Dimensional Speech Emotion Recognition by Fusing Acoustic and Linguistic Information

Bagus Tris Atmaja　（バグストリスアトマジャ）

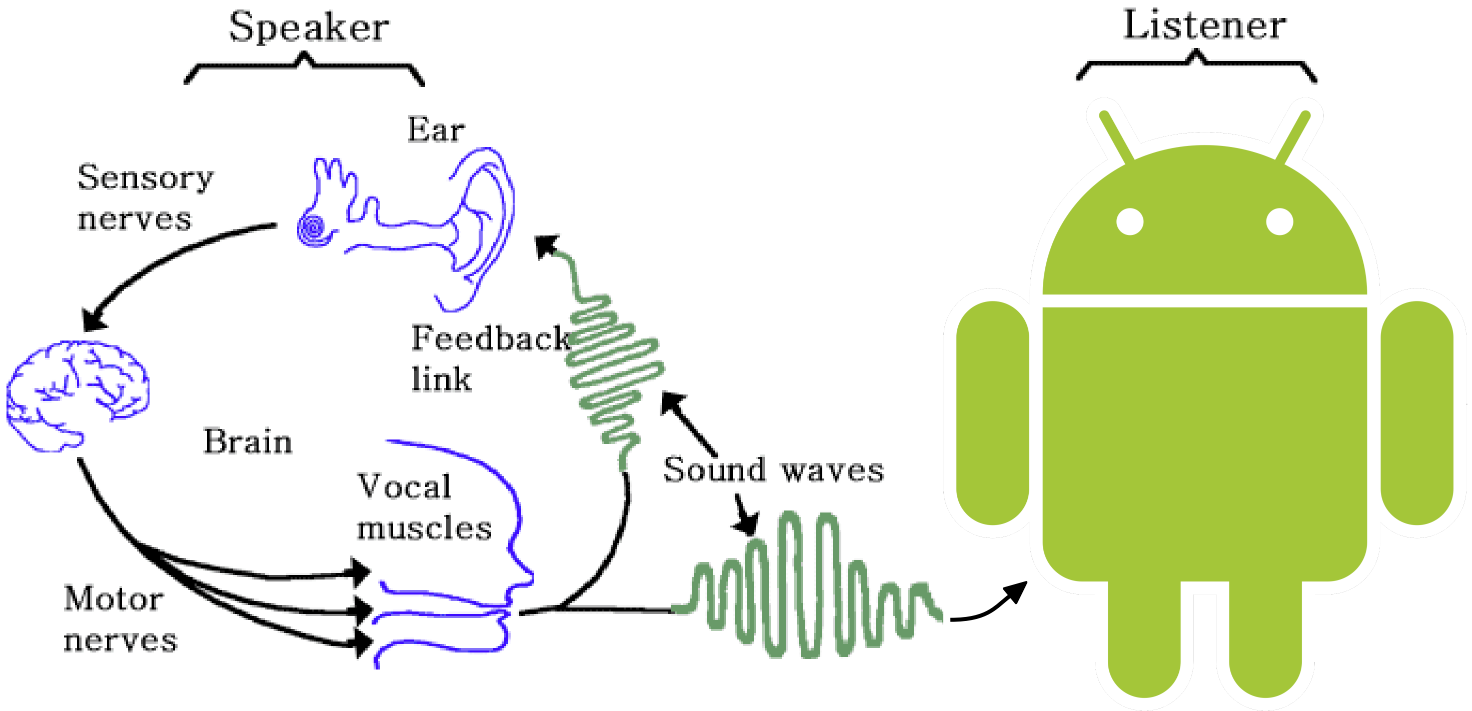
JAIST-RIEC Meeting
26 February 2021

北陸先端科学技術大学院大学
情報科学系　音情報処理分野　赤木・鵜木研究室

Outline

1. **Introduction:**
   Background, Aims & Issues, Applications
2. Research Methodology:
   Motivation, Previous work, Concept, Strategy, Datasets, 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:
   Comparative analysis, Summary, Contributions, Future research

Human-machine communication

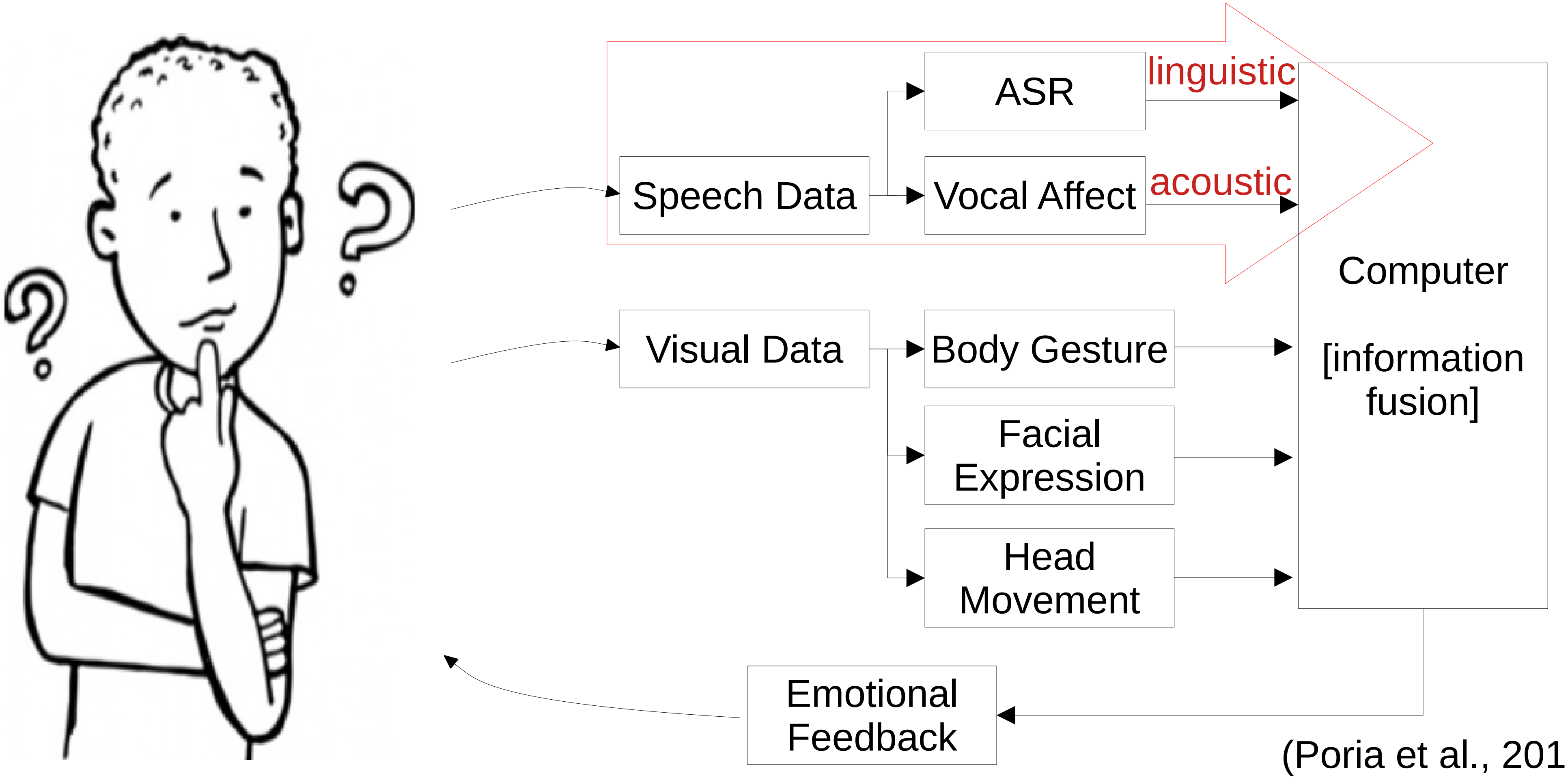


Speech Emotion Recognition

In **speech chain**, acoustic and linguistic are connected by physiological function; fusing both information may improve emotion recognition rate by **machine**

Multimodal affective computing

Affective computing: *computing that relates to, arises from, or influences emotion (Picard, 1995)*



(Poria et al., 2017)

