RIT

Fusion Strategy for Prosodic and Lexical Representations of Word Importance

Sushant Kafle sushant@mail.rit.edu
Cecilia O. Alm coagla@rit.edu
Matt Huenerfauth matt.huenerfauth@rit.edu

20th Annual Conference of the International Speech Communication Association
INTERSPEECH 2019

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.

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.

: *it was really not very good uh-*

Image Source: https://www.writermag.com

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.

1 : *it was really ~~not~~ very good uh-*

: *it was really not very good uh-*

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.

1 : ***it was really*** *~~not~~* ***very good uh-***

: ***it was really not very good uh-***

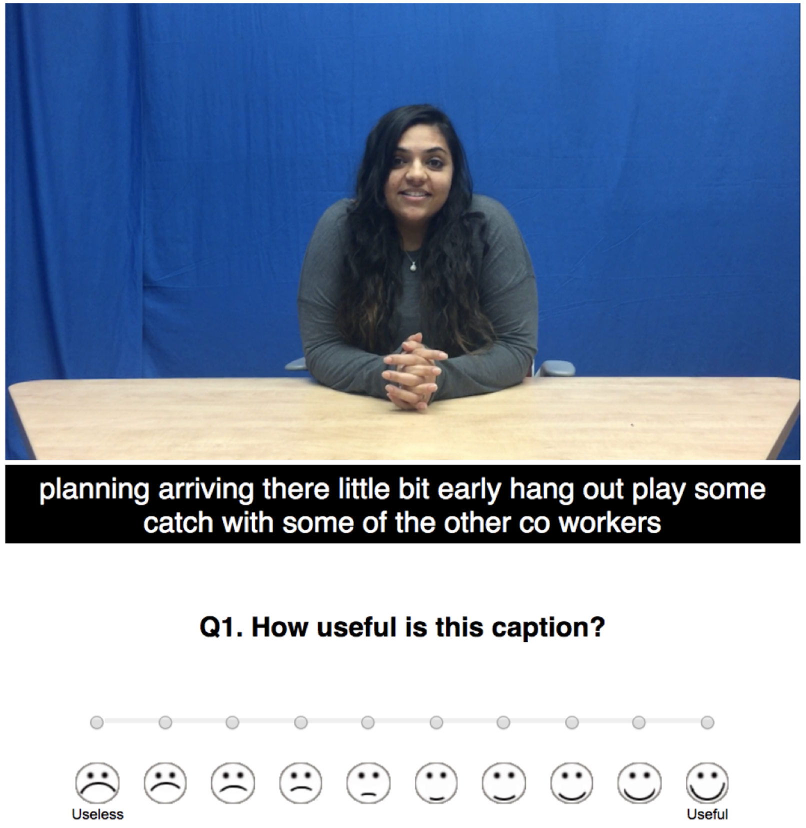
2 : *~~it~~* ***was*** *~~really~~* ***not very good*** *~~uh-~~*

Image Source: https://www.writermag.com

Motivation

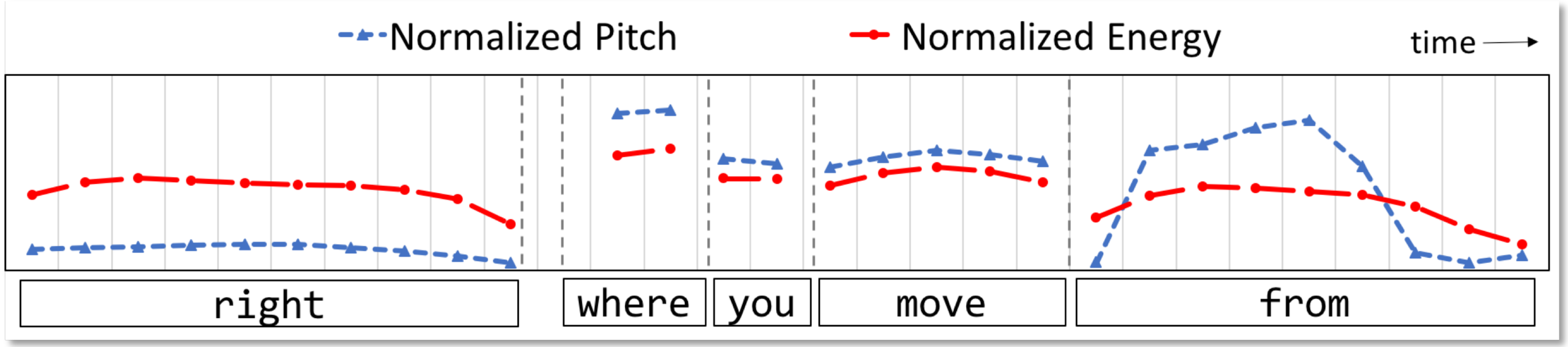
- Automatically predicting the importance of words in spoken language is useful for tasks such as:
  - Speech Recognition (ASR) evaluation
  - Text Classification, and,
  - Summarization.

