sequences given the input speech. The researchers used an external language model to re-score the output from the CTC based character model. Several other researchers proposed refinements to the CTC-based ASR systems which further enhanced the performance of these systems [4, 107]. However, CTC-based ASR systems had several intrinsic challenges: For instance, CTC makes an important independence assumption which treats the network outputs at difference frames as conditionally independent outputs. Further, CTC-based systems require the use of an external language model as direct greedy decoding does not perform well [134].

Recently, attention-based encoder-decoder models have been popular which were first applied to an ASR application by Chan et al. [23] and Chorowski et al. [30]. These model have three main components:

- Encoder Layer: The function of the encoder is to transform the input speech into a higher-level representation. This can be thought of as the acoustic model in conventional ASR.

- Attention Layer: The attention layer identifies encoded frames that are relevant to producing the current output. This can be thought of as an alignment model that identifies alignments between the input and the output for making the prediction.

- Decoder Layer: The decoder layer operates by predicting each output token as a function of the previous predictions and the contextualized
representation from the attention layer.

These attention-based encoder-decoder models have achieved state-of-the-art performance in various large-scale ASR tasks, however researchers have identified that these models are not very good for streaming applications. Consequently, several other interventions have been proposed, which includes combining various end-to-end approaches (like CTC and Attention-based models [59]), to other structural improvements (like using multi-headed attention [27]), etc.

2.3 Other Terminology

2.3.1 Confidence Scores

In speech recognition, confidence scores are usually represented as a numeric score between 0 and 1 that is used to evaluate the reliability of the recognition results. For example, a confidence score for every recognized word would indicate how likely it has been correctly recognized by the ASR system.

Confidence scores are often estimated based on a combination of the features collected during the decoding process of ASR (which includes acoustic as well as language information about recognition decisions). Generally, a classifier is then trained with these features to generate a single score to indicate the correctness of the recognition decision. Another common approach to estimating confidence scores is based on the posterior probability $P(W|X)$ in the standard maximum a posteriori (MAP) decision rule (shown in Eq. 2.1). Since this is an absolute measure of trust of the model in its decision, this approach to
confidence estimation is also very popular. However, as shown in Eq. 2.1, a challenge is that the posterior probability estimated for MAP decision does not include the normalization term $P(X)$ in the denominator, which needs to be approximated.

### 2.3.2 Word Error Rate

Accurate, large-vocabulary, continuous speech recognition is still considered an unsolved problem. Although there have been recent leaps in the performance of these systems, ASR performance is generally not on par with humans, who currently provide most caption text for DHH users. Noise in the input audio, the ambiguity of human speech, or unforeseen speaker characteristics (e.g. a strong accent) can lead to ASR errors. As researchers continue to improve ASR accuracy, they generally report the performance of their systems using a metric called Word Error Rate (WER).

WER is calculated by aligning the hypothesis (ASR output) with a human-generated reference transcript using Levenshtein (minimum edit) distance. Given the ubiquity of this metric, it is reasonable that reducing WER may be a goal of many ASR research efforts (implicitly, if not overtly).

\[
WER = \frac{S + D + I}{N} \tag{2.2}
\]

where WER represents the Formula for Word Error Rate, based on $S$ (number of erroneous substitutions of one word for another), $D$ (number of deletions, i.e. erroneous omissions of words that were spoken), $I$ (number of insertions of
spurious words in the ASR output), and N (number of words actually spoken).

As shown in Eq. 2.2, WER is calculated by comparing the “hypothesis
text” (the output of the ASR system) to the “reference text” (what the human
actually said in the audio recording). The metric considers the number of
misrecognitions in the hypothesis text, normalized by the word-length of the
reference text. Notably, WER does not consider whether some words may be
more important to the meaning of the message or whether some words might
be more predictable than others in a text. In fact, researchers have previously
found that humans perceive different ASR errors as having different degrees
of impact on a text – some errors might distort the meaning of the text more
harshly than others [116]. Others have found that the impact of errors may be
dependent upon the specific application in which ASR is used [41, 122].
PART I: WORD IMPORTANCE
MODELING
Many speech-based models consider words as a fundamental unit of meaning and prosody. However, words contribute differently to the meaning of an utterance; some words may be crucial for understanding a turn while others may be less so. This differential importance of words in a spoken language context has benefited various tasks, from speech recognition (ASR) evaluation [116] to text classification [85,176] and summarization [67,177]. The overall objective of research on identifying important words in spoken dialogues is to tackle the usability challenges of ASR systems when envisioning them in captioning applications for DHH users. For instance, Part II of this work discusses word-importance based metrics of quality for accurately evaluating the usability of automatic captioning systems.

With this goal, in Part I of this thesis, we explore approaches to predict the importance of words to the meaning of the spoken dialogues for various readers. Our work begins by investigating prior work on identifying the importance of words for various applications in Chapter 3. Taking inspiration from prior studies on reading mechanisms of DHH readers, Chapter 4 presents our unsu-
pervised method of estimating the importance of a word as a measure of its predictability in the context. Next, in Chapter 5, we describe our method for building the corpus of word importance. Subsequently, Chapter 6 describes our methods towards more accurate word importance modeling through supervised training of statistical models, based on the human-labelled data on importance of words collected in Chapter 5.

Specifically, Part I of this thesis will explore research question **RQ1** (as presented in 1.2), which states:

**RQ1**: How can we identify words in a spoken message that are important to its understandability for DHH readers? To answer this research question, we formulate the following four sub-research questions that we’ll look to investigate in the subsequent chapters:

**RQ1.1**: Does measuring the predictability of word (given context) help measure the importance of that word when focusing on applications for DHH users? (We will examine RQ1.1. in Chapter 4).

**RQ1.2**: Do supervised models based on textual features from a spoken language transcript accurately predict word importance? (We will examine RQ1.2. in Chapter 6).

**RQ1.3**: Do acoustic-prosodic cues in spoken dialogues help identify important words in the dialogue? (We will examine RQ1.3. in Chapter 6).
CHAPTER 2. AUTOMATIC SPEECH RECOGNITION

RQ1.4: Do models trained on both textual and speech features from spoken dialogues outperform word importance models trained on a single type of feature? (We will examine RQ1.4. in Chapter 6)
Chapter 3

Prior Methods of Word Importance Estimation

There is an increasing interest among researchers of speech and language technology applications to identify the importance of individual words, for the overall meaning of the text. Depending on the context of how the importance of a word is defined, this task has found to be useful in varieties of applications such as text summarization [67], text classification [100], or speech synthesis [117]. This chapter discusses methods that have been previously investigated for the task of word importance estimation, for various applications.
3.1 Word Importance Estimation as a Keyword Extraction Problem

Prior research on identifying and scoring important words in a text has largely focused on the task of keyword extraction, which involves identifying a set of descriptive words in a document that serves as a dense summary of the document. Several automatic keyword extraction techniques have been investigated over the years, including unsupervised methods, e.g. Term Frequency Inverse Document Frequency (TF-IDF) weighting [61] – as well as supervised methods that leverage various semantic features from text for prediction [69,100,102,156]. More recently, several neural network architectures have also been proposed that aim at learning how one specific word contributes in the discriminative task that the network is trying to accomplish [29,157,169]. In the coming sections 3.1.1 – 3.3 we will investigate each of these methods in detail by discussing them in two broad categories.

3.1.1 Frequency-based Keyword Extraction

One of the most commonly used techniques for identifying important words (keywords) in a text is the Term-Frequency Inverse Document Frequency (TF-IDF) measure. This strategy is often used to identify relevant words in a document which is based on observation from a larger collection of documents. Much like the word predictability score, TF-IDF scoring is also an unsupervised measure, thereby eliminating the need to collect subjective scores from humans, which can be both resource-intensive and time-consuming. In this work, the
CHAPTER 3. PRIOR WORD IMPORTANCE ESTIMATION

TF-IDF score for a word \((w)\), referred to as a term, in a text document \((D)\) is computed in reference to a collection of reference text documents \((\mathcal{D})\), such that \(D \in \mathcal{D}\), as follows:

\[
\text{TF-IDF}(w, D) = tf(w, D) \times idf(w, \mathcal{D})
\]  

(3.1)

\[
idf(w, \mathcal{D}) = \log\left(\frac{|\mathcal{D}|}{n_w}\right)
\]  

(3.2)

where, \(\text{TF-IDF}(w, D)\) is the TF-IDF measure of a word \(w\) in a document \(D\). Then, \(tf(w, D)\) computes the frequency of the term \((w)\) in the document \((D)\). \(idf(w, \mathcal{D})\) is the inverse-document frequency which measures the presence of the term \((w)\) in the collection document \((\mathcal{D})\), where \(|\mathcal{D}|\) is the number of reference documents used and \(n_w\) is the the number of document that contains the term.

This strategy of ranking words based on their frequency of occurrence in documents have been commonly studied in the literature [61]. Similar methods such as word occurrence frequency and other variants have also been popular [69, 111, 156].

3.1.2 Supervised Methods of Keyword Extraction

Besides the unsupervised methods of keyword extraction, researchers have also investigated several supervised methods based on the semantic features of words (and their context) for identifying important words in the text [69, 100, 102, 121, 156, 174]. Following sections will discuss research work in this avenue, which we have categorized into two big sub-groups:
Utilizing Linguistic Features for Keyword Extraction

Hulth [69] utilized various syntactic features from text, besides the traditional statistical features of word frequency, to extract the relevant keywords from text abstracts. In their work, the author showed that simple syntactic features extracted from the text, such as parts-of-speech tags of words and noun-phrase chunks, resulted in greater performance when utilizing these features as input to supervised learning algorithms. More similar to our work, Hong et al. [67] investigated ways to score words in text based on their likelihood to be included in a human-summarized version of the text. They investigated various unsupervised methods of scoring the importance of words, such as word probability (measures the frequency with which the word occurs in the input text), log-likelihood ratio (compares the occurrence of the word in the input text against a large reference corpus) and Markov random walk model (considers the importance of words as the weights of the vertices in a graph where each node represents a word in the input text and the vertices represent their syntactic relationship – the Page rank algorithm is used to update the weights of the vertices in the graph). Researchers then utilized other features from text such as positional features of word, their part of speech information, and more to train a supervised model for predicting the importance of words. Similar work has also been done by other researchers who have focused on a different genre of texts e.g., conversational style in meeting for keyword identification [102,156].

More recently, several neural network architectures have shown to be useful in the task of language modeling [114,147] and, subsequently, researchers have
explored these architectures for the task of importance scoring of words for various applications \[29, 157, 169\]. For example, Chopra \textit{et al.} \[29\] present a neural encoder-decoder architecture based on Recurrent Neural Network (RNN) units to generate an abstractive summarized representation of a sentence. As discussed in Section 2.1.2, researchers use the RNN-units to generate a context-based representation of words in the text which has been shown to be a useful feature in many linguistic applications, including text summarization. However, rather than using features learned from every word in the sentence, researchers used attention-based filter in their neural architecture that ensures that only important input words are selected for further processing \[29\]. With this setup, researchers demonstrate significant improvement in the sentence summarization task. A similar methodology was utilized by Wang \textit{et al.} \[169\] who used an attention-based encoder-decoder architecture to generate an abstract from multiple sources of opinions and argument in the form of text. Beyond the attention-based networks, Sheikh \textit{et al.} \[157\] demonstrated a Neural Bag-of-Words model where the model uses a weighted bag of words architecture to get a summative representation of a text. In this setup, each feature representation of word is weighted (by a learned parameter \((\alpha)\)) which is used to sum each individual word features to get a total representation of the text. This final representation of text when trained to produce specific application-oriented predictions, such as sentiment prediction in text sentiment analysis \[157\], would in turn learn the weighing parameters \((\alpha)\) for each word, which are shown to correspond with the importance of each word in that application.
3.1.3 Limitations and Challenges

While this conceptualization of word importance as a keyword-extraction problem has led to positive results in the field of text summarization [67, 99, 167], this approach may not generalize to other applications. For instance, given the sometimes meandering nature of topic transition in spontaneous speech dialogue [156], applications that process transcripts of such dialogue may benefit from a model of word importance that is more local, i.e. based on the importance of a word at sentential, utterance, or local dialogue level, rather than at a document-level. Furthermore, the dyadic nature of dialogue, with interleaved contributions from multiple speakers, may require special consideration when evaluating word importance.

To address these challenges, in later sections, we present new methods of word importance scoring that provide importance of word based on its local context. We also present a corpus with annotation of word importance that is used to support research into these complex issues of word importance measurement at a more granular level.

In the coming sections, we discuss prior research that provides groundwork for building unsupervised and supervised models of word importance. More specifically, Section 3.2 explores prior work on reading strategies of deaf individuals where we discuss the effects of several features of text (such as frequency and predictability) on text comprehension for these readers. Inspired by this work, Chapter 4 describes our work on modeling predictability of a word given its context and its use in word-importance prediction for DHH users. Similarly,
Section 3.3 discusses prior work on harnessing acoustic-prosodic cues from speech for semantic modeling for various natural language processing tasks. This work is later referenced in Chapter 6 which discusses various supervised models of word importance including unimodal-feature-based models (e.g., models trained only on acoustic-prosodic features from speech) and multimodal-feature-based models (e.g., models trained on both text- and speech-based features).

### 3.2 Reading Strategies of Deaf Individuals

In prior work, research have hypothesized that deaf readers use a sentence-understanding strategy in which they seek content words in order to derive a representation of sentence meaning, potentially ignoring other information, e.g. morpho-syntactic relationships between words [34,35]. Research on the eye movements of deaf readers has also revealed that deaf readers visually fixate on approximately 30% of the words in a text. The skipped words were largely determined by lexical factors, such as how frequent a word is, the length of the word, and the predictability of the word in that sentence [10]. Similarly, Keith et al. [142] found that both the length of the word and predictability of the word in context were related to whether readers skip over a word and to the amount of time readers spent on non-skipped words. In general, highly predictable words have been shown to be read faster and skipped more often than unpredictable words by most readers [139], and especially by less-skilled readers [10].
Furthermore, word predictability has been a common theme in prior research on assessing the readability of a text or the reading comprehension skills of a participant [37, 84, 140, 141]. For instance, the Cloze procedure is an assessment methodology that has been around for many years, and it is one of the most common ways of evaluating both the readability of a text and the reading skills of participants. In this task, the participant is given a text with one word omitted, and they must guess the missing word. Most standardized English-language tests (e.g., TOEFL, GRE, WRAT) utilize some variation of the Cloze procedure to evaluate participants’ reading skills. The predictability of a word refers to the degree to which a reader can use the context to guess the word. For example:

The ________ was barking at the mail-man.

The predictability of the word “dog” is high given the context. The context of the word is powerful enough to provide a hint as to what the word is. Conversely, in the sentence:

The meeting is scheduled on ________.

The predictability is very low – suggesting that the readers might not be able to rely on the context to predict the word.

Given the use of word predictability in reading assessment (Cloze tests) and given the aforementioned eye-tracking research (indicating that DHH readers are more likely to skip over highly predictable words), this linguistic property could be useful in measuring the importance words in text, especially for DHH users. In contrast to the popular frequency-based approaches (like TF-IDF measure discussed in Section 3.1.1), this approach of importance scoring of words based
on their predictability (given context) have not been fully investigated in the literature. As we discussed earlier, given the possible differences in the reading strategies of readers who are deaf compared to hearing readers, it might be useful to investigate this measure when focusing on applications for this group of users. Section 4 discusses our methodology in realizing this property of text as a word importance measure.

3.3 Acoustic-Prosodic Cues for Semantic Knowledge

Previous researchers have modeled prosodic cues in speech for various applications [19,163,174]. For instance, in automatic prominence detection, researchers predict regions of speech with relatively more spoken stress [19,161,168]. Identification of prominence aids automatically identifying content words [168], a crucial sub-task of spoken language understanding [9,118]. Moreover, researchers have investigated modeling prosodic patterns in spoken messages to identify syntactic relationships among words [136,163]. In particular, Tran et al. [163] demonstrated the effectiveness of speech-based features in improving the constituent parsing of conversational speech texts. In other work, researchers investigated prosodic events to identify important segments in speech, useful for producing a generic summary of the recordings of meetings [121,174]. At the same time, prosodic cues are also challenging in that they serve a range of linguistic functions and convey affect. We investigate models applied to spoken messages at a dialogue-turn level, for predicting the importance of words for understanding an utterance.
Chapter 4

Unsupervised Models of Word Importance

In section 3.2, we discussed a special reading mechanism employed by deaf readers, where readers tend to skip words that are frequent and also words more predictable given the context. While models based on word frequency have been previously explored in the literature for estimating the importance of words in the text, the measure of predictability of a word in a context as an estimate of word importance remains unexplored. In this chapter, we investigate various methods to quantify the measure of predictability of a word in a text, and we evaluate such a method when focusing on applications for DHH users.
4.1 Defining the Word Predictability Measure

We define word predictability as a measure that estimates how predictable a word is in a context. It can be thought of as a measure which estimates the total effort required to make a prediction about the word in that context. The intuition is that, if a word is highly important in the text it would be difficult for a reader to compensate if the word were missing in the text. Hence, words that have stronger contexts, which allow readers to easily infer the word, could be considered unimportant words in the text. To effectively formulate this measure, we consider several language models that are able to make predictions about words given the context. With the help of the language models, we compute the predictability of word based on how difficult it is for the model to make inference about the word given its context.

Figure 4.1: Figure showing how the language model is used to make inference about the predictability of a word given its context in an example sentence. Reader can refer to Section 4.2 for mathematical detail on how this score is computed.

Figure 4.1 shows how we make an estimate of the difficulty of predicting a word given a context. In the figure, the sentence “The _________ is barking at the mail-man.”
at the mail-man” is missing a word. The language model used in the example considers various candidate words for the context, e.g., “dog”, “tree”, etc. Based on how confident the model is across its the various candidates, we estimate how difficult it would be to make an inference from the model. It should be noted that our measure doesn’t directly consider the confidence of the model in predicting the correct word in the context, but rather looks to model the overall difficulty in making a guess using the context. This is a contrasting difference from other measures of predictability of words that have been discussed in the past, like surprisal [55], which is more model-specific.

4.2 Methods for Computing Word Predictability

We investigate two approaches for computing the predictability of a word in a context: First, we investigate the n-gram model which is used to model the word co-occurrence statistics (based on a large reference text) for making predictions on the predictability of a context. Next, we present a neural-language model which makes use of the neural network architecture to model longer dependencies in text for making more informed decision about the predictability of the word. Section 4.2.1 and 4.2.2 provides details on each of these methods, respectively.

4.2.1 N-gram Language Model

N-gram models are one of the popularly used approaches for language modeling that consider the relative frequency counts of words and word phrases from a very large corpus to make estimates about the language. As discussed in
Section 2.1.2, one of the key tasks of a language model is to be able to make predictions about the word \(w\) given its context \(c\), i.e., computing \(P(w|c)\). Rather than using the entire context (history) to make the prediction, n-gram models consider approximating the context using only the last \(n-1\) words. For example, a bigram model \((n=2)\) approximates the prediction of a word using only the preceding word, i.e., \(P(w_t|w_{t-1})\). Similarly, a trigram model \((n=3)\) makes prediction about a word using the last two words from the context, i.e., \(P(w_t|w_{t-1},w_{t-2})\). Theoretically, as the value of \(n\) increases, the model is able to make more accurate predictions. However, in reality using a longer context also increases data sparsity as it becomes increasingly difficult to find longer common sequences of words in the training corpus, hence making accurate predictions more difficult.

**Methodology: Estimating Word Predictability Using N-grams**

To compute the predictability score of a word, we utilized several n-gram language models; these models are based on how frequently certain sequences of words, of various length, have appeared in large collections of text. Similar models are commonly employed in word-prediction systems for text-entry applications, e.g., [46]. Based on the probability score assigned to the predictions by the language model, we compute the predictability score for the word. We trained our n-gram models \((n = 1 \text{ to } 5)\) on the Switchboard [51], the English CALLHOME [21], and the TEDLIUM [152] corpora, which contain a total of 1.9 million word tokens. These corpora were selected because they closely represent conversational speech dialogues (similar to the one-on-one meeting
context in which we are considering the use of ASR for real-time captioning). The n-gram models were used bi-directionally, to make predictions using both left and right word-sequence contexts, independently. To rank the possible word candidates using each context, a so-called Stupid Back-off [17] mechanism was utilized. For ranking predictions from the left context, the following scoring function was used:

\[
S(w_t | w_{t-n+1}^{t-1}) = \begin{cases} 
\frac{\text{count}(w_{t-n+1})}{\text{count}(w_t)} & \text{if count}(w_{t-n+1}) > 0 \\
\lambda S(w_t | w_{t-n+2}^{t-1}) & \text{otherwise}
\end{cases}
\] (4.1)

where \( w_t \) represents word indexed at position \( t \), and \( w_x^y \) represents a word sequence beginning at position \( x \) and ending at \( y \) such that \( y > x \). Then, \( n \) represents the n-gram size and we utilized a value of 0.4 for lambda (\( \lambda \)), as recommended by Brants et al. [17]. A similar scoring function was used to rank the candidates from the right context. The predictions from both the right and left contexts were combined and then ranked for later use.

To obtain a predictability score from these predictions, we first selected the top \( N = 20 \) ranked unique candidates and transformed their count probabilities to normalized probabilities (that sum to 1). For instance, in the sentence: “The meeting is scheduled on _________” example, the language model might predict various possible words, e.g. “Monday, Friday, Tuesday, etc.” The model will predict that each of these words has some probability of appearing in that context. Based on the distribution of probability among the candidates, an entropy score was calculated as follows:
CHAPTER 4. UNSUPERVISED MODELS OF WORD IMPORTANCE

\[ E(w) = \sum_{i=0}^{N_c} -P(w_c(i)) \cdot \log(P(w_c(i))) \]  \hspace{1cm} (4.2)

where \( E(w) \) represents the entropy of a word \( w \) (at a unique location in the text). \( w_c(i) \) is a candidate of the word \( w \) predicted by the language model and \( P(w_c(i)) \) is the probability of the candidate \( w_c(i) \) as determined by the language model.

The entropy score calculated in Equation 4.2 is a measure often used in information theory to calculate the unpredictability of a state. In our application, it is the measure of the degree of unpredictability of a word given the context. A higher value indicates that the chances of picking the right word from the list of candidates are low - meaning it is difficult to predict the word. Whereas, a lower value would indicate that some words in the list of candidates clearly have a higher probability than others - meaning it is easier to predict the word. The entropy is normalized to get a predictability score within a (0, 1) range.

**Challenges and Limitations**

The current n-gram based word importance prediction model does not generalize well to unseen data (e.g., texts containing out-of-vocabulary words), as these models are based on exact-search and match strategy. Further, there are issues of long text dependencies that are often ignored when using the n-gram models. To overcome these limitations, in the coming section (Section 4.2.2) we discuss a neural architecture for language modeling.
4.2.2 Neural Language Model

In the previous section, we utilized several n-gram models to measure how easy (or difficult) it is to make predictions of words given the context. However, n-gram based models have some inherent challenges, as discussed in section 4.2.1, especially when encountering out-of-vocabulary words.

To tackle these limitations, we investigated neural-network-based language models to estimate the predictability of a word in a context. Neural network based language models have been quite popular and have been shown to be useful in varieties of applications. Therefore, the next logical step towards improving our word predictability model is to investigate the neural-language model based estimation of word predictability. However, for the purposes of our application we need to make some architectural changes to the standard neural language-model setup – the methodological details of which are described in the following section.

Figure 4.2: Diagram of neural word predictability model demonstrating how the context of a word \( w(i) \) is captured using bi-directional recurrent units.
Model Architecture and Training

We utilize a bi-directional RNN to build our language model: Our model uses pre-trained GLoVE\(^1\) embedding representation for words as input, which is then processed by Long-short Term Memory (LSTM) units for context modeling. To make prediction for a word \((w(t))\) at time \((t)\), the model uses the hidden representation of the context of the word from both directions (i.e., \(h_{fw}(t-1)\) from the forward moving LSTM and \(h_{bw}(t+1)\) from the backward-moving LSTM) skipping the hidden representations \((h_{fw}(t)\) and \(h_{bw}(t)\)) of the word, as seen in the Fig. 4.2. A similar setup for language modeling has been discussed previously by Rei [143].

The hidden representations from the forward and backward LSTMs are passed through softmax layers in order to make the prediction about the word \((w(t))\):

\[
P(w|h_{fw}(t-1)) = \sigma(W_{fw}h_{fw}(t-1)) \tag{4.3}
\]

\[
P(w|h_{bw}(t+1)) = \sigma(W_{bw}h_{bw}(t+1)) \tag{4.4}
\]

For training, the objective function for both components is then constructed as a regular language modeling objective which calculates the negative log-likelihood for the prediction, as:

\(^1\)GLoVE is an unsupervised training algorithm based on word to word co-occurrence statistics from a large corpus that is used to obtain vector representation for words. The pre-trained GLoVe embedding used in our analysis was obtained from: https://nlp.stanford.edu/projects/glove/
CHAPTER 4. UNSUPERVISED MODELS OF WORD IMPORTANCE

\[ E_c = - \sum_t P(w(t)|c) \]  \hspace{1cm} (4.5)

where \( c \) represents the context for the word \( (w(t)) \) which is either \( h_{fw}(t - 1) \) or \( h_{bw}(t + 1) \). The total loss of the model is obtained by calculating the sum of the loss due to both (forward and backward) contexts. To make predictions, we use combine the probability score assigned using both the forward and the backward context.

The neural-network was implemented using Tensorflow [162]. The LSTM hidden layers used for context-modeling of words were set to size 650 in each direction. We replaced all the digits in the text with the character 0, and the vocabulary size was fixed to 23,000. The utterances were grouped into batches of size 50, for optimization. For training, the Adam optimizer [82] with an initial learning rate of 0.001 and decay of 0.9 was used. During training, the model operated on utterance units without maintaining previous dialogue contexts.

Methodology: Estimating Word Predictability Using Neural Language Model

To compute the word predictability scores for the word, we utilized the combined probability scores from the model which considers predictions from both the forward- and the backward-moving LSTMs for each word position. Using this combined probability score, we computed entropy using the equation described in Equation 4.2. As before, this entropy score was normalized to get
4.3 Evaluation and Conclusion

Motivated by the effect of word predictability on the reading patterns of various readers, this chapter discussed methods for estimating the predictability of words given its context using various types of language models. In prior work, eye tracking studies revealed that readers spend more time (as defined by the eye gaze-duration during reading) on words that are difficult to predict given their context, as compared to time spent on words that are easily predictable; this suggested the importance of a word with this linguistic property in the comprehension of the text. With this premise, this chapter investigated several language models (e.g., n-gram based models and DNN-based models) for estimating the predictability of words given context, as a measure of their importance.

Rather than presenting an intrinsic evaluation of these models at this time, we instead embed them on a useful application, and in a later chapter, we present our evaluation of our word-importance models when applied to the task of ASR-generated error impact prediction in text understandability for DHH users. Details of this evaluation are described in Section 9.4.1.

To summarize the results that will be formally presented in Chapter 9, we found that these methods of modeling the predictability of words and using them as a word importance measure in an error-impact prediction task led to improved performance when compared to the traditional frequency-based
word importance estimation methods like TF-IDF. In particular, we found that a neural-language model based word predictability estimation (as a word importance measure) yielded superior performance when compared to the other models in the task.
Chapter 5

Building the Word Importance Annotation Corpus

Previously in Chapter 4, we discussed methods for estimating predictability of words as our first attempt at considering importance of words. However, this measure is not based on the actual empirical evidence of importance from our users, therefore it can’t be fully validated as such. For the purpose, we need to gather data from people and train more sophisticated models of word importance. This chapter describes efforts at gathering data about importance of words in spoken conversations, which will be used in a later chapter (Chapter 6) for training (and evaluating) supervised models of word importance.

More specifically, we setup the task of collecting information about the importance of words in spoken dialogues as an annotation task. We begin by quantifying the word importance information in spoken dialogues in Section 5.1, then describing the annotation task in Section 5.2 and analyzing the validity of
the annotations in Section 5.3.

5.1 Defining Word Importance

![Figure 5.1: Visualization of importance scores assigned to words in a sentence by a human annotator on our project, with the height and font-size of words indicating their importance score (and redundant color coding: green for high-importance words with score above 0.6, blue for words with score between 0.3 and 0.6, and gray otherwise).]

We discussed in Section 3.2 that eye-tracking studies have revealed the reading patterns of various readers. While these studies suggest some features that may relate to readers’ judgments of word importance, at least as expressed through their choice of eye fixations, we needed to develop a specific definition of word importance in order to develop annotation guidelines for our study. Rather than ask annotators to consider specific features, e.g. word length, which
may pre-suppose a particular model, we instead took a functional perspective, with our application domain in mind. That is, we define word importance for spontaneous spoken conversation as the degree to which a reader of a transcript of the dialogue would be unable to understand the overall meaning of a conversational utterance (a single turn of dialogue) if that word had been dropped or omitted from the transcript. This definition underlies our data acquisition strategy and also suits our target application, e.g., evaluating ASR for real-time captioning of meetings.

5.2 Word Importance Annotation Task

The Switchboard corpus consists of audio recordings of approximately 260 hours of speech consisting of about 2,400 two-sided telephone conversations among 543 speakers (302 male, 241 female) from across the United States [51]. In January 2003, the Institute for Signal and Information Processing (ISIP) released written transcripts for the entire corpus, which consists of nearly 400,000 conversational turns. The ISIP transcripts include a complete lexicon list and automatic word alignment timing corresponding to the original audio files.

5.2.1 Annotation Scheme

For our annotation project, we defined word-importance as a single-dimensional property, which could be expressed on a continuous scale from 0.0 (not important

\[1https://www.isip.piconepress.com/projects/switchboard/\]
at all to the meaning of the utterance) to 1.0 (very important). Figure 5.1 illustrates how numerical importance scores can be assigned to words in a sentence – in fact, this figure displays actual scores assigned by a human annotator working on our project. Of course, asking human annotators to assign specific numerical scores to quantify the importance of a word is not straightforward. In this section, we discuss how we attempt to overcome the subjective nature of this task, to promote consistency between annotators, as we developed this annotated resource. Section 5.3 characterizes the level of agreement between our annotators on this task.

