# Emotion, Sentiment, and Keyword Search

Zixiaofan Yang and Julia Hirschberg

COMS 6998

April 19, 2019

# Outline

- Emotion recognition in speech
- Sentiment and emotion in text
- Situation Frame (SF) detection
- Homework 4: emotion recognition

# Emotion Recognition in Speech

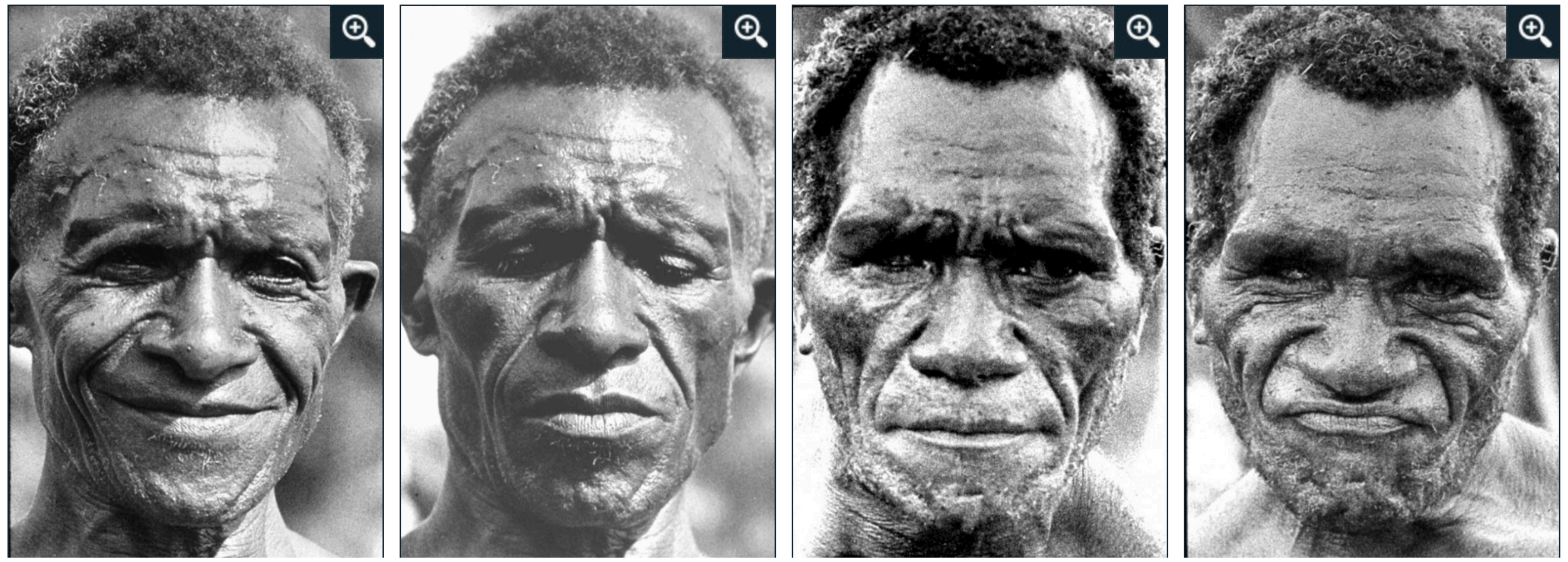
# What is Emotion?

- 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



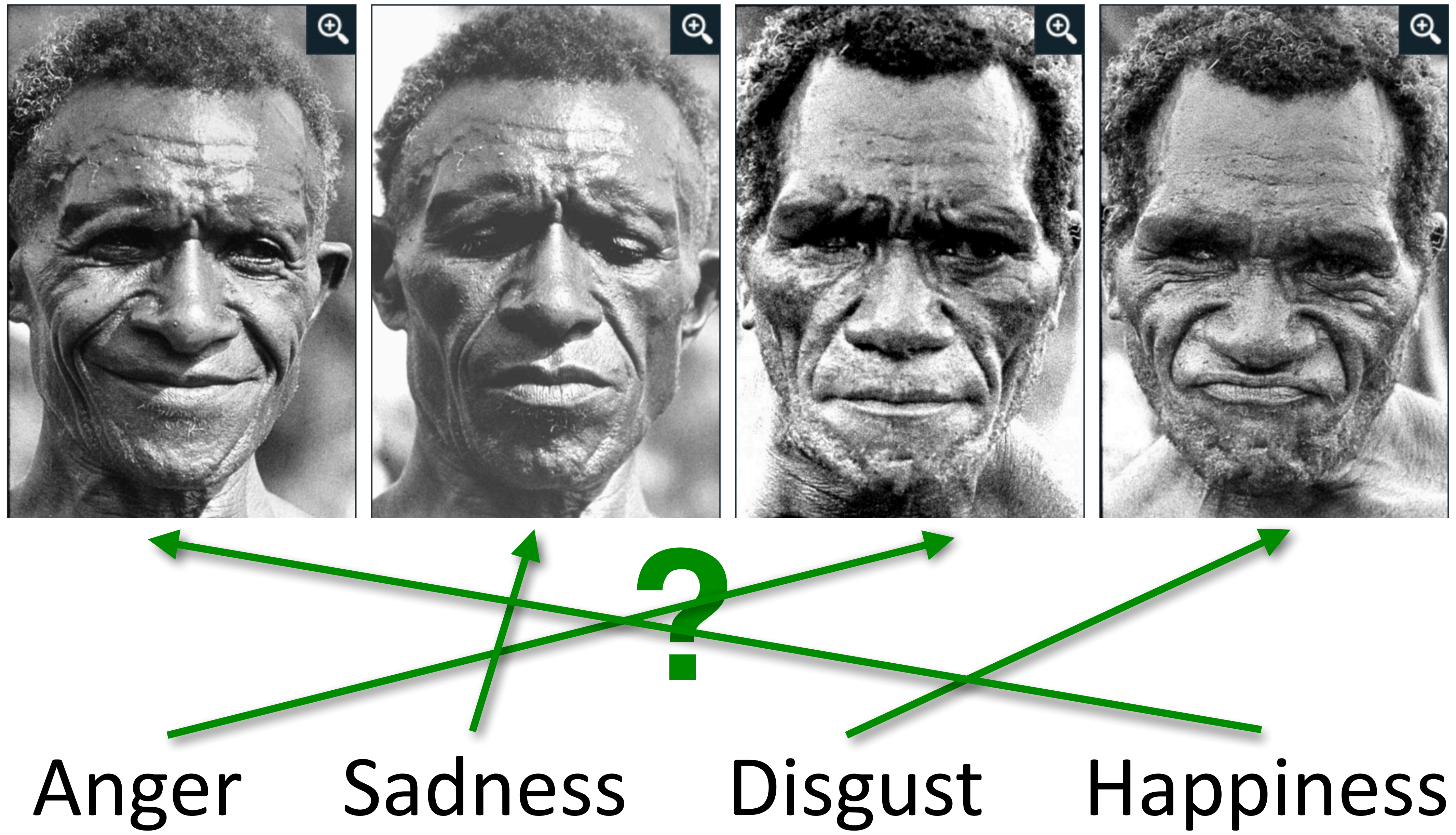
?

Anger Sadness Disgust Happiness

# Emotion - Categorical Approach

[Ekman et al., 1987]

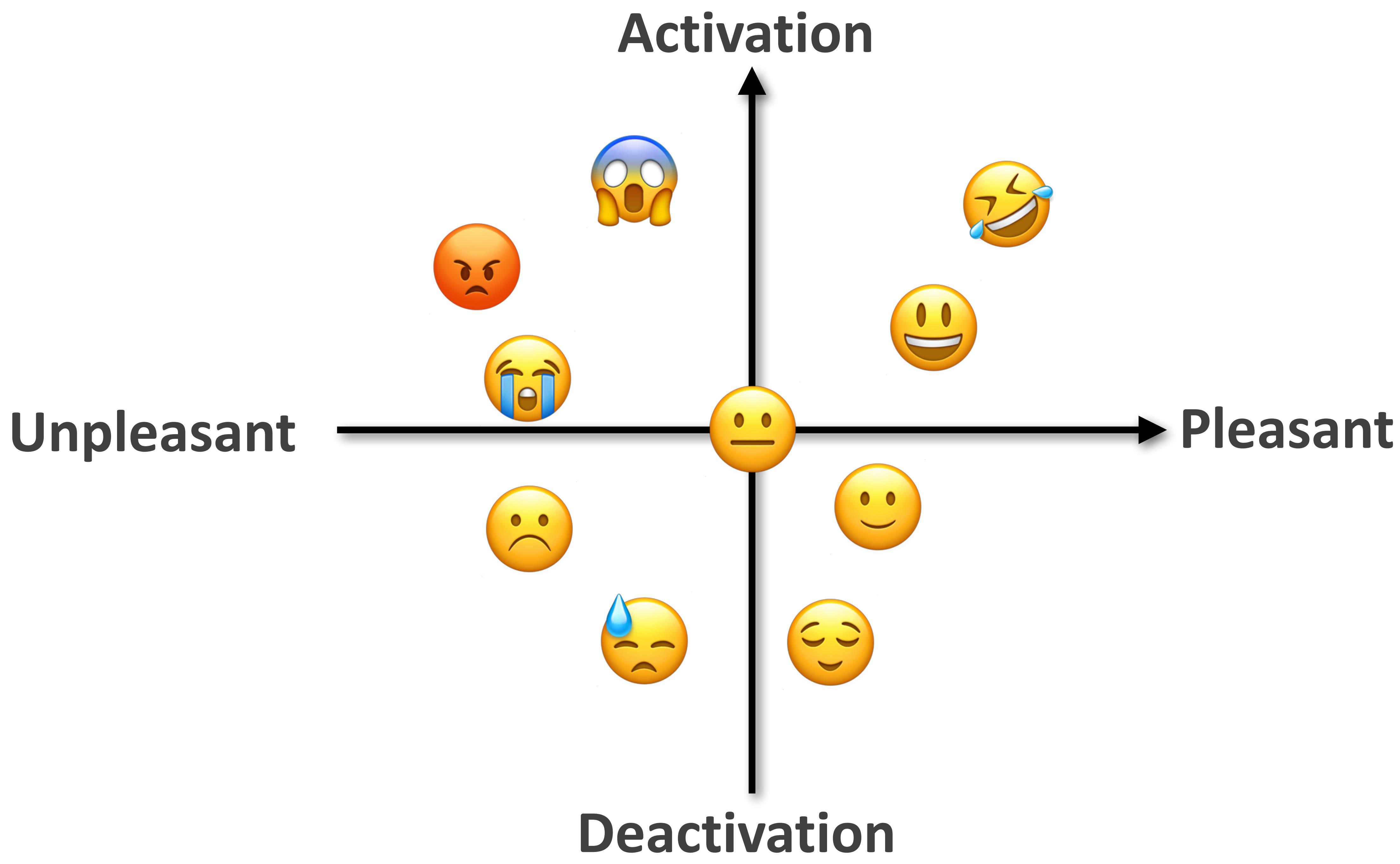
- Discrete ‘basic emotions’
- Originate from facial expressions



# Emotion - Dimensional Approach

[Russell and Barrett, 1999]

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



# Why Study Emotional Speech?

- Recognition
  - 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
  - Deception, Charisma, Sleepiness, Interest…

- Some emotional clues are only in speech

# Emotion in Speech

**Acted speech**

✓ Easier to collect & control

✗ Extreme emotions

- Mostly categorical approach
- Examples: (Emotional Prosody Speech)
  - Happy, Sad, Angry, Bored

**Spontaneous speech**

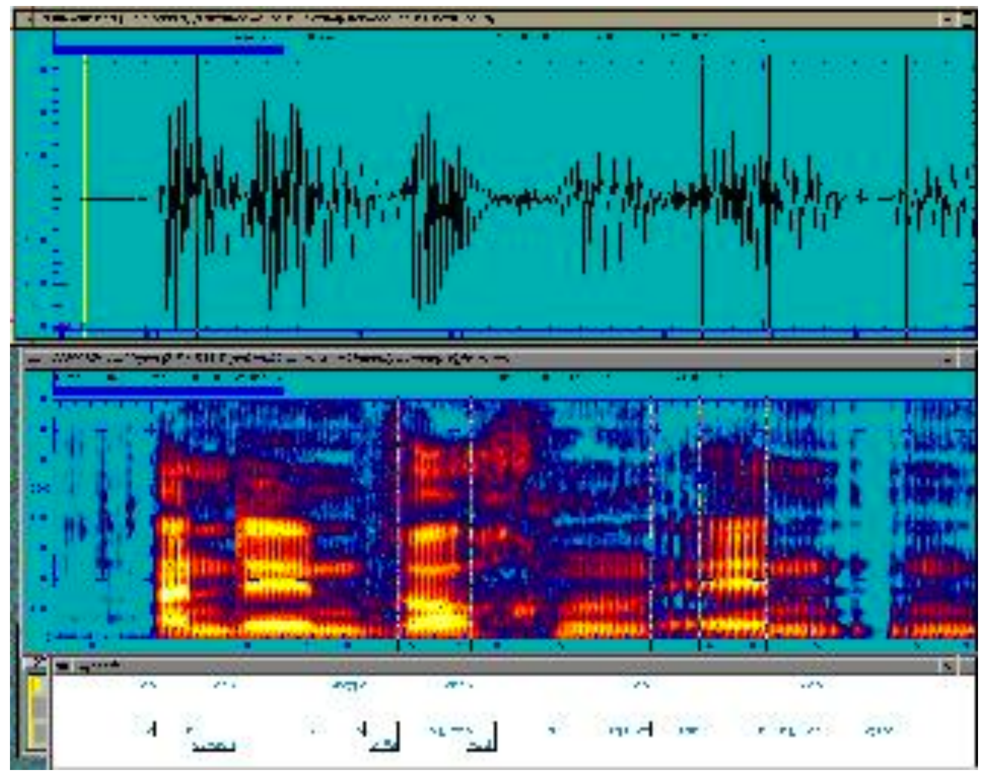
✗ Harder to collect & annotate

✓ Subtle changes in emotion

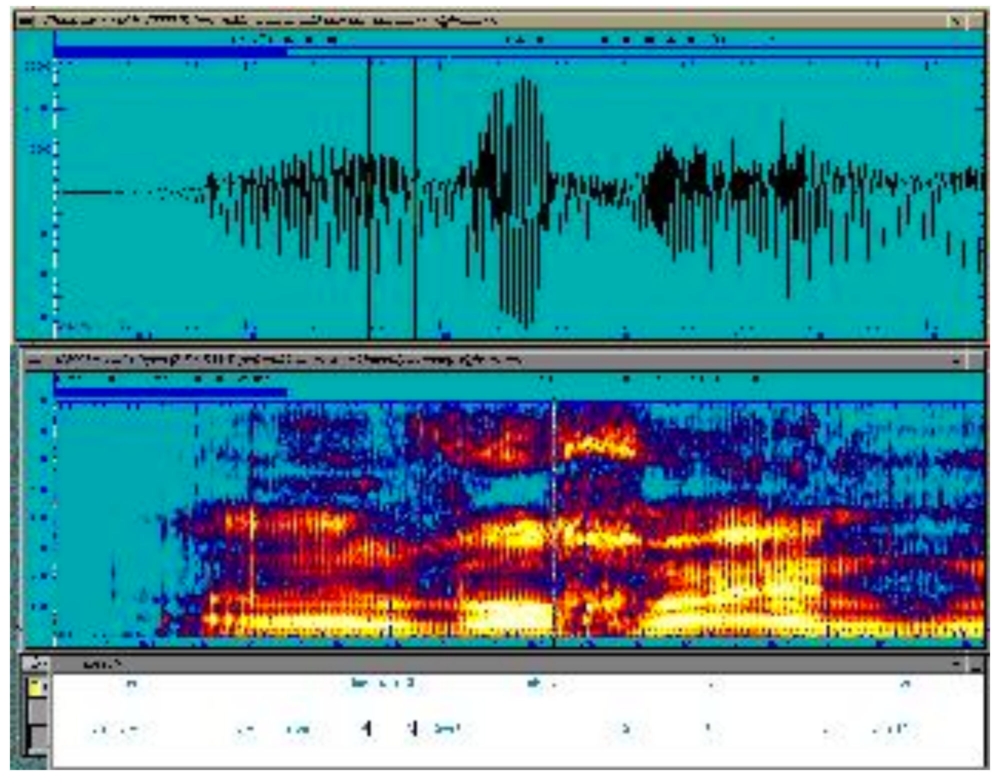
- Both categorical & dimensional approach
- Example: (AT&T "How May I Help You?" System)
  - Neutral -> frustrated -> angry
  - Arousal ↑, Valence ↓

