INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Fusion Strategy for Prosodic and Lexical Representations of Word Importance

Sushant Kafle, Cecilia O. Alm, Matt Huenerfauth

Rochester Institute of Technology
1 Lomb Memorial Drive, Rochester, NY
sushant@mail.rit.edu, coagla@rit.edu, matt.huenerfauth@rit.edu

Abstract

We investigate whether, and if so when, prosodic features in spoken dialogue aid in modeling the importance of words to the overall meaning of a dialogue turn. Starting from the assumption that acoustic-prosodic cues help identify important speech content, we investigate representation architectures that combine lexical and prosodic features and evaluate them for predicting word importance. We propose an attention-based feature fusion strategy and additionally show how the addition of strategic supervision of the attention weights results in especially competitive models. We evaluate our fusion strategy on spoken dialogues and demonstrate performance increases over state-of-the-art models. Specifically, our approach both achieves the lowest root mean square error on test data and generalizes better over out-of-vocabulary words.

Index Terms: prosody modeling, fusion strategy of prosodic and lexical representations, word importance in dialogues.

1. Introduction

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 [1, 2] to text classification [3, 4] and summarization [5, 6]. In particular, researchers in [2] found that when captioning individual conversation turns during a live meeting for people who are deaf or hard-of-hearing (DHH), differential weighting of automated recognition errors based on the importance of words correlates better with the human judgment of ASR quality than the traditional word error rate (WER) metric.

While prior models of word importance considered text features [7, 8], speech-based features hold promise when analyzing conversational speech [9]. Speakers often use prosodic cues to help listeners discern spoken messages; however, these cues are omitted from an automatically generated text transcript [10]. 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. Our 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 a corpus of word importance [7], 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`.

2. Related Work

Joint modeling of lexical and prosodic features has benefited various applications, such as constituent parsing of conversational speech texts [11], and summarization of recordings of meetings [12, 13]. The most common strategy for joint representation of features is through concatenation. Despite the popularity of this strategy, it has been shown to fail to fully capture cross-modal interactions [14, 15]. Consequently, several multimodal feature representation strategies have been proposed for various applications [16, 14, 15, 17]. Our 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 [18, 19], with a goal of extracting relevant keywords from a text. Other supervised measures of keyword extraction have been proposed [20, 21, 22, 23] 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. [2].

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. Recently, several models that consider contextualized word representations have been proposed [7, 9]. However, as discussed in Section 1, linguistic models based on text-only features or on speech-only features may be insufficient for conversational speech-based applications.

3. 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 sig-

http://dx.doi.org/10.21437/Interspeech.2019-1898

nal 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 [17]. 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 [24, 25, 26]. We investigate an attention-based feature fusion architecture that considers the effect of prosodic cues on the lexical meaning of a spoken message.

3.1. 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 [11], as well as its relation to its context [27]. Our speech-feature sub-network aims to learn a feature representation for a spoken word that encapsulates this information.

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 ($w_i \simeq [a_1^i, a_2^i, .., a_T^i]$) and passed as a sequential input to our RNN, as:

$$\overrightarrow{h_t} = \text{RNN}(a_t^i, \overrightarrow{h_{t-1}}) \quad (1)$$

$$\overleftarrow{h_t} = \text{RNN}(a_t^i, \overleftarrow{h_{t-1}}) \quad (2)$$

where a_t^i 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 $S_i = [\overrightarrow{h_T}; \overleftarrow{h_T}]$.

3.2. 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:

$$h_{s_i} = tanh(W_1 \cdot S_i + b_1) \quad (3)$$

$$\alpha_i = tanh(W_2 \cdot [h_{s_i}; E_i] + b_2) \quad (4)$$

$$Z_i = E_i + \alpha_i \cdot h_{s_i} \quad (5)$$

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 dimension(h_{s_i}) = 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 α_i.

The intuition is to learn an appropriate composition vector ($\alpha_i \cdot hs_i$) 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. For instance, inherently neutral words like *dog* can bear both positive and negative sentiment as a part of a discourse. The shift in connotative meaning is often conveyed through prosody that informs listeners about the relation of the word to the discourse and to the *mutual belief* built up by interlocutors during the course of the discourse which might influence its connotative meaning in context [28, 29].

3.3. 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 [30, 31, 32], 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.

$$\tilde{L} = L + \lambda \begin{cases} \sum_{w_i} -log(|\alpha_i|), & \text{if } w_i \notin V \\ 0, & \text{otherwise} \end{cases} \quad (6)$$

where L represents the training loss and $\tilde{L}$ represents the new loss (with regularization constraint as determined by α_i) that the model is optimizing, w_i is the word, α_i represents the attention weights for the prosodic features of the word, and V is the vocabulary of the model. Additionally, λ 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 (α_i), meaning higher reliance on speech features for the prediction. The motivation behind this supervision technique is discussed in Section 5.2 – where we found that in general, prosodic features are less prone to OOV errors as compared to text-based lexical features.