Research aims

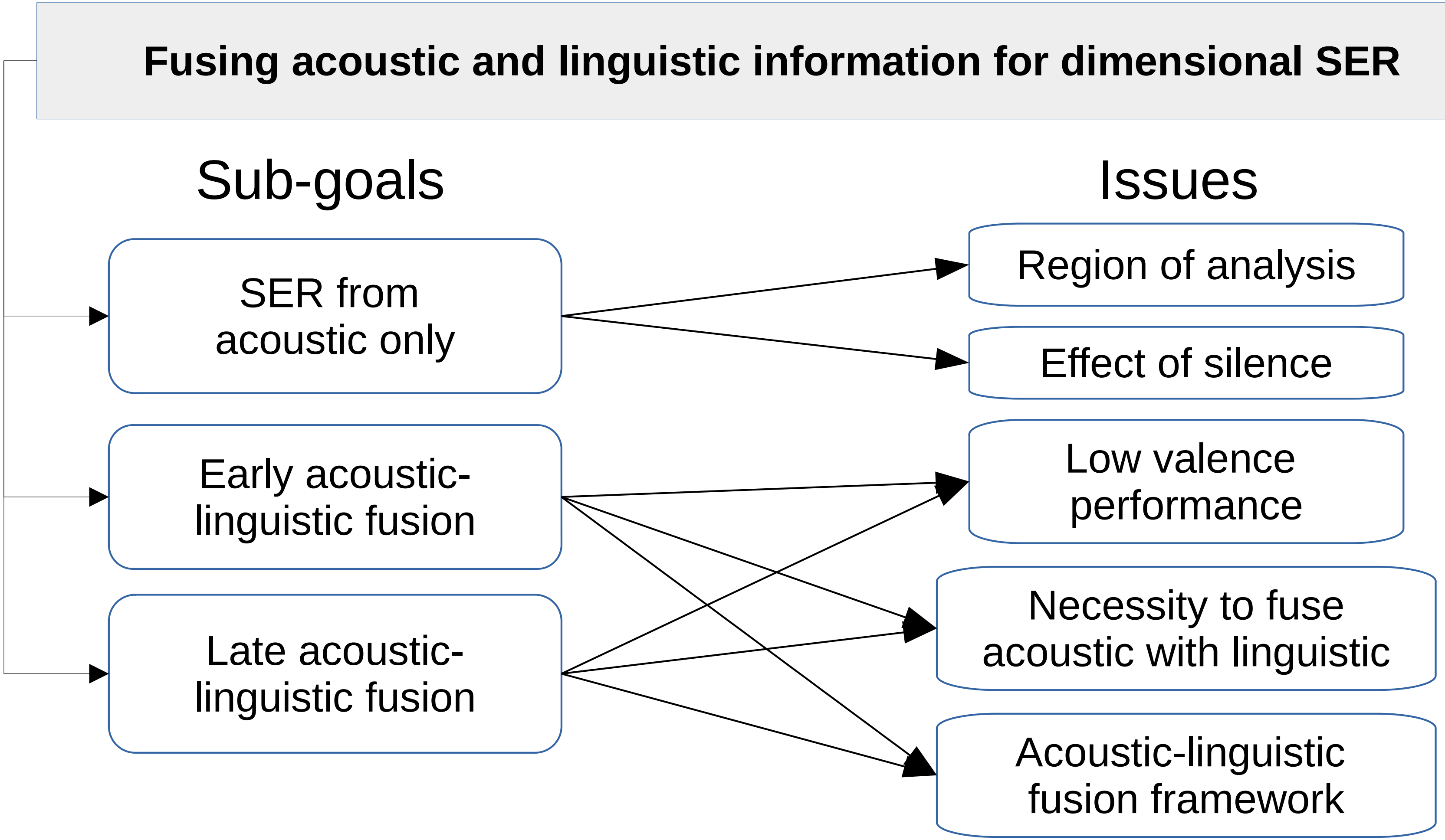
- The goal of this research is *to investigate the necessity of **fusing acoustic information with linguistic information** for dimensional speech emotion recognition (SER)*
- To achieve this goal, three sub-goals were addressed:
  1) Maximizing the potency of **acoustic-only** SER
  2) Fusing acoustic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

Research issues

1. Which region of analysis to extract acoustic features for SER (El-Ayadi, 2011)
2. The effect of post processing in SER (El-Ayadi, 2011)
3. Low valence prediction performance in dimensional SER (Li, 2019; El-Barougy, 2013)
4. The necessity to fuse acoustic information with other modalities (El-Ayadi, 2011)
5. The fusion framework for fusing acoustic and linguistic information

Correlation between aims and issues

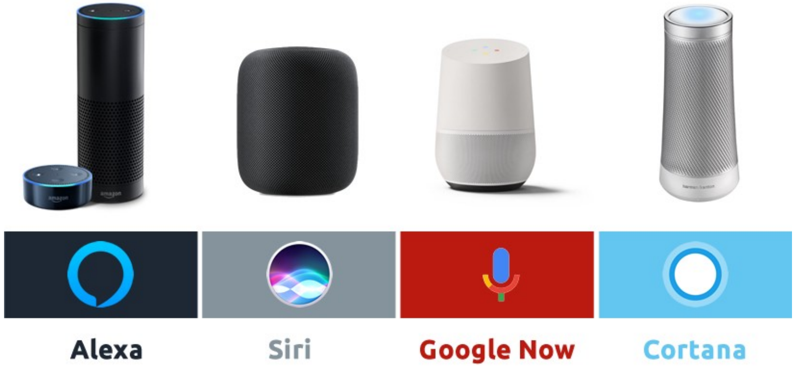
Main goal



Possible applications

- Contact/Call center application

- Voice assistant

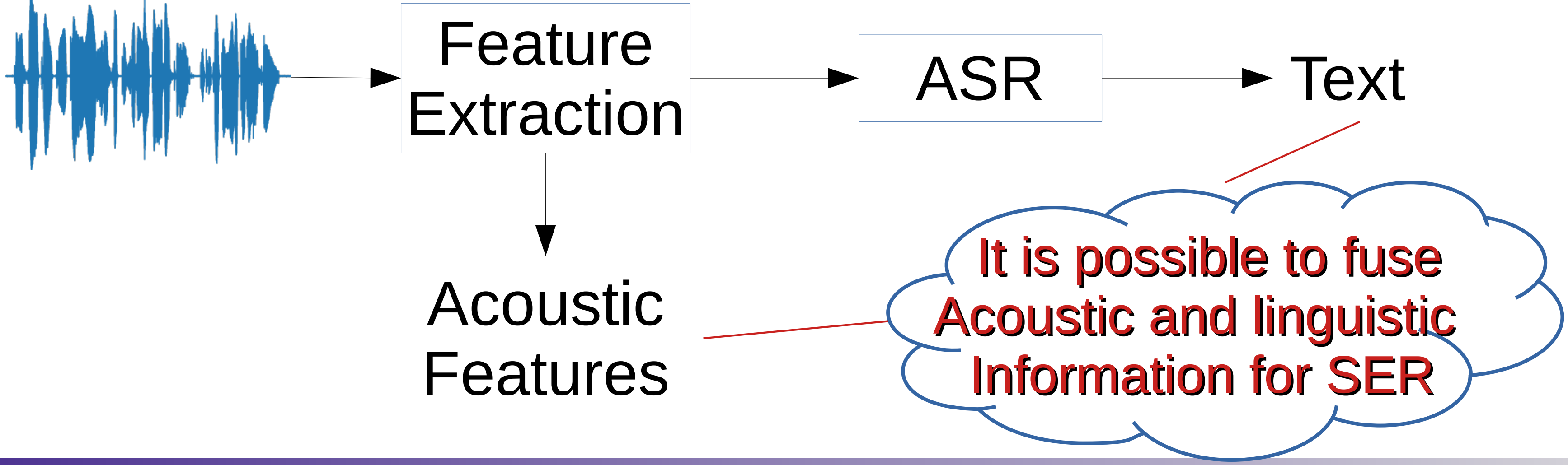


Outline

1. Introduction:
   Background, Aims & Issues, Applications
2. **Research Methodology:**
   Motivation, Previous work, Concept, Datasets, 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:
   Comparative analysis, Summary, Contributions, Future research

