Emotion and Sentiment Detection

COMS 6998

FALL 2022

Outline

- Emotion in speech
  - Emotion theory
  - Emotional speech corpora
  - Features for emotional speech
  - Models for emotion recognition
  - Expressive synthetic speech
- Sentiment and emotion in text

Emotion in Speech

Outline

- Emotion in speech
  - Emotion theory
  - Emotional speech corpora
  - Features for emotional speech
  - Models for emotion recognition
  - Expressive synthetic speech
- Sentiment and emotion in text

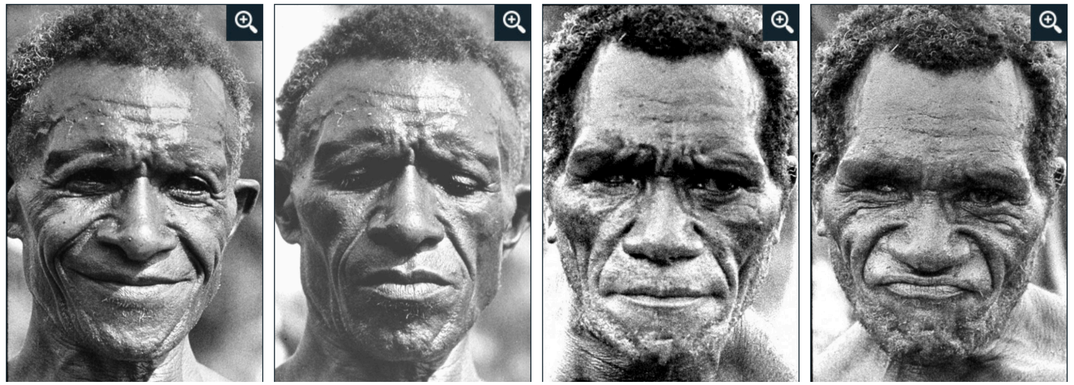
What is Emotion?

- Two families of theories of emotion
  - **Categorical** approach
    - Emotions are categories
    - Limited number of basic emotions
  - **Dimensional** approach
    - Emotions are dimensions
    - Limited number of labels but unlimited number of emotions

Emotion - Categorical Approach

(Ekman et al., 1987)

- Discrete ‘basic emotions’
- Originate from facial expressions



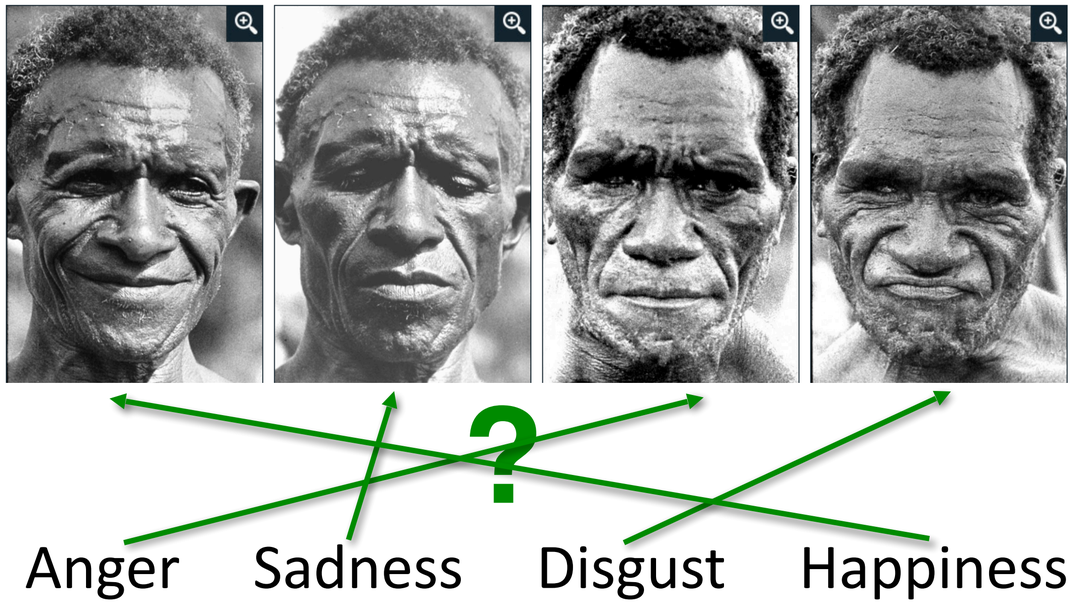
?

Anger Sadness Disgust Happiness

Emotion - Categorical Approach

(Ekman et al., 1987)

- Discrete ‘basic emotions’
- Originate from facial expressions



Emotion - Dimensional Approach

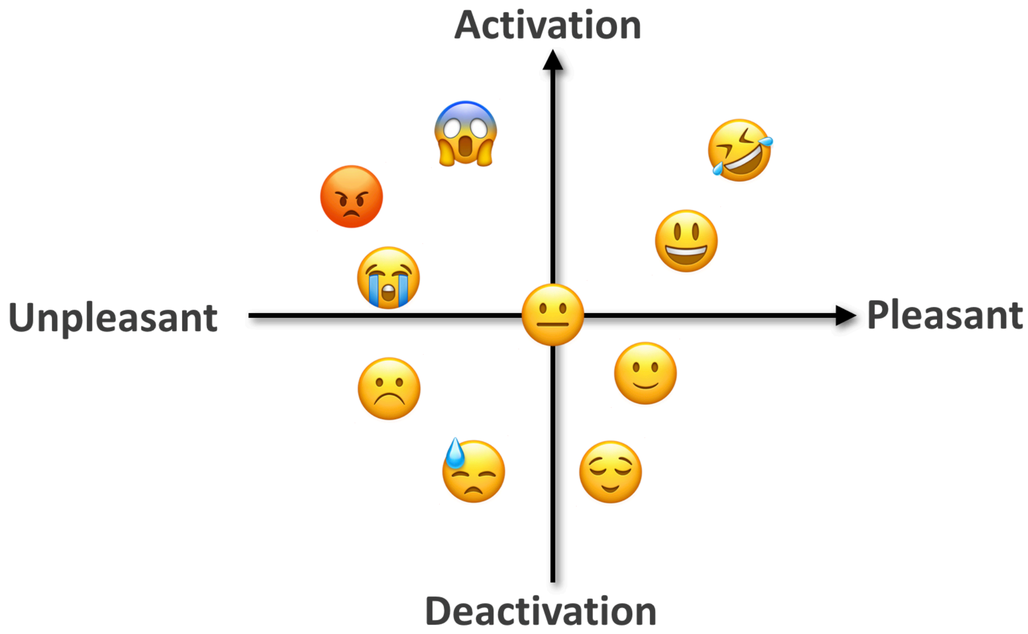
(Russell and Barrett, 1999)

- Continuous multi-dimensional space
- Common physiological system

Emotion - Dimensional Approach

(Russell and Barrett, 1999)

- Continuous **Arousal-Valence** space
- Common physiological system



Why Study Emotional Speech?

- Recognition
  - Customizing virtual assistants
  - Anger/frustration in call centers
  - Confidence/uncertainty in online tutoring systems
  - “Hot spots” in meetings
- Generation
  - TTS for virtual assistants, computer games, etc.
- Other applications: Speaker state identification
  - Deception, charisma, sleepiness, interest, humor…

- Some emotional clues are only in speech

Outline

- Emotion in speech
  - Emotion theory
  - Emotional speech corpora
  - Features for emotional speech
  - Models for emotion recognition
  - Expressive synthetic speech
- Sentiment and emotion in text

Emotion in Speech

Acted speech

- ✓ Easier to collect & control
- ✗ Extreme emotions
- Mostly categorical approach
- Examples: (Emotional Prosody Speech)
  - *Which emotion do you hear?*

Spontaneous speech

- ✗ Harder to collect & annotate
- ✓ Subtle changes in emotion
- Both categorical & dimensional approach

Emotion in Speech

Acted speech

- ✓ Easier to collect & control
- ✗ Extreme emotions
- Mostly categorical approach
- Examples: (Emotional Prosody Speech)
  - Happy, Sad, Angry, Bored

Spontaneous speech

- ✗ Harder to collect & annotate
- ✓ Subtle changes in emotion
- Both categorical & dimensional approach

Emotion in Speech

Acted speech

- ✓ Easier to collect & control
- ✗ Extreme emotions
- Mostly categorical approach
- Examples: (Emotional Prosody Speech)
  - Happy, Sad, Angry, Bored

Spontaneous speech

- ✗ Harder to collect & annotate
- ✓ Subtle changes in emotion
- Both categorical & dimensional approach
- Example: (AT&T “How May I Help You?” System)
  - Categorical emotion(s)?
  - Arousal and Valence?

Emotion in Speech

Acted speech

- ✓ Easier to collect & control
- ✗ Extreme emotions
- Mostly categorical approach
- Examples: (Emotional Prosody Speech)
  - Happy, Sad, Angry, Bored

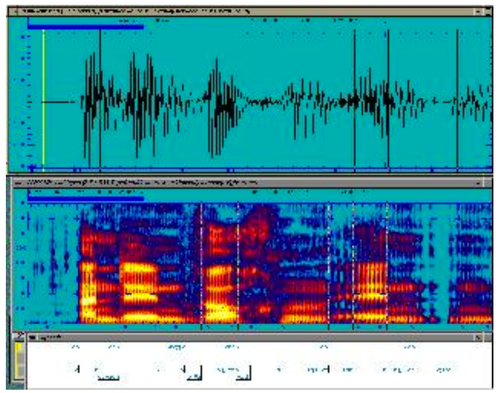
Spontaneous speech

- ✗ Harder to collect & annotate
- ✓ Subtle changes in emotion
- Both categorical & dimensional approach
- Example: (AT&T “How May I Help You?” System)
  - Neutral -> frustrated -> angry
  - Arousal ↑, Valence ↓

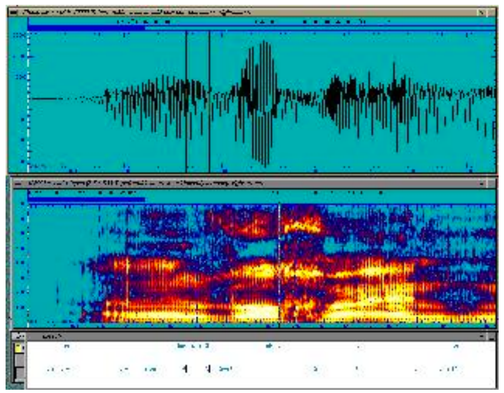
Emotional Speech Corpora - Acted & Categorical

(EmoDB, German)

