JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY
1990

PhD Thesis Defense

Dimensional Speech Emotion Recognition by Fusing Acoustic and Linguistic Information

Bagus Tris Atmaja

1 December 2020

Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology

Information Science

Outline

1. Introduction: background, aims, contributions, applications
2. Research Methodology:
   - Motivation
   - Problems
   - Concept
   - Strategy
   - Datasets and evaluation metric
3. Dimensional SER using acoustic features
4. Early Fusion of Acoustic and Linguistic Information
5. Late Fusion of Acoustic and Linguistic Information
6. Conclusions: summary, future research directions

1. INTRODUCTION

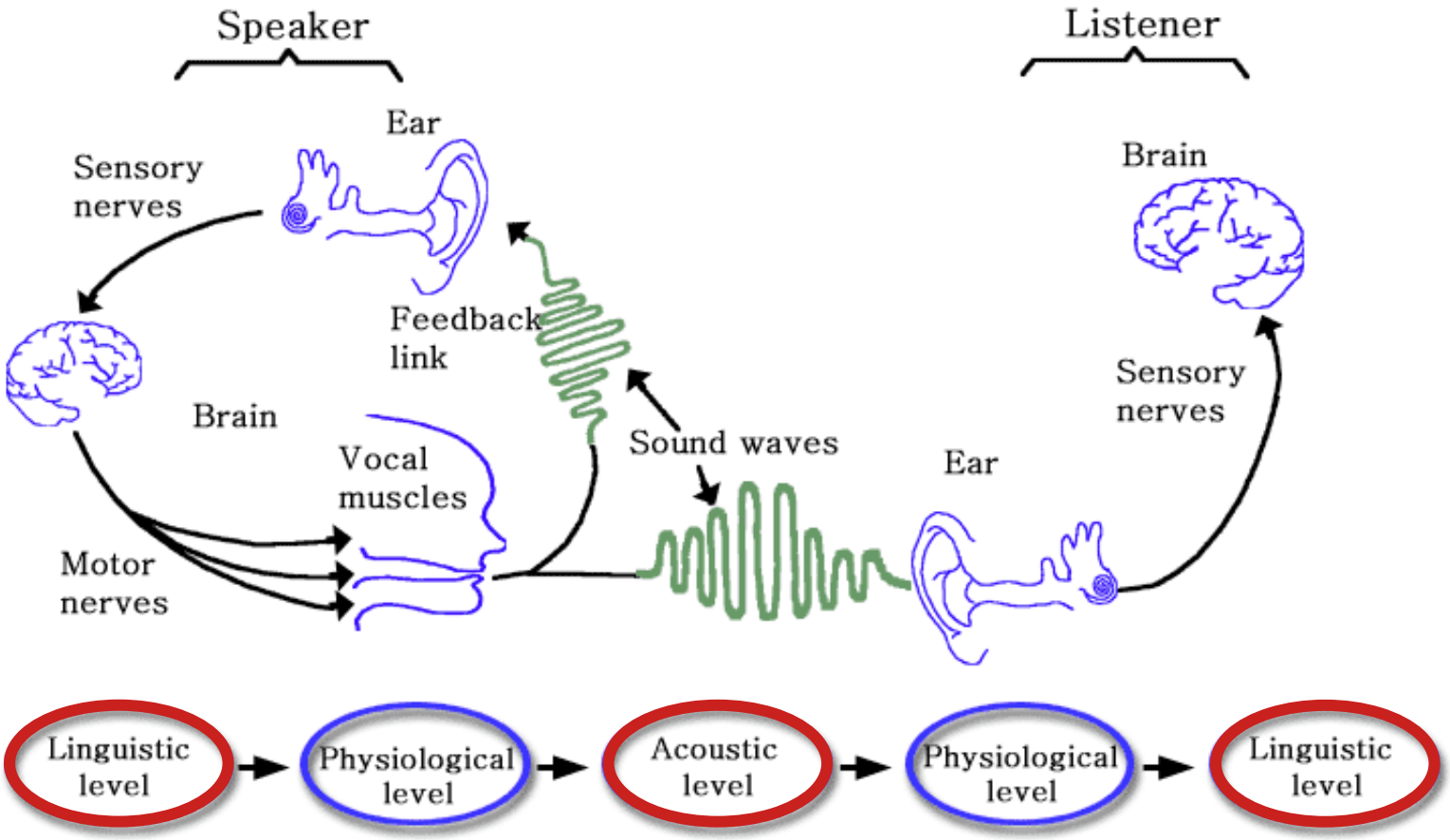
Background

Research aims

Contributions

Applications

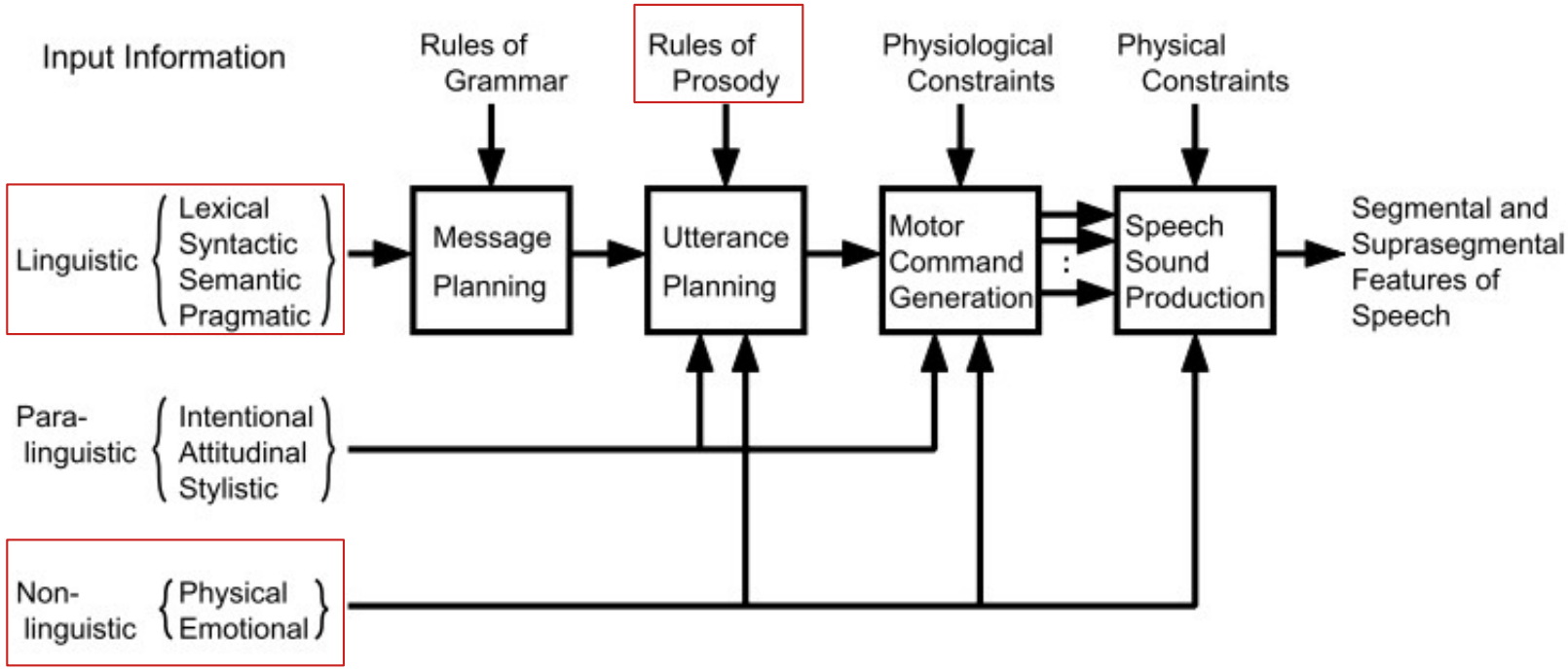
The Speech chain (Denes and Pinson, 1993)



Acoustic and linguistic are connected by physiological function; linguistic information may contribute to expressive speech aside from acoustic information.

Speech information

Information manifested in speech (Fujisaki, 2003)



Emotion information is embedded in speech with other information including linguistic

Research aims

- The goal of this research is *to investigate the necessity of fusing linguistic information with acoustic information for dimensional speech emotion recognition (SER)*.
