Speaker and Emotion Recognition of TV-Series Data Using Multimodal and Multitask Deep Learning

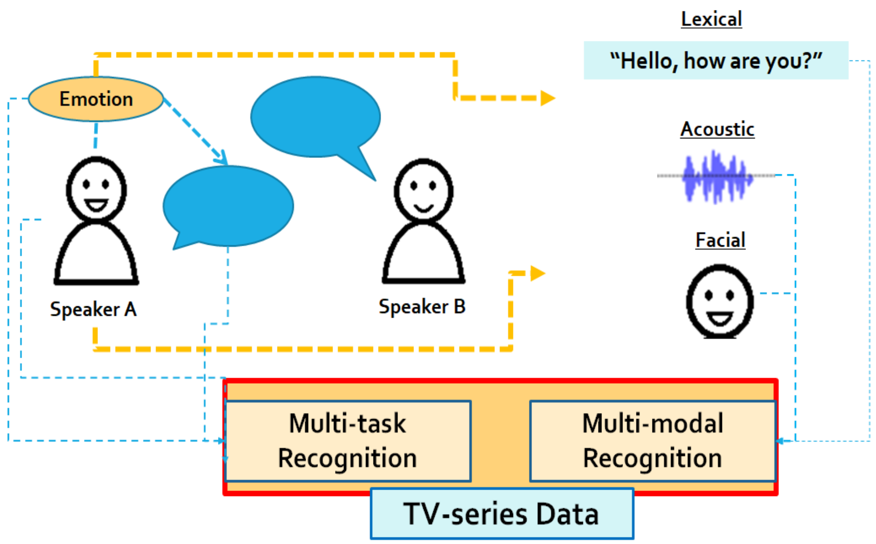
Sashi Novitasari[1], Quoc Truong Do[1], Sakriani Sakti[1,3], Dessi Lestari[2], Satoshi Nakamura[1,3]
[1] Graduate School of Information Science, Nara Institute of Science and Technology
[2] Department of Informatics, Bandung Institute of Technology
[3] RIKEN AIP
[1]{sashi.novitasari.si3, do.truong.dj3, ssakti, s-nakamura}@is.naist.jp
[2]{dessipuji}@informatika.org

Outline

1. Introduction
2. Data
3. Model Architectures
4. Features
5. Experiment
6. Conclusion

I. Introduction

- Real-life communication involves linguistic and paralinguistic aspects
- Multimodal and multitask recognition of non-verbal aspects of speech
- Recognition of speech's speaker and emotion from emotion-rich data



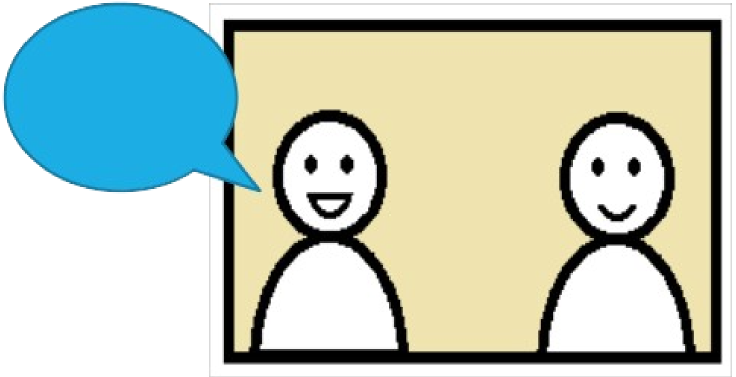
- Previous works:

  - Multimodal or multitask emotion-speaker recognition (not integrated)

  (Tang et al., 2016; Tian et al., 2016; Vallet et al., 2013)

II. Data

- TV-series data → expressive conversation
  - Video graphic: Facial features
  - Audio : Acoustic features
  - Subtitle : Lexical features
- English
- Utterance-level annotation
  - Speaker : 57 names
  - Emotion - valence : 3 classes (negative - neutral – positive)
  - Emotion - arousal : 3 classes (negative - neutral - positive)



Hello!