# Emotional Speech Corpora - Acted & Categorical (EmoDB)

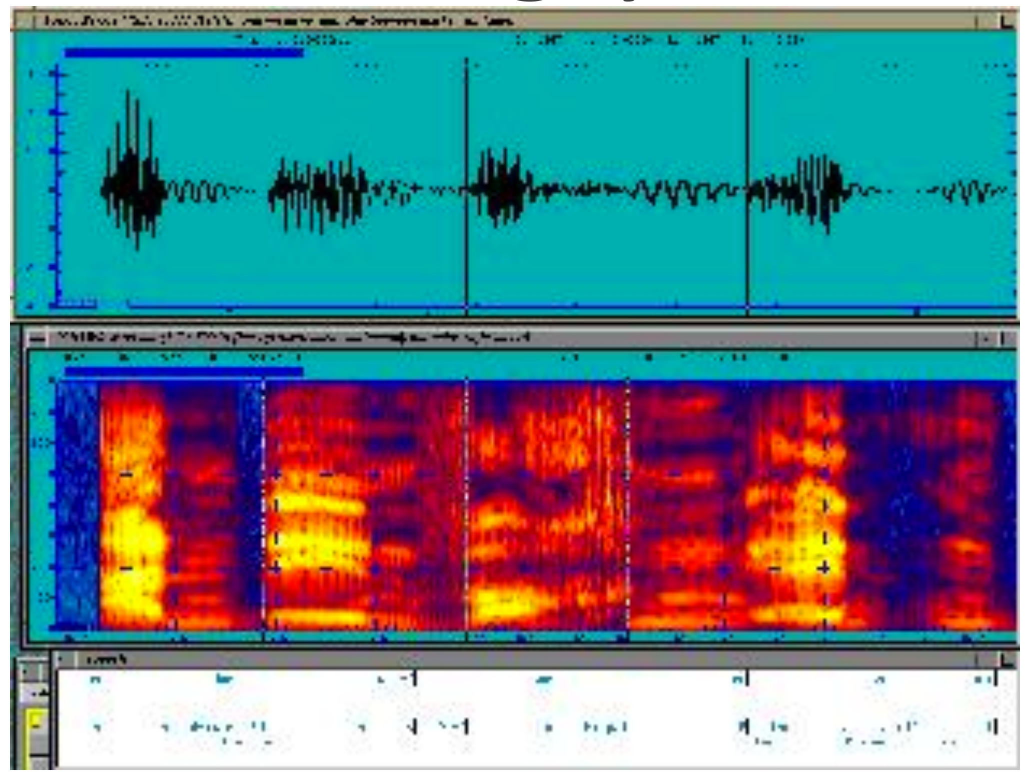
Neutral



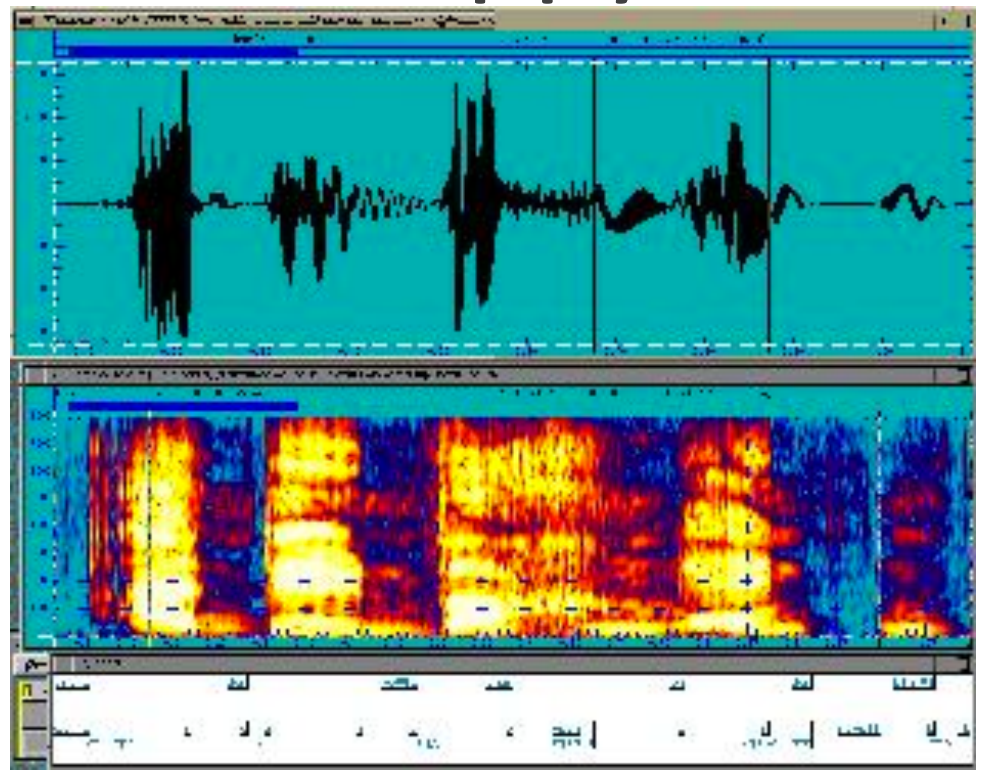
Bored



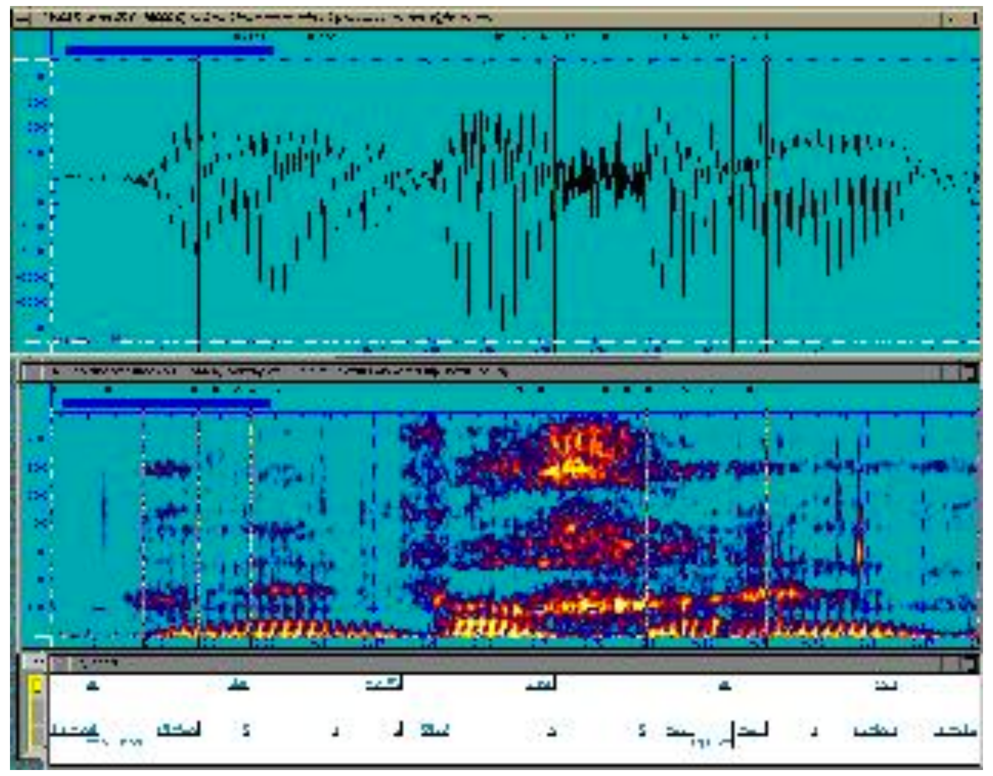
Angry



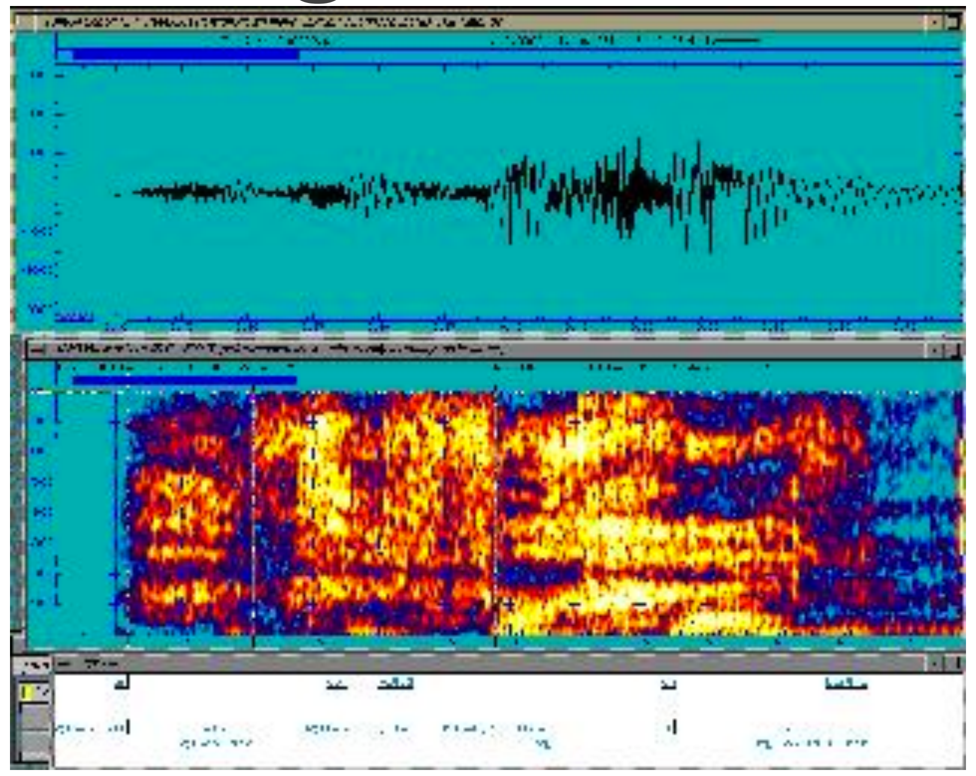
Happy



Sad



Frightened



# Acted & Categorical Speech: Actors vs Students

(Emotional Prosody Speech) (Mandarin Affective Speech)

Sad
Happy
Angry
Bored
Interested
.......

Anger
Elation
Neutral
Panic
Sadness

# Spontaneous Speech with Dimensional Annotations
## (SEMAINE database)

- The goal of the operator is to engage the user in emotional conversations
- 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.

# Partial List of the Existing Emotion Corpora

- Lack of naturalness
- Unbalanced emotional content
- Limited size of corpora, limited number of speakers

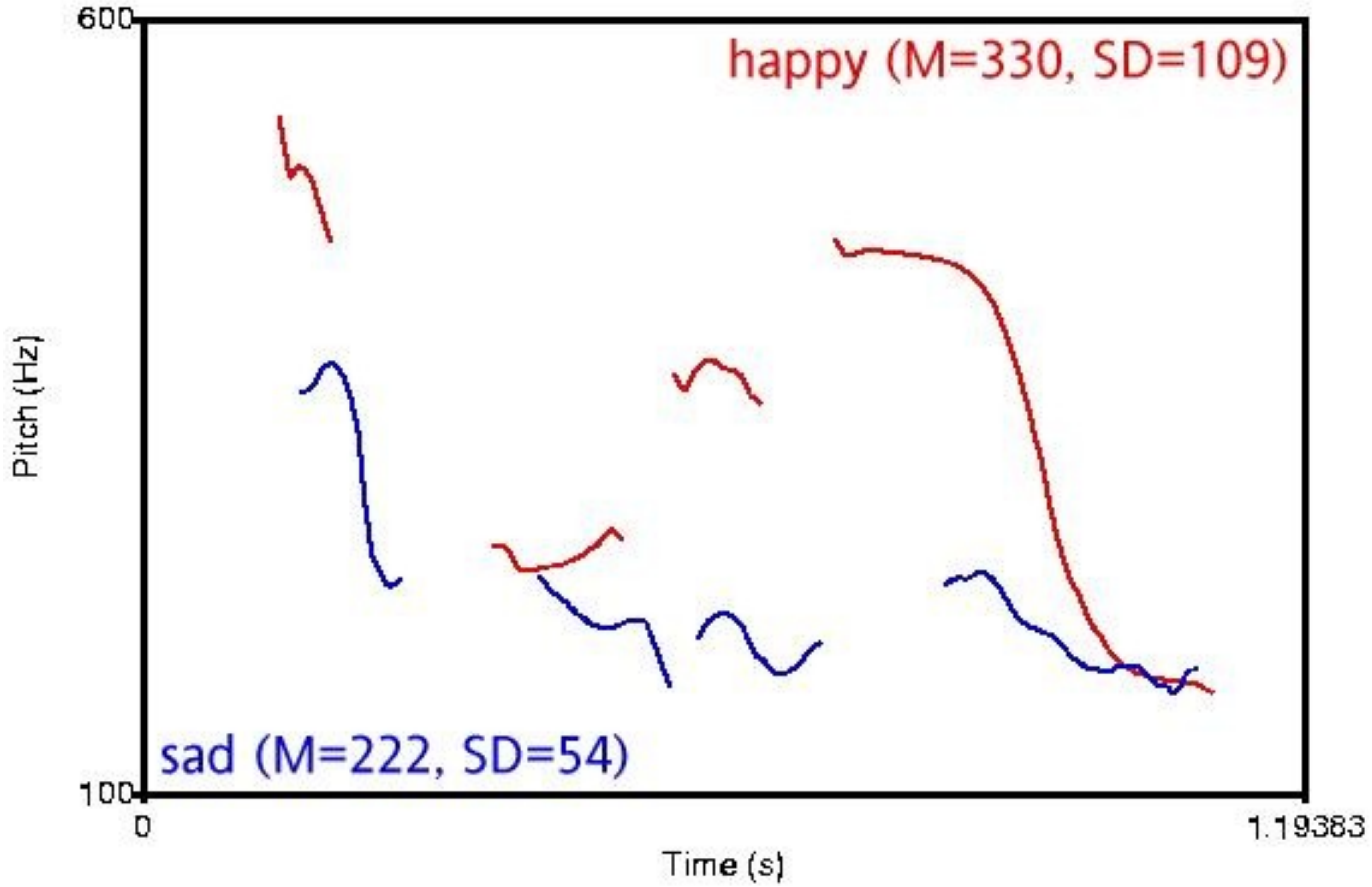
| Corpus | Size | # Spkr | Type | Lang. |
|---|---|---|---|---|
| IEMOCAP [10] | 12h26m | 10 | acted | English |
| MSP-IMPROV [19] | 9h35m | 12 | acted | English |
| CREMA-D [2] | 7,442 samples | 91 | acted | English |
| Chen Bimodal [20] | 9,900 samples | 100 | acted | English |
| Emo-DB [6] | 22m | 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 |

# MSP-Podcast corpus

- Retrieve potential segments from podcast recordings
- Annotations
  - Dimensional descriptors
    - Activation, dominance and valence
  - Categorical labels
    - Anger, happiness, sadness, disgust, surprised, fear, contempt, neutral and other
- Version 1.1 has 22,630 speaking turns (data collection is still ongoing)
- The largest speech emotional corpus in the community