4. Experimental Methodology

4.1. Dataset

We used the Word Importance Corpus for the training and evaluation of our word importance prediction models [7]. 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 [33], 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.

4.2. Unimodal Representations

We make use of the 6-billion-token-based 300-dimension pretrained GloVe [34] embeddings as our lexical representation for the word. To get a word vector representation for speech, we utilized the network described in Section 3.1. 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 [9], 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 speaker-normalized to account for inter-speaker variations.

4.3. Comparison Models

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

Concatenation (CONCAT) [35, 11, 36]: The model creates a multimodal representation of words by simply concatenating the unimodal features at the word level.

Attention-based Weighted Sum (ATTN) [37, 16]: 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) [14]: 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) [15]: 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) [17]: 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.

4.4. 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) [38] as RNN cell[1] 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), we found a loss weighting factor (λ) 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.

[1]We used GRU rather than LSTM units due to better performance observed during our initial set analysis.

4.5. 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 4.1. We used the RMS error as the primary measure of performance, comparing predictions against the gold standard corpus. As described by [7], the annotators of the corpus were asked to consider three ordinal ranges {LOW: [0 - 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 (τ-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.

5. Results and Discussion

5.1. Error Analysis of Unimodal Models

The performance of the two unimodal-feature (lexical and speech) models in Table 1 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 1: *Comparative performance of lexical and prosodic unimodal models. Bold font shows the best scores.*

Models	RMS	RMS (OOV words only)
prosodic-only	21.5	**27.0**
lexical-only	**16.84**	27.35

Each word in the Word Importance Corpus is annotated with an importance score between 0 and 1 [7]. Error analysis revealed that the lexical-only model trained on transcripts had a lower percentage of highly deviated predictions (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.

5.2. 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 above.

Tables 2 and 3 summarize the performance of all models. Notably, the results in Table 2 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 3 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

Figure 1: *Visualization of the lexical and prosodic distribution of words **love**, **night**, **cold** in different spoken contexts. The blue (top) and red (bottom) contours represent the distribution of all positive and all negative sentiment words, respectively. For each word, the black contour shows its distribution for different spoken contexts. Individual instances of the spoken word are shown as dots. Proximity to the blue or red contours illustrates how a word adopts positive (love) or negative (cold) connotations, or straddle both (night).*

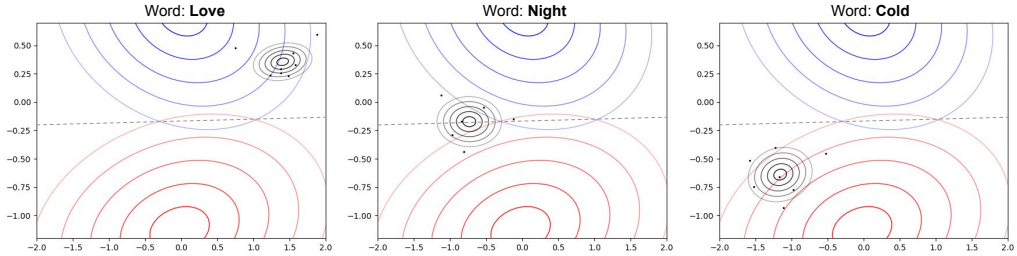

Table 2: *Comparison of different models combining lexical and prosodic cues. Per column, the top two results are marked with ⋆ & † symbols, respectively. Our proposed model demonstrates lower RMS error both overall as well as for OOVs specifically.*

Models	**RMS**	**RMS** (OOV words only)
CONCAT	15.64†	23.20†
ATTN	16.08	23.84
TNF	17.14	29.08
LMF	16.59	27.02
RAVEN	17.0	28.5
Proposed ($\lambda = 0$)	15.80	23.65
Proposed ($\lambda = 0.8$)	14.75⋆	21.71⋆

Table 3: *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.*

Models	**RMS** (across **ranges**)			τ-**b**
	HI	**MID**	**LOW**	
CONCAT	21.81†	13.07†	10.85	59.02
ATTN	25.87	13.44	10.77	58.41
TFN	26.0	13.71	11.34	58.17
LMF	27.56	13.53	10.31⋆	60.04†
RAVEN	29.04	12.50⋆	11.65	59.77
Proposed ($\lambda = 0$)	25.13	13.29	10.85	59.80
Proposed ($\lambda = 0.8$)	22.4⋆	13.27	10.60†	61.35⋆

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 $\tau-$b scores indicate that our model is also better at capturing the overall rank distribution of words in a turn.

