Modeling Acoustic-Prosodic Cues for Word Importance Prediction in Spoken Dialogues

Sushant Kafle, Cecilia O. Alm, Matt Huenerfauth

Rochester Institute of Technology, Rochester NY

{sxk5664,matt.huenerfauth,coagla}@rit.edu

Abstract

Prosodic cues in conversational speech aid listeners in discerning a message. We investigate whether acoustic cues in spoken dialogue can be used to identify the importance of individual words to the meaning of a conversation turn. Individuals who are Deaf and Hard of Hearing often rely on real-time captions in live meetings. Word error rate, a traditional metric for evaluating automatic speech recognition (ASR), fails to capture that some words are more important for a system to transcribe correctly than others. We present and evaluate neural architectures that use acoustic features for 3-class word importance prediction. Our model performs competitively against state-of-the-art text-based word-importance prediction models, and it demonstrates particular benefits when operating on imperfect ASR output.

1 Introduction

Not all words are equally important to the meaning of a spoken message. Identifying the importance of words is useful for a variety of tasks including text classification and summarization (Hong and Nenkova, 2014; Yih et al., 2007). Considering the relative importance of words can also be valuable when evaluating the quality of output of an automatic speech recognition (ASR) system for specific tasks, such as caption generation for Deaf and Hard of Hearing (DHH) participants in spoken meetings (Kafle and Huenerfauth, 2017).

As described by Berke et al. (2018), interlocutors may submit audio of individual utterances through a mobile device to a remote ASR system, with the text output appearing on an app for DHH users. With ASR being applied to new tasks such as this, it is increasingly important to evaluate ASR output effectively. Traditional Word Error Rate (WER)-based evaluation assumes that all word transcription errors equally impact the
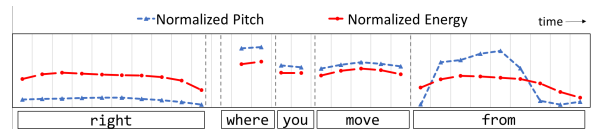

Figure 1: Example of conversational transcribed text, *right where you move from*, that is difficult to disambiguate without prosody. The intended sentence structure was: *Right! Where you move from?*

quality of the ASR output for a user. However, this is less helpful for various applications (McCowan et al., 2004; Morris et al., 2004). In particular, Kafle and Huenerfauth (2017) found that metrics with differential weighting of errors based on word importance correlate better with human judgment than WER does for the automatic captioning task. However, prior models based on text features for word importance identification (Kafle and Huenerfauth, 2018; Sheikh et al., 2016) face 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 (McKeown et al., 2005).

- **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 (Frazier et al., 2006),

such information may be omitted from a text transcript, as in Figure 1, 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 on 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. Our 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.

2   Related Work

Many researchers have considered how to identify the importance of a word and have proposed methods for this task. Popular methods include frequency-based unsupervised measures of importance, such as Term Frequency-Inverse Document Frequency (TF-IDF), and word co-occurrence measures (HaCohen-Kerner et al., 2005; Matsuo and Ishizuka, 2004), which are primarily used for extracting relevant keywords from text documents. Other supervised measures of word importance have been proposed (Liu et al., 2011, 2004; Hulth, 2003; Sheeba and Vivekanandan, 2012; Kafle and Huenerfauth, 2018) for various applications. Closest to our current work, researchers in (Kafle and Huenerfauth, 2018) described a neural network-based model for capturing the importance of a word at the sentence level. Their setup differed from traditional importance estimation strategies for document-level keyword-extraction, which had treated each *word* as a *term* in a document such that all *words* identified by a *term* received a uniform importance score, without regard to context. Similar to our application use-case, the model proposed by Kafle and Huenerfauth (2018) identified word importance at a more granular level, i.e. sentence- or utterance-level. However, their model operated on human-generated transcripts of text. Since we focus on real-time captioning applications, we prefer a model that can operate without such human-

produced transcripts, as discussed in Section 1.

