PhD Thesis Defense

Dimensional Speech Emotion Recognition by Fusing Acoustic and Linguistic Information

s1820002 - Bagus Tris Atmaja (バグストリスアトマジャ)

2 February 2021

北陸先端科学技術大学院大学

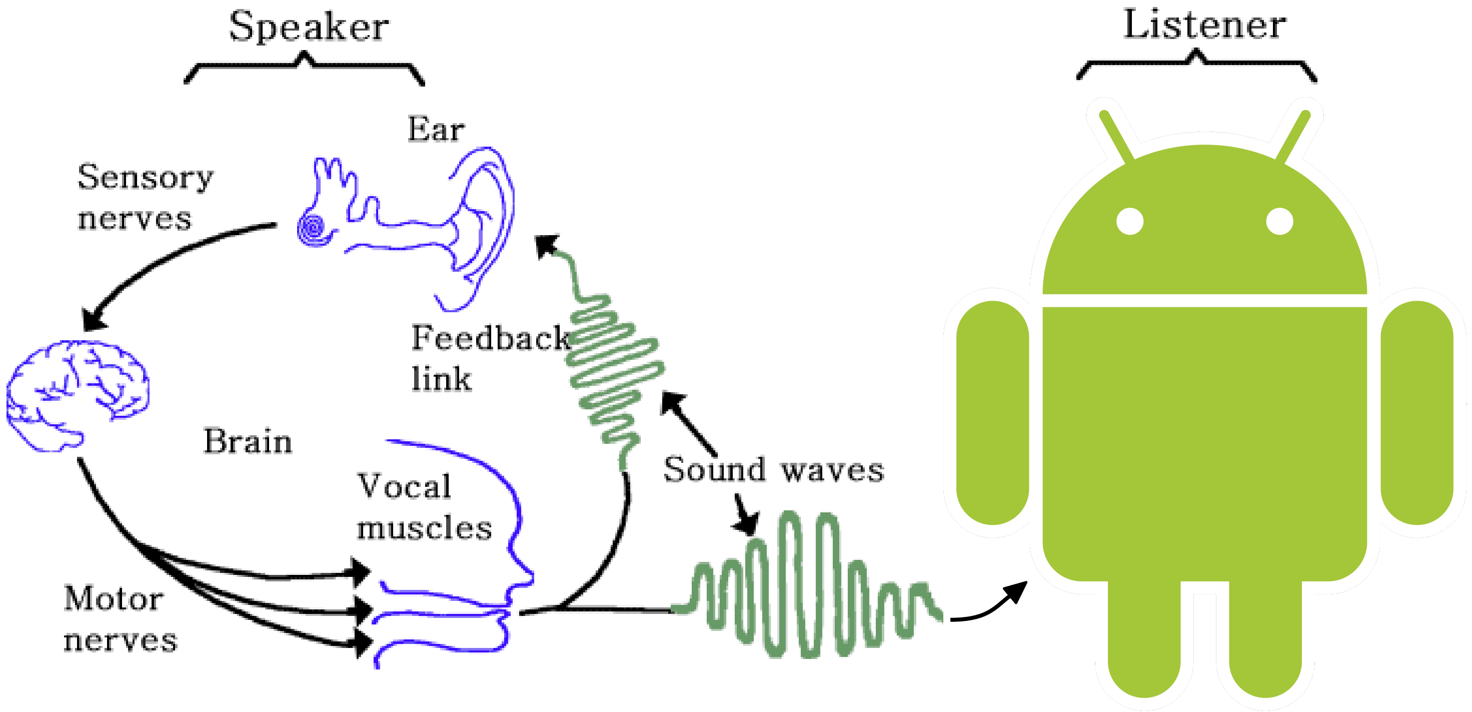
情報科学

音情報処理分野　赤木・鵜木研究室

Outline

1. **Introduction:**
   Background, Aims, Novelty, Significance, Applications
2. Research Methodology:
   Motivation, Problems, Concept, Strategy, Datasets and Evaluation metric, Previous Work
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

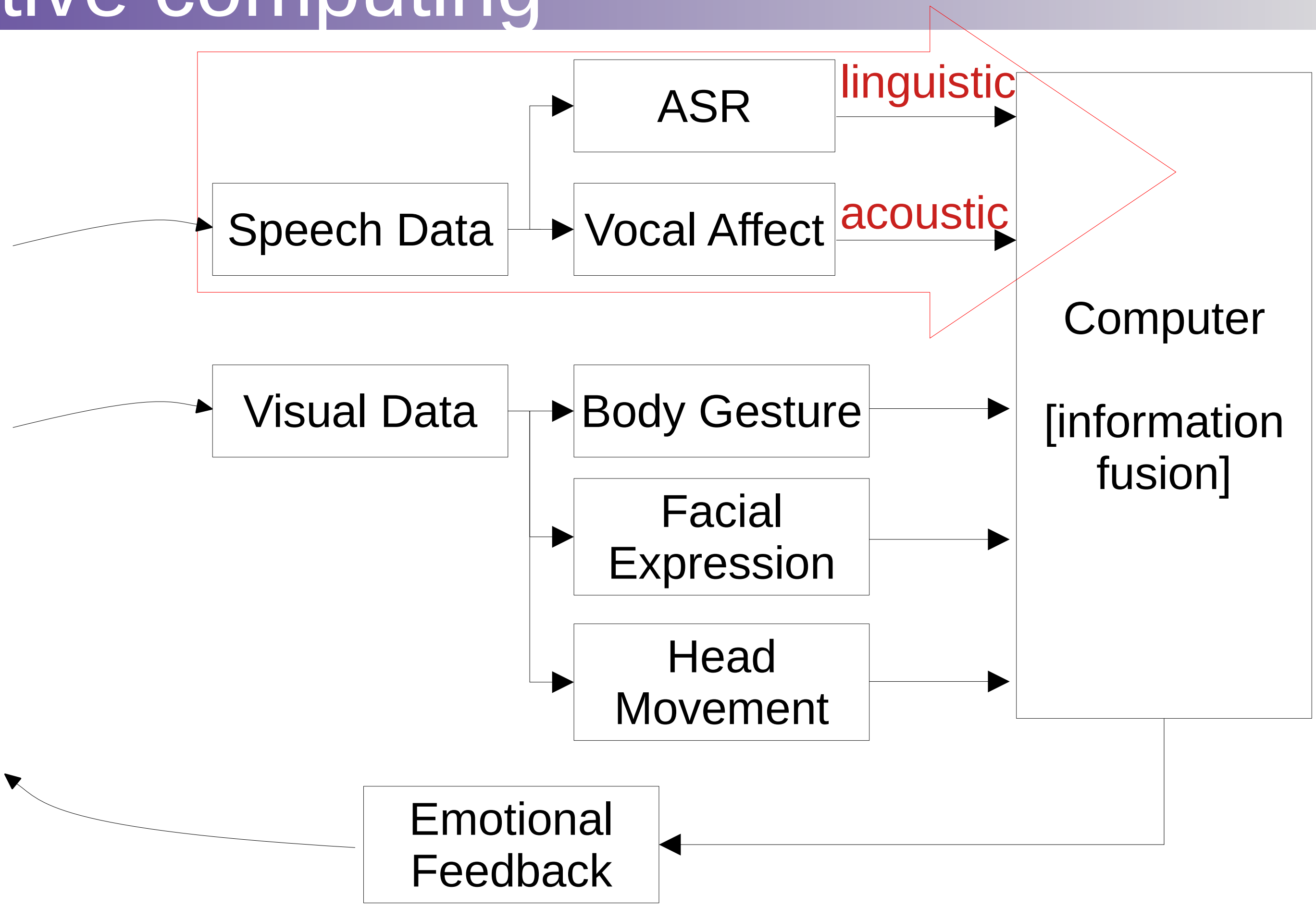
Human-machine communication



Linguistic level → Physiological level → Acoustic level → Speech Emotion Recognition

In speech chain, acoustic and linguistic are connected by physiological function; fusing both information may be useful for emotion recognition by machines

Multimodal affective computing



Combining more than one modality from different sensors for affective computing – computing that relates to, arises from, or influences emotion

(Poria et al., 2017)
(Picard, 1995)

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)

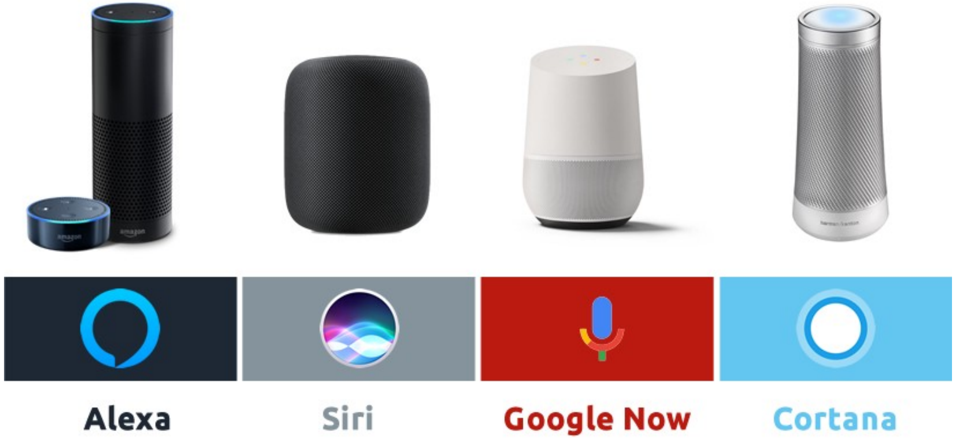
Novelty & significance

- SER from acoustic information only
  - **Silent feature calculation based on ratio of silent frames and total frames**
  - **Acoustic features aggregation to map acoustic features from 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
  - Effect of removing ‘target sentence’ from lexical controlled dataset

Possible applications

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

- Voice assistant



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

Outline

1. Introduction:
   Background, Aims, Novelty, Significance, Applications
2. **Research Methodology:**
   Motivation, Problems, Concept, Strategy, Datasets and Evaluation metric, Previous Work
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 researching SER
  - In some cases, only speech data could be obtained.
  - There is strong correlation between speech and emotion:

Pitch → valence

Rhythmic + Spectral → arousal (Mairano et al., 2019)

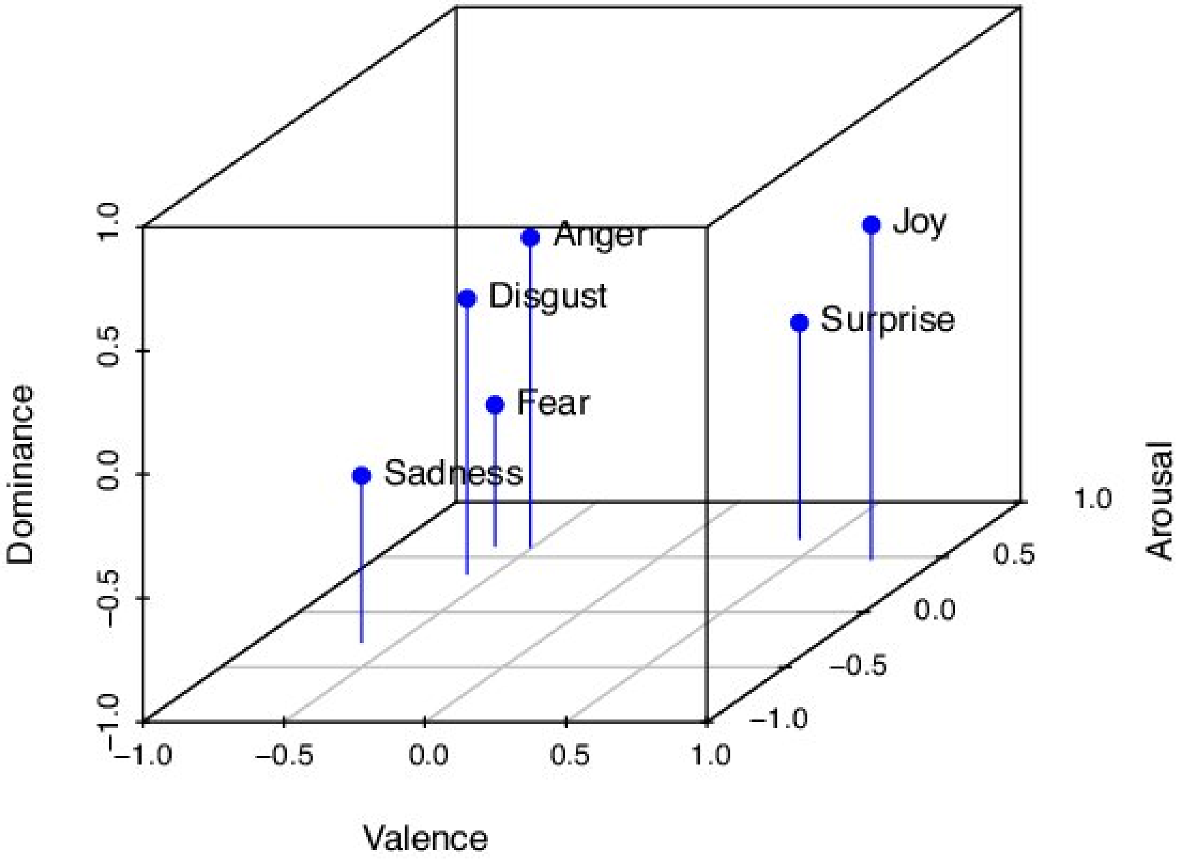
- Why researching SER is difficult
  - The labels are given by annotators; no exact values (cf. digits).