- Differential treatment of errors, based on word importance, is shown to **correlate better** with human subjective judgement of ASR quality in captioning applications for d/Deaf and Hard-of-hearing users. (Kafle and Huenerfauth, 2017)



(Figure from: Kafle and Huenerfauth, 2017)

Importance of Prosody

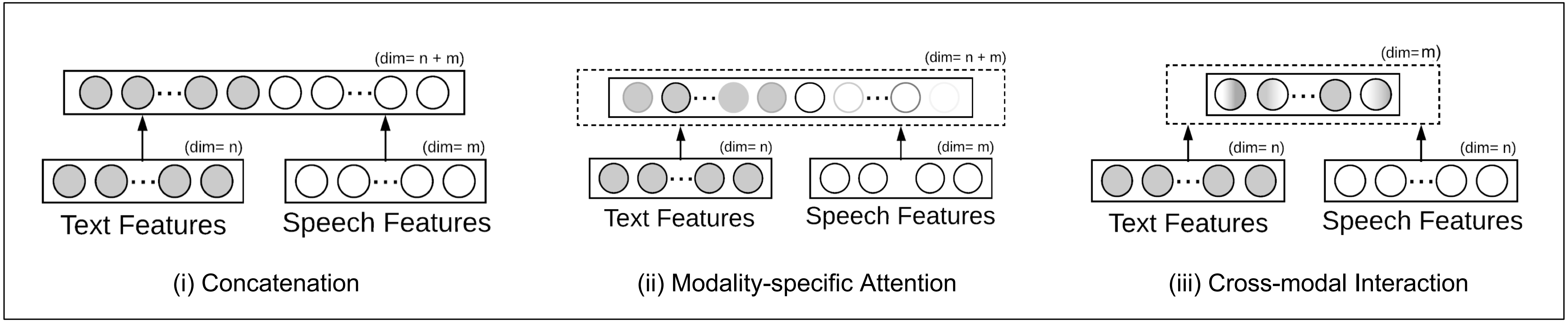


(Figure from: Kafle et. al, 2019)

- Spoken messages include prosodic cues that focus a **listener's attention** on the most important parts of the message to help disambiguate meaning.
- It also 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.

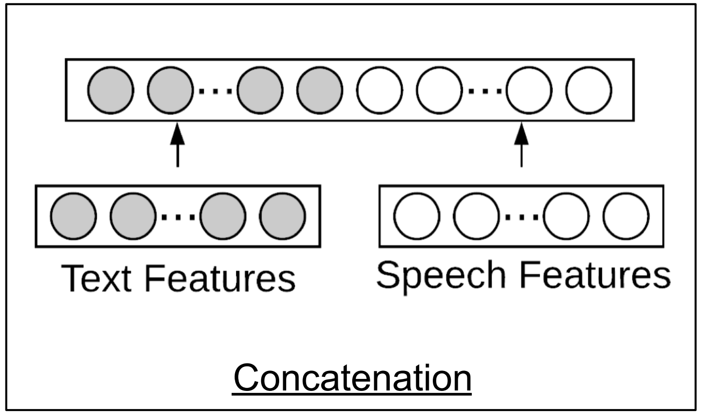
Goal of this work

- Starting from the assumption that acoustic-prosodic cues help identify important speech content, this investigates:
  - Representation strategies for combining lexical and prosodic features at the word-level
  - Performance of each when predicting word importance



Prior Work: Joint Feature Representation

- The most common strategy for joint representation of features is through concatenation. However, it fails to fully capture **cross-feature** (cross-modal) interactions. (Zadeh et. al., 2017; Liu et. al., 2018)



- Consequently, several other feature representation strategies, that consider cross-modal interaction, has been investigated. (Zadeh et. al., 2017; Liu et. al., 2018; Wang et. al.)

