Deep learning for speech emotion recognition: Combining acoustic and linguistic information

Bagus Tris Atmaja
bagus@ep.its.ac.id

VibrasticLab, ITS
Acoustic Information Science Lab., School of Information Science, JAIST

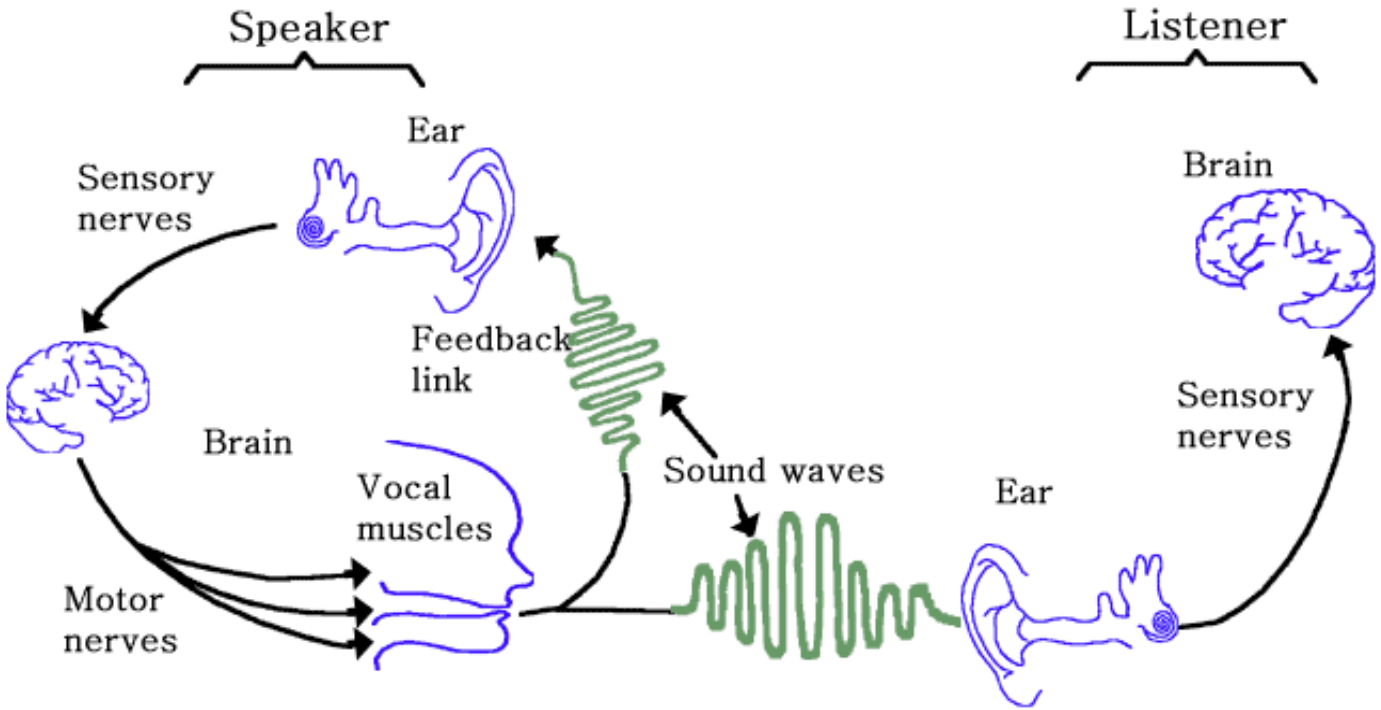
This slide is available at https://intip.in/ser01

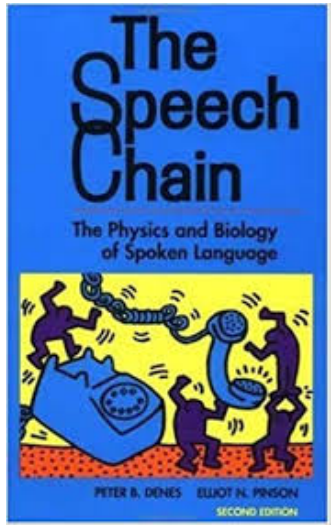
Outline

1. **Introduction:**
   Background, Applications, Companies
2. Emotion Recognition:
   Emotion models, Datasets, Evaluation metric, Tools
3. Acoustic and Linguistic Features
4. Classifiers
5. Fusion methods
6. Conclusions: Summary, Future research