5.3. Combined Representation and Prosodic Deviation

Our integrated lexical and prosodic representation attempts to encode the influence of prosody into the lexical meaning of the word. With this setup, the same word spoken differently (on different contexts) would have different feature representations. This is indeed the case as demonstrated in Figure 1. In the figure, the fused feature representation of words have been projected into a two-dimensional vector space using Principal Component Analysis (PCA). The two axes represent a new projection space that explains maximum variance in the sentiment-bearing words in our corpus. The blue (top) and red (bottom) Gaussian contours represent the distribution of all the positive and negative sentiment words in the corpus respectively, showcasing that the variance of sentiment is primarily along the Y-axis. From this perspective, the feature variation of words due to prosody can be interpreted as a change in sentiment.

Table 4: *The word **night** in different spoken contexts with corresponding positioning in the contour plot (Figure 1).*

Conversational Context	Positioning
stealing cars like at **night** breaking into ...	bottom-half
you have a good **night** we'll see you ...	top-half
last **night** i did thirty minutes of riding ...	middle

Figure 1 shows the feature distribution of three words *love*, *night* and *cold* in different spoken contexts. The sentiment-bearing words *love* and *cold* lie on the upper and lower parts of the graph respectively, indicating that their sentiment is quite stable. However, as seen in Table 4, a neutral word like *night* overlaps in positive and negative sentiment and varies by spoken contexts; this could reflect varying prosodic renderings.

6. Conclusion

We have shown that by incorporating features from speech into the lexical embeddings, we can enhance the performance of word-importance prediction systems. We proposed an attention-based multimodal feature representation strategy that learns to adjust the text-based feature representation of spoken words to reflect the post-lexical meaning conveyed through prosody. Further, we demonstrate that incorporating modality-specific heuristics into training helps our model perform better. Our evaluations 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. In future work, we will investigate speech- and text-based features for modeling the importance of larger semantic units.

7. Acknowledgements

This material is based upon work supported by the National Science Foundation under Award No. 1462280, by the Department of Health and Human Services under Award No. 90DPCP0002-01-00, by a Microsoft AI for Accessibility (AI4A) Award, and by a Google Faculty Research Award.

8. References

[1] T. Mishra, A. Ljolje, and M. Gilbert, "Predicting human perceived accuracy of ASR systems," in *Proc. Interspeech 2011*, 2011, pp. 1945–1948.

[2] S. Kafle and M. Huenerfauth, "Evaluating the usability of automatically generated captions for people who are deaf or hard of hearing," in *Proc. ASSETS 2017*. ACM, 2017.

[3] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proc. NAACL-HLT 2016*, 2016, pp. 1480–1489.

[4] Y. Ko, J. Park, and J. Seo, "Automatic text categorization using the importance of sentences," in *Proceedings of the 19th international conference on Computational linguistics-Volume 1*. ACL, 2002, pp. 1–7.

[5] K. Hong and A. Nenkova, "Improving the estimation of word importance for news multi-document summarization," in *Proc. EACL 2014*, 2014, pp. 712–721.

[6] W. Yih, J. Goodman, L. Vanderwende, and H. Suzuki, "Multi-document summarization by maximizing informative content-words," in *Proc. IJCAI 2007*, 2007, pp. 1776–1782.

[7] S. Kafle and M. Huenerfauth, "A corpus for modeling word importance in spoken dialogue transcripts," in *Proc. LREC 2018*, 2018.

[8] I. A. Sheikh, I. Illina, D. Fohr, and G. Linarès, "Learning word importance with the neural bag-of-words model," in *Proc. Rep4NLP@ACL 2016*, 2016, pp. 222–229.

[9] S. Kafle, C. O. Alm, and M. Huenerfauth, "Modeling acoustic-prosodic cues for word importance prediction in spoken dialogues," in *Proc. SLPAT 2019*, 2019, pp. 9–16.

[10] L. Frazier, K. Carlson, and C. Clifton, "Prosodic phrasing is central to language comprehension," *Trends in cognitive sciences*, vol. 10, no. 6, pp. 244–249, 2006.

[11] T. Tran, S. Toshniwal, M. Bansal, K. Gimpel, K. Livescu, and M. Ostendorf, "Parsing speech: a neural approach to integrating lexical and acoustic-prosodic information," in *Proc. NAACL-HLT 2018*, 2018, pp. 69–81.

[12] S. Xie, D. Hakkani-Tur, B. Favre, and Y. Liu, "Integrating prosodic features in extractive meeting summarization," in *Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding, 2009*. IEEE, 2009, pp. 387–391.

[13] G. Murray, S. Renals, and J. Carletta, "Extractive summarization of meeting recordings," in *Proc. Interspeech 2005*, 2005, pp. 593–596.

[14] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proc. EMNLP 2017*, 2017, pp. 1103–1114.