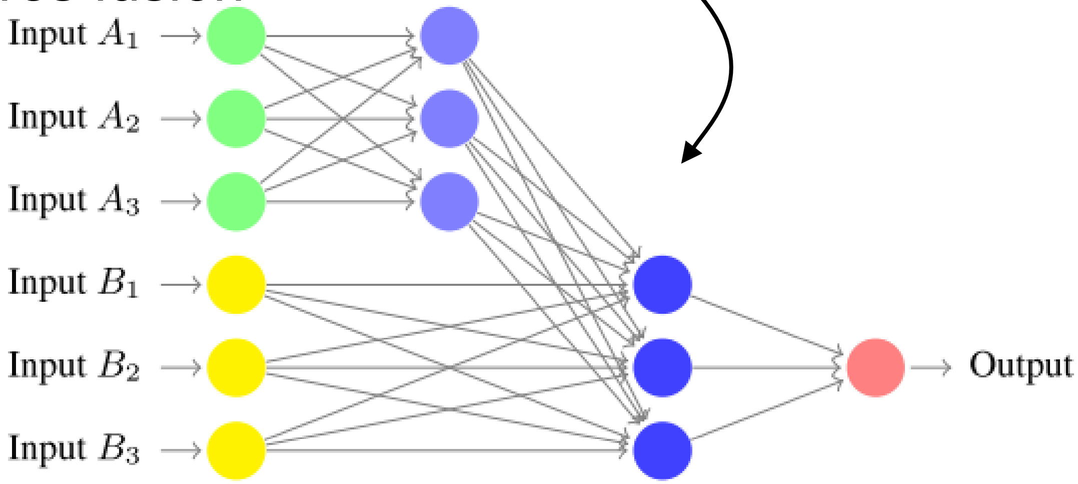
Motivation

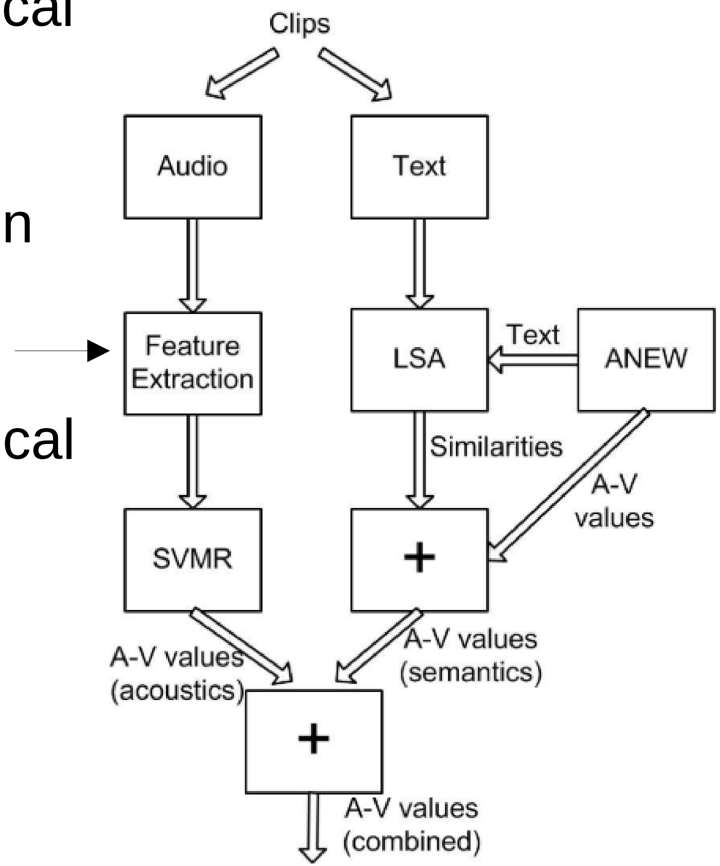
- Why fusing acoustic with linguistic information?
  - Speech can be transcribed into text using **A**utomatic **S**peech **R**ecognition (ASR)
  - Linguistic information can be extracted from transcription
  - Human communicate emotion through speech and language (Kotz et al., 2011)
  - More data tends to be more effective (Halevy et al., 2009)



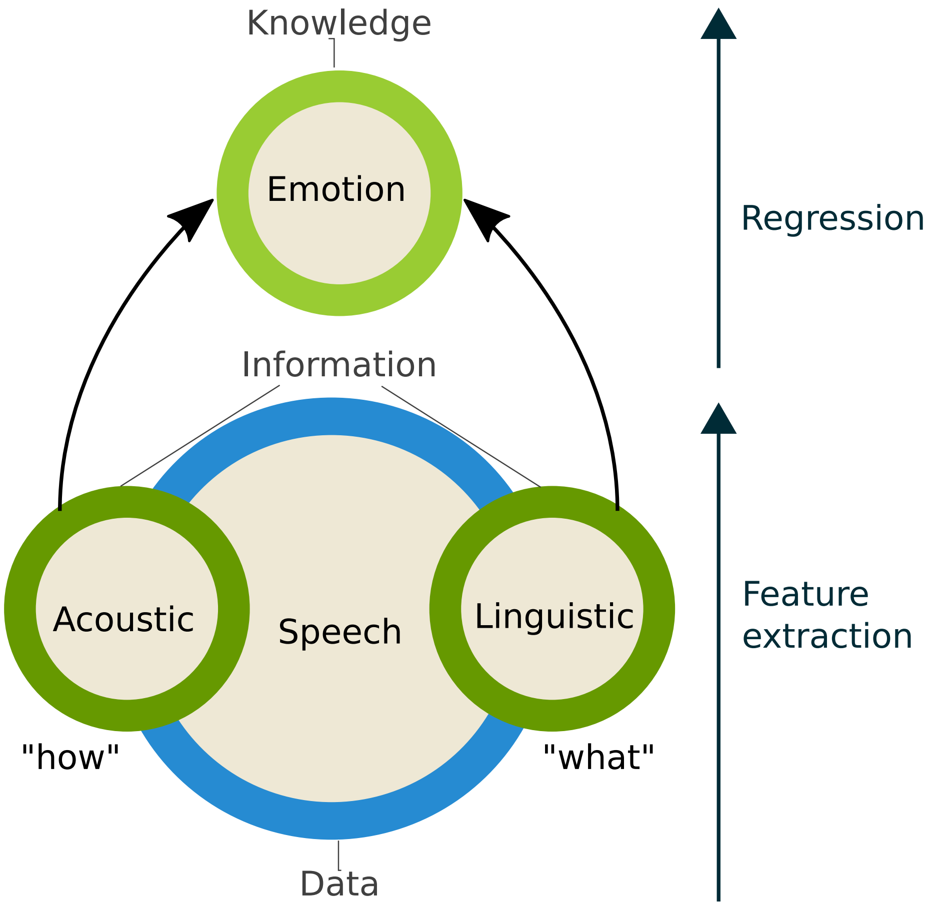
Previous work

- Lee et al. (2002): decision-based fusion using logical “OR” to predict negative/non-negative emotion by using acoustic features and spot keywords
- Karadogan & Larsen (2012): decision-based fusion using weighting function to fuse acoustic and semantic information
- Tian et al. (2016): hierarchical-based acoustic-lexical features fusion





Concept/Philosophy



“It is not only ***how*** things are said, but also ***what*** things are said”

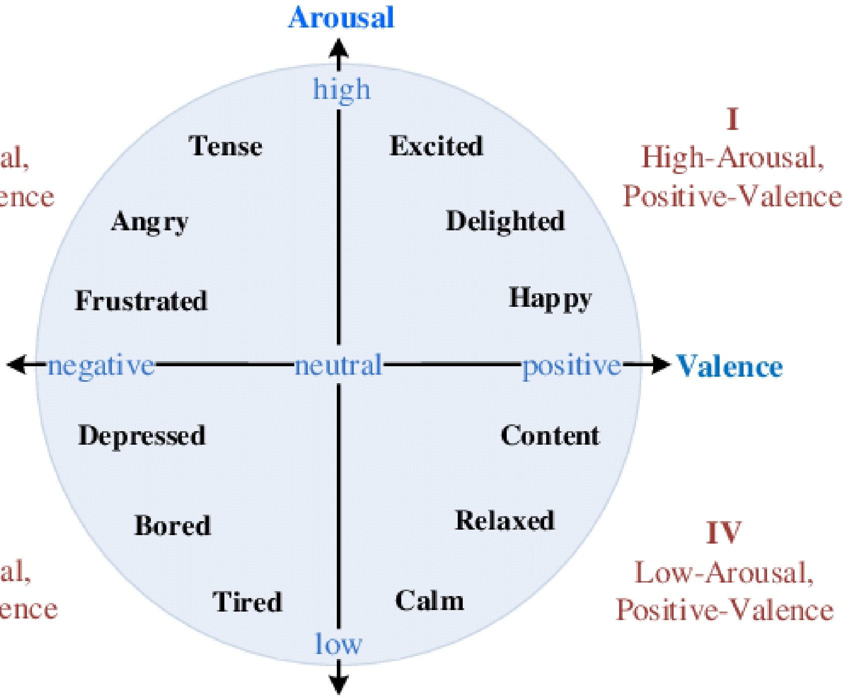
Emotion model

- Dimensional emotion: emotion as continuous degree in several attributes/dimensions
- Most common dimensions: Valence, Arousal, and Dominance

