Learning Spontaneity to Improve Emotion Recognition in Speech

Karttikeya Mangalam, Tanaya Guha

Indian Institute of Technology, Kanpur

Interspeech 2018

September 6, 2018

Interspeech 2018
Introduction
Motivation
Objective
Past Work
Our work
Dataset Description
Feature extraction
Spontaneity Detection
Emotion Recognition
Schematic Models
Results
4-emotion classification
Clustered classification
Conclusion

Problem Statement

Emotion recognition is the process of identifying human emotion using one or a combination of input signals such as speech, facial expressions, body movement and gestures.

Applications

- Call centers - monitoring, automatic answering etc.
- Personal Home Assistants - Google Home, Amazon Echo
- Music, movie and media streaming & recommendation
- Social robots incorporating face analysis
- Several other consumer facing HCI applications
- Social anxiety therapy and other behavioral disorders

Interspeech 2018

Introduction
Motivation
Objective
Past Work
Our work
Dataset Description
Feature extraction
Spontaneity Detection
Emotion Recognition
Schematic Models
Results
4-emotion classification
Clustered classification
Conclusion

Motivation: Speech Emotion Recognition

- Holistic speaker modeling and similar downstream tasks
- Indispensable for imparting a 'chatty' aspect to human-machine conversations
- Foundational for adding other modalities

Paper Objective

- Explore the effect of spontaneity on emotion recognition from speech
- Look into suitable speech features for spontaneity detection in an interpretable manner

Interspeech 2018

Introduction
Motivation
Objective
Past Work
Our work
Dataset Description
Feature extraction
Spontaneity Detection
Emotion Recognition
Schematic Models
Results
4-emotion classification
Clustered classification
Conclusion

Previous Work

- Two step approach[1]
  - Frame by frame extraction low and mid level acoustic and prosodic features from raw speech
  - Use ML classifiers for pattern recognition
- Detection of Fluency & spontaneity is well studied[2]
- But relation with emotion recognition is not explored
- Recent use of CNN and LSTM network with attention for detection emotion in speech with self-learnt features.

Note: In the entirety of this work, we use Support Vector Machines as classifiers for pattern recognition.

[1] Jin 2015, Abdelwahab & Busso 2017, Zong 2016, Nwe 2003, Schuller 2003

[2] Dufour 2009, 2014

University of Southern California

USC

The Interactive Emotional Dyadic Motion Capture (IEMOCAP) Database

USC Viterbi
School of Engineering

- We use USC-IEMOCAP database[3] for evaluation.
- 12 hours of audiovisual data with MOCAP recordings
- 5 different sessions
- 151 dyadic conversations
- Over 10,000 labeled sentences
- Well balanced in spontaneity labels
- Very skewed (long-tailed) in emotion labels.

[3] IEMOCAP: Interactive emotional dyadic motion capture database, Busso et al 2008

Interspeech 2018
Introduction
Motivation
Objective
Past Work
Our work
Dataset Description
Feature extraction
Spontaneity Detection
Emotion Recognition
Schematic Models
Results
4-emotion classification
Clustered classification
Conclusion

Table 1: Data distribution of different classes in IEMOCAP database

Emotion	#Examples	%age Data	Emotion Group
Frustration	2901	29.3	Negative
Anger	1199	12.11	Negative
Excited	1934	19.54	Positive
Fear	101	1.02	Negative
Happiness	652	6.58	Positive
Sadness	1249	12.62	Negative
Neutral State	1720	17.38	Neutral
Surprise	0100	1.02	Positive
Others	26	0.20	Positive

Few classes have most of the examples.
$\rightarrow$ Either cluster or re-balance dataset by pruning

- Clustered Data distribution - Negative ($\sim$ 4550), Positive ($\sim$ 2750), Neutral ($\sim$2900) examples

Interspeech 2018

Introduction
Motivation
Objective
Past Work
Our work
Dataset Description
Feature extraction
Spontaneity Detection
Emotion Recognition
Schematic Models
Results
4-emotion classification
Clustered classification
Conclusion