Previous researchers have modeled prosodic cues in speech for various applications (Tran et al., 2017; Brenier et al., 2005; Xie et al., 2009). For instance, in automatic prominence detection, researchers predict regions of speech with relatively more spoken stress (Wang and Narayanan, 2007; Brenier et al., 2005; Tamburini, 2003). Identification of prominence aids automatically identifying content words (Wang and Narayanan, 2007), a crucial sub-task of spoken language understanding (Beckman and Venditti, 2000; Mishra et al., 2012). Moreover, researchers have investigated modeling prosodic patterns in spoken messages to identify syntactic relationships among words (Price et al., 1991; Tran et al., 2017). In particular, Tran et al. 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 (Xie et al., 2009; Murray et al., 2005). 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.

3   Word Importance Prediction

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[1], and assigns an importance label to every spoken word in the turn using a bi-directional LSTM architecture (Huang et al., 2015; Lample et al., 2016).

$$\overrightarrow{h_t} = LSTM(s_t, \overrightarrow{h_{t-1}}) \qquad (1)$$
$$\overleftarrow{h_t} = LSTM(s_t, \overleftarrow{h_{t-1}}) \qquad (2)$$

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

---

[1]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 our train/evaluation corpus; in the future, speech tokenization could be automated.
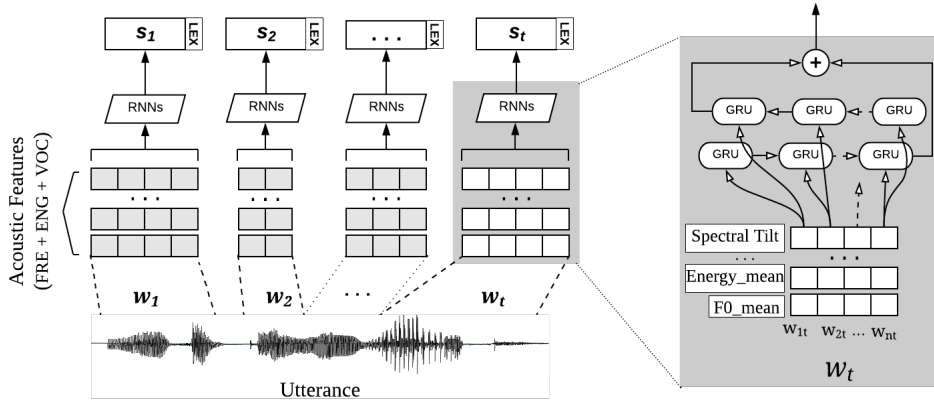
Figure 2: Architecture for feature representation of spoken words using time series speech data. For each spoken word (w) identified by a word-level timestamp, a fixed-length interval window (τ) slides through to get $n = time(w)/\tau$ sub-word interval segments. Using an RNN network, a word-level feature (s), represented by a fixed-length vector, is extracted using the features from a variable-length sub-word sequence.

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.

3.1 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 (HI), medium importance (MID) and low importance (LOW) (details on Section 5). These importance class labels have a natural *ordering* such that the cost of misclassification is not uniform e.g., incorrect classification of HI class for LI class (or vice-versa) will have higher error cost than classification of HI class for 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 (SOFTMAX), a relaxed softmax tailored for ordinal classification (ORD), and a linear-chain conditional random field (CRF) for making a conditioned decision on the whole sequence.

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

Relaxed Softmax Layer. In contrast, the 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 (Cheng et al., 2008) – 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 4 describes

our architecture for acoustic-prosodic feature representation at the word level, and Sections 5 and 6 describe our experimental setup and subsequent evaluations.

4 Acoustic-Prosodic Feature Representation

Similar to familiar feature-vector representations of words in a text e.g., word2vec (Mikolov et al., 2013) or GloVe (Pennington et al., 2014), various researchers have investigated vector representations of words based on speech. In addition to capturing acoustic-phonetic properties of speech (He et al., 2017; Chung et al., 2016), some recent work on acoustic embeddings has investigated encoding semantic properties of a word directly from speech (Chung and Glass, 2018). In a similar way, our 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.

4.1 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 (Brenier et al., 2005). 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 (Tamburini, 2003).

• **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 De Jong and Wempe (2009). 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 (H1 - H2), i.e. the difference between the amplitudes of the first harmonic (H1) and the second harmonic (H2) in the Fourier Spectrum. The spectral-tilt measure has been shown to be effective in characterizing glottal constriction (Keating and Esposito, 2006), which is important in distinguishing voicing characteristics, e.g. whisper (Ito et al., 2002). We also exmined other voicing measures, e.g. Harmonics-to-Noise Ratio and Voiced Unvoiced Ratio.

