

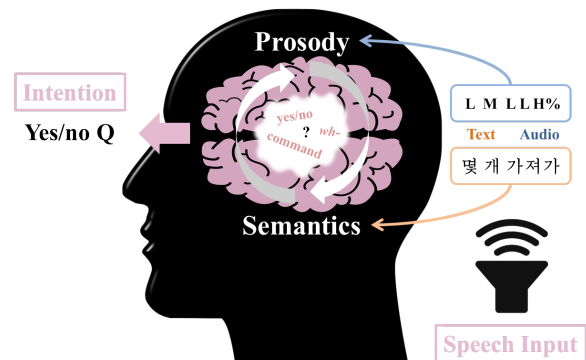
Won Ik Cho¹ (wicho@hi.snu.ac.kr)
Jeonghwa Cho² (jeonghwa@umich.edu)
Woo Hyun Kang¹ (whkang@hi.snu.ac.kr)
Nam Soo Kim¹ (nkim@snu.ac.kr)

Abstract

Keywords: syntactic ambiguity resolution; speech intention disambiguation; audio-text co-attention framework; prosody-syntax-semantics interface

Resolving syntactic ambiguity is a core task in spoken language analysis, since identifying the sentence type and understanding the intention of a text form utterance is challenging for some prosody-sensitive cases. Notably, in some *wh-in-situ* languages like Korean and Japanese, some uttered word sequences incorporate syntactic ambiguity, which leads to difficulties discerning directive speech from constative or rhetorical ones. For example, the following sentence in Seoul Korean can be interpreted differently depending on the prosody (Cho, Cho, Kang, & Kim, 2019):

where L, M, and H denote relative pitch of each syllabic block and USE denotes an underspecified sentence ender. Unlike English translations, if given only the text with periods or

¹Denotes an underspecified sentence ender.

question marks removed (usually provided as an output of automatic speech recognition (ASR)), the language understanding modules may not be able to determine if it is a statement or a question. Even with such marks, it is vague whether the question is *yes/no* or *wh*-.

As such, for an utterance that contains components whose roles are decided by prosody, it requires both the acoustic and textual data for spoken language understanding (SLU) modules (and even humans) to correctly infer the speech intention. In this process, the pitch sequence, the duration between words, and the overall tone all together decide the intention of an utterance. Thus, we concluded that introducing prosodic information is indispensable for resolving syntactic ambiguity, as depicted in Figure 1.

In this paper, we first review literature on prosody-syntax interface, syntactic ambiguity resolution, and machine learning-based speech analysis. Then, we introduce the architecture of speech intention classification systems that co-utilize the audio and text features, along with a comparison with speech-based and text-aided (self-attentive) recurrent neural network (RNN) models. Next, in the experiment section, the utilized corpus and the results are described, based on which we suggest the possibility of a connection between computational speech intention identification and human language processing mechanisms. Our contribution is as follows:

- We applied parallel bidirectional recurrent encoder (P-BRE), multi-hop attention (MHA), and cross-attention

(CA) to identify the speech intention of syntactically ambiguous utterances and scrutinized them in cross-modal perspective.

- We analyzed the result in view of experimental linguistics, linking the co-attention frameworks to neurological findings.

Related Work

The interaction of prosody and syntax has been long investigated regarding sentence types, especially for the question types including *wh*-intervention (Pires & Taylor, 2007) and declarative forms (Gunlogson, 2002). Moreover, for some head-final languages, the sentence-final intonation can play a significant role in clarifying the sentences. For instance, the prosody assigned to the final particle or word of non-scrambled Korean sentences usually decides the sentence form, e.g., declaratives or interrogatives (Yun, 2019).