[15] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. B. Zadeh, and L. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," in *Proc. ACL 2018*, 2018, pp. 2247–2256.

[16] M. Rei, G. K. O. Crichton, and S. Pyysalo, "Attending to characters in neural sequence labeling models," in *Proc. COLING 2016*, 2016, pp. 309–318.

[17] Y. Wang, Y. Shen, Z. Liu, P. P. Liang, A. Zadeh, and L.-P. Morency, "Words can shift: Dynamically adjusting word representations using nonverbal behaviors," *Proc. AAAI 19*, 2018.

[18] Y. HaCohen-Kerner, Z. Gross, and A. Masa, "Automatic extraction and learning of keyphrases from scientific articles," *Computational linguistics and intelligent text processing*, pp. 657–669, 2005.

[19] Y. Matsuo and M. Ishizuka, "Keyword extraction from a single document using word co-occurrence statistical information," *International Journal on Artificial Intelligence Tools*, vol. 13, no. 01, pp. 157–169, 2004.

[20] F. Liu, F. Liu, and Y. Liu, "A supervised framework for keyword extraction from meeting transcripts," *Proc. of IEEE transactions on audio, speech, and language processing, 2011*, vol. 19, no. 3, pp. 538–548, March 2011.

[21] B. Liu, X. Li, W. S. Lee, and P. S. Yu, "Text classification by labeling words," in *Proc. AAAI 2004*, vol. 4, 2004, pp. 425–430.

[22] A. Hulth, "Improved automatic keyword extraction given more linguistic knowledge," in *Proc. EMNLP 2003*. Association for Computational Linguistics, 2003, pp. 216–223.

[23] J. Sheeba and K. Vivekanandan, "Improved keyword and keyphrase extraction from meeting transcripts," *International Journal of Computer Applications*, vol. 52, no. 13, 2012.

[24] L. C. Nygaard, D. S. Herold, and L. L. Namy, "The semantics of prosody: Acoustic and perceptual evidence of prosodic correlates to word meaning," *Cognitive science*, vol. 33, no. 1, pp. 127–146, 2009.

[25] D. Wang and S. Narayanan, "An acoustic measure for word prominence in spontaneous speech," *Proc. of IEEE transactions on audio, speech, and language processing*, vol. 15, no. 2, pp. 690–701, 2007.

[26] P. Price, M. Ostendorf, S. Shattuck-Hufnagel, and C. Fong, "The use of prosody in syntactic disambiguation," in *Speech and Natural Language: Proceedings of a Workshop Held at Pacific Grove, California, February 19-22, 1991*, 1991.

[27] J. Kleinhans, M. Farrs, A. Gravano, J. M. Prez, C. Lai, and L. Wanner, "Using prosody to classify discourse relations," in *Proc. Interspeech 2017*, 2017, pp. 3201–3205.

[28] J. B. Pierrehumbert, "Exemplar dynamics: Word frequency, lenition and contrast," in *Typological studies in language, Vol. 45. Frequency and the emergence of linguistic structure (pp. 137-157)*, 2000.

[29] D. R. Ladd, *Intonational Phonology*, 2nd ed., ser. Cambridge Studies in Linguistics. Cambridge University Press, 2008.

[30] L. Liu, M. Utiyama, A. Finch, and E. Sumita, "Neural machine translation with supervised attention," *arXiv preprint arXiv:1609.04186*, 2016.

[31] Z. Fang, S. Kong, T. Yu, and Y. Yang, "Weakly supervised attention learning for textual phrases grounding," *arXiv preprint arXiv:1805.00545*, 2018.

[32] M. Nguyen and T. Nguyen, "Who is killed by police: Introducing supervised attention for hierarchical lstms," in *Proc. COLING 2018*, 2018, pp. 2277–2287.

[33] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Proc. ICASSP 1992*, vol. 1. IEEE, 1992, pp. 517–520.

[34] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. EMNLP 2014*, 2014, pp. 1532–1543.

[35] T. J. Park and P. Georgiou, "Multimodal speaker segmentation and diarization using lexical and acoustic cues via sequence to sequence neural networks," in *Proc. Interspeech 2018*, 2018, pp. 1373–1377.

[36] C. Liu, C. Ishi, and H. Ishiguro, "Turn-taking estimation model based on joint embedding of lexical and prosodic contents," in *Proc. Interspeech 2017*, 2017, pp. 1686–1690.

[37] S. Moon, L. Neves, and V. Carvalho, "Multimodal named entity recognition for short social media posts," in *Proc. NAACL-HLT 2018*, 2018, pp. 852–860.

[38] K. Cho, B. van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder–decoder approaches," *Syntax, Semantics and Structure in Statistical Translation*, p. 103, 2014.