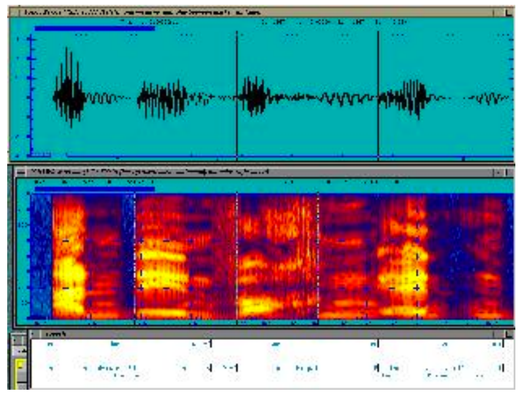
Neutral



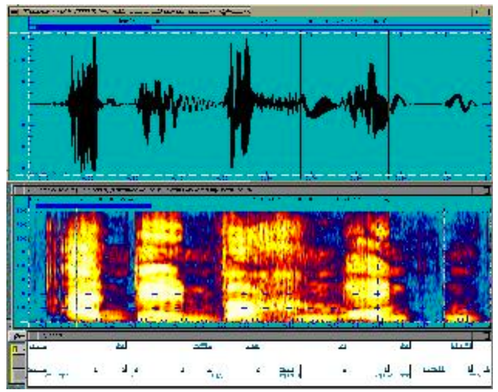
Bored



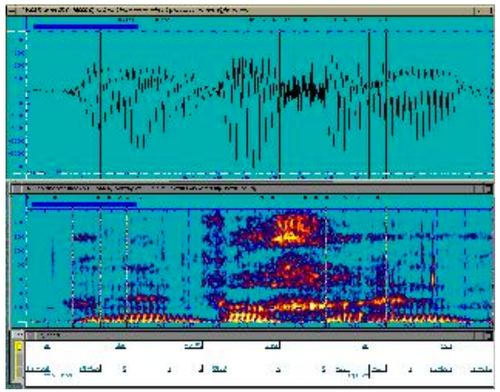
Angry



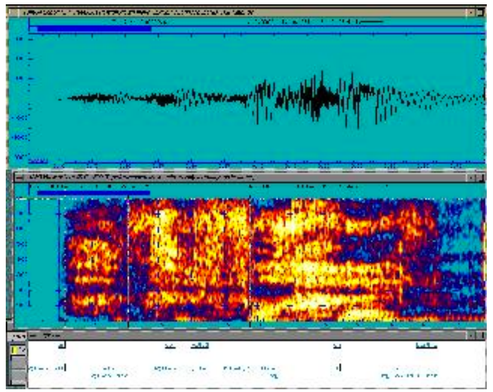
Happy



Sad



Frightened



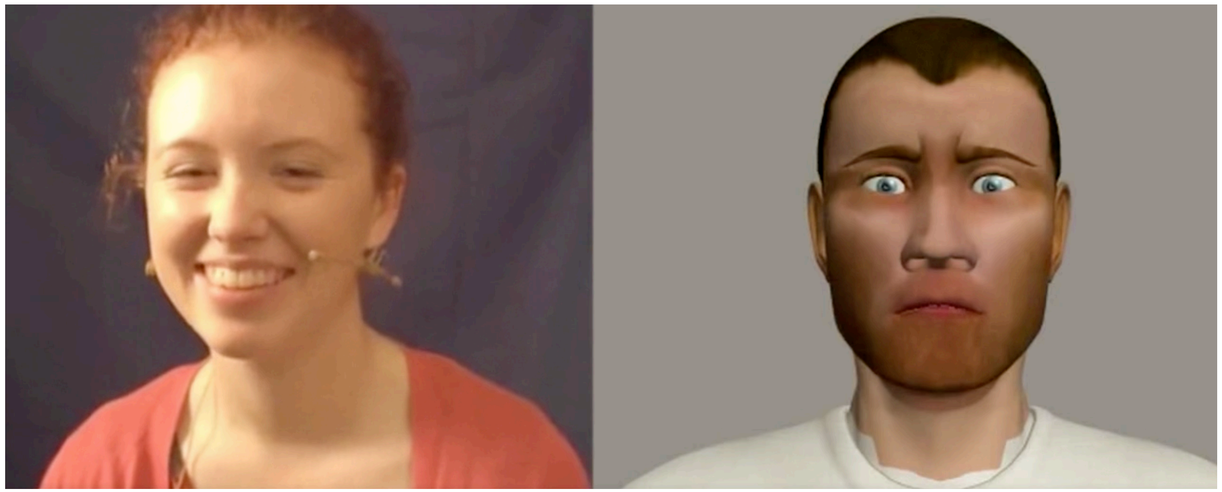
Acted & Categorical Speech: Actors vs Students

(Emotional Prosody Speech)	(Mandarin Affective Speech)
Sad	Anger
Happy	Elation
Angry	Neutral
Bored	Panic
Interested	Sadness

Spontaneous Speech with Dimensional Annotations
(SEMAINE database)

- The goal of the Sensitive Artificial Listener operator (right) is to engage the user (left) in emotional conversations
  - “Anything nice happened this week?” “It’s all rubbish.”
- 6-8 annotators. Annotations range from -1 to 1 with 20ms intervals.

  - Valence score : -0.88
  - Valence score : 0.58
  - Valence score : 0.83

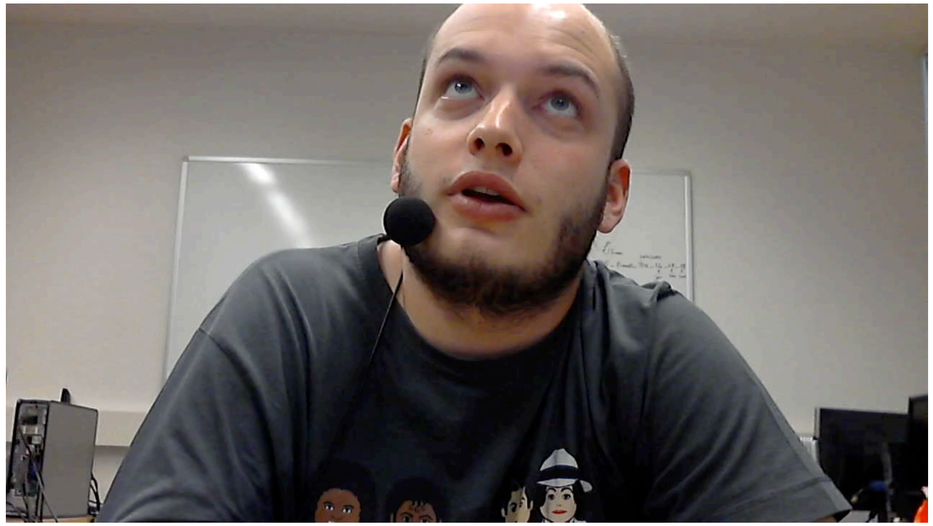


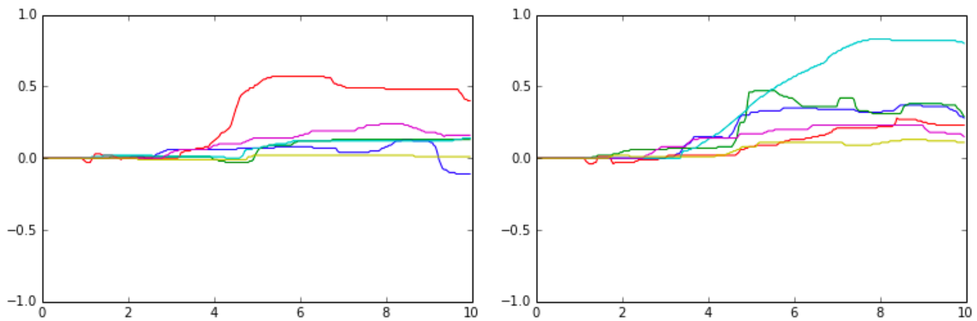
Spontaneous Speech with Dimensional Annotations
(RECOLA database)

- 3 hours of audio, visual, and physiological recordings of between 46 French speaking participants
- Participants were asked to reach consensus on how to survive in a disaster scenario
- 6 annotators. Annotations range from -1 to 1 with 40ms intervals.

Spontaneous Speech with Dimensional Annotations

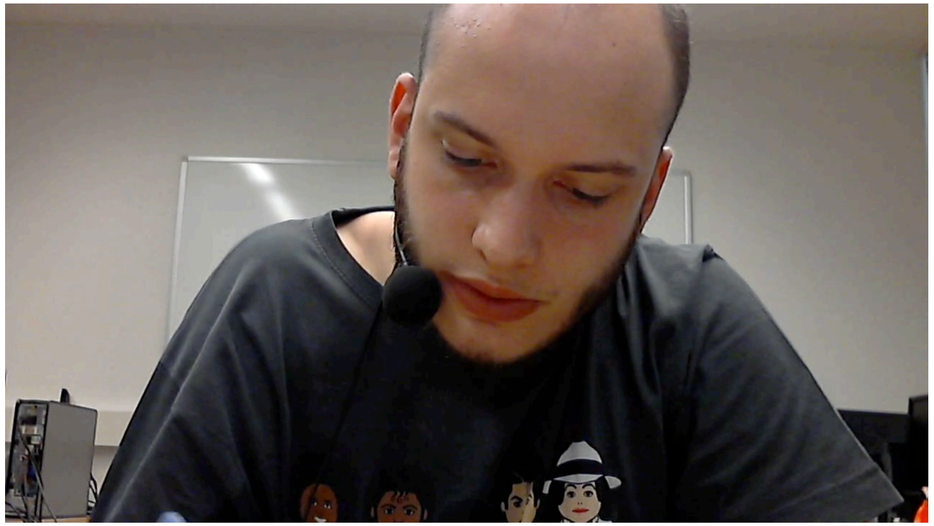
(RECOLA database)

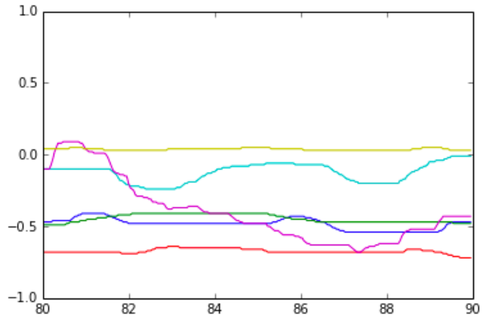




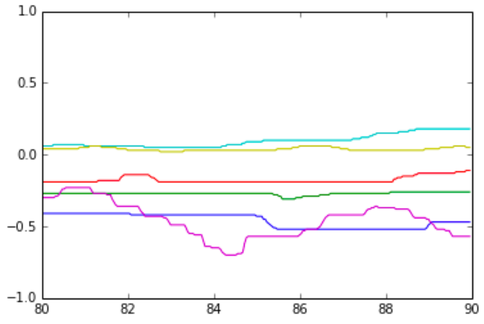
Arousal

Valence





Arousal



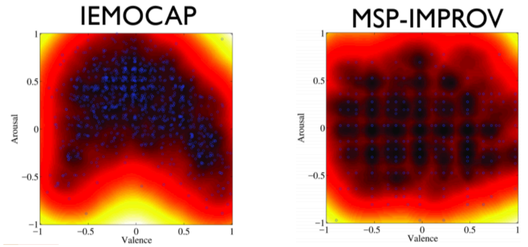
Valence

Partial List of the Existing Emotion Corpora

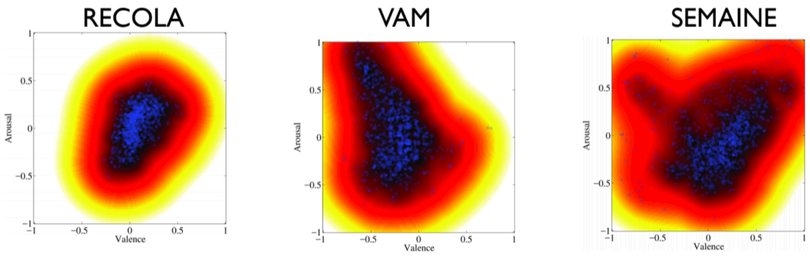
- Lack of naturalness (acted)
- Unbalanced emotional content (spontaneous)
- Limited in size, limited number of speakers

Corpus	Size	# Spkr	Type	Lang.
IEMOCAP [10]	12h26m	10	acted	English
MSP-IMPROV [19]	9h35m	12	acted	English
CREMA-D [2]	7,442 samples	91	acted	English
Chen Bimodal [20]	9,900 samples	100	acted	English
Emo-DB [6]	22m	10	acted	German
GEMEP [21]	1,260 samples	10	acted	-
VAM-Audio [15]	48m	47	spont.	German
TUM AVIC [22]	10h23m	21	spont.	English
SEMAINE [13]	6h21m	20	spont.	English
FAU-AIBO [14]	9h12m	51	spont.	German
RECOLA [11]	2h50m	46	spont.	French