- Speech features used in the Interspeech 2009 emotion challenge [Schuller 2009]
- Four low level descriptors (LLDs)
  - Mel-Frequency Cepstral Coefficients (MFCC)
  - Zero-Crossing Rate (ZCR)
  - Voice Probability (VP)
  - Fundamental Frequency (F0)

Interspeech 2018
Introduction
Motivation
Objective
Past Work
Our work
Dataset Description
Feature extraction
Spontaneity Detection
Emotion Recognition
Schematic Models
Results
4-emotion classification
Clustered classification
Conclusion

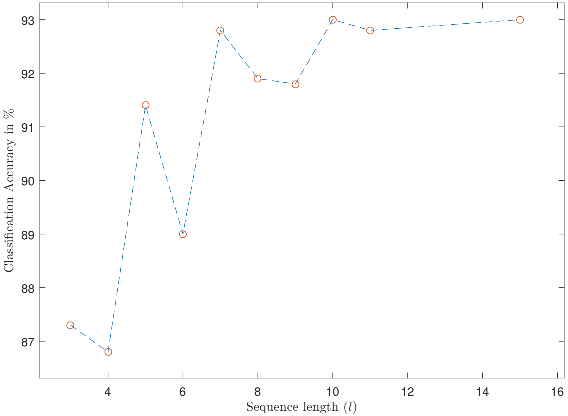
Key idea

Multitask learning to detect emotion and spontaneity simultaneously

Interspeech 2018

Introduction
Motivation
Objective
Past Work
Our work
Dataset Description
Feature extraction
Spontaneity Detection
Emotion Recognition
Schematic Models
Results
4-emotion classification
Clustered classification
Conclusion

- Identical spontaneity state across different sentences in the same conversation
- Use context for improving spontaneity detection
- Concatenate feature vectors from consecutive sentences.



Interspeech 2018

Introduction
Motivation
Objective
Past Work
Our work
Dataset Description
Feature extraction
Spontaneity Detection
Emotion Recognition
Schematic Models
Results
4-emotion classification
Clustered classification
Conclusion

Intermediate Conclusions

- Around $\sim 93\%$ accuracy on spontaneity detection!
- Good enough to use as an auxiliary task.
- But which features actually contribute?
- Are there some superfluous features confusing the classifier?

Interspeech 2018
Introduction
Motivation
Objective
Past Work
Our work
Dataset Description
Feature extraction
Spontaneity Detection
Emotion Recognition
Schematic Models
Results
4-emotion classification
Clustered classification
Conclusion

Intermediate Conclusions

- Around $\sim 93\%$ accuracy on spontaneity detection!
- Good enough to use as an auxiliary task.
- But which features actually contribute?
- Are there some superfluous features confusing the classifier?

$\Longrightarrow$ Feature Ablation Experiments.

Interspeech 2018

Introduction
Motivation
Objective
Past Work
Our work
Dataset Description
Feature extraction
Spontaneity Detection
Emotion Recognition
Schematic Models
Results
4-emotion classification
Clustered classification
Conclusion

Table: Effect of features on spontaneity classification accuracy (in %) for different sequence lengths

Feature(s) removed	$\ell = 5$	$\ell = 10$
None	91.4	93.0
ZCR	91.0	92.4
VP	90.6	92.6
F0	90.5	92.6
MFCC	83.4	85.5
VP, MFCC	80.7	83.8
F0, MFCC	83.2	84.9
ZCR, MFCC	78.8	82.3
VP, F0	90.6	91.5
VP, ZCR	90.2	92.1
F0, ZCR	90.6	92.2
VP, ZCR, F0	83.7	91.9
Any two, MFCC	< 76	< 80

Interspeech 2018

Introduction
Motivation
Objective
Past Work
Our work
Dataset Description
Feature extraction
Spontaneity Detection
Emotion Recognition
Schematic Models
Results
4-emotion classification
Clustered classification
Conclusion