- This work explores text-and-speech representations for word importance prediction.

Prior Work: Word Importance Prediction

- Portrayal of word importance prediction **as keyword extraction task**:
  - Considers importance of words at a document level rather than at a sentential or a phrase level. (Liu, 2011; Hulth, 2002; Sheeba, 2012)

- This setup treats each word as a *term* in a document such that all words identified by a *term* receive a uniform importance score, **without regard to their local context**.

- Recently, models that consider contextualized word representation has been proposed. However, they consider **unimodal features** (lexical or prosodic, not both) which may be insufficient for conversational speech-based application.

Lexical-Prosodic Representation

for word importance prediction

Attention-based Feature Fusion

- This feature fusion architecture captures how prosody impacts the lexical semantics of the spoken word.

- Uses architecture to learn a **composition vector** that controls the contribution of prosodic features on word meaning:

Attention-based Feature Fusion

- This feature fusion architecture captures how prosody impacts the lexical semantics of the spoken word.

- Uses architecture to learn a **composition vector** that controls the contribution of prosodic features on word meaning:

$$\alpha = tanh(W \cdot [S; L] + b)$$

S: Acoustic-prosodic feature representation.

L: Lexical feature representation.

Z: Lexical-Prosodic Representation

Attention-based Feature Fusion

- This feature fusion architecture captures how prosody impacts the lexical semantics of the spoken word.

- Uses architecture to learn a **composition vector** that controls the contribution of prosodic features on word meaning:

$$\alpha = tanh(W \cdot [S; L] + b)$$

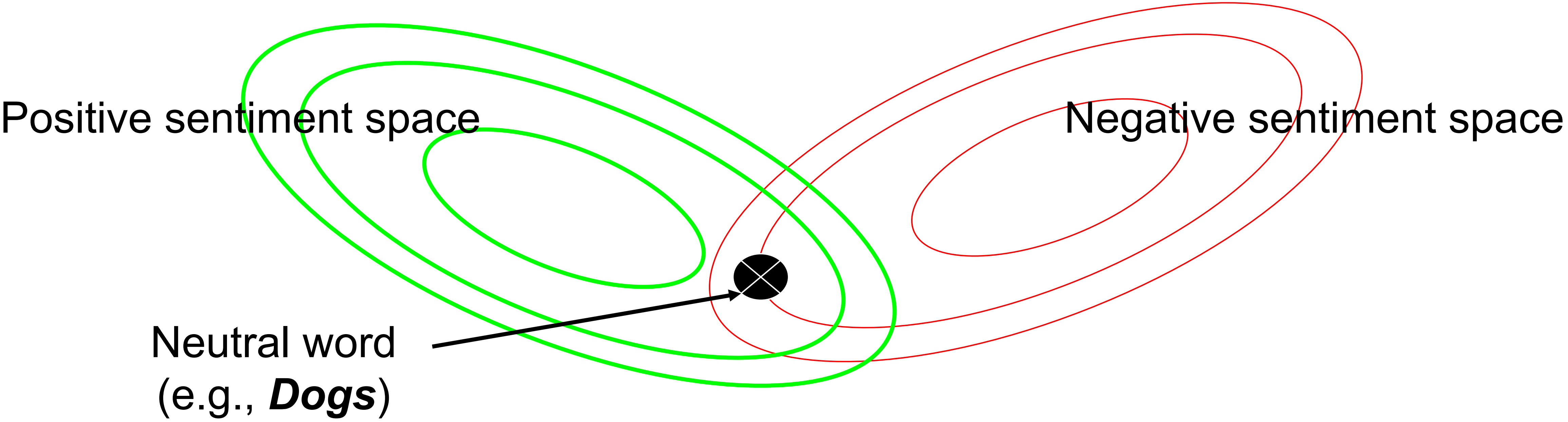
$$Z = L + \underbrace{\boxed{\alpha \cdot S}}_{\text{Lexical Shift}}$$

S: Acoustic-prosodic feature representation.