III. Model Architectures

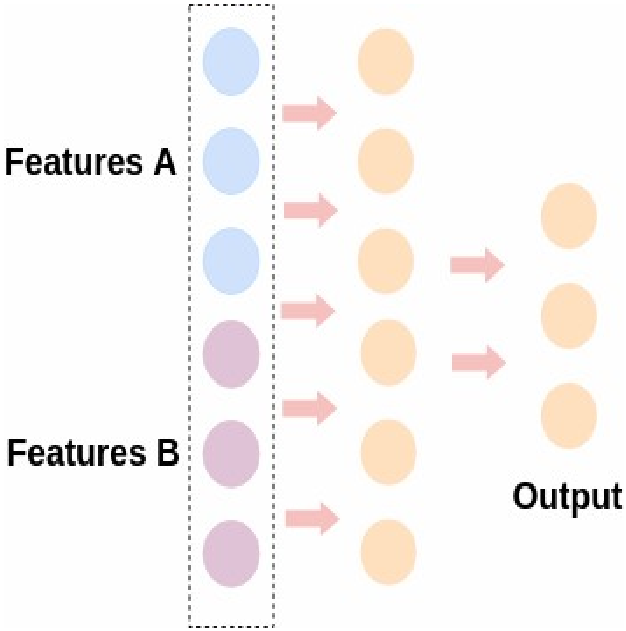
- Multilayer perceptron models (5 layers)
- Multimodal classification
- Multitask classification

III. Model Architectures

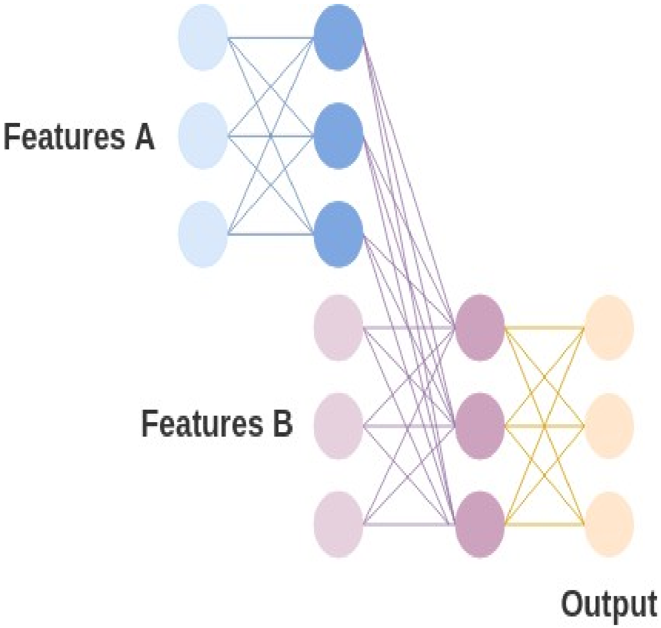
Multimodal Classification

2 evaluated approaches:

a. Features concatenation

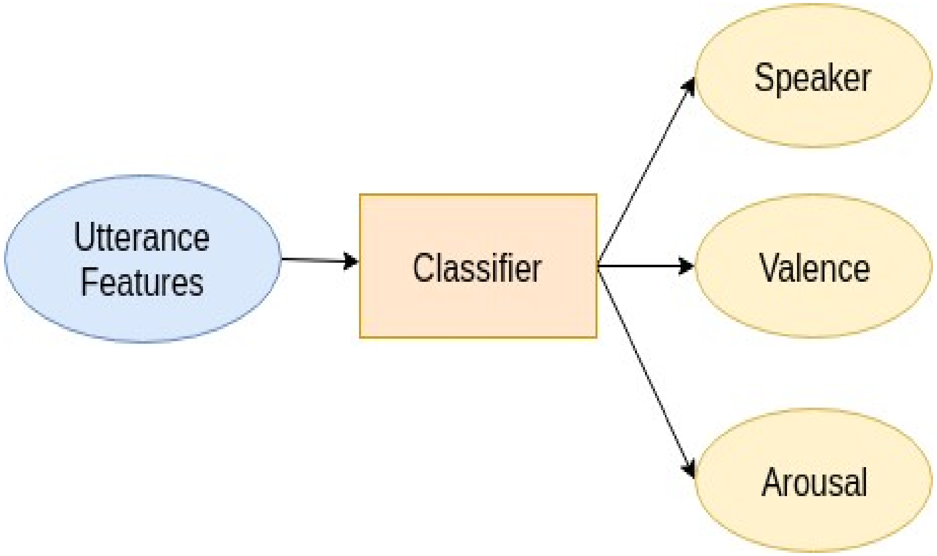


b. Features hierarchical fusion



III. Model Architectures
Multitask Classification

Perform classification on several tasks at once.