Classifier Design : Multitask-Multilabel learning

- Based on above, we propose two different emotion recognition models
  - Multi-label Hierarchical Emotion Recognition
  - Joint Emotion and Spontaneity Recognition
- Both utilize spontaneity info but take different assumptions on data

Interspeech 2018
Introduction
Motivation
Objective
Past Work
Our work
Dataset Description
Feature extraction
Spontaneity Detection
Emotion Recognition
Schematic Models
Results
4-emotion classification
Clustered classification
Conclusion

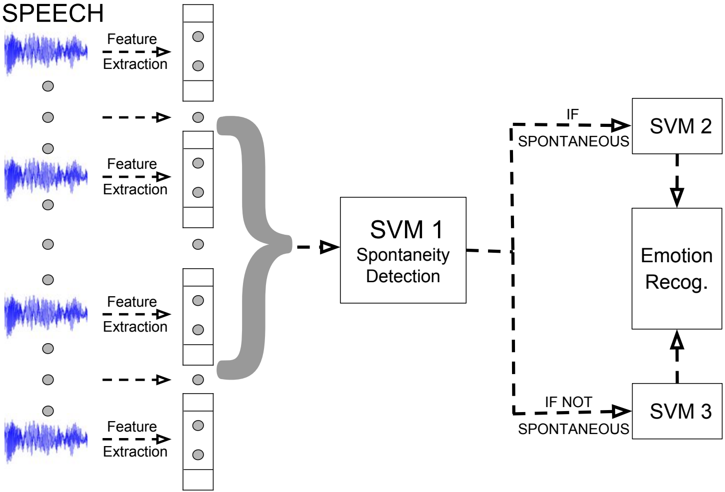


Figure: Multi-label Hierarchical Model

Interspeech 2018
Introduction
Motivation
Objective
Past Work
Our work
Dataset Description
Feature extraction
Spontaneity Detection
Emotion Recognition
Schematic Models
Results
4-emotion classification
Clustered classification
Conclusion

Loss Function Optimization

- weight matrix $\mathbf{W} \in \mathbb{R}^{|Y| \times d}$ containing a set of weight vectors $\mathbf{w}_{\{y^s, y^e\}}$

$$\mathcal{L}(\mathbf{W}, Y, \mathbf{F}) = \frac{1}{2} \sum_{(y^s, y^e) \in Y} ||\mathbf{w}_{\{y^s, y^e\}}||^2 + \mathbf{C} \sum_{j=1}^{N} \zeta_j$$

- $\mathcal{L}$ = regularization loss $\|\mathbf{w}_{\{y^s, y^e\}}\|$ + soft-margin loss ζ_j

Interspeech 2018

Introduction
Motivation
Objective
Past Work
Our work
Dataset Description
Feature extraction
Spontaneity Detection
Emotion Recognition
Schematic Models
Results
4-emotion classification
Clustered classification
Conclusion

Loss Function Optimization

- weight matrix $\mathbf{W} \in \mathbb{R}^{|Y| \times d}$ containing a set of weight vectors $\mathbf{w}_{\{y^s, y^e\}}$

$$\mathcal{L}(\mathbf{W}, Y, \mathbf{F}) = \frac{1}{2} \sum_{(y^s, y^e) \in Y} ||\mathbf{w}_{\{y^s, y^e\}}||^2 + \mathbf{C} \sum_{j=1}^{N} \zeta_j$$

- $\mathcal{L}$ = regularization loss $||\mathbf{w}_{\{y^s, y^e\}}||$ + soft-margin loss ζ_j