To reduce the cognitive load on annotators and to promote consistency, we created the following annotation scheme:

Range and Constraints. Each word is assigned a numeric score between [0, 1], where 1 indicates a high importance score; the numeric score has the precision of 0.05. Importance scores are not meant to indicate an absolute proportion of the utterance’s meaning represented by each word, i.e. the scores do not have to sum to 1.

Methodology. Given an utterance (a speaker’s single turn in the conversation), the annotator first considers the overall meaning conveyed by the utterance, with the help of the previous conversation history (if available). The annotator then scores each word based on its (direct or indirect) contribution to the utterance’s meaning, using the rubric described in the Interpretation and Scoring section below.

Rating Scheme. To help annotators calibrate their scores, Table 5.1 provides some recommendations for how to select word-importance scores in
CHAPTER 5. WORD IMPORTANCE ANNOTATION CORPUS

### Table 5.1: Guidance for the annotators to promote consistency and uniformity in the use of numerical scores.

<table>
<thead>
<tr>
<th>Range</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>[0 - 0.3)</td>
<td>Words that are of least importance - these words can be easily omitted from the text without much consequence.</td>
</tr>
<tr>
<td>[0.3 - 0.6)</td>
<td>Words that are fairly important - omitting these words will take away some important details from the utterance.</td>
</tr>
<tr>
<td>[0.6 - 1]</td>
<td>Words that are of high importance - omitting these words will change the message of the utterance quite significantly.</td>
</tr>
</tbody>
</table>

Various numerical ranges.

**Interpretation and Scoring.** Annotators should consider how their understanding of the utterance would be affected if this word had been *dropped*, i.e. replaced with a blank space (“____”). Since these are conversations between pairs of speakers, annotators should consider how much the other person in the conversation would have difficulty understanding the speaker’s message if that word had been omitted, i.e. if they had not heard that word intelligibly.

### 5.3 Inter-Annotator Agreement Analysis

There were 3,100 tokens in our *overlap* set, i.e. the subset of transcripts independently labeled by both annotators. This set was used as the basis for calculating inter-annotator agreement. Since scores were nearly continuous (ranges [0,1] with a precision of 0.05), we computed the Concordance Correlation...
Coefficient \((\rho_c)\), also known as Lin’s concordance correlation coefficient, as our primary metric for measuring the agreement between the annotators. This metric indicates how well a new test or measurement \((X)\) reproduces a gold standard or measure \((Y)\). Considering the annotations from one annotator as a gold standard, we can generalize this measure to compute the agreement between two annotators. Like other correlation coefficients, \(\rho_c\) also ranges from -1 to 1; 1 being the score of perfect agreement.

Concordance between the two measures can be characterized by the expected value of their squared difference as:

\[
E[(Y - X)^2] = (\mu_y - \mu_x)^2 + \sigma_x^2 + \sigma_y^2 - 2\rho\sigma_x\sigma_y
\]  
(5.1)

where, \(\rho\) is the correlation coefficient, \(\mu_x\) and \(\mu_y\) are the means of the population of the variables \(X\) and \(Y\), and \(\sigma_x\) and \(\sigma_y\) are their standard deviation. The expectation score coefficient (between -1 and 1) is calculated as follows:

\[
\rho_c = \frac{2\rho S_x S_y}{(Y - X)^2 + S_x^2 + S_y^2}
\]  
(5.2)

where, \(\rho_c\) is the correlation coefficient, \(\bar{X}\) and \(\bar{Y}\) are the mean of \(X\) and \(Y\), and \(S_x\) and \(S_y\) are standard deviations.

Using the concordance correlation measures, we obtained an agreement score \((\rho_c)\) of 0.89 between our annotators, which we interpret as an acceptable level of agreement, given the subjective nature of the task of quantifying word importance in spoken dialogue transcripts.
5.4 Summary of the Corpus

This section presented the Word Importance Annotation corpus, a new collection of annotation of transcripts of the Switchboard conversational speech corpus produced through human annotation of the importance of individual words to the meaning of each utterance. The corpus consists of over 25,000 terms that have been manually labelled with word importance information. The importance scores are represented on the scale of \([0, 1]\), with high importance word receiving an importance score of 1 and low importance word receiving a score of 0. In the corpus, the importance of a word is relative to its context of use in a conversation – meaning the importance of a word defined in a context is independent of its importance in other context(s).

In our project, a pair of hearing annotators have assigned word-importance scores to these transcripts. As of September 2017, they have annotated over 25,000 tokens, with the overlap of approximately 3,100 tokens. This covers 25,048 utterances spoken by 44 different English speakers. We made these annotations publicly available\(^2\) as a set of supplementary files, aligned to the ISIP transcripts for the Switchboard corpus.

Despite the subjective nature of this task, we found that with the creation of a detailed protocol for our annotation team (defined in Section 5.2), we can achieve a significant level of agreement in the word importance information among the annotators. We found that our concordance correlation score between our two annotators was 0.89.

\(^2\)http://latlab.ist.rit.edu/lrec2018
As discussed in Section 3.1, current methods of word importance estimation in text utilizes a document-level abstraction of importance. This means that all the words identified by a term in the document receives a similar importance score, regardless of where the words occur in the document. In contrast, the goal of this work has been to enable prediction of word importance at a more granular level. In the coming sections, we demonstrate the use of this corpus in training and evaluation of several supervised word-importance prediction models, which are capable for making word-importance prediction at a sentential level rather than at a document level.

Further, by overlaying word-level importance information on a conversational speech corpus, we promote the design of word importance models that are more useful in spoken contexts; this is relevant to the application use-case of captioning real-time communication discussed in this thesis.
Chapter 6

Supervised Models of Word Importance

6.1 Text-based Model of Word Importance

Our exploration into the word importance prediction model begins with the investigation into the text-based models for word importance. By “text-based” models we refer to models that are based on features derived from the written form of the text, rather than models which may use some acoustic information from a speech audio signal (see Section 6.2). With the recent success of deep neural networks in linguistic applications, we considered these approaches for this task. In this work, we will discuss the design and the development of the neural network architecture for our task (Section 6.1.1) and describe our training and evaluation methodology for word importance modeling (Section 6.1.2).
6.1.1 Model Architecture

We adopted the neural architecture described by Lample et al. [90] consisting of bidirectional LSTM encoders with a sequential Conditional Random Field (CRF) layer on top. Our input word tokens were first mapped to a sequence of pre-trained distributed embeddings [131] and then combined with the learned character-based word representations to get the final word representation. As shown in Figure 6.1, the bidirectional LSTM encoders are used to create a context-aware representation of each word. The hidden representations from each LSTM were concatenated to obtain a final representation, conditioned on the whole sentence. The CRF layer uses this representation to look for the most optimal state \( Y \) sequence through all the possible state configurations.

The neural framework was implemented using Tensorflow, and the code is publicly available\(^1\). The word embeddings were initialized with publicly available pre-trained GLoVE vectors [131]. The embeddings for characters were set to length 100 and were initialized randomly. The LSTM layer size was set to 300 in each direction for word- and 100 for character-level components. Parameters were optimized using the Adam [82] optimizer, with the learning rate initialized at 0.001 with a decay rate of 0.9, and sentences were grouped into batches of size 20. We applied a dropout with a probability of 0.5 during training on word embeddings.

We investigated two variations of this model: (i) a bidirectional LSTM model with sequential CRF layer on top (LSTM-CRF) treating the problem as a discrete classification task, (ii) a new bidirectional LSTM model with a

---

\(^1\)https://github.com/SushantKafle/speechtext-wimp-labeler
Figure 6.1: General unfolded network structure of our model, adapted from Lample et al. [90]. The bottom layer represents word-embedding inputs, passed to bi-directional LSTM layers above. Each LSTM takes as input the hidden state from the previous time step and word embeddings from the current step, and outputs a new hidden state. $C_i$ concatenates hidden representations from LSTMs ($L_i$ and $R_i$) to represent the word at time $i$ in its context.

sigmoid layer on top (LSTM-SIG) for a continuous prediction. The LSTM-CRF models the prediction task as a classification problem, using a fixed number of non-ordinal class labels. In contrast, the LSTM-SIG model provides a continuous prediction, using a sigmoid non-linearity to bound the prediction
scores between 0 and 1. Using a square loss, we train this model to directly learn to predict the annotation scores, similar to a regression task.

6.1.2 Experimental Setup

Partitioning the Word Importance Annotation corpus (described in Section 5.2) as 80% training, 10% development, and 10% test sets, we evaluated our model using two measures: (i) total root mean square error (RMS) - the deviation of the model predictions from the human-annotations and, (ii) $F_1$ measure in a classification task - the ability of the model to predict human-annotations categorized into a group of classes. To evaluate performance in terms of classification, we discretized annotation scores into 6 classes: $[0, 0.1)$, $[0.1, 0.3)$, $[0.3, 0.5)$, $[0.5, 0.7)$, $[0.7, 0.9)$, $[0.9, 1]$.

6.1.3 Experiment 1: Performance of the Models

Table 6.1 summarizes the performance of our models on the test set, presenting average scores for 5 different configurations, to compensate for outlier results due to randomness in model initialization. While the LSTM-CRF had a better (higher) F-score on the classification task, its RMS score was worse (higher) than the LSTM-SIG model, which may be due to the limitation of the model as discussed in Section 5.

Confusion matrices in Figure 6.2 provide a more detailed view of the classification performance of each model. Since the LSTM-SIG was trained to optimize the accuracy of its continuous predictions, rather than its discrete assignment of instances to classes, it is not surprising to see a wider diagonal
Table 6.1: Model performance in terms of RMS deviation and macro-averaged $F_1$ score, with best results in **bold** font.

<table>
<thead>
<tr>
<th>Model</th>
<th>RMS</th>
<th>$F_1$ (macro)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSTM-CRF</td>
<td>0.154</td>
<td><strong>0.60</strong></td>
</tr>
<tr>
<td>LSTM-SIG</td>
<td>0.120</td>
<td>0.519</td>
</tr>
</tbody>
</table>

Figure 6.2: Confusion matrices for each model for classification into 6 classes: $c_1 = [0, 0.1)$, $c_2 = [0.1, 0.3)$, and so forth.

in the confusion matrix in Figure 6.2(b), which indicates that the LSTM-SIG model was more likely to misclassify words to ordinally adjacent classes. The figure illustrates that both models were worse at classifying words with importance scores in the middle range $[0.3, 0.7)$.

### 6.1.4 Experiment 2: Comparison with Human Annotators

Treating our human-annotations as ground truth, we also computed the concordance correlation coefficient to measure the agreement between the human annotation and each model. The average correlation between the human an-
notator and the LSTM-CRF model was higher ($\rho_c = 0.839$), as compared to the LSTM-SIG model ($\rho_c = 0.826$). For comparison, the agreement analysis between the annotators was ($\rho_c = 0.89$).

### 6.1.5 Limitations of this Research

We presented a supervised model of word importance which operates over textual transcripts of spoken conversations at an utterance-level and makes prediction on the importance of each spoken word to the overall meaning conveyed by the utterance. The model makes use of bi-directional neutral network architecture for sequence modeling and tagging. For vectorized representation of spoken words, we utilized the pre-trained GLoVE vectors [131].

All of the models were trained and evaluated using the the Word Importance Annotation corpus (described in Section 5.2), where the best achieving model had an $F_1$ score of 0.60 on a 3-class word importance classification task. Further, we obtained a model-human agreement correlation of 0.839, in comparison of human-human agreement correlation of 0.89.

The results show the efficacy of Word Importance Annotation corpus, described in Section 5.2, in building neutral networks models for predicting word importance. However, the current model relies on text-based linguistic knowledge for modeling, ignoring the information existing in the form of speech (e.g., prosody). Hence, in the coming section (Section 6.2), we will investigate the usefulness of speech-based information in the task of word importance modeling in conversational speech.
6.2 Speech-based Importance Model

As discussed in Section 3.3, acoustic-prosodic cues contain important semantic information which has been found to be useful for a variety of tasks. This section investigates the usefulness of these speech-based features in the task of word importance prediction, especially because prior models based on text features for word importance identification [77,157] have faced challenges when applied to conversational speech:

- **Difference from Formal Texts**: Unlike formal texts, conversational transcripts may lack capitalization or punctuation, use informal grammatical structures, or contain disfluencies (e.g. incomplete words or edits, hesitations, repetitions), filler words, or more frequent out-of-vocabulary (and invented) words [113].

- **Availability and Reliability**: Text transcripts of spoken conversations require a human transcriptionist or an ASR system, but ASR transcription is not always reliable or even feasible, especially for noisy environments, nonstandard language use, or low-resource languages, etc.

While spoken messages include prosodic cues that focus a listener’s attention on the most important parts of the message [44], such information may be omitted from a text transcript, as in Figure 6.3, in which the speaker pauses after “right” (suggesting a boundary) and uses rising intonation on “from” (suggesting a question). Moreover, there are application scenarios where transcripts of spoken messages are not always available or fully reliable. In such cases, models based
Figure 6.3: Example of conversational transcribed text, *right where you move from*, that is difficult to disambiguate without prosody. The intended sentence structure was: *Right! Where you move from?*

with a speech signal (without a text transcript) might be preferred.

With this motivation, we investigate modeling acoustic-prosodic cues for predicting the importance of words to the meaning of a spoken dialogue. The goal is to explore the versatility of speech-based (text-independent) features for word importance modeling. In this work, we frame the task of word importance prediction as sequence labeling and utilize a bi-directional Long Short-Term Memory (LSTM)-based neural architecture for context modeling on speech.

### 6.2.1 Model Architecture

For the task of word importance prediction, we formulate a sequence labeling architecture that takes as input a spoken dialogue turn utterance with word-level timestamps\(^2\), and it assigns an importance label to every spoken word in the turn using a bi-directional LSTM architecture [68, 90].

\(^2\)For the purposes of accurately evaluating efficacy of speech-based feature for word importance, we currently make use of high-quality human-annotated word-level timestamp information in the train or evaluation corpus; in the future, speech tokenization could be automated.
Figure 6.4: Architecture for feature representation of spoken words using time series speech data. For each spoken word \((w)\) identified by a word-level timestamp, a fixed-length interval window \((\tau)\) slides through to get \(n = time(w)/\tau\) sub-word interval segments. Using an RNN network, a word-level feature \((s)\), represented by a fixed-length vector, is extracted using the features from a variable-length sub-word sequence.

\[
\overrightarrow{h_t} = LSTM(s_t, \overrightarrow{h_{t-1}}) \tag{6.1}
\]

\[
\overleftarrow{h_t} = LSTM(s_t, \overleftarrow{h_{t-1}}) \tag{6.2}
\]

The word-level timestamp information is used to generate an acoustic-prosodic representation for each word \((s_t)\) from the speech signal. Two LSTM units, moving in opposite directions through these word units \((s_t)\) in an utterance, are then used for constructing a context-aware representation for every word. Each LSTM unit takes as input the representation of the word \((s_t)\), along with the hidden state from the previous time step, and each outputs a new hidden state. At each time step, the hidden representations from both LSTMs are concatenated \(h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}]\), in order to obtain a contextualized
representation for each word. This representation is next passed through a projection layer (details below) to the final prediction for a word.

Importance as Ordinal Classification

We define word importance prediction as the task of classifying the words into one of the many importance classes, e.g., high importance (\text{Hi}), medium importance (\text{Mid}) and low importance (\text{Low}) (details on Section 6.2.3). These importance class labels have a natural ordering such that the cost of misclassification is not uniform e.g., incorrect classification of \text{Hi} class for \text{Li} class (or vice-versa) will have higher error cost than classification of \text{Hi} class for \text{Mi}. Considering this ordinal nature of the importance class labels, we investigate different projection layers for output prediction: a softmax layer for making local importance prediction (\text{SOFTMAX}), a relaxed softmax tailored for ordinal classification (\text{ORD}), and a linear-chain conditional random field (\text{CRF}) for making a conditioned decision on the whole sequence.

- \textbf{Softmax Layer}. For the \text{SOFTMAX}-layer, the model predicts a normalized distribution over all possible labels ($L$) for every word conditioned on the hidden vector ($h_t$).

- \textbf{Relaxed Softmax Layer}. In contrast, the \text{ORD}-layer uses a standard sigmoid projection for every output label candidate, without subjecting it to normalization. The intuition is that rather than learning to predict one label per word, the model predicts multiple labels. For a word with label $l \in L$, all other labels ordinally less than $l$ are also predicted. Both the softmax and the
relaxed-softmax models are trained to minimize the categorical cross-entropy, which is equivalent to minimizing the negative log-probability of the correct labels. However, they differ in how they make the final prediction: Unlike the softmax layer which considers the most probable label for prediction, the ord-layer uses a special scanning strategy [25] – where for each word, the candidate labels are scanned from low to high (ordinal rank), until the score from a label is smaller than a threshold (usually 0.5) or no labels remain. The last scanned label with score greater than the threshold is selected as the output.

- **CRF Layer.** The CRF-layer explores the possible dependence between the subsequent importance label of words. With this architecture, the network looks for the most optimal path through all possible label sequences to make the prediction. The model is then optimized by maximizing the score of the correct sequence of labels, while minimizing the possibility of all other possible sequences.

Considering each of these different projection layers, we investigate different models for the word importance prediction task. Section 6.2.2 describes the architecture we use for acoustic-prosodic feature representation at the word level, and Sections 6.2.3–6.2.6 describe the experimental setup and subsequent evaluations.

### 6.2.2 Acoustic-Prosodic Feature Representation

Similar to the familiar feature-vector representations of words in a text, e.g., word2vec [115] or GloVe [131], various researchers have investigated vector
representations of words based on speech. In addition to capturing acoustic-phonetic properties of speech [32,65], some recent work on acoustic embeddings has investigated encoding semantic properties of a word directly from speech [31]. In a similar way, this work investigates a speech-based feature representation strategy that considers prosodic features of speech at a sub-word level, to learn a word-level representation for the task of importance prediction in spoken dialogue.

Sub-word Feature Extraction

We examined four categories of features that have been previously considered in computational models of prosody, including: pitch-related features (10), energy features (11), voicing features (3) and spoken-lexical features (6):

- **Pitch (freq) and Energy (eng) Features**: Pitch and energy features have been found effective for modeling intonation and detecting emphasized regions of speech [19]. From the pitch and energy contours of the speech, we extracted: minimum, time of minimum, maximum, time of maximum, mean, median, range, slope, standard deviation and skewness. We also extracted RMS energy from a mid-range frequency band (500-2000 Hz), which has been shown to be useful for detecting prominence of syllables in speech [161].

- **Spoken-lexical Features (lex)**: We examined spoken-lexical features, including word-level spoken language features such as duration of the spoken word, the position of the word in the utterance, and duration of silence before the word. We also estimated the number of syllables spoken in a word, using the
methodology of Jong et al. [33]. Further, we considered the per-word average syllable duration and the per-word articulation rate of the speaker (number of syllables per second).

- **Voicing Features (voc):** As a measure of voice quality, we investigated spectral-tilt, which is represented as \((H_1 - H_2)\), i.e. the difference between the amplitudes of the first harmonic \((H_1)\) and the second harmonic \((H_2)\) in the Fourier Spectrum. The spectral-tilt measure has been shown to be effective in characterizing glottal constriction [81], which is important in distinguishing voicing characteristics, e.g. whisper [71]. We also examined other voicing measures, e.g. Harmonics-to-Noise Ratio and Voiced Unvoiced Ratio.

In total, we extracted 30 features using Praat [14], as listed above. Further, we included speaker-normalized \((\text{znorm})\) version of the features. Thereby, we had a total of 60 speech-based features extracted from sub-word units.

**Sub-word to Word-level Representation**

The acoustic features listed above were extracted from a 50-ms sliding window over each word region with a 10-ms overlap. In this model, each word was represented as a sequence of these sub-word features with varying lengths, as shown in Figure 6.4. To get a feature representation for a word, we utilized a bi-directional Recurrent Neural Network (RNN) layer on top of the sub-word features. The spoken-lexical features were then concatenated to this word-level feature representation to get the final feature vectors. For this task, we utilized Gated Recurrent Units (GRUs) [28] as the RNN cell, rather than LSTM units,
due to better performance observed during the initial analysis.

6.2.3 Experimental Setup

We utilized the Word Importance Annotation (described in Section 5.2) which consists of a portion of the Switchboard corpus [51] that had been manually annotated with word importance scores [77]. As discussed in previous sections, the annotation covers 25,048 utterances spoken by 44 different English speakers, containing word-level timestamp information along with a numeric score (in the range of [0, 1]) assigned to each word from the speakers. These numeric importance scores have three natural ordinal ranges [0 - 0.3), [0.3, 0.6), [0.6, 1] that the annotators had used during the annotation to indicate the importance of a word in understanding an utterance. The ordinal range represents low importance (LI), medium importance (MI) and high importance (HI) of words, respectively.

The models were trained and evaluated using this data, treating the problem as a ordinal classification problem with the labels ordered as (LI < MI < HI). We created a 80%, 10% and 10% split of the data for training, validation, and testing. The prediction performance of the model was primarily evaluated using the Root Mean Square (RMS) measure, to account for the ordinal nature of labels. Additionally, the evaluation includes F-score and accuracy results to measure classification performance. As the baseline, we used various text-based importance prediction models trained and evaluated on the same data split, as described in Section 6.2.6.

For training, we explored various architectural parameters to find the best-
working setup for the models: The input layer of GRU-cells, used as word-based speech representation, had a dimension of 64. The LSTM units, used for generating contextualized representation of a spoken word, had a dimension of 128. We used the Adam optimizer with an initialized learning rate of 0.001 for training. Each training batch had a maximum of 20 dialogue-turn utterances, and the model was trained until no improvement was observed in 7 consecutive iterations.

Tables 6.2, 6.3 and 6.4 summarize the performance of the models on the word importance prediction task. The performance scores reported in the tables are the average performance across 5 different trials, to account for possible bias due to random initialization of the model.

6.2.4 Experiment 1: Comparison of the Projection Layers

We compared the efficacy of the learning architecture’s three projection layers (Section 6.2.1) by training them separately and comparing their performance on the test corpus. Table 6.2 summarizes the results of this evaluation.

<table>
<thead>
<tr>
<th>Model</th>
<th>ACC</th>
<th>F1</th>
<th>RMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSTM-CRF</td>
<td>64.22</td>
<td>56.31</td>
<td>75.21</td>
</tr>
<tr>
<td>LSTM-softmax</td>
<td>65.66</td>
<td>57.34</td>
<td>74.08</td>
</tr>
<tr>
<td>LSTM-ORD</td>
<td>63.72</td>
<td>57.58</td>
<td>68.21</td>
</tr>
</tbody>
</table>

Table 6.2: Performance of the speech-based models on the test data under different projection layers. Best performing scores highlighted in bold.

Results and Analysis: The LSTM-softmax-based and LSTM-CRF-based projection layers had nearly identical performance; however, in comparison, the
LSTM-ORD model had better performance with significantly lower RMS score than the other two models. This suggests the utility of the ordinal constraint present in the ORD-based model for word importance classification.

### 6.2.5 Experiment 2: Ablation Study on Speech Features

To compare the effect of different categories of speech features on the performance of the model, we evaluated variations of the model by removing one feature group at a time from the model during training. Table 6.3 summarizes the results of the experiment.

<table>
<thead>
<tr>
<th>Model</th>
<th>ACC</th>
<th>F1</th>
<th>RMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>speech-based</td>
<td>63.72</td>
<td>57.58</td>
<td>68.21</td>
</tr>
<tr>
<td>- ENG</td>
<td>62.24</td>
<td>55.67</td>
<td>71.14</td>
</tr>
<tr>
<td>- FREQ</td>
<td>63.25</td>
<td>57.30</td>
<td>69.0</td>
</tr>
<tr>
<td>- VOC</td>
<td>62.90</td>
<td>56.84</td>
<td>70.5</td>
</tr>
<tr>
<td>- LEX</td>
<td>63.37</td>
<td>57.34</td>
<td>71.49</td>
</tr>
<tr>
<td>- ZNORM</td>
<td>62.04</td>
<td>53.86</td>
<td>72.0</td>
</tr>
</tbody>
</table>

Table 6.3: Speech feature ablation study. The minus sign indicates the feature group removed from the model during training. Markers (⋆ and †) indicate the biggest and the second-biggest change in model performance for each metric, respectively.

**Results and Analysis:** Omitting speaker normalized features (ZNORM), which included raw features normalized per speaker, resulted in the greatest increase in the overall RMS error (+5.5% relative increase in RMS respectively). Similarly, omitting the spoken-lexical features (LEX), which includes word-level spoken language features, results in the second greatest increase in the overall RMS errors (+4.8% relative increase in RMS respectively). This shows the
discriminative importance of these features for word importance prediction. Further, the results indicated the importance of energy-based (ENG) features, which resulted in a substantial drop (-2.4% relative decrease) in accuracy of the model.

6.2.6 Experiment 3: Comparison with the Text-based Models

In this analysis, we compare the best-performing speech-based model with a state-of-the-art word-prediction model based on text features; this prior text-based model did not utilize any acoustic or prosodic information about the speech signal. The baseline text-based word importance prediction model used in our analysis is Section 6.1 (also described in Kafle et al. [77],) and it uses pre-trained word embeddings and bi-direction LSTM units, with a CRF layer on top, to make a prediction for each word.

As discussed in Section 6.2, human transcriptions are difficult to obtain in some applications, e.g. real-time conversational settings. Realistically, text-based models need to rely on ASR systems for transcription, which will contain some errors. Thus, we compare our speech-based model and this prior text-based model on two different types of transcripts: manually generated or ASR generated. We processed the original speech recording for each segment of the corpus with an ASR system to produce an automatic transcription. To simulate different word error rate (WER) levels in the transcript, we also artificially injected the original speech recording with white-noise and then processed it again with our ASR system. Specifically, we utilized Google Cloud Speech\(^3\)

\(^3\)https://cloud.google.com/Speech_API
ASR with WER $\approx 25\%$ on our test data (without the addition of noise) and WER $\approx 30\%$ after noise was inserted. Given our interest in generating automatic captions for DHH users in a live meeting on a turn-by-turn basis (Section 1), we provided the ASR system with the recording for each dialogue-turn individually, which may partially explain these somewhat high WER scores.

The automatically generated transcripts were then aligned with the reference transcript to compare the importance scores. Insertion errors automatically received a label of low importance ($\text{li}$). The WER for each ASR system was computed by performing a word-to-word comparison, without any pre-processing (e.g., removal of filler words).

<table>
<thead>
<tr>
<th>Model</th>
<th>ACC</th>
<th>F1</th>
<th>RMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>speech-based</td>
<td>63.72</td>
<td>57.58</td>
<td>68.21</td>
</tr>
<tr>
<td>text-based</td>
<td>77.81</td>
<td>73.6</td>
<td>54.0</td>
</tr>
<tr>
<td>$+ \text{ wer: 0.25}$</td>
<td>72.30</td>
<td>69.04</td>
<td>65.15</td>
</tr>
<tr>
<td>$+ \text{ wer: 0.30}$</td>
<td>71.84</td>
<td>67.71</td>
<td>68.55</td>
</tr>
</tbody>
</table>

Table 6.4: Comparison of our speech-based model with a prior text-based model, under different word error rate conditions.

**Result and Analysis:** Given the significant lexical information available for the text-based model, it would be natural to expect that it would achieve higher scores than would a model based only on acoustic-prosodic features. As expected, Table 6.4 reveals that when operating on perfect human-generated transcripts (with zero recognition errors), the text-based model outperformed our speech-based model. However, when operating on ASR transcripts (including recognition errors), the speech-based models were competitive in performance with the text-based models. In particular, prior work has found
that WER of $\approx 30\%$ is typical for modern ASR in many real-world settings or without good-quality microphones \cite{8,93}. When operating on such ASR output, the RMS error of the speech-based model and the text-based model were comparable.

### 6.2.7 Limitations of this Research

In contrast to prior work, which had depended on text-based features, we have proposed a neural architecture for modeling prosodic cues in spoken messages, for predicting word importance. Our text-independent speech model had an F-score of 56 in a 3-class word importance classification task. Although a text-based model utilizing pre-trained word representation had better performance, acquisition of accurate speech conversation text-transcripts is impractical for some applications. When utilizing popular ASR systems to automatically generate speech transcripts as input for text-based models, we found that model performance decreased significantly. Given this potential we observed for acoustic-prosodic features to predict word importance our continued work involves combining both text- and speech-based features for the task of word importance prediction.

### 6.3 Text- and Speech-based Importance Model

While prior models of word importance considered text features \cite{77,157}, speech-based features hold promise when analyzing conversational speech \cite{74}. Speakers often use prosodic cues to help listeners discern spoken messages; however,
these cues are omitted from an automatically generated text transcript [44].
Automatically generated transcripts may also lack capitalization or punctuation
or use nonstandard grammar, and they contain more speech disfluencies, such
as hesitations, filler words, out-of-vocabulary words, and neologisms than in
formal writing.

We therefore investigate how to fuse acoustic-prosodic features from speech
with lexical features from transcripts, in order to achieve a more holistic
representation of a spoken word for the task of word importance prediction.
This work proposes and evaluates an effective attention-based early-feature
fusion strategy. We also demonstrate how strategic supervision of the learned
attention-weights during training can help our model achieve better performance
on the importance prediction task. We evaluate our method with experiments
on the Word Importance Annotation corpus (described in Section 5.2) [77],
comparing its performance to state-of-the-art methods. Further, we visualize
the connotative variation in the fused representation of spoken words in different
spoken contexts. We also release pre-trained models at https://github.com/
SushantKafle/feature-fusion-word-importance.