Acted

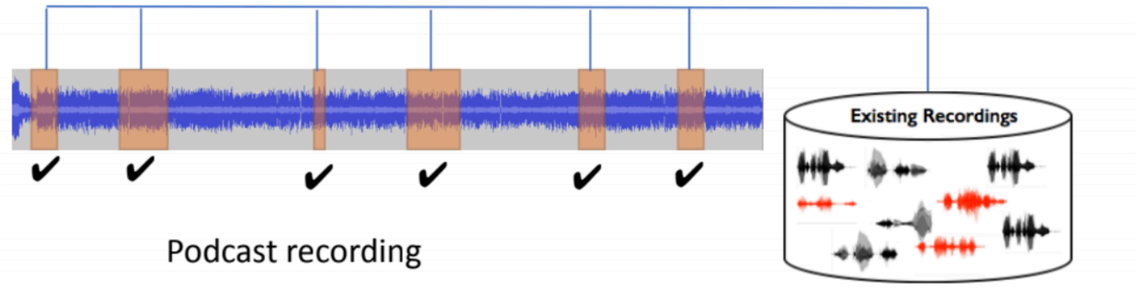


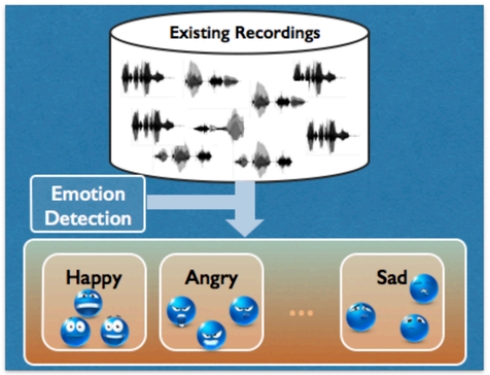
Spontaneous



MSP-Podcast Corpus

- Use existing podcast recordings, divided into speaker turns
- Emotion retrieval to balance the emotional content
- Annotate using crowdsourcing framework

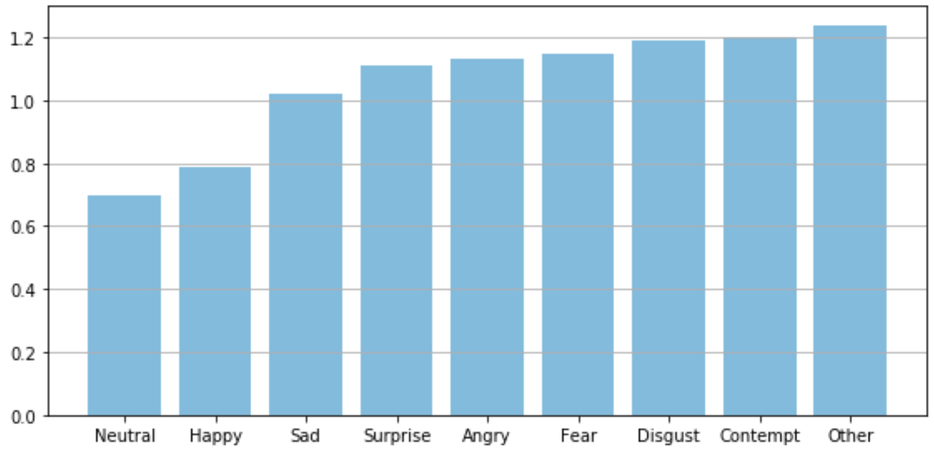




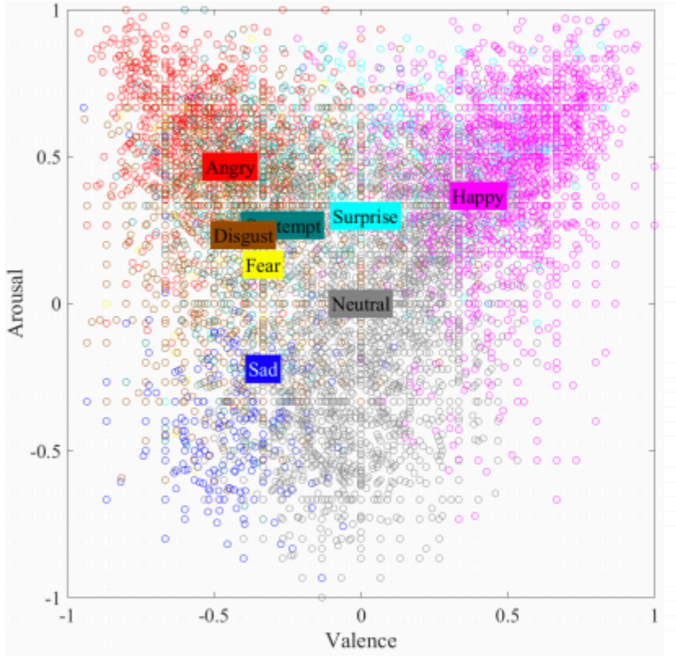
- 62140 speaking turns, ~100 hours of speech
- Similar approach: CMU-MOSEI dataset

MSP-Podcast Corpus

- Annotations
  - Dimensional: activation, valence, dominance
  - Categorical: anger, happiness, sadness, disgust, surprised, fear, contempt, neutral and other



Rater disagreement per emotion category



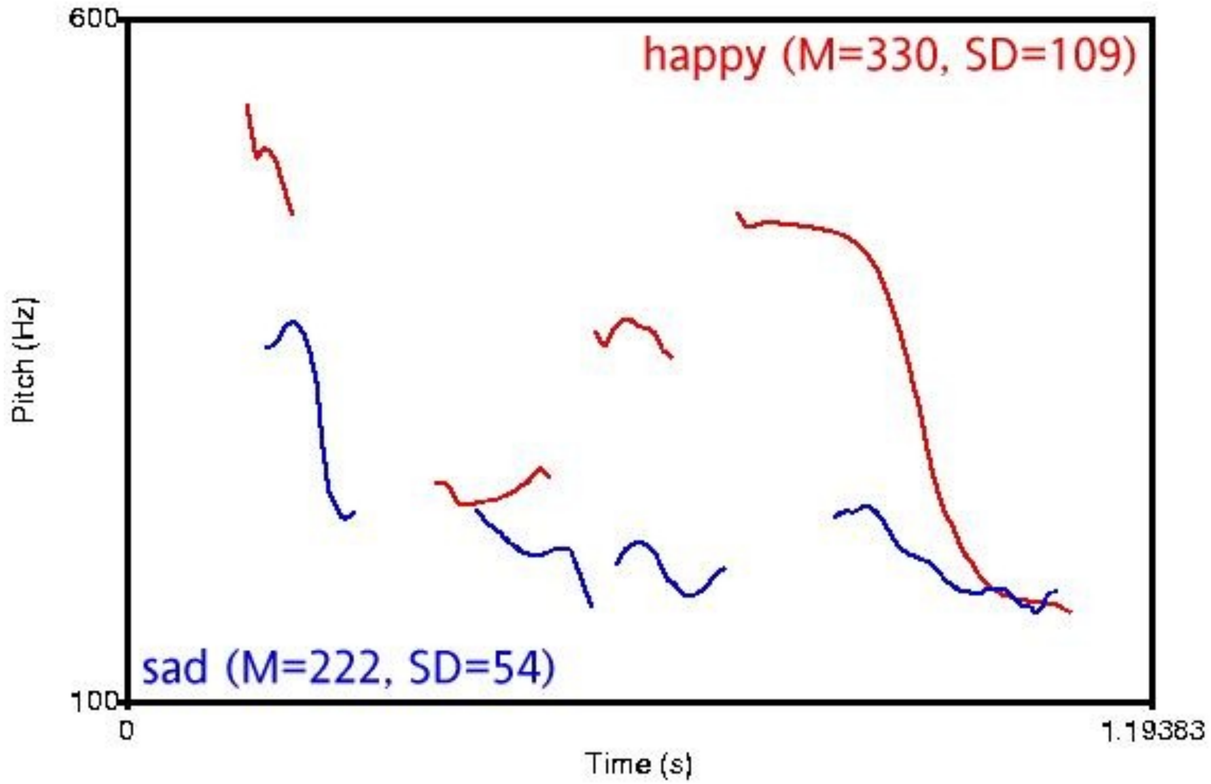
Emotion distribution in arousal/valence space

Outline

- Emotion in speech
  - Emotion theory
  - Emotional speech corpora
  - Features for emotional speech
  - Models for emotion recognition
  - Expressive synthetic speech
- Sentiment and emotion in text