IEMOCAP ID: Ses01F_Impro01_F001

Annotators	Valence	Arousal	Dominance
Annot. #1	3	2	2
Annot. #2	2	3	3
Annot. #3	2	3	2

Motivation

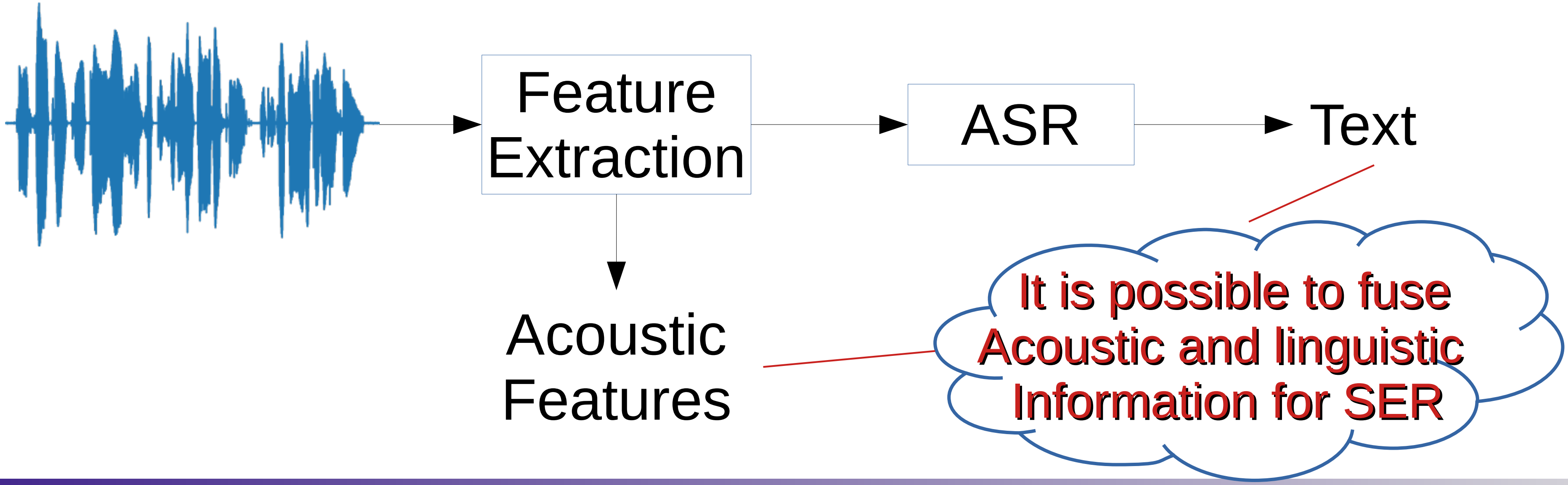
- Why dimensional SER
  - Human’s variability is high; hence, categorization doesn’t have an essence
  - Categorical emotion is not enough to describe affective state
  - Most previous SER works only focus on categorical emotion



Valence-arousal-dominance (VAD) model with Ekman’s six basic emotions (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
  - Human communicate emotion through speech and language (Kotz et al., 2011)
  - More data tends to be more effective (Halevy et al., 2009)

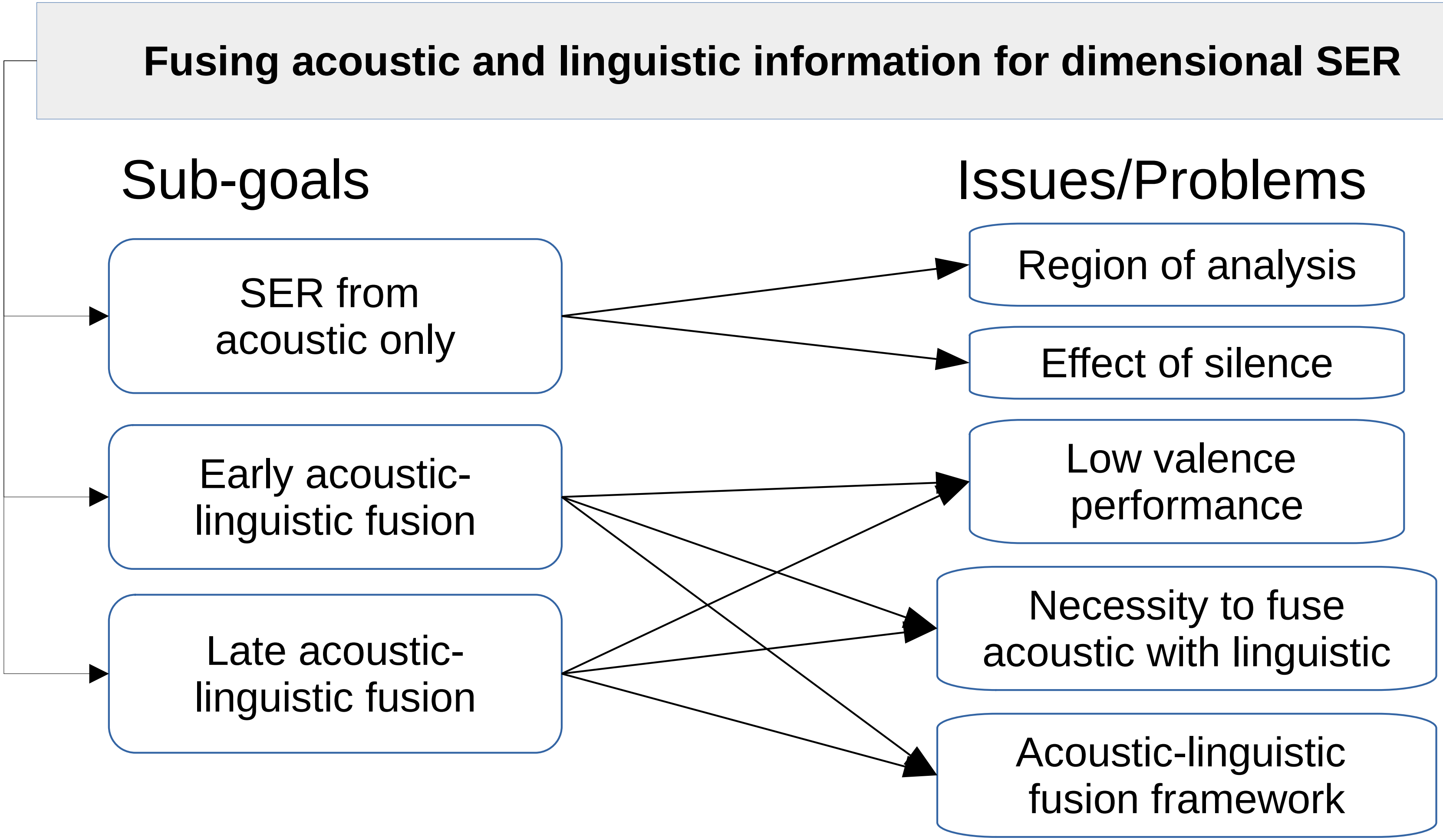


Research Issues

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 (Xingfeng 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 acoustic information

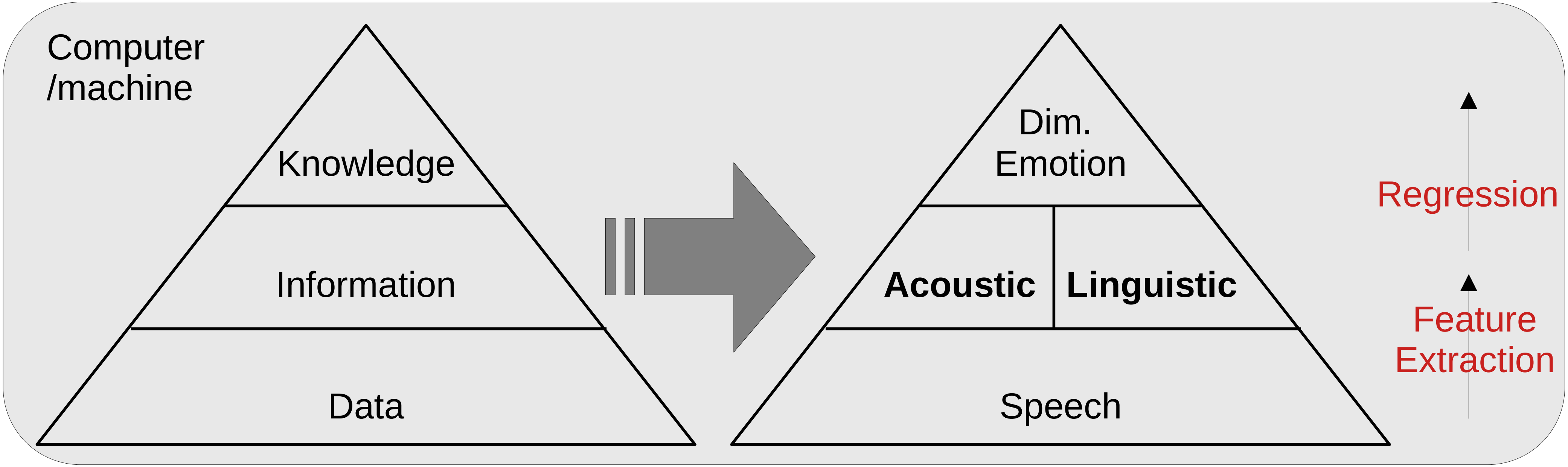
Correlation between aims and issues

Main goal



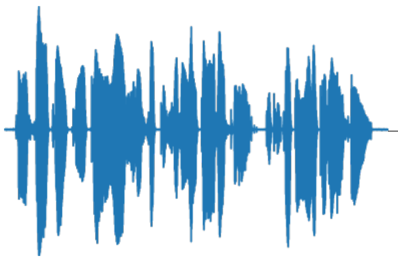
Concept/Philosophy

"It is not only ***how*** things are said, but also ***what*** things are said."



Information (*acoustic and linguistic*) is extracted from data (*speech*); knowledge (degree of dimensional emotion) is extracted from information (*acoustic and linguistic*).