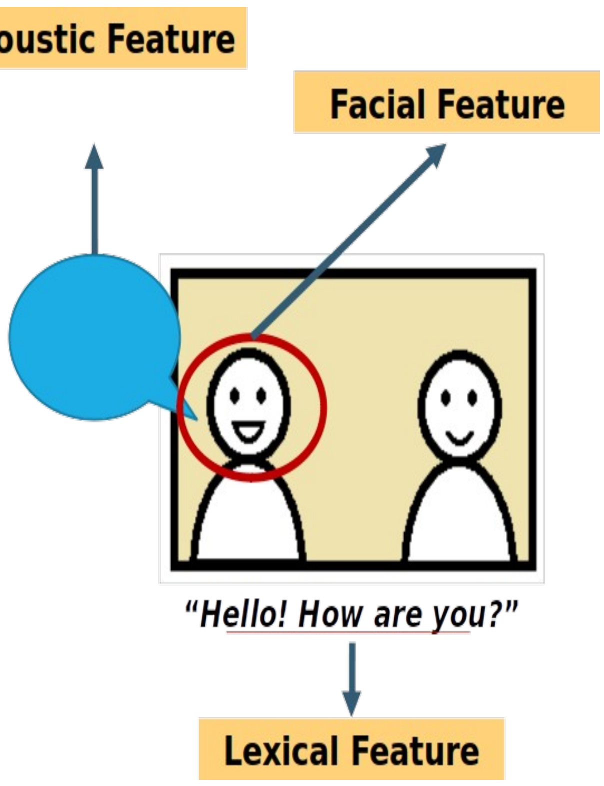


IV. Features

1. Acoustic (main)
   - INTERSPEECH 2010 feature conf.
   - openSMILE toolkit (Eyben et al., 2010)

2. Lexical
   - Word-vectors average
   - Pre-trained Google Word2Vec (Mikolov et al., 2013)

3. Facial
   - Facial contours and angles
   - openFace toolkit (Baltrusaitis et al., 2016)



V. Experiment

V. Experiment

- **Train set**: 2460 utterances
  - Speaker : 57 speaker (imbalanced)
  - Valence : Negative 31%, Neutral 60% , Positive 9%
  - Arousal : Negative 4%, Neutral 75% , Positive 21%

- **Evaluated** on 300 utterances
  - Speaker : 10 speaker, 30 samples each
  - Valence : Negative 32%, Neutral 57% , Positive 11%
  - Arousal : Negative 1%, Neutral 78% , Positive 21%

- Compared performance of unimodal, multimodal, single-task, and multitask models

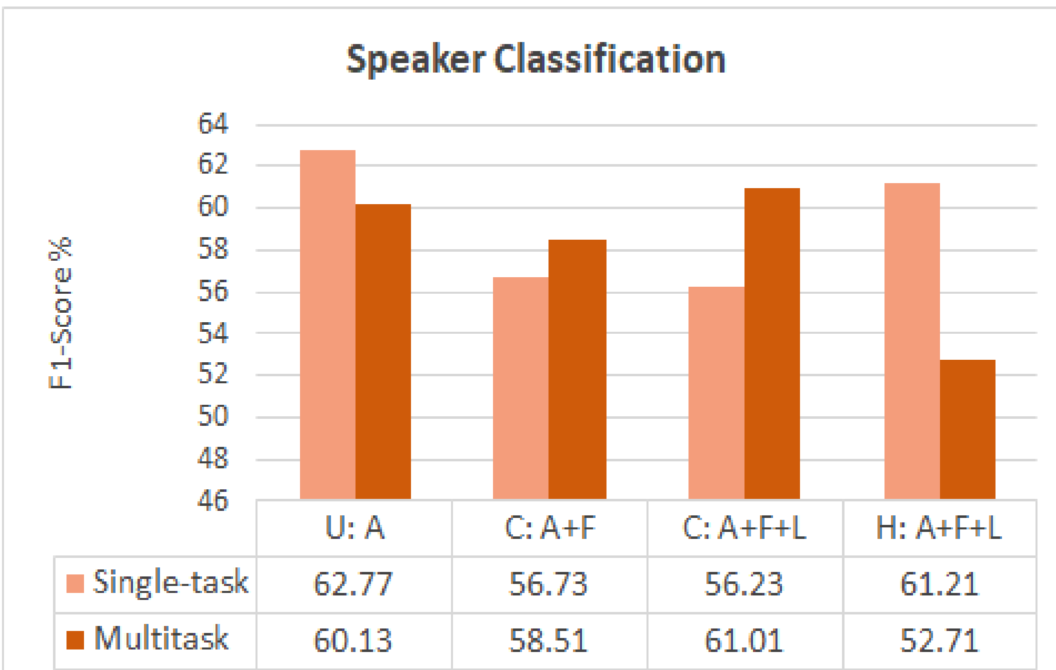
- Evaluated based on F1-score(%) on evaluation set