DEMO

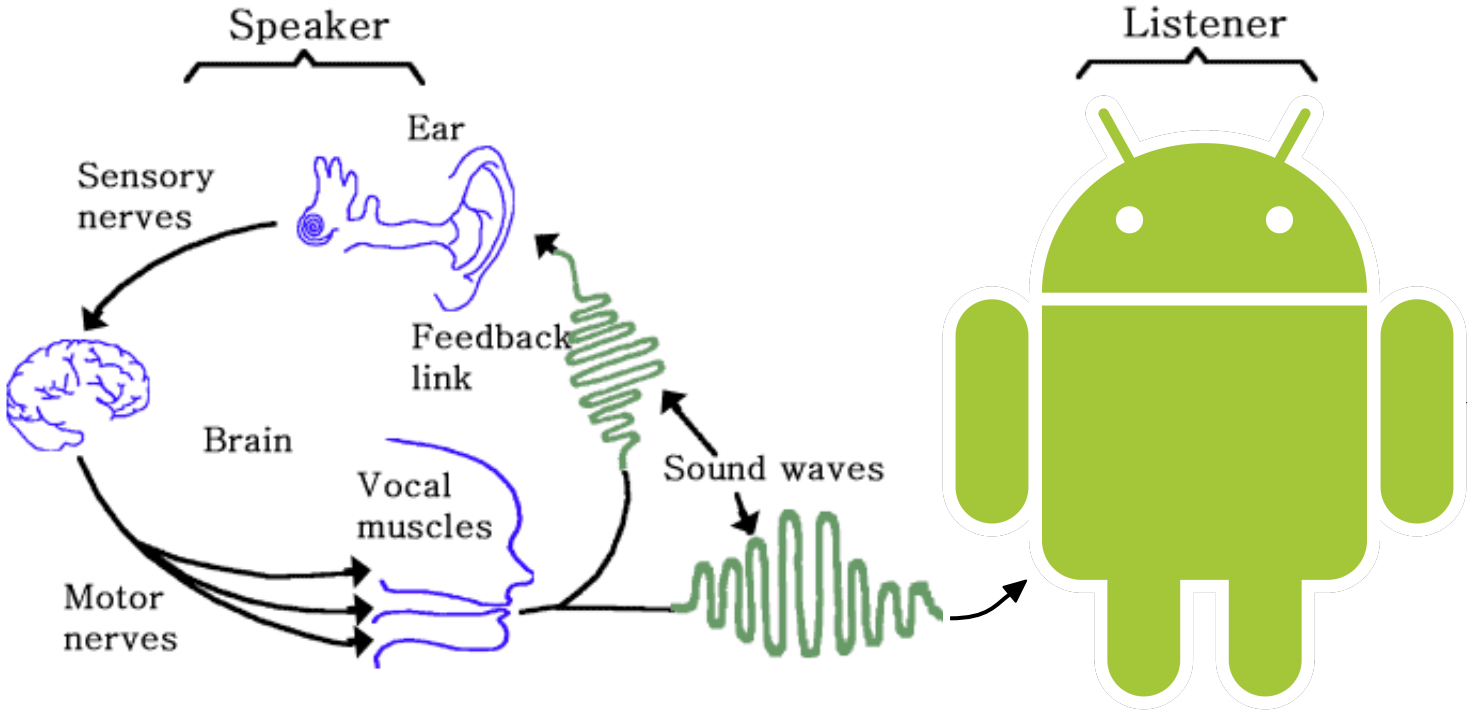
The speech chain (Denes and Pinson [Bell], 1993)





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

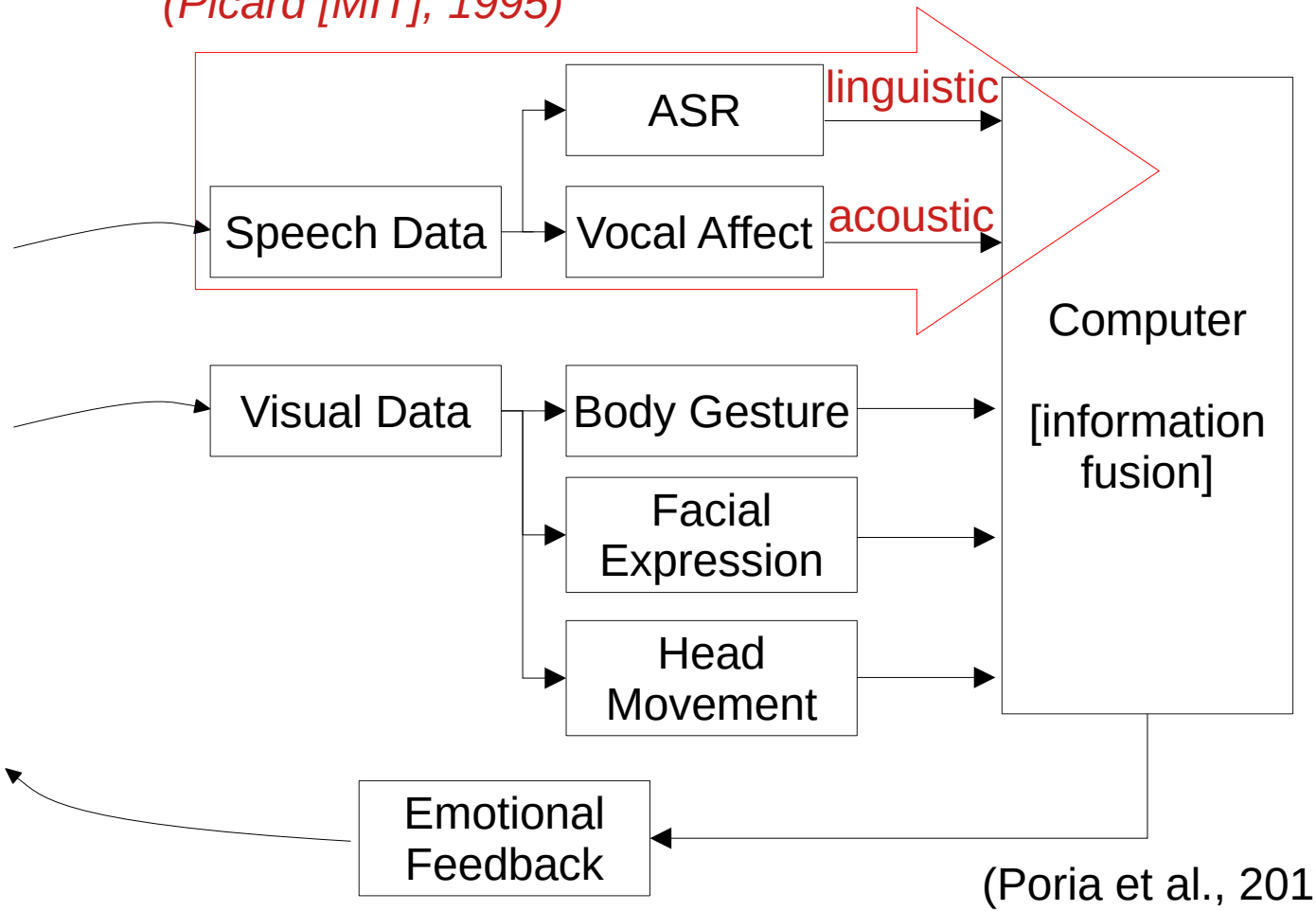
Human-machine communication



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 [MIT], 1995)