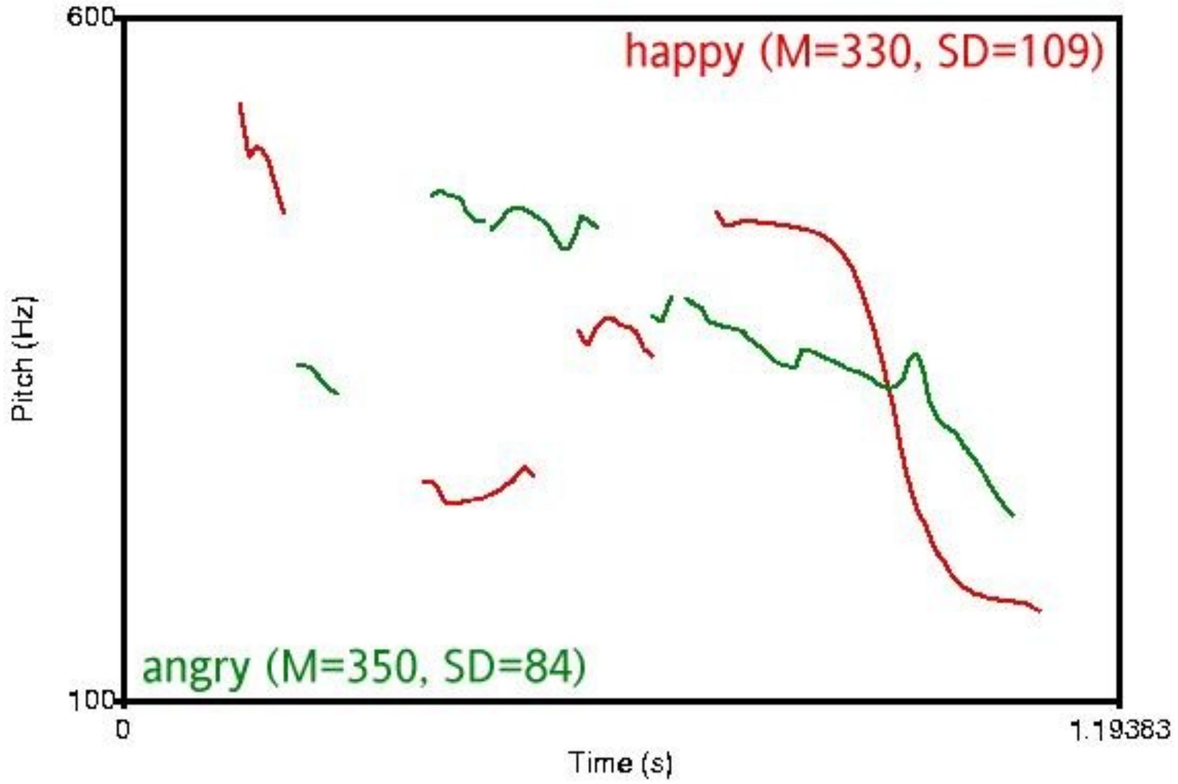
Features for Emotional Speech - Pitch

Different Valence / Different Arousal

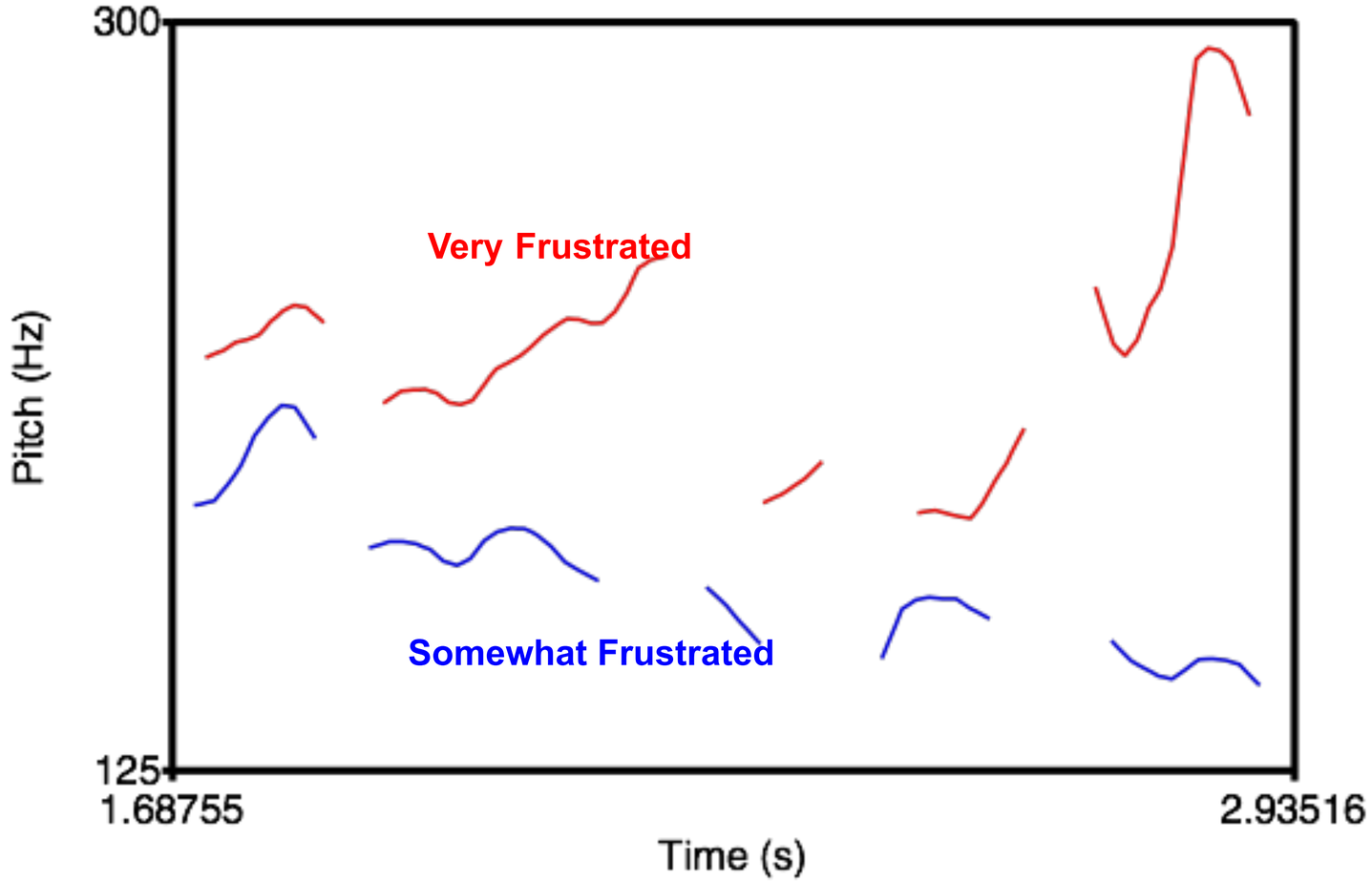


Features for Emotional Speech - Pitch

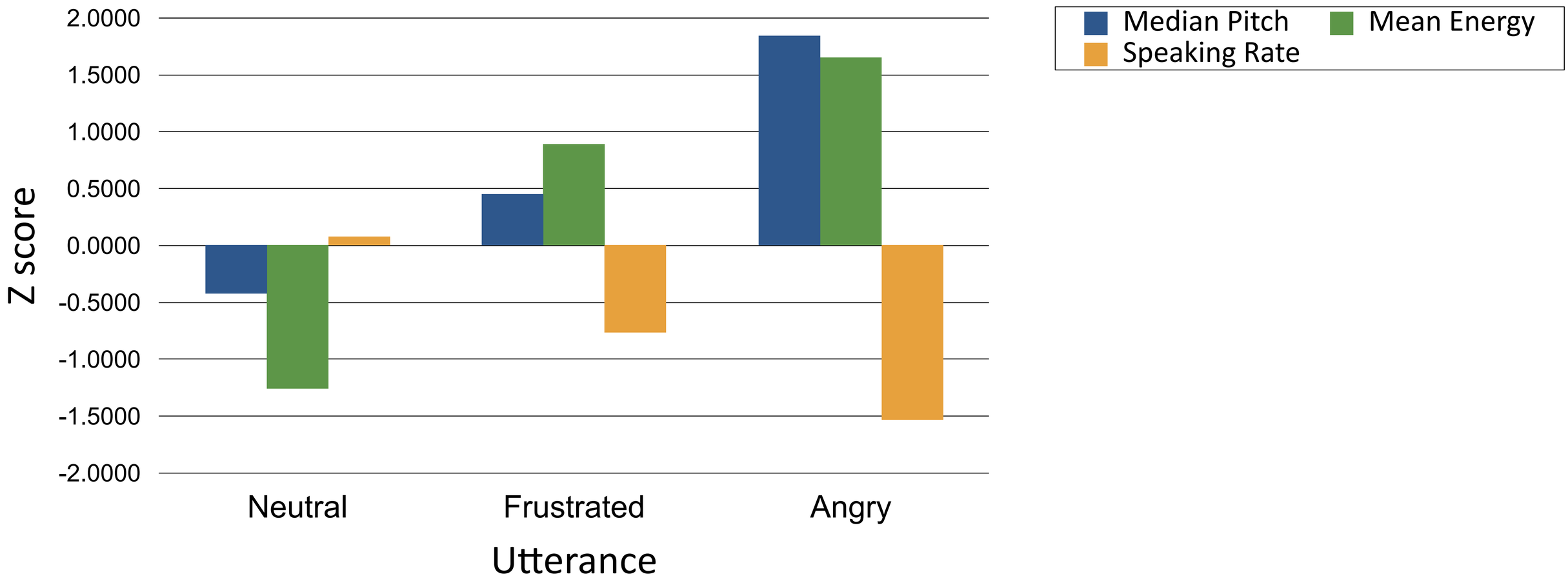
Different Valence / ***Same*** Arousal



Pitch Contour Differences



Features for Emotional Speech



Outline

- Emotion in speech
  - Emotion theory
  - Emotional speech corpora
  - Features for emotional speech
  - Models for emotion recognition
  - Expressive synthetic speech
- Sentiment and emotion in text

Emotion Recognition in Speech

Categorical Approach

- Discrete ‘basic emotions’
- Classification problem

Dimensional Approach

- Continuous Arousal - Valence space
- Regression problem

Emotion Recognition - Categorical

(Dellaert et al., 1996)

- Emotions: happiness, sadness, anger, fear, normal
- Data: 5 speakers * 5 emotions * 50 utterances = 1250
- Human performance on 4-way classification: 82% (anger easiest, fear hardest)

- Features: rhythm, smoothed pitch, individual voiced parts
- Best model (KNN with majority voting of specialists): 79.5%

Emotion Recognition - Categorical

(Petrushin, 1999)

- Dataset 1
  - Emotions: happiness, anger, sadness, fear, and normal
  - Data: **(30+5)** speakers * 5 emotions * 4 utterances = 700
  - Human performance: 63.5% (anger easiest, fear hardest)

  - Features: F0, energy, speaking rate, first three formants and their bandwidths
  - Best model: feature selection + ensembles of 15 neural networks
    - Normal: 60-75%, happiness: 60-70%, anger: 70-80%, sadness: 70-85%, fear: 35-55%
    - Average: ~70%

Emotion Recognition - Categorical

(Petrushin, 1999)

- Dataset 2 (call centers)
  - Distinguish between two states (**arousal**):
    - "Agitation": anger, happiness and fear
    - "Calm": normal state and sadness
  - Data: 56 telephone messages (15~90 seconds)
    - Automatically split into 1-3 second chunks
  - Model: feature selection + ensembles of neural networks
  - Average accuracy: 77%

Emotion Recognition - Categorical

(Liscombe et al., 2003)

- 10 emotions and neutral
  - ‘Positive’ **valence**: confident, encouraging, friendly, happy, interested
  - ‘Negative’ **valence**: angry, anxious, bored, frustrated, sad
- Subset: 4 speakers * (10+1) emotions = 44
- Human rating: Emotions in the same valence group are positively correlated; vise versa
- Features: Pitch, energy, speaking rate; nuclear accent, pitch contour
- Results:
  - Pitch, energy, and speaking rate are correlated with arousal
  - Spectral tilt and pitch contour are correlated with valence

Emotion Recognition - Categorical

(Liscombe et al., 2003)

- Full dataset (EPSaT): **1760** utterances
- Acoustic-prosodic features:
  - Pitch, energy, speaking rate; nuclear accent, pitch contour
- Accuracy: ~75%
- Feature selection:
  - Some single features performed as well or better than the entire feature set

Emotion Recognition - Categorical

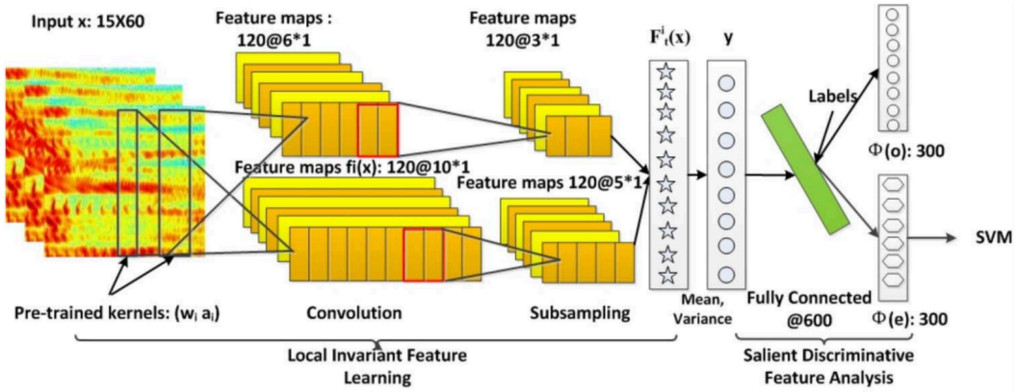
(Jin et al., 2015)

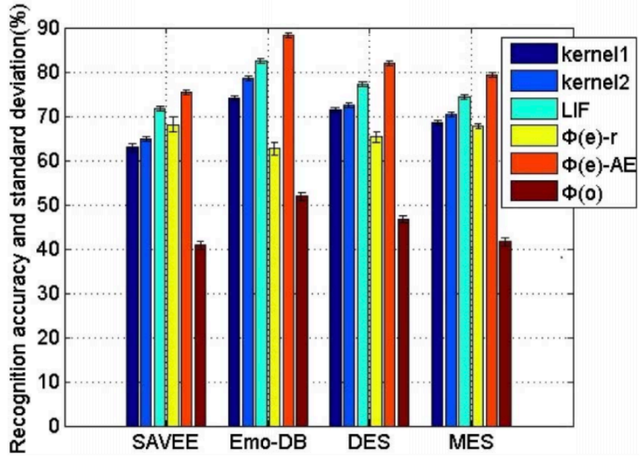
- Emotions: happy, angry, sad, and neutral
- Data (USC-IEMOCAP): **5531** utterances
- Features:
  - Acoustic: openSMILE (intensity, F0, jitter, shimmer and MFCCs)
  - **Lexical**: emotion vector (eVector), Bag-of-Words (BoW)
- Best model: SVM, late fusion of acoustic and lexical features
- Accuracy: 69.2%

Emotion Recognition - Categorical

(Mao et al. 2014)

- **Using neural networks on spectrograms**
- Evaluation on 4 datasets:
  - anger, disgust, fear, happiness, sadness, surprise, and neutral
  - anger, disgust, fear, joy, sadness, boredom, and neutral
  - anger, joy, surprise, sadness, and neutral
  - anger, joy, surprise, sadness, and disgust