In a broader perspective, syntactic ambiguity has been dealt with not only in studies on sentences but also phrases. In Korean, dative (Hwang & Schafer, 2009) and comparatives (Kim & Sells, 2010) terms have been mainly investigated. Also, linking syntax with phonetics, Baek and Yun (2018) demonstrated that syntactic ambiguity is resolved via prosodic information that can elaborate the low/highness of the attachment. They handled several cases in Korean where the syntax differentiates upon phonetic properties, especially among long phrases (e.g., *diligent boy's sister*). This approach is mainly concerned with contiguity theory (Richards, 2016), which claims that syntax can make reference to phonological structure, and that movement operations can be triggered by the need to produce phonologically acceptable objects.

Aforementioned studies on prosody/phonetics-syntax-semantics interface deal with various types of disambiguation, which incorporates the variance in topic, agent, experiencer, and object of syntactically ambiguous sentences. Among them, a few concerns questions, commands, and their directivity (see Yun (2019)). Cho, Cho, et al. (2019) suggested the seven-class categorization (namely statement, *yes/no* question, *wh*-question, rhetorical question, command, request, and rhetorical command), based on (i) sentence-middle intonation that affects topic and *wh*-intervention, (ii) sentence-final intonation that changes the sentence form, and (iii) the overall tone that influences rhetoricalness. The limitation is that, the analysis beyond the categorization has yet been performed. For now, the prosodic activeness-based disambiguation (Richards, 2016; Baek & Yun, 2018) well formulates the phonetic segments that clarify syntax. But we deemed it necessary to resolve the ambiguity within wider range of sentence types, promoting possibly an automatic management. In this regard, we take into account computational approaches that autonomously discover the latent and non-codified criteria.

Early studies on spoken language analysis adopt a simple concatenation of acoustic and textual features (Gu, Li, Chen,

Zhang, & Marsic, 2017), where parallel convolutional or recurrent neural networks (CNN/RNNs) were used to summarize each feature. A recent study includes hierarchical attention networks (HAN) (Gu et al., 2018) that point out the components which are essential for inferring the answer. In the related area of speech emotion recognition, multi-hop attention (MHA) (Yoon, Byun, Dey, & Jung, 2019) was introduced to encourage a comprehensive information exchange between textual and acoustic features. Nonetheless, since the experiments in literature generally utilize speech utterances with less confusing intention or emotion (e.g., syntactically non-ambiguous sentences or emotion utterances without semantic cue), there has been little study concentrating on the resolution of ambiguous sentences as in (s1).

In terms of the prosody-syntax-semantics interface, we concluded that the interaction between acoustic and textual information is required for such cases. We aim to materialize this approach in our co-attentional architectures in the form of MHA (Yoon et al., 2019) and cross-attention (CA) (Lee, Chen, Hua, Hu, & He, 2018), which have shown their power in the area of speech emotion recognition and image-text matching respectively.

Model

Here, we describe how the co-attention frameworks are constructed in terms of speech processing, self-attentive embedding, text-aided analysis, multi-hop attention, and cross-attention, as shown visually in Figure 2. In all models, the input is either audio-only (2.1-2) or audio-text pair (2.3-5). The text-only model is not taken into account since the text alone does not help resolve the syntactic ambiguity.

Audio-only Model (Audio-BRE)

The baseline model utilizes only audio input. Frame-level audio features are fed as an input to bidirectional long short-term memory (BiLSTM) (Schuster & Paliwal, 1997), for which the expression BRE (bidirectional recurrent encoder) is assigned (Yoon et al., 2019). The final hidden state is fully connected to a multi-layer perceptron (MLP) to yield a correct answer as a maximum probability output in the final softmax layer. Refer to (1) in Figure 2 for an illustration of the model's architecture.

Audio-only Model with Self-attentive Embedding (Audio-BRE-Att)

Since the audio-only BiLSTM² model lacks information regarding the identification of the core parts in analyzing an utterance, we augmented a self-attentive embedding layer as utilized in the sentence representation (Lin et al., 2017). In brief, a context vector, which has the same width as the hidden layers of the BiLSTM, is jointly trained to assign weight vector to the hidden layer sequence thereof.