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



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

Information: acoustic and linguistic features

- Acoustic is the *main information* to perceive emotion in speech, while linguistic is the *additional information*.
- Conceptual information in practice is implemented as ***features***

- Acoustic Features
  - Hand-crafted Features
    - LLD
      - Energy
      - f_o, etc.
    - HSF
      - Min/max
      - Std, var, etc.
  - Features Extracted by Machine/Deep Learning
    - BoAW
    - DeepSpectrum, etc
- Linguistic features
  - Hand-crafted Features
    - Emotional words and phrase
    - Exclamation, negation, word number
    - TF-IDF
    - BoW, etc
  - Machine/Deep Learning-based features
    - wor2vec
    - ELMo
    - GloVe
    - FastText
    - BERT, etc

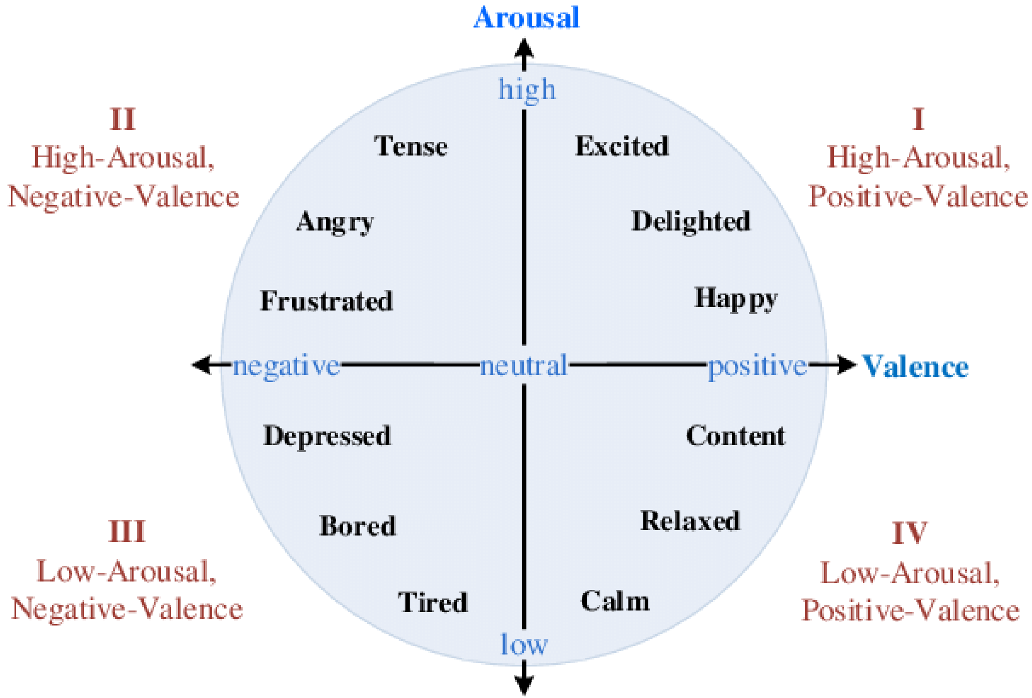
Knowledge: dimensional emotions degree

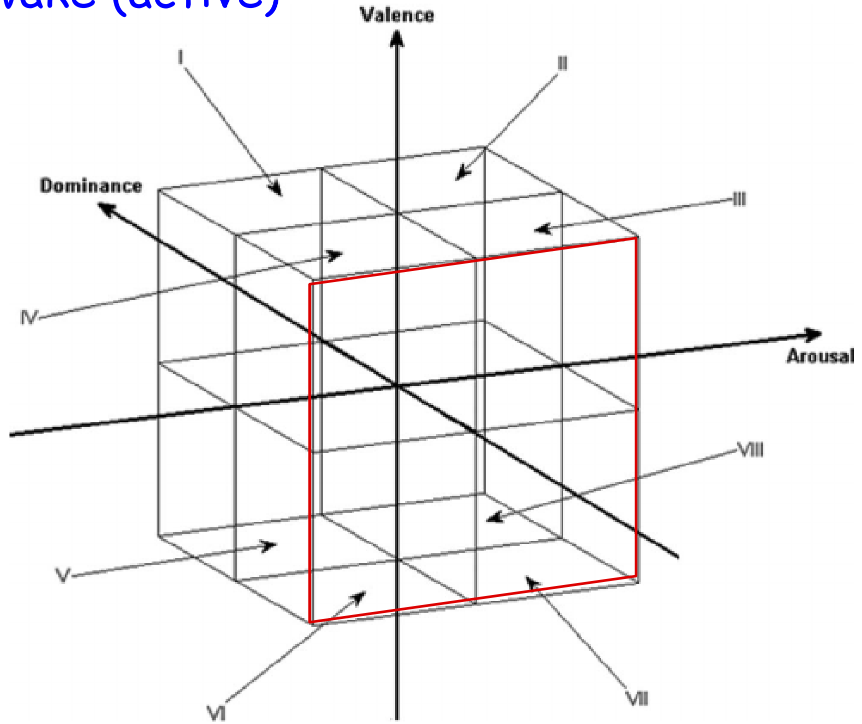
- Dimensional emotion: emotion as continuous degree in several attributes/dimensions
- Most common dimension: Valence, Arousal, and Dominance

Pleasantness, positive to negative

Power control, low to high

Activation, sleepy (passive) to awake (active)





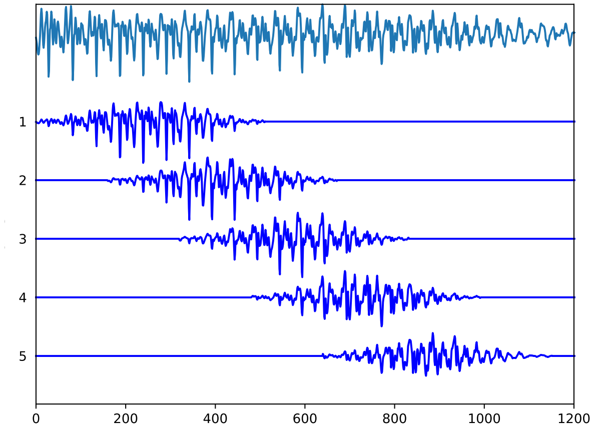
2D space (VA)

3D space (VAD)

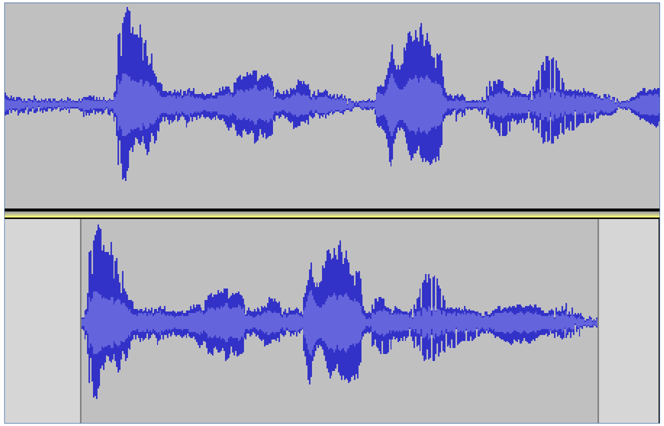
Research strategy

Is acoustic only enough for SER?

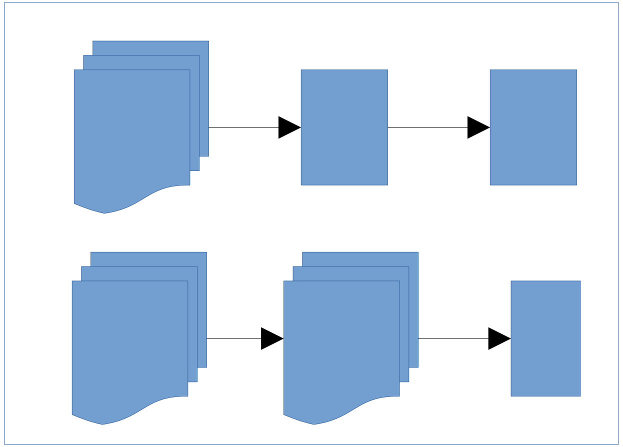
1) Dimensional SER using acoustic features:
   - Which region of analysis to extract acoustic features
   - Effect of silent pause regions
   - Aggregation methods for chunks to an utterance



LLD vs. HSF



Keeping vs. Removing vs. Using silence



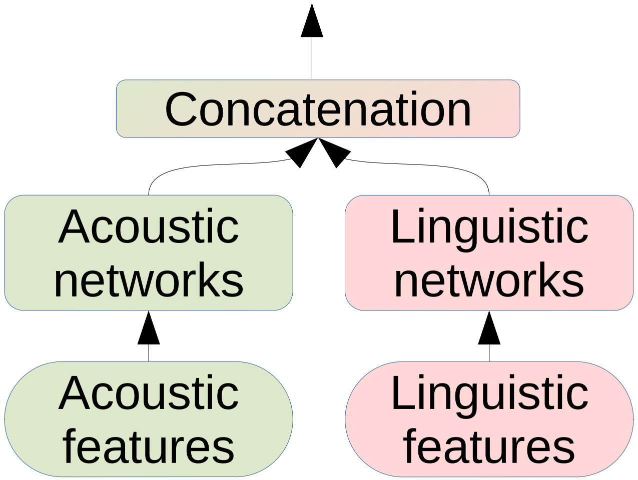
Input vs. output aggregation

Research strategy

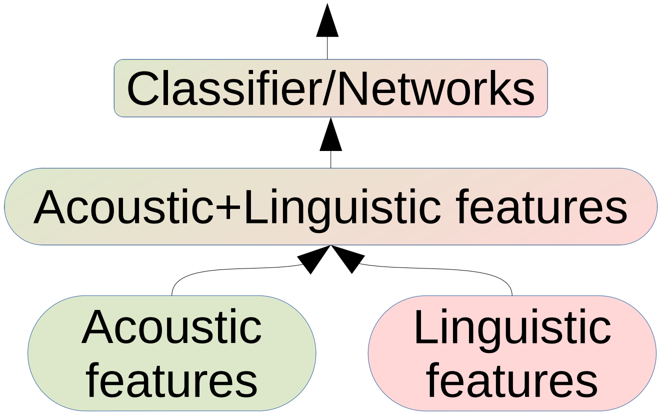
2) Early acoustic-linguistic fusion (feature level [FL]):

- Effect of different word embeddings
- Early fusion by network concatenation
- Early fusion by feature concatenation
- Using ASR outputs for linguistic input

Early fusion is the simplest *fusion* method



Network Concatenation



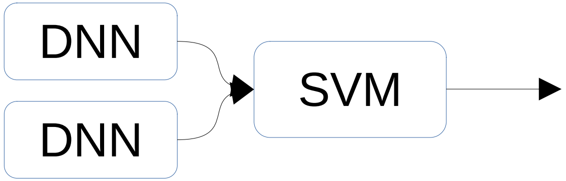
Feature Concatenation

Research strategy

Humans process linguistic and acoustic at different regions

3) Late acoustic-linguistic fusion (decision level [DL]):

- Late fusion approach by two-stage processing:
  - DNNs
  - SVM



- Results and discussion:
  - Result of two-stage processing
  - Speaker dependent (SD) vs. speaker independent (LOSO, leave-one-session-out)
  - Effect of removing ‘target sentences’

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 metric (CCC)

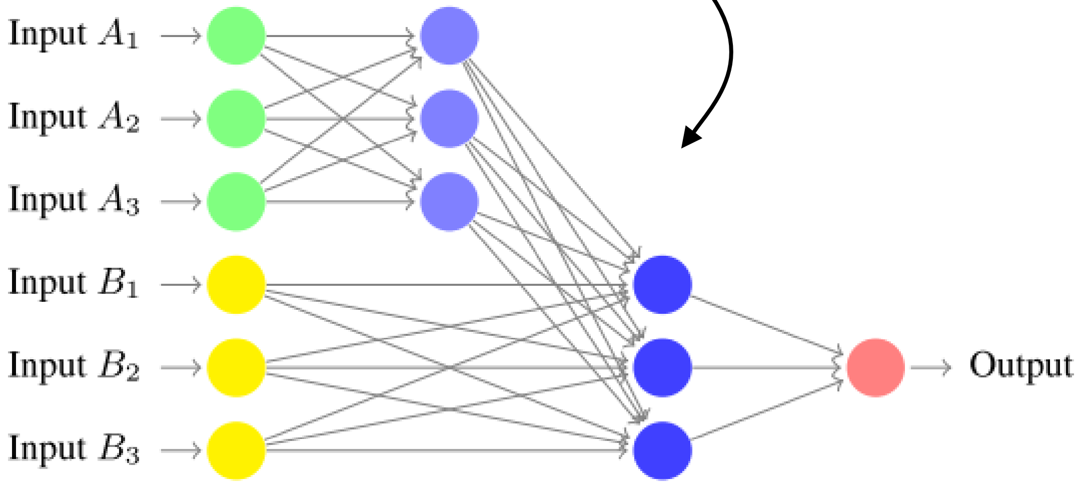
$$CCC = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2}$$