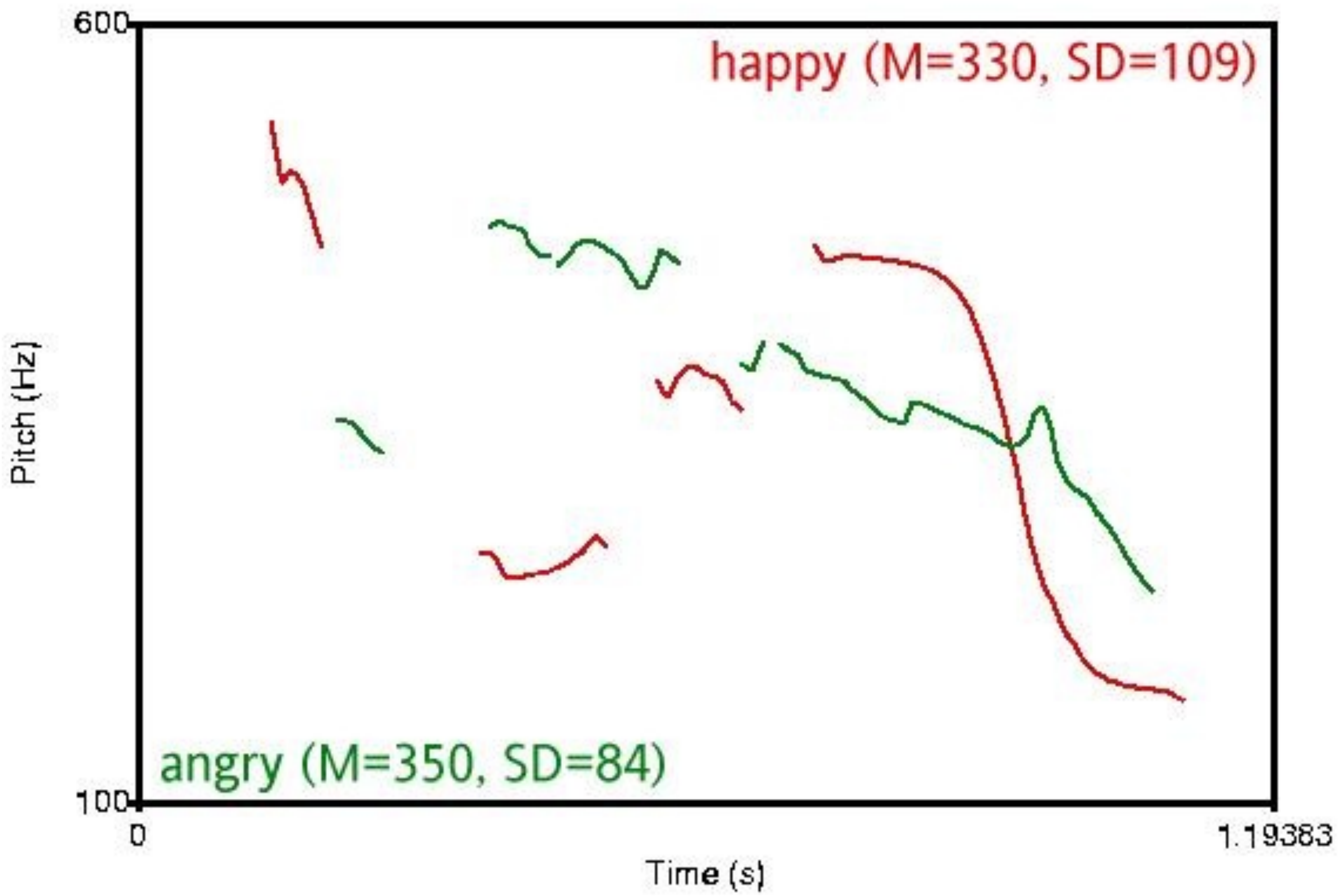
# Features for Emotional Speech - Pitch

Different Valence / Different Arousal

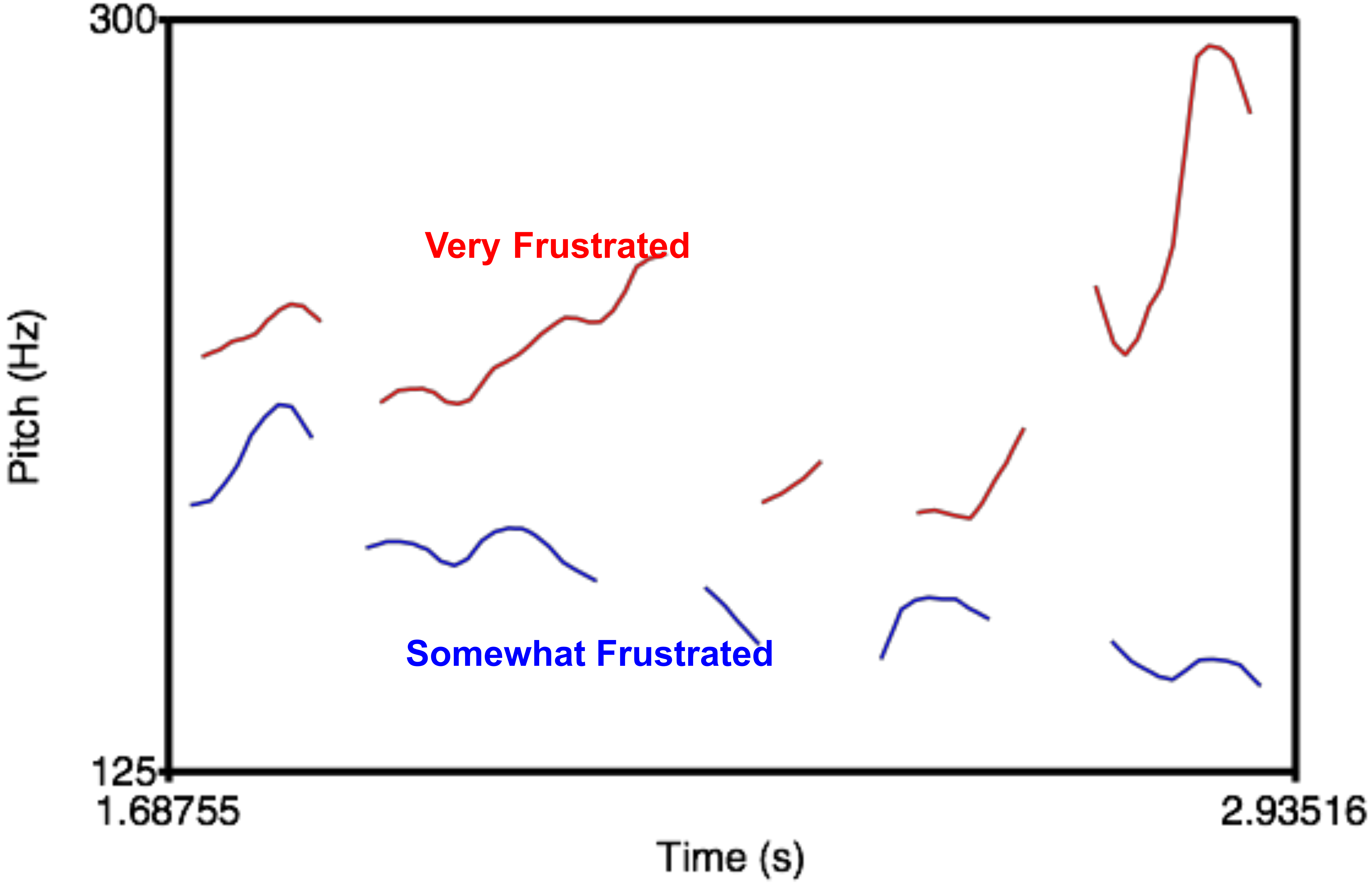


# Features for Emotional Speech - Pitch

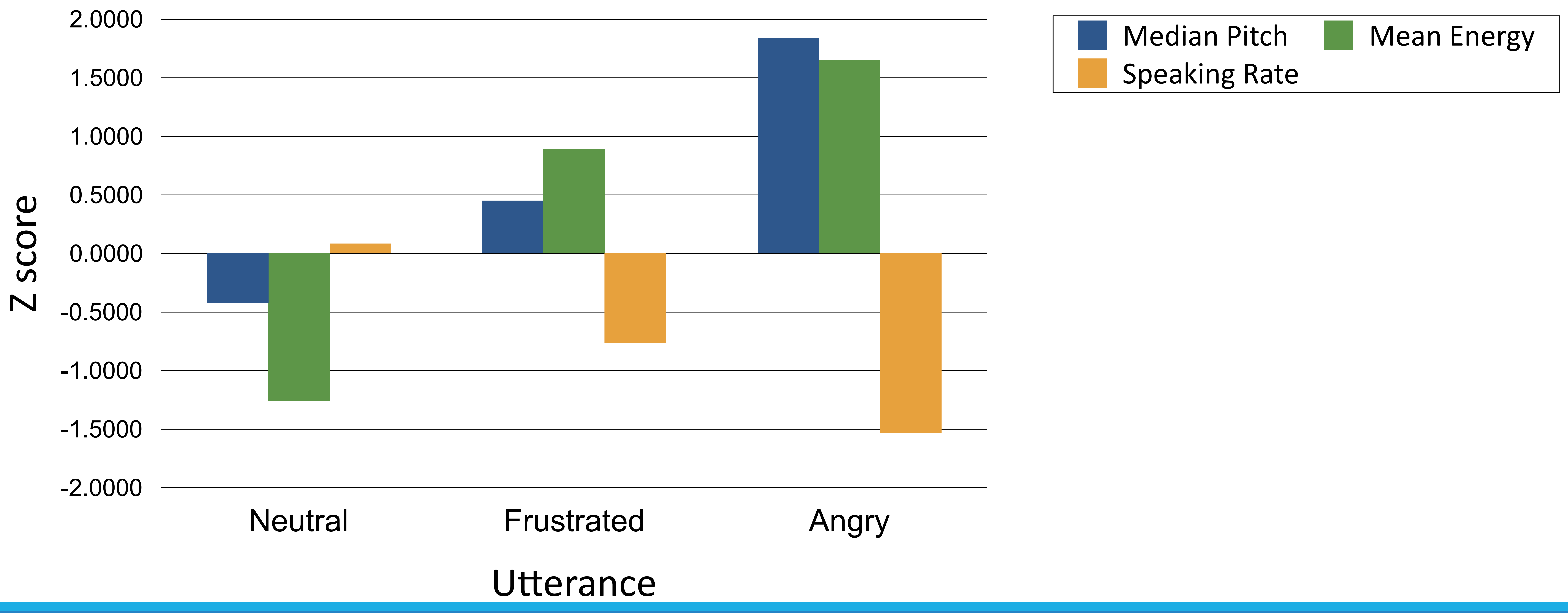
Different Valence / *Same* Arousal



# Pitch Contour Differences



# Features for Emotional Speech



# Emotion Recognition in Speech

**Categorical** Approach

- Discrete 'basic emotions'
- Classification problem

**Dimensional** Approach

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

# Emotion Recognition - Categorical

(Liscombe et al. 2003)

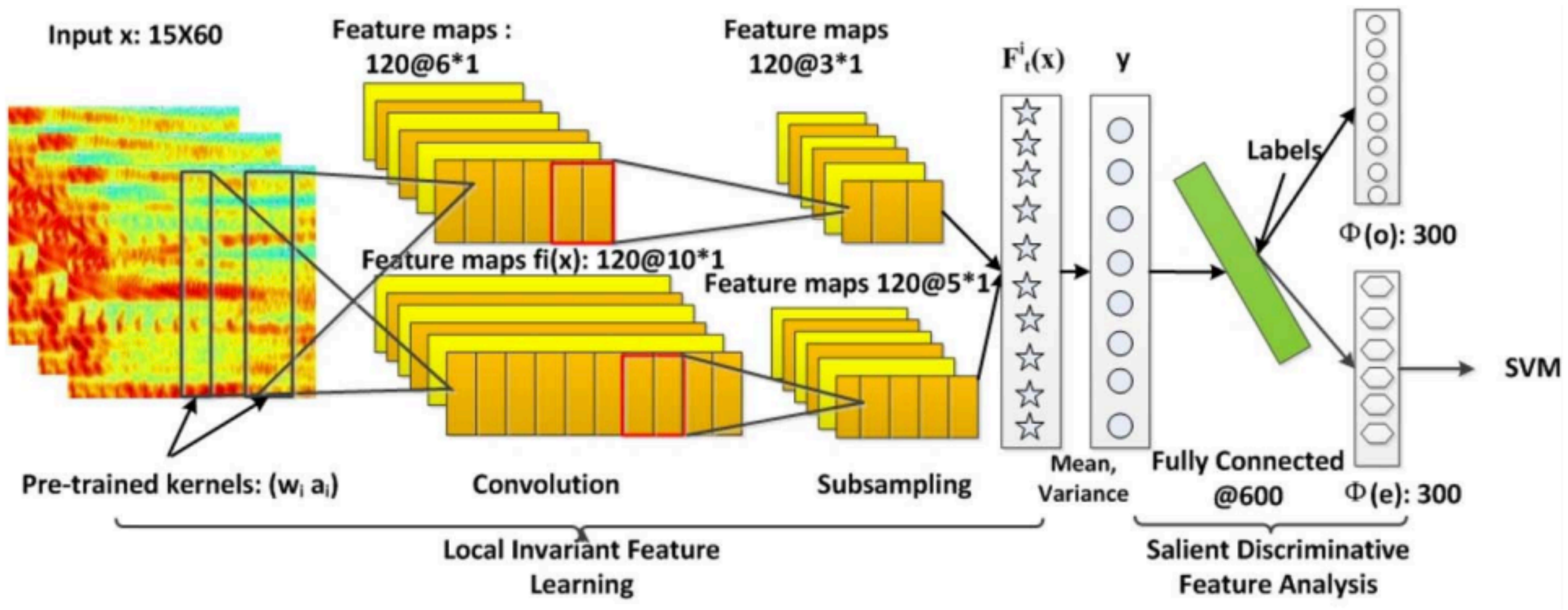
- Acoustic-prosodic features:
  - Pitch, energy, speaking rate
  - Nuclear accent, pitch contour

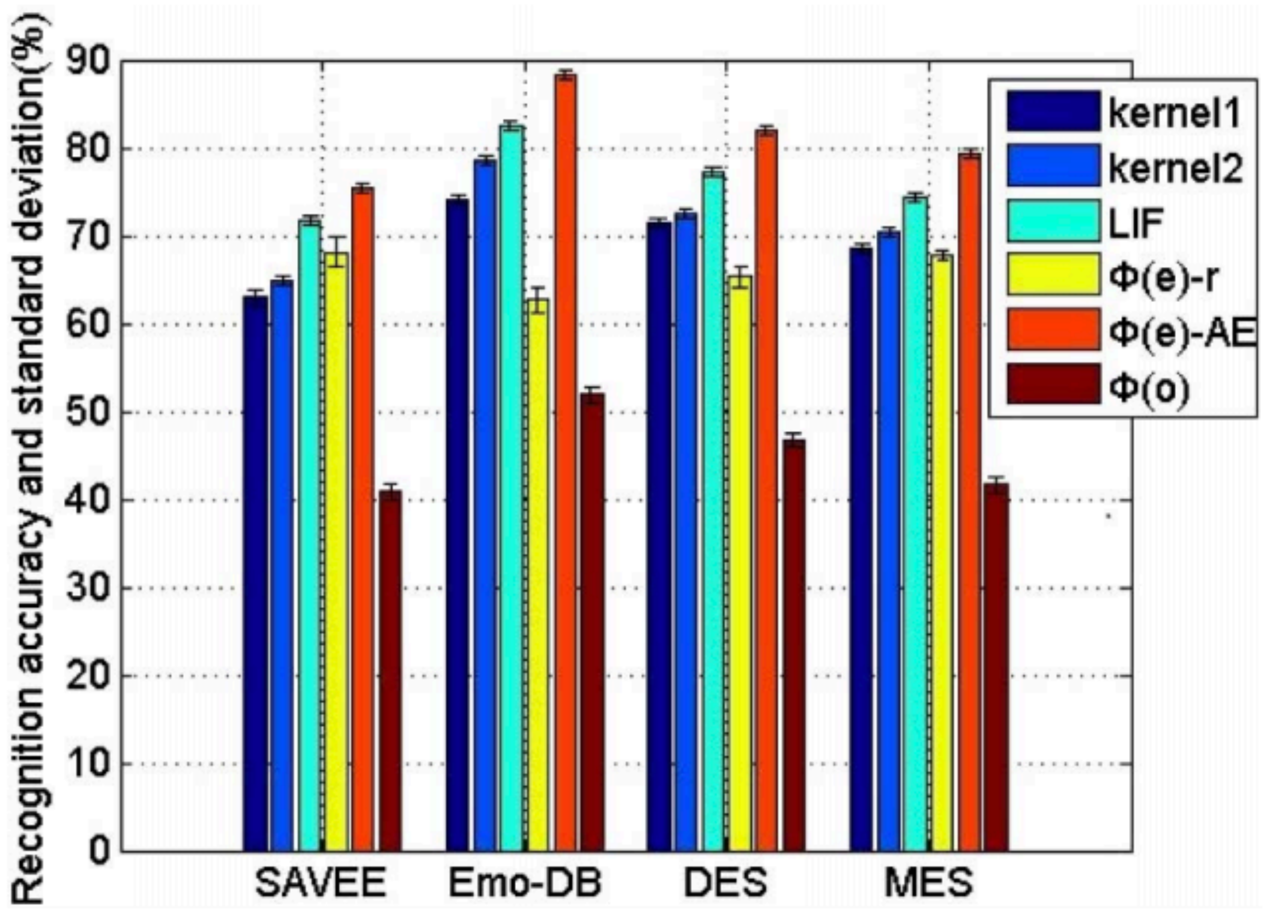
| **Emotion** | **Baseline** | **Accuracy** |
|---|---|---|
| angry | 69.32% | 77.27% |
| confident | 75.00% | 75.00% |
| happy | 57.39% | 80.11% |
| interested | 69.89% | 74.43% |
| encouraging | 52.27% | 72.73% |
| sad | 61.93% | 80.11% |
| anxious | 55.68% | 71.59% |
| bored | 66.48% | 78.98% |
| friendly | 59.09% | 73.86% |
| frustrated | 59.09% | 73.86% |

# Emotion Recognition - Categorical

(Mao et al. 2014)

- Learning emotion from 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 in Speech

**Categorical** Approach

- Discrete ‘basic emotions’
- Classification problem

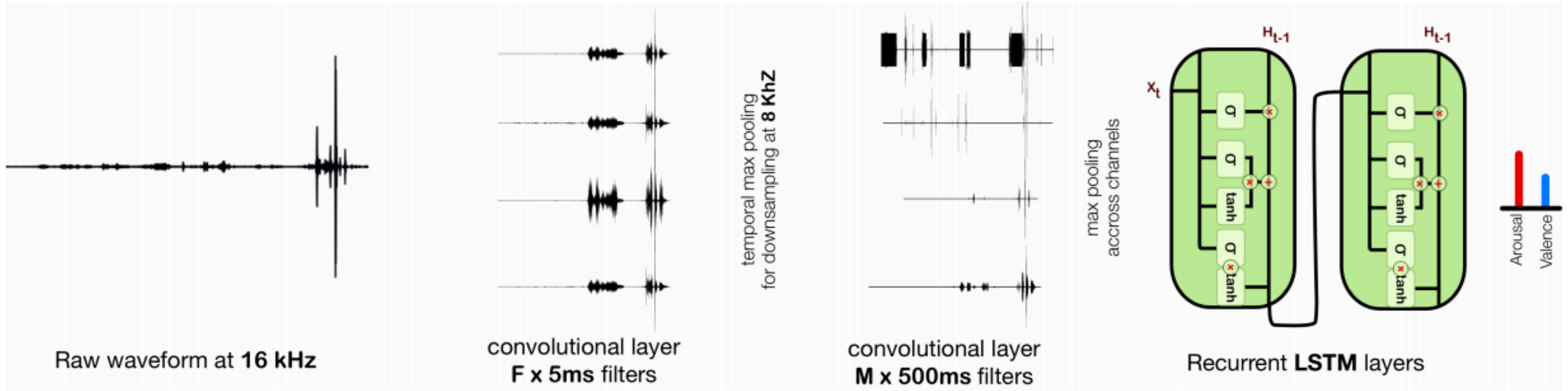
**Dimensional** Approach

- Continuous **Arousal - Valence** space
- Regression problem
  - High granularity in time and value
  - Suitable for deep learning models

# Emotion Recognition - Dimensional

(Trigeorgis et al. 2014)

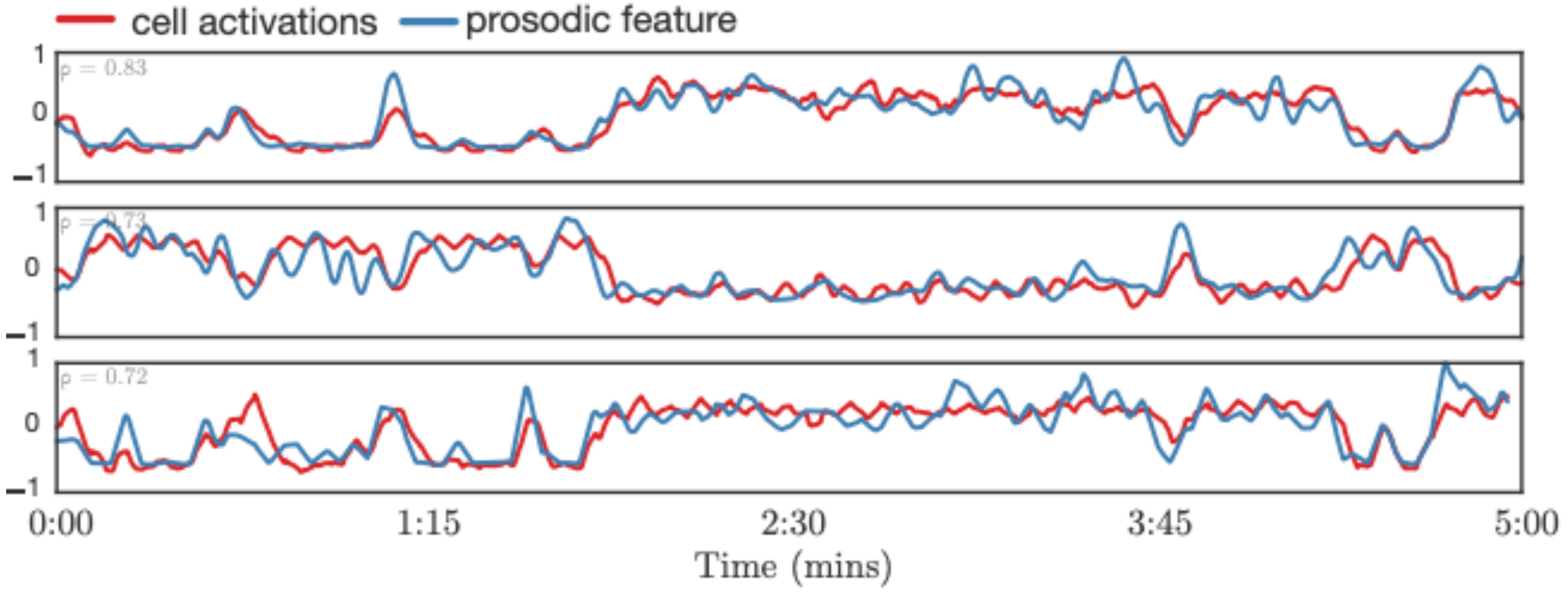
- Learning emotion (valence-arousal) from waveforms directly
- Convolutional layers:
  1. Extracting spectral information
  2. Extracting long-term characteristics
- Recurrent layers: modeling the context