The whole process, as in (2) of Figure 2, implies that the weight is decided upon the overall distribution of acoustic

²In this paper, we interchangeably use BRE and BiLSTM.

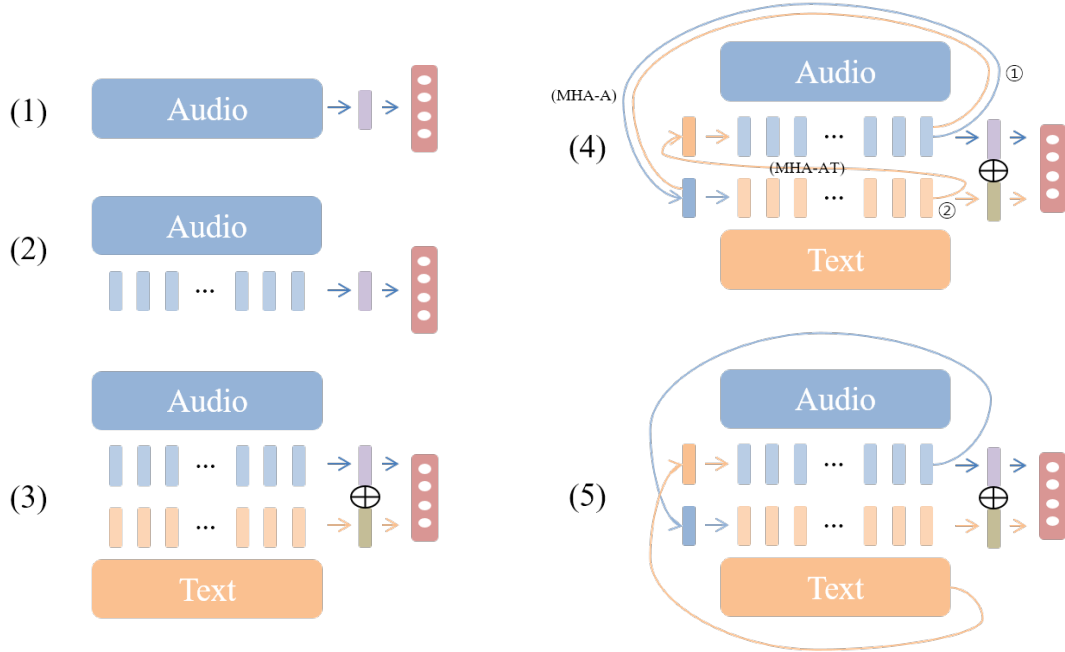


Figure 2: Block diagrams of the implemented models. Each number denotes the corresponding subsections in **Model**. Best viewed in color.

features. Since the acoustic feature reflects the lexicon and the syntactic property, the weight ends up playing a crucial role in predicting the intention of the speech.

Parallel Utilization of Audio and Text (Para-BRE-Att)

Unlike emotion analysis where either textual or acoustic features do not necessarily dominate, in intention analysis, obtaining textual information can bring a significant advantage (Gu et al., 2017), even when a period or a question mark is omitted as in our experiment. Here, text input for the ambiguous sentences are identical (without punctuation mark) for two to four different versions of the speech, but feeding them as an input of separately constructed *Audio-BRE* may provide supplementary information. The final hidden layer of *Audio-BRE-Att* is concatenated with that of the *Text-BRE-Att* (BRE-Att that exploits textual features) to make up a new feature layer, as illustrated in (3) of Figure 2.

Multi-Hop Attention (MHA)

In multi-hop attention (MHA) (Yoon et al., 2019), which is proposed for speech emotion recognition, textual and acoustic features interact by sequentially transmitting information to each other. This is the background of the expression ‘multi-hop’, where the hopping is performed by adopting the final representation of each feature as a context vector of the other as in (4) of Figure 2. The final output of the former and the latter are eventually concatenated.