6.3.1 Prior Work on Joint Modeling of Speech and Text

Joint modeling of lexical and prosodic features has benefited various applica-
tions, such as constituent parsing of conversational speech texts [163], and
summarization of recordings of meetings [121, 174]. The most common strategy
for joint representation of features is through concatenation. Despite the popu-
larlarity of this strategy, it has been shown to fail to fully capture cross-modal
interactions [104,178]. Consequently, several multimodal feature representation strategies have been proposed for various applications [104,144,170,178]. This work continues this line of research by investigating multimodal feature representation strategies for spoken words, as evaluated on the task of word importance prediction. Further, we aim to design a better feature-fusion strategy that exploits strengths (and weaknesses) of our unimodal features, and uncover modality-specific challenges in the prediction task.

The word importance prediction problem has similarities to familiar natural language problems, like keyword identification or summarization, where the goal is to identify a set of descriptive words from a large document of text. Several methods have been proposed, including frequency-based models like Term Frequency-Inverse Document Frequency (TF-IDF), and word co-occurrence measures [61,111], with a goal of extracting relevant keywords from a text. Other supervised measures of keyword extraction have been proposed [69,100,102,156] for a range of applications. All of these methods, however, consider the importance of words at a document level rather than at a sentential or a phrase level – limiting their generalizability to applications that consider word importance at a more granular level, e.g. [76].

Importance prediction of words in sentences requires consideration of both the lexical nature of the word and also its context of use. This differs from traditional setups that treat each word as a term in a document such that all words identified by a term receive a uniform importance score, without regard to context. In previous section, we have discussed several models that consider contextualized word representations [74,77]. However, as discussed in Section
6.3.2 Lexical-Prosodic Feature Representation

Our work considers two modalities of speech to obtain a feature representation of a spoken word $Z_i$: the acoustic-prosodic signal and the textual transcript. Rather than considering these two modalities as independent observations of speech, we focus on their cross-modal interaction to obtain a unified representation. We recognize that non-verbal cues during face-to-face communications contribute to influencing how humans understand spoken words [170]. Prosody is one such channel in spoken dialogue that is important in conversational speech, where speakers attach prosodic prominence to words (or sub-word components) to help listeners disambiguate meaning [125,136,168]. We investigate an attention-based feature fusion architecture that considers the effect of prosodic cues on the lexical meaning of a spoken message.

Speech Feature Sub-network

Every utterance has a unique phonetic (or phonological) realization which may differ from its lexical form. These phonetic variations often encode information about the organization of the utterance [163], as well as its relation to its context [83]. Our speech-feature sub-network (described in Section 6.2.1) aims to learn a feature representation for a spoken word that encapsulates this information.

As described in Section 6.2.1, we utilize a bi-directional recurrent neural
network (RNN)-based model to represent variable length spoken words into a fixed-length vector. Our network operates over the spoken words independently using word-level timestamp information. Each word region in speech is first partitioned into fixed-length sub-word intervals \( w_i \approx [a^i_1, a^i_2, ..., a^i_T] \) and passed as a sequential input to our RNN, as:

\[
\begin{align*}
\overrightarrow{h}_t &= \text{RNN}(a^i_t, \overrightarrow{h}_{t-1}) \\
\overleftarrow{h}_t &= \text{RNN}(a^i_t, \overleftarrow{h}_{t-1})
\end{align*}
\]

where \( a^i_t \) represents the sub-word interval segment of word \( w_i \), and \( \overrightarrow{h}_t \) and \( \overleftarrow{h}_t \) refer to the RNN hidden states at time \( t \). Finally, two RNN layers operating over the sub-word interval sequence in opposite directions summarize the interval-level features into a word-level representation \( S_i = [\overrightarrow{h}_T; \overleftarrow{h}_T] \).

**Attention-based Feature Fusion**

The goal of our attention-based feature fusion network is to capture the influence of prosody on the lexical semantics of the spoken word. Formally, our model uses an attention architecture dependent on both lexical and prosodic features in order to learn a composition vector that controls the contribution of prosodic features on the semantics of a word:
where $S_i$ and $E_i$ represent the speech-based and lexical representation of the word, and $h_{s_i}$ represents the non-linear projection of $S_i$, such that $\text{dimension}(h_{s_i}) = \text{dimension}(E_i)$ to facilitate composition. $W_1$, $W_2$ and $b_1$, $b_2$ are the weight and bias vectors to be learned during training, and $Z_i$ represents the final feature representation which is the weighted sum of the lexical and prosodic features using the attention weight vector $\alpha_i$. The intuition is to learn an appropriate composition vector $(\alpha_i \cdot h_{s_i})$ that can be used to project lexical embeddings into an appropriate semantic space, based on their prosodic character. This results in a meaning representation that considers both the lexical and prosodic meaning in combination.

**Attention Supervision**

Since we are using the attention-based weight vector to regulate the prosodic influence, we can also supervise the attention vector to match an expected distribution, to help with convergence during training. Supervising attention weights has been found useful previously [40,103,124], enabling the incorporation of heuristic constraint into a model. Here, we supervise attention weights to rely on prosodic features when the word is an out-of-vocabulary (OOV) word,
as shown in Equation 6.8.

\[
\tilde{L} = L + \lambda \begin{cases} 
\sum w_i - \log(|\alpha_i|), & \text{if } w_i \notin V \\
0, & \text{otherwise}
\end{cases}
\]  

(6.8)

where \( L \) represents the training loss and \( \tilde{L} \) represents the new loss (with regularization constraint as determined by \( \alpha_i \)) that the model is optimizing, \( w_i \) is the word, \( \alpha_i \) represents the attention weights for the prosodic features of the word, and \( V \) is the vocabulary of the model. Additionally, \( \lambda \) is the loss weighting factor, such that if \( \lambda = 0 \) no supervision will be enforced.

The negative log-likelihood loss will encourage the model to assign higher absolute weights to the speech features (\( \alpha_i \)), meaning higher reliance on speech features for the prediction. The motivation behind this supervision technique is discussed in Section 6.3.5 – where we found that in general, prosodic features are less prone to OOV errors as compared to text-based lexical features.

### 6.3.3 Experimental Setup

**Dataset**

We used the Word Importance Corpus (described in Section 5.2) for the training and evaluation of our word importance prediction models [77]. It consists of over 25,000 unique words (types), and each token has been manually annotated with importance information. The annotation covers a subset of conversations in the Switchboard corpus [51], which consists of about 25,048 utterances spoken by 44 different English speakers, with word-level timestamp information and a
numeric score of importance (in the range of $[0, 1]$) assigned to each spoken word. We created an 80%, 10% and 10% split of the corpus for training, validation, and testing. All the experiments were set up such that each speaker is only present in one of the data partitions. Otherwise, models trained and tested on the same set of speakers might not be generalizable to unseen speakers.

Unimodal Representations

We make use of the 6-billion-token-based 300-dimension pre-trained GloVe [131] embeddings as our lexical representation for the word. To get a word vector representation for speech, we utilized the network described in Section 6.3.2. As an input to this model, we partitioned the spoken word into fixed-length sub-word intervals and extracted prosodic features that have been previously considered for modeling word importance. As described in Section 6.2.1, a total of 30 prosodic feature were considered, which included pitch-related features (20), energy features (22), voicing features (6) and spoken-lexical features (12). All of the features were normalized (z-normalized) on a per speaker basis to account for inter-speaker variations.

Comparison Models

We compared against models based on different multimodal feature representation strategies:

- **Concatenation (CONCAT) [101, 129, 163]**: The model creates a multimodal representation of words by simply concatenating the unimodal features at the word level.
• **Attention-based Weighted Sum (ATTN)** [119, 144]: Instead of concatenating the unimodal signals as alternative feature vectors, the model uses an attention network to decide how to combine the information for the final representation.

• **Tensor Fusion Network (TFN)** [178]: This strategy models both the modality-specific and cross-modal interactions by computing an outer product over a set of unimodal vectors (with an extra constant dimension 1) rather than just the concatenation. Lastly, using a high-dimensional weight vector the outer product is projected into the final multimodal vector representation.

• **Low-rank Multimodal Fusion (LMF)** [104]: Drawing from the success of TNF networks, LMF proposes a more efficient version: It has fewer learnable parameters and an efficient computational setup through decomposition of the high-dimensional weight vectors into lower rank factors. This allows the estimation of a multimodal representation directly from the unimodal representations and their modality-specific decomposition factors.

• **Recurrent Attended Variation Embedding Network (RAVEN)** [170]: The model considers the sub-word structure of non-verbal behaviors to learn a multimodal-shifted representation for words. The non-verbal behaviors may be inferred from different multimodal channels such as a visual and/or an acoustic signal; our work only considers the latter for comparison.

**Model Architecture and Training**

As the prediction model, we utilized a bi-directional LSTM-based sequence-labeling architecture of word importance prediction. The sequence of word
representations (both unimodal or multimodal) was processed by the bidirectionally moving LSTM layers, to obtain a contextual representation of the word at each time step. This representation was passed through the final projection layer (sigmoid) for word importance prediction.

We used Gated Recurrent Units (GRUs) [28] as RNN cell\(^4\) for our speech-based sub-network. We used a GRU cell of dimension 64, and each word-level LSTM unit was of size 128. The lexical \((E_i)\) and speech \((S_i)\) dimensions were 300 and 30 respectively. All our models were trained to minimize the Root Mean Square (RMS) loss. For attention supervision (described in Equation 6.8), we found a loss weighting factor \((\lambda)\) of 0.8 to be best-suited for our task. For our comparison models, we used the best working setup based on their performance on the validation split. We used Adam optimizer with an initialized learning rate of 0.001 for training. Each training batch had a maximum of 20 sentences, and the model was trained until no improvement was observed in 7 consecutive iterations. A dropout of 0.5 was applied at the input layer for all models.

**Evaluation Metrics**

To compare the various models, we evaluated their predictions on word importance with the test set of the Word Importance corpus, described in Section 6.3.3. We used the RMS error as the primary measure of performance, comparing predictions against the gold standard corpus. As described by [77], the annotators of the corpus were asked to consider three ordinal ranges \{LOW: [0 -

\(^4\)We used GRU rather than LSTM units due to better performance observed during our initial set analysis.
0.3), mid: [0.3, 0.6), hi: [0.6, 1.0}] when they selected a numerical value in the range [0 - 1] to represent the semantic importance of each word. Thus, we also compared the performance of the models at predicting the importance of words belonging to each of these ordinal ranges. Further, we used the Kendall-Tau ($\tau$-b) correlation measure to compare the rank distribution of words, according to their predicted and their actual importance in a dialogue-turn. We report mean results in percent from 5-fold cross-validation evaluation.

### 6.3.4 Experiment 1: Error Analysis of Unimodal Models

The performance of the two unimodal-feature (lexical and speech) models in Table 6.5 indicates that although the model based only on lexical text features had a lower RMS error when predicting the importance of words in our test dataset, it performed poorly when operating over OOV words, as compared to the unimodal model based on speech features only.

<table>
<thead>
<tr>
<th>Models</th>
<th>RMS</th>
<th>RMS (OOV words only)</th>
</tr>
</thead>
<tbody>
<tr>
<td>prosodic-only</td>
<td>21.5</td>
<td>27.0</td>
</tr>
<tr>
<td>lexical-only</td>
<td>16.84</td>
<td>27.35</td>
</tr>
</tbody>
</table>

Table 6.5: Comparative performance of lexical and prosodic unimodal models. RMS column represents the overall RMS score, whereas RMS (OOV words only) represents the RMS deviation of the prediction on OOV words only. Bold font shows the best scores.

Each word in the Word Importance Corpus is annotated with an importance score between 0 and 1 [77]. Error analysis revealed that the lexical-only model trained on transcripts had a lower percentage of highly deviated predictions
CHAPTER 6. SUPERVISED WORD IMPORTANCE MODELS

<table>
<thead>
<tr>
<th>Models</th>
<th>RMS</th>
<th>RMS (OOV words only)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CONCAT</td>
<td>15.64†</td>
<td>23.20†</td>
</tr>
<tr>
<td>ATTN</td>
<td>16.08</td>
<td>23.84</td>
</tr>
<tr>
<td>TNF</td>
<td>17.14</td>
<td>29.08</td>
</tr>
<tr>
<td>LMF</td>
<td>16.59</td>
<td>27.02</td>
</tr>
<tr>
<td>RAVEN</td>
<td>17.0</td>
<td>28.5</td>
</tr>
<tr>
<td>Proposed (λ = 0)</td>
<td>15.80</td>
<td>23.65</td>
</tr>
<tr>
<td>Proposed (λ = 0.8)</td>
<td>14.75*</td>
<td>21.71*</td>
</tr>
</tbody>
</table>

Table 6.6: Comparison of different models combining lexical and prosodic cues. Per column, the top two results are marked with † & symbols, respectively. Our proposed model demonstrates lower RMS error both overall as well as for OOVs specifically.

(cases where the importance-score prediction differed from ground truth by more than 0.2, as determined by inspecting errors), compared to the speech-only model (18% vs. 26%). However, the lexical-only model was less robust for OOV words. Such words accounted for 49% of the highly deviated errors from the lexical-only model, compared to 27% from the prosody-only model.

6.3.5 Experiment 2: Comparison of Fusion Strategies

The difference in performance in the previous experiment between the two unimodal-feature-based models on OOV words inspired the design of our fusion strategy for a new integrated prosodic and lexical representation. Since the speech-based model showed better performance on OOV words (lower percentage of highly deviated errors) compared to the text-based model, we investigated encoding this as our feature combination heuristic, as shown in Equation 6.8 above.
CHAPTER 6. SUPERVISED WORD IMPORTANCE MODELS

Table 6.7: Comparison of models on ordinal-range classes, and Kendall-tau (τ-b) rank-prediction correlation. The top two results per column are marked with ⋆ & † symbols. Our proposed model performs better for high and low importance words.

<table>
<thead>
<tr>
<th>Models</th>
<th>RMS (across ranges)</th>
<th>τ-b</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HI</td>
<td>MID</td>
</tr>
<tr>
<td>CONCAT</td>
<td>21.81†</td>
<td>13.07†</td>
</tr>
<tr>
<td>ATTN</td>
<td>25.87</td>
<td>13.44</td>
</tr>
<tr>
<td>TFN</td>
<td>26.0</td>
<td>13.71</td>
</tr>
<tr>
<td>LMF</td>
<td>27.56</td>
<td>13.53</td>
</tr>
<tr>
<td>RAVEN</td>
<td>29.04</td>
<td>12.50*</td>
</tr>
<tr>
<td>Proposed (λ = 0)</td>
<td>25.13</td>
<td>13.29</td>
</tr>
<tr>
<td>Proposed (λ = 0.8)</td>
<td>22.4*</td>
<td>13.27</td>
</tr>
</tbody>
</table>

Tables 6.6 and 6.7 summarize the performance of all models. Notably, the results in Table 6.6 show that our feature-fusion strategy with attention supervision achieved the lowest RMS error compared to a range of other comparison models. In addition, Table 6.7 reports the performance of models when the predicting importance of words belonging to different importance categories (low, mid, high). Notably, our approach was better at identifying the high and low importance words, i.e., the two edges of the importance scale, in dialogue turns. Intuitively, this relates to natural speech patterns – a speaker is likely to render essential words more prominently than low-importance words, and accordingly this is when prosodic features can be most effective in modeling word importance. Further, higher τ−b scores indicate that our model is also better at capturing the overall rank distribution of words in a turn.

Table 6.6 also shows that while our proposed model is better at predicting the importance of OOV words in comparison to the other models, the performance
on the OOV words is generally poor, as indicated by higher RMS scores for these words. This performance gap shows the need of additional research on improving the generalizability of these models.

Further, Table 6.7 also demonstrates that the performance of the model drops when making predictions on the high importance words. We hypothesize that high importance words generally include proper nouns (e.g., names of the people and places) that tend to be OOV words as well – which could partially explains the high RMS scores for these words.

6.4 Conclusions

With the help of the Word Importance Annotation corpus described in Chapter 5, this chapter demonstrated different types of word importance prediction models ranging from unimodal-feature-based models (e.g. text-based models or speech-based models) to multimodal-feature-based models that utilize both the text-based features and the speech-based features. We evaluated all of these models using the test-set of the Word Importance Annotation corpus, and discussed their efficacy in predicting the importance of words in spoken dialogues.

We have shown that by incorporating features from speech-based features (acoustic-prosodic representation) into the text-based features (lexical embeddings), we can enhance the performance of word-importance prediction systems, in comparison to text-only and speech-only feature based models. Specifically, we developed an attention-based multimodal feature-fusion strategy that learns
to adjust the text-based feature representation (e.g., pre-trained GLoVE embeddings [131]) of spoken words to reflect the post-lexical meaning conveyed through prosody (or other acoustic-prosodic cues in speech).

By analyzing the individual performance of the unimodal-feature-based models, we were able to identify their strengths and weaknesses. For instance, we found that text-based models of word importance show better performance (lower RMS in test-set) compared to speech-based models of word importance. However, when operating over OOV words (words in test-set that are not observed during training), the performance of the text-based models dropped significantly, in comparison to the speech-based models. We hypothesized that because the text-based features are directly dependent on the lexical property of the word, an OOV word poses a generalization issue. However, since the speech-based features are lexically invariant, they could therefore be better at handling these cases.

Consequently, we were able to demonstrate that by incorporating these modality-specific heuristics into training of our multimodal feature based models, we can help our model perform even better. Specifically, we showed that our multimodal-feature-based model achieves the lowest RMS score on the word importance prediction task, compared to other state-of-the-art models.
EPILouGEl FOR PaRT I

Part I of this work described our investigation into the task of predicting the importance of words in spoken dialogues, especially when realizing application for the DHH readers. With this aim, we explored several strategies of word importance modeling, which included: unsupervised methods of word importance prediction through the estimation of the predictability of words given their context and, various supervised models that explore other linguistic features from spoken dialogues for estimating the word-level importance. For our predictability measure, we took inspiration from studies done in understanding the reading strategies of DHH readers, and we evaluate this method of importance prediction in an automatic caption evaluation task (described in Part II of this work) for DHH users.

To summarize, Part I of our work looked to answer the following research question:

RQ1: How can we identify words in a spoken message that are important to its understandability for DHH readers? We presented several research works that describe our methods for modeling the importance of words
in spoken dialogues, based on various features from text and/or speech. We
sub-divided this RQ into sub-problems that we investigated:

**RQ1.1:** Does measuring the predictability of word (given context)
help measure the importance of that word when focusing on
applications for DHH users? (We described our word predictabil-
ity based importance estimation of words throughout Chapter 4. We
discuss evaluation strategies for these methods for providing applica-
tions for DHH users in Section 4.3, with actual evaluations described
in Section 9.4.1.)

**RQ1.2:** Do supervised models based on textual features from a spo-
ken language transcript accurately predict word importance?
(In Chapter 6.1, we described our text-based word importance model.
Section 6.1.2 reflects the performance of the model which is well above
the probabilistic baseline of a random model.)

**RQ1.3:** Do acoustic-prosodic cues in spoken dialogues help identify
important words in the dialogue? (We showed that acoustic-
prosodic cues in the human speech were, in fact, useful in predicting
the importance of words. Section 6.2 shows the results of this analysis.)

**RQ1.4:** Do models trained on both textual and speech features from
spoken dialogues outperform word importance models trained
on a single type of feature? (We showed that combining text-based
features (lexical representation) and speech-based features (acoustic-
prosodic representation) improves the prediction of the importance of words. Section 6.3 shows the results of this analysis.)
Part II: Automatic Caption

Quality Evaluation
When researchers are considering using an ASR system for a captioning task, they must evaluate the quality of the system in that environment to decide if it will work well. Unfortunately, the metrics currently available for this evaluation are too simplistic: Currently, researchers who evaluate ASR performance often focus on improving the Word Error Rate (WER) metric, however, it has been found to have little correlation with human-subject performance for many applications. Thus, Part II of our work describes the development of several new captioning-focused evaluation metrics for predicting the impact of ASR errors on the usability of automatically generated captions for people who are DHH. We design experimental studies with DHH users where we collect their subjective judgments on the quality of automatically generated captions. With the collected data, we design and evaluate various metrics of ASR quality that are more in-tune with the actual human judgments in captioning applications for DHH users.

In order to create better metrics for estimating ASR quality, Part II of work utilizes the word importance prediction models discussed in Part I of
this research. The word importance models help estimate the quality of the automatic captions by evaluating whether importance-bearing words in a spoken dialogue are recognized by the captioning systems. With this strategy of evaluating ASR systems’ quality, the new metrics of ASR quality aim to predict the understandability of the automatic captions rather than just counting the number of errors in them.

This part of the document is organized as follows: The survey of prior work in Chapter 7 provides motivation for this work, highlighting the limitations of the current metrics for evaluating the quality of ASR systems for various applications. Chapter 8 provides our research in understanding (and collecting the data on) the effect of various kinds of errors on the understandability of text for DHH users. Next, we present our work in designing and validating a metric for evaluating ASR-based captioning systems for DHH users in Chapter 9.

More precisely, Part II our work will explore the following research question RQ2 (as presented in 1.2), which states:

**RQ2:** Do our models of estimating the quality of ASR systems for generating captioning for DHH users accurately predict the quality of the output? To answer this research question, we formulate additional sub-research questions that we will answer in Part II of our research:

**RQ2.1:** In a simulated two-person business meeting scenario, does our metric for predicting quality of automatically generated captions for DHH users outperform the WER metric? (We
RQ2.1: In a simulated business meeting scenario, does our metric for predicting the quality of automatically generated captions outperform prior metrics published in the literature, in terms of correlation between DHH human judgments and the predicted score from the metrics? (We will examine RQ2.1 in Phase 1 of the work in Chapter 9).

RQ2.2: In a simulated business meeting scenario, does our metric for predicting the quality of automatically generated captions outperform prior metrics published in the literature, in terms of correlation between DHH human judgments and the predicted score from the metrics? (We will examine RQ2.2 in Phase 3 of the work in Chapter 9).

RQ2.3: In a realistic two-person conversational settings, would our metric for predicting the quality of automatically generated captions outperform the WER metric? (We examine RQ2.3 in Phase 4 of the work in Chapter 9).
Chapter 7

Prior Approaches to ASR Evaluation

7.1 Limitations of the Word Error Rate Metric

While WER is the most commonly used metric for evaluating speech recognition performance, researchers have argued for alternative evaluation measures that would better predict human performance on tasks that depend on ASR text output usability [112, 120]. There have also been concerns about the nature of the metric: Researchers have criticized that while WER has a lower bound of zero (indicating that a hypothesis text is a perfect match for a reference text), WER lacks a proper upper bound, making it difficult to evaluate WER scores in an absolute manner [112]. Further, researchers have also argued that WER is ideally suited to evaluation of ASR quality only for those applications in which the human can correct errors by typing, since the WER metric is
based upon counting errors – which directly relates to the cost of restoring the output word sequence to the original input sequence [112]. In other applications, researchers have observed a weak relationship between WER and human task performance. For example, in the task of spoken document retrieval (in which a human is searching for a speech audio file, which has been transcribed by ASR, by typing search terms for desired information), researchers have found that the WER of the ASR system has little correlation with the retrieval system performance [47, 56]. Moreover, other researchers [171] saw improvements in a spoken language understanding task, even during a significant increase in WER.

### 7.2 Other Methods of ASR Evaluation

Several researchers have proposed alternative metrics to WER for evaluating the performance of ASR for specific applications. Nanjo and Kawahara in [122] have weighted errors based on the Term Frequency-Inverse Document Frequency (TF-IDF) measure, in the context of a keyword-based open-domain speech understanding application. As introduced in Section 3.1.1, TF-IDF is commonly used by researchers studying information retrieval; it assigns high scores to words that are generally rare but which appear in great frequency in a particular document, e.g., if a rare word like “daffodil” appears very frequently on a particular webpage, then it is reasonable to think that the word “daffodil” is an important keyword for that webpage. Specifically, these researchers have used TF-IDF as a loss function during the decoding step of their ASR system [122].
(As discussed in Section 2.1.3, during the decoding process ASR system aims to determine the most likely sequence of words that corresponds to speech information.) The loss function penalized errors on keywords more heavily than errors on other words, when choosing from a list of output candidates. The authors explored using this metric as a weighting factor in a Boolean fashion (i.e. is something a keyword or a non-keyword) or by using the actual numerical TF-IDF scores as weights.

Garofolo et al. attempted to modify WER to weight content words more heavily than other words [47]. Generally speaking, content words include nouns, verbs, adjectives, and adverbs that convey semantic meaning, rather than function words, e.g. determiners, that convey grammatical information. The authors used ASR in an information retrieval application; users searched for excerpts in large spoken audio recordings. The authors found a nearly linear relationship between their proposed metric and retrieval performance across different systems: i.e., ASR systems that recognized content words more accurately provided the best input for their retrieval task. To summarize, both Garofolo et al. [47] and Nanjo and Kawahara [122] found that keyword identification (to differentially weight specific kinds of errors) led to useful ASR metrics for applications related to information search.

Some researchers have considered applications of ASR that are even closer to our focus on automatic captioning: For instance, some have proposed a metric for evaluating ASR output on a speech transcription task [116]; their metric was based on opinion scores collected from humans who judged the quality of ASR-generated voicemail-to-text transcripts. Scores from their metric correlated
with the human judgments better than WER did. Their metric learned the cost of different error types (namely, insertion, deletion, and substitution) and learned a weight factor called the saliency index for words to predict their contribution in text understandability. While not focused on creating a fully automatic metric, Apone et al. [2] investigated different categories of captioning errors (e.g., substitution of a word with an incorrect tense) and weighted each category to design a weighted WER metric. This metric was proposed for evaluating the accuracy of captions for television.

The match error rate (MER) and word information loss (WIL) metrics were introduced by Morris et al. [120], as replacements for WER in settings where high error rates are common. The MER metric is similar to WER except that it is properly normalized and thus computes the probability of a given match (between the reference text and the hypothesis text) being incorrect. Similar to MER, WIL is a probabilistic approach that approximates the proportion of the word information lost due to the presence of errors.

Our work is also inspired by the work of McCowan et al. [112], which discussed the challenges of application-oriented evaluation of ASR systems and proposed a generic framework to evaluate the ASR output based on information retrieval concepts like precision and recall. Their framework treated the speech recognition task as analogous to an information retrieval task, i.e. the goal for transcription is to retrieve all the relevant information (i.e. the spoken word) from in the original speech signal. In their framework, they provided room to incorporate application-dependent importance weights for words and for different ASR error types. However, for our application of real-time captioning
for DHH users, the assumptions made in their framework are less appropriate: They treated words as independent units of information, without considering their position in a sentence, i.e. under this assumption, identical words located at different positions in a sentence will have identical weights. Later, in Chapter 9, we provide a comparison of performance these metrics on a caption usability prediction task for DHH users.

In contrast, in Chapter 9.1, we propose a new captioning-focused evaluation framework called the Automatic-Caption Evaluation (ACE) framework to accurately model the impact of an error in the understandability of a caption-text. To measure the impact of an error in a caption-text, the framework considers the importance of words and the semantic deviation due to the error. With the help of this framework, this chapter will discuss the design of several caption quality evaluation metrics and provides evaluations of their performance, through studies with DHH users.

### 7.3 Metric of ASR Quality for DHH users

The premise of our ASR evaluation approach is that rather than simply counting the number of errors, it would be better to consider which words are incorrect or where they occur in the sentence when evaluating ASR text output for captioning applications for DHH users. As discussed in Section 7.2, some researchers have previously examined the limits of the WER metric and have considered some alternatives. Our research is novel in that we are specifically interested in measuring the quality of ASR output for a captioning application
for DHH individuals, and we evaluate our proposed metric in a user-study with DHH participants.

There are reasons to believe that it is important to create and evaluate metrics for measuring ASR output quality specifically targeted for DHH users. Anecdotally, some accessibility researchers have argued that ASR-generated errors on captions are more comprehension-demanding than human produced errors [7, 89]. Further, prior research has characterized differences in literacy rates and reading mechanisms between DHH readers and their hearing peers: Standardized testing in the U.S. has measured lower English literacy rates for deaf adults [72, 106]. Furthermore, literacy researchers have hypothesized that the basic mechanism employed by many deaf adults to understand written sentences differs from that of hearing readers: Specifically, deaf readers may identify the most frequent content words and derive a complete representation of the meaning of the sentence, ignoring other words [10, 34]. This reading strategy is often referred to as a keyword strategy, and it suggests that a subset of the words in a caption text might be of very high importance to DHH users (for text understandability). Following this same reasoning, it might be disadvantageous to penalize each error in a caption text equally. Some errors may be very consequential to the understandability of the text (with the potential to mislead or confuse the readers), while other errors may have little impact (perhaps easily ignored by readers). Our goal is to develop a metric that can predict the quality of an ASR text output based on the usability of the text as a caption for DHH users. Unlike WER, we want our metric to distinguish between harmful errors in the caption (likely to degrade the quality
of caption for DHH users) and less harmful errors; the metric should use this distinction when penalizing a text for each type of error.

With this aim, we seek to identify an ASR evaluation metric that is more captioning-focused, to measure the impact of errors on the understandability of a caption for DHH users. Specifically, we investigate a new ASR evaluation metric that considers the importance of the spoken word for understanding the meaning of the spoken message – and the semantic deviation in the meaning due to each error. In the coming sections, we will discuss how we design such a metric and how we evaluate it in a caption quality evaluation study with DHH users.
Chapter 8

Collection of Understandability Scores from DHH users for Text with Errors

A key to evaluating the quality of automatically generated caption output is to understand what impact an error has on the understandability of the text. To gather evidence of different types of errors in text and their impact of text understandability of DHH users, this chapter presents our work in the design and setup of our data-collection study for this purpose.

8.1 Understanding the Effect of Recognition Errors

With the motivation to study the effect of different ASR-generated errors on the understandability of a text for DHH users, we formulate a user study with
DHH users when given imperfect English texts (containing ASR errors) and asked to answer some questions requiring inferences from the text. The data collected from the user study helps us understand the relationship between ASR errors and the impact they have on the understandability of a text for DHH users. In later sections, we discuss how this data is used for designing metrics of ASR systems quality for captioning applications for DHH users.