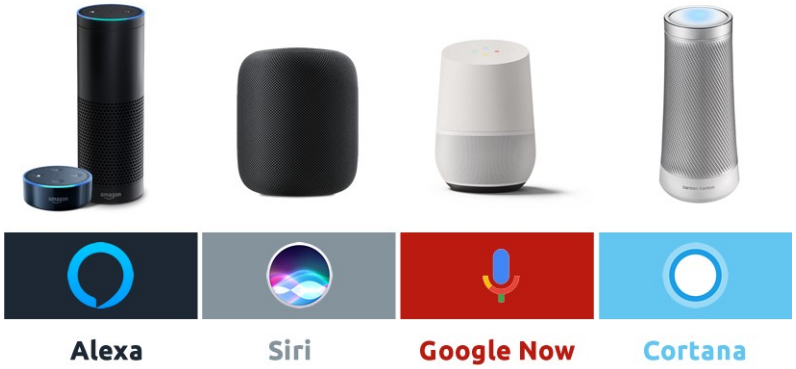


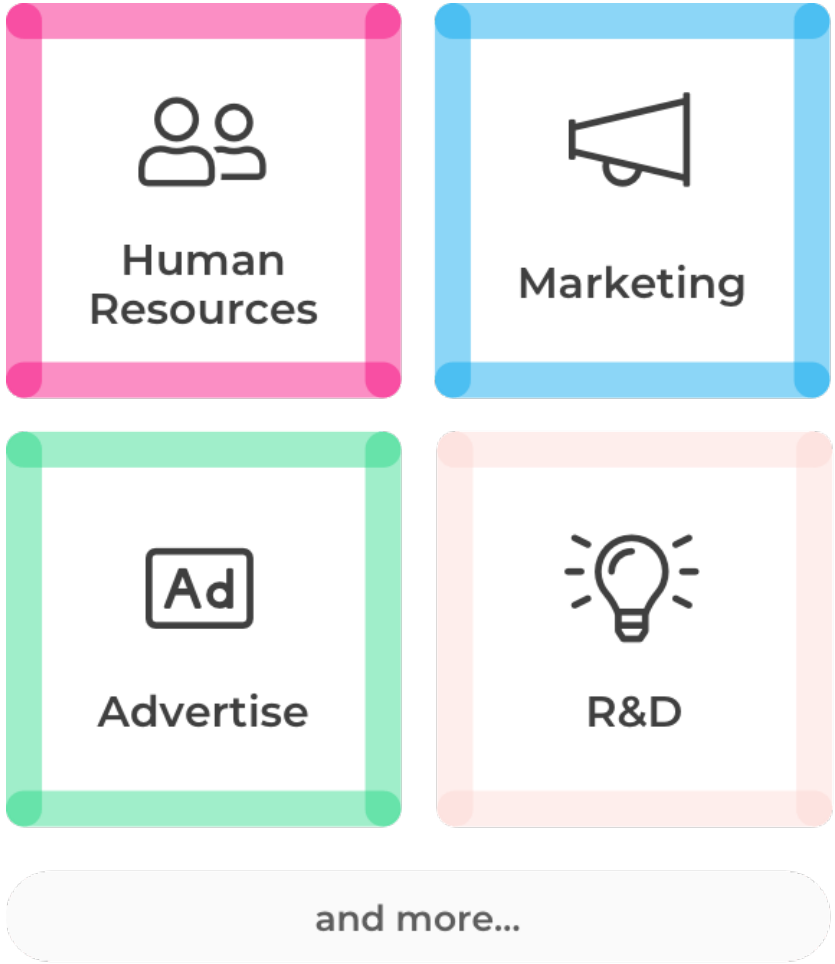
(Poria et al., 2017)

Possible applications

- Contact/Call center application

- Voice assistant





Some companies working in emotion recognition

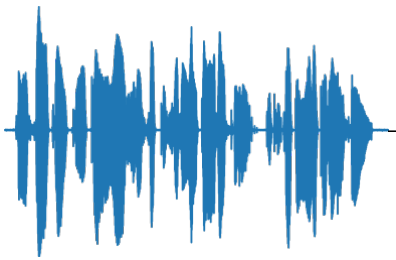
Outline

1. Introduction:
   Background, Applications, Companies
2. **Emotion Recognition:**
   Emotion models, Datasets, Evaluation metric, Tools
3. Acoustic and Linguistic Features
4. Classifiers
5. Fusion methods
6. Conclusions: Summary, Future research