# Emotion Recognition - Dimensional

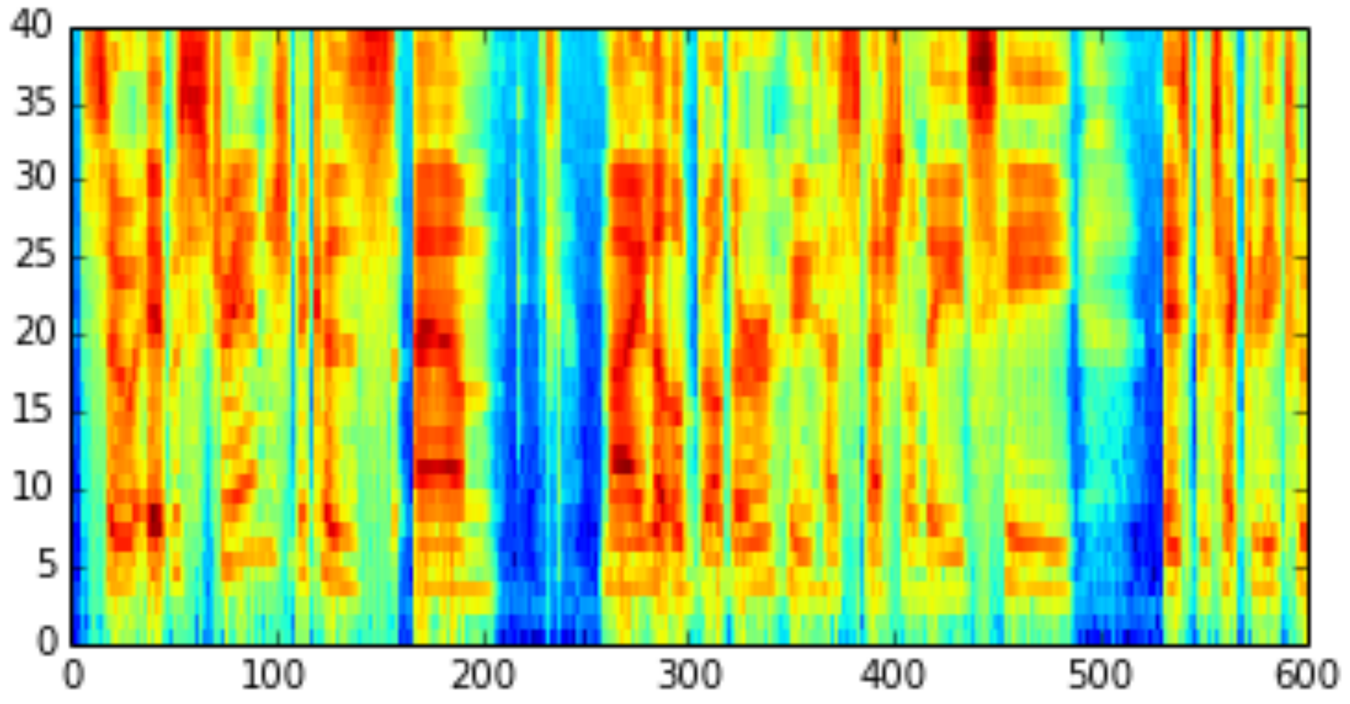
(Trigeorgis et al. 2014)

- Evaluation metric: Concordance correlation coefficient
  - Valence: 0.686, arousal: 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)

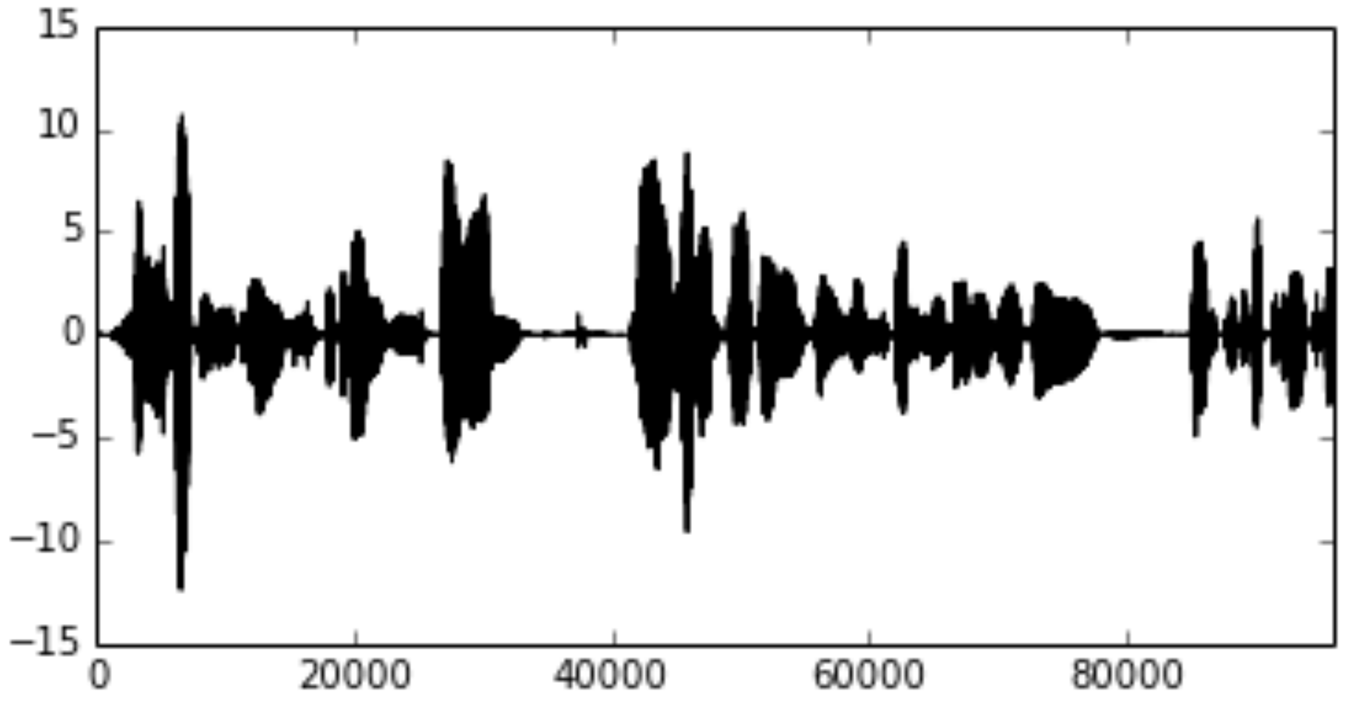


# Emotion Recognition - Dimensional

**Spectrogram**




**Waveform**

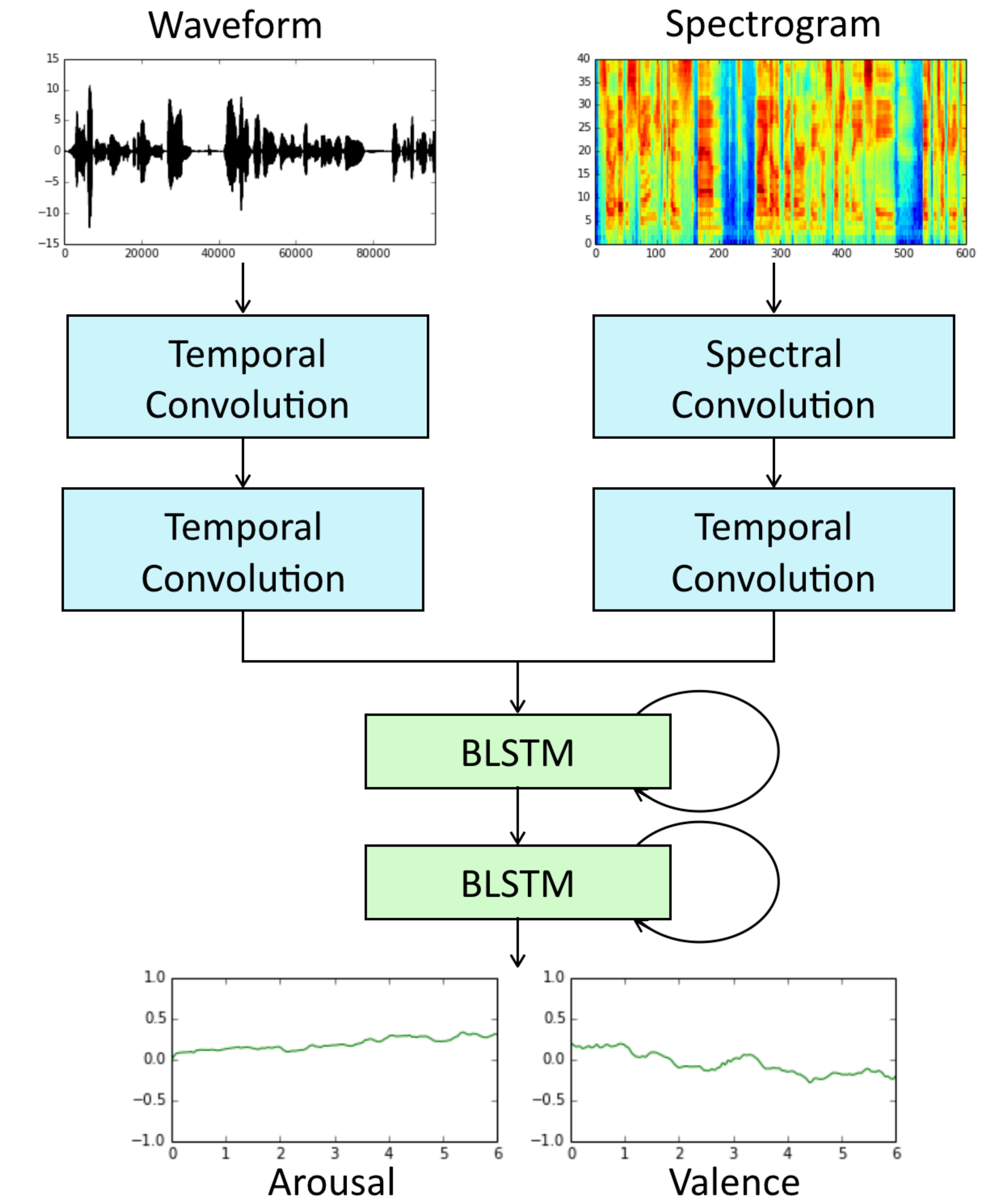



Do spectrograms and waveforms contain complementary information for emotion recognition in speech?

# Emotion Recognition - Dimensional

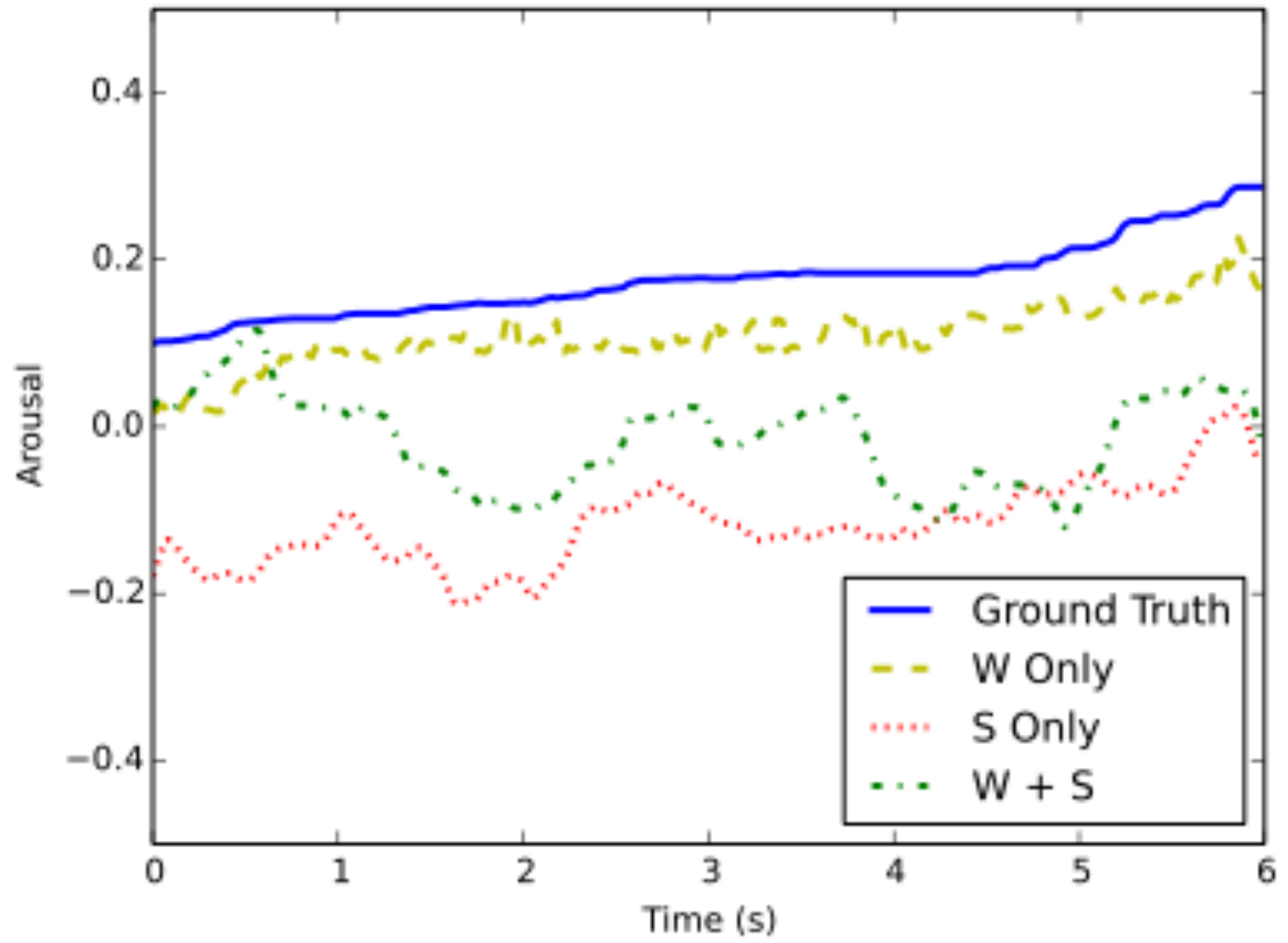
- Input: raw waveform and spectrogram
- Model: convolutional recurrent neural networks
- Task: Predict arousal and valence
  - Continuous in both time and value
- Results:

| 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** |

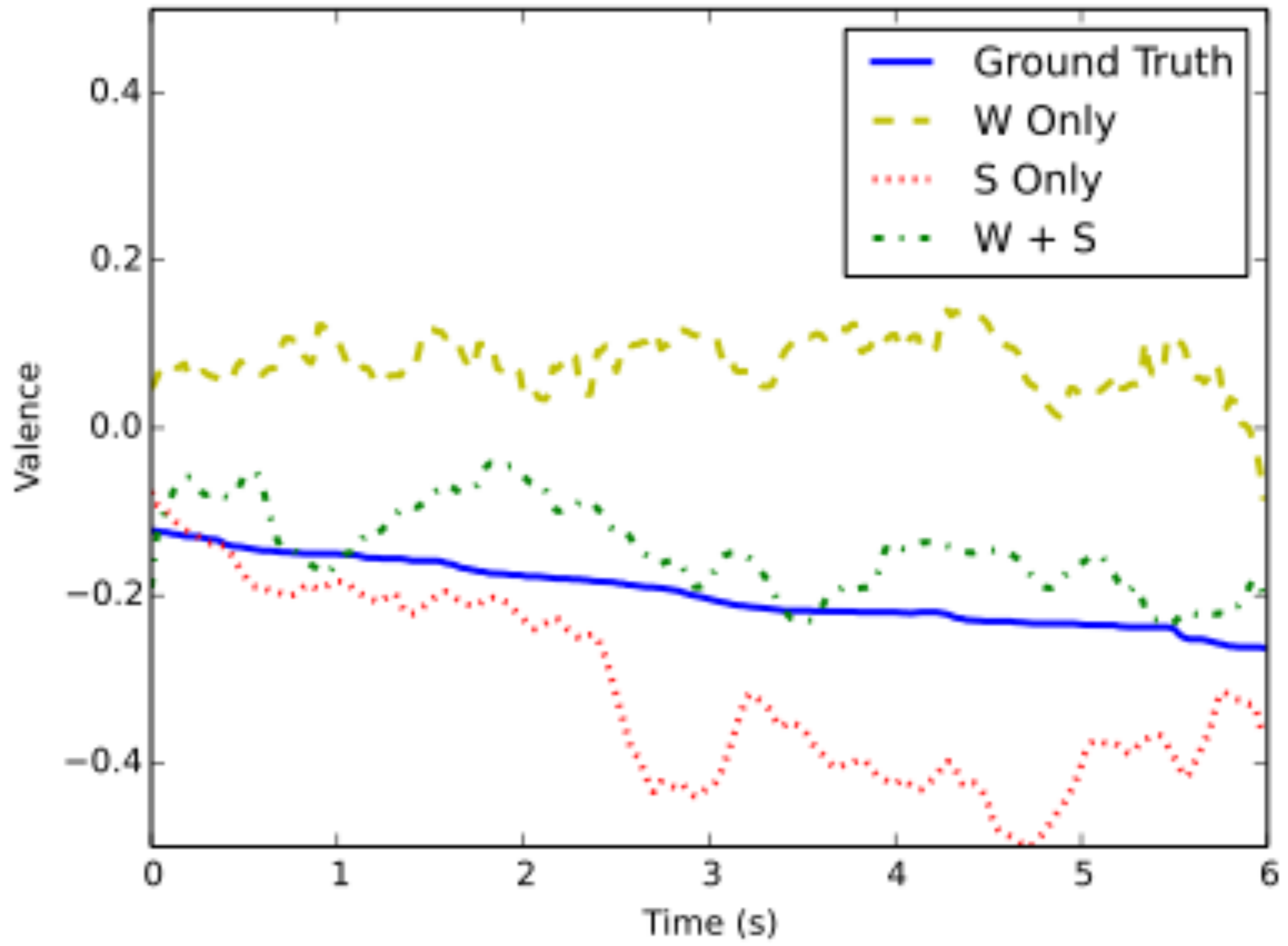


# Example Analysis - Dimensional

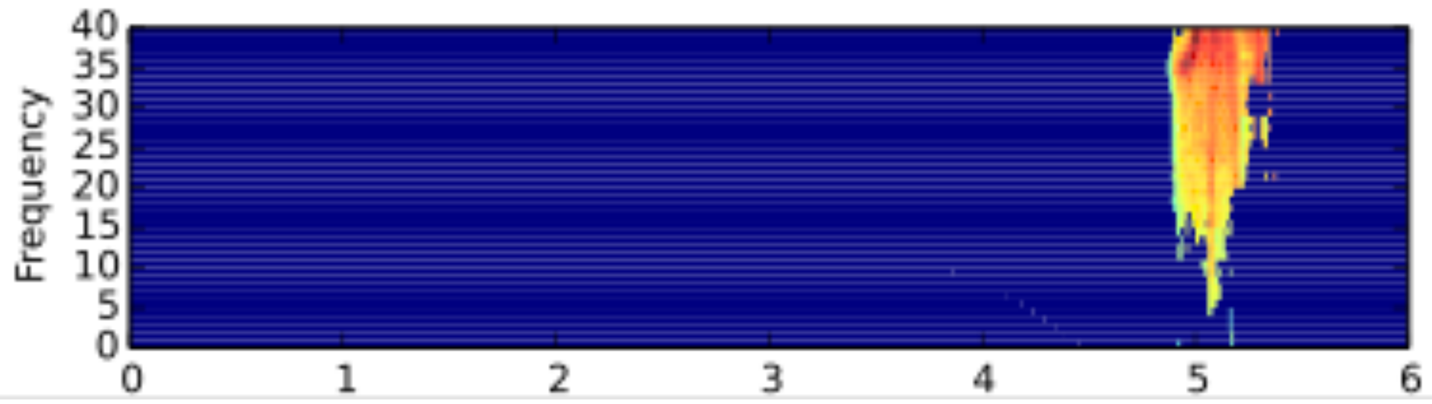
**Arousal**




**Valence**




**"…cos she's so frigging superior"**




Local Interpretable Modelagnostic Explanations (LIME)