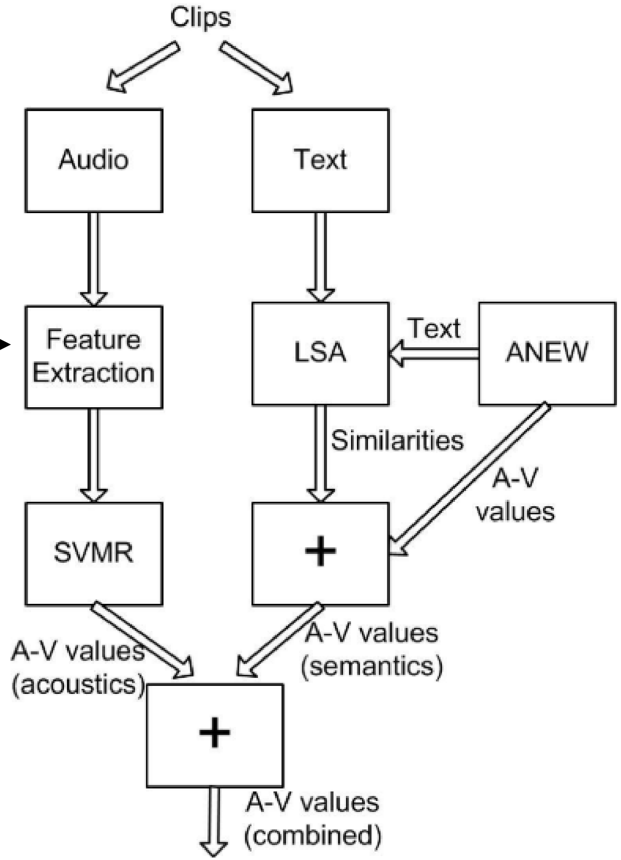
  - A step further than (Pearson) correlation coefficient
  - Penalizes any deviation from the identity relationship (both scale and location/shift)
  - Captures both accuracy and precision
  - Mathematically and experimentally superior to error-based loss functions (Pandit and Schuller, 2020; Atmaja and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

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

Previous Work

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

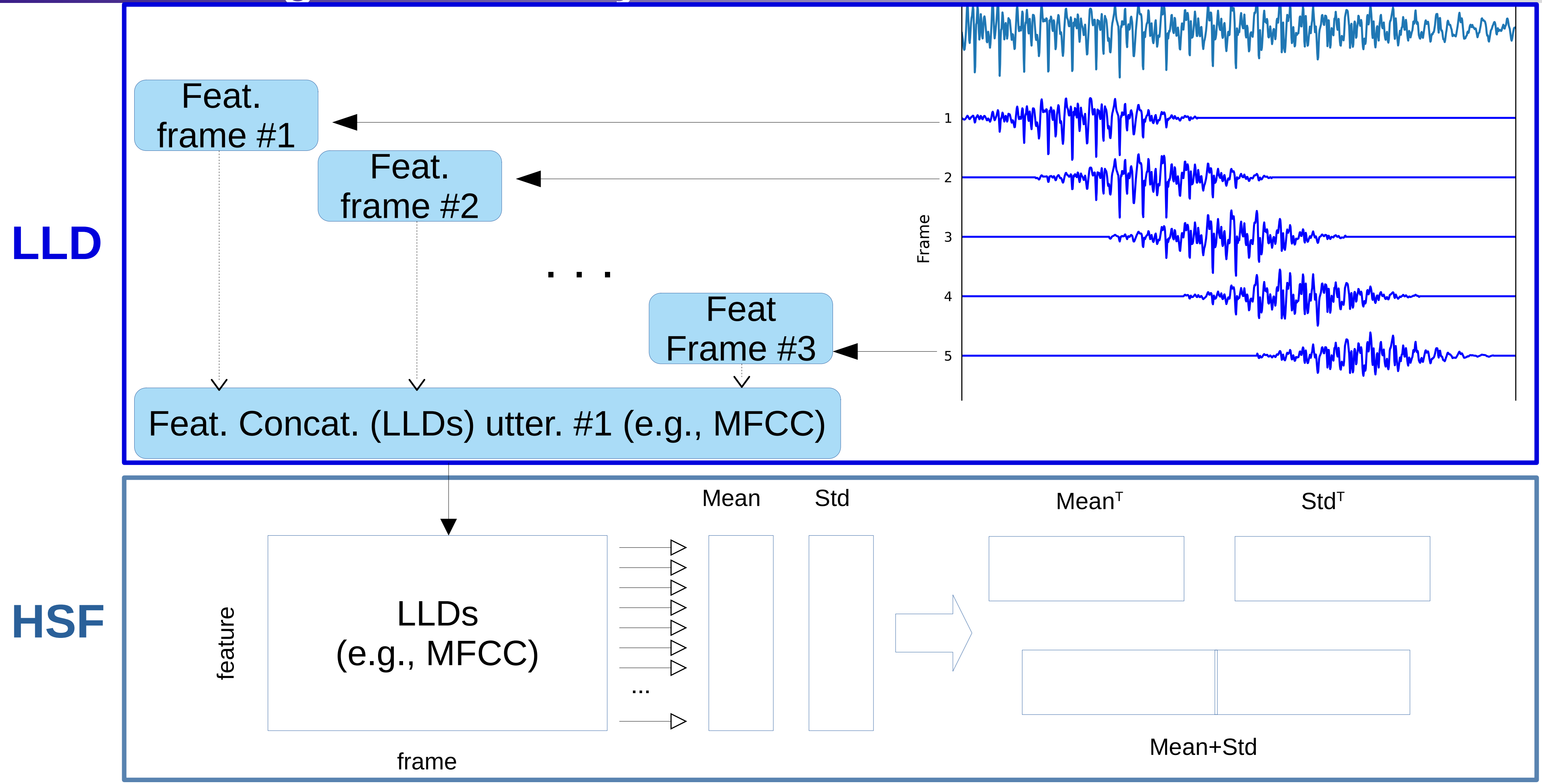




Outline

1. Introduction:
   Background, Aims, Novelty, Significance, Applications
2. Research Methodology:
   Motivation, Problems, Concept, Strategy, Datasets and Evaluation metric, Previous Work
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. HSFs (IEMOCAP data)

LLD

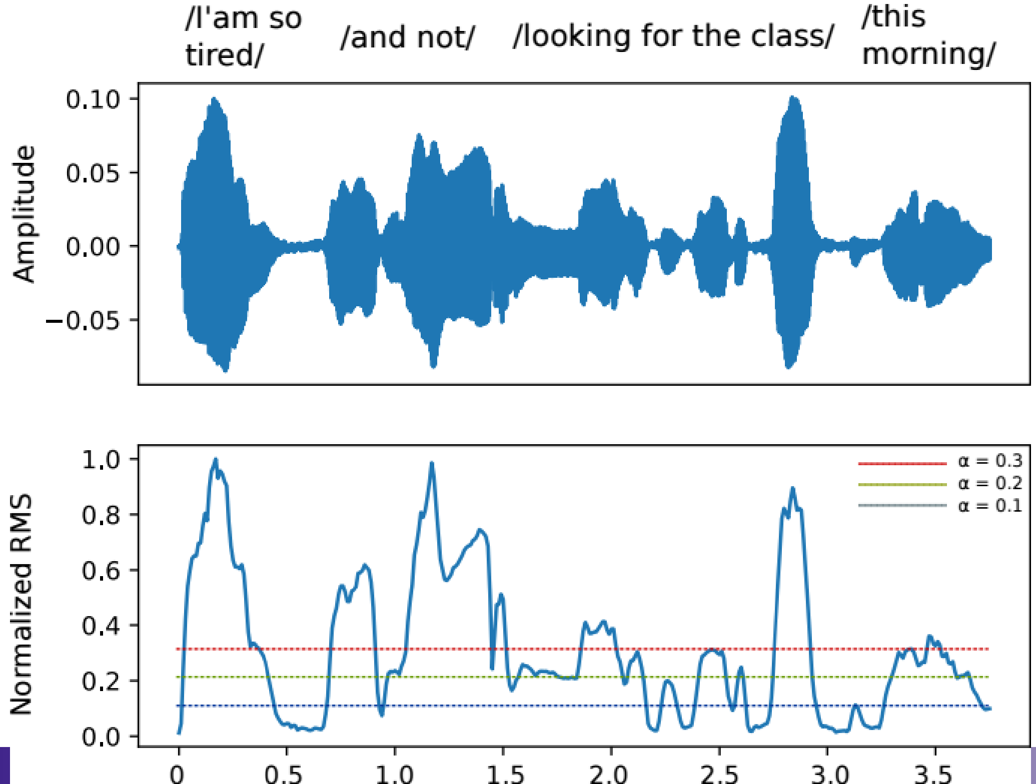
Feature	Dim	Val	Aro	Dom	Mean
MFCC	(3414, 40)	0.148	0.488	0.419	0.352
Log mel	(3414, 128)	0.103	0.543	0.438	0.362
GeMAPS	(3409, 23)	0.164	0.527	0.454	0.382
pAA	(3412, 34)	0.130	0.513	0.419	0.354
pAA_D	(3412, 68)	0.145	0.526	0.439	0.370

HSF
mean+std

Feature	Dim	Val	Aro	Dom	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	136	0.173	0.612	0.455	0.413

Effect of silent pause regions

- Three different treatment 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 silence can be done by using such methods, e.g., 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.

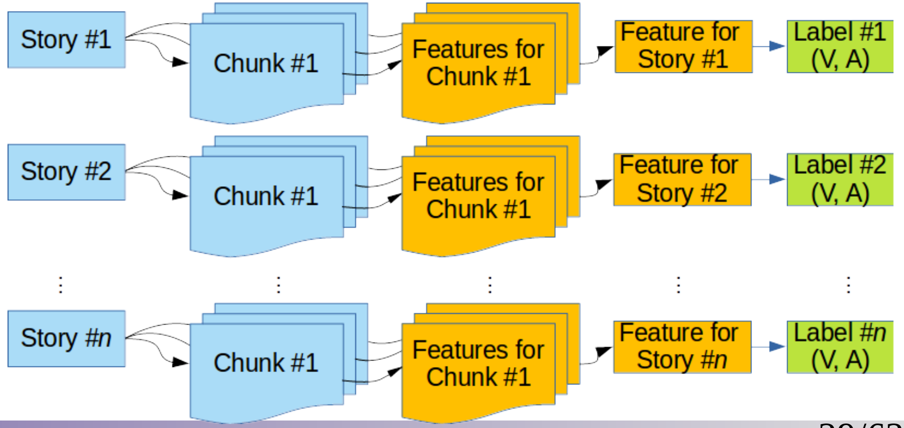
Results: effect of silent pause features

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.

Aggregation methods

- The common methods to aggregate results from many-to-one is by output aggregation, i.e., majority voting.
- However, it is necessary to investigate that other aggregation may perform better, aside from ***aiming at fusing acoustic features with other features (linguistic)***.
- Human may perceive emotion from chunks to utterance based on information aggregation (not decisions/outputs aggregation).
- Thus, two aggregation methods can be evaluated:
  - Acoustic features (input) aggregation:
    - Mean values
    - Max. values
  - Output aggregation:
    - majority voting



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

Summary of Part III

- Proposed solution for the several issues in acoustic-based dimensional SER:

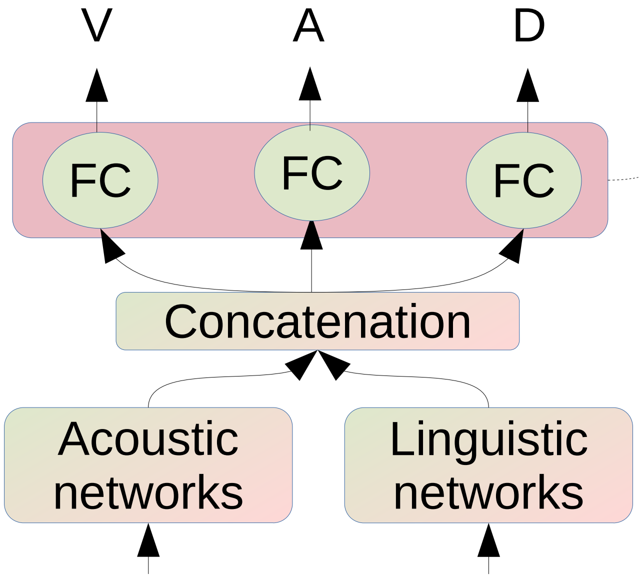
Issue	Proposed method		
Region of analysis	frames		**utterance/fixed length**
Silence region	**removing silence**	keeping silence	**utilizing silence**
Aggregation method	**input aggregation**		output aggregation

- 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, Novelty, Significance, Applications
2. Research Methodology:
   Motivation, Problems, Concept, Strategy, Datasets and
   Evaluation metric, Previous Work
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 MTL



Acoustic features:
pAA LLD
pAA HSF
GeMAPS LLD
GeMAPS HSF

Linguistic features:
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:

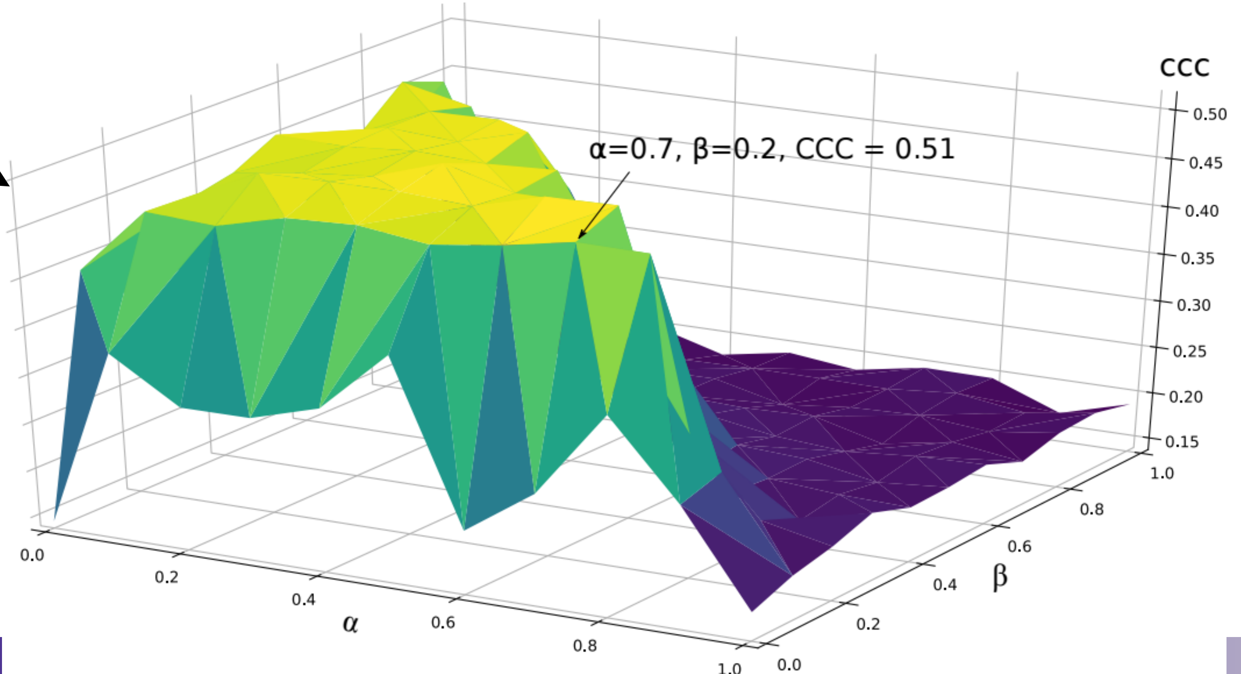
$$\begin{aligned}\mathrm{CCCL}_{tot} = \; & \alpha\; CCCL_V + \beta\; CCCL_A \\ & + (1 - \alpha - \beta)\; CCCL_D\end{aligned}$$

Total loss function with 3 parameters:

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

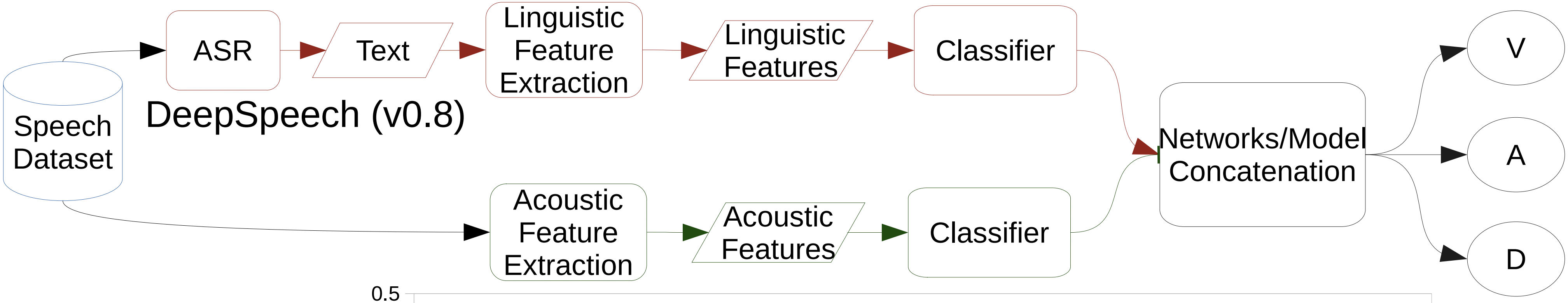
Result: networks concatenation with MTL

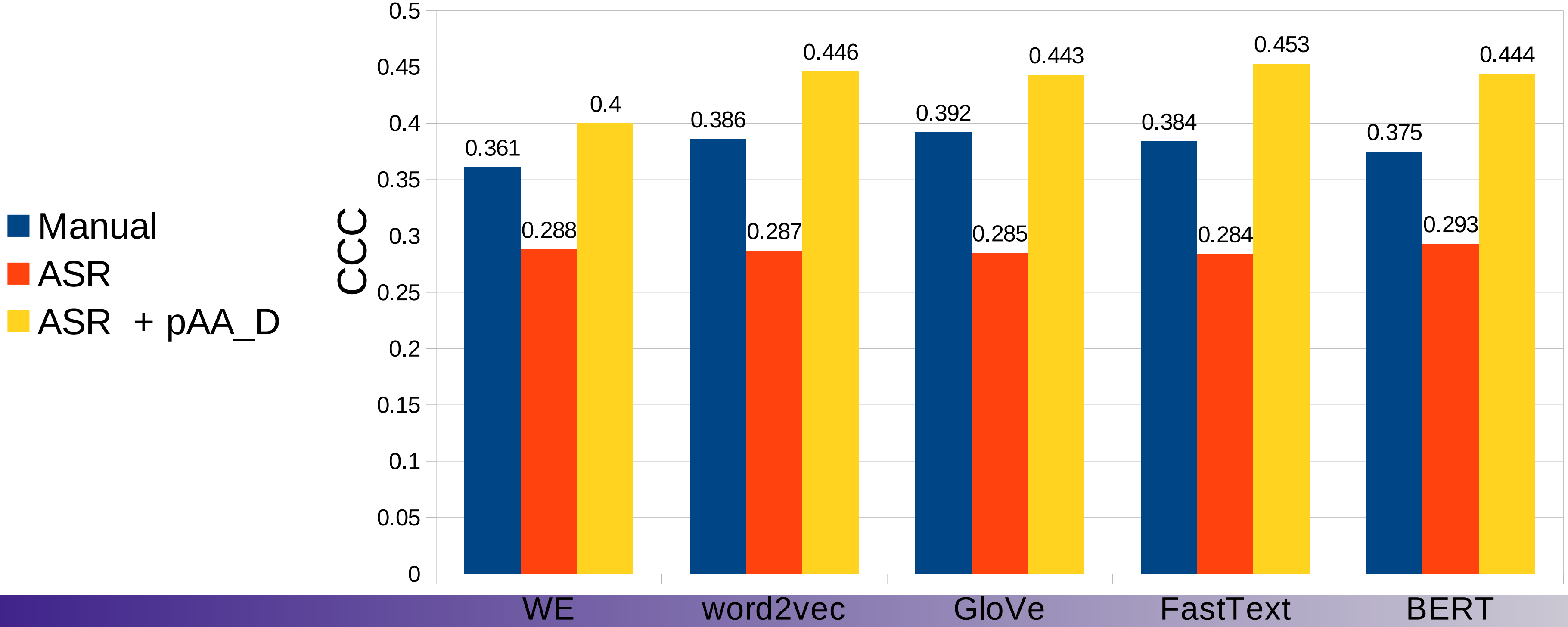
MTL method	V	A	D	Mean
No parameter	0.409	0.585	**0.486**	0.493
2 parameters	**0.446**	**0.594**	0.485	**0.508**
3 parameters	0.419	0.589	0.483	0.497



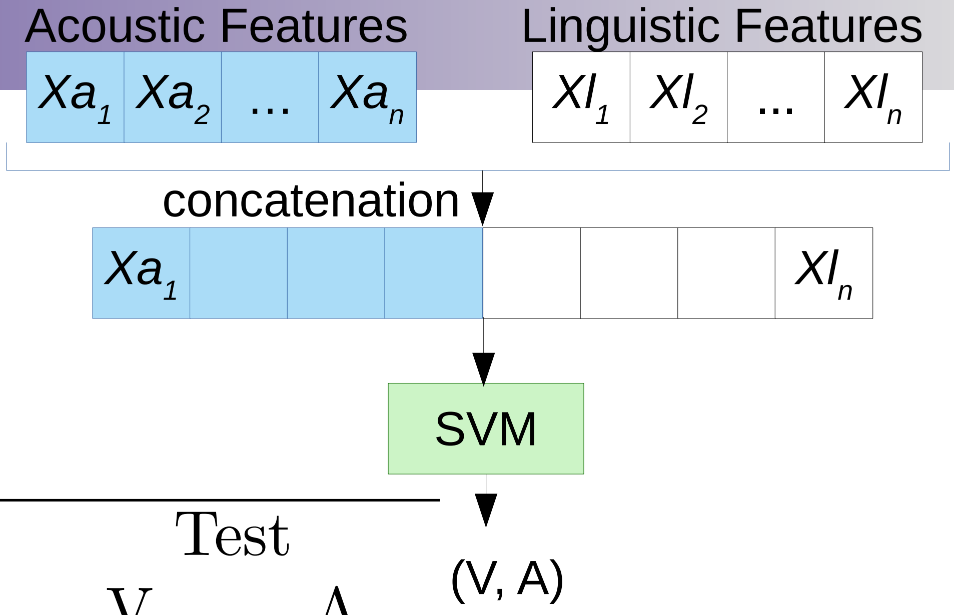
Unimodal parameters:
Aco, α=0.1, β=0.5
Ling, α=0.7, β=0.2

Dimensional SER from ASR outputs (WER=40%)





Feature concatenation



Accuracy (%) 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

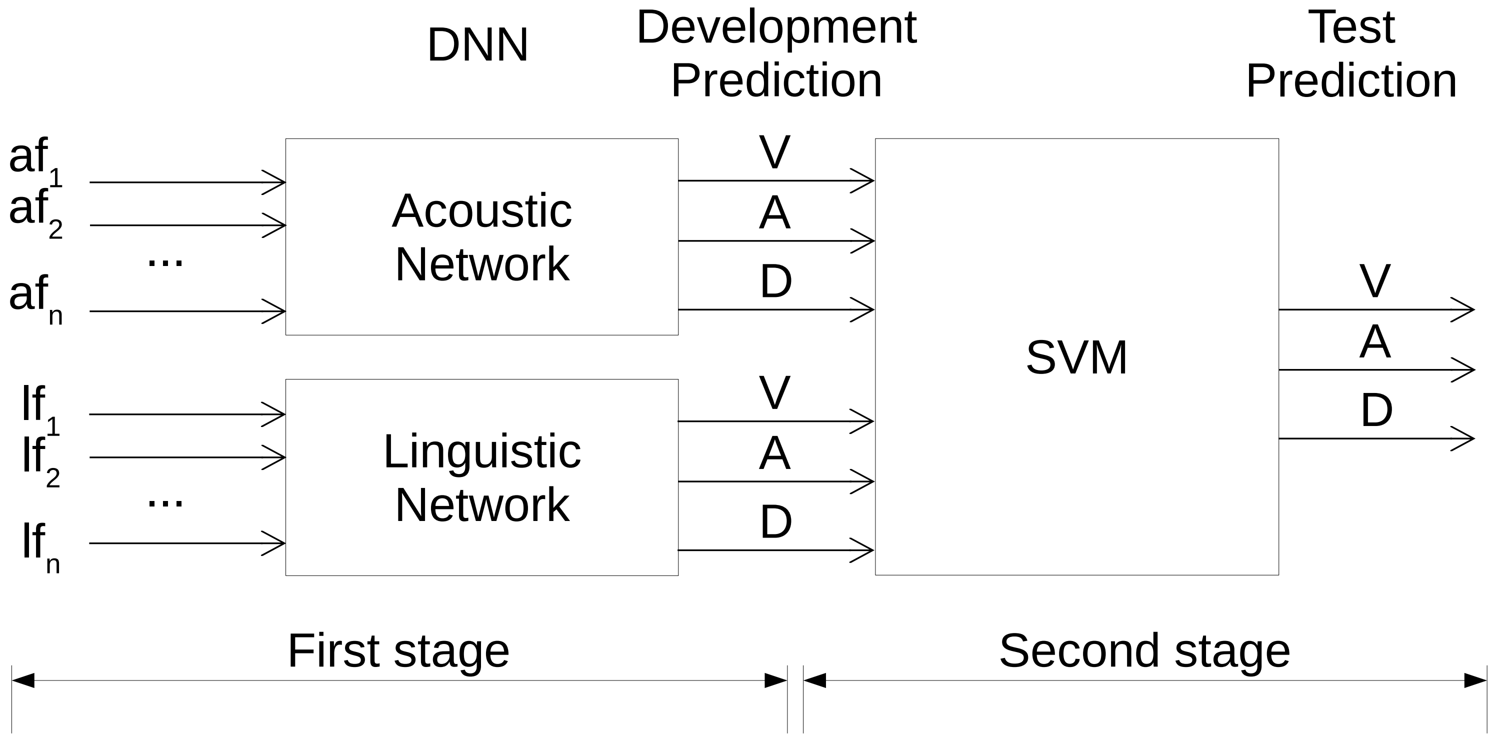
Summary of Part IV

- Fusing acoustic and linguistic information by network concatenation improves dimensional SER in several ways:
  - Linguistic information improves dimensional SER prediction particularly on valence prediction
  - Multitask learning could predict valence, arousal, and dominance simultaneously; the best score was achieved using MTL with two parameters
  - Dimensional SER from ASR outputs resulting in lower performance than manual transcription; pre-trained linguistic model didn't help much in this case
- Feature (input) concatenation improves unimodal emotion recognition on valence prediction