DEMO

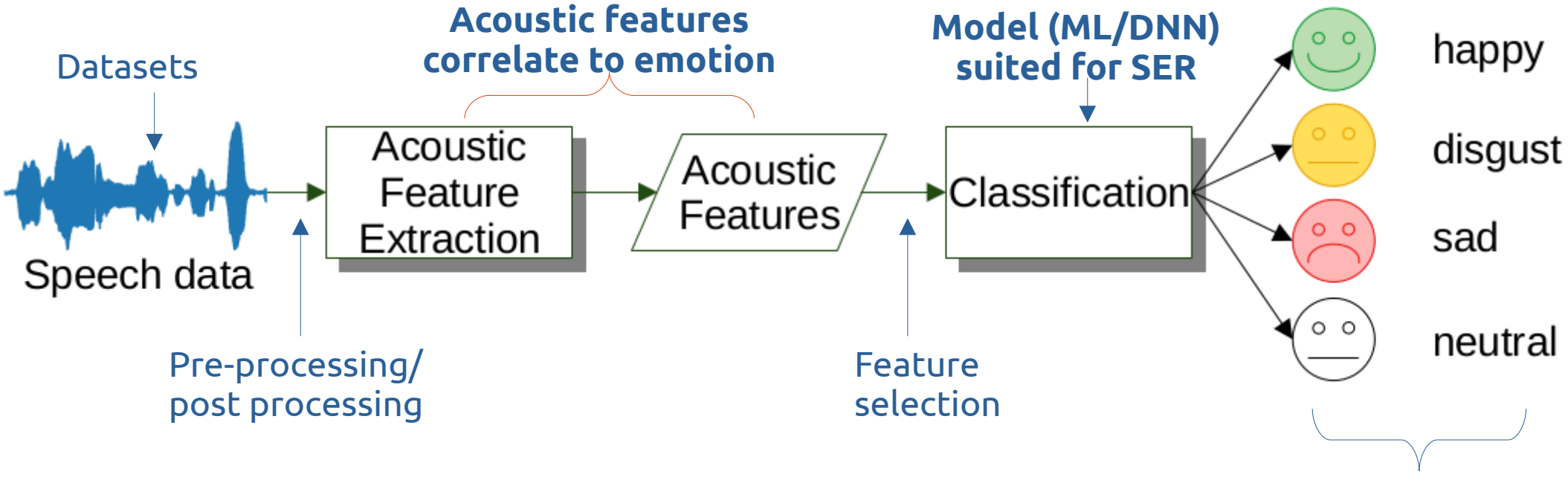
Speech emotion recognition (SER)

- SER: part of affective computing which focuses on *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)
Gender	Female
Speaker	Speaker #2
Age	25 – 54 years old
Emotion	Angry (V: 1, A:4, D:4)

Speech emotion recognition (SER) work flow



Typical workflow of SER Research

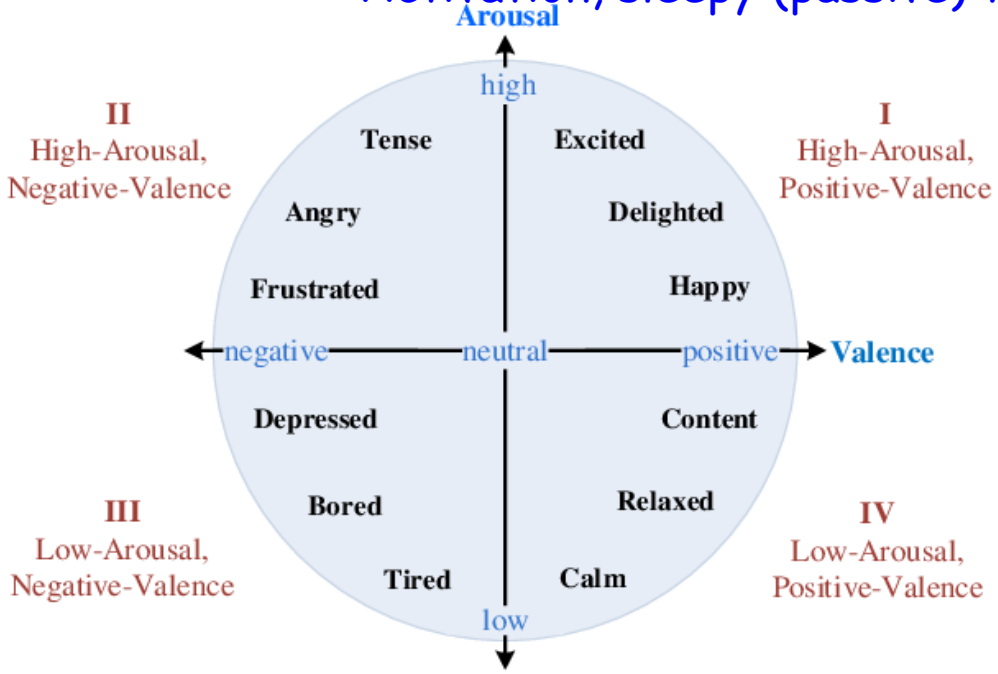
Emotion model

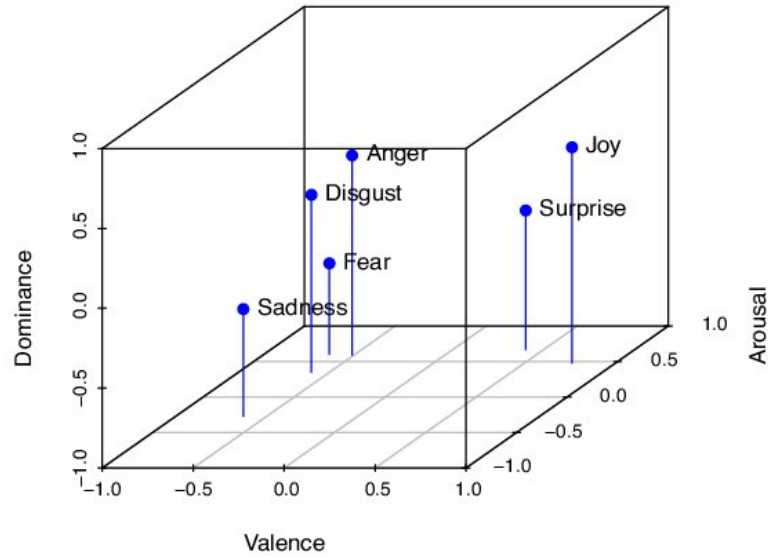
- Categorical emotion: *discrete* emotions e.g happiness, sadness, surprise, fear & anger
- Dimensional emotion: emotion as *continuous* degree in several attributes/dimensions
- Most common dimensions: Valence, Arousal, and Dominance