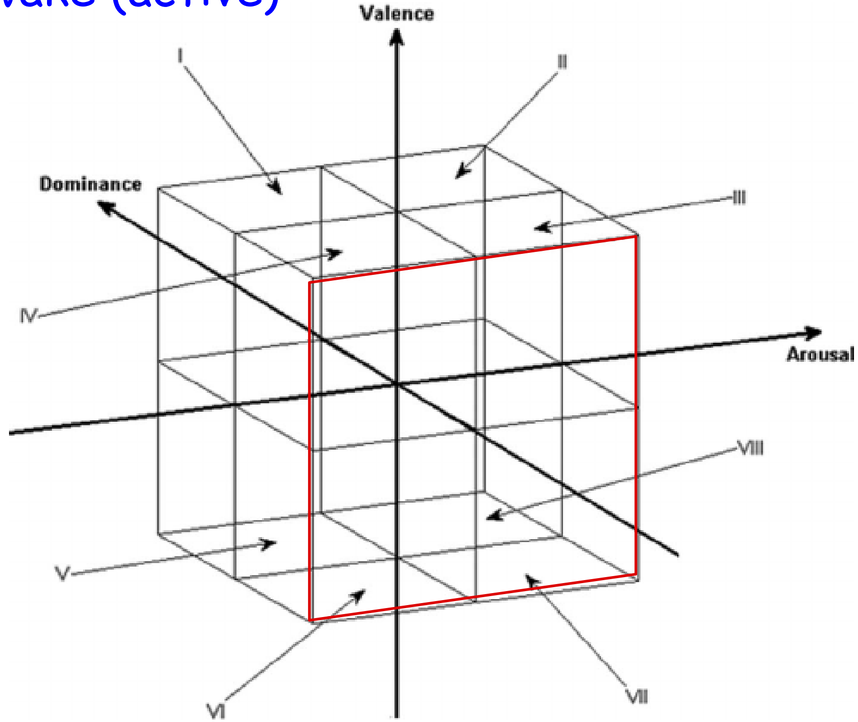
Pleasantness, positive to negative

Activation, sleepy (passive) to awake (active)

Power control, low to high



2D space (VA)



3D space (VAD)

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]

Previous research (in Akagi-lab) used small datasets and unsupervised learning which is hard to implement DNN methods and compare the results on these datasets

Evaluation metric

- Concordance correlation coefficient (CCC)

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

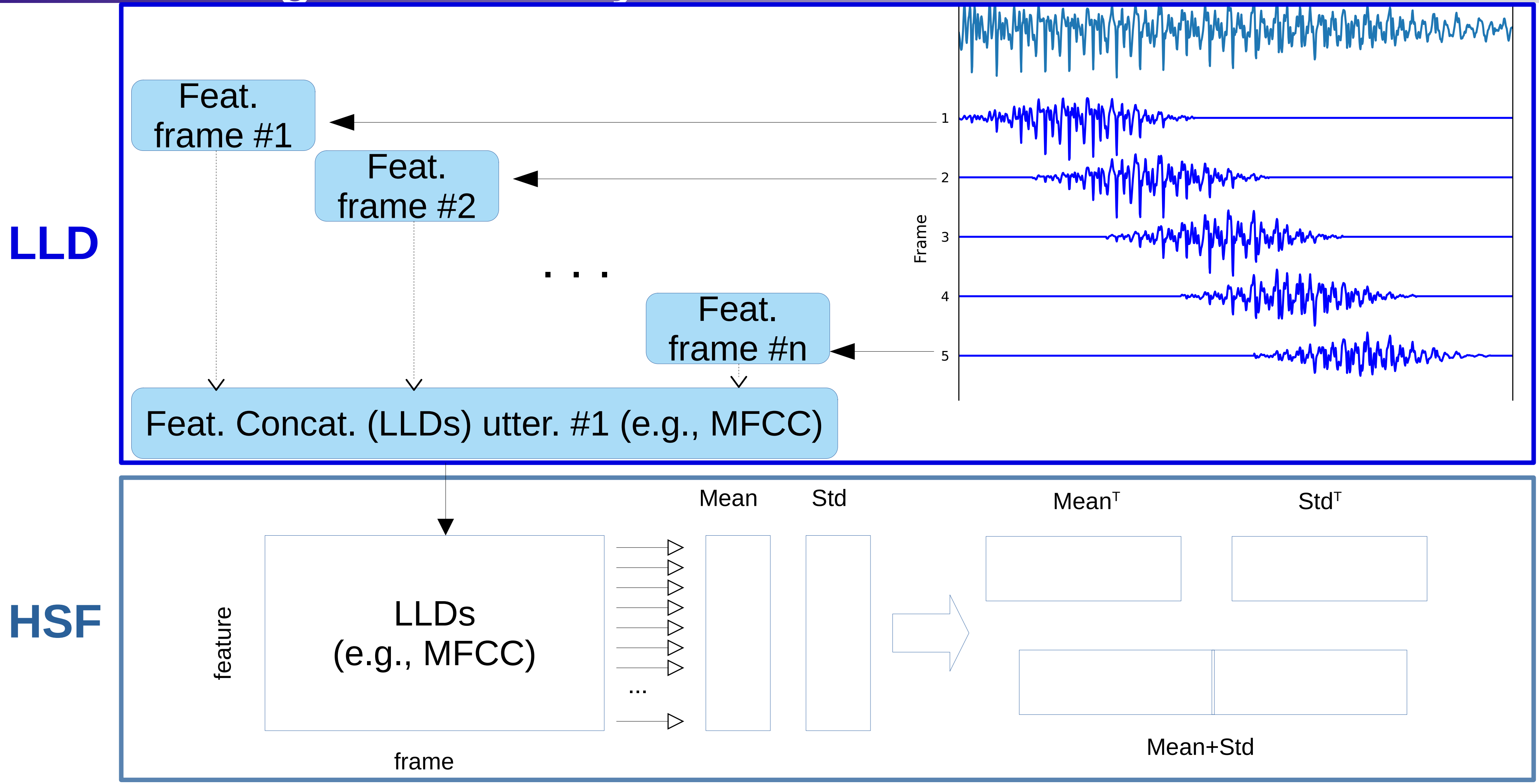
  - A step further than (Pearson) correlation coefficient
  - Penalizes any deviation from the identity relationship (both scale and location/shift)
  - Captures both accuracy and precision
  - Mathematically and experimentally superior to error-based loss functions (Pandit and Schuller, 2020; Atmaja and Akagi, 2020)
  - Interpretation (Altman, 1991):

    CCC < 0.2 (poor); 0.2 < CCC < 0.8 (moderate); CCC > 0.8 (good)

Outline

1. Introduction:
   Background, Aims & Issues, Applications
2. Research Methodology:
   Motivation, Previous work, Concept, Strategy, Datasets, 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:
   Comparative analysis, Summary, Contributions, Future research

Which region of analysis to extract features



Results: LLD vs. HSF (IEMOCAP data, in CCC)

LLD

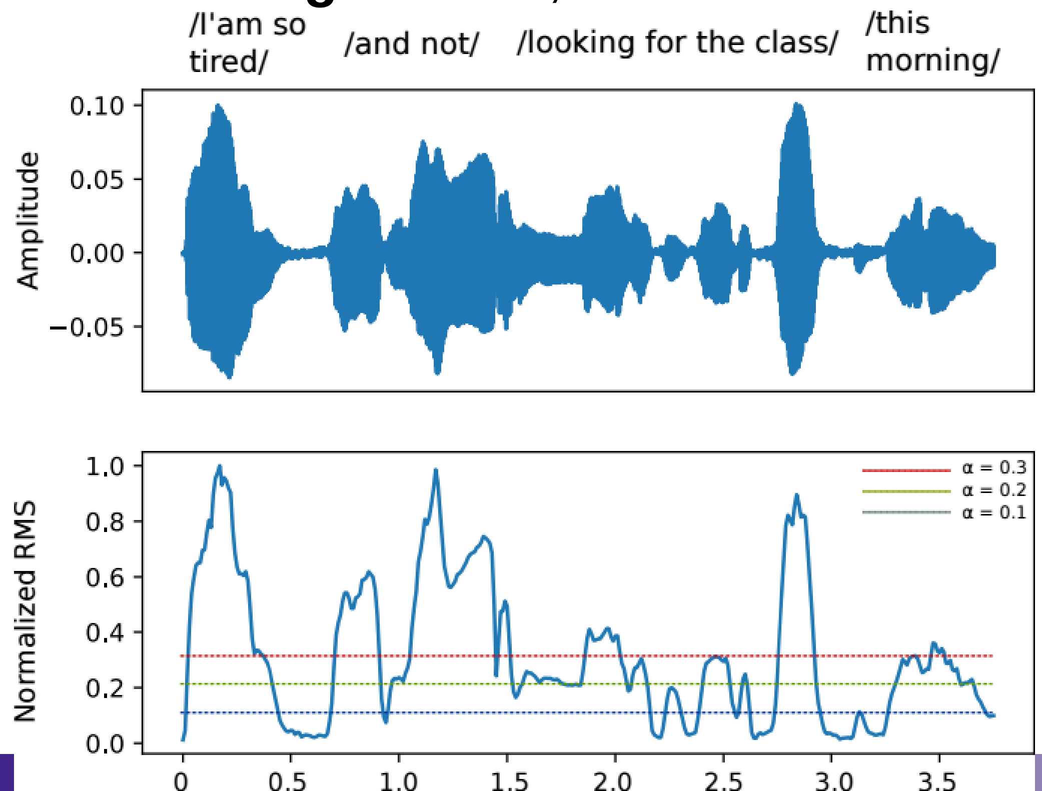
Feature	Dim	V	A	D	Mean
MFCC	(3414, 40)	0.148	0.488	0.419	0.352
Log mel	(3414, 128)	0.103	0.543	0.438	0.362
GeMAPS	(3409, 23)	0.164	0.527	0.454	0.382
pAA	(3412, 34)	0.130	0.513	0.419	0.354
pAA_D	(3412, 68)	0.145	0.526	0.439	0.370