V. Experiment Result

V. Experiment

Result: Speaker

F1-scores (%) on evaluation set



	U: A	C: A+F	C: A+F+L	H: A+F+L
Single-task	62.77	56.73	56.23	61.21
Multitask	60.13	58.51	61.01	52.71

***Multimodal approaches**

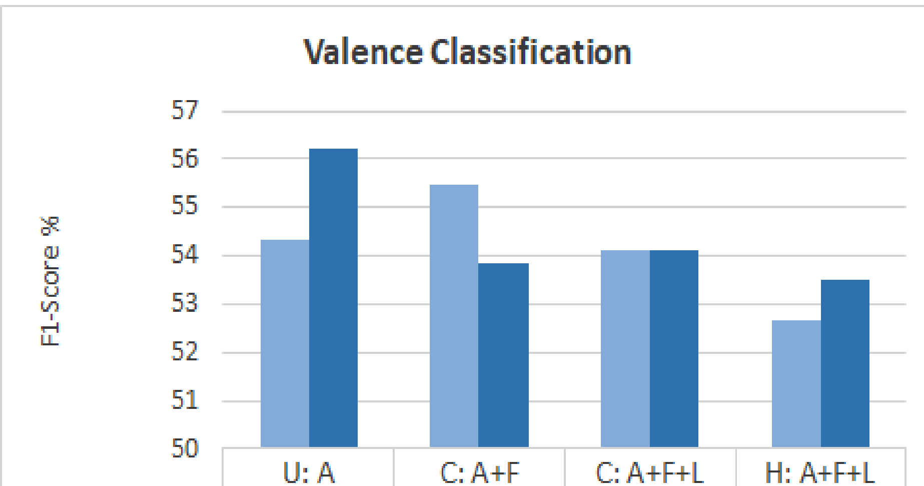
- U - Unimodal
- C - Features concatenation
- H - Features hierarchical fusion

Feature types

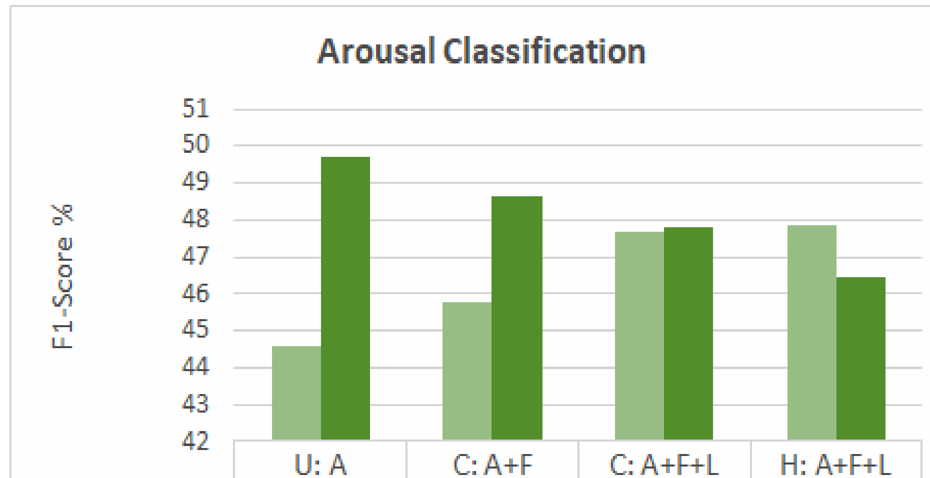
- A - Acoustic
- F - Facial
- L - Lexical