Pleasantness, positive to negative

Activation, sleepy (passive) to awake (active)

Power control, low to high





Datasets

IEMOCAP

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

MSP-IMPROV

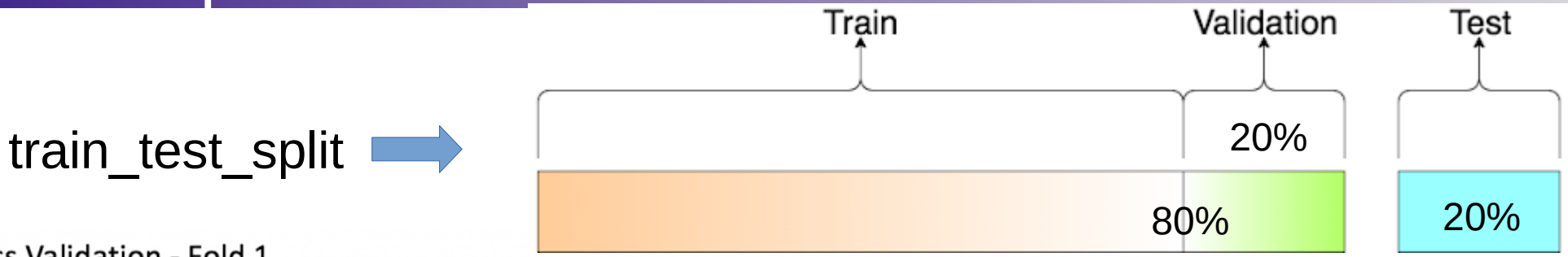
> 9 hours long
8438 turns
12 speakers
6 sessions
4 emotions
V, A, D [1-5]

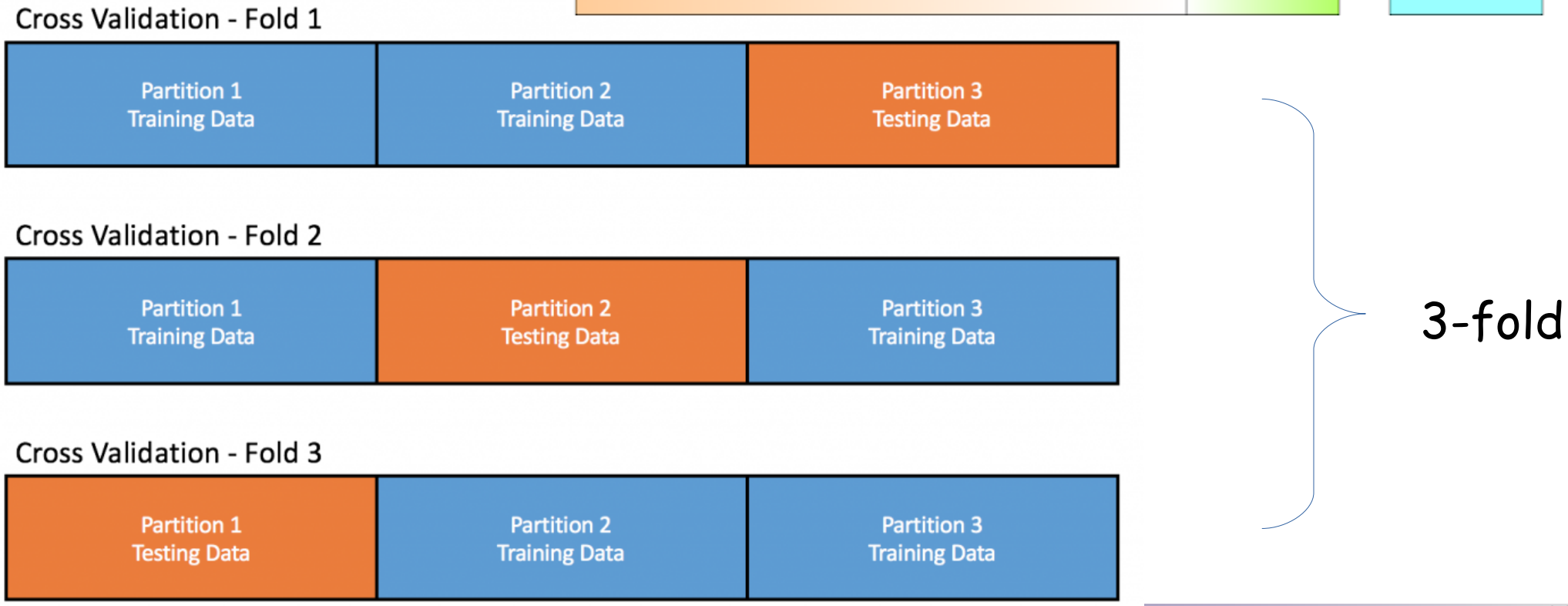
USOMS-e

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

Other datasets: EmoDB (Germany), **RAVDESS** (US), Fujitsu DB (JP), Keio DB (JP), SEWA (Multi), SEMAINE (En), etc.