Here, we first implement hopping only *from audio to text* (4a) (MHA-A), and then augment *from text to audio* to make up (4b) (MHA-AT). They showed better performance than the

further hopped model (i.e., MHA-ATA) in the original study (Yoon et al., 2019). Also, it is empirically more acceptable than the reverse case (e.g., MHA-T/TA) since auditory sensory first faces acoustic data than semantic information.

Cross-Attention (CA)

From the perspective of another co-attention framework, we adopt cross-attention (CA) that fully utilizes the information flow exchanged simultaneously by both acoustic and textual features, as depicted in (5) of Figure 2. In the preceding paper on image-text matching (Lee et al., 2018), image segments are utilized in determining the attention vector for the text, and similarly in reverse. Thus, not limited to using the representation regarding one feature as a context vector of the other’s attention weight, we assumed it also plausible to utilize the final representation of *Audio-BRE-Att* in making up a weight vector for *Text-BRE* and vice versa. In this case, self-attentive embedding was not applied to the textual features, in order to reflect the auditory-first nature³.

Experiment

Corpus Description

The dataset for the analysis of ambiguous speech, which requires prosodic disambiguation, is a corpus that contains

³Since auditory sensory meets the speech before the audio is encoded to lexical components, we considered it fair to assign different levels of representation and weight regarding both modalities. Here, we implement it in the way of giving self-attentive embedding only to the audio features. In fact, providing attention to the textual features as well, resultingly degraded the performance; we also tried to avoid that case.

	<i>S</i>	<i>YN</i>	<i>WH</i>	<i>RQ</i>	<i>C</i>	<i>R</i>	<i>RC</i>
<i>Who</i>	547	544	446	202	112	26	18
<i>What</i>	294	283	186	64	32	14	4
<i>Where</i>	64	64	49	6	11	4	1
<i>When</i>	37	54	40	22	0	4	15
<i>How</i>	59	62	28	8	6	0	0
<i>How much</i>	84	40	100	0	14	8	0

Table 1: A frequency matrix on *wh*- particles and the intention types, excerpted from (Cho, Cho, et al., 2019). For male and female cases each, the total speech files were randomly organized and split into train and test set with the ratio of 9:1.

about 1.3K sentences with two to four different types of prosody (and corresponding intention). Specifically, each sentence (i) starts with a *wh*-particle, (ii) incorporates a predicate made up of general verbs and pronouns, and (iii) ends with underspecified sentence enders so that the overall prosody varies according to intention (and sometimes with politeness suffix). All the sentences received consensus of three Korean native speakers, and the total number of speech utterances reaches 3,552. A male and a female speakers recorded each utterance with an appropriate prosody for each intention, to obtain a dataset of size 7,104. The number of intentions is seven, namely statement (*S*), yes/no question (*YN*), *wh*-question (*WH*), rhetorical question (*RQ*), command (*C*), request (*R*), and rhetorical command (*RC*). The categorization is slightly modified from the publicly available Korean corpus (Cho, Lee, Yoon, Kim, & Kim, 2018) to reflect *wh*- intervention as illustrated in (s1). The specification of the corpus is given in Table 1. The corpus⁴ and its detailed generation scheme are published as a separate article (Cho, Cho, et al., 2019).

Features and Architecture

For acoustic features, mel spectrogram (MS) and root mean square energy per frame (RMSE) were obtained by Librosa and were concatenated frame-wisely. For textual features, two types of character-level representation were adopted, namely sparse and dense, as they show best performance for classification tasks (Cho, Kim, & Kim, 2019). For sparse vectors, multi-hot encodings of the Korean characters were used (Song, Han, Cho, & Lee, 2018). These features display conciseness and also preserve the property of the blocks as a conjunct form. For dense features that regard distributional semantics, the recently disclosed fastText (Bojanowski, Grave, Joulin, & Mikolov, 2016)-based word vector dictionary was exploited (Cho, Cheon, Kang, Kim, & Kim, 2018).