HSF
mean+std

Feature	Dim	V	A	D	Mean
MFCC	80	0.155	0.580	0.456	0.397
Log Mel	256	0.151	0.549	0.455	0.385
GeMAPS	46	0.191	0.523	0.452	0.389
pAA	68	0.145	0.563	0.445	0.384
pAA_D	128	0.173	0.612	0.455	0.413

Effect of silent pause regions

- Three different treatments to evaluate silent pause regions:
  - **Removing silence**, and extract acoustic feature (AF) from these regions
  - **Keeping silence**, and extract AF from whole regions
  - **Utilizing silence**, as additional features to AF

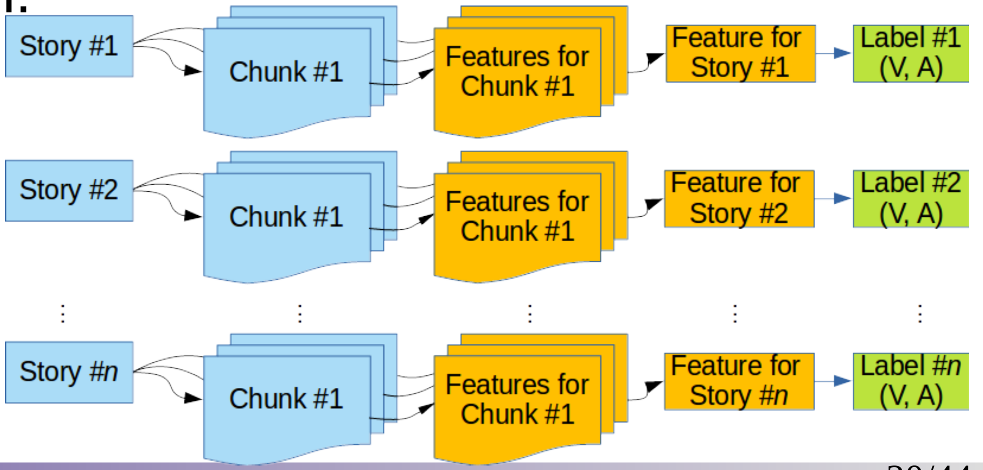


Removing & Utilizing silence:

- Removing silence can be done by using voice activity detection with RMS energy.
- If the RMS energy of particular frames lower than threshold (α), then these regions are removed.
- In contrast, those regions can be used to calculate silent pause features.

Acoustic feature aggregation

- Common methods to aggregate results for many-to-one problem
  ⇒ output aggregation, i.e., majority voting
- ***Initial aim: for fusing acoustic features with other features***
- Human may perceive emotion from chunks to utterance based on information aggregation (*not decisions/outputs aggregation*)
- Two aggregation methods are evaluated:
  - Acoustic feature (input) aggregation:
    - Mean values
    - Max. values
  - Output aggregation:
    - majority voting



Summary of Part III

- Proposed solutions for several issues in acoustic-based dimensional SER:

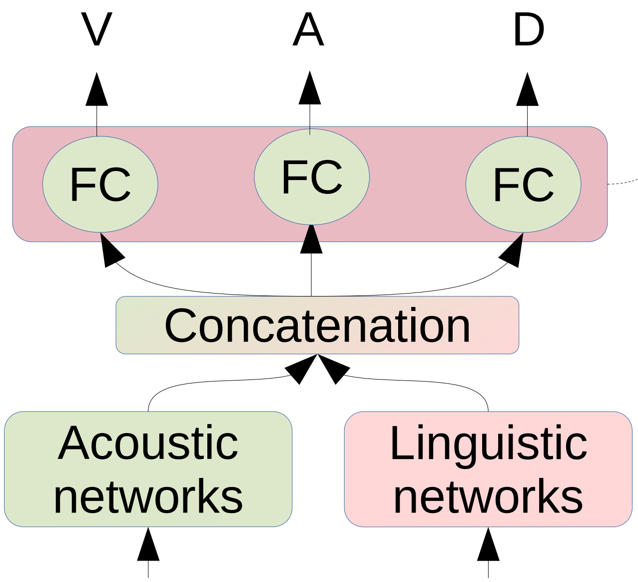
Issue	Proposed method		
Region of analysis	frames		**utterance/fixed length**
Silence region	**removing silence**	keeping silence	**utilizing silence**
Aggregation method	**input aggregation**		output aggregation

- Acoustic-based dimensional SER still suffers from low performance of valence prediction
- **Using acoustic features only for SER is not enough!**

Outline

1. Introduction:
   Background, Aims & Issues, Applications
2. Research Methodology:
   Motivation, Previous work, Concept, Strategy, Datasets, 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:
   Comparative analysis, Summary, Contributions, Future research

Network concatenation with multitask learning (MTL)



Acoustic <u>features:</u>
pAA LLD
pAA HSF
GeMAPS LLD
GeMAPS HSF

Linguistic <u>features:</u>
WE
word2vec
GloVe
FastText
BERT

- Loss function:

$$CCCL = 1 - CCC$$

Total loss function (with no parameter):

$$CCCL_{tot} = CCCL_V + CCCL_A + CCCL_D.$$

Total loss function with 2 parameters:

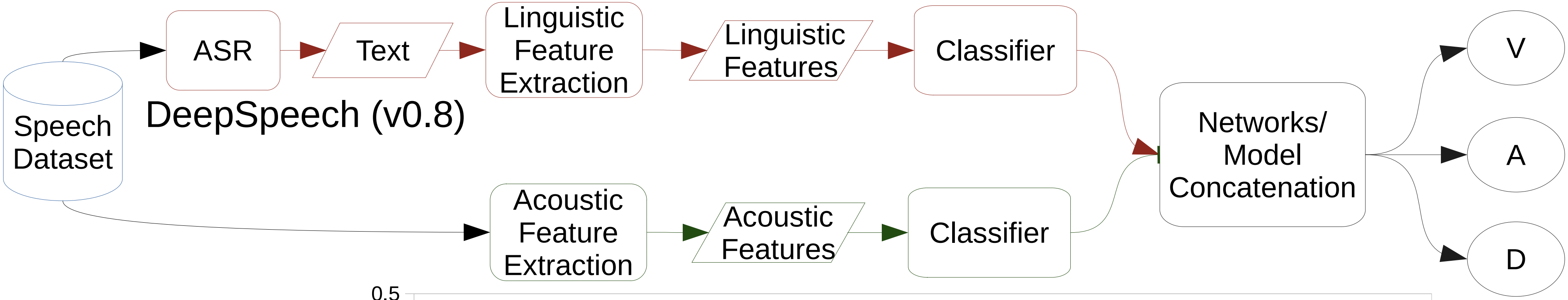
$$\begin{aligned} \mathrm{CCCL}_{tot} = \ & \alpha\ CCCL_V + \beta\ CCCL_A \\ & + (1 - \alpha - \beta)\ CCCL_D \end{aligned}$$