L: Lexical feature representation.

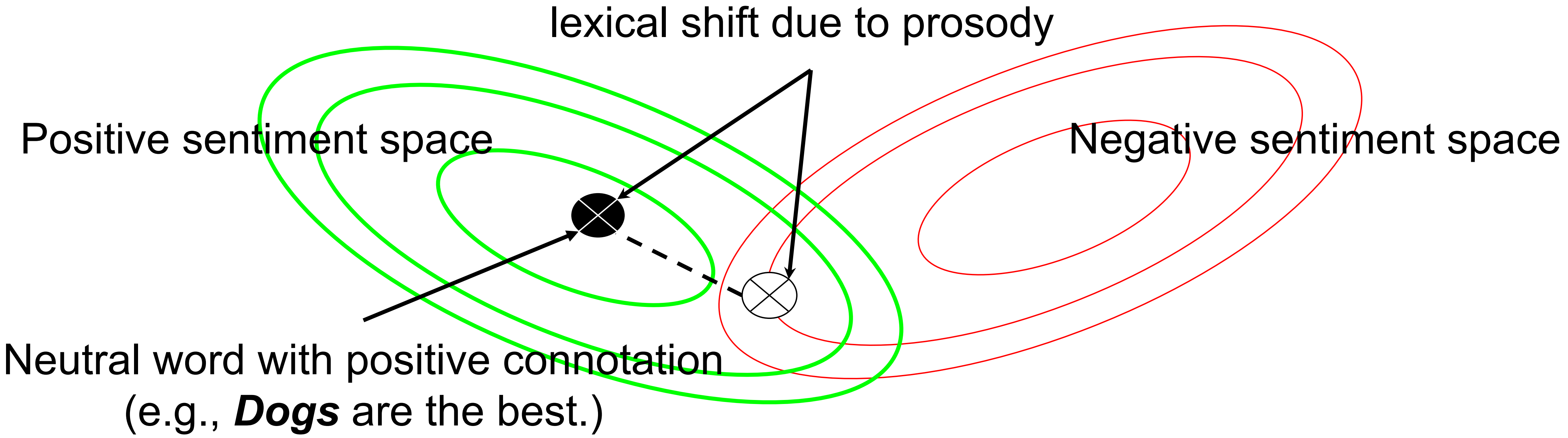
Z: Lexical-Prosodic Representation

Attention-based Feature Fusion



- Composition vector projects lexical embeddings into an appropriate semantic space, based on their prosodic character.

Attention-based Feature Fusion



- Composition vector projects lexical embeddings into an appropriate semantic space, based on their prosodic character.

Experimental Setup

- **Dataset:** Word Importance Corpus (Kafle et. al, 2018)
  - Consists of over 25k unique words with manually annotated importance information on a dialogue turn label.

- **Lexical Representation:** GloVe (Pennington et. al., 2014)

- **Acoustic-Prosodic Representation:** bi-RNN based subnetwork (Kafle et. al, 2019) operating over features such as:
  - Energy-related features (RMS min, max, mean, median, time of max, etc.)
  - Frequency-related features (F0 min, max, mean, median, time of max, etc.)
  - Voicing features (HNR, VUR, Spectral-tilt, etc.)
  - Spoken-lexical features (word duration, articulation rate, etc.)

Exp. 1: Error Analysis of Unimodal Models

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

- Lexical-only model had a lower RMS error when predicting word importance, but it performed poorly for OOV words. For OOVs, the prosodic-only model did better.

Intervention: Attention Supervision

- Allows incorporation of heuristic constraints into a model.
- We supervised attention during training to rely on prosodic features when the word is an out-of-vocabulary (OOV) word.