- To achieve this goal, three sub-goals were addressed:
  1) Maximizing the potency of acoustic-only SER
  2) Fusing acoustic and linguistic information at feature level (early fusion)
  3) Fusing acoustic and linguistic information at decision level (late fusion)

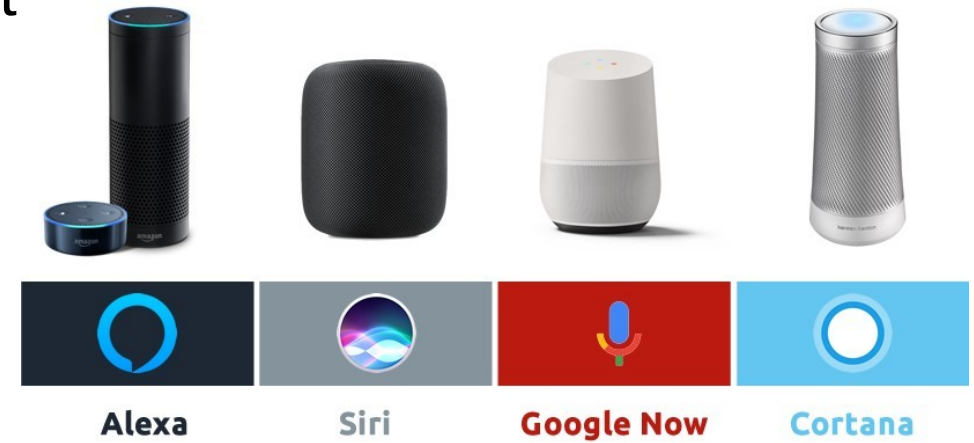
Research contributions

1) Acoustic feature extraction
   - Evaluation of Mean+Std from LLD for dimensional SER
   - Contribution of silent pause region in dimensional SER
   - Acoustic feature aggregation
2) Information fusion
   - Early fusion acoustic-linguistic information fusion for dimensional SER
   - Late fusion acoustic-linguistic information fusion for dimensional SER
   - Contribution of linguistic information for valence performance improvement
3) Classification methods
   - Deep multilayer perceptron (MLP) for dimensional SER
   - Correlation-based loss functions
   - Multitask learning for dimensional SER

Possible applications

- Call center applications
  - Emotion of caller
  - Emotion of operator

- Voice assistant



- Other speech-based technologies (voice message, voice mail, etc.)

2. RESEARCH METHODOLOGY

Motivations

Issues

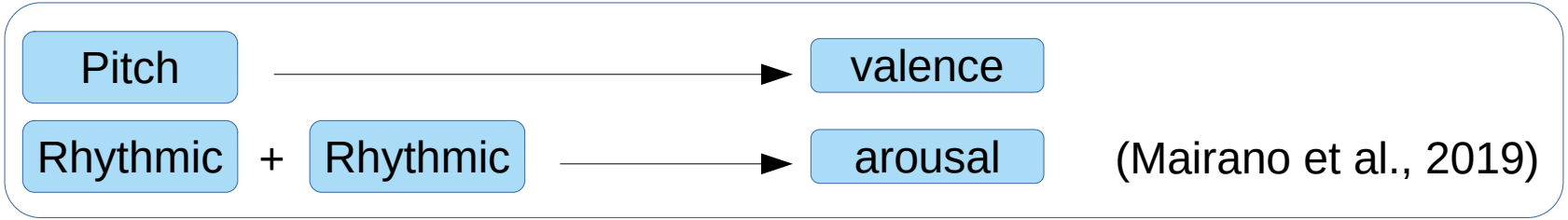
Philosophy

Strategies

Datasets and Evaluation Metric

Motivation

- Why researching SER
  - In some cases, only speech data could be obtained.
  - There is strong correlation between speech and emotion:



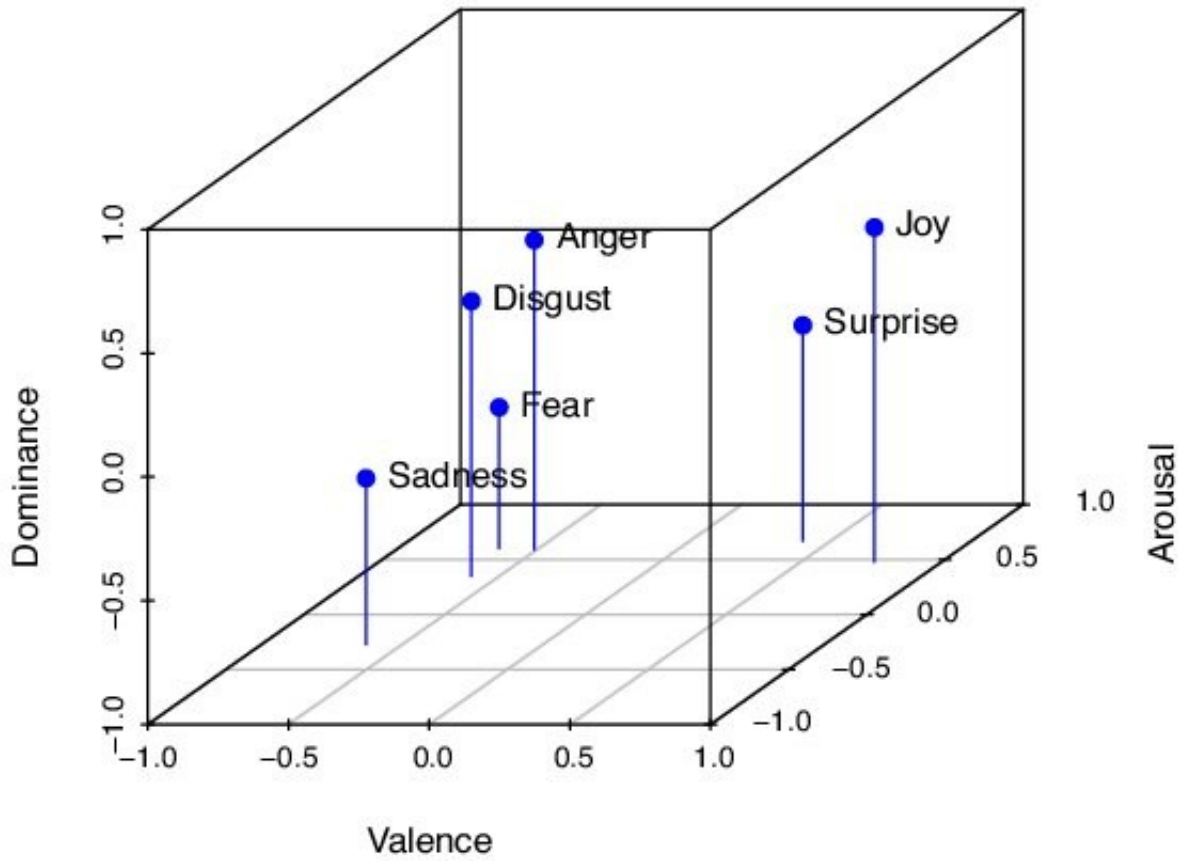
- Why researching SER is difficult
  - The labels are given by annotators; no exact values (cf. digits).

IEMOCAP ID: Ses01F_Impro01_F001

Annotators	V	A	D
Annot. #1	3	2	2
Annot. #2	2	3	3
Annot. #3	2	3	2

Motivation

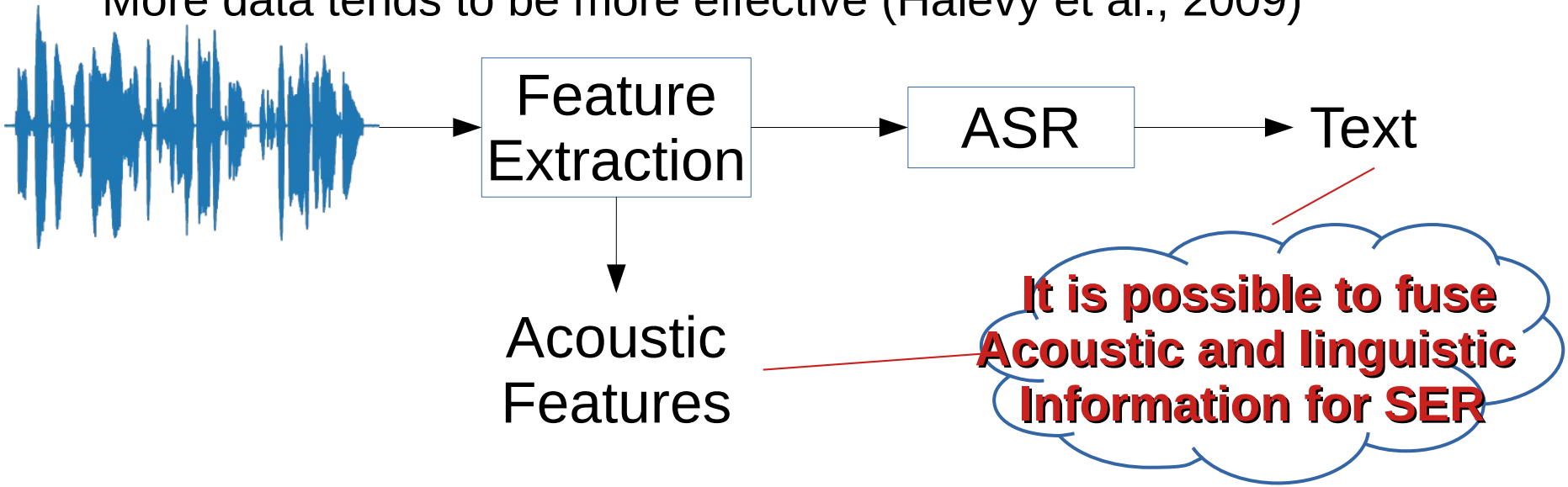
- Why dimensional SER
  - Human's variability is high; hence, categorization doesn't have an essence.
  - Categorical emotion is not enough to describe affective state
  - Most previous SER works only focus on categorical emotion