# 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

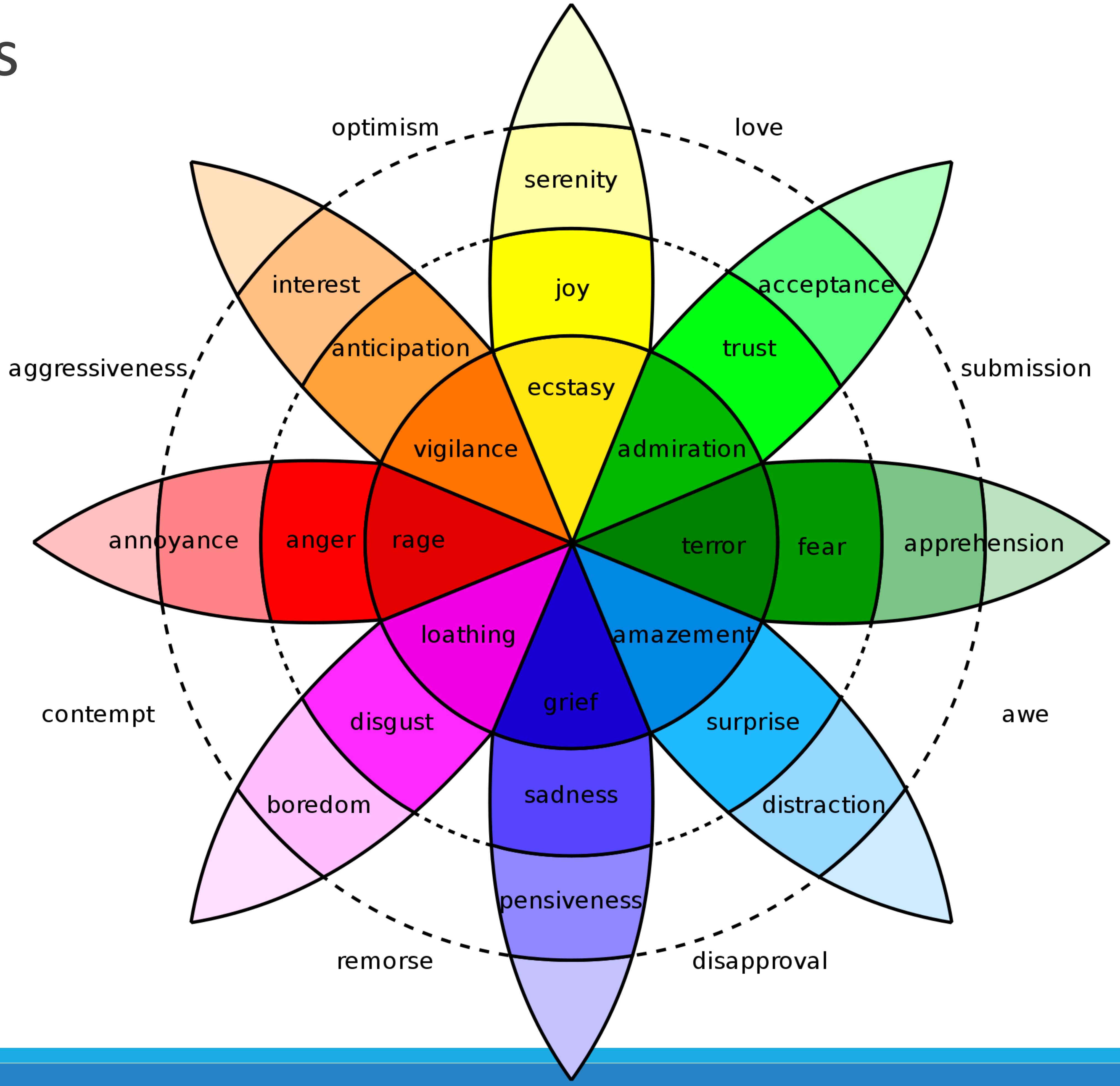
# Polyglot (Multilingual text processing toolkit )

- 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

# Plutchick'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)

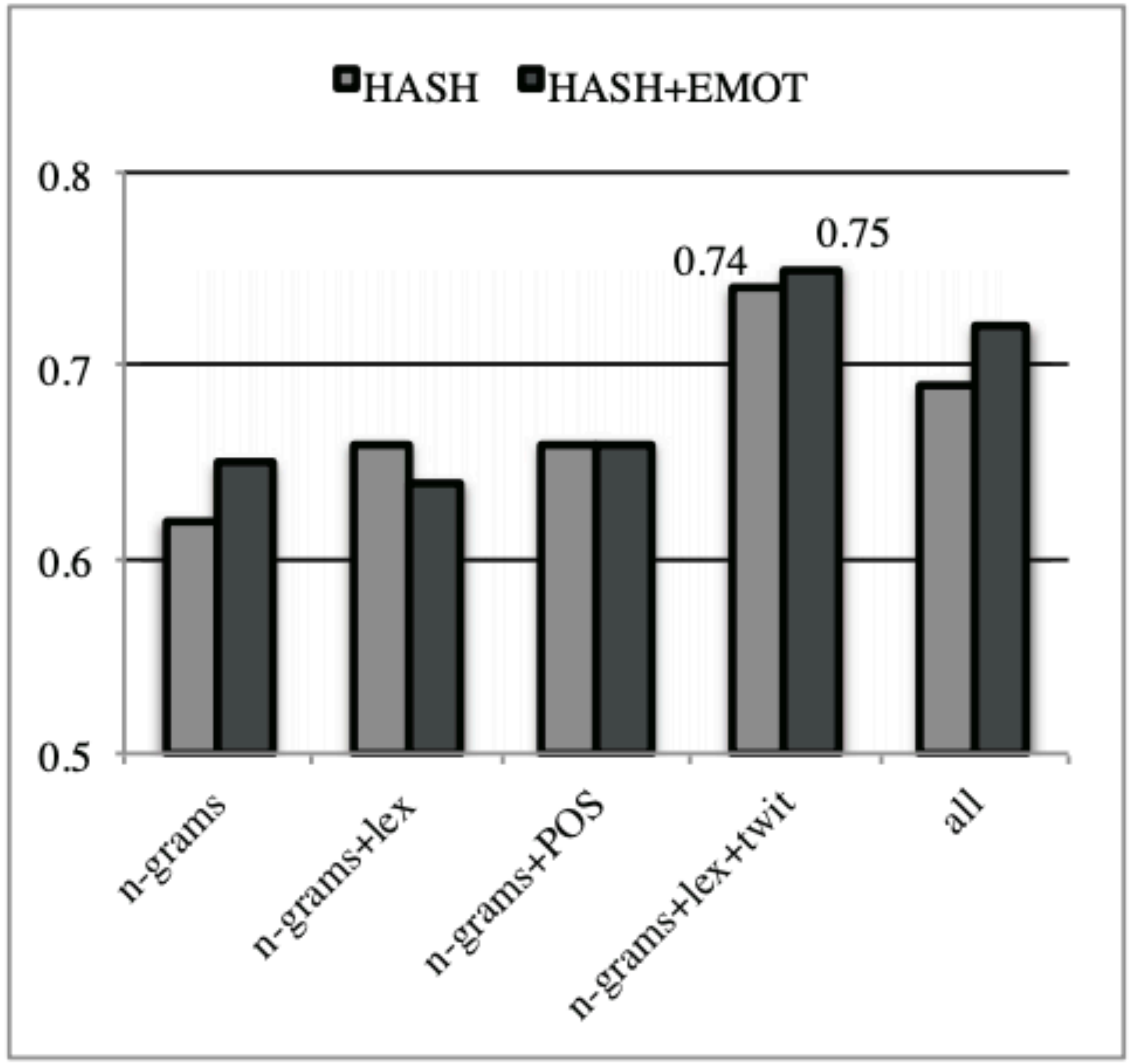
- Use emoticons to find tweets with sentiment

| 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%

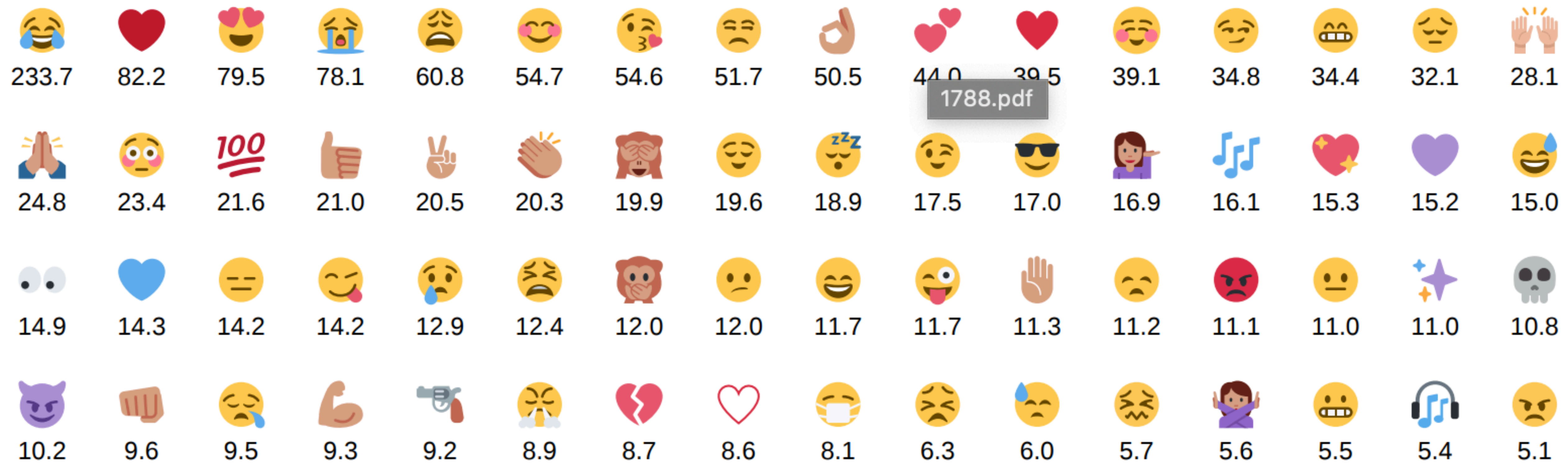
# Sentiment in Twitter #thingsilike (Kouloumpis et al. 2011)

| | |
|---|---|
| **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 |

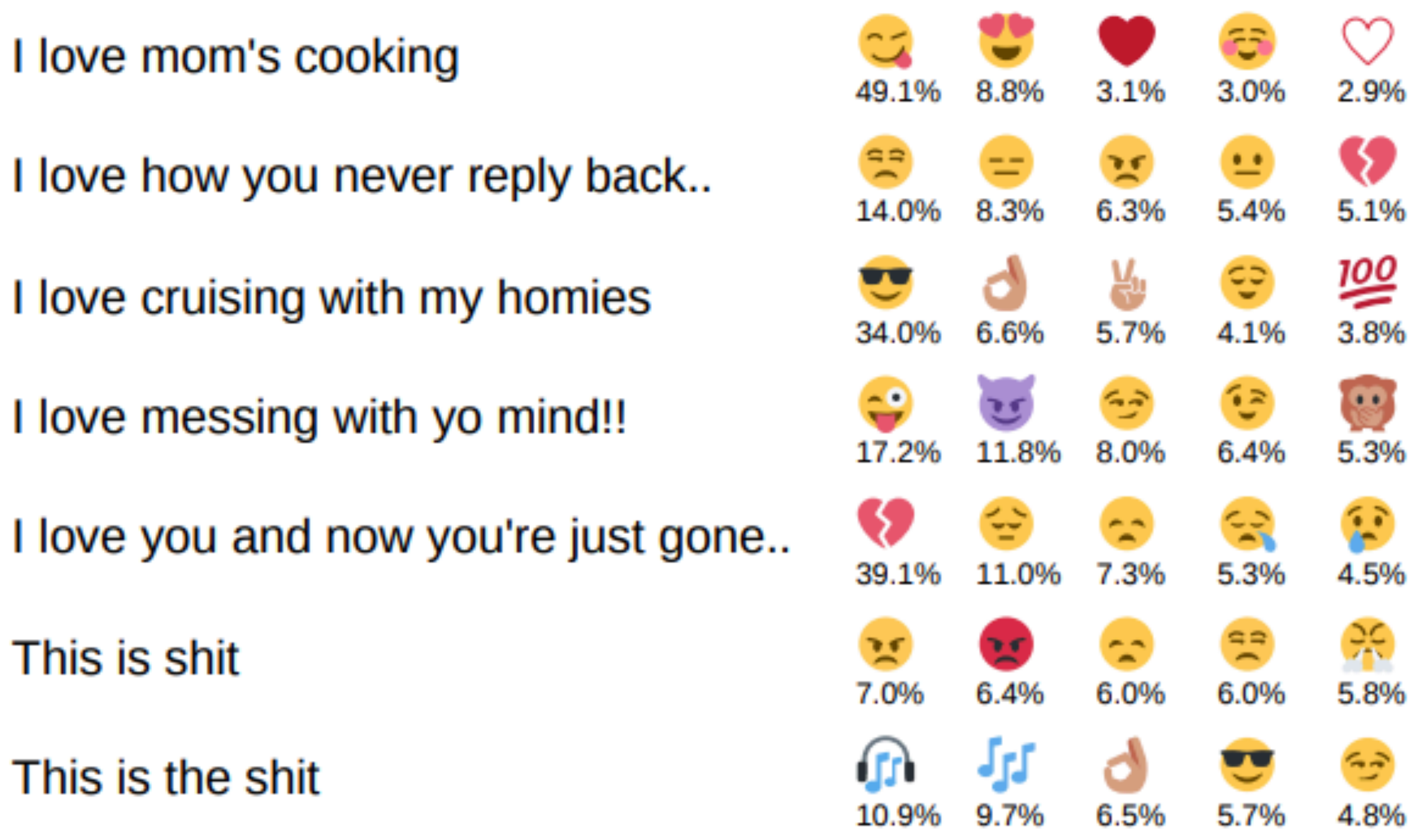


# Emoji in Twitter 😊 (Felbo et al. 2017)

- Number of training data (in *millions*)