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

Exp. 2: Comparison of Fusion Strategies (1 of 2)

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*

- Comparison of different models combining lexical and prosodic cues. Per column, the top two results are marked with (∗) and (†) symbols. Our model has lower RMS error overall AND for OOVs.

Exp. 2: Comparison of Fusion Strategies (1 of 2)

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*

wo/ Attention Supervision

- Comparison of different models combining lexical and prosodic cues. Per column, the top two results are marked with (∗) and (†) symbols. Our model has lower RMS error overall AND for OOVs.

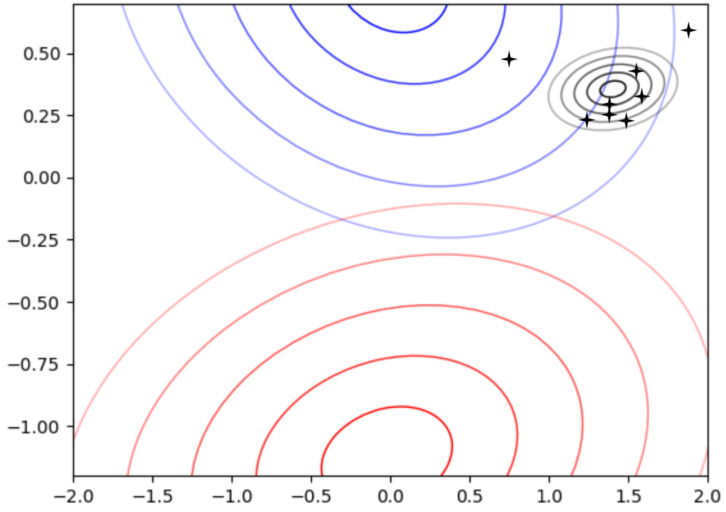
Exp. 2: Comparison of Fusion Strategies (2 of 2)

Models	RMS (across ranges)			τ-b
	HI	MID	LOW	
CONCAT	$22.81^{\dagger}$	$13.07^{\dagger}$	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	$10.31^{\star}$	$60.04^{\dagger}$
RAVEN	29.04	$12.50^{\star}$	11.65	59.77
Proposed ($\lambda = 0$)	25.13	13.29	10.85	59.80
Proposed ($\lambda = 0.8$)	$22.4^{\star}$	13.27	$10.60^{\dagger}$	$61.35^{\star}$

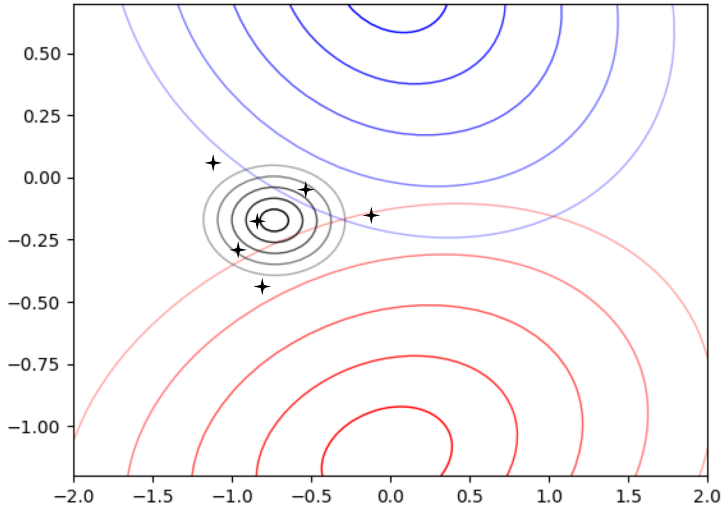
- Comparison of models on ordinal-range classes, and Kendall-tau (**τ**-b) rank prediction correlation. The top two results per column are marked with ($*$) and (†) symbols. Our proposed model performs better for high and low importance words.