8.2 User Study (QUESTION-ANSWER STUDY)

We performed a user study with a goal of understanding how ASR errors affect DHH users’ performance on a comprehension task, given that a text contains some ASR generated errors. In this study, users were presented with imperfect English text passages (containing artificially inserted errors, actually observed on real ASR errors for that passage) and were asked to answer questions that required understanding the information content of those passages. Based on the answers, we collected Comprehension Scores for the respective questions, which we subsequently used to model the relationship between errors in the text and its comprehensibility. This study has been approved by the Institutional Review Board (IRB) on October 28, 2015. Principal Investigators: Michael Stinson and Matt Huenerfauth.

8.2.1 ASR Error Category

Errors in the output of an ASR system may differ widely as to how much they affect the user’s ability to understand the message. For all users, one
can imagine that if a semantically important word was incorrectly identified by the ASR system (e.g. the word *not* was missed in a sentence), then the meaning may be strongly affected. For DHH users specifically, given differences in literacy skills, it may be the case that the way in which ASR errors affect their comprehension of a text will differ from other users.

To guide our creation of stimuli for the user study, we established a hierarchical classification of various sub-types of ASR errors in the context of temporal information about the alignment of the gold-standard. Broadly, ASR errors can be categorized into three types: substitution, deletion and insertion errors. Further, we divided substitution errors into four types: one to one substitution, one to many substitution, many to one substitution and many to many substitution. One to one substitution refers to the errors when one word is substituted by the other. One to many substitution errors are the error due to substitution of one word by many (for e.g., *undistinguished* substituted by *on distinguished*). Similarly, many to one errors are the errors when many words are substituted by a single word. Many to many errors corresponds to a multi-word span of text in the reference transcript with inaccurate recognition such that none of the word boundaries within the span align with those within the corresponding span of ASR output. We further sub-categorized one to one substitution errors into three types namely, morphologically similar substitution, phonetically similar substitution and remaining other types of substitution errors. The morphologically similar errors are the errors where the actual word is substituted by another word with an inflectional or derivational morphological relationship to the first (for e.g., *developed* substituted by *develop*). The
CHAPTER 8. COLLECTION OF UNDERSTANDABILITY SCORES

phonetically similar errors are the errors due to the substitution of a word by another word with similar phoneme representation; for example, the words *table* (T EY B AH L) and *stable* (S T EY B AH L) have a very close (≥ 60% match) phoneme structure so they are considered as a phone neighbor of each other.

These categorization of different error types were meant to serve as a coarse categorization of the errors and was used as a basis for ensuring that the stimuli presented in our user study contained a good mixture of different error types.

8.2.2 Study Resources

For the user study, we created a dataset of around 20 short passages, with each passage containing three sentences marked as our Region Of Interest (ROI). For example, the text below shows a sample text passage used in the study with three bold sentences representing the three ROIs in the text.

*People who study film music often complain about the lack of recognition their field receives. The study of film music is an interdisciplinary field, falling in between cinema studies and musicology. This is one of the reasons why it receives so little attention. For example, when film music scholars, who often do not have music-degree credentials on par with the pure musicologists, write about film soundtracks, their articles are often ignored by the musicologists. Conversely, when the work of film music scholars touches on the visual aspects of film, the cinema studies people often treat it as the work of amateurs. So with the members of the two fields most closely related to it ignoring it, it is easy to understand why*
members of the film music field feel a degree of frustration.

The questions for passages was designed in such a way that each question would only require inference from one of each ROI sentence in that passage. In total, each passage had three questions where each question was text explicit (TE); as described by Jackson et al. [72], TE questions measures exact recall from the text without requiring any inferential use of information from the reader's memory. For each ROI sentence, an average of 8 different variations was generated where each variation was produced by inserting at most one category of ASR error into the ROI sentence. To produce each variation of the ROI, we began with a perfect text and inserted one of those errors. The sentence below is one ROI sentence for which variations were generated:

Conversely, when the work of film music scholars touches on the visual aspects of film, the cinema studies people often treat it as the work of amateurs.

We produced different variations of this ROI text by adding ASR generated errors in the sentence. ASR generated errors were collected by manually recording the ROI text (multiples times) and running it against the ASR system. Some variations of the ROI text are shown below:

• Conversely, when the work of film music scholars touches on the visual aspects of film, the cinema studies people often cricket as the work of amateurs.
• Conversely, when the work of film music scholars touches on the visual aspects of film, the cinema studies people often treat it as the work of amateurs.

• Conversely, when the work of film music scholars touches on the regional aspects of film, the cinema studies people often treat it as the work of amateurs.

• Conversely, when the work of film music scholars touches on the visual aspects of film, the cinema studies people often treat it has worked amateurs.

• Conversely, when the work of film music scholars touches on the visual aspects of film, the studies people often treat it as the work of amateurs.

• Conversely, when the work of film music scholars touches on the visual aspects of film, the cinema study people often treat it as the work of amateurs.

• Good whiskey, when the work of film music scholars touches on the visual aspects of film, the cinema studies people often treat it as the work of amateurs.

• Conversely, when the work of film music scholars touches on the visual aspects of film, the cinema studies people often treat it as the work of amateurs.

This procedure ensured that the artificially created variations of the ROI sentence agreed with the actual imperfect output produced by the ASR.
8.2.3 Recruitment and Participants

Participants for the study were recruited from among associate degree students at the National Technical Institute for the Deaf (NTID) at Rochester Institute of Technology (RIT). We collected data from 30 DHH participants with age distribution of ($\mu=22.63$, $\sigma=2.63$), including 12 men and 18 women, where 26 participants self-identified as Deaf and 4 of participants as Hard-of-Hearing.

8.2.4 Study Procedure

Each participant was given 10 different comprehension passages to read, each containing three multiple choice questions that needed to be answered in a time period of 70 minutes. A pilot test with a DHH member of our research team helped us to determine an appropriate number of question items for the 70-minute experiment. The comprehension passages given to the participants were generated by replacing each ROI sentence by its erroneous counterpart (one of the variations). The number of errors of each category that were displayed to each participant was balanced among all participants in the study to ensure that individual human differences in task performance did not disproportionately affect the scores for any one category of error. Further, each ROI appeared several times throughout the entire study in a form without any errors inserted so that we could obtain baseline measurements for the difficulty of the particular comprehension question, to enable subsequent normalization of the collected scores. Scores of answers from each question were binary with correct answer receiving the Comprehension Score of 1 and incorrect answer receiving the score
8.3 Summary of the data

To summarize, we described our method for collecting data on the effect of speech recognition errors on text understandability for DHH users. For the purpose, we designed a study where we asked our DHH participants to read imperfect English text paragraphs (containing recognition errors) and asked comprehension questions the required understanding of certain portions of text in the imperfect English paragraph. The participants’ would get a score of 1 if they answered the question correctly and 0 if they didn’t answer the question correctly. The average score received by the participant in the task represented the average impact of the recognition errors on the understandability of the text. However, to keep things tractable we only introduced one error per sentence in the text paragraph and each question asked were text-explicit, meaning they only required understanding a single specific sentence in that paragraph for answering the questions – which would be the stimuli of our study with the recognition errors. Therefore, our data collected included English sentences with recognition errors and the understandability score from DHH participant for each of those sentences.

In the coming sections, we will see how such data will be useful in building a caption-quality evaluation metric, which is the overall goal of Chapter 9. Section 9.1.3 and 9.4.1, in particular, will make use of this data for the development of a metric of ASR quality for DHH users.
Chapter 9

Metric for ASR Evaluation for Captioning Applications

9.1 Automatic-Caption Evaluation Framework

As discussed in Section 7.3, we are interested in the potential for ASR systems to be used as a real-time captioning tool for impromptu meetings. There are many commercial and research ASR systems available, each with different capabilities, e.g., adapting to the voice of specific speakers, operating in contexts with different types of background noise, or recognizing different vocabulary or genres [60,96,98]. A natural question is how to compare ASR systems to determine their suitability for use in this context.

Given the limitations of WER discussed in Section 7.1, we therefore present a new framework, called the Automatic-Caption Evaluation (ACE) framework, which aids the design of a better evaluation metrics for carefully assessing
the efficacy of these tools. The framework considers two primary factors for evaluating the impact of an error in a caption text: (a) the importance of the spoken word (reference) in understanding the meaning of the message and (b) the semantic deviation between the error word and reference word. These two factors are used to predict the impact of an error in a caption text as follows:

\[ I(w_r, w_h) = \alpha \times IMP(w_r) + (1 - \alpha) \times D(w_r, w_h) \]  

(9.1)

where the \((w_r, w_h)\) pair represents a recognition pair obtained after comparing (aligning) the automatic caption text with the actual human transcription of the spoken message, such that \((w_r \neq w_h)\). \(IMP(w_r)\) represents the importance score of the reference word \((w_r)\) in the meaning of the spoken message, \(D(w_r, w_h)\) represents the semantic distance of the aligned pair \((w_r, w_h)\), and \(I(w_r, w_h)\) represents the impact due to the error. Alpha \((\alpha)\) represents the interpolation weight, which determines how much each of the two factors (word-importance or semantic-distance) contributes to the overall impact score. In other words, the overall impact of an error is determined by the weighted combination of the importance of the reference word and the semantic distance between the error word and the reference word. The weighting factor is determined by the value of alpha \((\alpha)\).

It should also be noted that this framework operates on a per-error basis, meaning it considers single error at a time. However, as we discuss in later sections, the errors caused by ASR are not always isolated to a single word, but instead, an error may be best understood as having affected an entire phrase,
depending on how the alignment is performed. In the coming sections, we provide a more detailed explanation of the key components of this framework and how we use them to create different automatic caption evaluation metrics.

9.1.1 Word Importance Sub-score

The word importance sub-score measure attempts to quantify the semantic contribution of a word to the user’s understanding a text. Inspired by the reading strategies of deaf readers (discussed in Section 3.2), we had previously formulated other unsupervised and supervised measures of word importance, based on word predictability, which are discussed in Sections 4 and 6.

In this work, we are currently making use of the word predictability models presented in Chapter 4 as our word importance sub-score model.

9.1.2 Semantic Distance Sub-score

Misrecognition errors in automatic captioning systems may be identified by comparing the caption text with a human transcription of what was actually said by the human speaker. This comparison is typically conducted through a process called *text-alignment*. The semantic distance sub-score is used to measure the quality of an aligned-unit by measuring how far a prediction is from the actual message.

Notably, compared to the aforementioned word importance sub-score, the semantic distance sub-score considers the quality of the transcription itself, without regard to its importance in the context. For example, an error on an important word could be even more harmful to a text’s understandability, if
the erroneous word that is displayed (in place of the correct word) is especially misleading or confusing. The intuition is that when an incorrect word is shown, then we must not only consider whether the original word was important or not, but we also have to consider how far-away (in meaning) is the erroneously shown word, as compared to the correct word.

9.1.3 The Weighting Variable

The word importance sub-score (Section 9.1.1) and the semantic distance sub-score (Section 9.1.2) are combined using a weighted sum to produce an error impact score (as shown in Equation 1), but that equation requires us to select a tuning parameter alpha (\(\alpha\)) to specify how much each sub-score contributed to the overall error impact score.

Selecting the Weighting Variable

To learn the appropriate value of alpha (\(\alpha\)), we fit the value of this parameter, using a dataset of texts that have been labeled as to the overall understandability of each text. In this case, we calculated these understandability labels based on response data we had collected during a prior study with DHH participants [75]; this is referred to as the QUESTION-ANSWER study in Section 8.2. From this dataset we had previously conducted, we examined the subset of question-responses that corresponded to English sentences that contained ASR errors. This data was used to calculate an aggregate comprehension score for each sentence, by averaging the scores from the 30 participants on questions about that sentence.
9.2 Research Methodology and Hypotheses

While the goal of our overall project is to develop a metric that could automatically evaluate the usability of a caption for DHH users, we must select a more specific scope in order to design a study to evaluate the efficacy of this new metric. We have selected to focus on measuring the efficacy of ASR for providing captions during a business meeting between a hearing person (speaking English) and a DHH participant. The rationale for this focus is that impromptu one-on-one meetings in a workplace may be a situation in which it is unlikely for professional captioning services to be scheduled or available. Therefore, there is an opportunity for using automatic methods like ASR technology as a captioning tool.

9.2.1 Four Phases of this Research

Our research methodology consists of four phases, which are described briefly below and depicted in Fig.9.1. Each phase is described more thoroughly in Sections 9.3 – 9.6.

**Phase 1:** We present our first metric, called the ACE metric, based on the automatic-caption evaluation framework (described in Section 3) and evaluate its efficacy in predicting the quality of automatically generated captions for DHH users. We predict DHH users will subjectively prefer ASR text output that is predicted as less erroneous by our ACE metric (as compared to ASR text output predicted as being less erroneous by the traditional WER metric). Specifically, if we ask DHH users to evaluate
Figure 9.1: Graphical illustration of research activities presented in this article.

the quality of ASR text output, we wish to answer the following research question (RQ2.1):

**RQ2.1.** In a simulated two-person business meeting scenario, can we design a better metric for predicting quality of automatically generated captions for DHH users as compared to the WER metric?

To answer the question, we hypothesize the following:

- **H1:** If we compare ASR output texts predicted as better by WER (i.e. with low WER-to-ACE ratio) to ASR output predicted as better
by ACE (i.e. with high WER-to-ACE ratio), DHH participants will subjectively prefer texts preferred by ACE.

– H2a: The subjective preference judgments of DHH participants on ASR output texts will correlate significantly with ACE.

– H2b: There will be a significantly higher correlation between DHH human judgments and ACE, as compared to the correlation between DHH human judgments and WER.

**Phase 2:** We explore different ways to improve the ACE metric and evaluate different improvement strategies to propose an improved version of the metric, called the ACE2 metric.

**Phase 3:** We compare the performance of our metric with other published metrics in prior work on evaluating ASR text output or on caption-quality prediction for DHH users. More specifically, we look to answer the following research question (RQ2.2):

**RQ2.2.** In a simulated business meeting scenario, would our metric for predicting the quality of automatically generated captions outperform other metrics in the prior literature, in terms of correlation between DHH human judgments and the predicted score from the metrics?

**Phase 4:** We conduct a new user-study with additional DHH users and new text stimuli, to compare our original ACE and our new ACE2 metrics, where we answer the following research question (RQ2.3):
RQ2.3. In a realistic two person conversational settings, would our metric for predicting the quality of automatically generated captions outperform the WER metric?

To answer this question, we hypothesize the following:

– H3: If we present ASR output texts with varying ACE and ACE2 scores, there will be a significantly higher correlation between DHH human judgments and the ACE2 metric, as compared to the correlation between DHH human judgment and the original ACE metric.

9.3 Phase 1: Designing and Evaluating the ACE Metric

In the previous sections, we discussed the ACE framework and its various components. This section describes the methodological details of the first metric we develop with the help of this framework – we call it the ACE metric. Sections 9.3.1 – 9.3.3 describes the how different components of the metric are formulated and computed. Section 9.3.4 details on how we set up a user-study for evaluating the efficacy this metric with DHH participants. Section 9.3.6 discusses the results of our analysis, and section 9.3.7 discusses the limitations and future work.
9.3.1 Computing the Word Importance Sub-score

We discussed in Part I of our research the different word importance estimation models. Chapter 4, in particular, described the word predictability based word importance estimation in spoken dialogues.

As an initial investigation, in Phase 1 of our research we explore the efficacy of the n-gram language model based word predictability measure (see Section 4.2.1). Therefore, this section demonstration of n-gram language model based word importance estimation as our word importance sub-score in the ACE framework. Readers may refer to Section 4.2.1 for methodological details on the predictability of the word as estimated using this model.

![Figure 9.2](image.png)

Figure 9.2: Visual illustration of n-gram-based word-importance scoring, based on the predictability of words in the context of a sentence, with higher bars indicating less predictable words.

Fig. 9.2 provides a visual example of how this metric assigns importance scores to a text. As shown in the figure, some words in the example text are highly unpredictable (with higher entropy score) - like ‘omaha’, ‘winters’, etc.,
while some are fairly predictable - like ‘i’, ‘in’, etc. in the text.

In conclusion, for each error word, the entropy score of the corresponding reference word is calculated to estimate the word predictability score. For insertion errors, where there could be two adjacent reference words that could be responsible for the error, an average entropy score is computed based on the adjacent reference words.

### 9.3.2 Computing the Semantic Distance Sub-score

The second measure that we consider is the degree to which the meaning of a word in the output text differs from the meaning of the actual word that was spoken.

**Vector Space Representation of Words for Semantic Distance**

In order to compute this semantic disagreement between the error word and the actual reference word, we utilized a pre-trained word2vec model from Google\(^1\). The word2vec model provides a vector representation of a word which can be subsequently used in many natural language processing applications and research. In this work, we compute the semantic distance between words by computing the cosine distance between their word2vec representation.

As shown in Fig. 9.3, each error receives a semantic distance score between 0 – 1: for e.g., substitution of “winters” → “windows” receives a high semantic distance score of 0.874. For insertion and deletion errors, the length of the word is used to approximate the distance. For insertion and deletion errors,

\(^1\)https://code.google.com/archive/p/word2vec/
Figure 9.3: Visual Illustration of word2vec based semantic distance scoring of different alignment pairs (reference word $\rightarrow$ hypothesis word) in an example sentence. The height of the black bar indicates the semantic distance between the words.

the semantic distance was based on the length of the hypothesized word and the reference word, respectively. A scaling factor of 0.05 was used to get the semantic distance score ($0.05 \times \text{length}$). The scaling factor decided empirically from the analysis of errors from ASR system on a separate dataset.

9.3.3 From Individual-Error Impact Scores to an Overall Sentence Error Score

So far, we have defined our methodology for computing the word-importance and semantic-distance sub-scores which, together with the weighting variable alpha, is used to compute the impact score of an error in an automatic caption. To get the overall captioning quality score of a sentence$^2$, ACE makes use of these individual impact scores due to errors in the text.

One way to formulate our metric would have been to compute this score is by summing over the impact scores of each error and normalizing it by the

---

$^2$A sentence in this document refers to any linguistically complete unit of text or, in the context of conversational speech text, a single unit of spoken utterance (as defined in [164]).
length of the reference text, as had been done for WER in Eq. 9.2, e.g.

\[
\text{Version1}_\text{ACE} = \sum I(w_r(i), w_h(i)) / N
\]  \hspace{1cm} (9.2)

where is the impact of error on the usability of the captions and, is the length of the reference text. This style of formulation for a metric is quite common [117, 120, 160], and its value would be equal to WER if all the error impact scores were to equate to 1. However, this above formulation of the metric has some limitations:

- For each error in the hypothesized text, we compute an impact score by considering a single error at a time in the reference text – consequently ignoring any non-linear impact due to a cascade of errors (effect of an error in presence of others). Thus, a sum of impact scores of individual errors is insufficient for capturing the true impact of such situations.

- Since the impact score of each error lies between the 0 and 1, the metric will always be less than or equal to WER, i.e.

\[
0 \leq \text{Version1}_\text{ACE} \leq \text{WER}
\]  \hspace{1cm} (9.3)

This is because, (the length of the reference text) is not the true normalizing constant for our metric.

Due to these limitations above, we invented a new (and final version) of the ACE metric as follows:
\[ ACE = \max I(w_r, w_h) \frac{\log(N) - \log(n)}{\log(N)} \]  

where \( I(w_r, w_h) \) represents the impact due to an error \((w_r \neq w_h)\), defined in Equation 9.4. \( N \) is the length of the reference text, and \( n \) is the total number of errors in the hypothesized text.

With this setup, the total impact due to errors in a sentence unit is represented as the maximum score among all the individual impact scores for all errors in the sentence. The intuition here was that if a sentence contains a major error, the overall effect on the sentence understandability can also be major. There are limitations of this approach, e.g. extending this approach to longer, multi-sentential texts; Section 9.3.7 enumerates some of these limitations.

As the length of the reference text increases, it slowly mitigates the impact of individual errors – the rationale being that as readers have more context (more surrounding words), it is easier to decipher the text’s true meaning. But, if the number of errors increases with the reference text, the impact of errors is counterbalanced (note the subtraction of a \( n \) term in the denominator). Using a sub-linear function (log), the rate of this change is regulated such that the effect is not always linear (e.g., the effect of large \( n \) might not be linearly reduced by a larger \( N \)). Like WER, ACE is also an error measure, meaning that a lower ACE score indicates a better caption text. Similar to WER, ACE does not have an upper bound, but it would be trivially possible to modify the metric to prevent it from exceeding some limit, e.g. establishing a ceiling value of 1.

In summary, our intention when designing this new ACE metric was to
penalize ASR output texts that contain errors that are likely to lead to misunderstanding; specifically, the ACE metric considers errors at locations in a text that are less predictable and errors that deviate semantically from the actual word.

9.3.4 Designing Stimuli for Metric Evaluation (PREFERENCE-2017 Study)

We conducted a study (referred to as the “PREFERENCE-2017” study in Fig. 9.1) with DHH participants to evaluate whether our ACE metric correlated to the subjective judgments of these users as to the overall understandability of a caption text. To create texts to display in this study, we used some staged and prerecorded videos from colleagues at our lab [12]. These videos display one side of a two-person business meeting communication - the speaker leading the conversation in the video is made to look like he is interacting to the participant who is watching the video, as shown in Fig. 9.4.

We extracted the verbatim script of what the human actor said during the videos, and we used the entire text as a potential source of stimuli sentences for inclusion in this study (see Section 9.3.4 below). Next, we processed the original audio recording from these videos using an ASR system that we expected to make a large number of errors (it is important for our stimuli selection process for us to have many possible errors to choose from). For this processing, we used the CMU Sphinx 4 system with its off-the-shelf US English acoustic and language models which have been previously disseminated to the research community.
While a simplistic approach for creating stimuli for the study would have been to simply display the raw output of the ASR system to users, we were interested in obtaining judgments from participants on texts that had a variety of ACE metric scores. Furthermore, to investigate research question RQ2.1, we were interested in presenting users with some pairs of ASR text output that displayed multiple hypotheses (i.e. two different guesses from the ASR system about what it heard), with one of the texts having a low WER-to-ACE score ratio (indicating that WER believed the text to be good, but ACE did not) and the other with a high WER-to-ACE ratio. Since ASR systems actually consider a wide variety of hypotheses when they analyze a speech audio file (with one hypothesis correct, and the remainder containing some variety of errors), we wanted to search the space of ASR output candidate hypotheses to select texts
to display in our study with various WER-to-ACE ratios. The following section describes our procedure for identifying ASR output hypotheses to display in our study with diverse WER-to-ACE ratios. Rather than inventing artificial errors to insert into the texts, our procedure obtains a large number of real ASR errors on a text and selects a subset of these errors to include in the texts displayed.

After we prepared the meeting script and ran it against our low-accuracy ASR system, the next step was to align the reference text (the verbatim script of what the human actually said) and the hypothesis text (the output of the ASR system) to obtain a list of all the errors in the ASR output. We performed a time-based alignment of the hypothesis text to the reference text to correctly identify all the errors in the hypothesis text and generated a list of confusion pairs: for each incorrectly recognized word/phrase was paired with the reference word/phrase. For additional details on our time-based alignment method, the readers can refer to our prior conference paper [76].

**Stimuli Selection**

The word-alignment of the low-quality hypothesis output from the ASR system and the reference text transcripts (in Section 9.3.4 above) identified non-overlapping aligned sub-strings of these texts – with the pair of aligned substrings being non-identical if an ASR error had appeared within a particular region of the text, as shown in Fig. 9.5.

Thus, this alignment represented a set of possible confusion pairs, with each pair corresponding to an independent error (no overlap in the time frames)
the ASR system had made. We note that the reference text and the list of confusion pairs can be thought of as specifying an entire *space* of possible ASR outputs: Considering the reference text as a starting point and considering each confusion pair as an *insert an error* operator, one can imagine an entire network of possible ASR text outputs that are possible. Each ASR output contains some subset of the errors from the list of confusion pairs.

Given this space of possible ASR outputs, our goal was to identify two output texts for each reference text, with these properties:

- The output texts should reflect the reasonable performance of a commercial ASR system in noise typical of a workplace setting when the speaker is not wearing a special headset microphone; so, we wanted to identify text candidates with WER of approximately 0.25 (ranging between 20-30%); as supported by prior published results evaluating modern ASR accuracy in realistic settings [8, 94].

- We wanted to identify one text candidate that had a low WER-to-ACE ratio and another candidate with a high WER-to-ACE ratio. Specifically, we selected two candidates with identical WER: one with a high ACE
score, and the other with a low ACE score.

We wrote code to execute a search procedure through the space of possibilities to identify a pair of text candidates that fit the above criteria. We executed this code on 45 sentences that had been extracted from the verbatim script of what the human spoke in our business meeting videos, and we thereby obtained 45 pairs of ASR text output candidates (two per sentence). Thus, the two text candidates identified represented two possible outputs from an ASR system. The errors that appear in the texts are realistic: They were actual errors made by an ASR system, and the overall WER error rate for the sentences is approximately 0.25. We can think of one of these text candidates as being preferred by WER (the one with the low WER-to-ACE ratio), and the other as being preferred by ACE (with the high WER-to-ACE ratio).

9.3.5 Experimental Study Setup and Procedure

During the study, each participant was presented with 45 pairs of text output; each pair was displayed simultaneously, as shown in Fig. 9.6. The reference text (what the human actually said) was provided at the top of the screen, and the two text candidates were presented on the left and right side of the screen (positioned as captioning text in a black box below the video image).

The participant was asked to provide an individual subjective quality rating for each of the two videos, using a ten-level scale (frown face to smiley face) with endpoints labeled as “Useless” and “Useful”. At the beginning of the study, participants were provided with instructions on the study procedure and a practice item, prior to being presented with the 45 sentence pairs.
Figure 9.6: Screenshot from the study, with side-by-side comparison of caption-text automatically generated by ASR. Each pair of texts (left and right) have identical WER scores, but one text in each pair was preferred by our ACE metric.

The WER score was identical for the two text candidates that were shown in each pair; across all 45 pairs, the WER was in the range of 0.25 to 0.3. The two versions of text differed in their ACE score; one had a higher ACE score while other had a lower score. The presentation of text candidates on the left or right side was randomized throughout the study.

We recruited participants from the Rochester Institute of Technology and surrounding campus community. We collected data from 30 DHH participants (age distribution with mean = 23.53 and standard deviation = 4.92), which included 17 men and 13 women. Among our participants, 14 people self-identified as deaf, 8 people identified themselves as Deaf, and 8 people of the
participants as hard-of-hearing. All of the participants reported that they were familiar with the use of captioning technology, and they regularly used captioning when watching television programming.

This experiment has been approved by the Institutional Review Board (IRB) on October 28, 2015. Principal Investigators: Michael Stinson and Matt Huenerfauth.

9.3.6 Results and Discussion

We collected 2,700 responses in total from our 30 participants (subjective scores for each sentence, for 45 stimuli pair per participant). Fig. 9.7 presents the average subjective judgment rating for each participant in the study, displaying their average score across all text candidates that they evaluated: the text output preferred by the ACE score (with high WER-to-ACE ratio) and the text output not preferred by ACE. Fig. 9.8a presents the summarized response data across all participants, and Fig. 9.8b visualizes a linear correlation best-fit line for the relationship between the ACE scores of each sentence and the participants’ subjective judgment rating.

![Figure 9.7: Average usability rating variation among participants.](image)
Hypothesis H1 considered whether DHH users reported a subjective preference for captions that were predicted as better by our ACE metric, as compared to the captions that were predicted as worse by our ACE metric (high WER-to-ACE ratio vs. low WER-to-ACE ratio). The median difference (subjective score on better texts - subjective score on worst texts) was 2.5; the DHH users had higher subjective ratings for texts that had been preferred by the ACE metric. A boxplot summarizing the subjective rating scores from the participants for each stimulus (text predicted as better by ACE vs text predicted as worse by ACE) is shown in Fig. 9.8a. The distribution of the two groups differed significantly (Wilcoxon signed-rank \( W = 643394.00, N = 1350, N_{\text{est}} = 1226, p\text{-Value} < 0.0001 \)). Thus, hypothesis H1 was supported: DHH users preferred predictions from the ACE metric.

For Hypothesis H2a, we considered if the usability scores from DHH users correlated with the ACE score significantly. We computed Spearman correlation score for the two scores. The correlation coefficient was found to be \( \rho = 0.742791 \) with a p-value < 0.0001. Fig. 9.8b shows the corresponding correlation graph.
This supports H2a: DHH user’s judgments on the usability of the caption text correlated with scores from ACE metric.

For Hypothesis H2b, we performed significant difference testing on the correlations between 1) the human subjective preferences and the ACE score ($r_{ha}$) and, 2) the human subjective preferences and the WER score ($r_{hw}$). We performed a Fisher r-to-z transformation in order to perform an asymptotic z-test. For $r_{ha}$ and $r_{hw}$, we found a significant difference between the two coefficients (z-score = 5.771, 1-tail p-value < 0.0001). Thus, hypothesis H2b was supported: The subjective judgment of DHH participants about the quality of ASR captions was more highly correlated with ACE, as compared to their correlation with WER.

9.3.7 Summary and Discussion of Limitations of ACE

In our Phase 1 research, we investigated the design and evaluation of a new caption quality evaluation metric, called ACE, that analyzed the output of ASR systems to predict the impact of various ASR recognition errors on the usability of automatically generated captions for DHH users. Further, we compared the performance of this new ACE metric to the traditional WER metric in a study with DHH participants. In a side-by-side comparison of pairs of ASR text output with identical WER score, the texts favored by our new metric were also preferred by DHH participants. Our metric also had a significantly higher correlation with DHH participants’ subjective scores on caption usability, as compared to the correlation between WER and their scores.