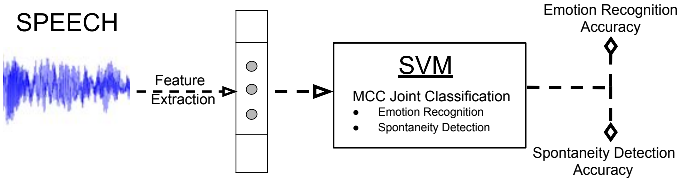


Figure: Joint Model for Emotion and Spontaneity Recognition

Interspeech 2018

Introduction
Motivation
Objective
Past Work
Our work
Dataset Description
Feature extraction
Spontaneity Detection
Emotion Recognition
Schematic Models
Results
4-emotion classification
Clustered classification
Conclusion

Results

Interspeech 2018

Introduction
Motivation
Objective
Past Work
Our work
Dataset Description
Feature extraction
Spontaneity Detection
Emotion Recognition
Schematic Models
Results
4-emotion classification
Clustered classification
Conclusion

Table: Emotion recognition results for all classes together in terms of weighted accuracy (in %) for pure 4-class classification.

	Scripted	**Spontaneous**	**Overall**
SVM baseline	56.8	73.0	65.4
RF baseline	62.1	66.0	64.1
CNN-based [10]	53.2	62.1	56.1
Rep. learning [11]	-	52.8	50.4
Spontaneity-aware methods			
LSTM [12]	-	-	56.7
Hierarchical	**64.2**	**74.0**	**69.1**
Joint	63.2	69.8	66.1

Interspeech 2018

Introduction
Motivation
Objective
Past Work
Our work
Dataset Description
Feature extraction
Spontaneity Detection
Emotion Recognition
Schematic Models
Results
4-emotion classification
Clustered classification
Conclusion

Table: Emotion recognition results for individual classes in terms of weighted accuracy (in %) for pure 4-class classification.

	Anger	Joy	Neutral	Sadness
SVM baseline	69.2	37.0	62.9	**76.9**
RF baseline	73.1	6.1	**78.8**	64.6
CNN-based [10]	58.2	**51.9**	52.8	66.5
Rep. learning [11]	53.5	36.9	52.6	64.3
Spontaneity-aware methods				
Hierarchical	**80.2**	37.5	65.9	73.3
Joint	71.2	13.1	75.9	76.3

Interspeech 2018
Introduction
Motivation
Objective
Past Work
Our work
Dataset Description
Feature extraction
Spontaneity Detection
Emotion Recognition
Schematic Models
Results
4-emotion classification
Clustered classification
Conclusion

Figure: Emotion recognition results for individual clusters in terms of weighted accuracy (in %) for clustered classification.

	Positive	Neutral	Negative	Spontaneous	Scripted	Overall
Baseline (SVM)	54.9	42.8	73.1	60.7	64.2	62.6
Baseline (RF)	53.6	48.3	67.3	58.6	61.6	60.1
Hierarchical	**66.9**	**48.9**	73.7	63.9	**67.5**	**65.7**
Joint	57.8	46.7	**74.3**	**64.1**	66.9	65.5

- Significantly poorer performance!
- Supplements training data
- Possible Reason: Confuses classifier with heteroskedastic feature vectors

Interspeech 2018

Introduction
Motivation
Objective
Past Work
Our work
Dataset Description
Feature extraction
Spontaneity Detection
Emotion Recognition
Schematic Models
Results
4-emotion classification
Clustered classification
Conclusion

Summary

- Detecting spontaneity in a multi-task approach helps emotion detection

Interspeech 2018

Introduction
Motivation
Objective
Past Work
Our work
Dataset Description
Feature extraction
Spontaneity Detection
Emotion Recognition
Schematic Models
Results
4-emotion classification
Clustered classification
Conclusion

Summary

- Detecting spontaneity in a multi-task approach helps emotion detection
- Hierarchical model for detection performs better than the joint model.

Interspeech 2018

Introduction
Motivation
Objective
Past Work
Our work
Dataset Description
Feature extraction
Spontaneity Detection
Emotion Recognition
Schematic Models
Results
4-emotion classification
Clustered classification
Conclusion

Summary

- Detecting spontaneity in a multi-task approach helps emotion detection
- Hierarchical model for detection performs better than the joint model.
- Grouping labels in Positive/Negative/Neutral clusters harms classification performance.