Dataset partition and cross validation





Evaluation metric

- Categorical model:

= weighted average recall (WAR)

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Recall = \frac{TP}{TP + FN}$$

	Emotion #1 Predicted	Emotion #2 Predicted
Emotion #1 Actual	True Positive (TP)	False Negative (FN)
Emotion #2 Actual	False Positive (FP)	True Negative (TN)

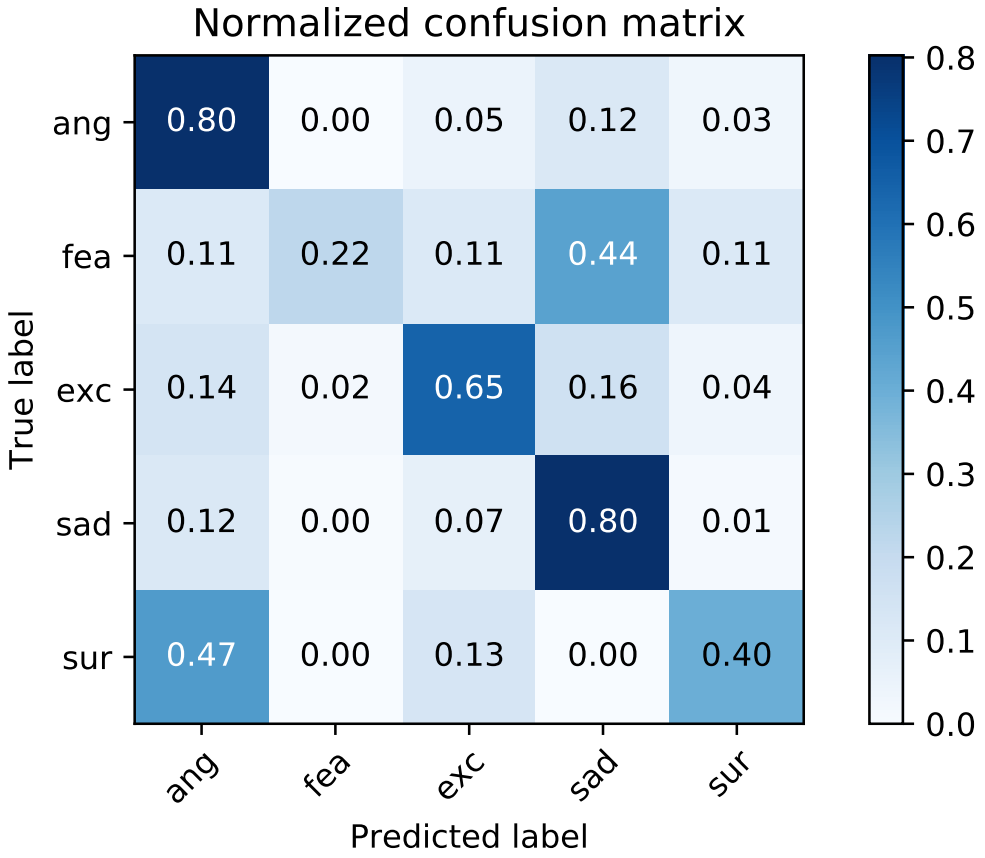
Unweighted Average Recall (UAR) = mean(Recalls)

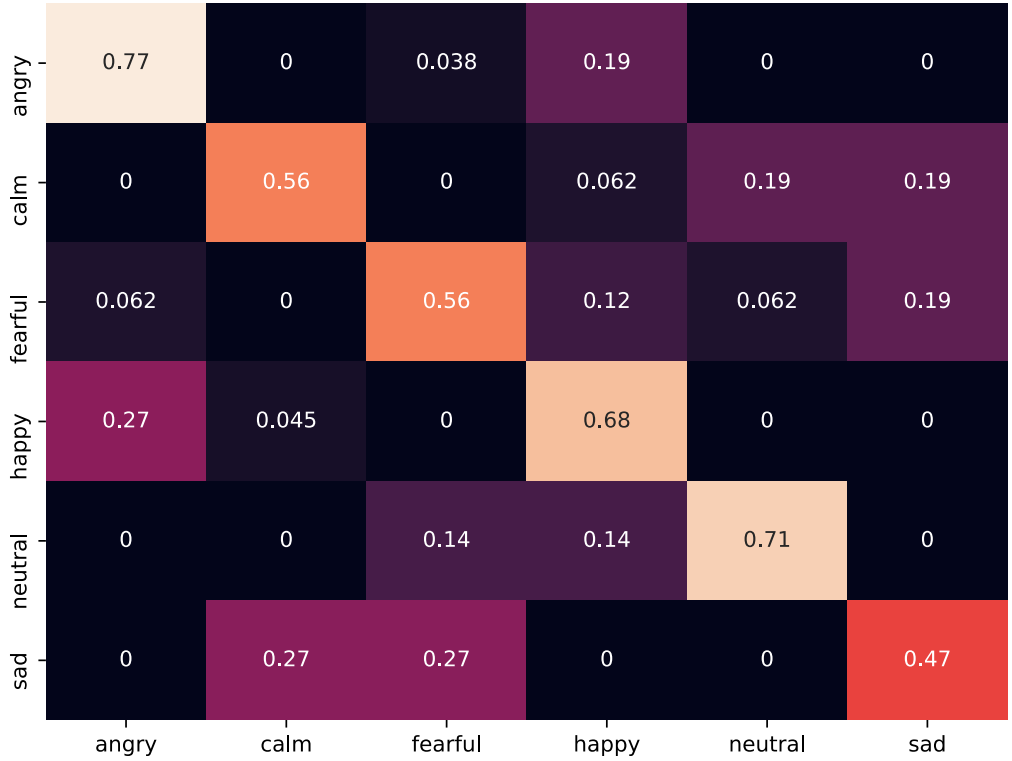
- Dimensional model: Concordance correlation coefficient (CCC)

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

from 0 to 1,
higher is better

Evaluation metric: Confusion matrix





Tools

- Acoustic feature extraction:
  - Librosa (https://librosa.org)
  - pyAudioAnalysis (https://github.com/tyiannak/pyAudioAnalysis)
  - openSMILE (https://audeering.github.io/opensmile-python/)
- Linguistic feature extraction
  - Keras/tensorflow
  - FastText (https://fasttext.cc/)
  - GloVe (https://nlp.stanford.edu/projects/glove/)
  - BERT (https://github.com/google-research/bert)
- Classifier
  - scikit-learn (www.scikit-learn.org)
  - keras (www.keras.io)