In total, we extracted 30 features using Praat (Boersma, 2006), as listed above. Further, we included speaker-normalized (ZNORM) version of the features. Thereby, we had a total of 60 speech-based features extracted from sub-word units.

4.2 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 our model, each word was represented as a sequence of these sub-word features with varying lengths, as shown in Figure 2. 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 our final feature vectors. For this task, we utilized Gated Recurrent Units (GRUs) (Cho et al., 2014) as our RNN cell, rather than LSTM units, due to better performance observed during our initial analysis.

5 Experimental Setup

We utilized a portion of the Switchboard corpus (Godfrey et al., 1992) that had been manually annotated with word importance scores, as a part of the Word Importance Annotation project (Kafle and Huenerfauth, 2018). That 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.

Our 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 our data for training, validation, and testing. The prediction performance of our model was primarily evaluated using the Root Mean Square (RMS) measure, to account for the ordinal nature of labels. Additionally, our evaluation includes F-score and accuracy results to measure classification performance. As our baseline, we used various text-based importance prediction models trained and evaluated on the same data split, as described in Section 6.3.

For training, we explored various architectural parameters to find the best-working setup for our models: Our 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.

6 Experiments

Tables 1, 2 and 3 summarize the performance of our 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.

6.1 Comparison of the Projection Layers

We compared the efficacy of the learning architecture's three projection layers (Section 3.1) by training them separately and comparing their performance on the test corpus. Table 1 summarizes the results of this evaluation.

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

Model	ACC	F1	RMS
LSTM-CRF	64.22	56.31	75.21
LSTM-SOFTMAX	**65.66**	57.34	74.08
LSTM-ORD	63.72	**57.58**	**68.21**

Table 1: Performance of our speech-based models on the test data under different projection layers. Best performing scores highlighted in **bold**.

the other two models. This suggests the utility of the ordinal constraint present in the ORD-based model for word importance classification.

6.2 Ablation Study on Speech Features

To compare the effect of different categories of speech features on the performance of our model, we evaluated variations of the model by removing one feature group at a time from the model during training. Table 2 summarizes the results of the experiment.

Model	ACC	F1	RMS
speech-based	63.72	57.58	68.21
− ENG	62.24[†]	55.67[†]	71.14
− FREQ	63.25	57.30	69.0
− VOC	62.90	56.84	70.5
− LEX	63.37	57.34	71.49[†]
− ZNORM	62.04[*]	53.86[*]	72.0[*]

Table 2: 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-based normalization (ZNORM) features and omitting spoken-lexical features (LEX) resulted in the greatest increase in the overall RMS error (+5.5% and +4.8% relative increase in RMS respectively) – suggesting the discriminative importance of these features for word importance prediction. Further, our results indicated the importance of energy-based (ENG) features, which resulted in a substantial drop (-2.4% relative decrease) in accuracy of the model.

6.3 Comparison with the Text-based Models

In this analysis, we compare our 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 described in Kafle and Huenerfauth (2018), 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 1, 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[2] 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 (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).

Model	ACC	F1	RMS
speech-based	63.72	57.58	68.21
text-based	77.81	73.6	54.0
+ WER: 0.25	72.30	69.04	65.15
+ WER: 0.30	71.84	67.71	68.55

Table 3: 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 3 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 (Lasecki et al., 2012; Barker et al., 2017). When operating on such ASR output, the RMS error of the speech-based model and the text-based model were comparable.

7 Conclusion

Motivated by recent work on evaluating the accuracy of automatic speech recognition systems for real-time captioning for Deaf and Hard of Hearing (DHH) users (Kafle and Huenerfauth, 2018), we investigated how to predict the importance of a word to the overall meaning of a spoken conversation turn. 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 continued work involves combining both text- and speech-based features for the task of word importance prediction.

8 Acknowledgements

---

[2]https://cloud.google.com/Speech_API

References

Jon P. Barker, Ricard Marxer, Emmanuel Vincent, and Shinji Watanabe. 2017. The chime challenges: Robust speech recognition in everyday environments. In Shinji Watanabe, Marc Delcroix, Florian Metze, and John R. Hershey, editors, *New Era for Robust Speech Recognition, Exploiting Deep Learning.*, pages 327–344. Springer.