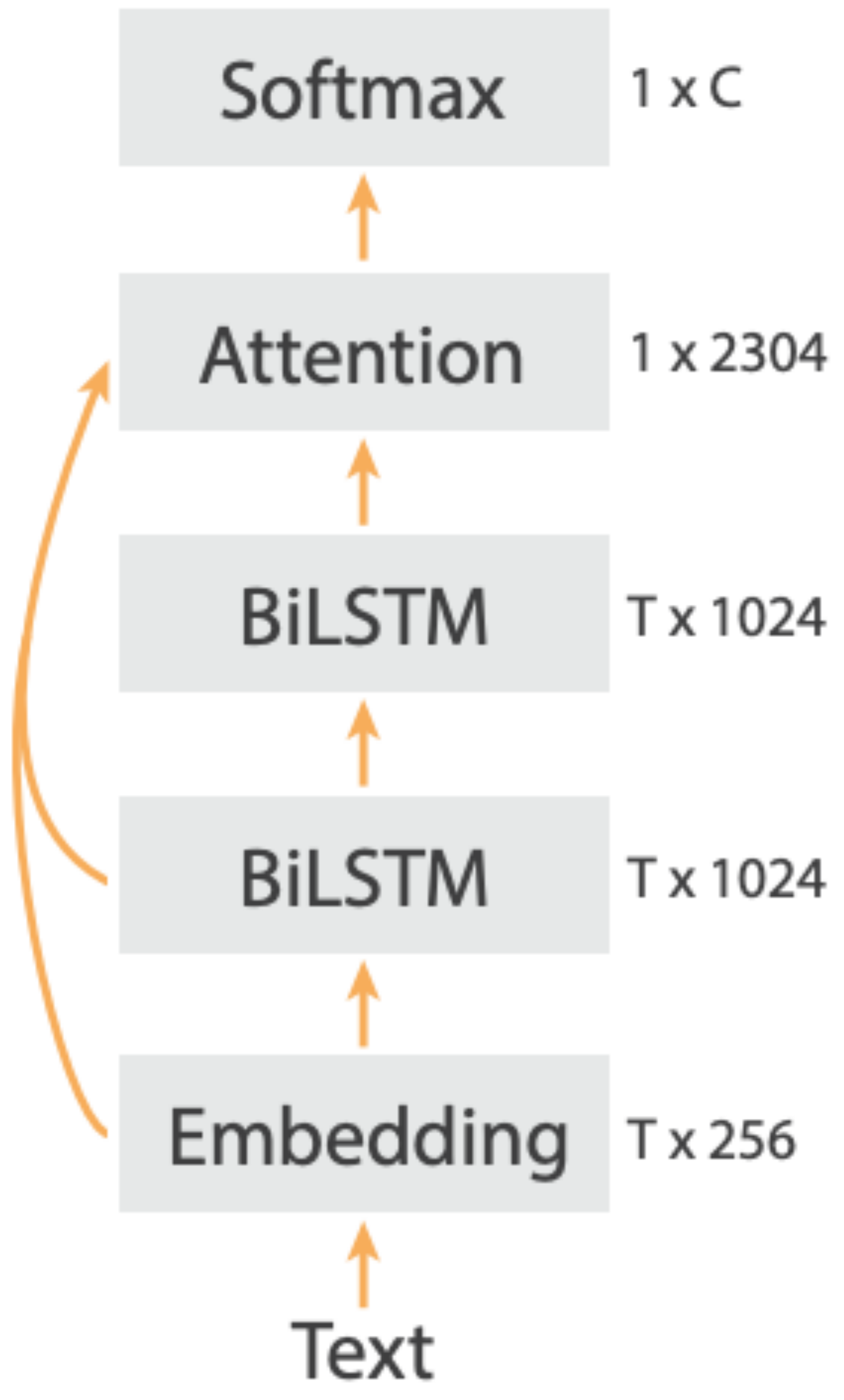



- Output: probability of emoji labels



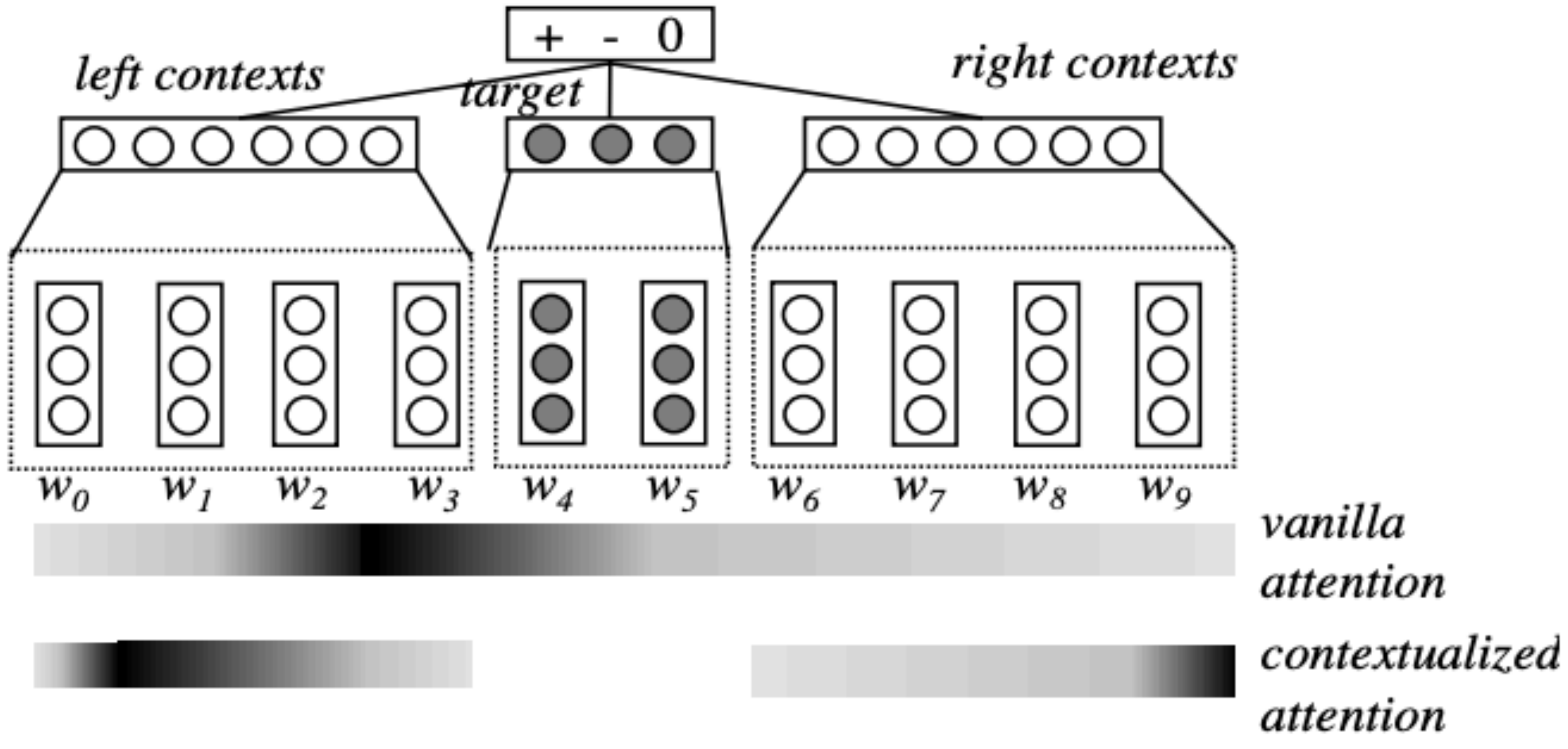
# Emoji in Twitter 😊 (Felbo et al. 2017)

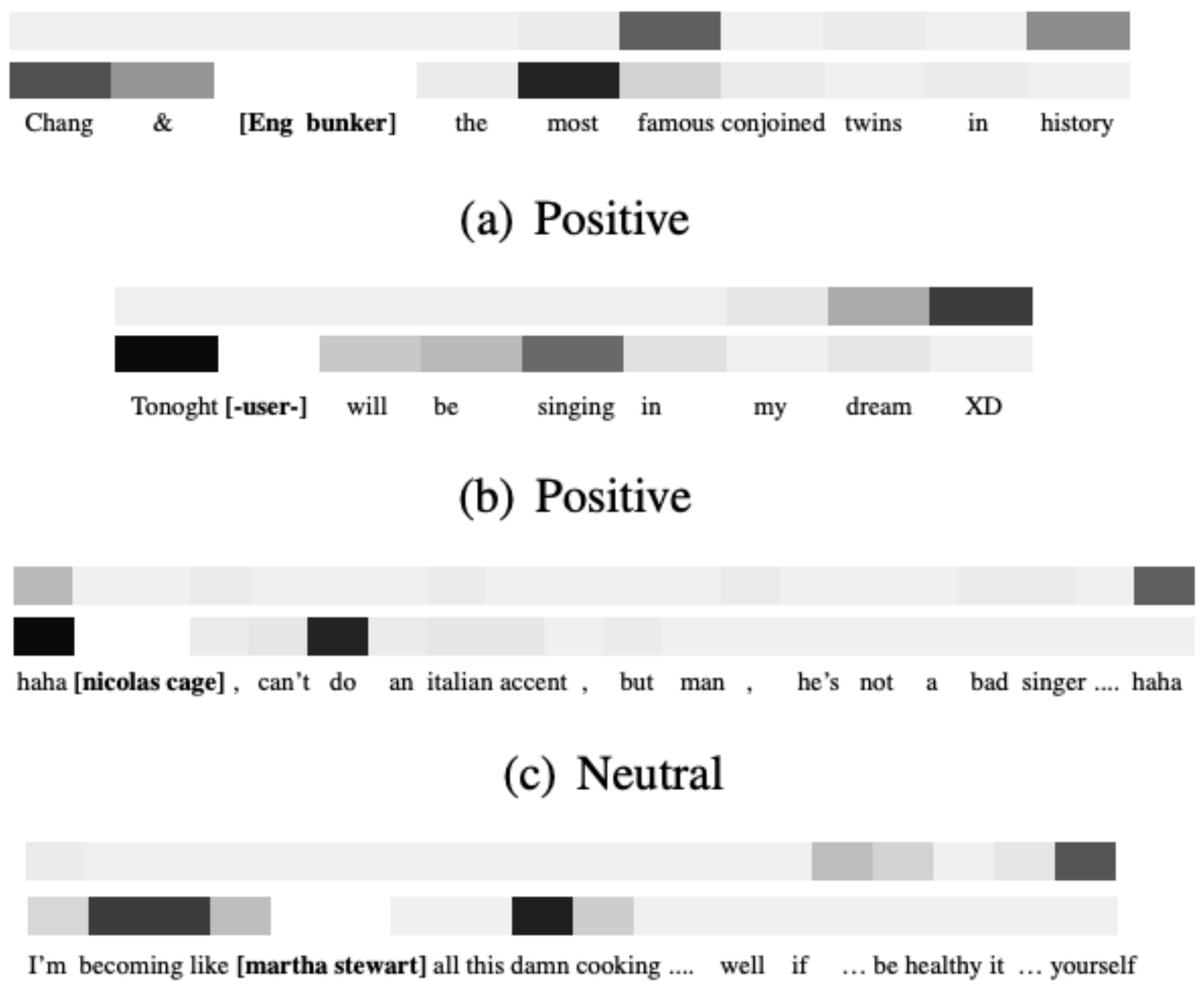
- DeepMoji model architecture



# 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

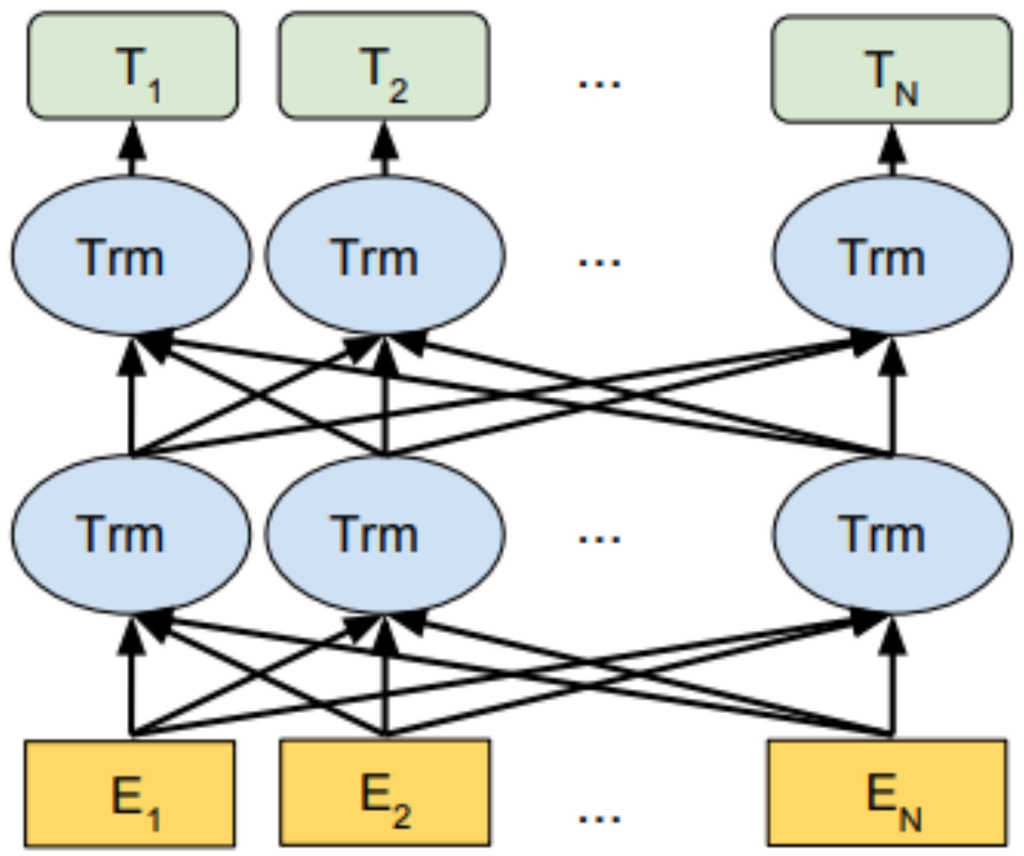
(b) Positive

(c) Neutral

(d) Negative

# BERT in Sentiment Analysis (Google AI Language)

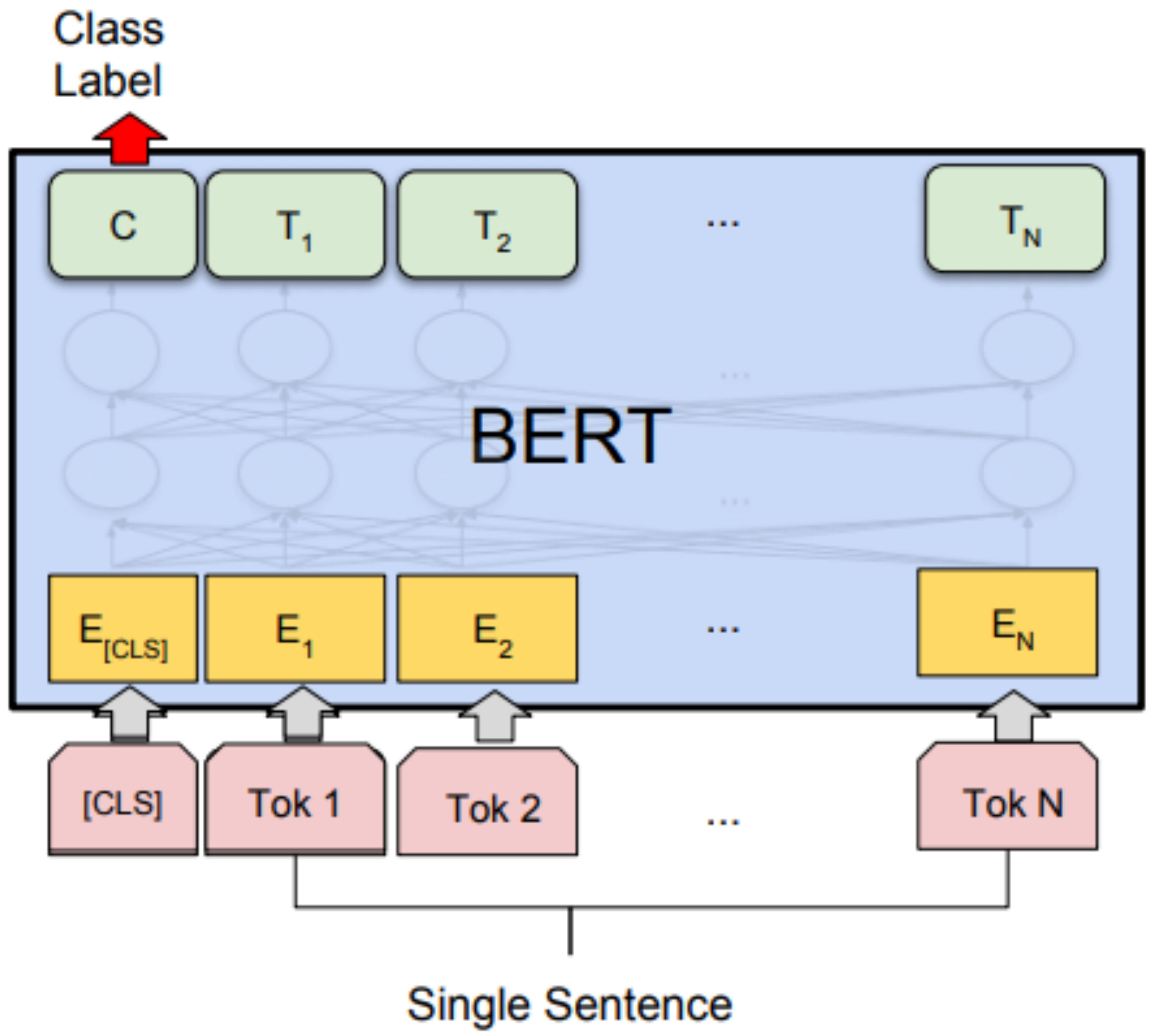
- BERT: Bidirectional Encoder Representations from Transformers
  - Transformer: stacked self-attention blocks




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

# BERT in Sentiment Analysis

- 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

# Situation Frame (SF) Detection

# LORELEI Project

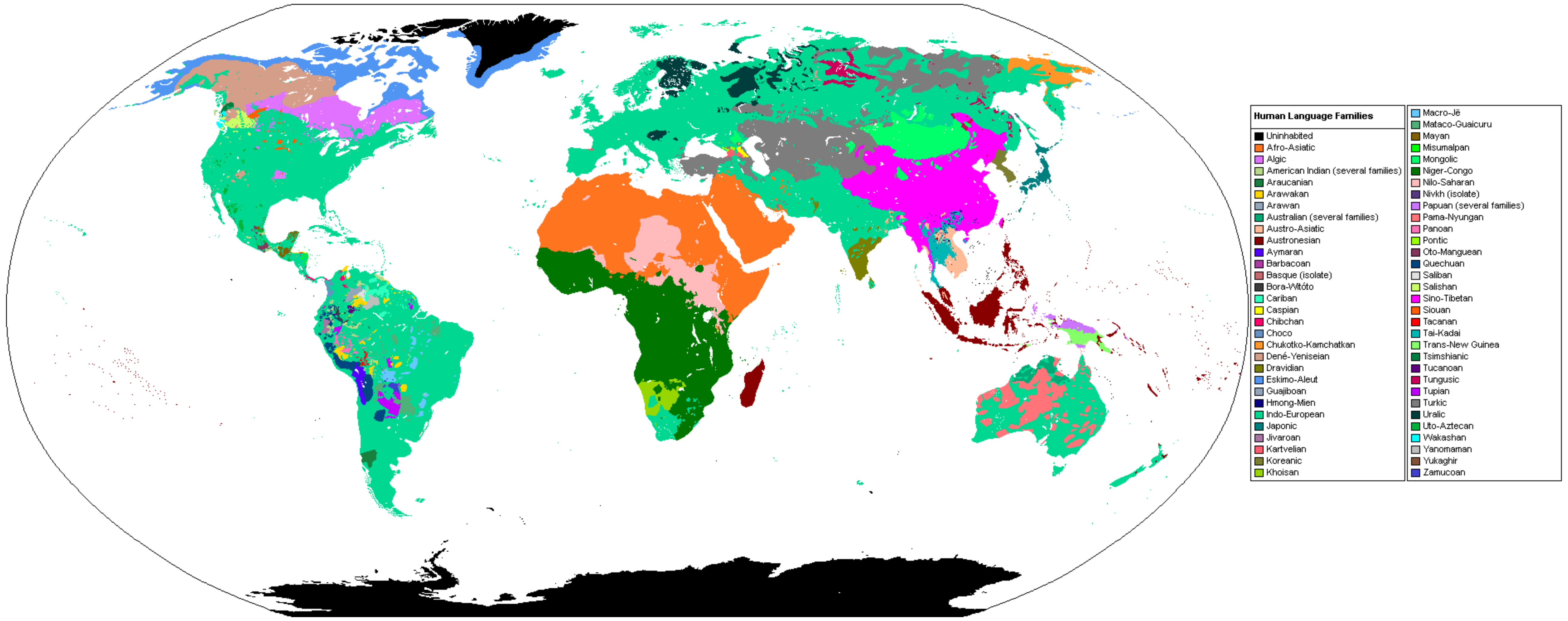
- **Low Resource Languages** for Emergent **Incidents** (LORELEI)
- Develop language technologies quickly to help first responders understand text and **speech** information
  - Using speech features to detect whether the speaker is talking about an incident
  - Keyword search in low-resource languages

# SF Speech - Overview

- Document-level situation frame (SF):
  - Type , Place , Status , and Confidence
- 11 SF Types:
  - Evacuation, food, water, medicine, infrastructure, shelter, rescue, utilities, crime, terrorism, regime change
- Two sub-tasks
  - Relevance layer: Does the segment contain at least 1 frame of any type?
  - Type layer: Which SF types (if any) are contained in the segment?