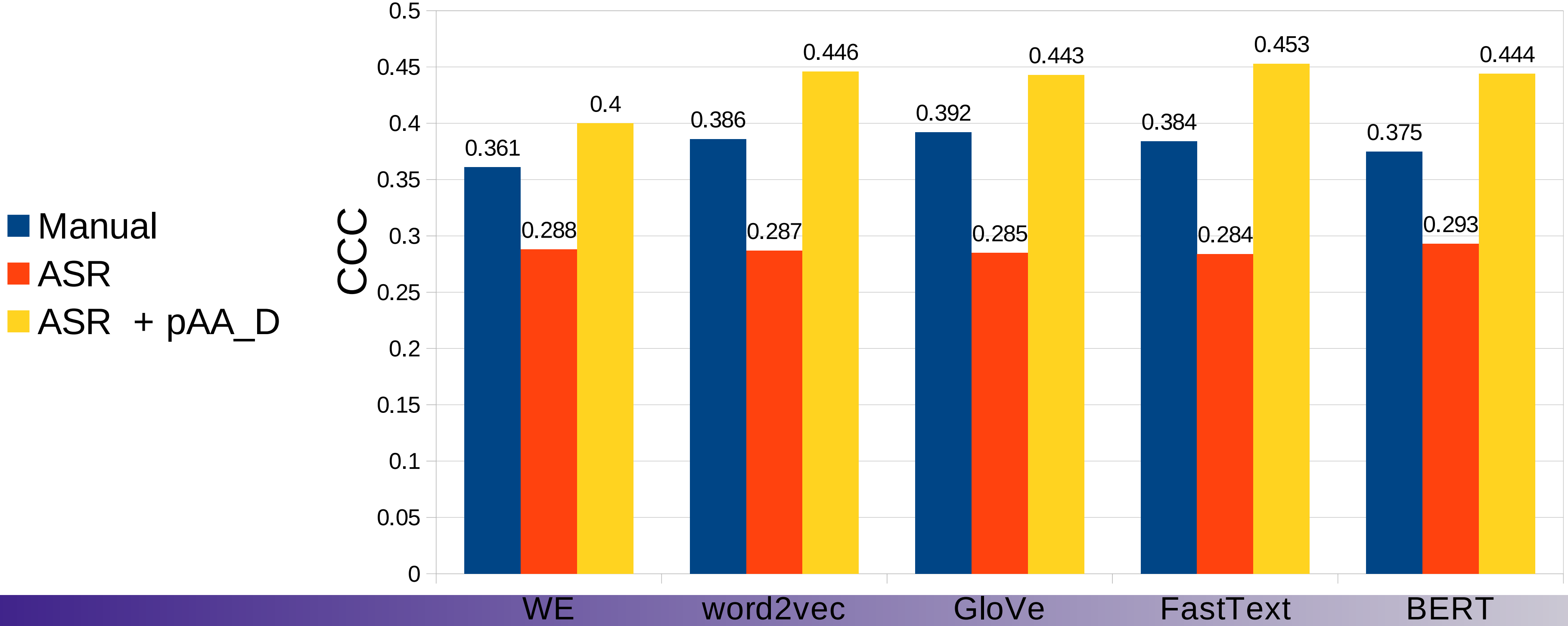
Total loss function with 3 parameters:

$$CCCL_{tot} = \alpha\ CCCL_V + \beta\ CCCL_A + \gamma\ CCCL_D$$

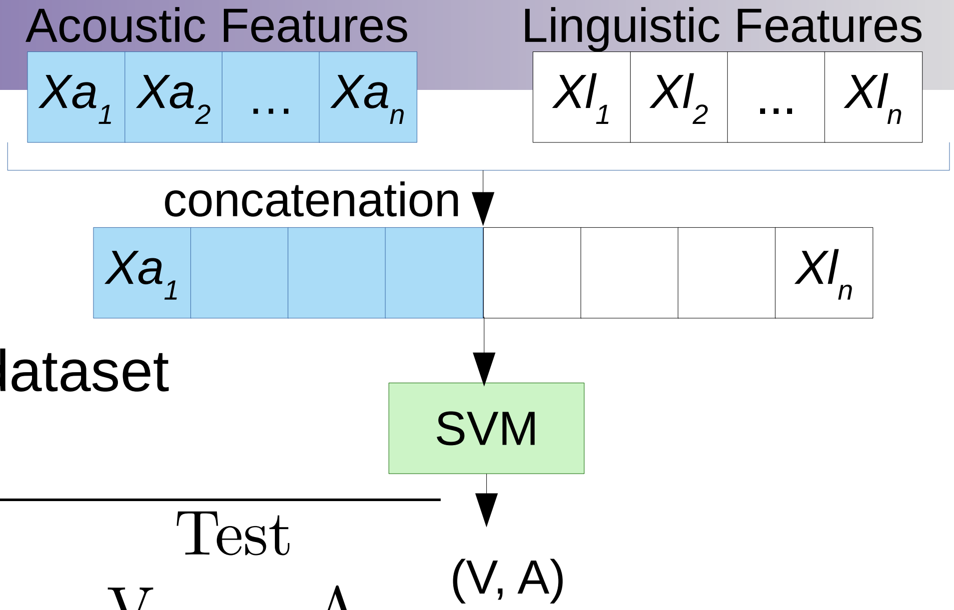
MTL method	V	A	D	Mean
No parameter	0.409	0.585	**0.486**	0.493
2 parameters	**0.446**	**0.594**	0.485	**0.508**
3 parameters	0.419	0.589	0.483	0.497

Dimensional SER from ASR outputs





Feature concatenation



Accuracy (UAR, %) from USOMS-e dataset (INTERSPEECH 2020)

Features		Dev		Test	
Acoustic	Linguistic	V	A	V	A
ResNet50	-	31.6	35.0	40.3	**50.4**
-	BLAtt	49.2	40.6	49.0	44.0
LibROSA	Gmax	**58.2**	34.6	40.5	34.8
ResNet50	Gmax	**58.2**	51.0	40.9	**50.4**
ResNet50	BLAtt	47.6	**52.5**	**56.3**	46.4
BoAW-250	BLAtt	**58.2**	44.4	49.0	47.4

Summary of Part IV

- *Fusing acoustic and linguistic information at feature level improves valence prediction*
- A proper choice of feature representation from linguistic information, i.e., using GloVe embedding, not only improves valence prediction but also improves arousal and dominance predictions
  - Multitask learning (MTL) models interrelation of emotion dimensions better than any other evaluated method
  - No significant different on using pre-trained linguistic models on ASR outputs, the different was observed in manual transcription

Outline

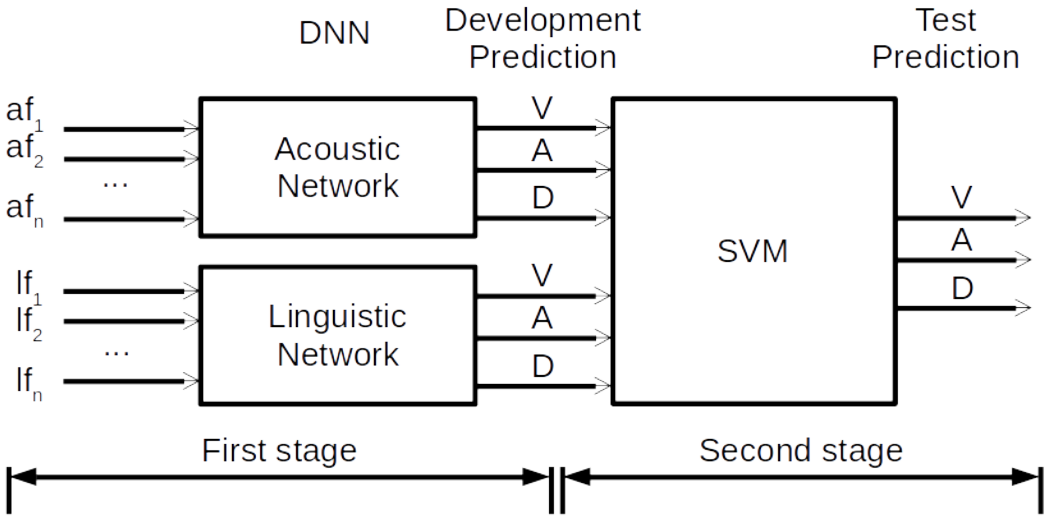
1. Introduction:
   Background, Aims & Issues, Applications
2. Research Methodology:
   Motivation, Previous work, Concept, Datasets, Metric
3. Dimensional SER Using Acoustic Features
4. Early Fusion of Acoustic and Linguistic Information
5. **Late Fusion of Acoustic and Linguistic Information**
6. Conclusions:
   Comparative analysis, Summary, Contributions, Future research

Result: late fusion

Input af:

- GeMAPS LLD,
- GeMAPS mean+std (HSF1)
- GeMAPS mean+std+sil (HSF2)

Input lf: WE, word2vec, GloVE



Dataset	Features (best)	V	A	D	Mean
IEMOCAP-SD	HSF2+word2vec	0.595	0.601	0.499	0.565
IEMOCAP-LOSO	HSF2+GloVe	0.553	0.579	0.465	**0.532**
MSPIN-SD	HSF2+word2vec	0.486	0.641	0.524	0.550
MSPIN-LOSO	HSF2+GloVe	0.291	0.570	0.405	0.422

MSPIN: Parts of MSP-IMPROV dataset excluding target sentence scenario (‘Target - improvised’ and ‘Target - read’)

Summary of Part V

- Late fusion of acoustic and linguistic information **improves valence prediction**
- Late fusion framework models the fusion of acoustic and linguistic information better than early fusion; better consistent results were obtained.
  - ***Linguistic information contributes to valence prediction while acoustic information contributes dominantly to arousal and dominance***
  - Results on speaker-independent is significantly different from speaker-dependent
  - Removing lexical-controlled utterances still shows some influence of those utterances; further investigation is needed