VAD model with Ekman's basic emotion (Buechel and Hahn, 2016)

Motivation

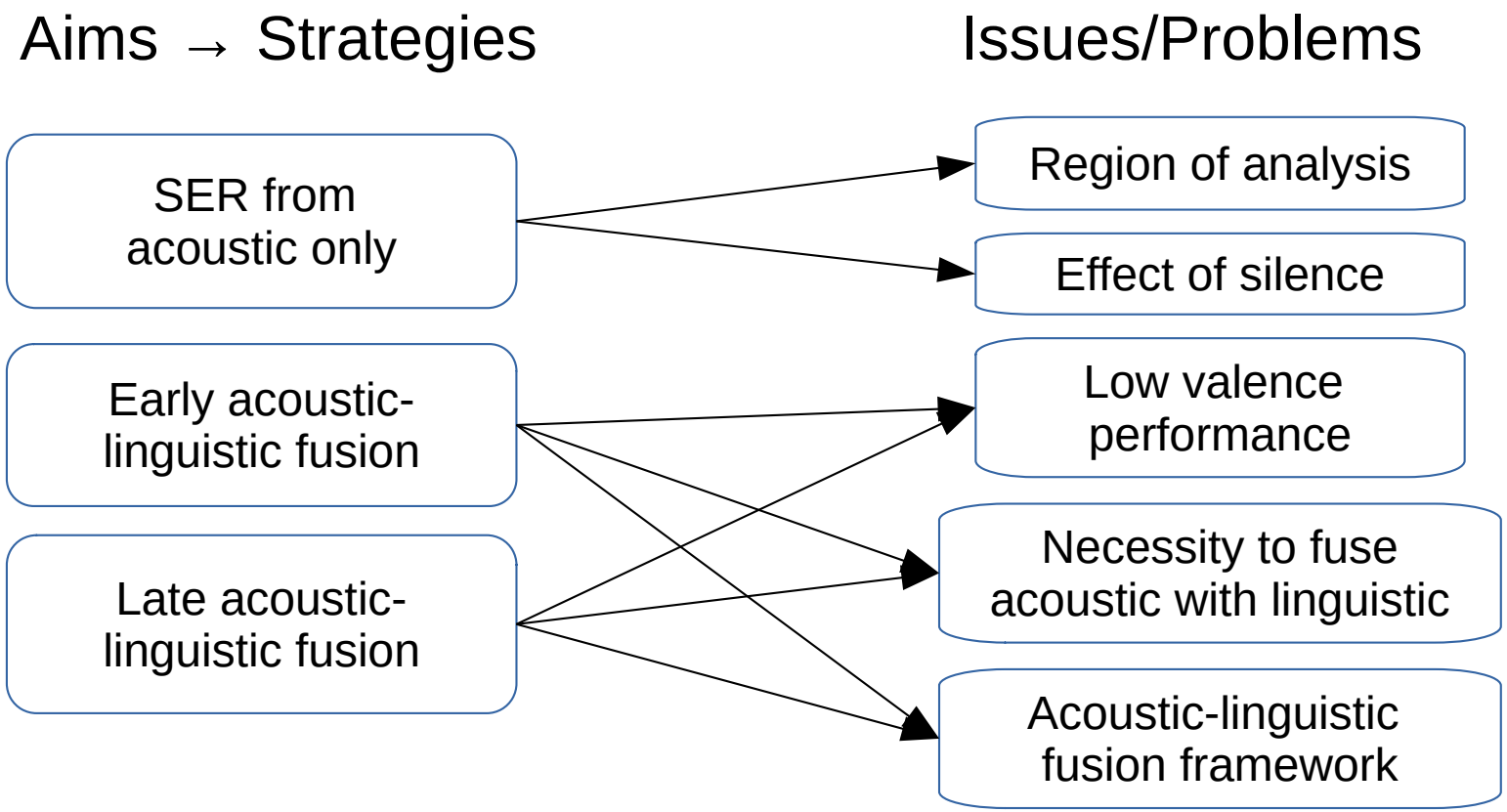
- Why fusing acoustic with linguistic information
  - Speech can be transcribed into text using speech-to-text system
  - Linguistic information can be extracted from transcription
  - Linguistic information is also related to human perceived emotion (humans may also use linguistic cues to perceive emotion)
  - More data tends to be more effective (Halevy et al., 2009)