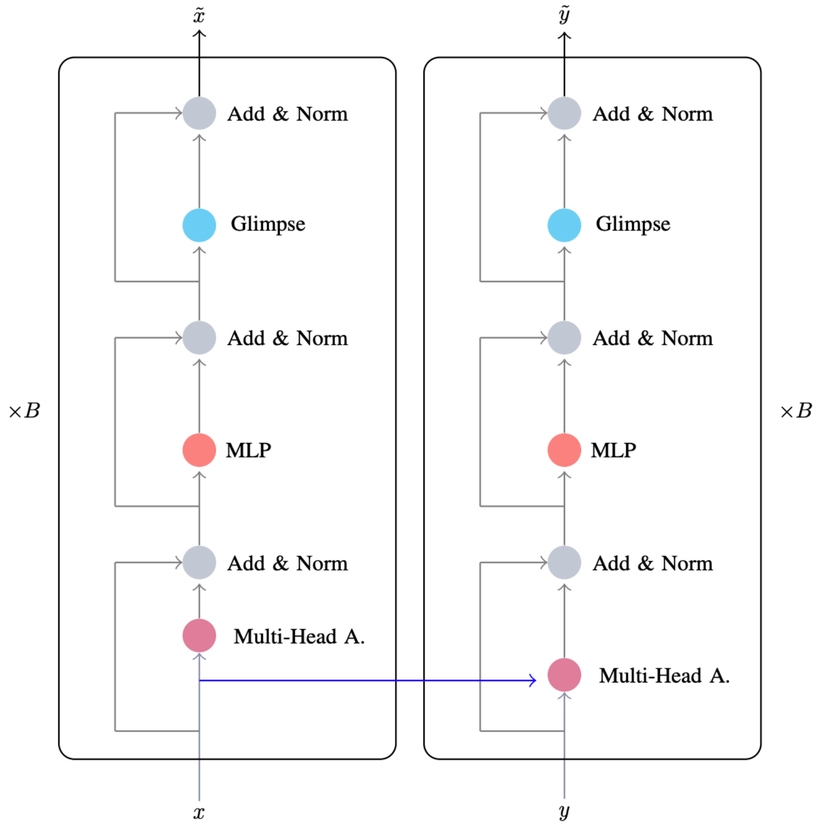




Emotion Recognition - Categorical

(Delbrouck et al. 2020)

- **Transformer-based multimodal** joint-encoding
- Dataset: CMU-MOSEI (Youtube video segments)
- Modalities:
  - Linguistic: GloVe word vector
  - Acoustic: mel-spectrograms
  - Visual: a pre-trained CNN



Emotion Recognition in Speech

Categorical Approach

- Discrete ‘basic emotions’
- Classification problem

Dimensional Approach

- Continuous **Arousal - Valence** space
- Regression problem

Emotion Recognition - Dimensional

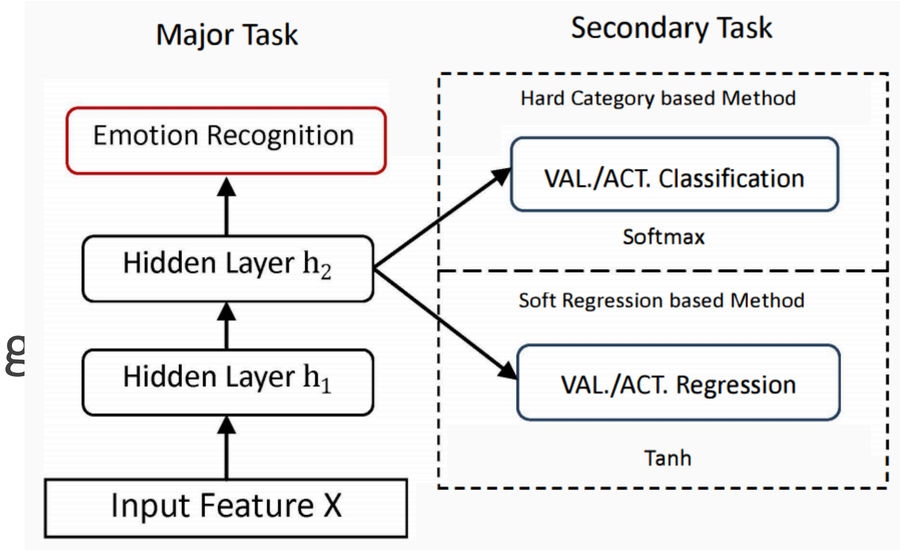
(Karadogan and Larsen, 2012)

- Emotion: arousal, valence (discrete value 1~9 for each dimension)
- Data: 59 short movie clips (5~25 seconds)
- Features and models:
  - Acoustic: openSMILE, feature selection, support vector regression
  - Lexical: affective norms for English words (ANEW) for keywords with arousal & valence scores + latent semantics analysis (LSA) to generate emotion scores for other words
- Results (mean absolute error): arousal: 1.28, valence: 1.40
- **Semantic features-> valence, acoustic features -> arousal**

Emotion Recognition - Dimensional

(Xia and Liu, 2015)

- **Major task: angry, happy, sad, and neutral**
- **Secondary task: valence, activation**
  - Classification
    - map the continuous labels into low, medium, hig
  - Regression
    - project the continuous labels into [-1,1] range
- Data (USC-IEMOCAP): 5531 utterances
- Features: openSMILE
- Model: Deep Belief Network (DBN) + SVM

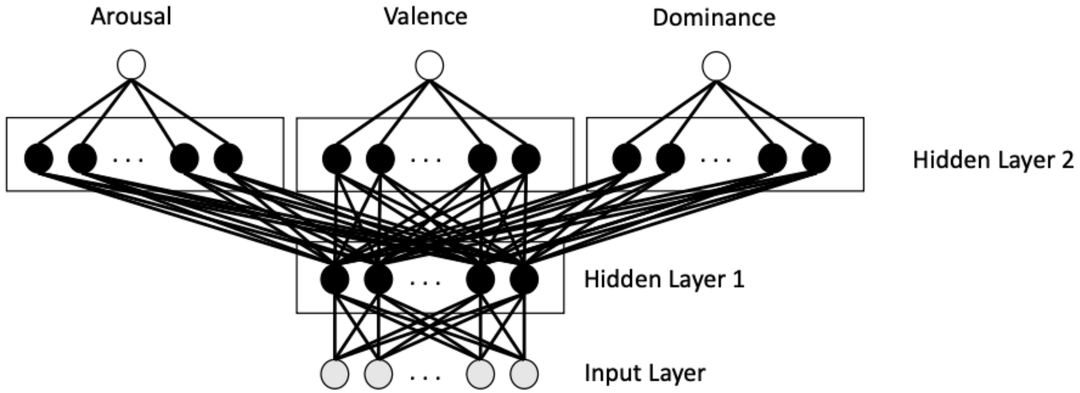


System		UA
Static features		59.7
DBN framework		60.5
$Static features + Pred._{act,val}$		60.7
$DBN features + Pred._{act,val}$		61.1
Multi-task learning	Hard category	**62.2**
	Soft regression	**62.5**

Emotion Recognition - Dimensional

(Parthasarathy and Busso, 2017)

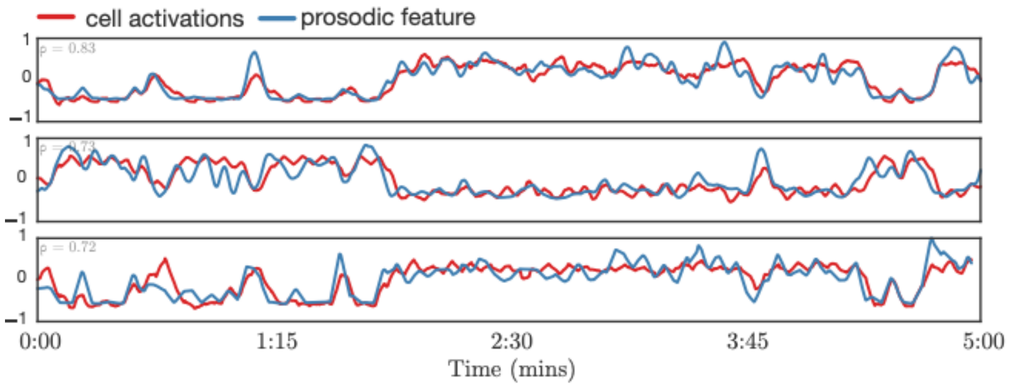
- Emotion: arousal, valence, dominance (discrete value 1~7 scaled to [-1,1])
- Data (MSP-PODCAST): 12,621 speech segments (2~11 seconds)
- Features: openSMILE
- Best multi-task learning model: shared first layer, individual second layer
- Results: A: 0.7635, V: 0.2894, D: 0.7130
- **Multi-task learning helps**
  - **Dominance > valence > arousal**
  - Learn better representations



Emotion Recognition - Dimensional

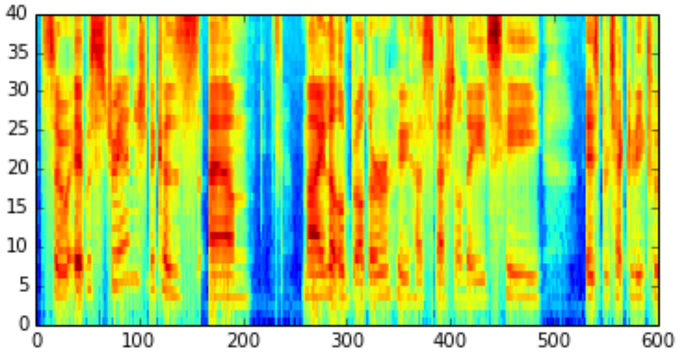
(Trigeorgis et al., 2016)

- Emotion: arousal, valence (continuous value [-1,1])
- Data (RECOLA): 46 French conversations, 5 min each
- Feature: **raw waveforms**
- Model: convolutional recurrent neural networks
- Results (Concordance correlation coefficient): arousal: 0.686, valence: 0.261
- **Some cells learn acoustic features automatically**
  - Range of RMS energy (ρ = 0.81)
  - Loudness (ρ = 0.73)
  - Mean of fundamental frequency (ρ = 0.72)

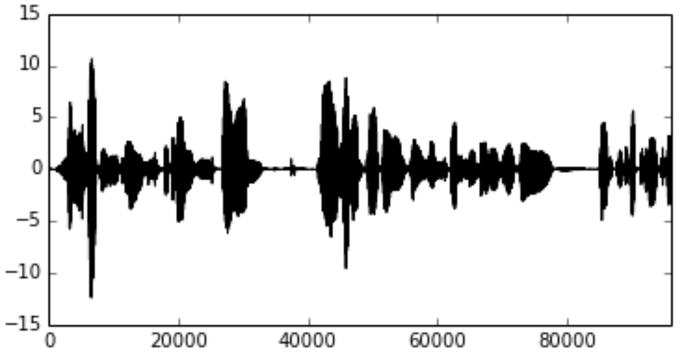


Emotion Recognition - Dimensional

Spectrogram



Waveform



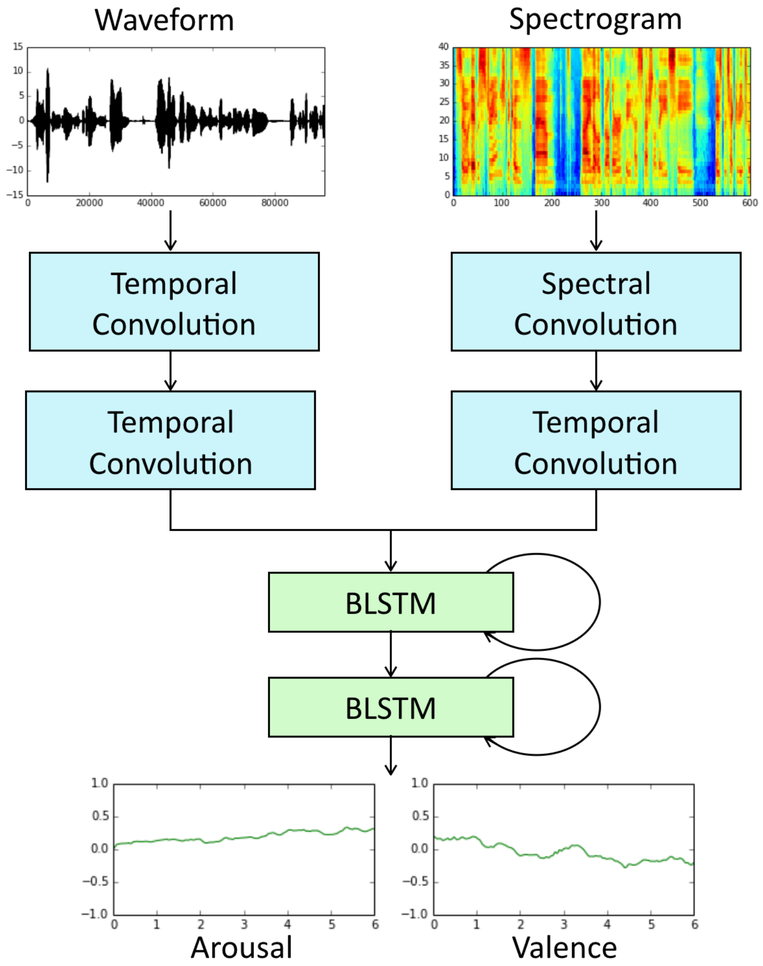
Do spectrograms and waveforms contain complementary information for emotion recognition in speech?