V. Experiment
Result: Emotion

F1-score (%) on evaluation set



	U: A	C: A+F	C: A+F+L	H: A+F+L
Single-task	54.31	55.45	54.1	52.65
Multitask	56.21	53.84	54.1	53.51

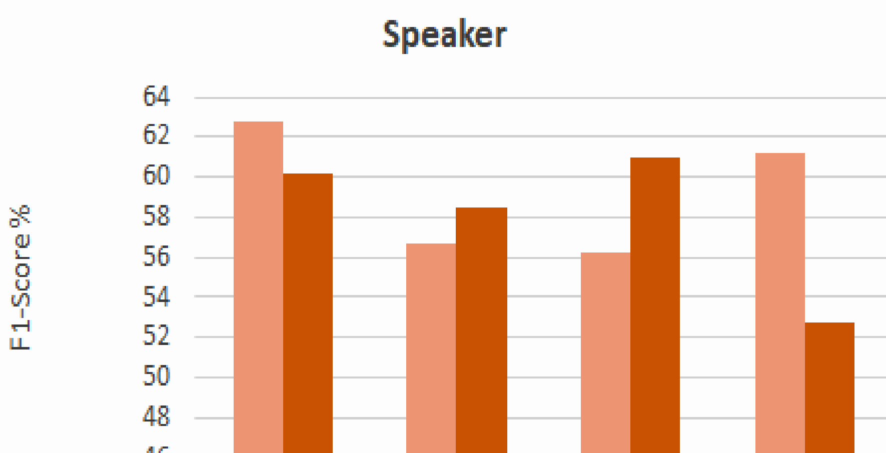


	U: A	C: A+F	C: A+F+L	H: A+F+L
Single-task	44.59	45.78	47.66	47.87
Multitask	49.71	48.66	47.77	46.44

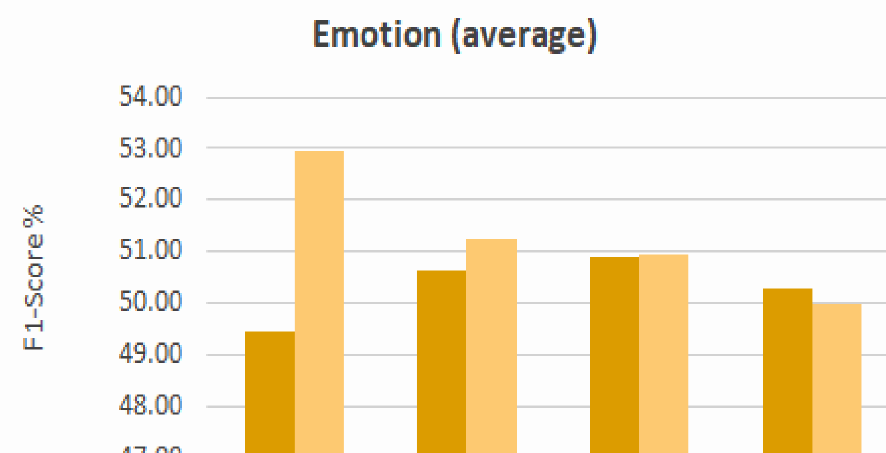
***Multimodal approaches**
U - Unimodal
C - Features concatenation
H - Features hierarchical fusion

Feature types
A - Acoustic
F - Facial
L - Lexical

V. Experiment
Result Summary



	U: A	C: A+F	C: A+F+L	H: A+F+L
Single-task	62.77	56.73	56.23	61.21
Multitask	60.13	58.51	61.01	52.71



	U: A	C: A+F	C: A+F+L	H: A+F+L
Single-task	49.45	50.62	50.88	50.26
Multitask	52.96	51.25	50.94	49.98

***Multimodal approaches**
U - Unimodal
C - Features concatenation
H - Features hierarchical fusion

Feature types
A - Acoustic
F - Facial
L - Lexical

VI. Conclusion

- We constructed the multimodal and multitask speaker-emotion recognition model by using deep learning and TV-series data
- Multitask model able to outperform single-task model, especially when recognizing emotion by using acoustic features only
- Multimodal-multitask model did not result in a significant improvement (larger data might be needed)

Thank You