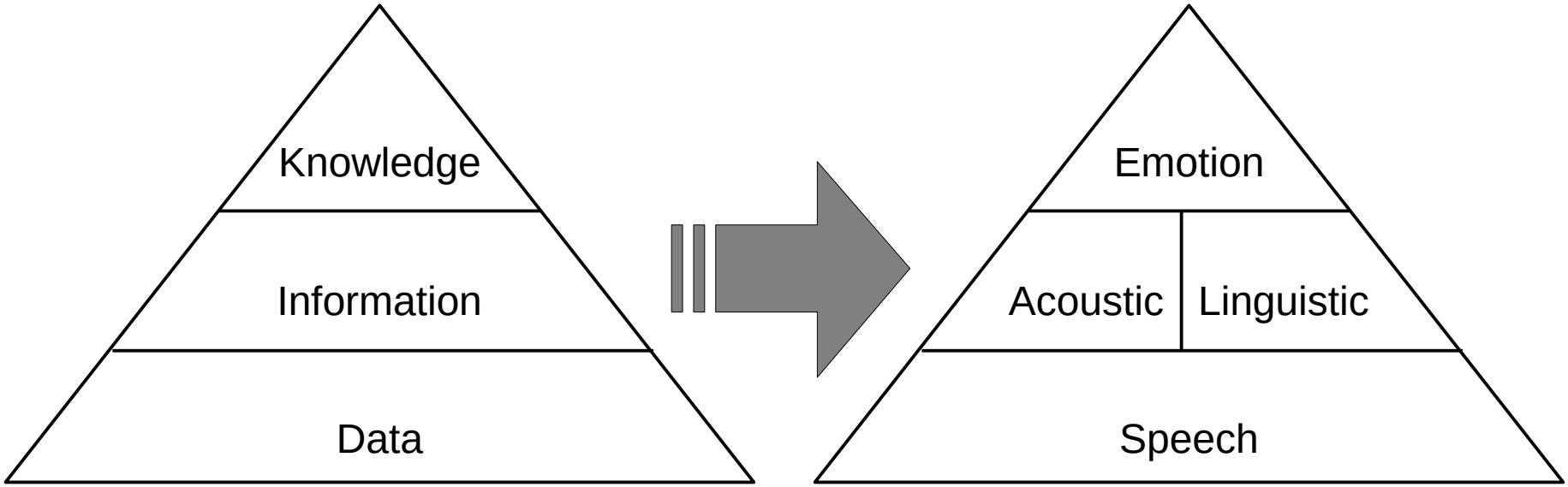
Research Issues

1. Which region of analysis to extract acoustic features for SER
2. The effect of silent pause regions in dimensional SER
3. Low valence prediction performance in dimensional SER
4. The necessity to fuse linguistic information with acoustic information
5. The fusion framework for combining acoustic and acoustic information

Correlation between aims and issues



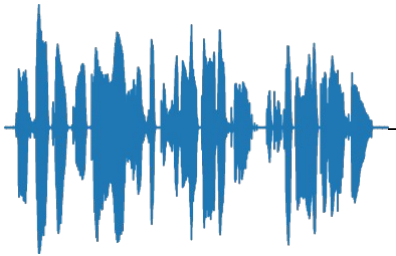
Concept/Philosophy



(acoustic and linguistic) Information is extracted from data (i.e., speech); knowledge (emotion) is extracted from (acoustic and linguistic) information.

Data: Speech

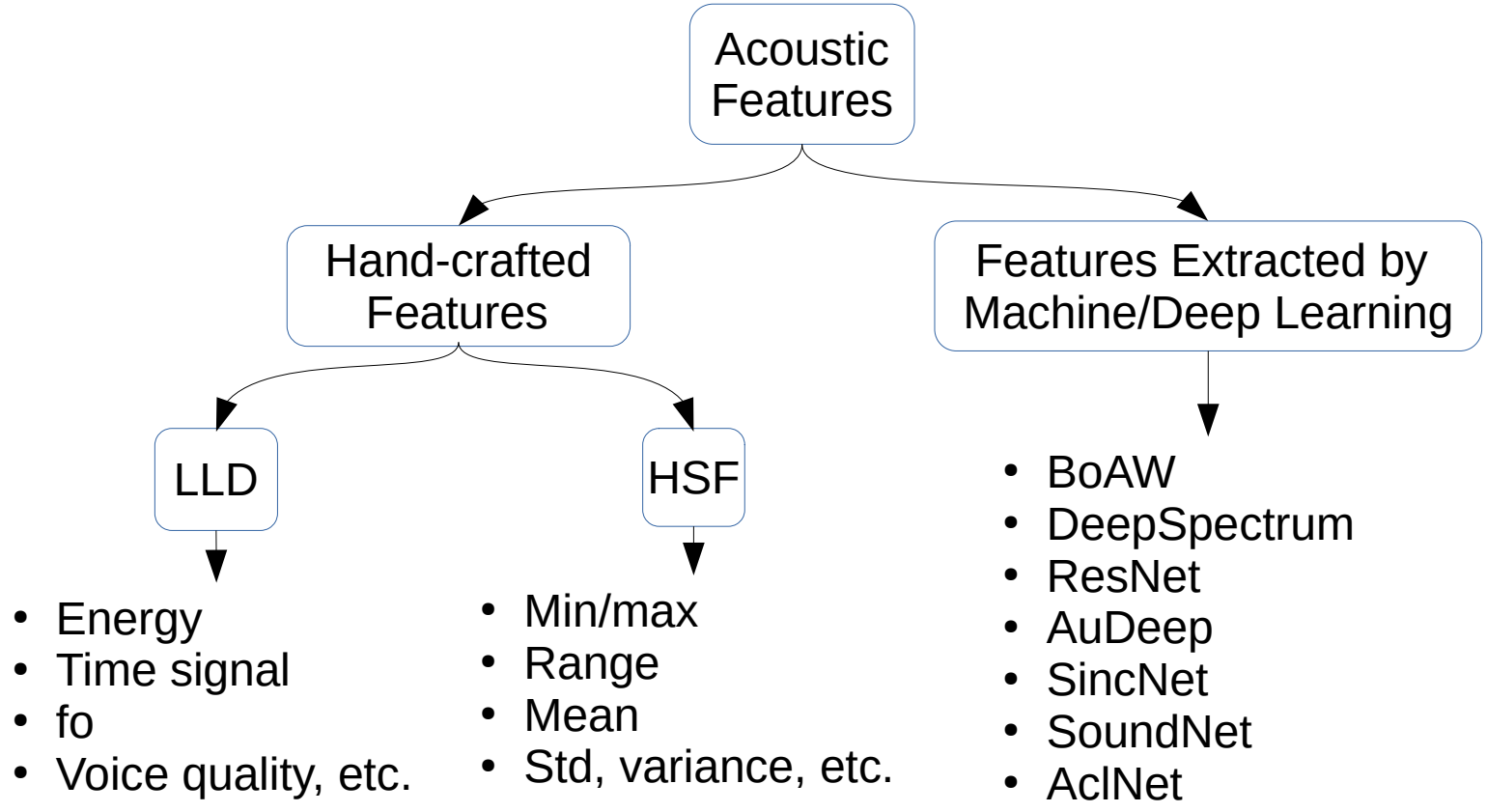
- Speech: the expression of or the ability ***to express thoughts and feelings by articulate sounds***.