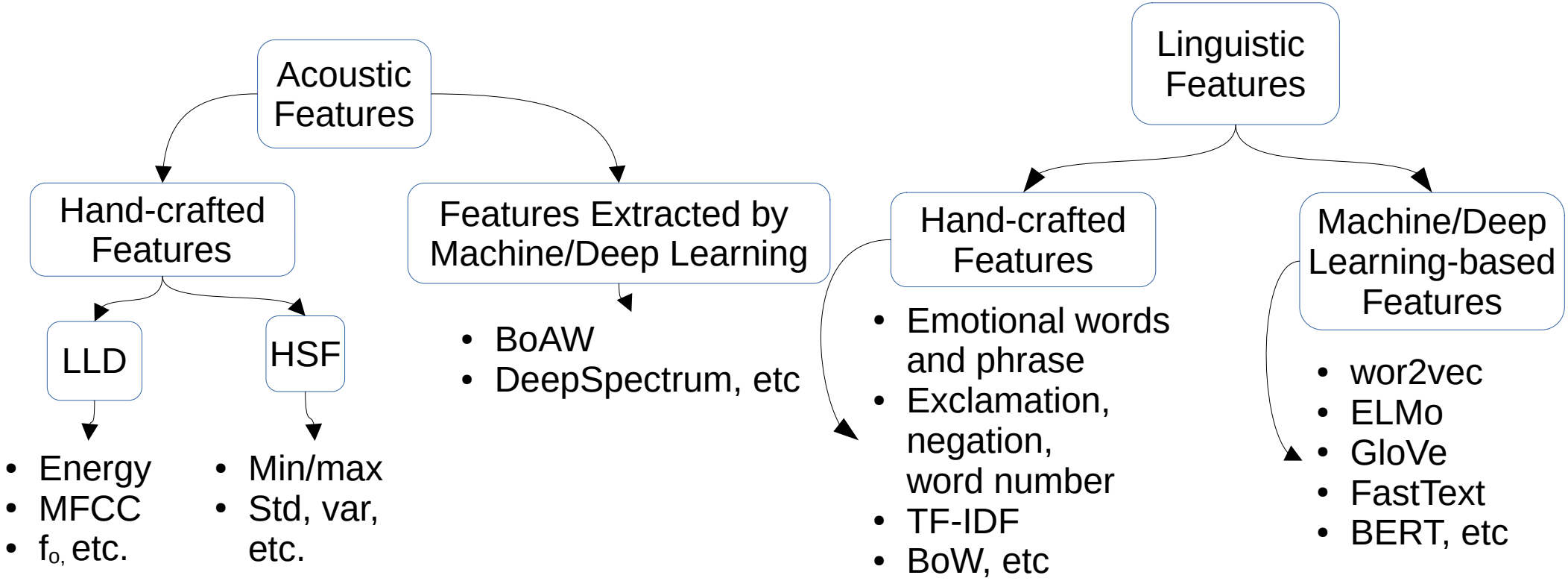
Outline

1. Introduction:
   Background, Applications, Companies
2. Emotion Recognition:
   Emotion models, Datasets, Evaluation metric, Tools
3. **Acoustic and Linguistic Features**
4. Classifiers
5. Fusion methods
6. Conclusions: Summary, Future research