Mary E Beckman and Jennifer J Venditti. 2000. Tagging prosody and discourse structure in elicited spontaneous speech. *Proc. Science and Technology Agency Priority Program Symp. on Spontaneous Speech.*

Larwan Berke, Sushant Kafle, and Matt Huenerfauth. 2018. Methods for evaluation of imperfect captioning tools by deaf or hard-of-hearing users at different reading literacy levels. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 91:1–91:12, New York, NY, USA. ACM.

Paul Boersma. 2006. Praat: doing phonetics by computer. *http://www. praat. org/.*

Jason M. Brenier, Daniel M. Cer, and Daniel Jurafsky. 2005. The detection of emphatic words using acoustic and lexical features. In *INTERSPEECH 2005 - Eurospeech, 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, September 4-8, 2005*, pages 3297–3300. ISCA.

Jianlin Cheng, Zheng Wang, and Gianluca Pollastri. 2008. A neural network approach to ordinal regression. In *Proc. of IJCNN 2008, Hong Kong, China, June 1-6, 2008*, pages 1279–1284. IEEE.

Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. In *Proc. of SSST@EMNLP 2014, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, Doha, Qatar, 25 October 2014*, pages 103–111. Association for Computational Linguistics.

Yu-An Chung and James R. Glass. 2018. Speech2vec: A sequence-to-sequence framework for learning word embeddings from speech. In *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2-6 September 2018.*, pages 811–815. ISCA.

Yu-An Chung, Chao-Chung Wu, Chia-Hao Shen, Hung-yi Lee, and Lin-Shan Lee. 2016. Audio word2vec: Unsupervised learning of audio segment representations using sequence-to-sequence autoencoder. In *Interspeech 2016, 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, September 8-12, 2016*, pages 765–769. ISCA.

Nivja H De Jong and Ton Wempe. 2009. Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior research methods*, 41(2):385–390.

Lyn Frazier, Katy Carlson, and Charles Clifton. 2006. Prosodic phrasing is central to language comprehension. *Trends in cognitive sciences*, 10(6):244–249.

John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, volume 1, pages 517–520. IEEE.

Yaakov HaCohen-Kerner, Zuriel Gross, and Asaf Masa. 2005. Automatic extraction and learning of keyphrases from scientific articles. In *Proc. of CICLing 2005, Mexico City, Mexico, February 13-19, 2005*, volume 3406, pages 657–669. Springer.

Wanjia He, Weiran Wang, and Karen Livescu. 2017. Multi-view recurrent neural acoustic word embeddings. In *Proc. of ICLR 2017, Toulon, France, April 24-26, 2017*. OpenReview.net.

Kai Hong and Ani Nenkova. 2014. Improving the estimation of word importance for news multi-document summarization. In *Proc. of EACL 2014, April 26-30, 2014, Gothenburg, Sweden*, pages 712–721. The Association for Computer Linguistics.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *CoRR*, abs/1508.01991.

Anette Hulth. 2003. Improved automatic keyword extraction given more linguistic knowledge. In *Proc. of EMNLP 2003, Sapporo, Japan, July 11-12, 2003*.

Taisuke Ito, Kazuya Takeda, and Fumitada Itakura. 2002. Acoustic analysis and recognition of whispered speech. In *Proc. of ICASSP 2002, May 13-17 2002, Orlando, Florida, USA*, pages 389–392. IEEE.

Sushant Kafle and Matt Huenerfauth. 2017. Evaluating the usability of automatically generated captions for people who are deaf or hard of hearing. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility, ASSETS 2017, Baltimore, MD, USA, October 29 - November 01, 2017*, pages 165–174. ACM.

Sushant Kafle and Matt Huenerfauth. 2018. A corpus for modeling word importance in spoken dialogue transcripts. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018.* European Language Resources Association (ELRA).

Patricia A Keating and Christina Esposito. 2006. Linguistic voice quality. *UCLA Working Papers in Phonetics*, 105(6).

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proc. of NAACL HLT 2016, San Diego California, USA, June 12-17, 2016*, pages 260–270. The Association for Computational Linguistics.