Text	“That’s out of control”
Language	English
Accent	English US (NA)
Sex	Female
Speaker	speaker #2
Age	25 – 54 yo
Emotion	angry (V: 1, A:4, D:4)

Information: acoustic

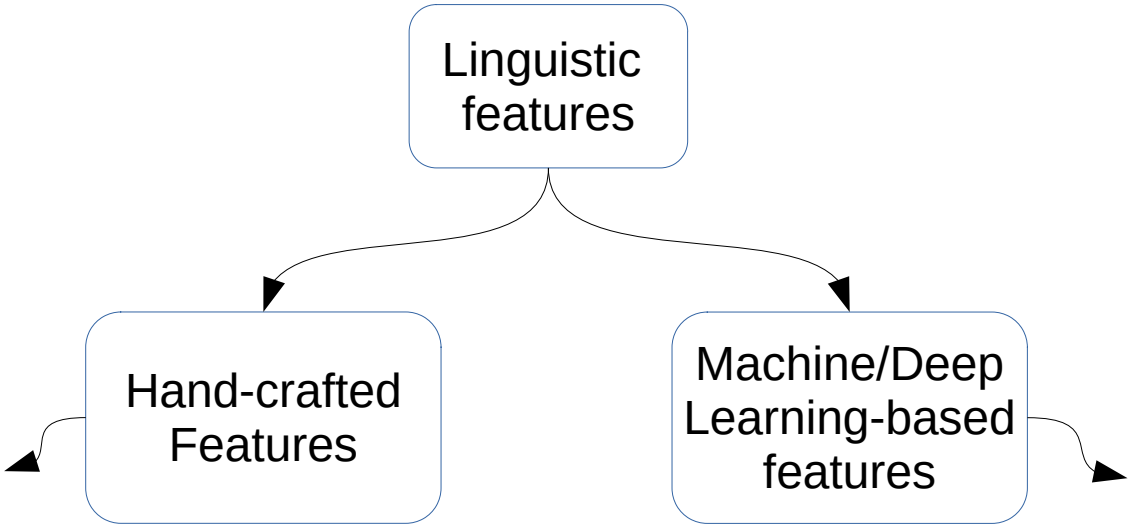
- Acoustic is the main information to perceive emotion in speech
- Conceptual information in practice is implemented as ***features***



Information: linguistic

Linguistic can be regarded as additional information (cue) to perceive emotion

- Emotional words and phrase
- Exclamation, negation, word number
- TF-IDF
- BoW
- LSA
- Lexicon-based features (ANEW, VADER, etc)
- phonemes, etc.



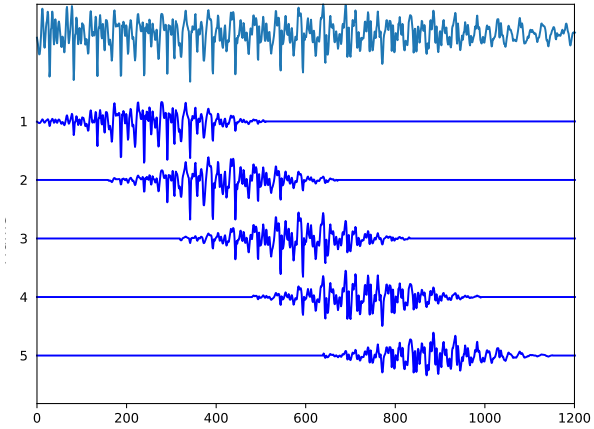
- wor2vec
- sentence2vec
- doc2vec
- ELMo
- GloVe
- FastText
- BERT
- LiFE

Knowledge: Emotion

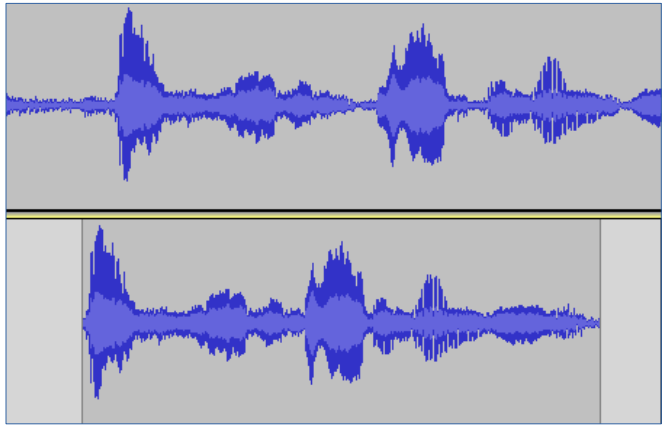
- Emotion is episode of interrelated, synchronized changes in the states of all or most of the five organismic subsystems in response to the evaluation of an external or internal stimulus event as relevant to major concerns of the organism (Scherer, 1987)
- Emotion as knowledge → emotional knowledge : one's ability to define and label emotions in oneself and in others (Rossi, 2016)

Research Strategy

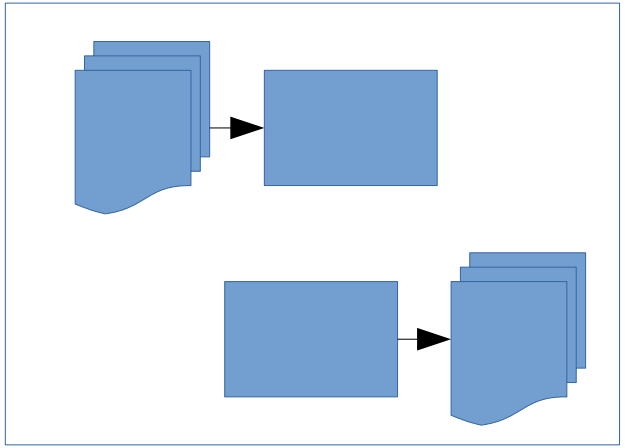
- Dimensional SER by acoustic information
  - Which region of analysis to extract acoustic features
  - Effect of silent pause features
  - Aggregation methods for chunks to an utterance