Interspeech 2018

Introduction
Motivation
Objective
Past Work
Our work
Dataset Description
Feature extraction
Spontaneity Detection
Emotion Recognition
Schematic Models
Results
4-emotion classification
Clustered classification
Conclusion

Summary

- Detecting spontaneity in a multi-task approach helps emotion detection
- Hierarchical model for detection performs better than the joint model.
- Grouping labels in Positive/Negative/Neutral clusters harms classification performance.
- Spontaneity detection as a standalone task is solvable to high accuracies ($\sim 93\%$) with the use of context and also boosts the performance for emotion recognition systems.

Interspeech 2018

Introduction
Motivation
Objective
Past Work
Our work
Dataset Description
Feature extraction
Spontaneity Detection
Emotion Recognition
Schematic Models
Results
4-emotion classification
Clustered classification
Conclusion

Thank you for your attention!

Questions?

Interspeech 2018

Introduction
Motivation
Objective
Past Work
Our work
Dataset Description
Feature extraction
Spontaneity Detection
Emotion Recognition
Schematic Models
Results
4-emotion classification
Clustered classification
Conclusion

Q. Jin, C. Li, S. Chen, and H. Wu, "Speech emotion recognition with acoustic and lexical features," in ICASSP 2015. IEEE, 2015, pp. 4749–4753.

M. Abdelwahab and C. Busso, "Ensemble feature selection for domain adaptation in speech emotion recognition," in ICASSP 2017, 2017.

Y. Zong, W. Zheng, T. Zhang, and X. Huang, "Cross-corpus speech emotion recognition based on domain-adaptive least-squares regression," IEEE Signal Processing Letters, vol. 23, no. 5, pp. 585–589, 2016.

T. L. Nwe, S. W. Foo, and L. C. De Silva, "Speech emotion recognition using hidden markov models," Speech communication, vol. 41, no. 4, pp. 603–623, 2003.

B. Schuller, G. Rigoll, and M. Lang, "Hidden markov model-based speech emotion recognition," in Multimedia and Expo, 2003. ICME'03. Proceedings. 2003

Interspeech 2018

Introduction
Motivation
Objective
Past Work
Our work
Dataset Description
Feature extraction
Spontaneity Detection
Emotion Recognition
Schematic Models
Results
4-emotion classification
Clustered classification
Conclusion

International Conference on, vol. 1. IEEE, 2003, pp. I–401.

R. Dufour, V. Jousse, Y. Estève, F. Béchet, and G. Linarès, "Spontaneous speech characterization and detection in large audio database," SPECOM, St. Petersburg, 2009.

R. Dufour, Y. Estève, and P. Deléglise, "Characterizing and detecting spontaneous speech: Application to speaker role recognition," Speech communication, vol. 56, pp. 1–18, 2014.

C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee, and S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," Language Resources and Evaluation, vol. 42, no. 4, pp. 335–359, 12 2008.

B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge," in Tenth Annual Conference of

Interspeech 2018

Introduction
Motivation
Objective
Past Work
Our work
Dataset Description
Feature extraction
Spontaneity Detection
Emotion Recognition
Schematic Models
Results
4-emotion classification
Clustered classification
Conclusion

the International Speech Communication Association, 2009.

M. Neumann and N. T. Vu, “Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech,” in Interspeech, 2017.

S. Ghosh, E. Laksana, L.-P. Morency, and S. Scherer, “Representation learning for speech emotion recognition.” in INTERSPEECH, 2016, pp. 3603–3607.

J. Kim, G. Englebienne, K. Truong, and V. Evers, Towards Speech Emotion Recognition ”in the wild” using Aggregated Corpora and Deep Multi-Task Learning. International Speech Communication Association (ISCA), 2017, pp. 1113–1117.