# SF Speech - Overview

- Available speech packs in 27 languages
  - Afro-Asiatic: AMH, SOM, ARA, HAU, IL5(Tigrinya), IL6 (Oromo)
  - Turkic: TUR, UZB, IL3(Uyghur)
  - Austronesian: TGL, IND
  - Niger–Congo: AKA, SWA, WOL, YOR, ZUL
  - Indo-European: BEN, FAS, HIN, RUS, SPA, USE
  - Sino-Tibetan: CHN
  - Uralic: HUN
  - Austroasiatic: VIE
  - Dravidian: TAM
  - Tai–Kadai: THA




- Incident languages (IL) for SF evaluation in 2018
  - IL9(Kinyarwanda), IL10(Sinhala)

# SF Speech – Relevance layer

- Binary classification
- Baseline model
  - openSMILE feature set
    - 384 hand-engineered features
  - Random forest model
    - limit the maximum depth to prevent overfitting
- End-to-end deep neural networks
  - CNN + LSTM
    - Adapt the model from speech emotion recognition task

# Cross-Language Experiments

- Higher accuracy for language pairs within the same language family

| | Afro-Asiatic | | | | | | Turkic | | | Austronesian | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | **AMH** | **SOM** | **ARA** | **HAU** | **IL5** | **IL6** | **TUR** | **UZB** | **IL3** | **IND** | **TGL** |
| **AMH** | \ | **0.62** | **0.62** | **0.59** | 0.56 | **0.66** | 0.62 | 0.67 | 0.66 | 0.66 | 0.58 |
| **SOM** | **0.65** | \ | **0.61** | **0.56** | **0.59** | 0.61 | 0.64 | 0.68 | 0.64 | 0.61 | 0.53 |
| **ARA** | **0.65** | **0.58** | \ | **0.59** | **0.58** | **0.65** | 0.72 | 0.73 | 0.62 | 0.67 | 0.63 |
| **HAU** | **0.68** | **0.59** | **0.65** | \ | **0.64** | 0.6 | 0.67 | 0.65 | 0.54 | 0.58 | 0.58 |
| **IL5** | 0.53 | **0.56** | 0.57 | **0.6** | \ | **0.65** | 0.67 | 0.62 | 0.56 | 0.56 | 0.49 |
| **IL6** | 0.63 | 0.54 | **0.61** | 0.55 | **0.6** | \ | 0.75 | 0.71 | 0.61 | 0.64 | 0.62 |
| **TUR** | 0.64 | 0.57 | 0.65 | 0.57 | 0.6 | 0.68 | \ | **0.74** | **0.6** | 0.65 | 0.62 |
| **UZB** | 0.59 | 0.55 | 0.65 | 0.53 | 0.59 | 0.65 | **0.76** | \ | **0.63** | 0.65 | 0.6 |
| **IL3** | 0.69 | 0.57 | 0.61 | 0.56 | 0.59 | 0.64 | **0.73** | **0.72** | \ | 0.64 | 0.64 |
| **IND** | 0.62 | 0.58 | 0.64 | 0.56 | 0.57 | 0.67 | 0.76 | 0.72 | 0.61 | \ | **0.65** |
| **TGL** | 0.63 | 0.52 | 0.61 | 0.53 | 0.58 | 0.63 | 0.69 | 0.63 | 0.61 | **0.66** | \ |

# SF Speech – Relevance layer

- Challenges
  - Coarse-grained annotation
    - 1 label for each utterance(up to 2 minutes)
  - Data from different sources in different languages
    - Tigrinya – VOA ; Oromo – local news
    - Hard to learn useful pattern across languages
- End-to-end deep neural networks
  - Tend to overfit training data
  - No significant improvement over baseline model

# SF Speech – Type layer

- Traditional method
  - Generate ASR transcript in the incident language
  - Translate into English
  - SF type detection in English
- Error propagation through the stages
  - English translation might be unintelligible
- Our method
  - Skip the ASR part
  - Query-by-example spoken term detection

# SF Speech – Type layer

- Step 1
  - Generate English keywords for each SF type
- Step 2
  - Ask the NI to translate and read the keywords in IL
  - Or use CMU TTS in IL to synthesize pronunciation
- Step 3
  - Find the IL keywords from speech segments
  - Calculate confidence scores for each SF type by the keyword search result

# SF Speech – Type layer

- Step 1 : Generate keywords for each SF type
- Method
  - Collect high frequency words for each type from SF annotated text data
  - Select related words manually
    - Remove incident-specific words in the training data
      - e.g. September (time), Turkey (place)
    - Delete overlapping words between types (e.g. injury appears in medicine, crime, rescue, etc.)
  - NI has to translate and read the words in 2 hours
    - 75 words in English

# SF Speech – Type layer

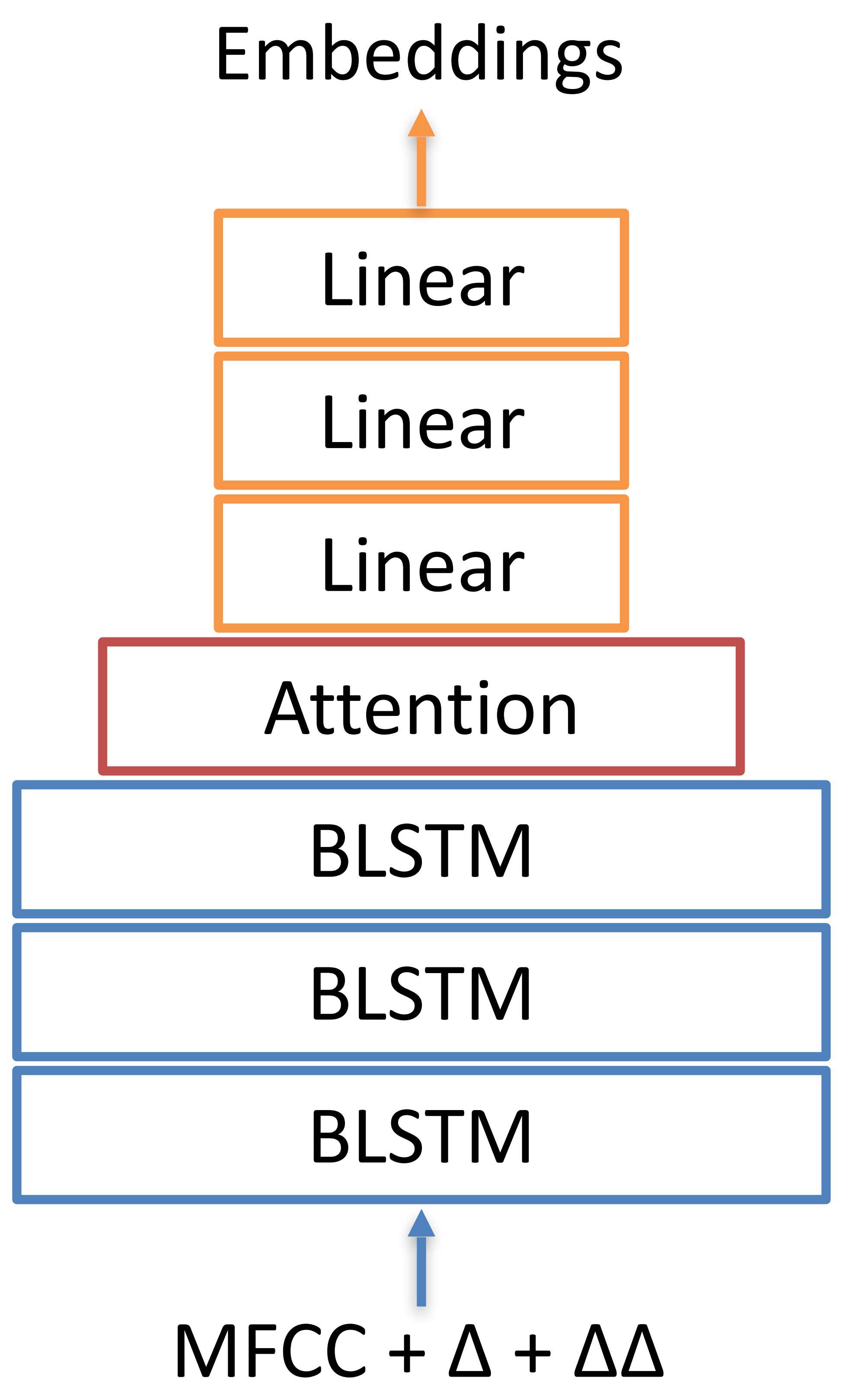
- Step 2 : Collect spoken keywords in IL from NI
- Method
  - 1 or 2 translations in IL for each English word
    - 108 words for Kinyarwanda; 122 words for Sinhala
  - Read/record the list 5 times
- Issue
  - Prosody, rising tone in list intonation
    - Ask NI:  try to pretend this is not a list; multiple reminders
  - Background sounds
    - The NIs in both ILs have babies crying, people walking, cooking? in background

# SF Speech – Type layer

- Step 3: Find keywords from speech
- Method
  - **Generate acoustic embeddings for spoken words**
  - Calculate the similarity between the embeddings of IL keywords and the embeddings of evaluation utterances
    - 2s sliding window, 0.5s stride on evaluation utterances
  - The confidence score of each SF in each utterance is the aggregation of similarity scores of all keywords that are related to that SF

# Siamese Neural Networks

- Base structure: generate embeddings for spoken words

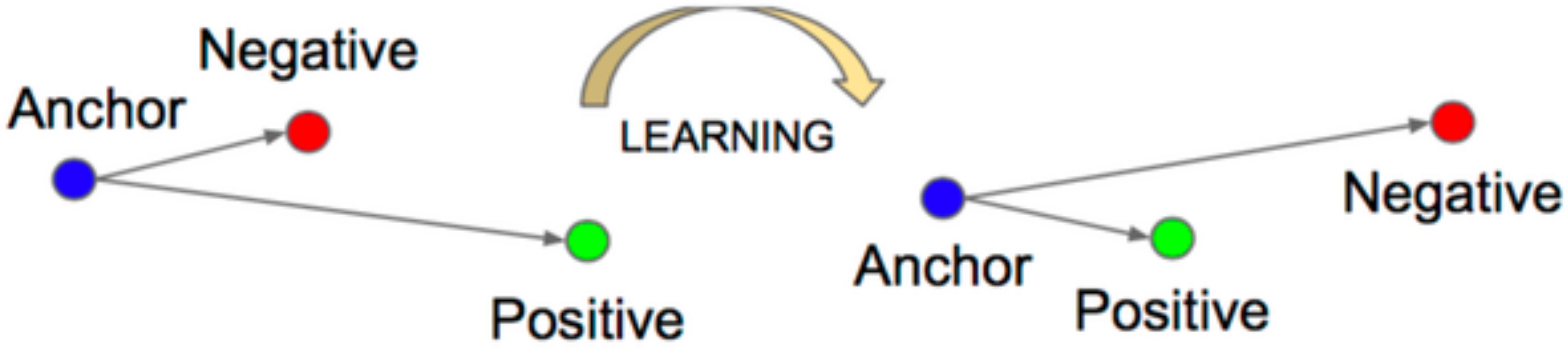


# Siamese Neural Networks

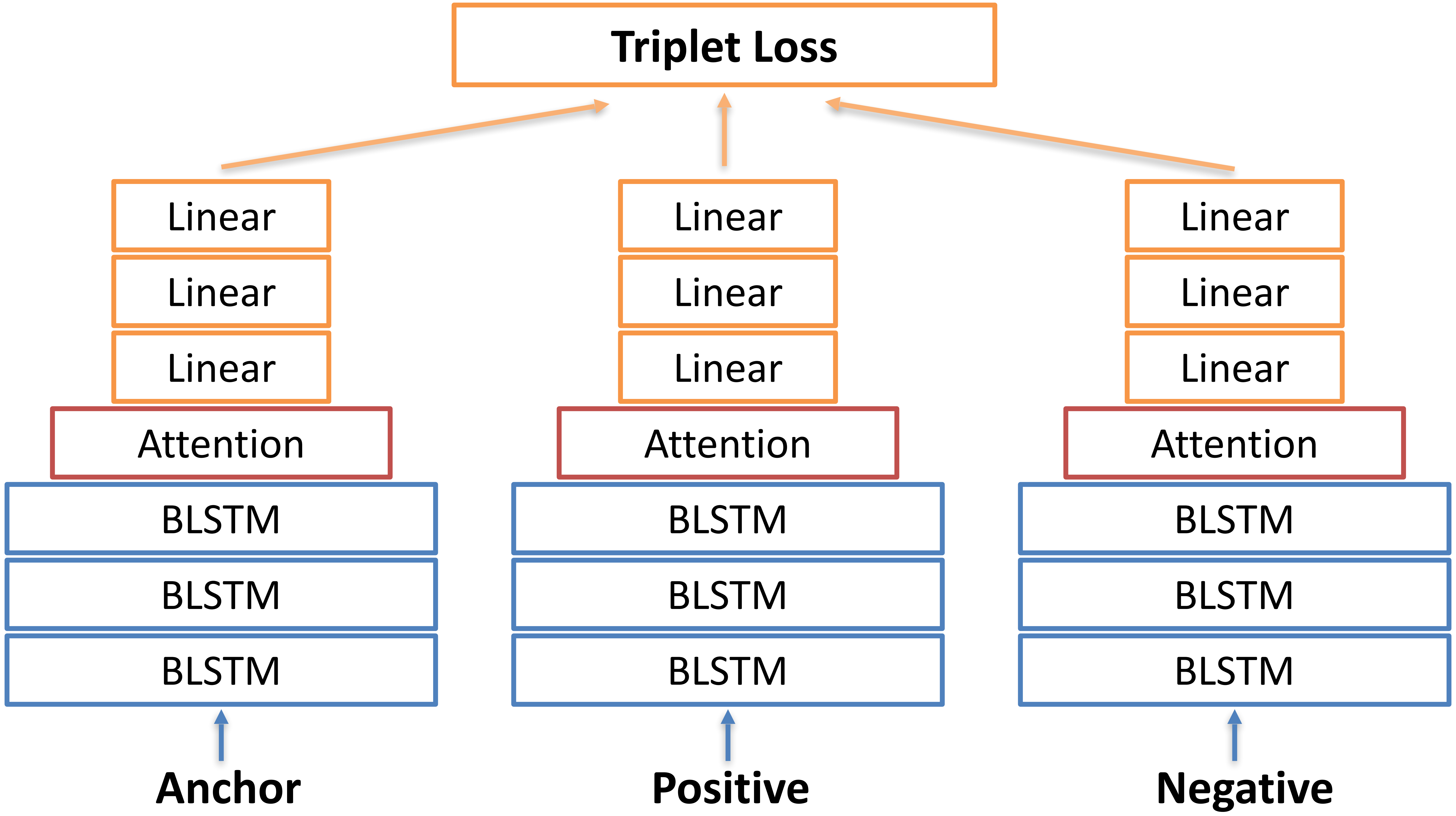
- Triplet Loss Function: (anchor, positive, negative)

$$Loss(x_a, x_p, x_n) = max\{0, m + d(x_a, x_p) - d(x_a, x_n)\}$$

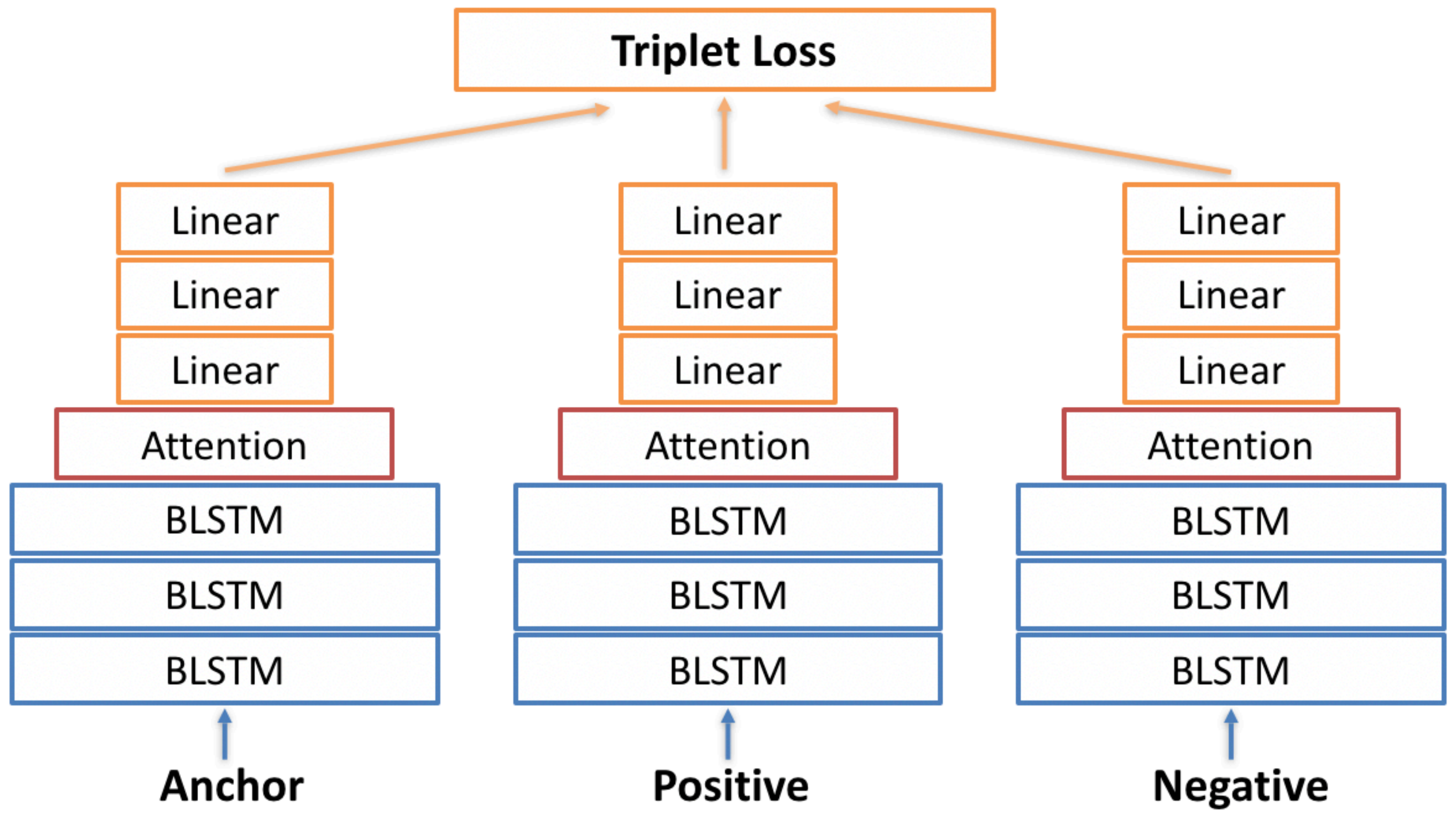
  – Bring the Anchor (current instance) close to the Positive (another instance of the same word) as far as possible from the Negative (an instance of a different word)