Outline

1. Introduction:
   Background, Aims, Novelty, Significance, Applications
2. Research Methodology:
   Motivation, Problems, Concept, Strategy, Datasets and Evaluation metric, Previous work
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

Two-stage dimensional SER



Input af: GeMAPS LLD, GeMAPS mean+std (HSF1), Gemaps mean+std+sil (HSF2)
Input lf: WE, word2vec, GloVE

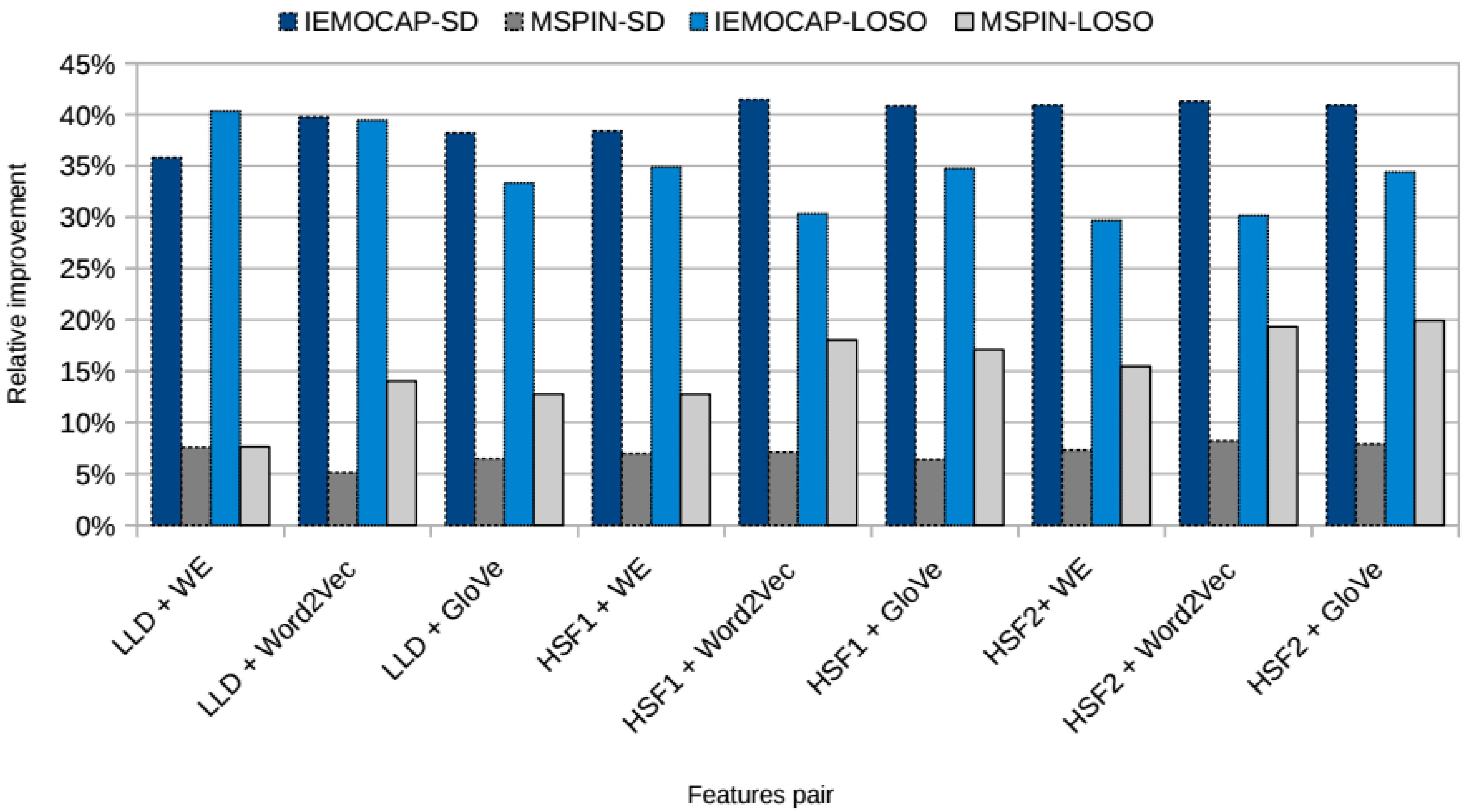
Result: late fusion

CCC scores on different dataset partitions

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’)

Result: relative improvement



Some discussions

- Speaker-dependent vs. speaker-independent
  - The results shows that speaker-dependent and speaker-independent emotion recognition with acoustic-linguistic fusion statistically different ($p < 0.05$)
  - SD scenario can not be used to predict real case scenario (which is speaker-independent)
- Effect of removing target sentence from MSP-IMPROV
  - Removing target sentence still resulted low score of CCCs.
  - There is possibility that the speakers are influenced by target-sentences scenario.
  - Further studies are needed to investigate the influence of lexical content in dimensional SER in different scenarios (when linguistic information is needed and what is the cue).

Summary of Part V

- Late fusion approach improves the performance of the previous early fusion approach in all dimensional emotions.
- As in previous early fusion, linguistic information contributes to valence prediction improvement while acoustic information contributes dominantly to arousal and dominance scores.
- Results on speaker independent is significantly different than speaker dependent.
- Removing lexical-controlled utterances still shows some influence of those utterances; further investigation is needed.

Outline

1. Introduction:

   Background, Aims, Novelty, Significance, Applications

2. Research Methodology:

   Motivation, Problems, Concept, Strategy, Datasets and Evaluation metric, Previous work

3. Dimensional SER Using Acoustic Features
4. Early Fusion of Acoustic and Linguistic Information
5. Late Fusion of Acoustic and Linguistic Information
6. **Conclusions:**

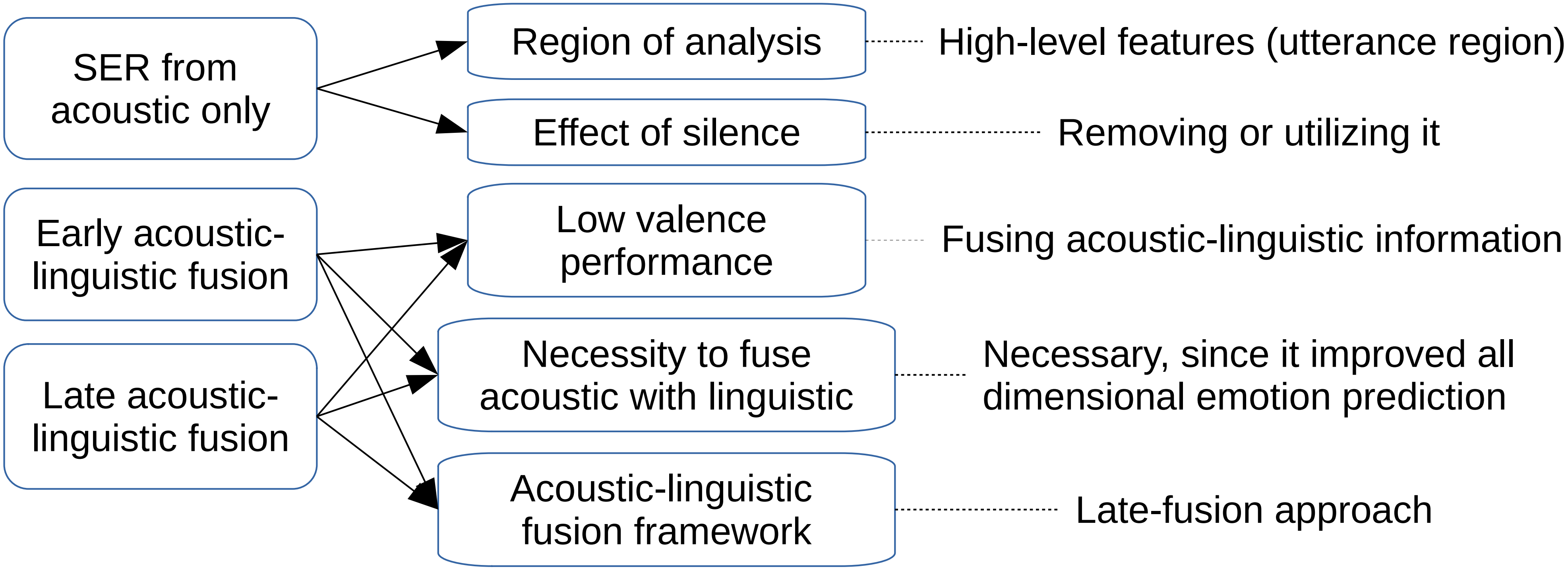
   Comparative analysis, Summary, Contributions, Future research

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

- Dimensional SER from acoustic features
  - ✓ Statistical features representation (particularly Mean+Std) shows meaningful impact in general acoustic feature set
  - ✓ Silent pause regions is predicted to contribute to dimensional SER; either removing silence or utilizing it as additional features slightly improve the performance
  - ✓ Mapping many-to-one from short terms (chunks) for long term (story) is better modeled by feature aggregation than output aggregation

  – Remains:
    - A method to calculate silent pause features that discriminate significantly among removing, keeping, and utilizing silence (e.g., TFS-ENV method)
    - The contribution of fusing LLD and HSF compared to individual region of analysis, the thresh-hold, and its complexity

Contributions (Cont'd)

- Dimensional SER using acoustic-linguistic information fusion
  - ✔ Dimensional SER can be performed simultaneously by early fusion multitask learning based on CCC loss; CCCL with two parameters is the most accurate method to model interrelation among emotion dimension
  - ✔ Late fusion approach is better to model fusion of acoustic and linguistic information, which also is perceptually closer to human multimodal processing than early fusion approach
  - ✔ Linguistic information improves prediction of valence while acoustic information dominantly influences prediction of arousal and dominance

  - Remains:
    - Fine-tuned BERT on acoustic-linguistic dimensional SER
    - Fully lexical-controlled dimensional SER

Future research direction

- Accelerating high-level feature extraction for speech emotion recognition
- Bimodal late-fusion approach by output aggregation (manually determined or majority voting)
- Bimodal acoustic-linguistic emotion 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

Publications

- Journals (3):
  1) 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.
  2) 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.
  3) B.T. Atmaja, and M. Akagi. “Two-stage dimensional emotion recognition by fusing predictions of acoustic and text networks using SVM,” Speech Communication, vol 126, February, 2021, pp 9-21. doi:10.1016/j.specom.2020.11.003.