While we have identified word predictability and semantic distance as useful
predictors of the usability of an automatically generated caption text, there are still limitations in our metric, which we address in the next phase of this research. Some of the limitations we have identified include:

L1. The current n-gram based word importance prediction model does not generalize well to unseen data (e.g., texts containing out-of-vocabulary words\(^3\)), as these models are based on exact-search and match. Furthermore, other unsupervised methods of word importance scoring, such as the ubiquitous TF-IDF scoring metric, have not been fully explored.

L2. While we had identified challenges in aggregating the individual error impact scores in a text into an overall score for a sentence, we had not fully explored alternatives approaches. There could be better, more simplistic methods for calculating this sentence-level error score.

9.4 Phase 2: Improving the ACE Metric to Create ACE2

The next step of our research included a closer evaluation of other unexplored strategies for building the automatic caption evaluation metric. The data collected during the PREFERENCE-2017 study in Phase 1 of our research will be utilized again in this phase of the project, to enable us to perform empirical evaluations, to explore various trade-offs between our original ACE metric and other, alternative metric designs for evaluation of ASR caption texts.

\(^3\)Words in a text that had not been present in the texts upon which the n-gram models were trained.
9.4.1 Improving the Word Importance Sub-score

To address limitation (L1) for the ACE metric, as discussed in Section 9.3.7, we explore two different approaches to improve the word importance sub-score:

Neural Word Predictability as Word Importance Measure

In Phase 1 of our work (refer to Section 9.3.1), we shared our rationale for using the word predictability estimation models as our primary measure of word importance in this task.

Since we experimented with the n-gram language model based word predictability measure earlier in Phase 1 of our work, we now look to investigate the efficacy of other measures of word importance estimation. In Section 4.2.2, we described a neural language model based architecture for estimating the word predictability. As a logical next step, we consider evaluating the use of this approach of word importance estimation of words in our task.

Fig. 9.9 provides a visual walk-through of how our neural language based word predictability estimation model assigns importance scores to a text. As shown in the figure, some words in the example text are found to be highly unpredictable - like ‘omaha’, ‘winters’, etc., while some are fairly predictable - like ‘i’, ‘in’, etc. in the text.

The example also demonstrates how the model is able to generalize on the rare word “Omaha”. If we compare the performance of this model with the n-gram based word predictability estimation model (see Fig. 9.2), we see that the n-gram based model assigns lower predictability (high importance) scores
Figure 9.9: Visual illustration of Neural-based word-importance scoring, based on the predictability of words in the context of a sentence, with higher bars indicating less predictable words.

to words after the word “Omaha”. As we discussed in Section 4.2.1, the n-gram model is based on search-and-match strategy, and a rare occurring word can affect the probability estimation of other neighbouring words, as the probability of matching the rare word would be low, which explains the high predictability scores assigned to words like “for” in the example in Fig. 9.2.

**Term Frequency Inverse Document Frequency as Word Importance Measure**

We introduced the Term Frequency Inverse Document Frequency (TF-IDF) measure in Section 3.1.1, which is a popularly used measure in information retrieval for extracting descriptive keywords in large document of text. This
strategy of statistically identifying importance of words in a text has been utilized in prior work on ASR evaluation [116, 122]. In this section, we look to compare the performance of the TF-IDF measure as a word importance sub-score in our new metric.

Figure 9.10: Visual illustration of TF-IDF based word-importance scoring, based on the predictability of words in the context of a sentence, with higher bars indicating less predictable words.

Following the methodological details provided in Section 3.1.1, we formulate our TFIDF measure. Specifically, we utilized the Switchboard corpus [51] as our reference corpus ($D$) for computing inverse document frequency of a word. Each two-person conversation in the corpus is treated as a “document”, thus the corpus contains a total of 2,438 documents. To compute the importance of a word ($w$) in a dialogue, we treat the corresponding dialogue text as the observed document ($D_w$) and calculate the TF-IDF score of the word ($w$) in
the document ($D_w$). Fig. 9.10 shows how our TF-IDF-based word importance model scores words in an example sentence. Notably, keywords like “omaha” or “winters” are scored higher than more common words like “in”, “for”, etc.

**Evaluating the Word Importance Models**

To compare these alternative methods of word-importance modeling, we compared our original ACE metric (which had used n-gram word-importance modeling) to two other ‘pseudo-ACE’ metrics, in which the word-importance sub-score was instead implemented using the neural-network-based or TF-IDF-based approached described above. For this analysis, we compared each metric on the data collected in our QUESTION-ANSWER study (described in Section 8.2), in which DHH participants answered comprehension questions for sentences containing errors.

<table>
<thead>
<tr>
<th>Word Importance Model</th>
<th>Semantic Distance Model</th>
<th>MSE Loss</th>
</tr>
</thead>
<tbody>
<tr>
<td>ngram_LM</td>
<td>word2vec</td>
<td>0.317</td>
</tr>
<tr>
<td>Neural_LM</td>
<td>word2vec</td>
<td>0.255</td>
</tr>
<tr>
<td>TFIDF</td>
<td>word2vec</td>
<td>0.257</td>
</tr>
</tbody>
</table>

Table 9.1: Comparison of error-impact prediction models (based on three different word-importance models) for predicting the comprehensibility of error-containing texts for DHH users. Note: The ngram_LM model corresponds to our original ACE metric from Phase 1.

Table 1 summarizes the performance of the candidate models on an error impact prediction task: “ngram_LM” represents the n-gram based word predictability measure as the word importance model (described in Section 4.2.1), “Neural_LM” is the neural language model based word importance
model (described in Section 4.2.2) and “TFIDF” is the TF-IDF based word importance model (described in Section 3.1.1), and finally “word2vec” represents the word2vec based semantic distance model (described in Section 9.3.2).

For the Neural\_LM-based and TFIDF-based models shown in Table 9.1, we recalculated the optimal \( \alpha \) coefficient for the weighted sum, as discussed in Section 9.1.3, with \( \alpha = 0.64 \) for the Neural\_LM-based metric and \( \alpha = 0.48 \) for the TFIDF-based metric.

For our evaluation, we considered the Mean Square Error (MSE) loss for the models which measures the deviation of the prediction of the model with the actual error impact score received from the participants. As shown in Table 9.1, the neural-network-based language model was the best performing error impact prediction model, with the lowest MSE. This model out-performed our original ACE metric on this data from the QUESTION-ANSWER study. Thus, we shall use this new Neural\_LM-based metric during the analysis in Section 9.4.2 below, and the Neural\_LM word-importance model will be utilized in our new ACE2 metric, promised in Section 9.4.

### 9.4.2 Alternatives for Combining Individual Error Scores into a Sentence Score

Calculating an overall score for a text, based on the individual error impact scores (for specific errors that occur within the text) can be challenging because our automatic-caption evaluation framework calculates the impact score for an error – as if it had been the only error the sentence. This approach makes it difficult to model cascading effects on comprehensibility of a text, due to
multiple errors – e.g. the effect of one error may affect (add to) the effect of another. For example, in Fig. 9.11, the impact score due to error “carrier” (for actual word “area”) does not consider that the previous word “kitchen” has also been misrecognized, when making the evaluation on the impact of the error. Since “kitchen” was misrecognized, the reader loses an important context clue as to the meaning, thereby making the word “area” even more important than before.

Figure 9.11: An example of error impact scoring in a sentence, with “c” indicating a correct word was recognized, and “i” indicating an incorrect word.

When designing our original ACE metric in Phase 1 of this research, we had contemplated a few possible methods for calculating a sentence-level score (given the scores for individual errors within the text) as discussed in Section 9.3.3, but we had not systematically compared a variety of methods for this calculation (a limitation of our work that we discussed in Section 9.3.7, see L2). The sentence-level score calculation strategy used in our original ACE metric
would assign the transcript in Fig. 9.11 an error score of 0.72. In this section, we propose few other aggregation strategies and evaluate their performance using the data collected in the PREFERENCE-2017 study in Phase 1 of our work.

Below, we discuss some of the alternative aggregating functions that we consider for our evaluation:

(a) **Mean, Median and Max**: The function takes the arithmetic mean of the individual impact scores due to errors in a text. It should be noted that this function does not consider the number of correctly recognized words when making the evaluation, but only considers the number of errors (\(num_{\text{errors}}\)) and their respective impact scores \(I(w_r, w_h)\). Thus, the \(ACE_{\text{mean}}\) metric below does not consider the number of correctly recognized words when making its evaluation.

\[
ACE_{\text{mean}} = \frac{\sum I(w_r, w_h)}{num_{\text{errors}}} \quad (9.5)
\]

Similarly, we consider using a median or the max function (to obtain the median among the error scores or the highest error score, respectively). For the example in Fig. 9.11, the sentence score using these methods (mean, median, and max) would give results (0.6, 0.6, and 0.8) respectively.

(b) **Position-based Weighted Average**: We also consider weighting the impact scores based on the position of the words; a similar idea was discussed in [70]. The intuition is that position may influence error
severity, e.g., errors at the end of the text might be more prominent than those at the start. We formulated five different weighting schemes, defined by the following distributions: Ramp Weight Distribution, Inverse Ramp Weight Distribution, Gaussian Weight Distribution, Inverse Gaussian Weight Distribution and Constant Weight Distribution.

\[
ACE_{pos\_weighted} = \sum_{i=1}^{N} W(i) \cdot I(w_r(i), w_h(i)) \quad (9.6)
\]

where \(N\) is the total number of alignment pairs during the comparison of the reference text and the hypothesis text. Then, \(i\) represents the position of the alignment in the text and \(W(i)\) represents the weight of the impact due to error based on its position \(i\).
(c) **Error-Spread Model**: Rather than treating an error impact score as an independent score representing the quality of a transcribed unit (or word), we wanted to consider the actual region of influence of an error in the text, when generating a quality score for the text. To realize this, we invented the Error-Spread model – which essentially *spreads* the impact of an error on to its nearby words. Ideally, such a region of spread could be linguistically informed (e.g. based on semantic boundaries), however, in our initial investigation we represent this as a constant parameter (learned) for all errors. Fig. 9.13 shows how the error-spread model operates on the individual error impact scores from the same example sentence as in Fig. 9.11, as we calculate an overall sentence score.

![Figure 9.13](image)

The figure illustrates how the impact of the misrecognition of the word “spread” as “split” indeed influences the nearby words, i.e. “i”, “the” and “newspapers.” Notably, the spread of the gray region visualizes the segment
of the text primarily affected by these errors.

To implement this, for each aligned pair \((w_r(i), w_h(i))\) representing an error \((w_r(i) \neq w_h(i))\), we use 1-dimensional Gaussian model centered at \(i\) with a fixed standard deviation \(\sigma\) (defining the spread of errors to the nearby words), and the height controlled by the impact score of the error \(I(w_r(i), w_h(i))\), to get the an error impact distribution function \(I_{dist}(w_r(i), w_h(i))\). With this representation, the region of impact is represented by the sum of all impact distribution functions. Equation 9.7 and 9.8 show how the final caption quality score is estimated, with the help of these error impact distribution functions.

\[
I_{dist}(w_r(i), w_h(i)) = I(w_r(i), w_h(i)) \ast \exp\left(\frac{-(x-i)^2}{2 \ast \sigma}\right) \tag{9.7}
\]

\[
ACE_{\text{-error}\_\text{spread}} = \frac{\sum_{i=1}^{N} I_{dist}(w_r(i), w_h(i))}{N} \tag{9.8}
\]

Finally, now that we have enumerated various methods for aggregating individual error impact scores to calculate a sentence level score, we evaluate the performance of the various strategies. For this analysis, we have utilized the data from the PREFERENCE-2017 study in Phase 1, where DHH users subjectively rated the quality of various automatically generated caption texts. Table 9.2 summarizes the results of our analysis. In the table: Group (a) shows the common statistical approaches mean, median and max. Group (b) shows the position-based weighting approaches. Group (c) is Error-Spread model and
lastly (d) is the aggregation function used in the original ACE metric. However, it should be noted the error impact model used in every row of this table is the new Neural_LM-based model, which was the best in our analysis in Section 9.4.1.

<table>
<thead>
<tr>
<th>Aggregate Function</th>
<th>Correlation ($\rho$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>0.797</td>
</tr>
<tr>
<td>Median</td>
<td>0.719</td>
</tr>
<tr>
<td>Max</td>
<td>0.837</td>
</tr>
<tr>
<td>Ramp</td>
<td>0.854</td>
</tr>
<tr>
<td>Inverse Ramp</td>
<td>0.705</td>
</tr>
<tr>
<td>Gaussian</td>
<td>0.851</td>
</tr>
<tr>
<td>Inverse Gaussian</td>
<td>0.851</td>
</tr>
<tr>
<td>Constant</td>
<td>0.859</td>
</tr>
<tr>
<td>Error-Spread</td>
<td>0.866</td>
</tr>
<tr>
<td>Neural_LM-based error impact model</td>
<td>0.861</td>
</tr>
<tr>
<td>(d)</td>
<td></td>
</tr>
<tr>
<td>impact model using the aggregation method from the original ACE metric (see 9.3.3)</td>
<td></td>
</tr>
</tbody>
</table>

Table 9.2: Comparison of different methods for calculating a sentence score, based on individual error scores contained within the text; each model below utilizes the Neural_LM-based error impact model (for calculating individual error scores) discussed in Section 9.4.1.

Thus, we see that the aggregation approach based on the Error-Spread method achieved the highest correlation with DHH users’ subjective judgements of sentence quality. Based on this analysis, we can now officially define our new (and improved) version of the automatic-caption evaluation metric, which we shall refer to as “ACE2” for the remainder of this article. This new version of the metric consists of:

- The word-importance sub-score of the model is implemented using the
Neural_LM approach described in Section 9.3.3.

- The semantic distance sub-score of the model is identical to the word2vec-based approach used in our original ACE metric.

- As discussed in Section 9.3.3, the new $\alpha$ coefficient used to calculate the weighted sum of the two sub-scores above was re-estimated, and it now has a value of 0.64.

- To calculate a sentence-level score based on the individual error impact scores, we utilize the Error-Spread approach described above.

### 9.5 Phase 3: Comparison with Prior Metrics

The formative evaluations presented above in Phase 2 of our research have largely consisted of comparisons among various alternative metric designs that we proposed, which has enabled us to invent the ACE2 metric defined above. In contrast, the focus of Phase 3 of our research is more outward-looking. That is, we identify various metrics for evaluating ASR output text or caption-quality that have been created by other researchers in prior work, and we make use of our data collected during the PREFERENCE-2017 study to compare our original ACE and new ACE2 metrics to these prior metrics. To begin, we have conducted a literature search to identify several other metrics that have been shown to be successful in predicting the quality of the ASR transcription for various applications:
9.5.1 Human Perceived Accuracy (HPA)

The HPA metric is designed to predict the human-perceived accuracy of ASR systems [116]; it uses learned weights to differentially penalize different error types, namely insertion, deletion and substitution errors. The metric also uses a measurement called word saliency which measures the semantic significance (or importance) of the spoken words. Upon setting up the metric for the captioning application (by learning the errors weights using our collection of data), we evaluated its performance in the caption usability prediction task. This metric had a correlation of 0.730 (Spearman’s correlation measure) with the DHH participants’ judgments in our PREFERENCE-2017 study data, from Phase 1 of our research.

9.5.2 Information Retrieval Based Evaluation Metrics

The general idea of this metric is to treat an automatic transcription as an information retrieval task, where the goal of a transcription is to retrieve all of the spoken messages as the output. Given this portrayal, researchers in [112] suggested using standard information-retrieval-based evaluation measures like precision, recall, and F-score to report an ASR system’s performance. Below, we share results using two variations of this metric – displaying Spearman correlation score ($\rho$) with the DHH participants’ judgements in our PREFERENCE-2017 study data:

- F-score (macro-averaged): $\rho = 0.418$
- F-score (micro-averaged): $\rho = 0.778$
CHAPTER 9. ASR EVALUATION FOR CAPTIONING

9.5.3 Word Information Lost (WIL)

This metric measures the proportion of word information communicated (or, inversely, lost) [120]. It is based on Mutual Information (MI) which measures the statistical dependence between the input words and output words. However, WIL provides a simple easy-to-implement probabilistic interpretation to the MI-based theory. The WIL metric had a correlation of 0.789 (Spearman correlation) with DHH participants’ judgments in our PREFERENCE-2017 study data.

9.5.4 Weighted Word Error Rate (WWER)

This metric considers each word to have importance weights (much like our word importance sub-score) such that when a word is recognized an appropriate penalty defined by the importance of the word is assigned to the error [122]. We observed a correlation of 0.742 (Spearman correlation) with DHH participants’ judgements in our PREFERENCE-2017 study data.

9.5.5 Weighted Keyword Error Rate (WKER) and Keyword Error Rate (KER)

Similar to the WWER metric, this metric penalizes errors based on the importance of the word in the text [122]. However, instead of weighting all the words based on the importance score, this metric follows the strategy of weighting only the keywords while all of the other non-keywords are simply ignored during evaluation. In contrast, KER weighs all the keywords with the score of 1 ignores all the non-keywords. On evaluating the performance of these metrics, the
KER metric performed slightly higher ($\rho = 0.757$) than the WKER metric ($\rho = 0.727$), in its correlation with the DHH participants' judgements in our PREFERENCE-2017 study data.

In summary, our ACE metric had a correlation of $\rho = 0.742$ with the DHH users' judgements of text quality in the PREFERENCE-2017 study, and our new ACE2 metric had a correlation of $\rho = 0.866$. Thus, ACE2 had the highest correlation of any metric in this comparative analysis. For comparison, the ubiquitous Word Error Rate (WER) metric was found to have a Spearman's correlation score of 0.108 with the DHH participants' judgements in this same dataset.

9.6 Phase 4: User-Based Evaluation of ACE and ACE2 (PREFERENCE-2018 Study)

As our final summative evaluation, we now look to compare the performance of our original ACE metric (originally published in [76]) with the newly proposed ACE2 metric, on a new dataset, which we have collected in a new study with DHH participants. Our goal is to definitively compare our old and new versions of this metric, as to their efficacy in predicting the quality of caption text for DHH users.

To set up this study, we followed a similar design as in our original PREFERENCE-2017 study (described in Section 9.3.5). Thus, we shall refer to this new study as “PREFERENCE-2018” for the remainder of this article.
Specifically, we presented DHH participants with automatically generated caption texts based on an audio recording of a two-person conversation, and we asked participants to provide their opinion as to the usability of each caption.

As in the PREFERENCE-2017 study, participants were able to see the reference text (the intended message of what the human was saying, as transcribed accurately by a human), and participants were able to consider this, when they were judging the quality of the automatic captions they were asked to evaluate.

Section 9.6.1 describes our methodology for designing the stimuli for this new study, Section 9.6.2 discusses the set-up and conduct of the study, and Section 9.6.3 provides the results of our evaluation, which compares the ACE and ACE2 metrics.

### 9.6.1 Designing Stimuli

The stimuli text that we showed to our DHH participants were generated by transcribing speech recordings of a conversation between two hearing people, using a modern ASR system. The recordings were the snippets from the aforementioned Switchboard corpus, which contains recordings of over 2,400 pairs of strangers having a casual conversation over a telephone line.

In our PREFERENCE-2017 study, we had generated stimuli by *engineering* realistic recognition errors in caption texts, to compare the WER metric and the ACE metric. The idea had been to select a pair of text with the same WER score but with different ACE scores (one high and other low), so that when the two captions were presented to users side-by-side, we could see which caption
participants preferred. Subsequently, we used the data from that original PREFERENCE-2017 study in Phase 2 and 3 of our research.

Although that previous study design had enabled us to gather a useful dataset of DHH participants’ judgments, for our new PREFERENCE-2018 study, we wanted to use more naturalistic set of caption texts for evaluation. Thus, rather than engineering the quality of a caption text (by selecting a subset of possible ASR errors from a larger set of possible ASR errors, as we had done in the prior study), we wanted to directly utilize the output from ASR systems as stimuli for our PREFERENCE-2018 study. This meant that our prior strategy of fixing the score from one metric and varying the score from the other metric was no longer possible, since we were limited to using the exact output from ASR systems.

Thus, we automatically processed the recordings from the Switchboard corpus using a variety of ASR systems: Google’s Cloud Speech, IBM’s Watson Speech to Text, and Sphinx ASR \(^4\). Given the output from these ASR system, we had to select the specific pairs of caption texts to show in the study. We therefore gave preference to selecting pairs of ASR hypothesis texts from different ASR systems that had greater differences (absolute value) between their ACE and ACE2 score. This approach to stimuli selection ensured that we had the potential to compare the correlations between the human subjective scores and each of our metrics. We also selected a mix of texts with low, medium and high ACE and ACE2 scores, so as to introduce sufficient variability for the

correlation measure. With this approach, we selected 180 stimuli texts for our study; the average WER of our resulting stimuli in the study was 26%.

9.6.2 User Study Setup

During the study, each participant was presented with 60 automatic caption text outputs (30 pairs, displayed side-by-side), along with the reference text displayed above the pair (as shown in Fig. 9.14). The participants were asked to provide a subjective judgment on the usability of each caption text, by indicating their choice on a ten-level scale (frown face to smiley face) with endpoints labeled as “Useless” and “Useful”. Aside from sharing the same reference text, the two ASR output texts displayed side-by-side did not share any specific metric value. (This was a notable difference from our previous PREFERENCE-2017 study design, where the two texts displayed side-by-side had identical WER scores but contrasting ACE scores.)

This experiment has been approved by the Institutional Review Board (IRB) on October 28, 2015. Principal Investigators: Michael Stinson and Matt Huenerfauth.

To recruit participants for the study, we reached out to students from the Rochester Institute of Technology and other people who were DHH in surrounding the Rochester community. We collected data from 12 DHH participants (age distribution with mean = 21.67 and standard deviation= 2.534), which included 7 men and 5 women. Among our participants, 6 people identified themselves as Deaf, 3 people self-identified as deaf, and 3 people as hard-of-hearing. All of the participants reported that they were familiar with the use
of captioning technology, and they regularly used captioning when watching television programming.

### 9.6.3 Results and Discussion

In the study, we collected a total of 720 responses from our 12 participants. We created three versions of the stimuli set, with each set containing 60 different caption-text outputs. Each participant in the study judged one version of the stimuli set.

Fig. 9.15 visualizes a linear correlation best-fit line for the relationship between the: (a) participants’ subjective judgment rating and ACE scores of each caption-text stimulus and (b) participants’ subjective judgment rating
and ACE2 scores of each caption-text.

As discussed in Section 9.2.1, the goal of this study was to investigate our Hypothesis H3 which considered whether the preference scores collected from DHH participants correlated more strongly with the ACE2 metric, as compared to how strongly their correlated with the ACE metric. Therefore, we began our analysis by computing the Spearman correlation scores to measure: (a) the correlation between the DHH participants’ judgments and the ACE score ($r_{ace}$) and (b) the correlation between the DHH participants’ judgments and the ACE2 score ($r_{ace2}$). The correlation coefficients were found to be: $r_{ace2} = (0.5519, \text{p-value} < 0.0001)$ and $r_{ace} = (0.3927, \text{p-value} < 0.0001)$. Fig. 9.15 shows the corresponding correlation graph.

To investigate Hypothesis H3, we next performed significance difference testing on the correlations between $r_{ace}$ and $r_{ace2}$. Specifically, we performed Fisher r-to-z transformation to calculated the difference. We found a significant
difference between the two coefficients (z-score $= 1.818$, 1-tail p-value $< 0.05$).
Thus, Hypothesis H3 was supported: If we present ASR output texts with varying ACE and ACE2 scores to DHH participants, there will be a significantly higher correlation between DHH participants' judgments of text-usability and the ACE2 metric, as compared to the correlation between DHH participants' judgments of text-usability and ACE.

Thus, we have found that ACE2 out-performs our original ACE metric for the task of predicting DHH users' subjective judgment of the quality of a caption text.

9.7 Conclusions

In Part II of this work, we identified that the current metrics used for evaluating (and sometimes optimizing) the performance of ASR systems rely on counting the number of recognition errors without regard to what the errors are and where they occur in the transcription. However, this approach of evaluating the performance of the ASR systems had previously been shown to be loosely connected with the actual opinions of human participants in various application settings.

Consequently, this work investigated whether traditional measures of ASR quality based on automatic metrics (such as the WER metric) correlated with DHH users' subjective judgement of quality. Our results indicated that that prior automatic metrics are poor predictors of DHH users' subjective judgement about the quality of ASR output.
Therefore, we presented the design and evaluation of new metrics for predicting the quality of an ASR system, in regard to whether the system’s output would produce understandable captions for DHH users. Thus, one contribution of this work is in predicting the effect of recognition errors on DHH users’ understanding of a caption-text. More specifically, we found two main factors that we found useful in predicting the impact of an error on caption understandability:

- **Word Importance**: This measure estimated the importance of the error word in the caption text. (Part I of this research discussed several models regarding this.)

- **Semantic Distance**: This measure computes the semantic deviation between the error word displayed in the caption text and the actual word that had been spoken.

As a part of our analysis, we have examined various approaches to estimate these measures for predicting the impact of errors in a caption-text. Further, measuring the overall quality of a text based on individual error impact scores required additional analysis on the efficacy of various methods for combining these individual error impact scores, to produce a single aggregate score of the quality of an entire sentence.

Specifically, Phase 1 of research in this chapter had proposed a (baseline) automatic caption evaluation metric, called the ACE metric, and compared the performance of our ACE metric with the WER metric in a caption usability prediction task, when used in a business meeting scenario. This study was
referred to as the PREFERENCE-2017 study in this thesis, which was discussed in Section 9.3.4. In a side-by-side comparison of pairs of ASR generated caption texts with identical WER, we found: (a) subjective preference of the users on the caption quality judgments from the ACE, and, (b) significantly higher correlation of the DHH users’ subjective usability rating with ACE score as compared to the correlation of this rating with the WER score.

In Phase 2 of this research, we addressed some limitations of the original ACE metric and empirically evaluated various possible improvement strategies. These improvement strategies were evaluated based on the subjective usability scores collected from the DHH users in the PREFERENCE-2017 study (Section 9.3.4) conducted in Phase 1 research. This lead to the development of a new metric, called the ACE2 metric, and in Phase 3 of our research, we compared ACE and ACE2 to previously published metrics from other researchers for evaluating ASR text quality.

Finally, in Phase 4 of our research, we collected additional subjective preference data from DHH participants in a new study. However, to diversify the evaluations, the captions used in this study were generated from a more informal two-person conversational dialogue (compared to the previous fake business meeting setup), and the output of various commercial ASR systems was used to produce stimuli directly, without engineering stimuli to contain specific subsets of ASR errors, as in the prior PREFERENCE-2017 study (Section 9.3.4).

Ultimately, the findings of this study revealed that users’ subjective evaluation of captions is better correlated with the new ACE2 metric (as compared
to WER or ACE), and we have described and evaluated a metric that can be used by future researchers for evaluating the suitability of ASR systems for generating captions for DHH users. Such a metric could be used as an initial investigation of caption quality under various environmental conditions or speakers, and it could be used to compare various ASR systems for this application – prior to conducting a study with DHH users. We also see potential for such metrics to be used to drive the development of ASR-based captioning systems, rather than the use of currently popular metrics, such as WER, which had very little correlation to DHH users’ judgments of text quality.
The Part II of the work discussed the need of a metric of ASR quality when focusing on captioning application for DHH users. Our work identifies the inefficacy of current measures of evaluating the quality of ASR systems, especially in captioning applications. We present the Automatic Caption Evaluation (ACE) framework to design better metrics of ASR quality and we validate them through studies with the real users.

We designed experiments to elicit accurate feedback on the quality of automatically generated captions for evaluation. Using the data collected from the user-based studies, we compared the various ASR evaluation metrics, and provided evidence of their efficacy in captioning applications. We showed that traditional metrics of ASR evaluation, such as the Word Error Rate (WER), are not well-suited for evaluation as it doesn’t reflect the users’ perception of ASR quality, as compared to our proposed metrics.

Specifically, this work aimed at answering the following research question:

**RQ2:** Do our models of estimating the quality of ASR systems for generating captioning for DHH users accurately predict the quality
In Part II of the work, we investigated various new metrics for predicting the quality of automatically generated captions for DHH users, and we validated their performance through evaluation of the new metrics with other candidate metrics by designing comparative studies with the DHH users. More specifically, we developed additional sub-research questions to answer this question more precisely:

**RQ2.1:** In a simulated two-person business meeting scenario, does our metric for predicting quality of automatically generated captions for DHH users outperform the WER metric? (Our work in Phase 1 of this research provided results (discussed in Section 9.3.6) that supports this research question. We had formulated three hypotheses (H1, H2a and H2b) to answer this question; refer to section 9.2.1 for the hypothesis investigated for the RQ).

**RQ2.2:** In a simulated business meeting scenario, does our metric for predicting the quality of automatically generated captions outperform prior metrics published in the literature, in terms of correlation between DHH human judgments and the predicted score from the metrics? (This research question was investigated in Phase 3 of the work. Results in section 9.5 support this RQ).

**RQ2.3:** In a realistic two person conversational settings, would our metric for predicting the quality of automatically generated captions outperform the WER metric? (We examined this re-
search question in Phase 4 of the work. Results in section 9.6.3 support this RQ).
PART III: CAPTION USABILITY ENHANCEMENTS
To enhance the usability of captioning systems, we present Part III of our work where we investigate importance-based highlighting in captions. Motivated by the accessibility challenges of captions for DHH readers (e.g., need to split visual attention between the caption text and other sources of information) we explore the usefulness of highlighting in captions. We speculate that this may allow readers to attend to the most important bits of information in the caption quickly which may reduce their reading times and enhance their understandability of these users.