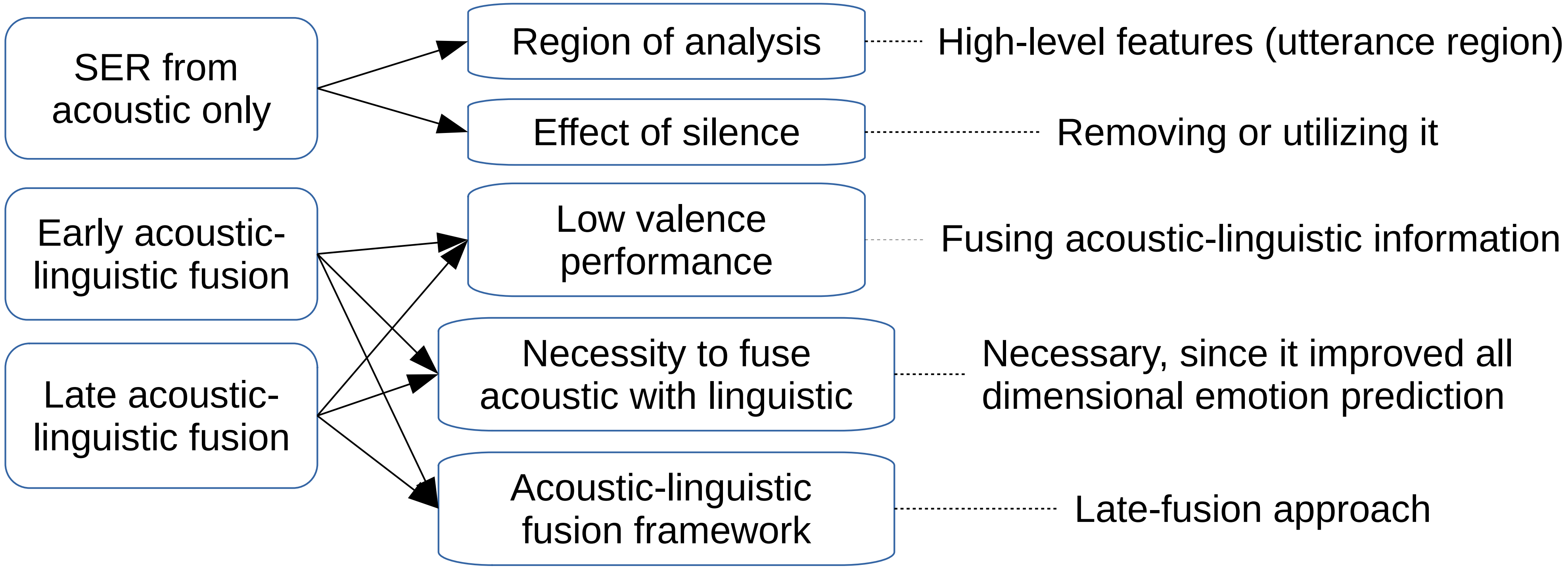
Outline

Comparative analysis

Dataset	Authors	Modalities	V	A	D
IEMOCAP SI	This study (Aco)	Ac	0.298	0.641	0.460
IEMOCAP SI	This study (FL)	Ac+Li	0.446	0.594	0.508
IEMOCAP SI	This study (DL)	Ac+Li	0.553	0.579	0.550
IEMOCAP SD	Zhao et al. (2018)	Ac+Age+G	0.715	0.392	0.539
IEMOCAP SD	Zhao et al. (2019)	Ac+Age+G	0.590	0.689	0.591
IEMOCAP+Podcast	Abdelwahab	Ac	0.140	0.305	0.181
Podcast+IEMOCAP	Partasarathy	Ac	0.235	0.623	0.441
MSP-Podcast SI	Sridhar et al.	Ac	0.291	0.711	0.690
SEMAINE	Yang	Ac	0.506	0.680	-
RECOLA	Bakshi et al.	Ac	0.314	0.660	-
SEWA (DE)	Schmitt et al.	Ac	0.489	0.499	-
SEWA (DE+HU)	Atmaja & Akagi	Ac+Vi	0.656	0.680	-
SEWA (DE+HU)	Chen et al.	Ac+Vi+Li	0.755	0.672	-

Summary

- **This study shows the necessity of fusing acoustic with linguistic information for dimensional SER**; the late fusion method models dimensional SER better than early fusion and unimodal acoustic analysis
- Potential solutions for the issues:



Contributions

- SER from acoustic information only
  - **Silent feature calculation based on ratio of silent frames and total frames**
  - **Acoustic feature aggregation to aggregate chunks to a story (long utterance) [many-to-one problem]**
  - Generalization of Mean+Std impact to other feature sets
  - Experimental evaluation of correlation- vs error-based loss functions for dimensional SER
- Early acoustic-linguistic information fusion
  - **Multi-task learning based on CCC loss with different number of parameters**
  - Contribution of different linguistic information
  - Evaluation of manual transcription and ASR outputs
- Late acoustic-linguistic information fusion
  - **Two-stage processing dimensional SER using DNNs and SVM**
  - Discussion about speaker-dependent vs. speaker-independent results
  - Effect of removing ‘target sentence’ from lexical controlled dataset

Future research direction

- Accelerating high-level feature extraction for speech emotion recognition
- Bimodal acoustic-linguistic emotion recognition by two spaces resultant
- Fully lexical controlled vs. lexical uncontrolled emotion recognition
- Bottleneck between acoustic and linguistic processing
- Concurrent speech and emotion recognition
- Model generalization

References

- P. B. Denes and E. Pinson, The speech chain. Macmillan, 1993.
- S. Poria, E. Cambria, R. Bajpai, and A. Hussain, “A review of affective computing: From unimodal analysis to multimodal fusion,” Inf. Fusion, vol. 37, pp. 98–125, Sep. 2017.
- P. Mairano, E. Zovato, and V. Quinci, “Do sentiment analysis scores correlate with acoustic features of emotional speech?,” in AISV Conference, 2019.
- S. Buechel and U. Hahn, “Emotion analysis as a regression problem-dimensional models and their implications on Emotion representation and metrical evaluation,” Front. Artif. Intell. Appl., vol. 285, pp. 1114–1122, 2016.
- A. Halevy, P. Norvig, and F. Pereira, “The unreasonable effectiveness of data,” IEEE Intell. Syst., vol. 24, no. 2, pp. 8–12, 2009.
- K. R. Scherer, “What are emotions? And how can they be measured?,” Soc. Sci. Inf., vol. 44, no. 4, pp. 695–729, 2005.
- C. A. Rossi, "The development and validation of the emotion knowledge and awareness test." (2016).
- V. Pandit and B. Schuller, “The many-to-many mapping between concordance correlation coefficient and mean square error,” arXiv, pp. 1–32, 2019.

References (cont'd)

- B.T. Atmaja, M. Akagi. “Evaluation of Error and Correlation-based Loss Functions For Multitask Learning Dimensional Speech Emotion Recognition,” International Conference on Acoustic and Vibration, Bali, Indonesia, 2020.
- D.G Altman, Practical statistics for medical research. London: Chapman and Hall, (1991).
- M. Schmitt, N. Cummins, and B. W. Schuller, “Continuous Emotion Recognition in Speech - Do We Need Recurrence?,” in Interspeech 2019, 2019, pp. 2808–2812.
- M. El Ayadi, M. S. Kamel, and F. Karray, “Survey on speech emotion recognition: Features, classification schemes, and databases,” Pattern Recognit., vol. 44, no. 3, pp. 572–587, 2011.
- R. Elbarougy, “A Study on Constructing an Automatic Speech Emotion Recognition System based on a Three-Layer Model for Human Perception,” 2013.
- X. Li, “A Three-Layer Model Based Estimation of Emotions in Multilingual Speech,” Japan Advanced Institute of Science and Technology, 2019.
- S. A. Kotz and S. Paulmann, “Emotion, Language, and the Brain,” Language and Linguistics Compass, vol. 5, no. 3, pp. 108–125, mar 2011.