- International conferences (10):
  1) B.T. Atmaja, K. Shirai, and M. Akagi, ``Deep Learning-based Categorical and Dimensional Emotion Recognition for Written and Spoken Text,'' International Seminar on Science and Technology, Surabaya, 2019.
  2) B. T. Atmaja and M. Akagi, ``Speech Emotion Recognition Based on Speech Segment Using LSTM with Attention Model,'' in 2019 IEEE International Conference on Signals and Systems (ICSigSys), 2019, pp. 40--44
  3) B. T. Atmaja, K. Shirai, and M. Akagi, ``Speech Emotion Recognition Using Speech Feature and Word Embedding,'' in 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2019, pp. 519–523.
  4) B. T. Atmaja and M. Akagi, ``Multitask Learning and Multistage Fusion for Dimensional Audiovisual Emotion Recognition,'' in ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 4482--4486.

Publications (cont'd)

5) B. T. Atmaja and M. Akagi, ``The Effect of Silence Feature in Dimensional Speech Emotion Recognition," in 10th International Conference on Speech Prosody 2020, May, pp. 26.
6) B.T. Atmaja and M. Akagi, ``Improving Valence Prediction in Dimensional Speech Emotion Recognition Using Linguistic Information, " in 2020 Oriental COCOSDA, pp. 166-171. IEEE, 2020. [best student paper]
7) B.T. Atmaja and M. Akagi, ``On The Differences Between Song and Speech Emotion Recognition: Effect of Feature Sets, Feature Types, and Classifiers", TENCON 2020, Osaka, Japan, 2020.
8) B.T. Atmaja, Y. Hamada and M. Akagi, ``Predicting Valence and Arousal by Aggregating Acoustic Features for Acoustic-Linguistic Information Fusion" TENCON 2020, Osaka, Japan, 2020.
9) B.T. Atmaja and M. Akagi, 'Deep Multilayer Perceptrons for Dimensional Speech Emotion Recognition," in 2020 APSIPA ASC, Auckland, New Zealand, 2020.
10) 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. [best student paper & presentation]

- Domestic conferences (4):

1) R. Elbarougy, B.T. Atmaja, M. Akagi, ``Continuous Tracking of Emotional State from Speech Based on Emotion Unit," ASJ Autumn 2018.
2) B.T. Atmaja, A.N.F. Fandy, D. Arifianto, M. Akagi, ``Speech recognition on Indonesian language by using time delay neural network," ASJ Spring 2019.
3) B.T. Atmaja, R. Elbarougy, M. Akagi, ``RNN-based dimensional speech emotion recognition," ASJ Autumn 2019.
4) B.T. Atmaja, M. Akagi, ``Dimensional Speech Emotion Recognition from Acoustic and Text Features Using Multitask Learning," ASJ Spring 2020.

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

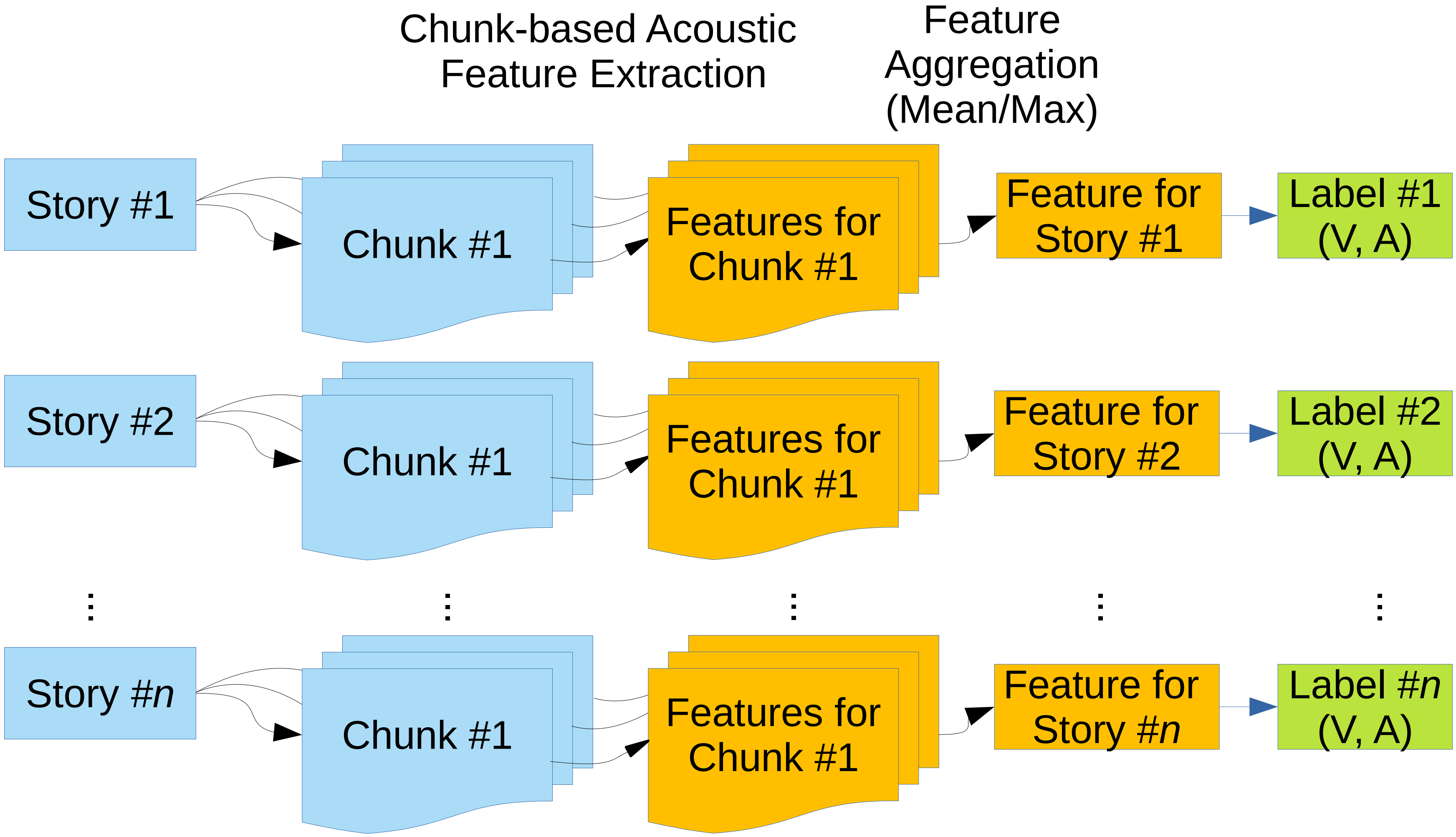
List of abbreviation

- 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

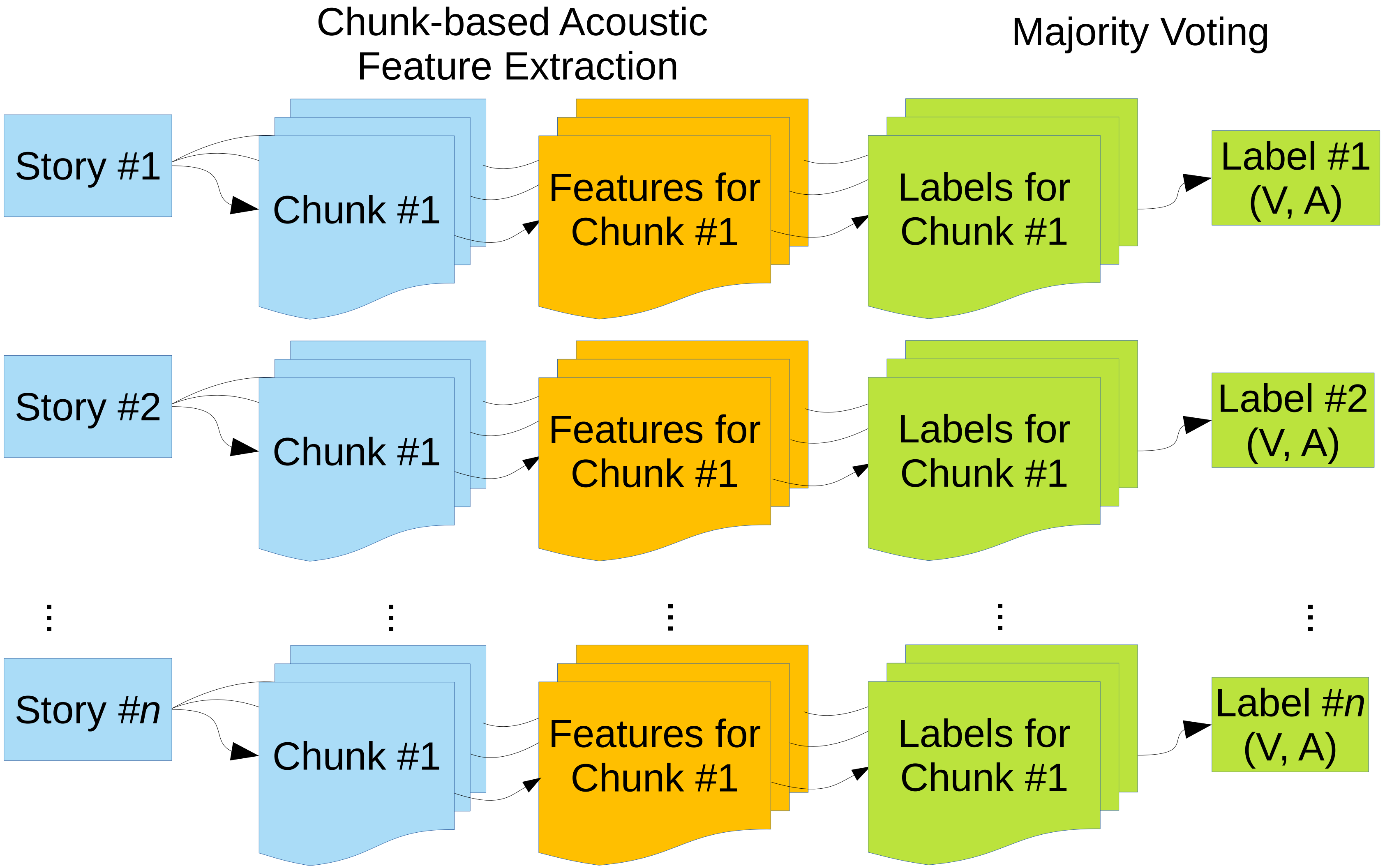
List of abbreviation (Cont’d)

- 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

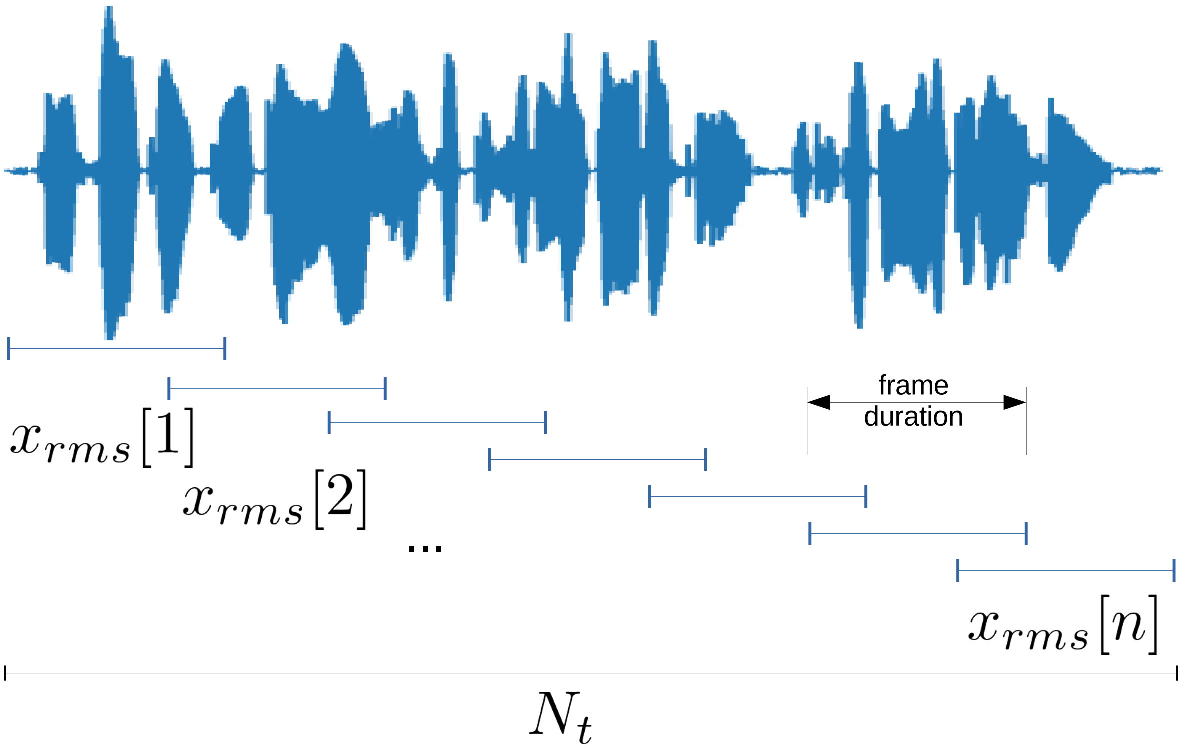
Acoustic feature aggregation



Outputs aggregation



Calculating silent pause features (sf)