Emotion Recognition - Dimensional

(Yang and Hirschberg, 2018)

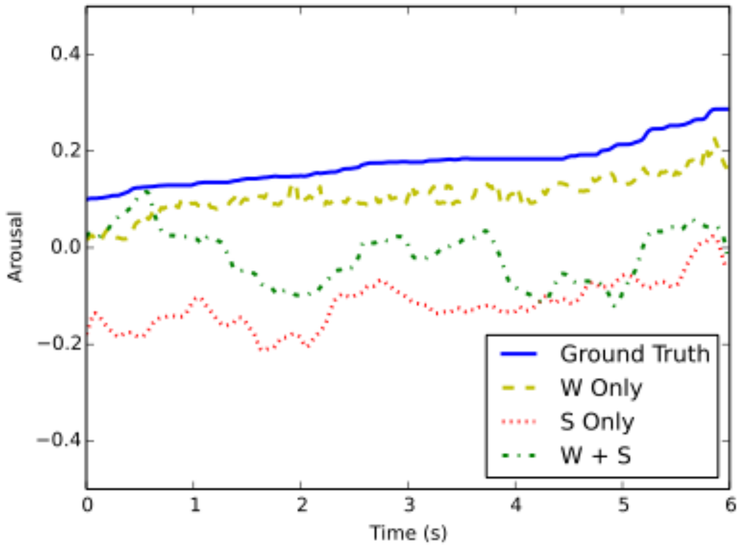
- Input: **raw waveform and spectrogram**
- Model: convolutional recurrent neural networks
- Task: Predict arousal and valence
  - Continuous in both time and value
- Results:

Corpus	Model	Results (CCC)	
		Arousal	Valence
SEMAINE	Baseline	0.376	0.177
	W Only	*0.675*	0.435
	S Only	0.656	*0.494*
	W + S	**0.680**	**0.506**
RECOLA	Baseline	0.317	0.162
	W Only	*0.674*	0.361
	S Only	0.651	*0.408*
	W + S	**0.692**	**0.423**

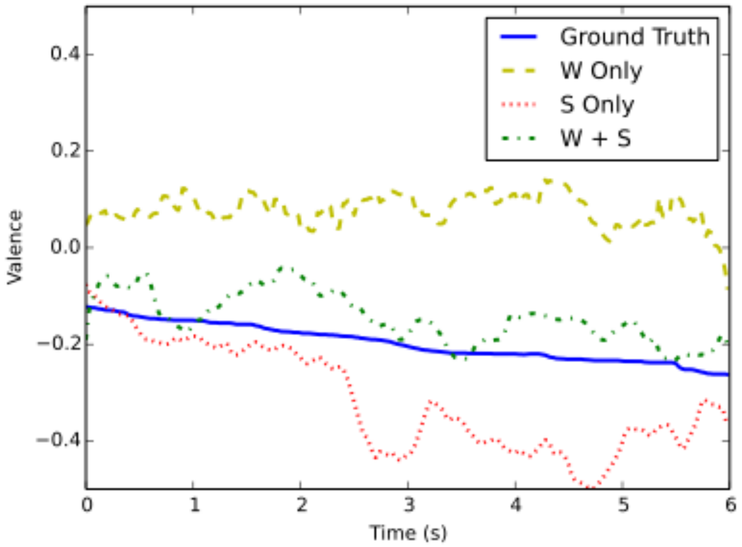


Example Analysis - Dimensional

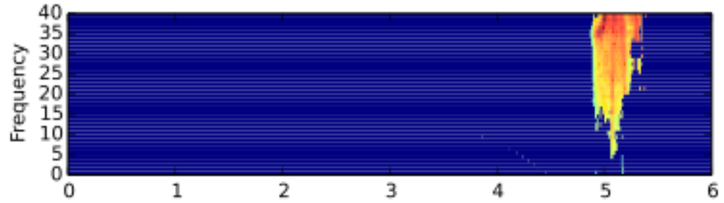
Arousal



Valence



"…cos she's so frigging superior"



Local Interpretable Modelagnostic Explanations (LIME)

Outline

- Emotion in speech
  - Emotion theory
  - Emotional speech corpora
  - Features for emotional speech
  - Models for emotion recognition
  - Expressive synthetic speech
- Sentiment and emotion in text

Expressive Synthetic Speech

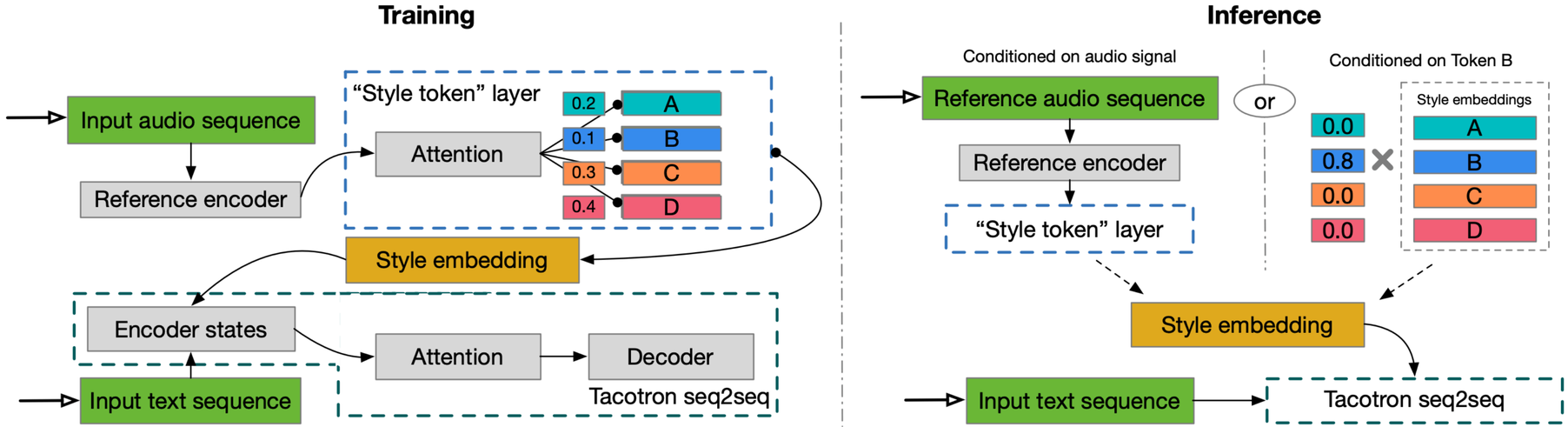
- http://emosamples.syntheticspeech.de/
- Murtaza Bulut (2005): neutral anger
- Greg Beller (2005): neutral sad

- Acapela (2013): neutral happy sad
  bad guy from afar

- Laughters: Greg Beller(2009)
  Shiva Sundaram(2007)

Expressive Synthetic Speech

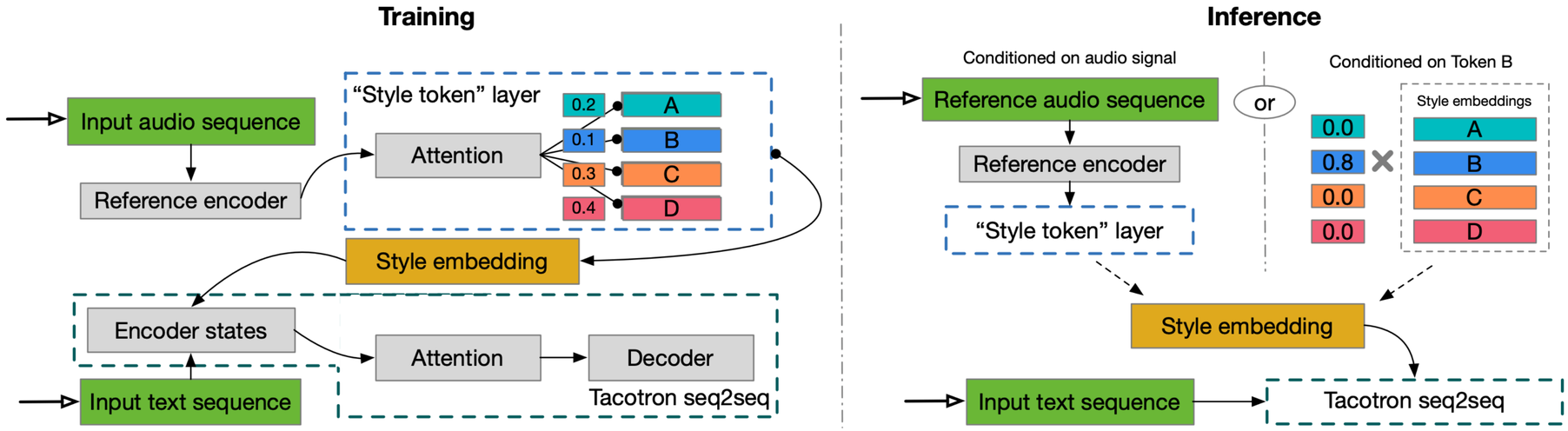
- Tacotron (/täkō ˌträn/): An end-to-end speech synthesis system by Google
- Tacotron + Style Tokens (Wang et al. 2018)



- *"United Airlines five six three from Los Angeles to New Orleans has Landed."*
  - *5 different "styles":*

Expressive Synthetic Speech

- Tacotron (/täkō ˌträn/): An end-to-end speech synthesis system by Google
- Tacotron + Style Tokens (Wang et al. 2018)



- *"Here you go, a link for Biondo Racing Products and other related pages."*
  - *Style token A (-0.3/0.1/0.3/0.5):*

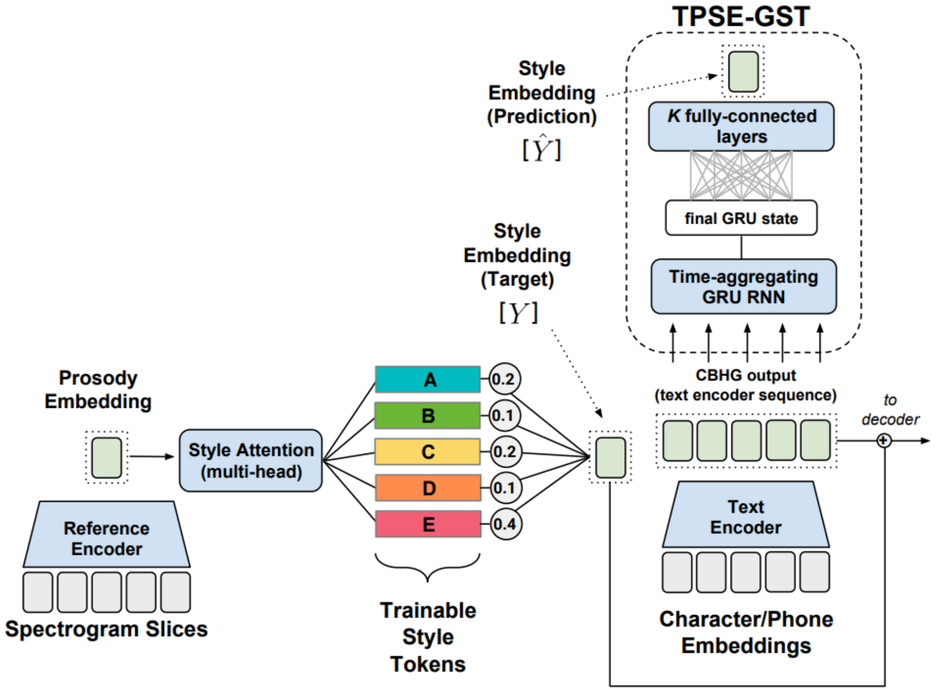
Expressive Synthetic Speech

- Tacotron (/täkō ˌträn/): An end-to-end speech synthesis system by Google
- Tacotron + predicting Style Tokens from text (Stanton et al. 2018)

"Thirty-six," he said, looking up at his mother and father. "That's two less than last year." "Darling, you haven't counted Auntie Marge's present, see, it's here under this big one from Mommy and Daddy."