To investigate this, we present our research in two chapters: First, in Chapter 11, we investigate the benefits of importance-based highlighting in captions for DHH readers. The research activity in this includes studies with DHH users to compare their usability ratings when viewing videos with and without caption highlighting. As a follow-up to this study, in the second part of this research, we focus on understanding DHH users’ preference for the different design choices for highlighting in captions. We compare prior recommendations of highlighting in static texts (e.g., textbooks and electronic documents) and explore their
efficacy in captioning applications for DHH users. Hence, Part III our work explores the following research question RQ3 (as presented in 1.2), which states:

RQ3: Do DHH users benefit from importance-based highlighting in captions? If so, what are their highlighting preferences? To study this research questions in more detail, we formulate the following sub-research questions:

RQ3.1: Are DHH users receptive to the premise of importance-based word highlighting in captions?

RQ3.2: When viewing captions of online lecture video, do DHH users subjectively prefer highlighting in captions?

We will examine RQ3.1. and RQ3.2. in Chapter 11. The research questions studied in this chapter are focused on exploring the benefits of highlighting in captions. However, prior research has shown that inappropriate design choices for highlighting, especially in captions, have detrimental effects on the experience of the readers. Therefore, as a preface to our main investigation in Chapter 11, we started with a small-scale formative study with a goal to establish a baseline on the highlighting preferences of DHH readers. Based on the results from the study, we designed our final study to investigate RQ3.2. While the formative studies established some ground-work on the preferred highlighting configuration in captioning applications, the results of the study were only preliminary without statistically significant results.
Hence, we formulate additional research questions that focused solely on the design preferences of users on highlighting in captions, which we investigate via an independent study discussed in Chapter 12. Additional research questions investigated in Part III of our research are as follows:

RQ3.3: What text decoration strategy would DHH users prefer for highlighting important words in captions?

RQ3.4: Do DHH users prefer highlighting important words in captions or highlighting entire sentences?

RQ3.5: When the same word is identified as important in multiple occurrences in the captioning for a single video, do DHH users prefer that the first occurrence be highlighted only or multiple occurrences?

RQ3.6: For what genres of video would DHH users also indicate an interest in seeing highlighting of important text in captions?

We will examine RQ3.3 – RQ3.6 in Chapter 12.
Chapter 10

Prior Work on Caption Accessibility

10.1 Caption Accessibility Challenges

Although services exist to provide access to spoken content for DHH users, e.g. sign language interpreters or captioning services, users face challenges in attending to multiple streams of visual information. Text captioning of video content is increasingly common, e.g. enabling educational institutions to satisfy legal requirements for making content accessible for DHH students [6, 42]. Yet, traditional text captions are not a complete solution to providing full access to video content for DHH users, especially when there are multiple concurrent visuals and/or visual-references within the captions [87, 95, 110]. DHH users who rely on visual information sources must strategically switch between the captions and other visual information in video content. Since
human cognition is a limited resource, with bounds on processing concurrent visual information sources, there can be a loss of information, even when high-quality accessibility services are provided. Consequently, research has found that DHH users typically get less out of even accessible mainstream classroom lectures than their hearing peers do [110].

Many interventions (discussed below) have been proposed for enhancing visual browsing of text through highlighting important words [87,95], but this work has looked at text (e.g. textbooks or web-pages). As online video has become an increasingly popular source of news, education, or entertainment among the general population [86,130], research is needed on whether highlighting important words would also be beneficial for caption text displayed during video, and whether this would benefit DHH users. There is reason to believe why research is needed that specifically focuses on DHH users in this context:

*Peripheral visual attention.* Research has found that DHH users, especially those who have used sign-language since an early age, have greater peripheral vision skills than hearing users [15,138]. Eye-tracking studies to understand DHH users' strategies when viewing captions have revealed that although DHH users spend a smaller amount of time reading captions (compared to hearing users viewing captions), the amount of time DHH users spend watching captions depends on the rate of change of captions and amount of motion in the images [24]. These findings suggest differences in how DHH users visually process a video caption text, as compared to general readers of text.
Reading literacy skills. Further, in standardized testing in the U.S., English literacy rates have been measured to be lower among adults who are deaf [72,106] – a result that may be due to reduced language exposure or other educational experiences during childhood. Lower reading literacy skills among DHH users could affect the usage of captions among these users [18,165]. Moreover, users with lower literacy can find it especially challenging to follow fast-moving captions [165]. This further suggests that research on the benefits of text highlighting is needed specifically among this group.

Errors and omissions in the caption text. Even when captions are produced by a human transcriptionist or captioning service, there can be errors in the text that is provided, or delays in when it is presented. Automatically produced captions, e.g. produced through automatic speech recognition (ASR), may have an even greater percentage of errors. Moreover, unlike human-generated errors, errors produced by automatic systems have been found to be even more cognitively demanding for users [77,89]. In addition, caption texts (especially if produced through an automatic method) customarily do not convey speaker traits like accents, vocal emphasis on words, or emotional subtext of speech, which could be useful for DHH users. Thus, while there has been prior work (discussed below) on highlighting words in static text, research is needed on captions.
10.2 Improving Caption Accessibility

Consequently, there has been a lot of work towards enhancing the accessibility of captions [16, 22, 78, 87, 95]. Several researchers have studied how to reduce users' cognitive load when viewing simultaneous visual streams, e.g. by combining them onto a single display [22, 87]. Some have investigated visual dispersion in classroom settings and how this affects DHH students [87]. In their work, researchers proposed a system to combine multiple video streams of classroom elements (e.g. lecturer, slides) into a single display. To notify the user of changes on various streams, e.g. when a new slide is displayed, they used visual cues on their user-interface, e.g. flashing outline of a video region. The authors discuss how the integration of different visual streams into a single display reduced the distance between the information streams, in support of context-switching and reducing access times. However, readers still had to integrate multiple sources of information, which is known to tax working memory and impact learning [5].

To tackle the challenges in accessing fast-moving captions, Lasecki et al. [95] investigated a display interface where students could pause the captions, to avoid falling behind when in classroom settings. The intuition was to reduce the visual dispersion of attention (between captions and other regions of the video), which could lead to DHH users losing track of the captions or the visual content. With the ability to pause the captions, users could follow the content at their own pace, even when the caption text contained references to the surrounding visual content. However, in a real-time classroom setting, pausing a caption could lead students to fall behind what the instructor speaks during
that interval. Other work has focused on improving the delivery of lectures in classrooms [16]. Researchers have reported that the pace at which content is presented affects learners' retention and understanding of content [91,92], especially when learners' working memory is overloaded with information. Brandao et al. [16] designed a system to reduce the lag between presenters and interpreters, to synchronize them into a single unit of content delivery, and to provide students the time to shift their attention and read visual material at their own pace.

10.3 Importance-based Highlighting in Text

Text highlighting provides a natural way of conveying important information in text, and research has found that readers are able to make use of this emphasis information, without special training as to its meaning [38]. In the education context, highlighting is a common strategy used by textbook authors, teachers, and students to indicate important concepts in a text, and this has been shown to enable faster browsing and recall of information by students [25, 45]. In general, researchers have also argued that strategic marking of words or phrases can enhance the reading experience [50,80]. Readers use highlighting as a way of functional coding, which helps with retention and faster browsing [50]. Similarly, readers perform better on comprehension tests when reading text with highlights [80]. Other work has focused on the effect of text highlighting on word-retention and learning [73,105,133]. When comparing keyword-highlighting and non-highlighting conditions, researchers found that students performed
better on a cloze task when a text-passage was highlighted [97]. With the increasing availability of digital text content, several highlighting interventions have been applied to such texts, including rendering text in a different color, with color backgrounds, or changing font decoration. Such special rendering enables readers to attend to the most important segments of the text or focus on relevant information quickly [26,80,133]. In addition, computing accessibility research has investigated text transformations to promote comprehension among struggling readers [146]. For instance, highlighting important words has been found to improve reading rate and comprehension among people with dyslexia [145]. In an eye-tracking study, researchers found that keyword highlighting improved comprehensibility and readability of onscreen texts for people with dyslexia [146].

10.3.1 Style Guidelines for Highlighting

While textual highlighting has known benefits, there are various ways an author could highlight a text. Depending on the context or application, some styles of highlighting might be preferred by readers. Research has revealed that readers tend to interpret various highlighting strategies in different ways. For instance, Acrey et al. [1] reported that italicizing (and boldfacing) is effective for conveying stress on some important words in a text, in comparison to underlining. In addition, these researchers also reported that UPPERCASE and boldface texts may be interpreted as being spoken loudly by the readers, while italicized text may have a softer emphasis [1]. Similarly, Strobelt et al. [159] ranked highlighting strategies according to the strength of each at
Table 10.1: Design aspects in prior text highlighting research, e.g. granularity (full sentences or individual words highlighted)

<table>
<thead>
<tr>
<th>Prior Work</th>
<th>Granularity</th>
<th>Decoration Style</th>
<th>How Words Selected</th>
<th>Source of Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hautasaari et al. [64]</td>
<td>Words / Phrases</td>
<td>Font Color</td>
<td>Automatic (Gensen Web)</td>
<td>ASR transcripts from audio conferences</td>
</tr>
<tr>
<td>Ponce et al. [133]</td>
<td>Words / Phrases</td>
<td>Font Color</td>
<td>Manual</td>
<td>Expository passages</td>
</tr>
<tr>
<td>Kawasaki et al. [80]</td>
<td>Words / Phrases</td>
<td>Font Color</td>
<td>Manual</td>
<td>Web content</td>
</tr>
<tr>
<td>Pan et al. [128]</td>
<td>Sentences</td>
<td>Underline</td>
<td>Manual (speakers selected the highlights)</td>
<td>ASR transcripts from multilingual communication</td>
</tr>
<tr>
<td>Hello et al. [140]</td>
<td>Phrases</td>
<td>Boldface</td>
<td>Automatic</td>
<td>Expository passages</td>
</tr>
<tr>
<td>Chi et al. [26]</td>
<td>Sentences</td>
<td>Background Color</td>
<td>Automatic</td>
<td>Electronic books</td>
</tr>
<tr>
<td>Berger et al. [11]</td>
<td>Words / Phrases</td>
<td>Background Color</td>
<td>Automatic (analyzed prosodic prominence)</td>
<td>Recordings of speakers reading expository passages</td>
</tr>
<tr>
<td>Vertanen et al. [166]</td>
<td>Words / Phrases</td>
<td>Underline</td>
<td>Automatic (based on ASR confidence)</td>
<td>ASR transcripts of participants reading expository passages</td>
</tr>
<tr>
<td>Seita et al. [154]</td>
<td>Words / Phrases</td>
<td>Underline and italics</td>
<td>Automatic (based on ASR confidence)</td>
<td>ASR transcripts from multiparty meeting</td>
</tr>
<tr>
<td>Kalle et al. [78]</td>
<td>Words / Phrases</td>
<td>Underline</td>
<td>Manual</td>
<td>Lecture transcripts</td>
</tr>
</tbody>
</table>

capturing a person’s attention, with yellow-colored background and boldfacing highly ranked. They reported that use of italics was less demanding of readers’ attention [159]. Table 10.1 summarizes the alternative highlighting strategies used in prior work on various sources of data.

Hence, it is common, especially for professionals, to refer to style manuals that specify guidelines for preparing documents for print and online publication, including best practices and recommendations for highlighting in texts. Some popular manuals include: Publication Manual of the American Psychological Association [3], the Chicago Manual of Style [127], and The Copyeditor’s Handbook [39]. These resources provide direction as to which words should be highlighted and how they should be styled or decorated in the text. Below are some recommendations from these style manuals regarding highlighting in texts:

the use of emphasis for full sentences, but instead recommends using it for a few individual words:

*In running text, italics and boldface should be used sparingly to set off specific terms and phrases. Most publishers discourage the use of italics or boldface for entire sentence and paragraphs: Long passages of wavy italic type are often difficult to read, and large patches of boldface (or frequency small patches of bold type) look unattractive.*

- [39], page 404

*Text Decoration Styles.* Style manuals also specify some form of distinctive treatment or appearance styling of important words. For example, the Chicago Manual of Style recommends using italics or boldface for highlighting key terms in a text. In American Psychological Association (APA) format, italics is the recommended decoration to be used to set words apart from the rest of the text in order to introduce new ideas or key words. However, the APA does not recommend italicizing a word only for emphasis, rather APA recommends writers to work with the bounds of syntax and sentence formation to exert emphasis in a text.

*Repeated Occurrences of Key Words.* The APA style manual specifically indicates that you should only use italics for the first time the word is used, but then you should use regular font for subsequent usage of the term. The Chicago Manual of Style includes a similar recommendation:

*Key terms in a particular context are often italicized on their first
10.3.2 Visual Markup of Text in Captions

Highlighting text in captions can be challenging, especially when considering the dynamic nature of the captions compared to static text. Users are engaged in an attention-demanding task of viewing a video with multiple sources of visual information in parallel to the caption-text stream, and if the choice of highlighting style or the frequency with which words are highlighted is suboptimal, then such visual decoration of the text could be distracting. This speculation is supported by prior research that has investigated the effect of different visual markup of caption texts in another context: to convey the confidence scores of a captioning service (e.g., ASR system) as to the accuracy of its caption output [12]. In a study with 107 DHH users, Berke et al. discovered that although participants were receptive to the idea of having visual indicators of the confidence of an automatic caption system, they were concerned about distraction from changes in text appearance. DHH users viewing video are sensitive to text-appearance changes in captions, and there is risk that highlighting text could actually be detrimental. Thus, empirical research is needed to determine the best choice of highlighting styles. Then, based on this set of preferred design parameters, we can then determine whether there are indeed benefits for DHH users from text highlighting.
Chapter 11

Evaluating the Benefits of Highlighting in Captions

11.1 Background and Introduction

Captioning services provide access to audio or audiovisual information to many people who are Deaf or Hard of Hearing (DHH). These services usually employ a human who transcribes the audio information to digital text. For video of live events or single-speaker lectures, this caption text is usually displayed below or alongside the visual source in 1 or 2 lines refreshing every 2 to 4 seconds [89]. Unlike the static text in books or on webpages, caption text during videos is often a representation of spoken language, which may be less formal and contain more disfluencies [113]. In addition, caption text is dynamic, meaning that the stream of language is moving at a speed that is not determined by the reader, but rather by the original video source. Readers are also limited in the degree
they can glance back to previous words in the text, since only a few lines of text may appear at a time, to avoid the caption from blocking too much of the video image. When reading caption text, users may also need to divide their visual attention between the words in the text and the visual information elsewhere in the video. These simultaneous demands on the user’s visual attention can be challenging for DHH individuals who are viewing videos, especially those in which there are multiple visual references or multiple concurrent sources of visual information [22]. Figure 11.1 shows an educational lecture video (a stimulus used in our study) with three concurrent visual sources: a lecturer in the middle, a projector display on top-right corner and the captions at the bottom of the screen.

While hearing individuals may be able to process audio speech and visual content concurrently, DHH individuals need to switch their attention between visual streams, which may result in missing some information. Researchers have attributed this need for DHH students to switch between the different visual content in a classroom setting to higher cognitive demands, leading to students lagging behind and under-performing, in comparison to their hearing peers [22].

Prior reading comprehension research has found that readers are often able to skim a text quickly for relevant information, rather than fixating on a portion for deep meaning [26,35]. Consequently, researchers have investigated methods for highlighting important words or phrases in text to enable faster browsing and to support the reading task [73,89,97,105,133]; that work did not examine captioning contexts. Although recent computing research has
investigated automatic methods for identifying important words in a caption text [76,77], there has been a lack of research on the usability of highlighting key words for users. Such highlighting in captions may require special consideration: Unlike text documents, captions are dynamic, with shorter text segments, which are usually shown in 1 or 2 lines, for 2 to 4 seconds [89]. Moreover, users are known to be sensitive to caption display parameters such as speed, font size, or decorations: Several researchers have measured the influence of such visual parameters of caption appearance on the readability of captions for DHH users [12, 89, 165].

Thus, our work investigates text-highlighting in captions for videos viewed by DHH users, and our contribution is threefold: First, we examine DHH users’
preferences for various visual parameters of highlighting important words in captions: In two rounds of in-person interview and prototype usability studies, DHH participants indicated when they would prefer to see word-highlighting, which style of visual markup for this highlighting they preferred, and what is the threshold percentage of words that should be highlighted. Secondly, to investigate the efficacy of text highlighting in captions, we present results from a larger study, in which DHH users responded to questions after viewing videos of two forms: with and without highlighting of important words in the captions, when viewing educational lecture videos. Thirdly, the question-types and empirical results in our larger study could be beneficial for future researchers when evaluating automatic methods for identifying important words for users in this educational-video context, with our results as a potential baseline.

11.1.1 Research Questions Investigated in this Chapter

Thus, our work investigates text-highlighting in captions for videos viewed by DHH users. To understand the preferences of DHH users who are viewing videos, in regard to highlighting important words in the caption text, this chapter investigates the following research questions:

RQ3.1. Are DHH users receptive to the premise of importance-based word highlighting in captions?

RQ3.2. When viewing captions of online lecture video, do DHH users subjectively prefer highlighting in captions?

We present the evaluation of our RQs in two phases: First, we conducted some
studies in which we gathered subjective preferences from a small number of DHH users about various display options for highlighting in captions. These smaller preliminary studies were not sufficiently powered to enable us to observe statistically significant differences in user preferences. Instead, the goal of these formative studies was to provide some preliminary answers to RQ3.1, so that we were not making arbitrary choices about our design. Later, we conducted a larger user study to compare the experience of DHH users when viewing online educational lecture videos under two conditions: with and without text-highlighting. This final summative study utilized the best configuration settings for caption-highlighting found in our initial formative studies, which had identified the most preferred application use-case (educational online-lecture videos), choice of highlighting style (underlining), and percentage of words to highlight (at most 15%). To answer RQ3.2 in our final study, we compared the two video conditions: with and without text highlighting.

11.2 Formative Studies: Method and Results

The goal of our formative studies was to understand DHH users’ interest in important-word highlighting in video captions, and their preferences among various visual markup strategies. Two studies, with 6 DHH participants each, helped us select display options for text-highlighting for our subsequent larger study. Rather than selecting design settings arbitrarily, we used this multi-study design to identify parameters for that final study.
11.2.1 Highlighting Configurations for Formative Studies

During this in-person interview and prototype-evaluation study, which was conducted in two rounds, we presented users with videos with different highlighting configurations. We were interested in two main design factors: the highlighting markup style and the percentage of words to highlight. One option for investigating these two factors would be to conduct a single study with a large number of users to investigate various possible combinations of both factors, within a single study. Because these initial studies were planned as preliminary formative studies, in support of our larger final evaluation study, we chose to instead devote more personnel and time resources toward conducting the final study with as many participants as possible. Thus, we decided to investigate these two design factors in a cascaded manner, through a two-round formative study design, with each factor investigated independently in each round: In round 1, to compare visual markup strategies, we conducted a within-subject study with 7 markup conditions previously shown in Table 10.1. In all of the video stimuli shown in this round-1 study, the percentage of words highlighted was kept constant at around 20%. In round 2, the stimuli videos included variations in the percentage of words highlighted in each caption. We investigated 4 conditions: low percentage (5%), medium (15%), high (25%) and very high (35%). At the end of the round-1 study, we had determined that underlining was the preferred method of visual highlighting of important words. Thus, all of the stimuli video in this round-2 study used underlining as the method of visual highlighting. As discussed in Section 11.5, this choice to
cascade the two small studies, each investigating a single factor, did not enable us to investigate interaction effects among variables. However, this tradeoff allowed us to devote more resources toward the larger final study.

These studies has been approved by the Institutional Review Board (IRB) on February 24, 2017. Principal Investigators: Matt Huenerfauth.

11.2.2 Stimuli Preparation for Formative Studies

As discussed below, in some open-ended interview questions conducted during this formative study, we asked our DHH participants about the types of videos for which they may be interested in text-highlighting. In responses to those questions, users expressed interest in highlighting for online educational lecture videos, but we had not anticipated that finding when we had launched this round-1 formative study. Thus, the stimuli video used to display various text markup styles and highlighted-word percentage in the round-1 study was from a non-education genre: Specifically, as stimuli for this formative study, we used videos of a fake business meeting which had previously been used in our earlier research in Section 9.3.4. The meeting video was first chopped in 12 smaller videos with an average duration of 30 seconds. Each participant was shown the 12 videos in order, each with a different display configuration-setup arranged in pseudo-randomized order (Latin-Square) for each condition.

In order determine which words should be highlighted in the caption text for these videos, we utilized automatic word-importance prediction system as our automatic word importance system, described in Section 6.1. In order to produce each video stimulus with a particular percentage of words highlighted
as important, we ranked words according to this score, and we highlighted a portion of the top-ranked words, to control the percentage of words with visual highlighting in the final caption text.

11.2.3 Recruitment and Participants for Formative Studies

For both rounds of formative studies, participants were recruited by e-mail and flyers at the Rochester Institute of Technology. Participants were eligible if they answered “yes” to both: Are you Deaf or Hard-of-Hearing? Do you use captions when viewing television? Participants met a DHH researcher fluent in both English and ASL in a private office to ensure a distraction-free environment. Participants were paid $40 for the 60-minute study.

11.2.4 Questionnaires for Smaller Studies

The questions asked and overall sequence of activities were identical for the round-1 and round-2 studies; the only difference was the video stimuli (either focusing on markup-style or percentage of highlighted words). At the beginning of each session, participants were informed that they would see captioned videos with some words shown differently, as a word-importance highlighting strategy. Before viewing video stimuli, participants answered open-ended items, on a pre-study questionnaire, regarding the usefulness of word-importance highlighting for captions.

Next, participants viewed stimuli videos; after each video they answered three questions, which had been used by prior researchers to successfully gather subjective responses from DHH participants about caption quality [88]:
Q1: *How easy were the captions to read?* (Five-point scale from very hard to very easy)

Q2: *How easy was it to follow the content in the video?* (Five-point scale from very hard to very easy)

Q3: *Did you find the caption distracting?* (Yes/No)

Towards the end of the study, the participants were asked some post-study questions to gauge their interest in various use-case scenarios in which word-highlighting of captions may be preferred. The rationale for asking these questions at the end of the study was so that the users would have had some initial experience at viewing captions with various types of highlighting. For these questions, participants were encouraged to propose any situations and/or genres of videos where captioning highlighting might be beneficial.
Figure 11.3: Round-1 Formative Study: Comparison of different visual markup-styles for highlighting in captions on *distracting* question.

Figure 11.4: Round-1 Formative Study: Comparison of different visual markup-styles for highlighting in captions on *easy to read* question.
11.2.5 Round-1 Results: Comparing Markup-Styles

A total of 6 DHH individuals participated in the round-1 study, with 3 males and 3 females, and self-identified hearing-status of 4 Deaf and 2 Hard of Hearing. Participants were shown captions with different visual markup (shown in Table 10.1). Figure 11.2 – 11.3 summarizes the responses of the participants for the different caption markup strategies. In particular, Figure 11.2 and 11.4 presents participants’ responses to Likert questions; diverging stacked bar graphs like these are recommended for presentation of Likert response data [151]. The segments of each bar indicate the percentage of responses for each Likert option, with the conditions of the study along the Y-axis. The neutral response is centered horizontally, with negative responses to the left and positive responses to the right. Participants preferred the underlining (ul) strategy for highlighting words in captions, and the bold strategy was a close second. Although the italics (it) markup was recognized as one of the least distracting strategies, it was harder to follow compared to other more distracting strategies like font color (color_red). Participants further reported that italicizing was harder to read. Notably, strategies like uppercasing (uc) and font size (size) changes were indicated as one of most distracting markups. The small sample size of this formative study did not support statistical significance testing.

11.2.6 Round-2 Results: Comparing Highlight Percentage

Based on the results of the round-1 study above, the video stimuli shown in the round-2 study used the underlining method of highlighting. In this
Figure 11.5: Round-2 Formative Study: Comparison of the percentage of words highlighted in captions on *easy to follow* question.

Figure 11.6: Round-2 Formative Study: Comparison of the percentage of words highlighted in captions on *distracting* question.
round-2 study, a total of 6 newly recruited DHH individuals participated, with 3 males and 3 females, and self-identified hearing-status of 4 Deaf and 2 Hard of Hearing. Participants in this round viewed videos using underlining as a highlighting strategy, with the video stimuli differing as to what percentage of words were highlighted. Participants preferred videos with 5% to 15% of the words highlighted. As shown in Figures 11.5 – 11.7, participants indicated that captions were most readable when 5% of the words were highlighted, and when 15% of the words were highlighted, participants found it easier to follow along with the captions. Participants found the 5% highlight condition the least distracting. The small sample size of this formative study did not support statistical significance testing.
11.2.7 Round-1 and Round-2 Results: Interest in Highlighting

Across both rounds, participants were asked identical questions about their interest in highlighting. Since we did not observe a difference in feedback comments across the two rounds, for brevity, responses from all 12 participants across both rounds are presented together below.

On the pre-survey questionnaires, participants shared their initial thoughts about importance-based highlighting for video captions. Participants were fairly open to the premise: 8 participants out of 12 welcoming the idea. When asked to elaborate, one participant responded as follows:

*I think important words being highlighted in captions should be worthwhile because it helps to get my attention in any matter. In my experience, sometimes I am too lazy to read all the captions, but I will be more attentive if there is something important to know.*

(P9)

Two participants expressed concerns about using such a feature, especially given its novelty and their lack of familiarity with this new form of text appearance. They were also concerned that visual distraction due to highlighting would decrease the readability of the captions:

*It might be hard to read; I haven’t seen that before. It could be useful, but I would have to see to make judgements.* (P6)

A few participants indicated doubts about the usefulness of this feature. Although they saw some potential, they were not sure if it was a silver-bullet
solution:

It depends, if it's in a classroom setting then, yeah, it sounds helpful but if it's based on the persons voice, etc. Then, no! If it's actually an important vocabulary, then yeah it would be nice. (P7)

Furthermore, in responses to our post-study questionnaire, many participants indicated that highlighting would be beneficial for online lecture videos, with some saying:

Highlighting helps me to keep track of the online materials and video content. (P9)

When the teacher talks too long, the deaf people have a hard time catching up. That's why the students need to know which words are important. (P7)

Other contexts in which participants indicated that text-highlighting may be useful were: meetings with hearing peers (mentioned by 5 of 12), classroom lectures (by 4 of 12) and news/political video announcements (by 3 of 12).

11.2.8 Discussion of Results from Round-1 and Round-2

The results from these round-1 and round-2 formative studies began to address research question RQ3.1. However, these small formative studies were conducted as a preliminary exploration of this design space, with a goal of informing the design of our final larger study below. Given the small number of participants in these formative studies, they were too underpowered to enable statistical
significance testing. While not yet providing a conclusive answer to the various
design questions raised by RQ3.1, these formative studies did enable us to avoid
making arbitrary design choices as to the appearance of word highlighting for
our final study.

11.3 Larger Study: Method and Results

The goal of our final study was to understand whether DHH users subjectively
prefer word importance highlighting in captions (RQ3.2). We utilized the
results from our two rounds of formative studies, as summarized in Table 11.1.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Markup Strategy</td>
<td>underline (ul)</td>
</tr>
<tr>
<td>Highlight Percent</td>
<td>5 – 15%</td>
</tr>
<tr>
<td>Video Genre</td>
<td>Online-lecture videos</td>
</tr>
</tbody>
</table>

Table 11.1: Results of preliminary studies used in final study.

11.3.1 Preparation of the Stimuli Video

Since online-lecture video was the most popular scenario suggested by par-
ticipants for word-importance highlighting, for our final study, we needed to
produce new video stimuli to match this context. When generating stimuli,
we used “underline” style highlighting, with 5-15% of words highlighted. As
discussed earlier, many online education platforms use a screen arrangement
with multiple concurrent visual streams, including an image of the instructor,
of slides, of captions, etc. As a basis for our stimuli, we made use of a public
dataset of educational video stimuli that had been produced in Kafle et al. [62], which possessed several desirable characteristics for use in experimental studies:

- Content Obscurity: The content in the videos had been engineered such that it was obscure, to remove content-bias from a participant having prior knowledge of a topic. The content was partially fictionalized, using fake names and other historical details wherever necessary.

- Content Homogeneity: As discussed in Kafle et al. [62], a similar density of information types (names, dates, etc.) was distributed throughout all the slides of the lecture and the script of the instructor spoke, to enable the videos to be partitioned into segments for experimental studies.

- Visual Homogeneity: As discussed in Kafle et al. [62], the videos have an onscreen layout with multiple regions typical of lecture videos on many education platforms. As shown in Figure 9.4, each contains: the instructor, the slides, a topic list for the lecture, and captions. The videos had been designed such that they promote visual homogeneity: with limited and consistent color use across the visual streams over time, and with limited upper body gestures by the instructor. This homogeneity enables the videos to be partitioned into segments with similar appearance.

The visual and content homogeneity of these videos allowed us to create a controlled setting for understanding the effects of text highlighting in captions. In total, this dataset contains four lessons [62], with each lesson being five minutes in duration, and containing exactly 10 presentation slides, each having
a time duration of 30 seconds. Each of the lessons was originally of duration 5-minutes. To generate our stimuli, we split each into sub-lessons of 1.5-minute duration each (discarding the final 0.5 minutes of each lesson). This yielded a set of 12 short videos, each of which was rendered in two conditions: with and without highlighting. While we had used our automatic system (Section 6.1) for identifying important words during our initial formative studies, that system had been designed to operate on conversational-style speech, rather than formal academic lectures. While it would be possible for researchers to re-train a word-importance system for this new genre of speech, building an automatic system for this task was beyond the scope of addressing our research questions in this study. Thus, instead of using an automatic system, we manually identified important words in the text that should be highlighted. To reduce individual bias, the words were identified based on a consensus labelling by a group of 3 researchers. We used the threshold criterion in Table 2 for selecting the number of words to be highlighted such that 15% of the words were highlighted in each stimulus video. Given the short video duration and our methodology of asking multiple researchers to agree upon the importance-labeling of words to be highlighted, practically it was easier to achieve a consensus at the 15% level.