Silent pause feature is calculated by

$$sf = \frac{N_s}{N_t}$$

where

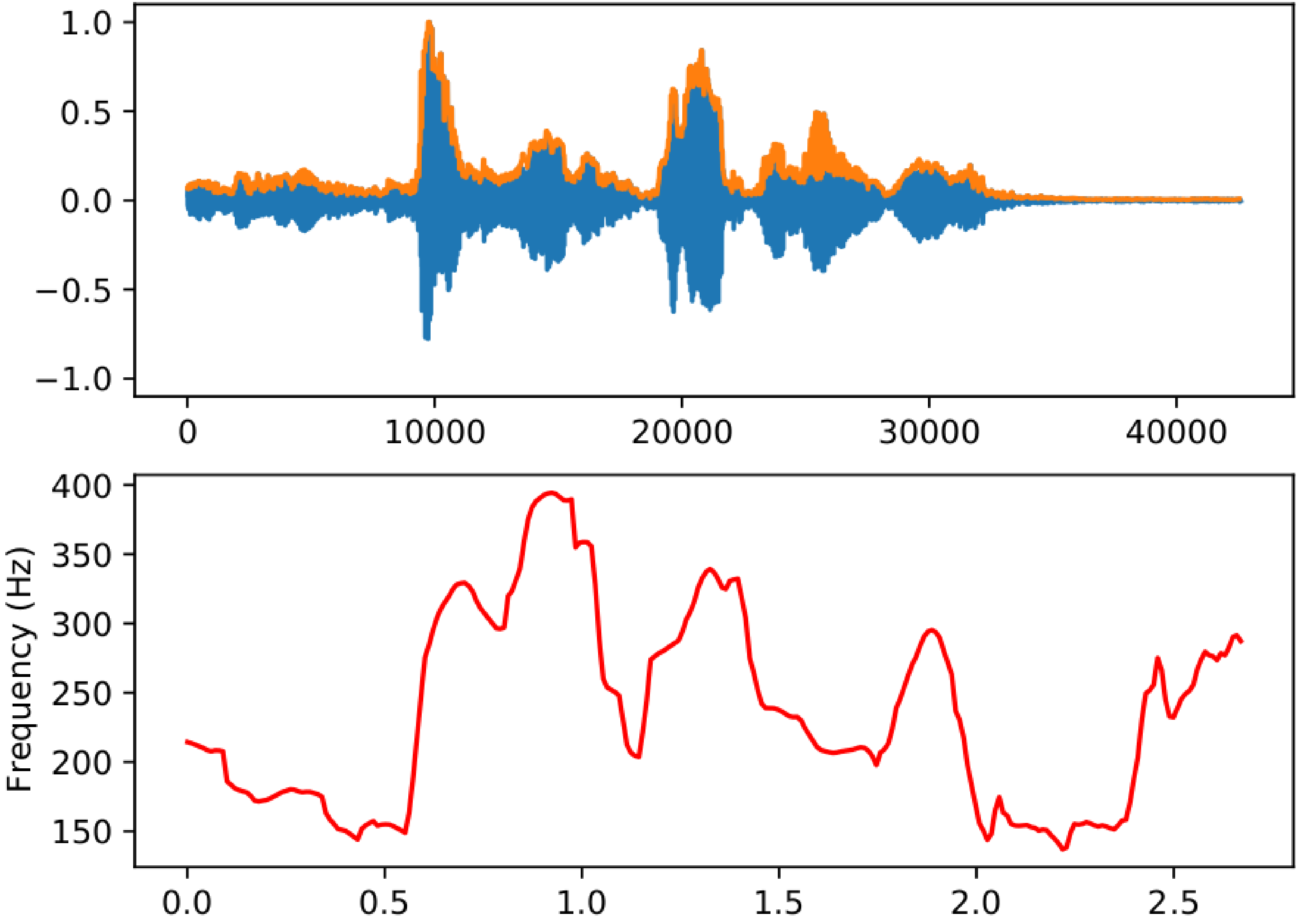
Ns: Number of silence frames

Nt: Total frames

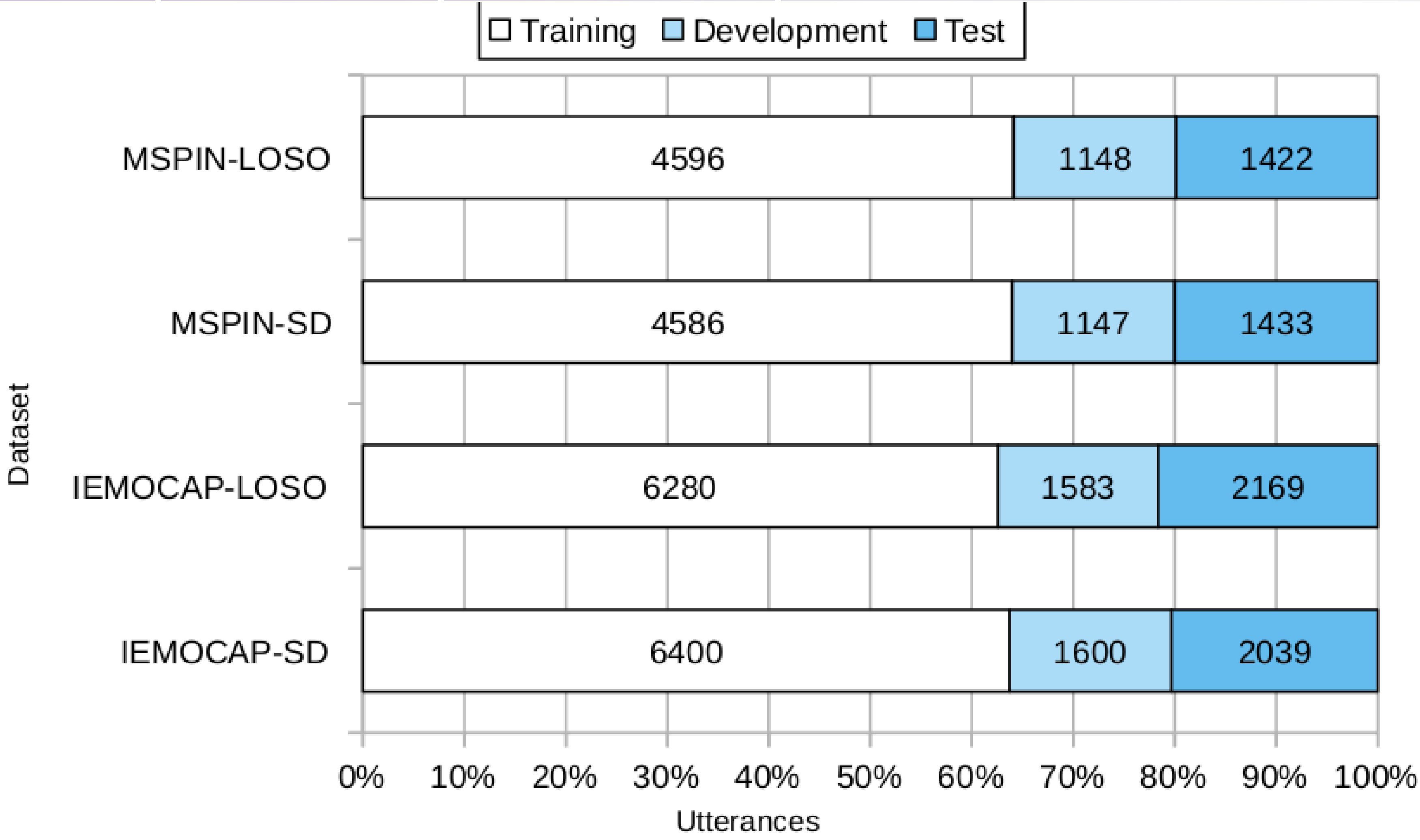
A frame is categorized as silence if the RMS is below threshold (th)

$$th = \alpha \times \tilde{x}_{rms}$$

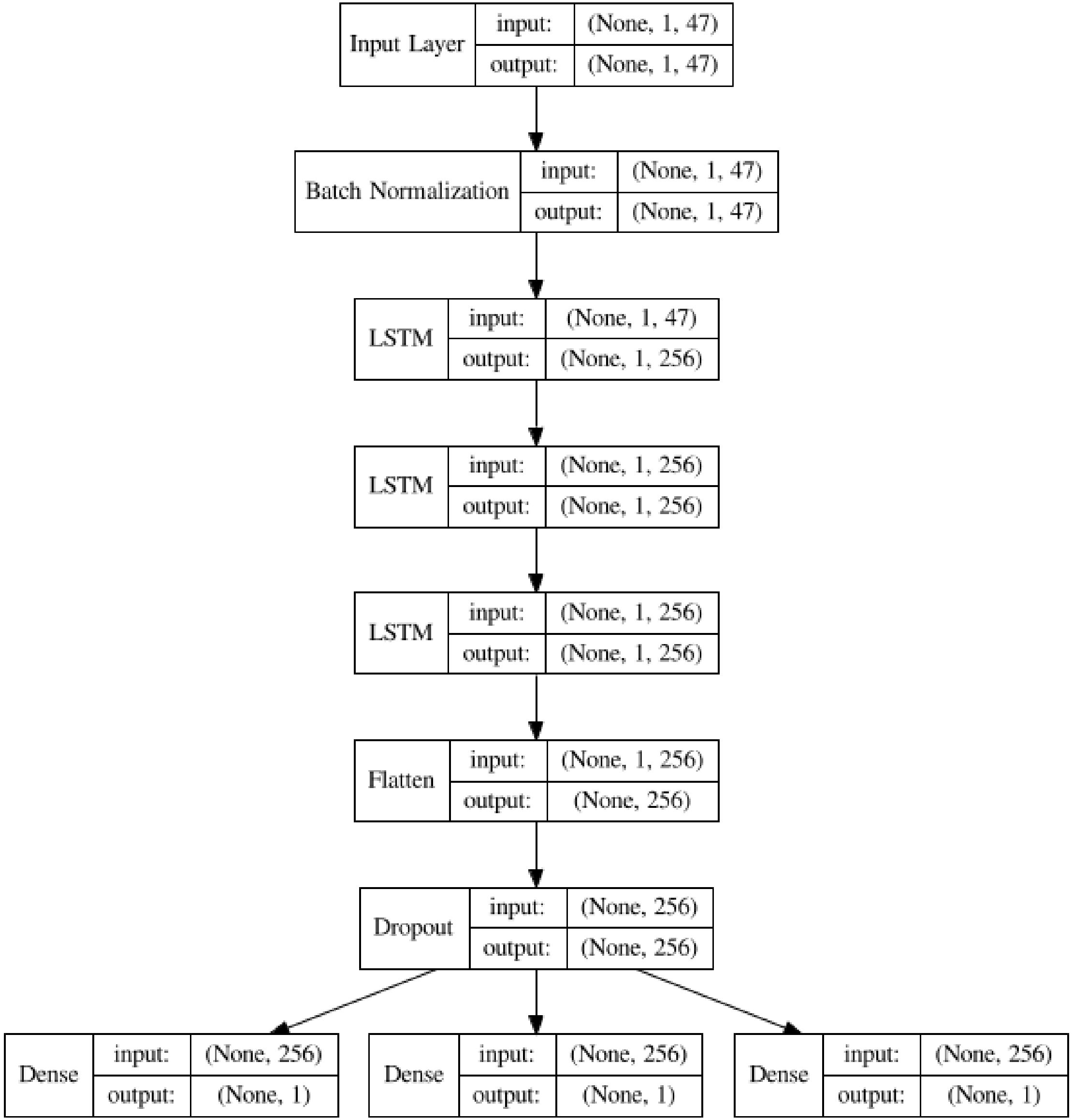
ENV and F0 contour



Dataset partition (late fusion)



DNN Model (Acoustic)



DNN Model (Linguistic)