# Siamese Neural Networks



# Siamese Neural Networks




- In each triplet:
  - Anchor: current word
  - Positive: another sample of the same word
  - Negative: the nearest among 5 randomly chosen different word
- A problem in this commonly used approach:
  - Whether two words are the ‘same word’ or ‘different word’ depends on their exact orthographic representations
  - ‘terrorist’ and ‘terrorism’ will be encourage to have dissimilar embeddings, even if they share the same stem and are pronounced similarly

# Improving Acoustic Word Embeddings

- Observation:
  - Both IL9 (Kinyarwanda) and IL10 (Sinhala) are morphologically rich languages
    - IL9: kwica (crime), kwicana (criminal)
    - IL10: ත්‍රස්තවාදියා terrorist<br>ත්‍රස්තවාදය terrorism

- If the embedding method can map words like this together, we may not need to collect all possible inflections

# Improving Acoustic Word Embeddings

1. Clustering words by their **stems**
   - In each triplet:
     - Anchor: current sample
     - Positive: another sample of the same **stem**
     - Negative: the nearest among 5 samples of different **stems**
2. Learning **pronunciation distance**

$$Loss(x_1, x_2) = (d(x_1, x_2) - d_{edit}(phone_1, phone_2))^2$$

# Low-resource Setting Experiments

- Using a subset of Switchboard (English)
  - 10k, 11k and 11k samples on the train, dev, and test
  - Less than 2 hours of speech for training
- Evaluation metrics: average precision on word-pairs (Word AP); average precision on stem-pairs (Stem AP); the correlation of embedding distance with phonetic similarity (Phonetic Sim).

| Model | Word AP | Stem AP | Phonetic Sim |
|---|---|---|---|
| Word Triplet | **44.5** | 47.8 | 23.3 |
| Stem Triplet | 42.3 | **54.1** | 21.7 |
| Pronunciation Dist | 26.8 | 27.3 | **38.8** |

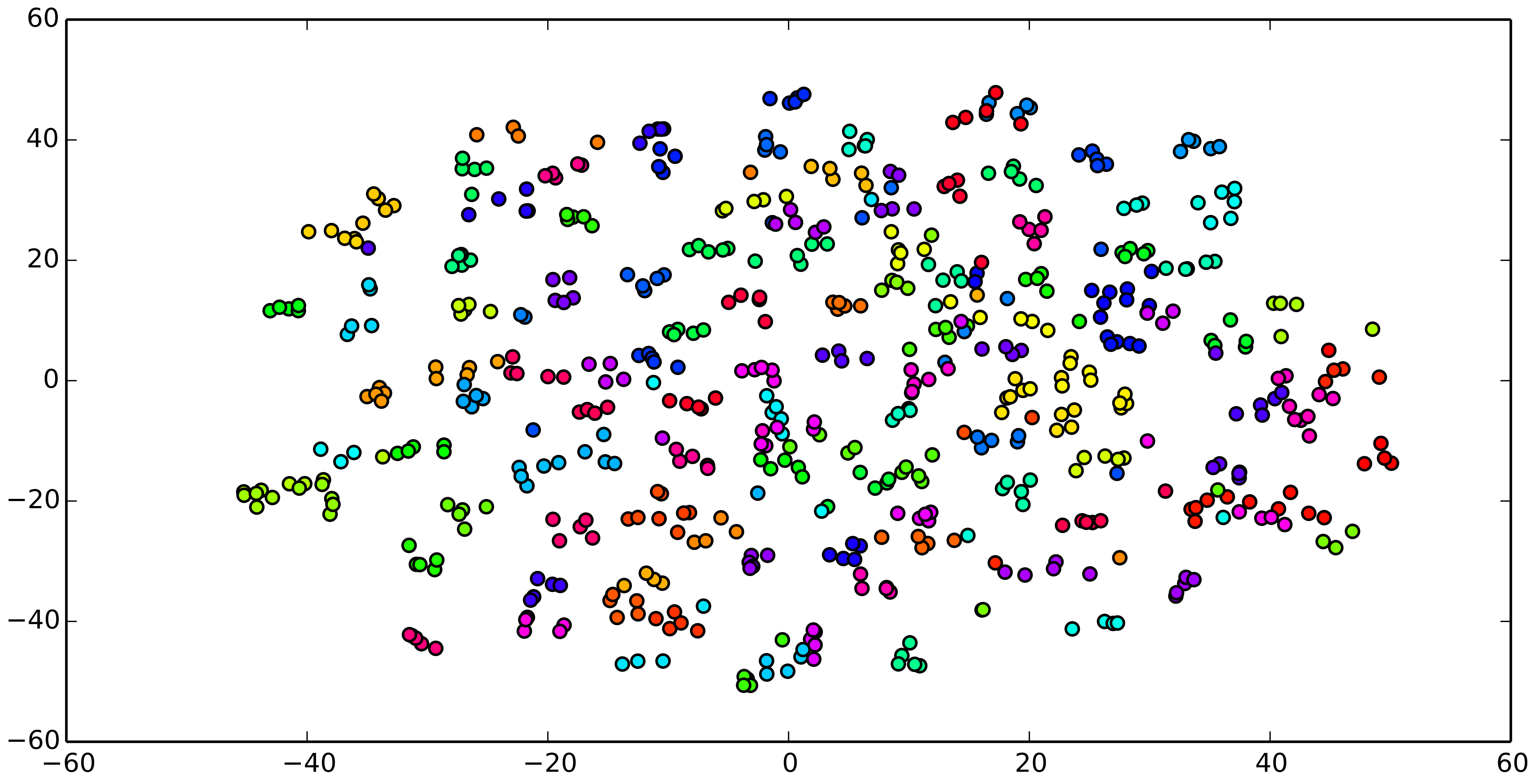
# Zero-resource Setting Experiments

- Train on full Switchboard dataset
  - Select: all words with duration 0.5s to 2.0s & appearing at least 2 times
  - 205270 samples, 11409 unique words
- Test on IL10 (Sinhala) keywords: 610 samples, 121 unique words
- Note: In these metrics, acoustically similar words in IL10 such as 'terrorist' and 'terrorism' are treated as **different** words

| Model | Word AP | Word P@4 |
|---|---|---|
| Word Triplet | 57.2 | **81.6** |
| Stem Triplet | **60.3** | 81.1 |
| Pronunciation Dist | 24.4 | 76.1 |

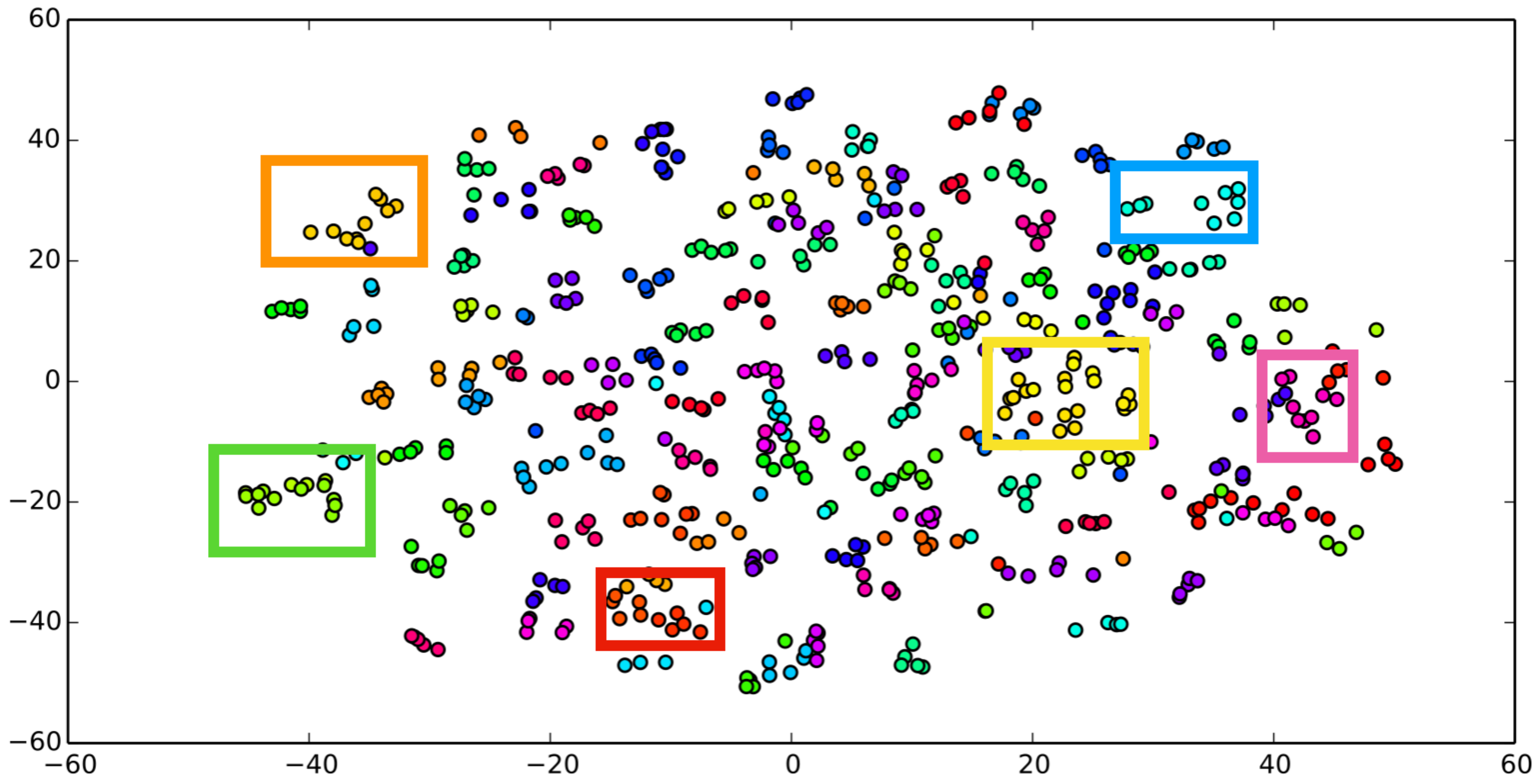
# Results on IL10 keywords

t-Distributed Stochastic Neighbor Embedding (t-SNE)



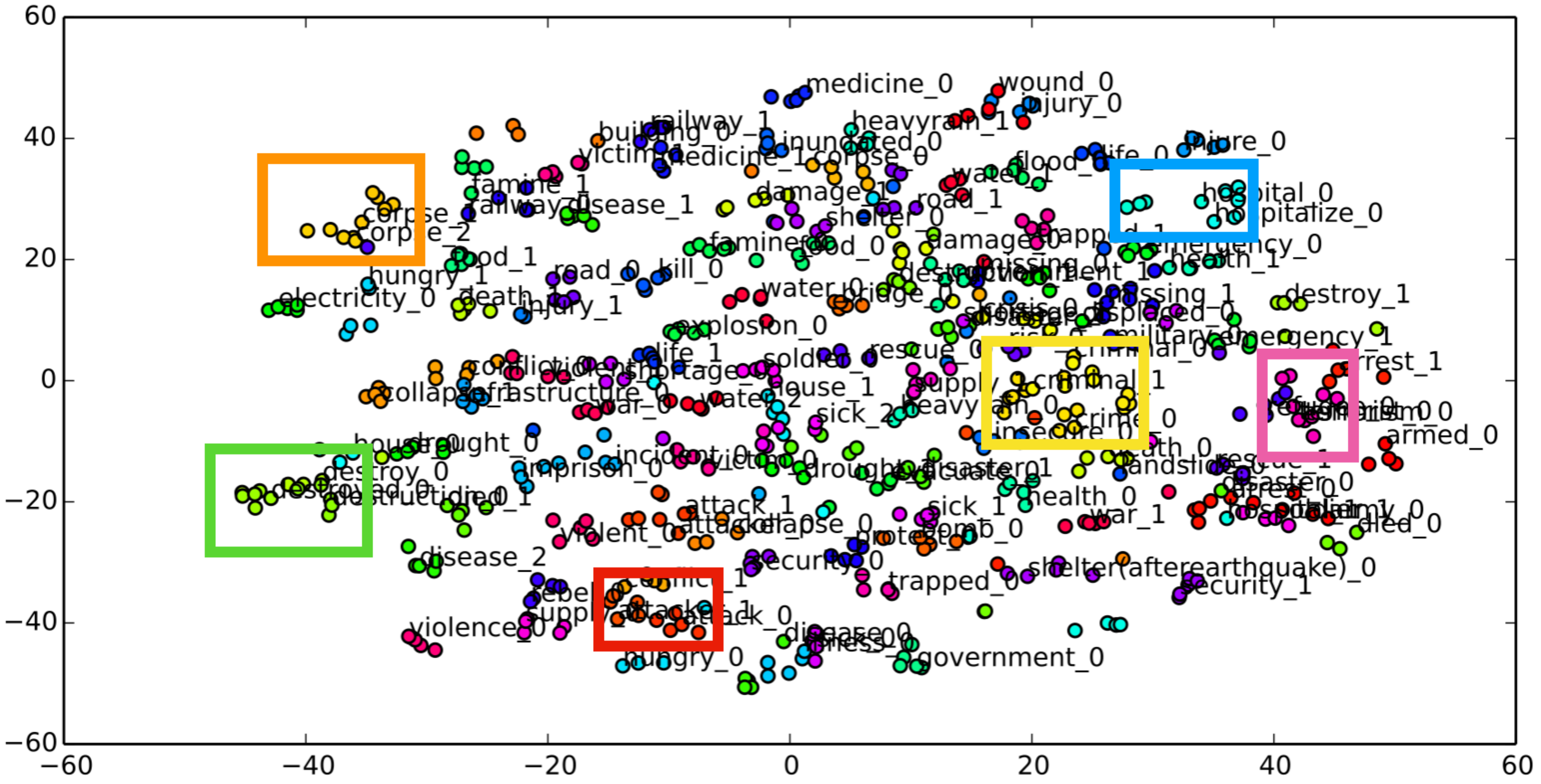
# Results on IL10 keywords

t-Distributed Stochastic Neighbor Embedding (t-SNE)



# Results on IL10 keywords

t-Distributed Stochastic Neighbor Embedding (t-SNE)



# Homework 4 - Emotion Recognition

# Homework 4 - Overview

- Emotion recognition in speech
- Dataset: the Emotional Prosody Speech and Transcript
  - 7 speakers: 4 female, 3 male
  - 15 emotions: neutral, interest, anxiety, pride, boredom, panic, cold-anger, hot-anger, contempt, elation, happy, shame, disgust, sadness, despair
  - 2324 speech utterances
  - Acted speech
  - Speech contents are semantically neutral

# Homework 4 - Feature Analysis

- Extract six features from each speech segment:
  - The min, max, mean of pitch
  - The min, max, mean of intensity
- Praat or Parselmouth
  - Pitch range 75~600 Hz; autocorrelation as pitch analysis method
  - Use only the left channel (channel 1)
- Normalization
  - Z-score normalization over the individual speaker
  - Normalizing by each speaker’s neutral utterances

# Homework 4 - Feature Analysis

- Plots of the mean and standard deviation of each feature across all emotion classes
  - 12 figures (6 before normalization, 6 after normalization)
  - 15 bars in each figure (with error bars for std)

- Report and discuss at least 5 observations

# Homework 4 - Classification Experiments

- Extract a feature set using openSMILE toolkit
  - *SMILExtract -C config/a_feature_set.conf -I speech.wav -O feature.csv*
  - No need to write your own configuration file
  - Use the provided configuration files in ./config
    - Recommend: The INTERSPEECH 2009 Emotion Challenge feature set (IS09_emotion.conf)
      - 384 features
      - Acoustic features (e.g. pitch, energy, voicing probability, MFCCs)
      - Functions (e.g. min, max, range, stddev, slope of linear approximation)

# Homework 4 - Classification Experiments

- Experiments
  - Leave-one-speaker-out cross validation
    - 7 multiclass classification experiments
  - Report the average of precision, average of recall, and average of F1 for each emotion class (averaging across experiments)
  - Also report the average score over all emotions and all experiments
- sklearn.metrics.classification_report()

| | precision | recall | f1-score |
|---|---|---|---|
| class 0 | 0.50 | 1.00 | 0.67 |
| class 1 | 0.00 | 0.00 | 0.00 |
| class 2 | 1.00 | 0.67 | 0.80 |

# Homework 4 - Error analysis

Analyze the errors made by your best performing experiment.

- Which class(es) were easiest to predict? Why do you think they were easy?
- Which were most difficult? Why do you think they were difficult?
- Based on this analysis, what ideas do you have to further improve your classifier?

# Homework 4 - What to submit

- **Code:** Feature extraction and classification experiments
- Data: You don't have to submit any data, but please make sure that all features used in the experiments can be reproduced by running the code.
- **Report:** (1) feature analysis, (2) classification experiments, (3) error analysis
- **README:** Documentation of your code

# Thank you!