LLD vs. HSF



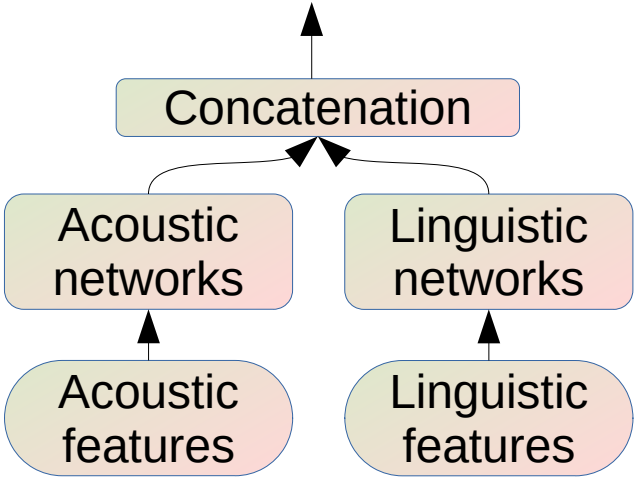
Keeping vs. Removing vs. Using silence



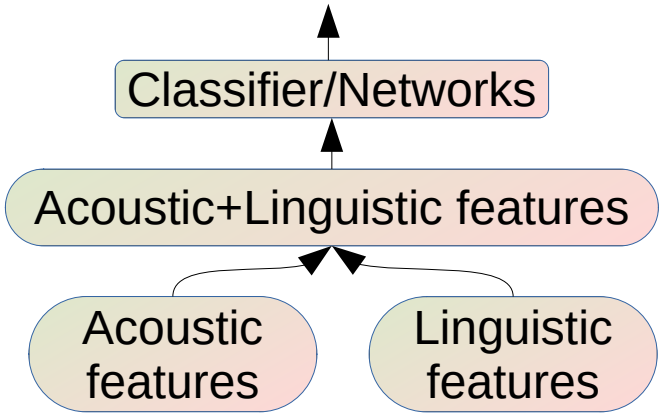
Input vs. output aggregation

Research Strategy

- Early acoustic-linguistic fusion (feature level [FL])
  - Word embeddings
  - Early fusion by networks concatenation
  - Early fusion by feature concatenation
  - Using ASR outputs for linguistic input



Network Concatenation



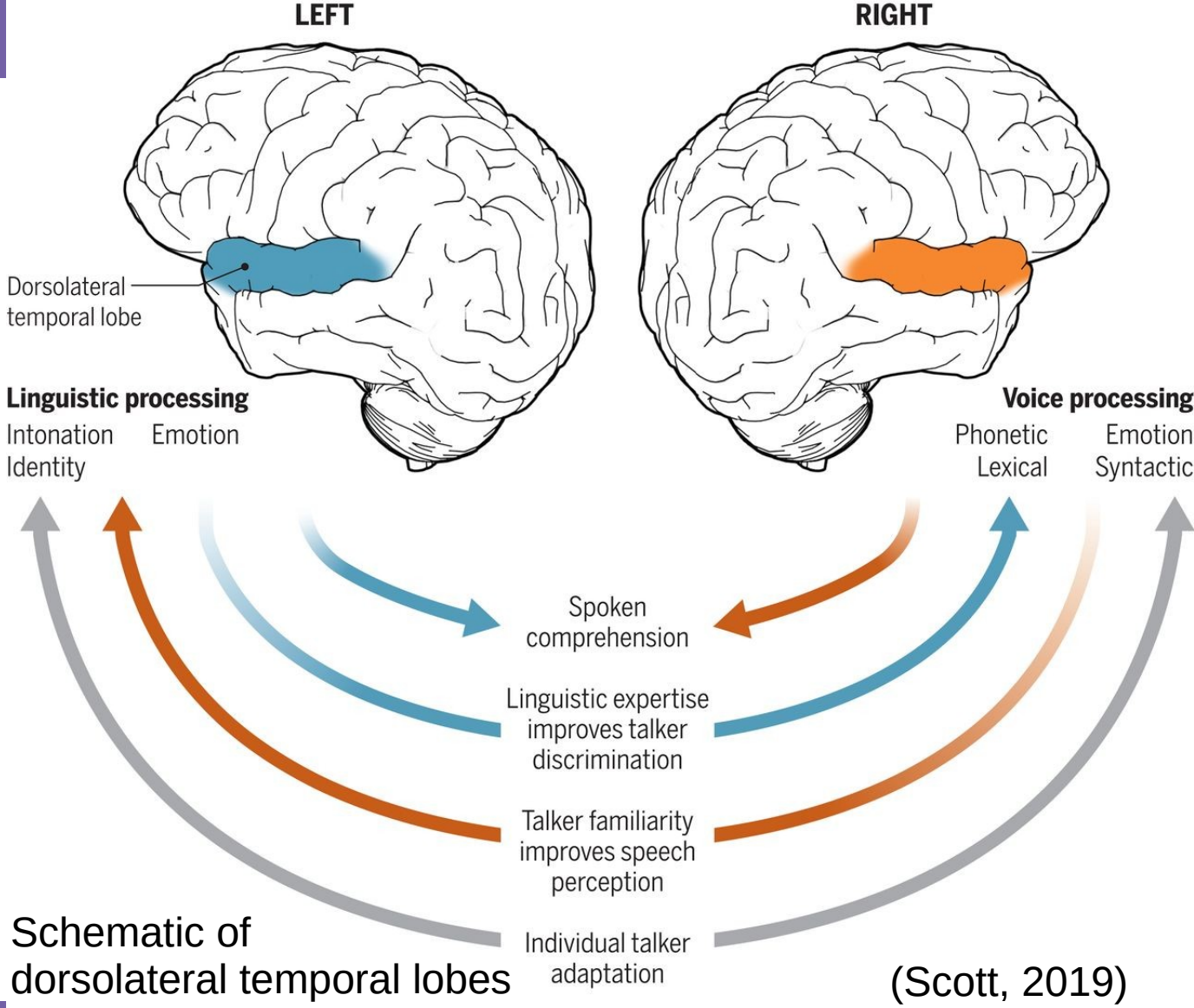
Features Concatenation

Research Strategy

- Late acoustic-linguistic fusion (decision level [DL]):
  - Datasets partitions
  - Two-stage dimensional SER
    - DNNs
    - SVM
  - Results
    - Result of two-stage processing
    - Speaker dependent vs. Speaker independent
    - Effect of removing target sentences