Considering the head-finality of the Korean sentences, we embedded the acoustic and textual features backward from the endpoint⁵. The maximum sequence length was fixed to

⁴<https://github.com/warnikchow/prosem>

⁵This was controversial, but was to assure that all the models can pay attention to the sentence end of each utterance, where the underspecified functional particles lie in. This does not mean *uni-LSTM*; this instead means that the characters are placed from the

	Accuracy (F1)		Param.s	Comp.
	Sparse	Dense		
(1) Audio-BRE	83.9 (0.652)		116K	65s
(2) Audio-BRE-Att	89.3 (0.759)		190K	67s
(3) Para-BRE-Att	93.2 (0.919)	92.8 (0.919)	260K	70s
(4a) MHA-A	93.8 (0.928)	93.5 (0.922)	266K	67s
(4b) MHA-AT	92.8 (0.909)	91.8 (0.904)	270K	67s
(5) CA	91.8 (0.884)	93.5 (0.919)	326K	65s
(3*) Para-ASR	90.0 (0.822)	-	-	-
(4a*) MHA-ASR	90.2 (0.799)	-	-	-

Table 2: Result on the 10% test set. For each model, where all reached the convergence, we chose the intersection among 5-best accuracy and F1 checkpoints that were yielded during the first 100 epochs of training.

cover the longest input, where all the features with shorter frame/character length were padded with zeros. Models were implemented with Keras (Chollet et al., 2015), using TensorFlow backend (Abadi et al., 2015). Architecture and hyperparameter specification are provided separately on-line, with all the codes for implementation⁶.

Result

Table 2 shows the comparison result utilizing the corpus dataset. Both train and test sets in (1-5) incorporate the scripts of ground truth, and for the others, the test set scripts were ASR result. Input materials are either sole audio or audio-text combined, both in the training and test phase.

Attention matters: First, in (1) and (2), we observed that audio itself incorporates substantial information regarding speech intention, and physical features such as duration, pitch, tone, and magnitude can help yield semantic understanding via attention mechanism. This seems to be related to the phenomenon where people often catch the underlying intention of a speech even when they fail to understand the whole words (Hellbernd & Sammler, 2016). Also, it was shown that attaching the attention layer guarantees stable convergence of the learning curve in the training phase.

Text matters: Next, as expected, the text-aided models (3-5) far outperform the audio-only ones (1, 2), notwithstanding bigger trainable parameter set size and the computation time. Although the character-level features we utilized do not necessarily represent semantic information (which is held at least at morpheme-level), this result can be interpreted as implying that utilizing textual features can help recognize the prosodic prominence within the audio features (Talman et al., 2019). It was beyond our expectation that the sparse vectors outperform the dense ones in general. The exception was in CA, which implies that CA takes more advantage from the distributional semantics within the text embedding. We infer that CA may exhibit significance if the utterances be-

end slot of the sequence in the input-level.

⁶<https://github.com/warnikchow/coaudiotext>

come more cumbersome, where pre-trained language models (LMs) prevalent these days might be helpful.

Co-attention framework helps: In (3-5), we noticed that co-utilizing both audio and text in making up the attention vectors as in (4) *MHA* or (5) *CA* shows better performance than a simple concatenation in *Para-BRE-Att*. Since the studies on speech emotion analysis (Pell, Jaywant, Monetta, & Kotz, 2011; Ben-David, Multani, Shakuf, Rudzicz, & van Lieshout, 2016) claim that prosody and semantic cues cooperatively affect inferring the ground truth, we suspect that a similar phenomenon takes place in the case of speech intention. That is, *acoustic and textual processing significantly benefits from a consequent or simultaneous interaction with each other*.

Over-stack may bring a collapse: We first hypothesized that (4b) or (5) would show better performance compared to (4a) due to a broader or deeper exchange of information between both sources. Instead, we observed performance degeneration, leading to the conclusion that the inference becomes unstable if too much information is stacked. It is assumed that speech intention analysis is *affected dominantly by the combination of speech analysis and a speech-aided text analysis* (4a, 5), preferably with the smaller contribution of text-aided speech analysis (4b), though the performance of the models may not be directly linked to actual human processing mechanism. This shows that *text matters but speech influences*, as will be discussed further below.