11.3.2 Study Setup and Questionnaires

In pilot testing, participants indicated that watching four 5-minute videos (in their original duration) was too tiring; so, each lesson was split into three shorter segments. Each participant saw 9 videos (segments of three lessons) during the study. However, each sub-lesson video within a lesson was always
presented in sequence, to preserve the original temporal flow of each lesson. The highlighting and non-highlighting conditions were presented in an alternate order in the videos, with the assignment of conditions to each stimulus video counterbalanced across participants.

Similar to the preliminary study, participants were asked pre-study and post-study questionnaires. During the study, participants viewed videos with and without highlighting and answered questions about the readability of the captions. In addition to the questions asked in the preliminary study, we also included questions about the user-perceived workload of the comprehension task. We asked about three dimensions of the NASA Task Load Index (NASA-TLX) [63]:

- **Mental Load:** This scale measures how much mental and perceptual activity was required for the task, e.g. thinking, decoding, remembering, looking, searching etc.

- **Temporal Demand:** This scale measures how much time pressure the user felt due to the pace of the task.

- **Effort:** This scale measures how hard the user had to work to accomplish their level of performance.

The wording of the TLX items was modified slightly to include the phrase “reading and understanding the captions in the video” after any mention of “the task.” The final set of questions as shown in this study appear in Table 11.2.

In prior methodological research [13], we found that in experiments with DHH participants (especially those with diverse levels of reading literacy)
Questions | Scale
---|---
Q1. It was easy to follow the content of the video and captions. | 5-point Likert Scale from Strongly Agree to Strongly Disagree.
Q2. It was easy to read the caption. | 
Q3. I was able to identify the important words and concepts. | 
Q4. I understood all of the content of the video and captions. | 
Q5. How mentally demanding was the task (reading and understanding the captions in the video)? | 21-point NASA-TLX scale from Very Low to Very High.
Q6. How hurried or rushed was the task (reading and understanding the captions in the video)? | 
Q7. How hard did you have to work to read and understand the captions in the video? | 

Table 11.2: List of questions used in the final study

evaluating English-text captions, comprehension-question probes were less discriminative than other question-types. Based on that finding, do not include comprehension-question probes in this study, instead we include the subjective-type questionnaires that were discussed above.

To summarize, each participant was shown videos of 1.5-minute duration each. Half of the videos contained captions with words highlighted in them, and half of the videos, without highlighting. For each video, our participants answered 7 questions in total (Table 11.2). Q1-Q4 were questions that were inspired from our earlier preliminary study, and Q5-Q7 were adopted from the NASA-TLX.
11.3.3 Recruitment and Participants

We recruited participants for this study using similar methods as in our formative studies. Participants were paid $40 for the 60-minute study. A total of 30 DHH individuals (age distribution with mean = 25 and standard deviation = 6.02) participated, with 15 males and 15 females, and self-identified hearing-status of 17 Deaf and 13 Hard of Hearing. To evaluate the English literacy skill of participants, we used Wide Range Achievement Test 4th edition (WRAT4), which had been previously validated with DHH users [132]. Our participants reported an average WRAT score of 82.6 ± 12.4, which is one standard deviation below the standard score (100) among adults in the U.S.

11.3.4 Results

For each question in Table 11.2, we collected 270 responses from the 30 DHH participants on the stimuli videos. This section presents a comparison of participants’ subjective preference of each of the two conditions: videos containing captions with highlighting (highlight) and videos containing captions without any highlighting (no_highlight).

Figure 11.8 compares responses when asked about the ease of following the content of the video under our two highlighting conditions, represented on the Y-axis of the chart. After conversion of scalar Likert responses to integer (e.g., “Strongly Disagree” = 1, “Disagree” = 2, etc.), a Wilcoxon rank-sum test indicated a significant difference for this question \( W = 576, p < 0.0001 \) with the average rating-score at \( \mu_1 = 3.32 \) and \( \mu_2 = 3.9 \) for the no_highlight and
CHAPTER 11. CAPTION HIGHLIGHTING

Figure 11.8: Percentage distribution of participants’ responses on the ease of following the content of the video and the caption.

highlight conditions respectively. In Figures 4 through 9, brackets indicate significant differences as follows: *** $p < 0.0001$, ** $p < 0.001$, * $p < 0.01$, or N.S. not significant.

Figure 11.9: Percentage distribution of participants’ responses on the readability of the caption.

Similarly, a Wilcoxon rank-sum test indicated significant differences on the general ease of reading the captions in videos [$W = 631$, $p < 0.0001$]
with the two means at $\mu_1 = 3.53$ and $\mu_2 = 4.09$ for no_highlight and highlight conditions respectively. Figure 11.9 summarizes the responses for the participants for this measure.

**Figure 11.10:** Percentage distribution of participants’ responses on being able to identify the important words and concepts.

Figure 11.10 shows the distribution of Likert-responses when participants indicated if they could identify the important words and concepts in the video. A Wilcoxon rank-sum test indicated significant differences for this question [$W = 246$, $p < 0.0001$] with means at $\mu_1 = 3.05$ and $\mu_2 = 4.18$ for no_highlight and highlight conditions respectively.

When asked to indicate the overall understandability of the captions, participants subjectively preferred the highlight condition over the no_highlight condition, with means at $\mu_1 = 3.9$ and $\mu_2 = 3.5$ respectively. A Wilcoxon rank-sum test indicated significant difference for this measure [$W = 759$, $p < 0.001$]. Figure 11.11 summarizes the percentage distribution of responses for this question.
Figure 11.11: Percentage distribution of participants’ responses on the understandability of the content of the video and the captions.

Participants also reported differences in the mental demand when reading and understanding the captions in the video.

Figure 11.12: Percentage distribution of participants’ responses on the mental demand when reading and understanding the captions in the video.
read and understand the captions in the video under the two highlighting conditions, as shown in Figure 11.12. The box plot reveals that the median score for the highlight condition was lower than that of the no_highlight condition. A Wilcoxon rank-sum test indicated a significant difference for this measure ($W = 2185, p < 0.01$).

![Box plot showing the median scores for the highlight and no_highlight conditions.](image)

Figure 11.13: Percentage distribution of participants' responses on the temporal demand of reading and understanding the captions in the video.

Figure 11.13 shows the box-and-whisker diagram summarizing the responses of the participants when asked about the temporal demand of the task. While the box plot may appear to show that the median score for the highlight condition was slightly lower than no_highlight condition, a Wilcoxon rank-sum test indicated no significant difference between the highlight and no_highlight conditions for this measure.
Lastly, there was a significant difference in responses to the question about the effort required to read and understand the captions in the video, under the two conditions, revealed through a Wilcoxon rank-sum test \( W = 1743, p < 0.001 \). Figure 11.14 shows the results of this analysis.

**Figure 11.14:** Percentage distribution of participants’ responses on the difficulty of reading and understanding the captions in the video.

### 11.4 Discussion and Conclusion

Although captioning of videos is essential for making content more accessible for DHH users, prior research has found that there is room for improvement, especially for videos with multiple channels of visual information. The additional demands of attention management and visual processing for such videos may
detract from an individual’s ability to comprehend the content. This effect would have particular significance in educational contexts, including for videos of lectures in which students must keep track of the instructor, slides, and other visual information sources.

Our results indicate that participants are open to the idea the highlighting in captions, and we measured statistically significant differences when comparing participants’ responses after they viewed captions with each highlighting condition, for online lecture videos. In particular, participants indicated that lecture videos containing highlighted words in captions were easier to read and follow, as compared to videos without any highlighting. This was an important finding because prior research on the incorporation of visual markup in captions (for conveying confidence of words in captions generated automatically) had revealed that participants had concerns about being distracted by text decoration. On the contrary, our results indicate that our markup strategy for highlighting in captions was preferred by DHH participants on this task, who reported a significantly higher ($p < 0.0001$) readability score on videos with captions containing highlighting. Although not distracting, the participants found the highlighting in captions to be noticeable enough to be able to identify important words and concepts in the video. Participants also reported an overall increase in the understandability of the content of the video and captions under highlighting.

Similarly, we observed a significant difference ($p < 0.01$) in participants’ mental and perceptual load required to read and understand the captions, with and without highlighting. Overall, they reported less mental load when viewing
videos with captions that contained highlighting. In addition, participants indicated that it required less effort to read and understand captions under the highlighting condition. However, we did not observe a difference in participants' rating of the temporal demand, under our two highlighting conditions. This result suggests that although the participants found highlighted captions easier to read, this did not influence their perception of time pressure from the pace of the video. It is important to note that in both highlighting conditions, participants indicated a relatively low degree of temporal demand (highlight: 8.46, no_highlight: 8.90), suggesting that participants were rather comfortable with the pace of the task in either case.

Although our investigation of the various design issues from RQ3.1 was only formative in nature, those preliminary studies had suggested that we should focus on the context of educational online lecture videos, with underlining markup, with 5%-15% of words highlighted. While we did not provide any statistically significant empirical evidence that these are the optimal settings of these design variables through the findings presented in this paper, our final study was able to confirm that for this specific combination of these variables, users did prefer captions with highlighting.

In conclusion, our findings demonstrate the potential benefit of caption highlighting in online educational videos for DHH users, which motivates additional research in this area, and we have provided a methodological foundation for the evaluation of such systems with DHH users.
11.5 Limitations of this Research and the Need for an Additional Study

There were limitations of our study that we would like to address: Our smaller formative studies were too under-powered (i.e. with too few participants) to enable us to investigate the design options in research question RQ3.1 conclusively. Although those studies served their formative purpose for this paper, we believe that future researchers and designers would benefit from a more conclusive investigation of those design options in a larger study. Consequently, the upcoming chapter (Chapter 12) investigates the design preferences of highlighting in captions for DHH users in more detail.
Chapter 12

Evaluating the Designs for Highlighting Captions

12.1 Background and Introduction

Prior research has investigated methods for reducing the visual cognitive load for DHH users when viewing captions via various interventions, e.g. delaying the delivery of the captions [95] or delaying information from a lecturer [16]. In the previous chapter, we investigated caption highlighting [78] to support DHH users when viewing educational videos. Much like highlighting of important words in static text, e.g. when key terms are shown in boldface in a student’s school textbook, caption highlighting involves visually enhancing important segments of text in captions, to enable faster browsing and to support the reading task [97,133].

As discussed in Section 10.3.1, although there exist a variety of guidelines
and standards for effective methods of highlighting key words in static texts, e.g. the style guide of the American Psychological Association (APA) [3] or Chicago Manual of Style [127], they have important limitations: (a) these guidelines were not produced for dynamic texts like captions, (b) they reflect an editorial standard but their efficacy is not research-based, and (c) they have not been evaluated with DHH readers. Given the differences between static text documents and dynamic captions, it is unknown whether such guidelines would directly transfer when highlighting in captioning contexts.

While our prior study (described in Chapter 11) found that DHH viewers of educational videos subjectively preferred videos with captions that had some important words underlined in the text, Section 11.5 discuss limitations of that study, namely a lack of empirical evidence to justify design choices made in their captioned video stimuli, e.g.: (a) Interviews with only a few DHH users guided those authors’ selection of education lecture videos as a suitable genre for use of highlighted captions. (b) Those few interviews were used to select the text-decoration approach (underlining individual words) used in the stimuli in their study.

The contribution of this chapter is empirical; specifically, we conducted a study with DHH individuals to gather their preferences in regard to various design options for highlighting the important words in video captions. Whereas our prior work (discussed in Chapter 11) had been under-powered and formative in nature, our study included 36 DHH participants, which enabled us to investigate their preferences in regard to:

- The granularity for selecting highlighting in captions; either highlighting
segments of text on a word-by-word basis or highlighting entire sentences at a time.

- The decoration to use when highlighting (underlining, italicizing, and bold-facing), which are commonly recommended markup styles for highlighting in static text documents [3, 127] or in video captions [78].

- Whether to highlight repeated occurrences of the same keyword in a caption text or only first occurrence. Finally, while our study focused on educational lectures, as recommended in Chapter 11, we asked participants to indicate their interest in highlighting for other video genres.

12.1.1 Harmful Effects of Inappropriate Highlighting

While the study by Berke et al. [12] revealed the risk if caption text is decorated incorrectly for DHH users, if we broaden our focus to consider research on hearing individuals reading static texts, we find prior work underscoring the risk of incorrectly selecting the words that should be highlighted in a text. For instance, prior educational research has revealed that inappropriate highlighting in text is detrimental to students' learning [48, 49]. Researchers found that students who read text passages that had been highlighted inappropriately were significantly less accurate when they later responded to reading-comprehension questions, in comparison to students who had read the same text passage with carefully selected highlighting of the text.

Researchers have therefore experimented with various approaches for selecting segments of text for highlighting. These include using an automatic system
for extracting important segments from a static text [26, 64, 76] to utilizing human annotators to manually review a text for highlighting [45, 80, 133].

While the static-text style guides discussed above advocated for highlighting individual words, rather than full sentences, prior educational research on highlighting of static texts for students has often presented students with text stimuli in which entire sentences have been highlighted, e.g. [26, 128]. The rationale is that when students use a highlighter marker to identify important content in a physical textbook, they often use the marker to highlight entire sentences. For this reason, we decided to also investigate this factor in our study: asking our participants to compare highlighted captions that use word-level or sentence-level granularity for highlighting. Given the risks revealed by prior research if texts are highlighted incorrectly, in our study, we will investigate design issues that relate to which words are highlighted or how that highlighting should appear.

12.1.2 Research Questions Investigated in this Chapter

To understand DHH users’ preferences in regard to various design options for highlighting keywords in captions, we conducted an experimental study in which DHH users viewed videos (with variations in the design of the caption highlighting), to investigate several research questions. Our research in Chapter 11 had identified educational lecture videos as a target genre for their investigation of highlighting important words in caption texts for DHH users. Participants in their formative interviews expressed strong interest for highlighting for this genre. Thus, we have decided to retain this focus on educational lecture videos,
and we have used such videos as the basis for stimuli generated for our study.

RQ3.3. What text decoration strategy would DHH users prefer for highlighting important words in captions?

Further, using underlining\(^1\) to highlight keywords, as suggested by our prior research in Chapter 11, we examine some additional questions:

RQ3.4. Do DHH users prefer highlighting important words in captions or highlighting entire sentences?

RQ3.5. When the same word is identified as important in multiple occurrences in the captioning for a single video, do DHH users prefer that the first occurrence be highlighted only or multiple occurrences?

Although our research in Chapter 11 had indicated that DHH users had a strong preference for educational lecture videos as a target genre for the use of keyword highlighting in captions, we wanted to understand whether DHH users may also be interested in such highlighting for other genres of captioned video:

RQ3.6. For what genres of video would DHH users also indicate an interest in seeing highlighting of important text in captions?

\(^1\)To avoid dividing participants among multiple studies, we investigated our research questions within a single experimental study with 36 DHH participants, a relatively large number of participants for an in-person study with this group. For this reason, we could not cascade the findings for one research question to guide the stimuli appearance for each subsequent one. Since our prior work in Chapter 11 had suggested that DHH users would prefer underlining, we used underlining in the video stimuli to investigate the later research questions. We determined that this approach would be more controlled than showing each participant different stimuli in later phases of the study, based on that individual’s preference for RQ3.3. However, the Results section will reveal that participants actually preferred **boldfacing**. We mention this issue again in our Limitations section.
12.2 Methodology

We conducted a laboratory-based experiment in which DHH participants viewed educational lecture videos with captions that included highlighting of important text, with a variety of designs related to our research questions. Participants responded to subjective questions about the quality or usability of the captions in each video stimulus.

This experiment has been approved by the Institutional Review Board (IRB) on February 24, 2017. Principal Investigators: Matt Huenerfauth.

12.2.1 Four Phases in the Study

Each session was divided into four phases, in which participants answered subjective questions about various captioned video stimuli. Examples are shown in Figure 12.1–12.3, and Table 12.1 summarizes the sources of video used.

Phase 1: RQ3.3. was investigated in a single-factor two-level within-subjects design, to compare DHH users’ preference for various text decoration styles for highlighting (underlining, **boldfacing**, and *italicizing*). These three styles had been the three most preferred in our prior underpowered formative interview study with DHH users (Chapter 11), which had asked if captions were easy to follow or distracting. Details of video stimuli creation are discussed below, but examples of Phase 1 stimuli appear in Figure 12.1. Each participant saw all six videos, each only once, with assignment of conditions (underlining, **boldfacing**, and *italicizing*) for each video rotated and balanced across participants. To
## Table 12.1: Sources of video content for stimuli in each phase.

<table>
<thead>
<tr>
<th>Phase</th>
<th>Source of Video for Producing Stimuli</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phase 1 6 videos</td>
<td>Two video lectures from the Harper Video Dataset [62] were split into sub-lessons of 1.5-minute duration each, following the same procedure as described in [78].</td>
</tr>
<tr>
<td>Phase 2 2 videos</td>
<td>40-second snippet from two TED Talks: <a href="https://www.ted.com/talks/prosanta_chakrabarty_four_billion_years_of_evolution_in_six_minutes">https://www.ted.com/talks/prosanta_chakrabarty_four_billion_years_of_evolution_in_six_minutes</a> <a href="https://www.ted.com/talks/manoush_zomorodi_how_boredom_can_lead_to_your_most_brilliant_ideas">https://www.ted.com/talks/manoush_zomorodi_how_boredom_can_lead_to_your_most_brilliant_ideas</a></td>
</tr>
<tr>
<td>Phase 3 4 videos</td>
<td>45-second snippets from four TED Talks, which included repetition of one or more key terms to be highlighted: <a href="https://www.ted.com/talks/tina_seelig_the_little_risks_you_can_take_to_increase_your_luck">https://www.ted.com/talks/tina_seelig_the_little_risks_you_can_take_to_increase_your_luck</a> <a href="https://www.ted.com/talks/jennifer_pluznick_you_smell_with_your_body_not_just_your_nose">https://www.ted.com/talks/jennifer_pluznick_you_smell_with_your_body_not_just_your_nose</a> <a href="https://www.ted.com/talks/adam_alter_why_ourScreens_make_us_less_happy">https://www.ted.com/talks/adam_alter_why_ourScreens_make_us_less_happy</a> <a href="https://www.ted.com/talks/mehdi_ordikhani_seyedlar_what_happens_in_your_brain_when_you_pay_attention">https://www.ted.com/talks/mehdi_ordikhani_seyedlar_what_happens_in_your_brain_when_you_pay_attention</a></td>
</tr>
</tbody>
</table>
make our results more comparable to our work in Chapter 11, we once again used the educational video dataset of Harper [62] for Phase 1 of our experiment. However, for Phases 2 and 3 (below) in our study, we used excerpts from TED Talk videos, which were more visually dynamic and may be more typical of the visual complexity of online video.

Figure 12.1: Samples of video stimuli in Phase 1 of the study with videos containing different text decoration styles for highlighting.

**Phase 2:** RQ3.4. was investigated in a 2-level single-factor within-subject design, to compare DHH users’ preference between sentence-level vs. word-level highlighting. For word-level, highlighting was selected with granularity being
the individual word; for the sentence-level condition, entire sentences were either highlighted or not. Example stimuli from this phase of the study appear in Figure 12.2. As source material for videos in Phase 2, two educational lecture videos from the TED Talk series were used (details in Table 12.1). During Phase 2, each participant viewed two conditions per video, presented side-by-side (with the right-left placement of each condition randomized for each participant). This side-by-side methodological approach was recommended in prior research for evaluating subtle appearance differences in video captions in studies with DHH users [77].

Figure 12.2: Samples of video stimuli in Phase 2 of the study with videos containing different levels of granularity for highlighting.
Phase 3: RQ3.5. was examined in a 2-level single-factor within-subjects design to compare DHH users’ preference between two strategies for handling when a word had been identified as important in multiple occurrences throughout a text: (a) highlighting the first occurrence only or (b) highlighting each occurrence. Stimuli are shown in Figure 12.3. Like Phase 2, videos of TED Talks were used as source material for stimuli (Table 12.1). As in Phase 2, participants saw both conditions for each video, presented side-by-side.

![Figure 12.3: Samples of video stimuli in Phase 2 of the study with videos containing different strategies for handling repeated keyword when highlighting.](image)

Phase 4: Unlike the earlier phases of the study, no video stimuli were shown in this final phase. Instead, participants responded to a question to investigate RQ3.6., and they completed demographic questionnaires and a literacy test.
12.2.2 Details of Video Stimuli Creation for Each Condition

We developed our own software to burn captions in videos which allowed us to set specific styles to the caption display format. For all the videos used in the study, we used Arial font with a font size of 15 to display our captions. The caption buffer-time, which controlled the amount of text displayed on the screen at a time, was set to 8000 milliseconds, and the text was displayed on 1 or 2 lines. The font color for the captions were set to white, which were drawn over a black background to avoid visual interference.

As discussed above and summarized in Table 12.1, we worked with 12 source videos to develop our captioned stimuli for the study. For instance, in Phase 1, we used the six videos to produce a total of 18 stimuli, with each video generated under each of the three display conditions (underlining, italics, and boldfacing). Similarly, a total of four stimuli were produced for Phase 2 (2 videos, 2 conditions) and eight stimuli for Phase 3 (4 videos, 2 conditions).

The caption text for the video segments shown in the study was produced by a human captionist and checked for accuracy by a researcher who reviewed the video. The text had Flesch Kincaid Grade Level reading complexity scores between 4th and 12th grade level, and the average Flesch Kincaid Reading Ease score was approximately 63.

As in Chapter 11, human annotators were asked to inspect the texts manually to identify the important words to be highlighted; for consistency, we followed the same approach as in Chapter 11 for how this should be done: The annotators were advised to select only 15-20% of the text as being important for highlighting.
To select important sentences for highlighting (for stimuli in Phase 2 of the study), sentences with more than 20% of words selected as keywords were marked as important sentences.

12.2.3 Questions Asked in the Study

To select the set of questions to ask in our study, we considered the questions used in prior research in which captions had been evaluated by DHH participants [13, 78, 88]. In particular, we were interested in using questions like those used in Chapter 11, to enable comparison between the studies.

We did not include comprehension-style questions in which participants would be asked to recall information from the videos. While this choice may seem counterintuitive at first blush, we wanted to follow the prior methodological recommendations of Berke et al. [13], which had reported that in experimental studies with DHH participants evaluating captioned videos in English, objective comprehension-question probes were less discriminative than other subjective question-types. Thus, for each video stimulus displayed in the study, participants were asked to indicate their agreement on a Likert-scale for the following three items:

a. *I was able to identify the important words and concepts.*

b. *I found the captions distracting.*

c. *It was easy to read the caption.*

Participants responded to each on a 5-point scale: strongly disagree, disagree, neither agree nor disagree, agree, or strongly agree. Questions a and b were
used in all three phases of the study, but question c was only used in phase 1. The reason was that during phases 2 and 3, participants view the same video side-by-side (with different highlighting conditions). Based on pilot testing prior to the launch of our study, we found that since participants were viewing the same video side-by-side, with identical caption text (albeit with differences in which portions were highlighted), they either assigned identical scores for both of the side-by-side stimuli or felt that the question implied that there must have been a difference in the words in each stimulus (which there was not). To avoid confusion, question c was only used in phase 1, in which stimuli were presented in a serial fashion.

In phase 4, before responding to some demographic and literacy-skill questions, participants were asked to indicate their agreement with six statements on a 5-point Likert-scale:

- I think highlighting important words in a caption would be useful for *online educational videos*.
- I think highlighting important words in a caption would be useful for *entertainment videos*.
- I think highlighting important words in a caption would be useful for *news and political videos*.
- I think highlighting important words in a caption would be useful for *live captions in a real-world classroom*.
- I think highlighting important words in a caption would be useful for *live
captions in a one-on-one meeting in the real world.

- I think highlighting important words in a caption would be useful for live captions in a meeting with many people in the real world.

12.2.4 Recruitment and Participants

Through emails and flyers on our university campus, we recruited DHH individuals who indicated that they regularly used captions when watching video. Our 36 participants included 16 males and 20 females, with mean age of 23.6 years (standard deviation of 5.14 years). Among the participants, 22 identified themselves as being d/Deaf and 14 as Hard of Hearing.

The participants were paid $40 for the 60-minute study. The study occurred in a lab setting using a laptop that displayed videos and questions onscreen. At the beginning of each phase, participants were provided with instructions, including an example video and the set of questions they could expect in the upcoming phase of the study.

At the end of the study (in Phase 4), participants were asked to complete the Wide Range Achievement Test 4th edition (WRAT4) [172]. Our participants had an average WRAT score of 85 ± 13.8, which is one standard deviation below the standard score (100) among all adults in the U.S., but a typical-to-high score among deaf adults in the U.S. [132].
12.3 Results

This section presents the results from the four phases in our study, which correspond to our four research questions. For statistical difference testing, Likert responses were converted to integers (e.g., Strongly Disagree: 1, Disagree: 2, etc.). In the graphs below, the p-values are indicated as “***” (0 to 0.001), “**” (<0.01), “*” (<0.05), or NS not significant. Participants’ responses to Likert questions are visualized using diverging stacked bar graphs [151]. The segments of each bar indicate the percentage of responses for each Likert option, with the conditions of the study along the Y-axis. The neutral response is centered horizontally, with negative responses to the left and positive responses to the right.

12.3.1 Text Decoration Style for Highlighting

Our RQ3.3. asked which text decoration style DHH users would prefer for highlighting important words in captions. Specifically, we compared three markup styles for this analysis: underlining, italicizing and boldfacing. For each video stimulus, participants responded to three questions, which were enumerated above as a, b, and c. For each question, we performed a Kruskal-Wallis test to determine if there was a statistical difference in participants’ responses across conditions. For post-hoc pairwise testing, Mann-Whitney U-tests with Bonferroni corrections were used.

Figure 12.4 displays results for I was able to identify the important words and concepts from phase 1 of the study. There was a significant difference
in responses across conditions: ($\chi^2(2) = 12.59; p=0.001$). Post-hoc pairwise analysis via Mann-Whitney U-tests revealed the following:

- **Boldface vs Italics**: Statistical difference [$U=1846, p=0.001$] with means $[\mu_b = 3.97; \mu_i = 3.5]$ respectively.

- **Boldface vs Underline**: No difference observed ($p>0.05$)

- **Underline vs Italics**: Statistical difference [$U=1978.5, p=0.011$] with means $[\mu_u = 3.8; \mu_i = 3.5]$ respectively.

Figure 12.5 displays results for *I found the captions distracting* from phase 1 of the study. There was a significant difference in responses across conditions: ($\chi^2(2) = 8.544; p=0.013$). Post-hoc pairwise analysis via Mann-Whitney U-tests revealed the following:

- **Boldface vs Italics**: No difference observed ($p>0.05$).
I found the captions distracting.

Figure 12.5: Participants’ responses to Phase 1 of the study, comparing decoration styles (italics, boldface and underline) for question *I found the captions distracting*, with significant differences marked with asterisks.

- **Boldface vs Underline**: Significant difference \([U=2036, \ p=0.027]\) with means \([\mu_b = 2.44; \ \mu_u = 2.93]\) respectively.

- **Underline vs Italics**: No difference observed \((p>0.05)\).

It was easy to read the captions

Figure 12.6: Participants’ responses to Phase 1 of the study, comparing decoration styles (italics, boldface and underline) for question *It was easy to read the caption*, with significant differences marked with asterisks.

Figure 12.6 displays results for *It was easy to read the caption* from phase
1 of the study. Statistical differences were found across the three conditions: 
\( \chi^2(2) = 9.590; p=0.008 \). Post-hoc pairwise analysis via Mann-Whitney U-tests revealed the following:

- **Boldface vs Italics**: Significant difference \([U=1945, p=0.005]\) with means \([\mu_b = 4.13; \mu_i = 3.77]\) respectively.

- **Boldface vs Underline**: Significant difference \([U=2097, p=0.031]\) with means \([\mu_b = 4.13; \mu_u = 3.86]\) respectively.

- **Underline vs Italics**: No difference observed \((p>0.05)\).

In summary, participants preferred boldface for indicating important words in captions, which is most clearly suggested by responses to the *It was easy to read the caption* question. This finding differs from the findings of our underpowered, formative studies in Chapter 11, which had recommended underlining for such highlighting.

### 12.3.2 Granularity for Highlighting

Similarly, RQ3.4. investigated whether users preferred highlighting at the word-level or the sentence level, with video stimuli of both conditions displayed side-by-side for comparison, with participants answering questions for each. Figure 12.7 and 12.8 displays the responses, and our analysis revealed:

1. For the *I was able to identify the important words and concepts* question, 
   a Wilcoxon signed-rank test revealed a significant difference across condi-
Figure 12.7: Participants’ responses to Phase 2 of the study, which compared the granularity of highlighting (at the sentence-level or the word-level), for question *I was able to identify important words and concepts*, with significant differences marked with asterisks.

Figure 12.8: Participants’ responses to Phase 2 of the study, which compared the granularity of highlighting (at the sentence-level or the word-level), for question *I found the captions distracting*, with significant differences marked with asterisks.

2. For the *I found the captions distracting* question, a Wilcoxon signed-
rank test revealed a significant difference across conditions: \( U=195.5 \), \( p=0.001 \), with means at \( \mu_{\text{word}} = 2.58; \mu_{\text{sentence}} = 3.16 \).

In summary, the results reported above for phase 2 suggest that participants preferred highlighting with granularity at the word-level, rather than highlighting entire sentences.

### 12.3.3 Handling Key Term Repetition

RQ3.5. investigated whether users preferred highlighting only the first occurrence of a repeating keyword or multiple occurrences for that keyword (anywhere in the text it had been indicated as being important). Similar to phase 2, video stimuli were shown side-by-side in both conditions, with participants answering two questions about each video.

Figure 12.9: Participants’ responses to Phase 3 of the study for whether repeated keywords should be highlighted only once (once) or every important occurrence (always), for question \textit{I found the captions distracting}, with significant differences marked with asterisks.

1. For the \textit{I was able to identify the important words and concepts} question,
a Wilcoxon signed-rank test did not reveal any significant difference across the conditions.