Why late fusion?

- Physiological evidence showed that linguistic and acoustic information are processed separately in different region of brain.
- Thus, decision level fusion by combining results from each modality may be better than feature level fusion.



Schematic of dorsolateral temporal lobes

(Scott, 2019)

Datasets

IEMOCAP

12 hours long
10039 turns
10 speakers
5 sessions
V, A, D [1-5]

MSP-IMPROV

> 9 hours long
8438 turns
12 speakers
6 sessions
V, A, D [1-5]

USOMS-e

261 stories
7778 chunks
87 speakers
V, A [L, M, H]

Evaluation metric

- Concordance correlation metric (CCC)

$$CCC = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2}$$

  - A step further than (Pearson) correlation coefficient
  - Penalises any deviation from the identity relationship (both scale and location/shift)
  - Captures both accuracy and precision
  - Mathematically and experimentally superior to error-based loss functions (Pandit and Schuller, 2020; Atmaja and Akagi, 2020)

3. SER Using Acoustic Features

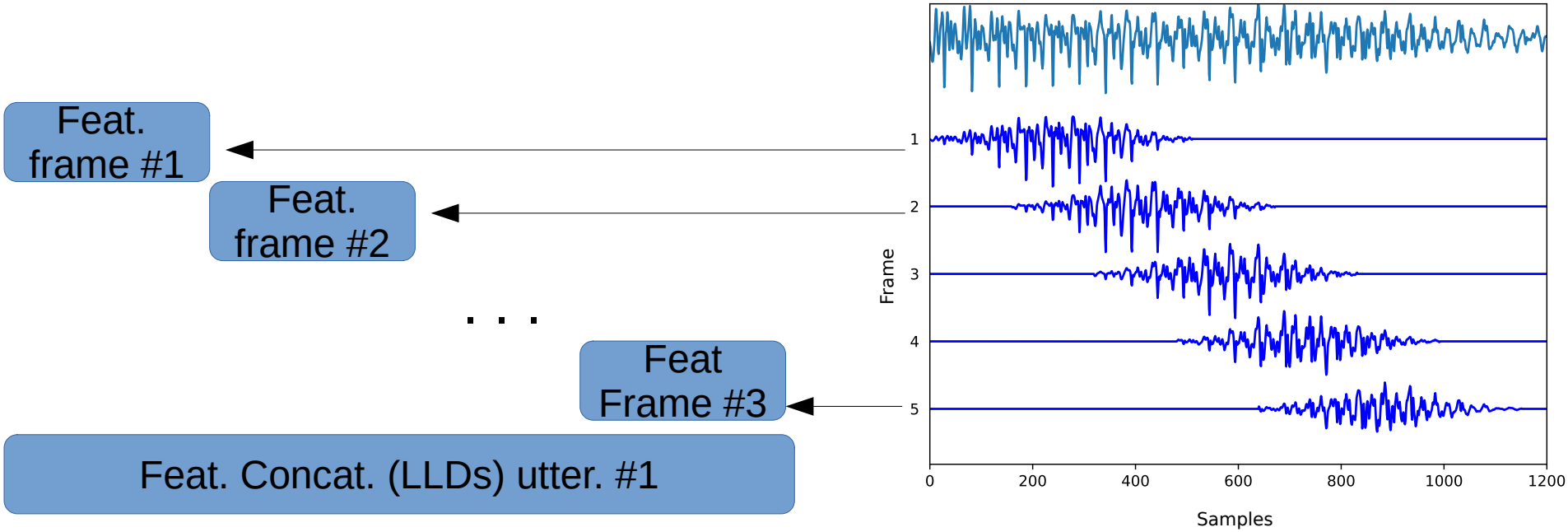
Which region of analysis to extract acoustic features?

Effect of silent pause regions

Acoustic feature aggregation

Which region of analysis to extract features: LLD

- Conventional methods divide speech signal into frames and apply feature extraction on these frames
- The acoustic features from all frames are concatenated (added with zeros if needed) to obtain information for a whole utterance



High-level statistical function (HSF)

4. EARLY FUSION OF ACOUSTIC AND LINGUISTIC INFORMATION (FEATURE LEVEL)

5. LATE FUSION OF ACOUSTIC AND LINGUISTIC INFORMATION (DECISION LEVEL)

6. CONCLUSIONS

General summary

Future research directions

Publications

- Journals (3):
  - B. T. Atmaja and M. Akagi, “Dimensional speech emotion recognition from speech features and word embeddings by using multitask learning,” APSIPA Trans. Signal Inf. Process., vol. 9, May 2020.
  - R. Elbarougy, B.T. Atmaja and M. Akagi, “Continuous Audiovisual Emotion Recognition Using Feature Selection and LSTM,” Journal of Signal Processing, Vol. 24, No. 6, November 2020
  - B.T. Atmaja, and M. Akagi. “Two-stage dimensional emotion recognition by fusing predictions of acoustic and text networks using SVM,” Speech Communication (Accepted, to appear).
- International conferences: 10 (ICASSP, APSIPA ASC, OCOCOSDA)
- Domestic conferences: 4 (ASJ meetings)