In-depth analyses: For a practicability of the systems, model parameter size and training time per epoch were recorded (Table 2). Taking into account that audio processing itself incorporates huge computation, co-utilizing the textual information seems to bring significant improvement.

Then, we performed an additional experiment on ASR result⁷ (3', 4a'), especially for the test utterances, where (3) and (4a) were chosen to observe how the degeneration differs in concatenation and co-attention frameworks. The training was performed with the ground truth, and the models for the sparse textual features were chosen upon the result with it as well (3, 4a). It is notable that both perform competitively with the case of perfect transcription, but the degeneration was more significant in the co-attention framework. This implies that the framework utilizing textual information more aggressively is ironically more vulnerable to errors. Thus, precise ASR and error-compensating text processing are both required for the improvement and application of the system.

Lastly, we observed that (i) the co-utilization of acoustic and textual features shows strength in identifying the intention classes that are highly influenced by prosody itself, e.g., distinguishing RQs from pure questions. Some cases deeply concerned the lexicon, e.g., distinguishing commands from statements or requests from *yes/no* questions. (ii) On the other hand, figuring out *wh*-intervention between *yes/no* and *wh*-questions, depended more on the interaction of audio and text processing, shown by a superior performance of *MHA* than

Para-BRE-Att. These two observations can be explained as follows: in Korean spoken language, identifying rhetoricalness often accompanies non-neutral emotion (as suggested for a syntactically similar language (Miura & Hara, 1995)), whereas *wh*-intervention mostly involves phonological properties. Thus, we assume that (i) the emotion-related identification concerns a comprehensive understanding of the utterance as in *Para-BRE-Att*, while (ii) the elaborate processing of verbal data requires an analysis that pays more attention to the details of audio and text.

Explainability

We want to claim that unlike emotion recognition, which is often dominated by the voice tone, intention analysis should rely on both acoustic and textual features. More precisely, the genuine intention cannot be inferred unless the audio and text are both given if the sentence incorporates syntactic ambiguity. This is the background for this study to incorporate various co-attention frameworks.

Neurological evidence also supports the idea of incorporating both acoustic and textual information in sentence comprehension. For example, when we try to comprehend a speech, the parts of the brain that deal with understanding the meaning (e.g., whether it is a statement, a question, or a command) include mainly the areas which take charge of syntax-semantics (Tanner, 2007) and some specific regions of the temporal lobe that incur disorder in phonological decoding if damaged (Boatman et al., 2000). At the outset, word forms are recognized based on the decoding process of auditory cortices (DeWitt & Rauschecker, 2012), and it is known that further language processing engages the Wernicke's area (DeWitt & Rauschecker, 2013). Thus, it is not unnatural to consider the further language processing as the pipeline, which corresponds with our observation that the attention regarding speech precedes that of lexicons. However, we want to claim that not only are such phonological and lexical processing subsequent, but also the related regions interchange the information with each other via various language pathways (Friederici, 2012), and possibly do so more intensively when faced with ambiguous utterances.

Moreover, an interaction between different lobes or hemisphere of the brain is also assumed, for example via dorsal and ventral pathways (Friederici, 2012) or corpus callosum (Sammler, Kotz, Eckstein, Ott, & Friederici, 2010). Here, the information flow is not only restricted to phonological information but can be extended to emotional information (Schmidt et al., 2013), as required for understanding rhetorical utterances. Supported by this biological phenomena and our experimental result, we can confirm that the correct identification of syntactically ambiguous sentences can benefit from the active co-utilization of acoustic and textual information.