APPENDIX

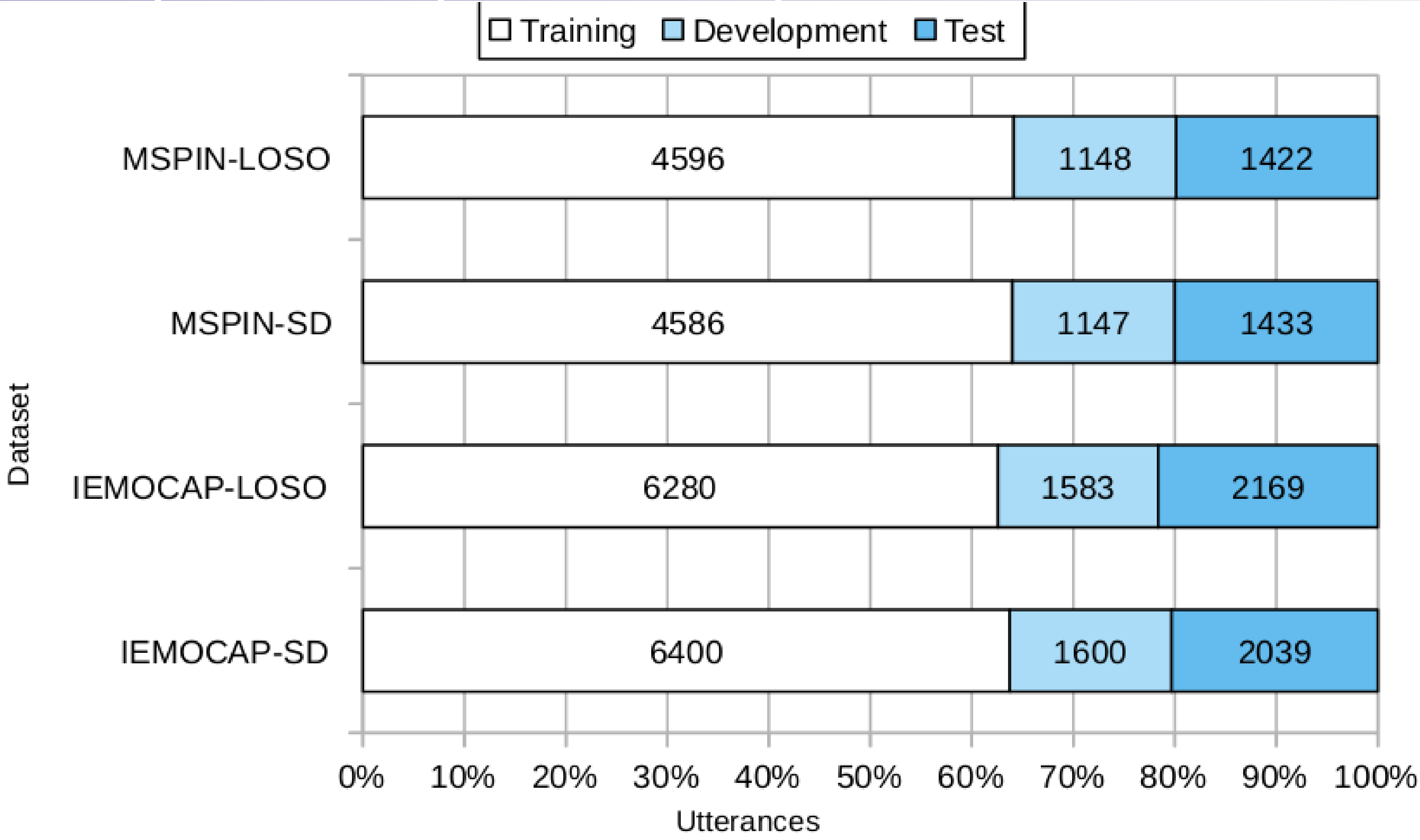
List of abbreviation

- ASR: Automatic Speech Recognition
- SER: Speech Emotion Recognition
- CCC: Concordance correlation coefficient
- DNN: Deep Neural Network
- SVM: Support Vector Machine
- FL: Feature-level fusion, DL: Decision-level fusion
- V: Valence, A: Arousal, D: Dominance
- VAD: Valence-arousal-dominance
- LLD: Low-level descriptor
- HSF: High-level statistical functions

List of abbreviation (Cont’d)

- SD: speaker dependent
- LOSO: leave one session out, SI: speaker independent
- WER: word error rate
- pAA: pyAudioAnalysis
- pAA_D: pyAudioanalysis with their deltas
- MTL: multi-task learning
- af: acoustic feature
- lf: linguistic feature
- WE: word embedding
- Std: standard deviation

Dataset partition (late fusion)



Results: effect of silent pause features

Strategy	V	A	D	Mean
IEMOCAP				
Removing silence	0.283	0.640	0.454	0.459
Keeping silence	0.268	0.641	0.458	0.456
Utilizing silence	0.298	0.641	0.460	0.466
MSP-IMPROV				
Removing silence	0.259	0.586	0.441	0.429
Keeping silence	0.217	0.586	0.425	0.409
Utilizing silence	0.227	0.601	0.443	0.424

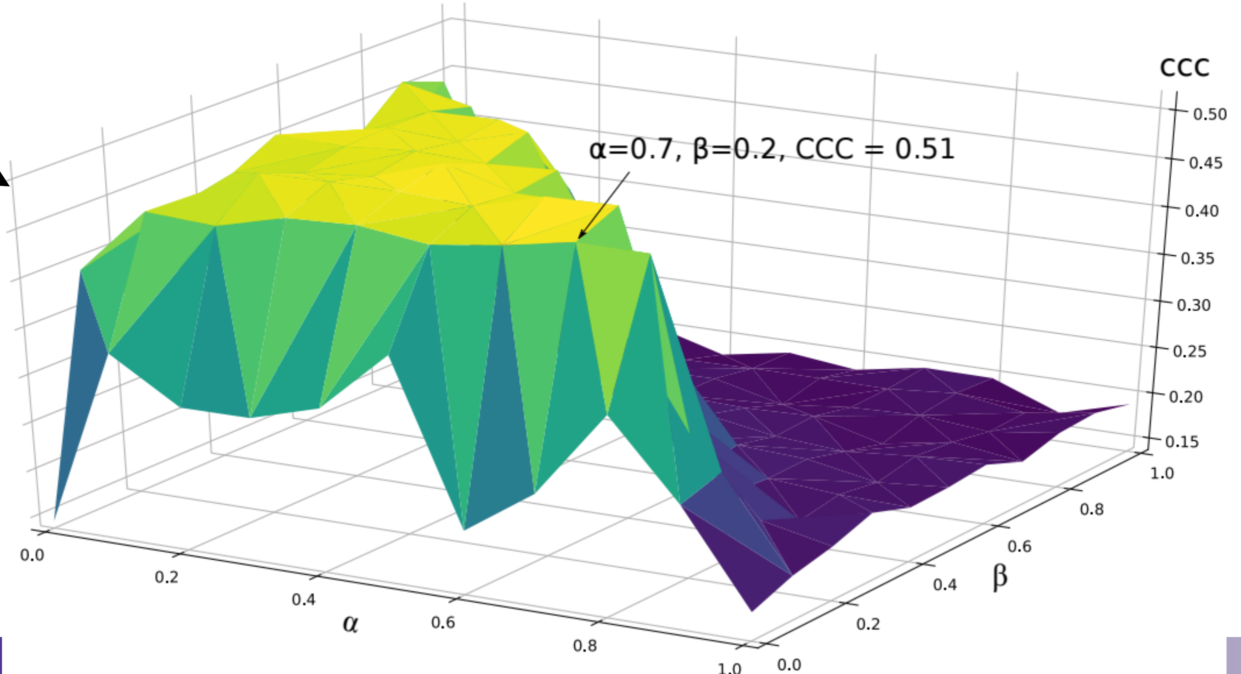
The improvement and correlation between removing, keeping, and utilizing silence is small; further studies (e.g., TFS-ENV) are needed to observe such improvements.

Result: feat. aggregation vs. majority voting

Features	Majority Voting [6]		Mean Input Agg.		Max Input Agg.	
	V	A	V	A	V	A
LibROSA HSF	-	-	**45.1**	38.3	42.7	39.7
ComParE	33.3	39.1	43.4	42.7	**45.3**	37.0
BoAW-125	38.9	42.0	44.6	**45.7**	44.6	40.1
BoAW-250	33.3	40.5	43.0	40.8	39.6	37.6
BoAW-500	38.9	41.0	42.6	41.0	42.9	37.9
BoAW-1000	38.7	30.5	43.5	41.5	40.2	39.8
BoAW-2000	**40.6**	39.7	41.9	44.8	43.4	40.1
ResNet50	31.6	35.0	36.5	36.7	37.1	39.0
AuDeep-30	35.4	36.2	38.4	42.1	42.8	35.6
AuDeep-45	36.7	34.9	39.5	40.5	39.3	33.3
AuDeep-60	35.1	**41.6**	43.4	42.1	40.7	41.4
AuDeep-75	32.7	40.4	41.9	44.4	40.9	**43.3**
AuDeep-fused	29.2	36.3	43.6	39.5	42.2	39.3

Result: networks concatenation with MTL

MTL method	V	A	D	Mean
No parameter	0.409	0.585	**0.486**	0.493
2 parameters	**0.446**	**0.594**	0.485	**0.508**
3 parameters	0.419	0.589	0.483	0.497



Unimodal parameters:
Aco, α=0.1, β=0.5
Ling, α=0.7, β=0.2

Result: relative improvement [Late fusion]