Exp. 3: Prosodic Deviation

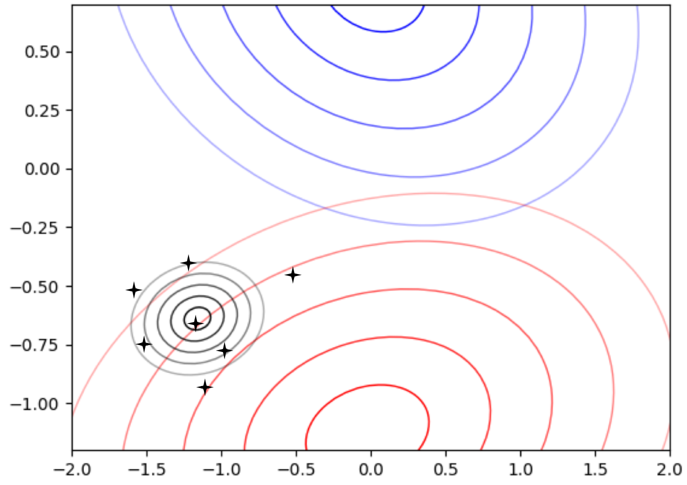
Word: **Love**



Word: **Night**



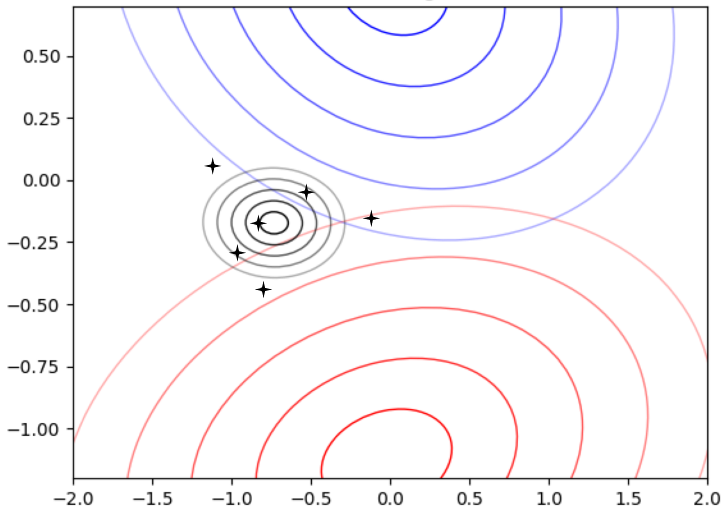
Word: **Cold**



- Visualization of the combined representation of words ***love, night, cold*** in difference spoken contexts. The blue (top) and red (bottom) contours represent the distribution of all positive and all negative sentiment words, respectively.

Exp. 3: Prosodic Deviation

Word: **Night**



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

- The word ***night*** in different spoken contexts with corresponding positioning in the contour plot.

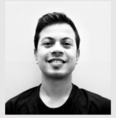
Conclusion

- Showed that by incorporating features from speech into the lexical embeddings, we can enhance the performance of word-importance prediction systems.
- Proposed an attention-based feature representation strategy that learns to adjust lexical feature representation of spoken words to reflect the post-lexical meaning conveyed through prosody.
- Demonstrate the utility of incorporating modality-specific heuristic into training.

RIT

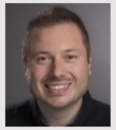
Any Questions?

This material was based on work supported 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.



Sushant Kafle
Ph.D. Student
Rochester Institute of Technology
Golisano College of Computing and Information Sciences
Computing and Information Sciences Ph.D. Program
Email: sxk5664@rit.edu

Cecilia O. Alm
Associate Professor
Rochester Institute of Technology
Comp Ling & Speech Proc Lab
Email: coagla@rit.edu

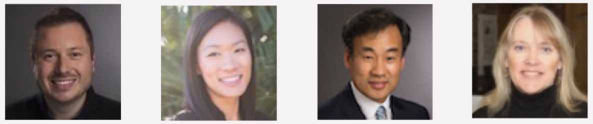


Matt Huenerfauth
Professor
Rochester Institute of Technology
School of information (iSchool)
Email: matt.huenerfauth@rit.edu

CAIR
Center for Accessibility and Inclusion Research

CAIR brings together researchers working on computer accessibility and assistive technology for people with disabilities, technology for older adults, and educational technologies. http://cair.rit.edu/

Faculty



Some CAIR Researchers