Walter S. Lasecki, Christopher D. Miller, Adam Sadilek, Andrew Abumoussa, Donato Borrello, Raja S. Kushalnagar, and Jeffrey P. Bigham. 2012. Real-time captioning by groups of non-experts. In *The 25th Annual ACM Symposium on User Interface Software and Technology, UIST '12, Cambridge, MA, USA, October 7-10, 2012*, pages 23–34. ACM.

Bing Liu, Xiaoli Li, Wee Sun Lee, and Philip S. Yu. 2004. Text classification by labeling words. In *Proc. of AAAI, July 25-29, 2004, San Jose, California, USA*, pages 425–430. AAAI Press / The MIT Press.

Fei Liu, Feifan Liu, and Yang Liu. 2011. A supervised framework for keyword extraction from meeting transcripts. *IEEE Trans. Audio, Speech & Language Processing*, 19(3):538–548.

Yutaka Matsuo and Mitsuru Ishizuka. 2004. Keyword extraction from a single document using word co-occurrence statistical information. *International Journal on Artificial Intelligence Tools*, 13(1):157–169.

Iain A. McCowan, Darren Moore, John Dines, Daniel Gatica-Perez, Mike Flynn, Pierre Wellner, and Hervé Bourlard. 2004. On the use of information retrieval measures for speech recognition evaluation. Idiap-RR Idiap-RR-73-2004, IDIAP, Martigny, Switzerland.

Kathleen R. McKeown, Julia Hirschberg, Michel Galley, and Sameer Maskey. 2005. From text to speech summarization. In *2005 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '05, Philadelphia, Pennsylvania, USA, March 18-23, 2005*, pages 997–1000. IEEE.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, pages 3111–3119.

Taniya Mishra, Vivek Kumar Rangarajan Sridhar, and Alistair Conkie. 2012. Word prominence detection using robust yet simple prosodic features. In *INTERSPEECH 2012, 13th Annual Conference of the International Speech Communication Association, Portland, Oregon, USA, September 9-13, 2012*, pages 1864–1867. ISCA.

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, Korea, October 4-8, 2004.*

Gabriel Murray, Steve Renals, and Jean Carletta. 2005. Extractive summarization of meeting recordings. In *INTERSPEECH 2005 - Eurospeech, 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, September 4-8, 2005*, pages 593–596. ISCA.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proc. of EMNLP 2014, October 25-29, 2014, Doha, Qatar.*, pages 1532–1543. ACL.

Patti Price, Mari Ostendorf, Stefanie Shattuck-Hufnagel, and Cynthia Fong. 1991. The use of prosody in syntactic disambiguation. In *Speech and Natural Language, Proceedings of a Workshop held at Pacific Grove, California, USA, February 19-22. 1991*. Morgan Kaufmann.

JI Sheeba and K Vivekanandan. 2012. Improved keyword and keyphrase extraction from meeting transcripts. *International Journal of Computer Applications*, 52(13).

Imran A. Sheikh, Irina Illina, Dominique Fohr, and Georges Linarès. 2016. Learning word importance with the neural bag-of-words model. In *Proc. of Rep4NLP@ACL 2016, Berlin, Germany, August 11, 2016*, pages 222–229. Association for Computational Linguistics.

Fabio Tamburini. 2003. Prosodic prominence detection in speech. In *Seventh International Symposium on Signal Processing and Its Applications, ISSPA 2003, July 1-4, 2003, Paris, France, Proceedings, Volume 1*, pages 385–388. IEEE.

Trang Tran, Shubham Toshniwal, Mohit Bansal, Kevin Gimpel, Karen Livescu, and Mari Ostendorf. 2017. Joint modeling of text and acoustic-prosodic cues for neural parsing. *CoRR*, abs/1704.07287.

Dagen Wang and Shrikanth Narayanan. 2007. An acoustic measure for word prominence in spontaneous speech. *IEEE Trans. Audio, Speech & Language Processing*, 15(2):690–701.

Shasha Xie, Dilek Hakkani-Tür, Benoît Favre, and Yang Liu. 2009. Integrating prosodic features in extractive meeting summarization. In *2009 IEEE Workshop on Automatic Speech Recognition & Understanding, ASRU 2009, Merano/Meran, Italy, December 13-17, 2009*, pages 387–391. IEEE.

Wen-tau Yih, Joshua Goodman, Lucy Vanderwende, and Hisami Suzuki. 2007. Multi-document summarization by maximizing informative content-words.