Application

As stated previously, ambiguous utterances are disturbing factors for speech intention understanding, which can mis-

⁷ASR was performed with a freely available API:
<https://aibril-stt-demo-korean.sk.kr.mybluemix.net/>

lead a computational model to provide a wrong intent or item. However, aggregating both audio and text actively in analyzing such utterances can help more precisely predict the intention, if given a transcription with high accuracy. We are optimistic that our approach will prove meaningful for solving intriguing problems. In real life, co-attention frameworks can help machines or aphasia patients understand speech. Following, the system users or social chatbots may be able to provide proper response/reaction in free-style or goal-oriented conversations with others.

Beyond the intention-related syntactic ambiguity, the implemented structures can be utilized in other kinds of natural language processing systems that incorporate multi-modal inputs that are expected to be interactive with each other. For example, the proposed network can be utilized in providing a proper translation in multi-modal context. Beyond just a text-to-text transformation, the system might capacitate abstracting and utilizing the source speech or image as an auxiliary input for the conventional machine translation process.

Conclusion

In this paper, we constructed a speech intention recognition system using co-attentional frameworks inspired by psycholinguistics and prosody-semantics interface of human language understanding. Multi-hop attention and cross-attention outperformed the conventional speech/attention-based and text-aided models, as shown by the evaluation using the audio-text pair recorded with manually created scripts. An additional experiment with ASR output was also conducted to guarantee real world usage. The implemented systems can help SLU modules correctly infer the intention of syntactically/semantically ambiguous utterances in Seoul Korean and possibly in a multi-lingual manner. Besides, we hope the results provide empirical evidence for finding out the language processing mechanism of ambiguous utterances.

Acknowledgments

This work was supported by the Technology Innovation Program (10076583, Development of free-running speech recognition technologies for embedded robot system) funded By the Ministry of Trade, Industry & Energy (MOTIE, Korea). The authors appreciate Jeemin Kang for the great support in making up the dataset. Also, we are thankful to the anonymous reviewers for providing helpful comments.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., ... Zheng, X. (2015). *TensorFlow: Large-scale machine learning on heterogeneous systems*. Retrieved from <http://tensorflow.org/> (Software available from tensorflow.org)

Baek, H., & Yun, J. (2018). Prosodic disambiguation of syntactically ambiguous phrases in korean. *MIT Working Papers in Linguistics*, 88, 89–100.

Ben-David, B. M., Multani, N., Shakuf, V., Rudzicz, F., & van Lieshout, P. H. (2016). Prosody and semantics

are separate but not separable channels in the perception of emotional speech: Test for rating of emotions in speech. *Journal of Speech, Language, and Hearing Research*, 59(1), 72–89.

Boatman, D., Gordon, B., Hart, J., Selnes, O., Miglioretti, D., & Lenz, F. (2000). Transcortical sensory aphasia: revisited and revised. *Brain*, 123(8), 1634–1642.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Cho, W. I., Cheon, S. J., Kang, W. H., Kim, J. W., & Kim, N. S. (2018). Real-time automatic word segmentation for user-generated text. *arXiv preprint arXiv:1810.13113*.

Cho, W. I., Cho, J., Kang, J., & Kim, N. S. (2019). Prosody-semantics interface in seoul korean: Corpus for a disambiguation of wh-intervention. In *Proceedings of the 19th international congress of the phonetic sciences (icphs 2019)* (pp. 3902–3906).

Cho, W. I., Kim, S. M., & Kim, N. S. (2019). Investigating an effective character-level embedding in korean sentence classification. *arXiv preprint arXiv:1905.13656*.

Cho, W. I., Lee, H. S., Yoon, J. W., Kim, S. M., & Kim, N. S. (2018). Speech intention understanding in a head-final language: A disambiguation utilizing intonation-dependency. *arXiv preprint arXiv:1811.04231*.

Chollet, F., et al. (2015). *Keras*. <https://github.com/fchollet/keras>. GitHub.

DeWitt, I., & Rauschecker, J. P. (2012). Phoneme and word recognition in the auditory ventral stream. *Proceedings of the National Academy of Sciences*, 109(8), E505–E514.

