Retrieving Speech Samples with Similar Emotional Content Using a Triplet Loss Function

UT Dallas MSP Multimodal Signal Processing Laboratory

THE UNIVERSITY OF TEXAS AT DALLAS

John Harvill, Mohammed AbdelWahab, Reza Lotfian, Carlos Busso

Erik Jonsson School of Engineering & Computer Science at the University of Texas at Dallas, Richardson, Texas 75080, USA

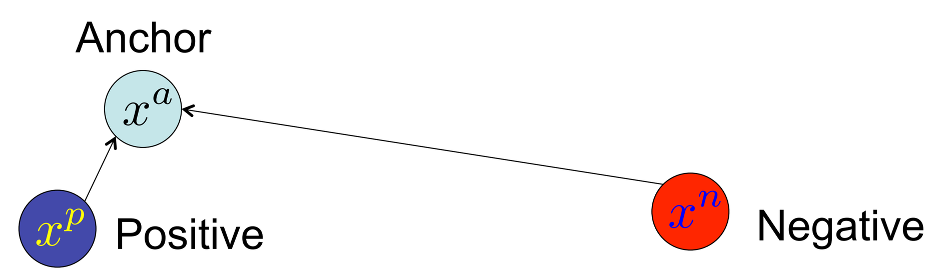
Motivation

Background:

- Identify speech with similar emotional content
  - Can a deep neural network learn to determine distance between expressive behaviors?
  - Can a given emotional descriptor facilitate this task?
  - How well can a computer perform this task?

Our Work:

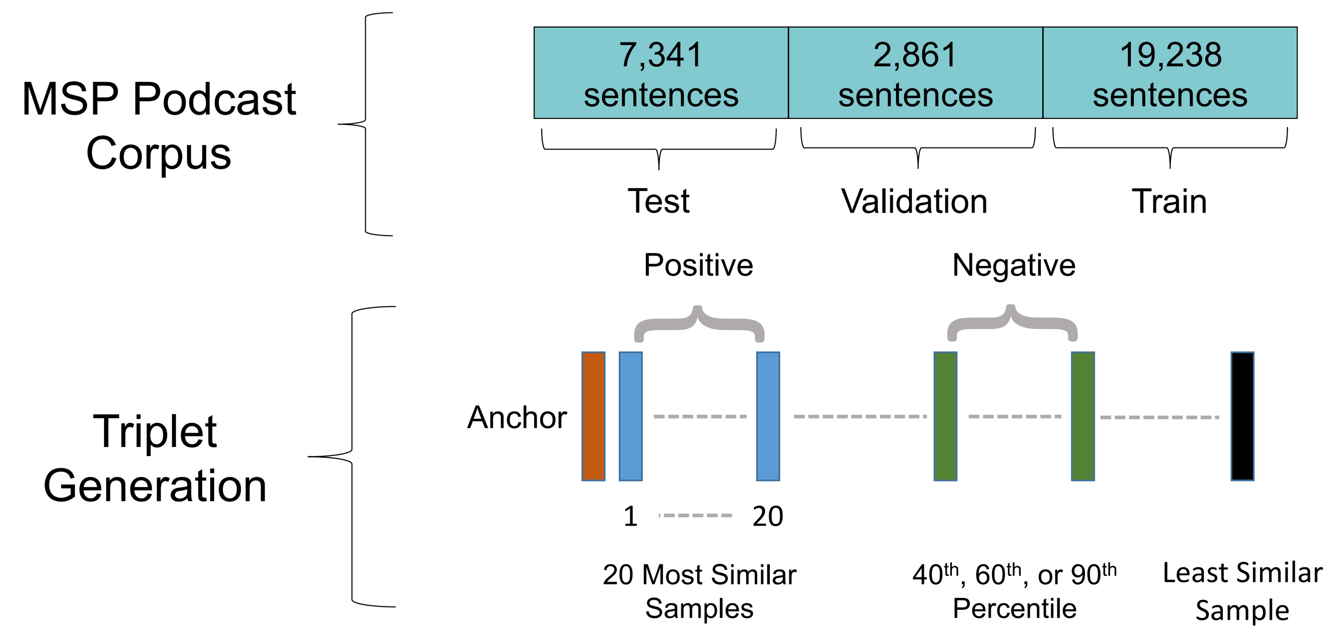
- Preference learning using triplet loss functions



- Compare emotional descriptors for this task:
  - Emotional attributes versus categorical emotions
- Compare results with human performance