2. For the *I found the captions distracting* question, a Wilcoxon signed-rank test revealed a significant difference across conditions: $[U=317.5, p=0.001]$ with means at $[\mu_{once}=2.40; \mu_{always} = 2.66]$. Participants’ responses are shown in Figure 12.9 for this question item.

In summary, participants found it less distracting when the system highlighted only the first keyword occurrence.

### 12.3.4 Interest in Highlighting Applications

![Highlighting Preferences across different genre of videos and captions](image)

Figure 12.10: Responses in Phase 4, on whether participants thought that highlighting important words in captions would be useful for six different applications, with significant pairwise differences marked with asterisks.

RQ3.6. explored the degree to which participants indicated their agreement with a statement that they *think highlighting important words in a caption would be useful for* various applications. A Kruskal-Wallis test indicated a significant
difference for participants’ responses across the six captioning application areas, and post hoc pairwise statistical difference testing was performed using Mann-Whitney U-tests with Bonferroni corrections. Figure 12.10 displays participants’ responses on the different applications and indicates significant pairwise differences.

12.4 Discussion of the Results

For RQ3.3., it is notable that prior style guidelines for authors of static text had recommended using italics for highlighting. While the underpowered, formative studies we conducted in Chapter 11 had led us to use underlining to highlight text in captions in educational videos for DHH users [78], the statistically significant results of this larger study contradict that earlier result. In contrast, our study, with a larger number of DHH participants, found that users favored boldface for highlighting in captions of such videos. Specifically, our participants reported that the boldface was easier to read in comparison to underlining or italicizing, and they reported that italicizing was less effective at enabling them to identify important words and concepts in the video. In open-ended discussion, some participants commented:

* I didn’t like italics because I have poor vision. If someone is on the spectrum of blindness, it would be difficult to distinguish the difference. I wouldn’t mind bold and underlined. - P28

* It is hard to tell the difference between italicized and normal. - P8
On the other hand, if a markup style is too noticeable (referred to as the pop-out effect in prior work \[159\]), there is a risk it may be too distracting, i.e. drawing too much attention from the reader. In our study, participants reported that when viewing captions containing underlining, captions were more distracting, compared to the boldface markup. Some participants also commented on this issue:

\textit{I think underlining is too distracting, bold and italicized is easier.} - P15

The underline makes it difficult to read. - P15

In addition, participants indicated that some forms of highlighting do more than just convey importance: Some also suggest an emotional subtext, which may unintended, e.g. with one participant commenting that italics \textit{“makes the speaker sound insincere or sarcastic.”} (P2)

For RQ3.4., when participants were asked about their preferred granularity for highlighting in captions, they indicated that captions highlighted at the word-level better enabled them to identify important words and concepts in the videos. Some participants referred to the sentence-level highlighting as \textit{“irregular”} (P23), and others disliked the long spans:

\textit{After the first underlined sentence, there’s a long period of time where the next sentences are not underlined; so, my brain automatically stopped paying attention and registering them.} - P2

Participants' subjective responses also indicated that sentence-level highlighting was more distracting than word-level highlighting, with some commenting about this issue:
A bit lost with the entire sentence underlined; focused on the lines rather than words. - P26

For RQ3.5., participants indicated their preference on highlighting repeating keywords in captions, with some indicating that repeated highlighting was “annoying” (P32).

For RQ3.6., participants provided subjective responses to a question about whether highlighting during captions would be useful in various application contexts. While many indicated that it would be useful for online educational videos (e.g. the stimuli in this study), participants felt it would be less useful for entertainment videos or one-on-one meetings in the real-world. Participants commented:

No, it depends! If it is work/academic-related, yes! But, [in] movies or entertainment, no! - P5

Yes, ...government videos like congress or weather - P21

Yeah, depends on the subject. If it’s more like an activity class, such as dance, gym then maybe not! Whereas, in an intense class it would be nice to underline the important information. - P19

Yes, definitely! Any situation where you would need to take lots of notes such as classroom, meetings, interviews, etc. - P28
12.5 Conclusions

Our initial study, discussed in Chapter 11, on the benefits of highlighting in captions for DHH individuals, had reported a usability benefit from highlighting important words in captions of educational videos for DHH users under a specific combination of design parameters in the stimuli presented. However, our research had not systematically investigated key design parameters for such highlighting. Considering recommendations in related work, including style guidelines for authors of static texts, we compared various design options for highlighting of important words in educational video captions in an experimental study with 36 DHH individuals, which has been presented in this chapter.

The main contributions of this work are empirical: We found that DHH users preferred boldface decoration and word-level granularity for highlighting important words in captions, and they preferred only highlighting the first occurrence of any recurring keywords in the text. We also presented some open-ended feedback from DHH participants who had experienced these caption highlighting variations, to contextualize these quantitative results. In addition to the specific guidance we provide for the design of keyword highlighting during captioned educational-lecture videos for DHH users, our study has also identified additional potential applications for caption highlighting in future work. More broadly, our study suggests that established style guidelines for text highlighting, originally intended for static texts and for general readers, may not translate to this user group nor the dynamic nature of captions, suggesting a need for further research into other groups and other dynamic text settings.
Part III of the work investigated highlighting in captions as a way to enhance the understandability of caption text for DHH users. We presented the research activities in the part in two phases: In the first phase (discussed in Chapter 11), we explored the benefits of importance-based highlighting in captions for DHH users, specifically when viewing education lecture videos. However, there were some limitations in that study. For instance, our interviews in the formative studies only consisted of a few DHH users who guided our selection of educational lecture videos as a suitable genre for use of highlighted captions. Further, those few interviews were also used to select the text-decoration style for highlighting. To overcome these challenges, the second part of our work (discussed in Chapter 12) investigated DHH users’ preference for the different design choices for highlighting in captions.

To summarize, Part III of our work proposes to explore the following research question:

RQ3: Do DHH users benefit from importance-based highlighting in
captions? If so, what are their highlighting preferences?

As discussed earlier, this research question was discussed in two separate chapters: Chapter 11 explored the benefits of importance-based highlighting in captions for DHH users through a user-based study where users were shown video with and without highlighting in caption, and were asked to provide subjective rating on different measures, such as readability of the captions, their understandability and several other task-load performance measurement question (e.g., NASA TLX [63]). Similarly, Chapter 12 evaluated several design configurations for highlighting in captions with a user-based studies with DHH participants. The different design configurations evaluated in the study included: text-decoration style for highlighting, the granularity of text selected for highlighting, and strategies for handling repetition of key terms selected for highlighting.

To answer this research question more thoroughly, we had designed several sub-research questions, the results to which are summarized below:

**RQ3.1: Are DHH users receptive to the premise of importance-based word highlighting in captions?** This research question was investigated in Section 11.2.7. The results of this investigation were a part of the pre-survey questionnaires where the participants were asked about experience with importance-based highlighting for video captions. Overall, participants reported that they were fairly open to the premise of highlighting in captions.

**RQ3.2: When viewing captions of online lecture video, do DHH
users subjectively prefer highlighting in captions? The results of our larger study in Chapter 11 discussed in Section 11.3.4 indicated that users’ found videos containing highlighted words to be easier to read and follow compared to the videos without any highlighting. Further, users reported that highlighting in caption allowed them to identify important words and concepts in the video, with lower mental and perceptual load reported when reading and understanding the captions.

RQ3.3: What text-decoration strategy would DHH users prefer for highlighting important words in captions? Based on prior recommendations, which included our formative studies in Chapter 11 and other style-guidelines for static text documents, we selected three markup styles of comparison for highlighting in captions. The results of this study was discussed in Section 12.3.1. In summary, participants preferred **boldface** markup for indicating important words in captions, which was reported to be *easier to read* in captions than other markup-styles in comparison.

RQ3.4: Do DHH users prefer highlighting important words in captions or highlighting entire sentences? This research question is also investigated by the user-study described in Chapter 12. Specifically, the results described in Section 12.3.2 shows that DHH users prefer word-level highlighting in captions that sentence-level highlighting. Participants in the study reported that sentence-level highlighting
was *distracting* and they were unable to identify the importance of words and concepts in the video with this choice of granularity for highlighting in captions.

**RQ3.5: When the same word is identified as important in multiple occurrences in the captioning for a single video, do DHH users prefer that the first occurrence be highlighted only or multiple occurrences?** Investigated in Chapter 12, the results of the research question (presented in Section 12.3.3) showed that users prefer highlighting only the first occurrence of the key terms.

**RQ3.6: For what genres of video would DHH users also indicate an interest in seeing the highlighting of important text in captions?** Lastly, the user study in Chapter 12 investigated DHH users’ preferred applications for highlighting in captions. The results of this investigation are presented in Section 12.3.4, which showed that users were excited about caption highlighting in the educational genre of videos, even more so than in entertainment-type video genres.
Chapter 13

Limitations and Future Work

This chapter discusses several limitations of this research, organized in sections below. As a future work, we provide possible opportunities in overcoming these limitations through additional exploration and research in these topics.

13.1 Word Importance Modeling

In Part I of the research, we presented several models for predicting the importance of words in spoken dialogues for understanding its meaning; ranging from unsupervised models based on the estimating predictability of words, to supervised models trained (and evaluated) on the Word Importance Annotation corpus. However, the proposed methods are not exhaustive. Below we discuss some avenues for future research that could help create better models for predicting the importance of words:
13.1.1 Modeling Importance at a Larger Semantic Units

Word is often considered a fundamental unit of meaning (and prosody) in many linguistic (and speech) based applications. However, it is not always a holistic unit of meaning in a real-world use. For instance, “New York” is a composition of two words in English which represents meaning at a phrasal-level than at the word-level. Therefore, the assumption of word-level granularity in meaning is not always be an ideal one. Perhaps, future research could investigate if considering larger semantic units in text (and speech) can help improve the importance prediction models.

13.1.2 Unsupervised (and Semi-supervised) Models of Word Importance

Unsupervised and semi-supervised approaches to build word importance models are important avenues for future research because these models are less dependent (or reliant) upon the human collected data of word importance. This is a useful property to have because collecting human-labelled data of word importance is a time and resource intensive task. While Section 4.2 presented word predictability based unsupervised estimation of word importance in spoken dialogues, research on to other possible methods of word importance estimation are unexplored – which includes: exploration of the effects of other psycho-linguistic properties of text such as surprisal (estimated using various types of language models) in the task of word importance prediction, semi-supervised approaches such as pre-training a larger-scale neural model
through a language-model objective and fine-tuning it for the word importance prediction task, task-specific word importance modeling, etc.

13.2 Automatic Caption Quality Evaluation

As speech technology researchers continue to improve the ASR technology, it becomes important to be able to evaluate (and compare) the performance of these systems effectively for various applications. With this goal, research on Part II of this thesis evaluated various existing metrics for evaluating ASR systems for captioning applications for DHH users. Motivated by the inefficacy of these metrics in predicting the usability of automatically generated caption for DHH users, we presented two new metrics (ACE and ACE2), and evaluated them through various studies with the users. While our new ACE2 metric outperformed pre-existing metrics at predicting the subjective judgments of DHH users as to the quality of caption texts, we still see room for improvement:

One limitation of our automatic caption evaluation framework is that the framework considers one error at a time (fixing all other errors) in order to measure each error’s individual impact on the understandability of the text. While for some applications, it may be useful to identify the individual contribution of specific errors in a text, in other contexts, it may be beneficial to consider models of error impact that represent more complex interactions among multiple errors in a text. To support such research, it would be valuable to collect more qualitative data on the impact of errors in text comprehension studies with DHH participants. Such a large data resource could enable
CHAPTER 13. LIMITATIONS AND FUTURE WORK

researchers to learn more complex inferences on impact due to multiple errors, without resorting to a more “controlled” single-error based analysis.

It should also be noted that the performance of each of the metrics have been evaluated based on their ability to predict the quality of the transcription of a full utterance unit. While this assumption is reasonable in a conversational setting, where the conversation is more dyadic, this approach would be less applicable if captioning technology were to be used to support a single-speaker channel, such as in the classroom or live lecture. Thus, a formal evaluation of the metric performance in measuring the quality of longer texts spans rather than individual conversational utterance units could also help take this research forward. Essentially, it would be necessary to perform automatic segmentation of the longer text transcript generated by the ASR and use the individual sentences/utterances identified in this longer text as the basis for evaluation.

We also foresee additional opportunities in boosting the performance of the ACE2 metric, through additional research on models of the importance of words in a text and of the semantic distance between error words and the intended word. More specifically, there are opportunities to investigate other supervised approaches for word-importance prediction, e.g. models talked about in Chapter 6 to improve the model. Several other semantic-distance models, e.g. based on additional semantic features (such as POS-tags, sentiment, polarity of words, etc.) or by identifying vector representations of words that are better suited to calculating semantic distance, e.g., Nguyen et al. [123] could also be explored as future research directions.

Further, a user-study with a larger number of users could be beneficial to
explore the generalizability of the results in our evaluation studies. This is because there is a huge diversity in the DHH population, from differences in the reading literacy skills to differences in various other life experiences such as the familiarity with captioning services – and a larger study could incorporate this diversity to a greater extent in our experiments.

Lastly, other hyper-parameter choices in the development of the evaluation metrics such as the choice of embeddings (e.g., word2vec) for word feature representation and handling insertion and deletion errors based on empirically determined scaling-factors could benefit from more rigorous exploration. Additional future research would be needed to determine whether these results are robust to variations in these technical choices.

13.3 Highlighting in Captions to Improve Caption Usability

Lastly, in Part III of our research, we investigated importance-based highlighting in captions for DHH users. Although captioning services provide access to spoken content for DHH users, caption texts demand visual attention for comprehension – which require DHH users to strategically switch between the captions and other visual information in video content. This can be especially challenging when there are multiple concurrent streams of visual information and/or visual references within the captions. To overcome these challenges, Part III of our research investigated highlighting in captions as a way to enhance the usability of the captions.
While research in highlighting in captions for DHH users yielded some interesting findings (see results in Section 11.3 and 12.2), our studies were not short of limitations:

One limitation of our studies is that we conducted our experimental evaluations (e.g. various design options for educational lecture videos) using only short segments of videos, from freely available sources of video content: the video research dataset of Harper [62] and/or TED Talks. There is a risk that our results may not generalize to videos with a different genre of educational content or with different visual complexity, e.g. lectures with complex animations or simulations that require careful visual attention simultaneous to the narration.

Another potential limitation is that the preferences of users during short videos may differ from their preferences from longer viewing experiences. We see risk, especially about this topic in regard to the issue of whether to highlight only the first occurrence of a keyword that had been marked as important multiple times in a video. While in this work, we found that users preferred marking only the first occurrence, after a long duration in a long video, they may tolerate a repeated highlighting of the same keyword. Both of these factors could be investigated in future work.

Another limitation is that we have investigated several research questions through a single study, rather than conducting independent studies with separate recruitment of participants for investigating each research question. For instance in Chapter 12, we investigate four separate research questions through a single study. While this choice was necessary for making recruiting sufficient participants feasible, it meant that we were not able to cascade findings from
one research question when planning the video stimuli for the next. For instance, the videos shown in phase 2 of our study, to investigate RQ3.5. (granularity) used underlining for text decoration, in accordance with the recommendations of our previous research study in Chapter 11. In fact, our participants preferred boldface captioning during phase 1 of our study. In future work, a study with a sufficiently large number of DHH participants, could compare the design factors we examined in a fully factorial design, to investigate the potential of interaction effects.

Lastly, future work could also investigate the generalizability of our results across different tasks and application contexts; we could investigate if highlighting would benefit other groups of users or would be useful for other video genres or other communication scenarios, such as live captioning in multi-party meetings.

13.4 Using Word-Importance Models during the Training or Decoding of ASR Systems

So far, in this research, we have discussed several approaches to predict semantic importance of words in spoken dialogues which has been beneficial in designing metrics of ASR output quality in captioning applications for DHH users and, also in providing informative highlights in captions for enhancing their usability for these users. These findings suggest that there could be value in building ASR systems that are designed to optimize for the accuracy of important words (rather than all the words), especially for enhancing the
understandability of the caption output for the users.

There could be several different opportunities during the training and the decoding pipeline of ASR systems to promote for understandability in the output of ASR systems. As discussed in Section 2, conventional ASR systems find hypotheses text i.e. sequence of word transcriptions from speech signals according to the following equation:

$$\hat{W} = \arg\max_W P(W|X) = \arg\max_W P(W)P(X|W)$$  \hspace{1cm} (13.1)

Equation 13.1 is also called the maximum a posteriori (MAP) estimation. However, the problem with MAP based approach to speech recognition is that it is suboptimal with respect to minimizing the number of word errors in the system output. Using a Bayesian Decision Theoretic framework, Stolcke et al. [158] showed that the MAP-based decision rule in Equation 2.1 minimizes the bayes risk under the 0-1 valued loss function, as shown in Equation 13.2 and 13.3.

$$\delta(X) = \arg\min_{W \in W} \sum_{W' \in W} l_{SER}(W, W')P(W'|X).$$  \hspace{1cm} (13.2)

$$l_{SER}(W, \delta(X)) = \begin{cases} 0 & \text{if } \delta(X) = W, \\ 1 & \text{otherwise.} \end{cases}$$  \hspace{1cm} (13.3)

This framework, however, also provides a flexible way to incorporate a custom loss function into the decoding process of ASR, by replacing the loss function $l_{SER}(., .)$ with a custom loss function. Since current ASR systems are evaluated
using the WER metric, researchers have investigated using this framework to explicitly improve the ASR systems to minimize the word error rate.

Below, we present two different research strategies, which have been discussed in the past, that offers enough flexibility to explore word-importance based loss functions (such as ACE and ACE2) for optimizing or re-ranking the ASR results – which may be fruitful avenue of research in future works.

13.4.1 N-best list Re-scoring Technique

In an ASR system, the function of the decoder is to find the most likely word sequence given the sequence of audio features. Although decoders are designed primarily to find a single solution, in practice, it is relatively simple to generate not just the most likely hypothesis but the n-best set of hypotheses. Therefore, in most ASR systems, along with the most likely word sequence, a list of n-best hypotheses can also be obtained as output. Other compact forms of representation of this n-best hypotheses list are also commonly used such as a word lattice representation [148] or a confusion network [109].

These representations have been popular especially because they provide a reduced search-space (out of all possible word sequence) that can be further decoded, with more flexibility, to improve the ASR output. This post processing technique also allows for general-purpose hypothesis to be tuned in a domain-specific or user specific way without having to design the whole ASR engine to do so [149]. Furthermore, the n-best hypotheses generated as an output from the ASR system can be processed with complete independence from the ASR system; thus, it can be treated as a separate stage in an ASR pipeline.
Researchers [20, 43, 126, 150, 158, 173] have utilized various rescoring techniques to select the best hypothesis from an ASR n-best hypotheses. This includes investigating the decision rule described in Equation 2.1 to re-score the n-best list output. More specifically, researchers have investigated replacing the loss function in Equation 2.1 with WER loss to explicitly minimize the word error rate, as follows:

\[
\text{loss}_{\text{WER}}(W, X) = \sum_{W' \in \mathbf{W}} \text{WER}(W, W') P(W' | X). \quad (13.4)
\]

This decoding strategy have been shown to provide statistically significant improvements in recognition task as compared to MAP based decoding as it explicitly incorporates task performance criterion to the decoding process of ASR [52, 158]. Successes of hypotheses scoring systems like ROVER [43] (and its variants) has been credited to this decoding strategy to directly improve WER. Several different contributions has been made to this approach of decoding in more recent years [36, 53, 54].

### 13.4.2 Improved Optimization Strategy (End-to-End Models)

Attention-based end-to-end models, discussed in Section 2.2, are typically trained by optimizing cross-entropy loss i.e. maximizing the log-likelihood probability of the training data. Equation 13.5 show the loss computed for training data sample.

\[
\text{loss}_{\text{CE}} = \sum_{n=1}^{N+1} -\log P(w_n | w_{n-1}, ..., w_0, X) \quad (13.5)
\]
However, this training criterion does not directly optimize the metric that is used to evaluate the ASR system, which is traditionally the WER metric. Therefore, researchers have investigated approaches to optimize ASR systems using a loss that directly minimizes or is correlated with minimizing the WER. Consequently, researchers have looked into the decision theoretic perspective provided by Stolcke et al. [158] to create appropriate training loss for these systems—similar to Equation 13.4.

Since, minimizing expected WER directly is intractable, as it involves a summation over all possible label sequences, researchers have approximated the expectation using either the n-best list [158] ($W \approx W_{N_{\text{best}}}$) or other sampling-based approaches [155].

Interestingly, both of these research work discussed above has only investigated the use of WER metric at their primary loss function for optimizing or re-ranking the ASR output. While this approach has resulted in positive results for various ASR models [27,52,135], experiments on other custom-loss functions that penalize ASR errors based on the importance of words have not been explored.
Chapter 14

Summary and Contributions

This dissertation has presented research that addresses the issues of trust and acceptance of Automatic Speech Recognition (ASR) systems especially when envisioning them as a captioning tool for supporting communication for people who are Deaf or Hard of Hearing (DHH). These issues mainly stem from the fact that these automatic systems have not been rigorously evaluated as to how well they provide access services for DHH users, and they are not fully dependable as they still produce errors in real-world use-case scenarios. We mainly explore two challenges, and we phrase them as two important questions listed below:

1. *How do we know when ASR systems are performing well for captioning applications for DHH users?* (We need better metrics for evaluating the quality of ASR-generated outputs for captioning applications for DHH users. Answering this question may reveal ways to *evaluate* and *compare*
the performance of ASR systems specifically for captioning applications for DHH users.)

2. How do we design to enhance the usability of ASR systems for captioning applications for DHH users? (Designing ASR systems to promote for their usability in captioning applications could help address the issues of trust and acceptance of these systems for the end-users.)

In this chapter, we summarize the contribution of the research activities presented in this work, their limitations, and potential directions for future research.

14.1 Summary of the Contribution of This Research

The research activities presented in this thesis are organized into three parts, as presented below. We summarize the contribution of each part of this research as follows:

Part I of this work investigated approaches to estimate the importance of words to the meaning conveyed in spoken dialogues. The main goal of this work was to facilitate research into the issue of the evaluation (discussed in Part II of this research) and usability (discussed in Part III of this research) of automatic captioning services. This work presented several statistical models that operated on features from text and/or speech for representing meaning for the task of importance prediction of words in spoken dialogues. The contribution of the research in Part I of this thesis are as follows:
1. **Dataset Contribution**: In Chapter 5, we presented the Word Importance Annotation corpus which contains over 25,000 tokens manually annotated with word importance information. We demonstrated the efficacy of this corpus in training and evaluating machine-learning models of word importance. This corpus has been made publicly available\(^1\) to enable future research in this field.

2. **Methodological Contribution**: As a part of the effort to build the Word Importance Annotation corpus, Section 5.2 presents the details of the Word Importance Annotation task consisting of several annotation schemes designed to help the development of the Word Importance Annotation corpus. These guideline will be useful for future researchers interested in extending this corpus for various purposes such as enabling the development of more powerful machine-learning models for importance prediction.

3. **Empirical Contribution**: Chapter 6 discusses several supervised models of word importance that are based on the Word Importance Annotation corpus. Performance evaluations of these models showed that incorporating speech-based features into text-based representation of meaning can improve the word importance prediction models. We compared several feature-fusion techniques for combining speech-based and text-based feature representations in spoken dialogues and evaluated them for the word importance prediction task. Our method of adjusting (shifting)

\(^1\)http://latlab.ist.rit.edu/lrec2018.
the text-based feature representation based on the semantic knowledge inferred from the acoustic-prosodic characteristics of speech showed the best performance for this task.

**Part II** of this work investigated challenges in the evaluation of ASR systems for captioning applications for DHH users. The research work in this part discussed the design and evaluation of automatic metrics of ASR success, especially when focusing on creating real-time communication applications for DHH users. Through studies with several DHH users, we collected subjective judgement on the quality of various automatically generated transcripts (containing errors) which allowed us to compare several automatic metrics for measuring ASR output quality, including the WER metric. In the study, our metric that considered the differential importance of words when computing the quality of the ASR-generated captions had a higher correlation with DHH users' subjective judgements on their quality in comparison to the other metrics including the WER metric. The contributions of the research work discussed in this part includes the following:

1. **Empirical Contribution:** As a part of the research effort in understanding the usefulness of automatic metrics of ASR quality in captioning applications, our study in Section 9.3 revealed that the traditional metrics of ASR output quality (such as the WER metric) do not correlate well with DHH users' subjective judgements of quality. Our empirical findings established the need for better metrics of ASR quality to evaluate these systems, especially when considering captioning applications for DHH
users.

2. **Theoretical Contribution**: In Section 9.1, we present the Automatic Caption Evaluation (ACE) framework which provides a way to measure the impact of a recognition errors on the understandability of a text by considering two sub-factors: the importance of word, and the semantic deviation due to errors. We demonstrated the utility of this framework by developing two metrics (ACE and ACE2) for evaluating the ASR output quality, and evaluated them against the DHH users' subjective judgement of quality on various ASR generated transcripts. Our results show that our proposed metrics show higher correlation scores with DHH users' judgement on the quality of transcripts containing errors, in comparison to the traditional evaluation metrics (e.g., WER).

**Part III** of this work investigated the word-importance-based highlighting in captions as a way to enhance their usability for DHH readers. While the benefits of highlighting in static text (e.g. textbooks or electronic documents) is widely known, the benefits of highlighting in captions had been relatively understudied. With this motivation, research work in this part investigated the benefits of highlighting in captions for DHH users, especially when viewing educational lecture-type videos. Further, we provide empirical results on the design preferences of highlighting in captions through experimental studies with DHH users. In summary, our results show that DHH users find highlighting key words in captions to be beneficial; DHH participants in studies reported higher subjective measurements on caption readability and understandability.
In addition, they reported lower perceptual and mental task-load performance scores when viewing videos with captions containing highlighting in comparison to videos without highlighting. Further, DHH participants indicated a preference for **boldface** decoration style of highlighting in captions with word-level and non-repeating highlights. The contributions of the research activities described in this part of this dissertation are as follows:

1. **Empirical Contributions**: With the goal of investigating the benefits of text highlighting in captions, the study in Chapter 11 collected DHH users’ responses to several caption (and video) usability questionnaires after viewing videos with and without highlighting of importance words in the captions. The results of this study showed that DHH individuals are open to the premise of highlighting in captions, with DHH participants subjectively reporting higher readability and understandability scores and lower task-load measurement scores when viewing videos with highlighting in captions. Further, as a follow-up study, we conducted another a study (described in Chapter 12) with DHH individuals to gather their preferences on various design options for highlighting the important words in video captions. This study provided empirical results on DHH users’ preferences in regard to: 1) the granularity for selecting highlighting in captions; 2) the decoration to use when highlighting; and, 3) whether to highlight repeated occurrences of the same keyword in a caption text or only first occurrence.

2. **Methodological Contributions**: The question-types and empirical re-
results presented in the studies discussed in this part of the research could be beneficial for future researchers when investigating other design parameters in captions, or when evaluating automatic methods for identifying important words for users in educational video contexts, with our results as a potential baseline.

14.2 Final Comments

This research will increase the usability of convenient, available, and low-cost ASR-based software for use in captioning for DHH users, for a more effective communication in the workplace or educational settings for these users. Our published research findings and disseminated resources (word importance corpus, models of word importance, stimuli presented in user studies, data collection tools and platforms) will also enable additional researchers to investigate ASR-based captioning for DHH users – and to investigate ASR or linguistic technologies in other applications. For instance, models of word-importance for conversation text transcripts may also inform the design or evaluation of real-time machine translation systems for supporting live conversations between users of different spoken languages.

A major barrier to the utilization of ASR systems for the task of captioning for DHH users has been a lack of trust and acceptance of such systems, especially given the sometimes low quality of ASR output. Although these technologies are improving in their quality, research is needed to understand how to best evaluate, optimize, and embed these technologies in useful acces-
sibility applications for DHH users. A backdrop to this work is that there are concerns among the DHH community that cost-savings could motivate automatic captioning to be deployed before they are fully usable, potentially replacing current services (e.g., ASL interpreting). For this reason especially, research is needed on the evaluation of ASR systems for the task of real-time captioning, in realistic meeting and communication environments. Further, research into the design of these systems specifically for captioning applications may add to the appeal of these systems for these tasks. Through computational linguistic and user-centered accessibility research, there may be an exciting future for these technologies, to benefit DHH users.
Bibliography


ACM SIGACCESS Conference on Computers and Accessibility, ASSETS 2016, Reno, NV, USA, October 23-26, 2016, Jinjuan Heidi Feng and Matt Huenerfauth (Eds.). ACM, 25–32. DOI:http://dx.doi.org/10.1145/2982142.2982177


BIBLIOGRAPHY


[120] Andrew Cameron Morris, Viktoria Maier, and Phil D. Green. 2004. From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition. In *INTERSPEECH 2004 - ICSLP, 8th International Conference on Spoken Language Processing, Jeju Island, Ko-


Appendices
Appendix A

Publications

Peer-Reviewed Journal Articles


Peer-Reviewed Conference Articles


APPENDIX A. PUBLICATIONS


Other Technical Papers and Pending Submissions


Appendix B

IRB Approval Forms

All of the studies presented in this thesis has been approved by the Institutional Review Board (IRB). Below, we provide the IRB decision form for two projects:

- *Creating the Next Generation of Live-Captioning Technologies*: This IRB covers the ASR evaluation studies presented in Part 2 for this work.

- *Identifying the Best Methods for Displaying Word-Confidence in Automatically Generated Captions for Deaf and Hard-of-Hearing Users*: This IRB covers the caption highlighting studies presented in Part 3 of this work.
Figure B.1: IRB Decision Form for "Creating the Next Generation of Live-Captioning Technologies".
Figure B.2: IRB Decision Form for “Identifying the Best Methods for Displaying Word-Confidence in Automatically Generated Captions for Deaf and Hard-of-Hearing Users”.