Tacotron

Tacotron + predicted style token



Emotional Voice Conversion

(Zhou et al., 2020)

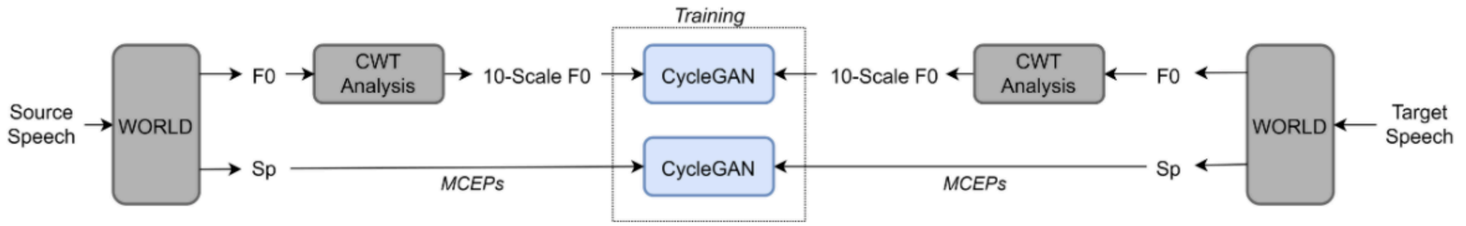


Fig.1 The training phase of the proposed CycleGAN-based emotional VC framework, where WORLD acts as the vocoder. CWT is used to decompose F0 into 10 scales. Blue boxes represent the training stage of the network, while grey boxes represent the blocks which do not need the training stage.

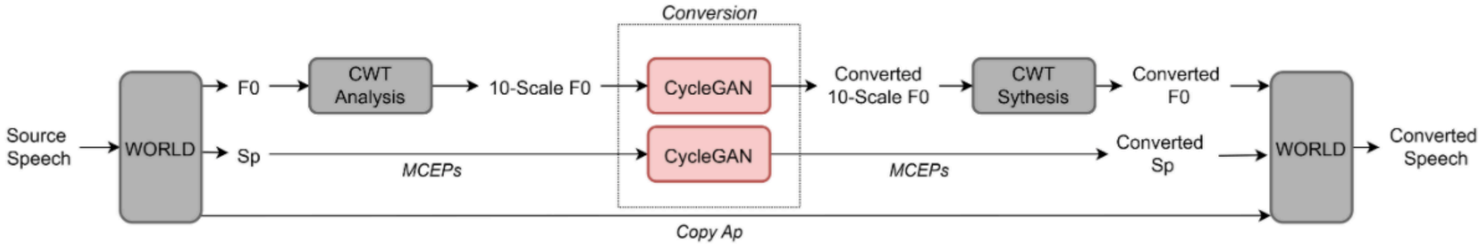


Fig.2 The run-time conversion phase of the proposed CycleGAN-based emotional VC framework. Pink boxes represent the network which are already trained.

Neutral Anger Converted

Neutral Sad Converted

Sentiment and Emotion in Text

English Sentiment Lexicon

- The General Inquirer (Stone et al. 1966)
  - Positive (1915), Negative (2291), Strong vs Weak, Pleasure, Pain, etc.
- LIWC (Linguistic Inquiry and Word Count)
  - Negative emotion (anxiety, anger, sadness); Positive emotion
- MPQA Subjectivity Cues Lexicon
  - 2718 positive, 4912 negative
- Bing Liu Opinion Lexicon
  - 2006 positive, 4783 negative
- SentiWordNet
  - WordNet synsets automatically labeled with positivity, negativity, and objectiveness

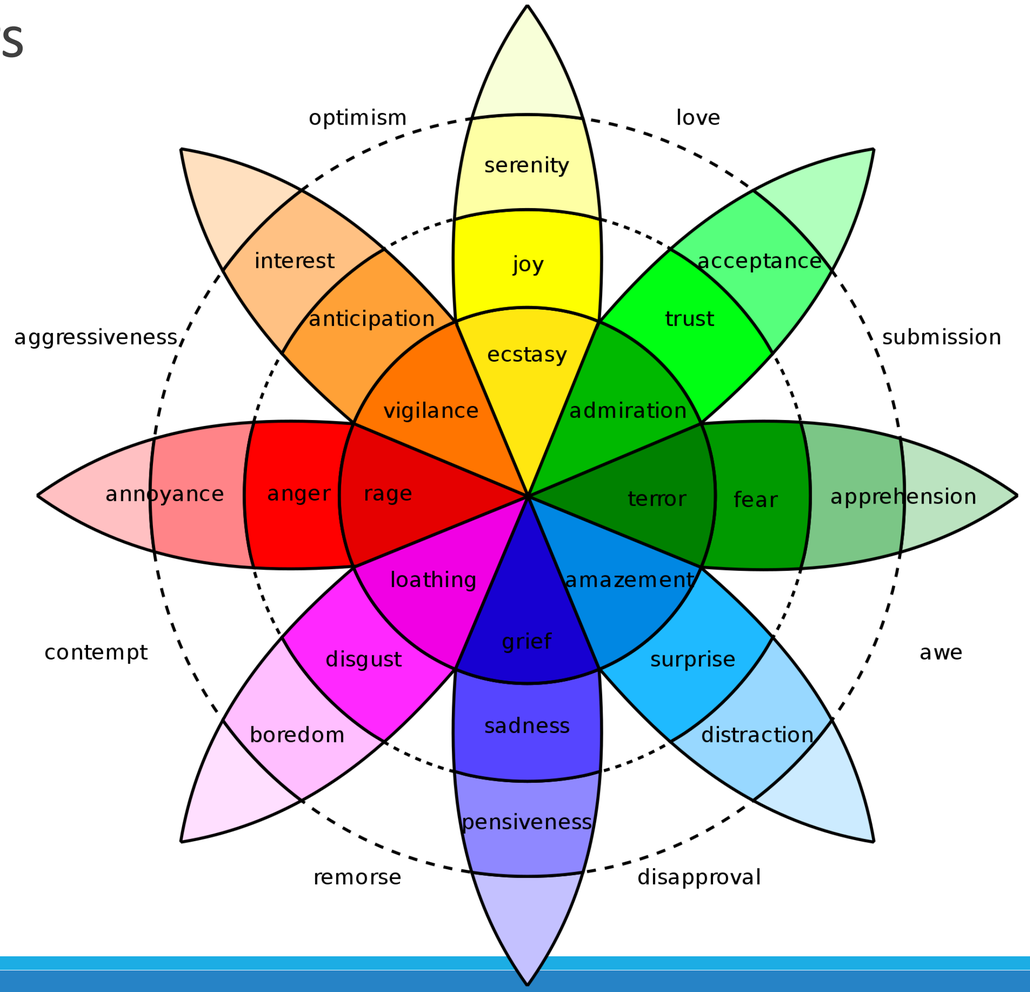
Polyglot (Multilingual text processing toolkit)

- Sentiment polarity lexicons for 136 languages
  - 7,741,544 high-frequency words from 136 languages in Wikipedia
  - Use Bing Liu Opinion Lexicon (English) as seed
  - Wiktionary + Google Translation + Transliteration + WordNet to generate edges between words
  - Propagate sentiment labels through the edges

1. Turkmen	2. Thai	3. Latvian
4. Zazaki	5. Tagalog	6. Tamil
7. Tajik	8. Telugu	9. Luxembourgish, Letzeb...
10. Alemannic	11. Latin	12. Turkish
13. Limburgish, Limburgan...	14. Egyptian Arabic	15. Tatar
16. Lithuanian	17. Spanish; Castilian	18. Basque
19. Estonian	20. Asturian	21. Greek, Modern
22. Esperanto	23. English	24. Ukrainian
25. Marathi (Marāṭhī)	26. Maltese	27. Burmese
28. Kapampangan	29. Uighur, Uyghur	30. Uzbek
31. Malagasy	32. Yiddish	33. Macedonian
34. Urdu	35. Malayalam	36. Mongolian
37. Breton	38. Bosnian	39. Bengali

Emotion Theory: Plutchik's wheel of emotion

- 8 basic emotions in four opposing pairs
  - joy–sadness
  - anger–fear
  - trust–disgust
  - anticipation–surprise



NRC Word-Emotion Association Lexicon

(Mohammad and Turney 2011)

- Categorical approach of emotion
- 10k words chosen mainly from earlier lexicons
- Labeled by Amazon Mechanical Turk
  - Joy, sadness, anger, fear, trust, disgust, anticipation, surprise; positive, negative

Q4. How much is *startle* associated with the emotion joy? (For example, *happy* and *fun* are strongly associated with joy.)

- *startle* is not associated with joy
- *startle* is weakly associated with joy
- *startle* is moderately associated with joy
- *startle* is strongly associated with joy

EmoLex	# of terms	% of the Union
EmoLex-Uni:		
Unigrams from Macquarie Thesaurus		
adjectives	200	2.0%
adverbs	200	2.0%
nouns	200	2.0%
verbs	200	2.0%
EmoLex-Bi:		
Bigrams from Macquarie Thesaurus		
adjectives	200	2.0%
adverbs	187	1.8%
nouns	200	2.0%
verbs	200	2.0%
EmoLex-GI:		
Terms from General Inquirer		
negative terms	2119	20.8%
neutral terms	4226	41.6%
positive terms	1787	17.6%
EmoLex-WAL:		
Terms from WordNet Affect Lexicon		
anger terms	165	1.6%
disgust terms	37	0.4%
fear terms	100	1.0%
joy terms	165	1.6%
sadness terms	120	1.2%
surprise terms	53	0.5%
Union	**10170**	**100%**

Lexicon of Valence, Arousal, and Dominance

(Warriner at al. 2013)

- Dimensional approach of emotion
- AMT Ratings for 14,000 words for emotional dimensions
  - Valence (the pleasantness of the stimulus)
  - Arousal (the intensity of emotion provoked by the stimulus)
  - Dominance (the degree of control exerted by the stimulus)
- Examples: (range 1-9)

Valence		Arousal		Dominance	
vacation	8.53	rampage	7.56	self	7.74
happy	8.47	tornado	7.45	incredible	7.74
whistle	5.7	zucchini	4.18	skillet	5.33
conscious	5.53	dressy	4.15	concur	5.29
torture	1.4	dull	1.67	earthquake	2.14

Detecting Sentiment/Emotion in Text

- Simplest unsupervised method
  - Sum the weights of each positive word in the document
  - Sum the weights of each negative word in the document
  - Choose whichever value (positive or negative) has higher sum
- Simplest supervised method
  - Use “counts of lexicon categories” as features (e.g. LIWC)
  - Baseline: use all unigram/bigram counts + POS tags
  - Hard to beat, but only works if the training and test sets are very similar

Sentiment in Twitter :) (Go et al. 2009)