MSP-Podcast Corpus

- Emotional corpus collected at UT-Dallas
  - Multiple sentences from speakers appearing in various podcasts (2.75s – 11s)
- Annotated on Amazon Mechanical Turk
  - VAD: Valence, arousal and dominance (Euclidean distance)
  - Primary emotions: anger, sadness, happiness, fear, surprise, disgust, contempt, neutral state and other (KL divergence)
- One triplets per sample within a given partition



Network Structure and Training

Acoustic Features

- Interspeech 2013 Computational Paralinguistic Challenge set (6,373D)
- calculated from low-level descriptors

Network Structure

- Trained, validated, tested on speaker independent sets
- 3 hidden layers, 1,024 nodes, ReLU activation
- 512 dimension embedding
- Dropout 0.2, batch normalization, 15 epochs
- 19,238 training triplets

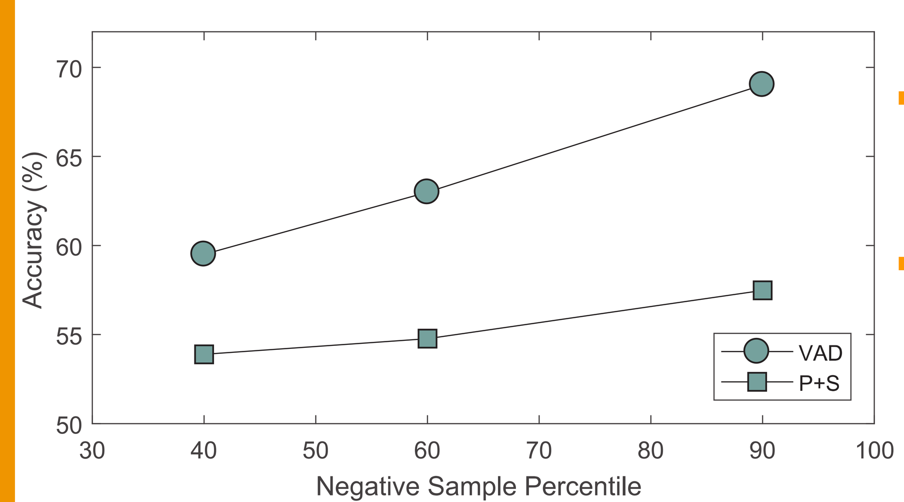
Desired Mapping

$$||f(x_i^a) - f(x_i^p)||_2^2 + \alpha < ||f(x_i^a) - f(x_i^n)||_2^2$$
$$\forall f(x_i^a), f(x_i^p), f(x_i^n) \in \Gamma$$

Loss Function

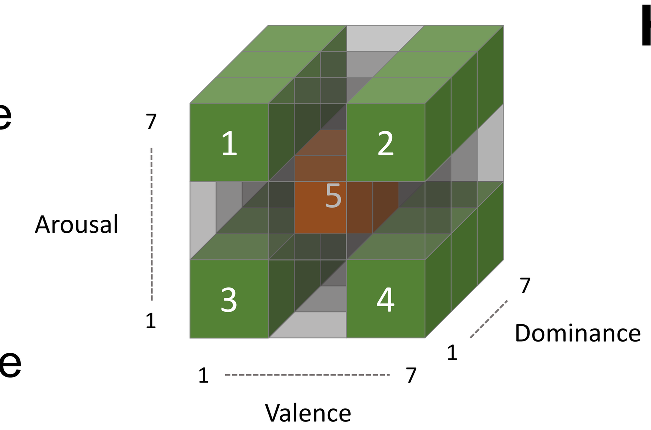
$$L = \max[0, \sum_i^N (||f(x_i^a) - f(x_i^p)||_2^2 - ||f(x_i^a) - f(x_i^n)||_2^2 + \alpha)]$$