DeWitt, I., & Rauschecker, J. P. (2013). Wernicke's area revisited: parallel streams and word processing. *Brain and language*, 127(2), 181–191.

Friederici, A. D. (2012). The cortical language circuit: from auditory perception to sentence comprehension. *Trends in cognitive sciences*, 16(5), 262–268.

Gu, Y., Li, X., Chen, S., Zhang, J., & Marsic, I. (2017). Speech intention classification with multimodal deep learning. In *Canadian conference on artificial intelligence* (pp. 260–271).

Gu, Y., Yang, K., Fu, S., Chen, S., Li, X., & Marsic, I. (2018). Multimodal affective analysis using hierarchical attention strategy with word-level alignment. *arXiv preprint arXiv:1805.08660*.

Gunlogson, C. (2002). Declarative questions. In *Semantics and linguistic theory* (Vol. 12, pp. 124–143).

Hellbernd, N., & Sammler, D. (2016). Prosody conveys speaker's intentions: Acoustic cues for speech act perception. *Journal of Memory and Language*, 88, 70–86.

Hwang, H., & Schafer, A. J. (2009). Constituent length affects prosody and processing for a dative np ambiguity in korean. *Journal of Psycholinguistic Research*, 38(2), 151.

- Kim, J.-B., & Sells, P. (2010). A phrasal analysis of korean comparatives. *Studies in Generative Grammar*, 20, 179–205.
- Lee, K.-H., Chen, X., Hua, G., Hu, H., & He, X. (2018). Stacked cross attention for image-text matching. In *Proceedings of the european conference on computer vision (eccv)* (pp. 201–216).
- Lin, Z., Feng, M., Santos, C. N. d., Yu, M., Xiang, B., Zhou, B., & Bengio, Y. (2017). A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130*.
- Miura, I., & Hara, N. (1995). Production and perception of rhetorical questions in osaka japanese. *Journal of Phonetics*, 23(3), 291–303.
- Pell, M. D., Jaywant, A., Monetta, L., & Kotz, S. A. (2011). Emotional speech processing: Disentangling the effects of prosody and semantic cues. *Cognition & Emotion*, 25(5), 834–853.
- Pires, A., & Taylor, H. (2007). The syntax of wh-in-situ and common ground. In *Proceedings from the annual meeting of the chicago linguistic society* (Vol. 43, pp. 201–215).
- Richards, N. (2016). *Contiguity theory* (Vol. 73). MIT Press.
- Sammler, D., Kotz, S. A., Eckstein, K., Ott, D. V., & Friederici, A. D. (2010). Prosody meets syntax: the role of the corpus callosum. *Brain*, 133(9), 2643–2655.
- Schmidt, A. T., Hanten, G., Li, X., Wilde, E. A., Ibarra, A. P., Chu, Z. D., . . . Levin, H. S. (2013). Emotional prosody and diffusion tensor imaging in children after traumatic brain injury. *Brain injury*, 27(13-14), 1528–1535.
- Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11), 2673–2681.
- Song, C., Han, M., Cho, H. Y., & Lee, K.-N. (2018). Sequence-to-sequence autoencoder based korean text error correction using syllable-level multi-hot vector representation. In *Proceedings of hclt [in korean]* (pp. 661–664).
- Talman, A., Suni, A., Celikkanat, H., Kakouros, S., Tiedemann, J., & Vainio, M. (2019). Predicting prosodic prominence from text with pre-trained contextualized word representations. *arXiv preprint arXiv:1908.02262*.
- Tanner, D. C. (2007). Redefining wernicke’s area: Receptive language and discourse semantics. *Journal of allied health*, 36(2), 63–66.
- Yoon, S., Byun, S., Dey, S., & Jung, K. (2019). Speech emotion recognition using multi-hop attention mechanism. In *Ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 2822–2826).
- Yun, J. (2019). Meaning and prosody of wh-indeterminates in korean. *Linguistic Inquiry*, 50(3), 630–647.