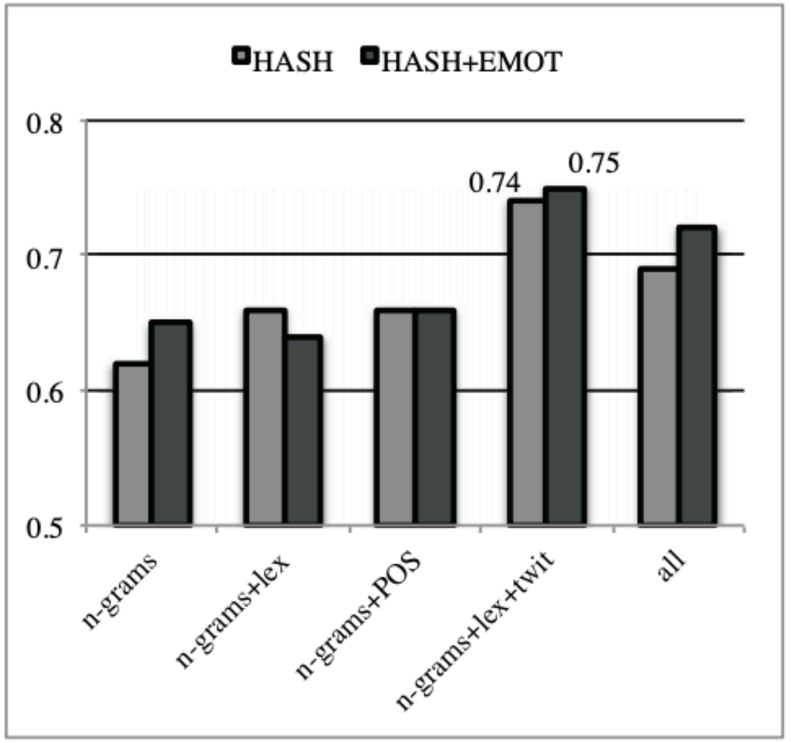
- Use emoticons to find tweets with sentiment

Emoticons mapped to :)	Emoticons mapped to :(
:)	:(
:-)	:-(
:)	: (
:D	
=)	

- Training set:
  - 800k tweets with positive emoticons, and 800k tweets with negative emoticons
  - Seed emoticons are stripped off before training
- Test set: 359 tweets manually annotated
- Accuracy: ~80%

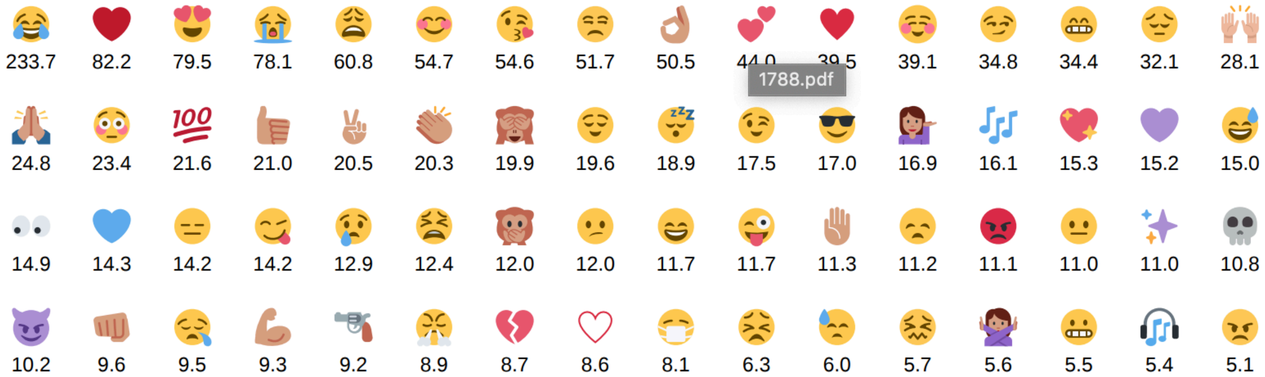
Sentiment in Twitter #thingsilike (Kouloumpis et al. 2011)

Positive	#iloveitwhen, #thingsilike, #bestfeeling, #bestfeelingever, #omgthatssotrue, #imthankfulfor, #thingsilove, #success
Negative	#fail, #epicfail, #nevertrust, #worst, #worse, #worstlies, #imtiredof, #itsnotokay, #worstfeeling, #notcute, #somethingaintright, #somethingsnotright, #ihate
Neutral	#job, #tweetajob, #omgfacts, #news, #listeningto, #lastfm, #hiring, #cnn

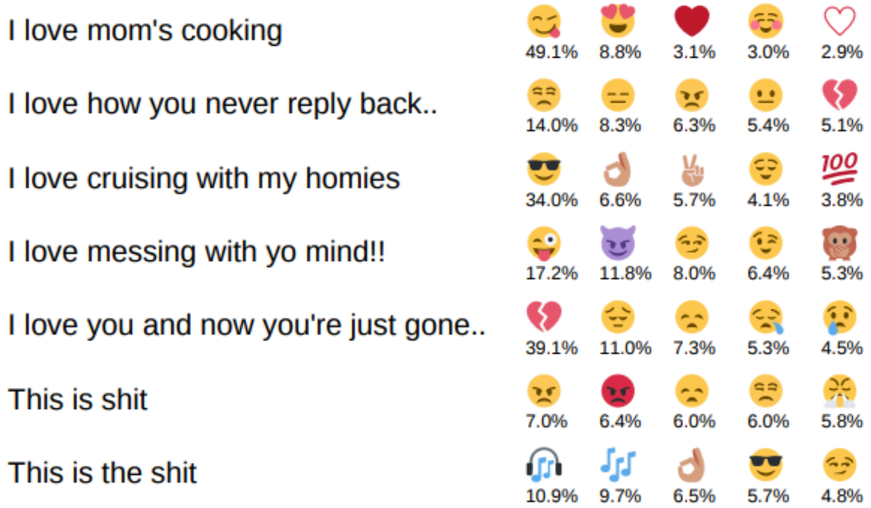


Emoji in Twitter (Felbo et al. 2017)

- Number of training data (in *millions*)



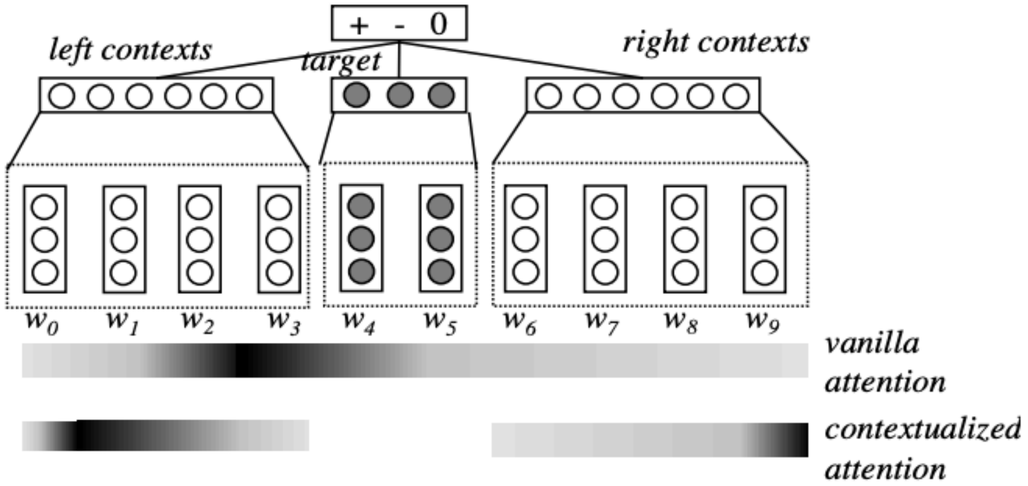
- Output: probability of emoji labels

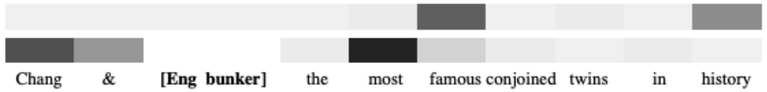


Attention Modeling for Targeted Sentiment

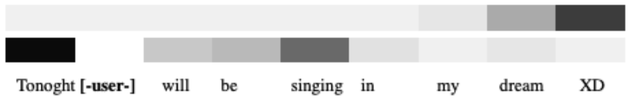
(Liu and Zhang 2017)

- Targeted Sentiment
  - ✔ “She began to love **miley ray cyrus** since 2013 :)”
  - ✘ “#nowplaying **lady gaga** - let love down”

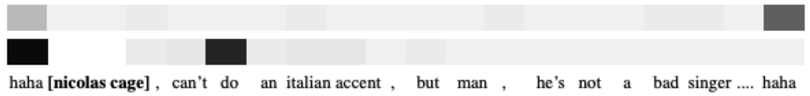




(a) Positive



(b) Positive

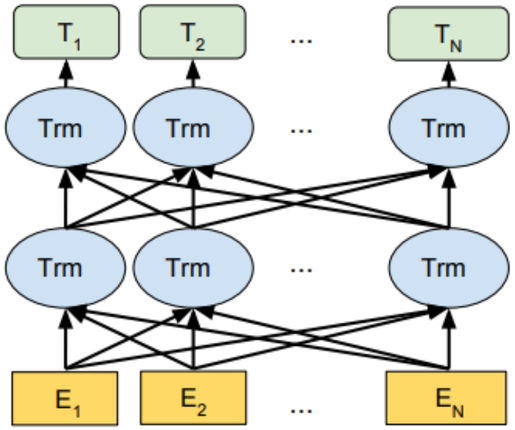


(c) Neutral

(d) Negative

BERT in Sentiment Analysis (Google AI Language)

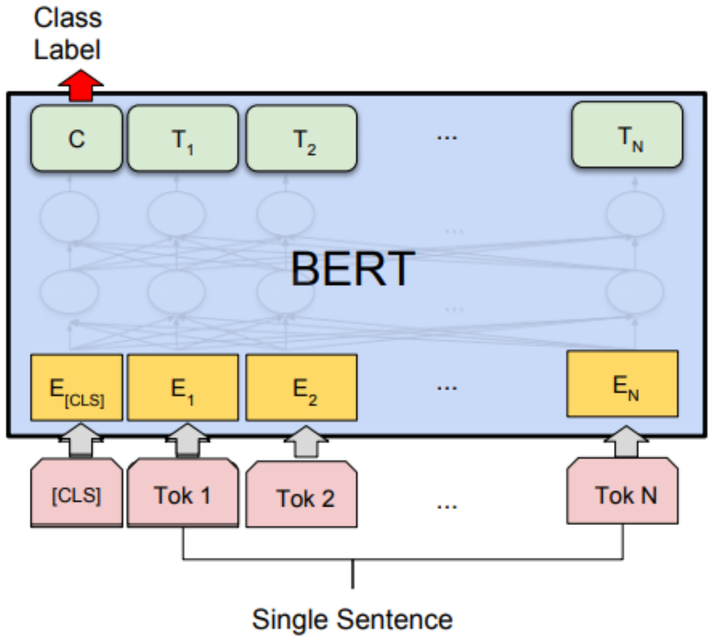
- BERT: Bidirectional Encoder Representations from Transformers
  - Transformer: stacked self-attention blocks



- Training: mask part of the input tokens at random, then predict those masked tokens

BERT in Sentiment Analysis (Google AI Language)

- Fine-tuning for single sentence classification task
  - Add a classification layer on the output of [CLS] token



- Accuracy on the Stanford Sentiment Treebank dataset: 94.9%

Text Sentiment Analysis Dataset

- Product reviews on Amazon
  - Multidomain sentiment analysis dataset
  - Amazon product data, 143 million reviews
- Movie reviews on IMDB
  - Cornell movie review data, labeled with sentiment polarity, scale, and subjectivity
  - Large Movie Review Dataset v1.0, 25k movie reviews
  - IMDB Movie Reviews Dataset, 50k movie reviews
  - Bag of Words Meets Bags of Popcorn, 50k movie reviews
- Reviews from Rotten Tomatoes
  - Stanford Sentiment Treebank, 11k reviews

Text Sentiment Analysis Dataset

- Tweets with emoticon
  - Sentiment140, 160k tweets
- Twitter data on US airlines
  - Twitter US Airline Sentiment, with negative reasons (e.g. “rude service”)
- Paper reviews
  - Paper Reviews

Thank you!

Questions?

brenda@cs.columbia.edu