Human and Machine Performance



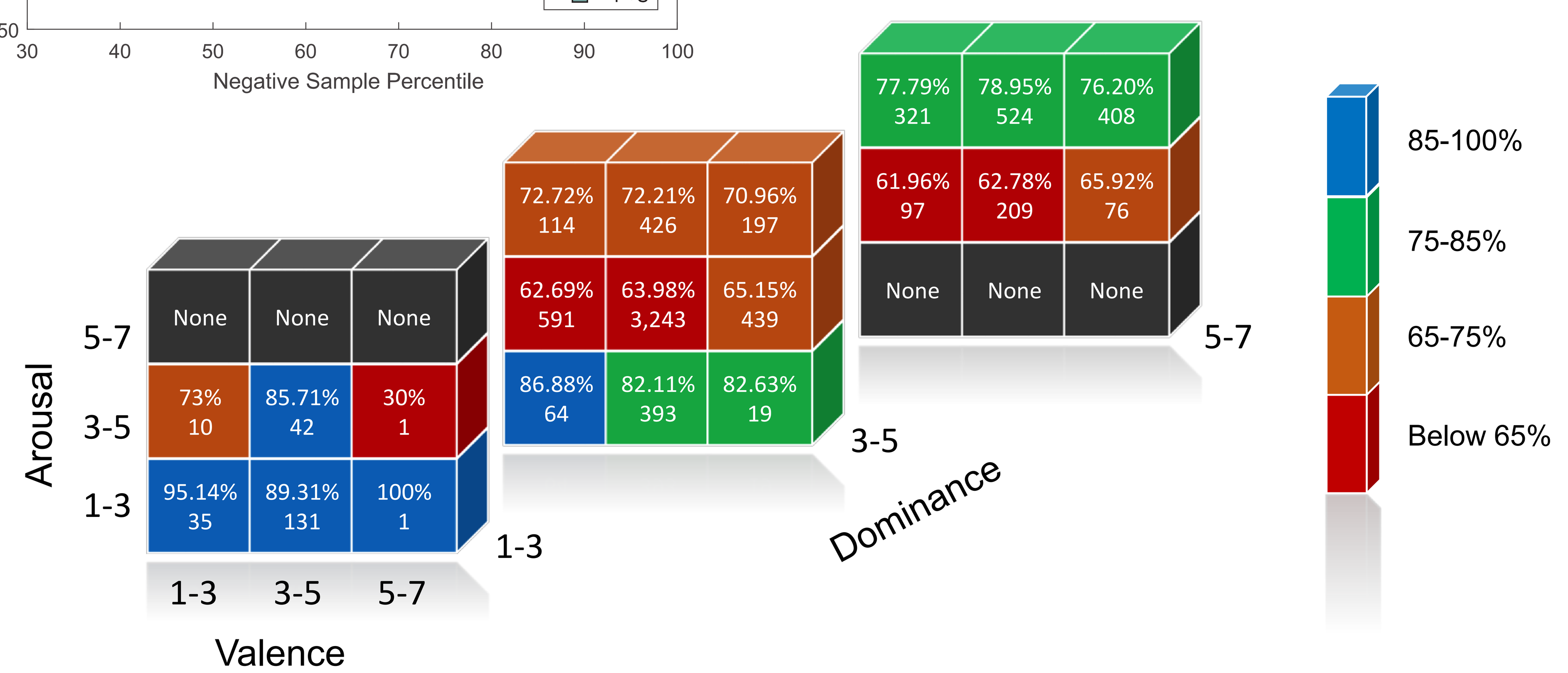
Global Performance

- Results per percentile used to get negative sample
  - VAD provides better representation for this task
- VAD results in terms of location of anchor
  - Extreme VAD regions lead to better performance



Human Performance (VAD)

- Perceptual evaluation
  - 60 triplets (5 regions in VAD)
  - Model performs better in 90%
  - Humans perform better in 40%



Region	Triplet Network	Triplet Network	Human Performance
Region	Entire Test Set	60 Triplets	60 Triplets
	90th Percentile	90th Percentile	90th Percentile
1	76.5%	82%	**86.7%**
2	74.5%	**96%***	73.3%
3	89.8%	**98%***	82.2%
4	83.5%	**74%**	66.7%
5	64.0%	65%	**75.3%**
	40th Percentile	40th Percentile	40th Percentile
1	66.7%	64%	**75.6%**
2	66.0%	64%	**80.0%***
3	78.8%	**78%**	65.6%
4	65.5%	**66%**	57.8%
5	56.6%	49%	**60.0%***

Conclusions

- Evaluating emotional similarity is better in the VAD space than in the categorical space
- Triplets with expressive anchors are easier to discriminate than triplets with neutral anchors
- Model performance is similar to human performance and superior in some regions of the VAD space

Future Work

- Improve accuracy for triplets with anchors in the middle of the VAD space
- Collect more perceptual evaluation data
- Perform similar study on data from one subject to learn that subject's emotional expression in depth

This work was funded by NSF CAREER award IIS-1453781