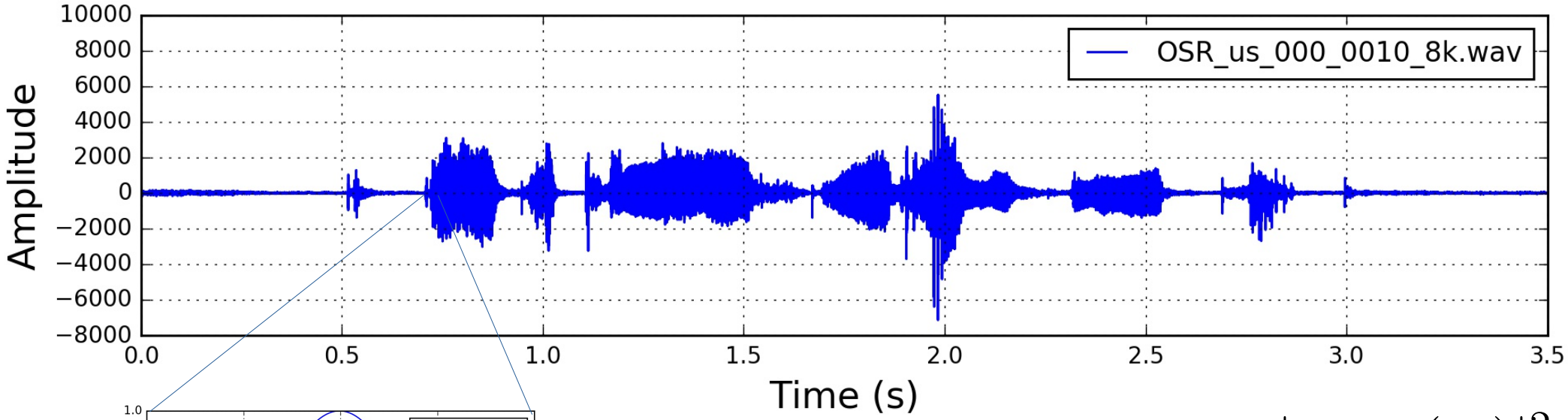
DEMO

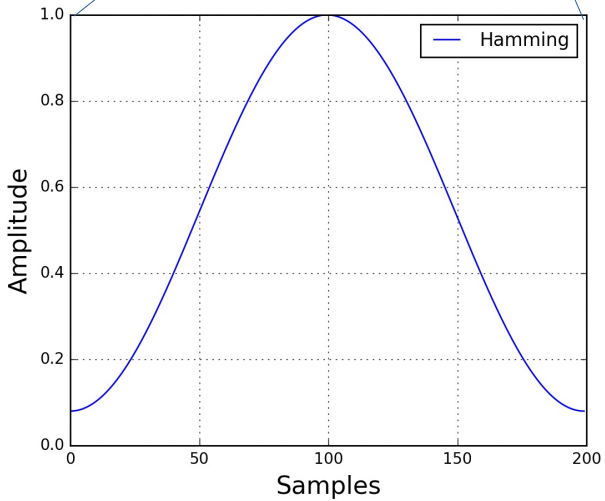
Acoustic and linguistic features

- 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: MFCC





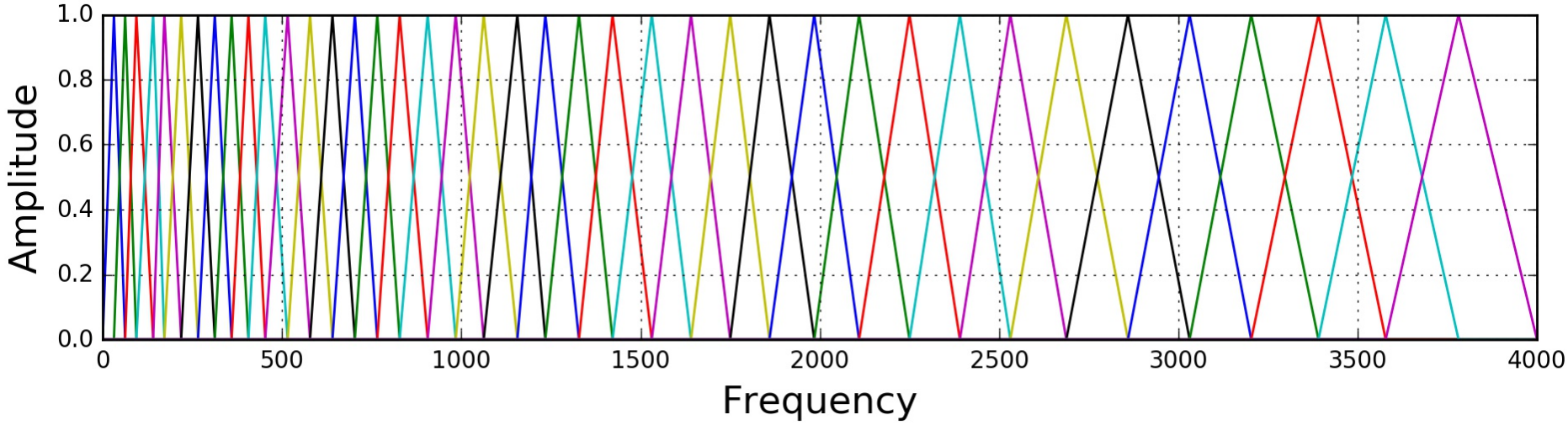
$$Powe\ Spectrum(P) = \frac{|FFT(x_i)|^2}{N}$$

$$mel(m) = 2595 \log_{10}(1 + \frac{f}{700})$$

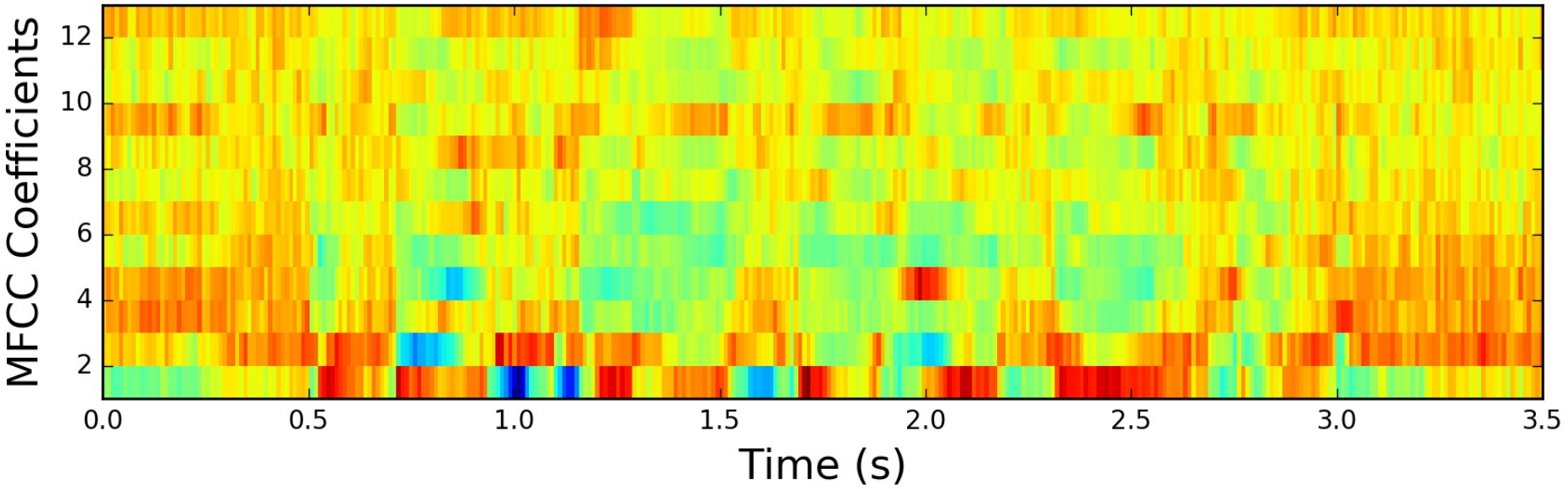
$$hertz(f) = 700(10^{m/2595} - 1)$$

https://haythamfayek.com/2016/04/21/speech-processing-for-machine-learning.html

Acoustic features: MFCC



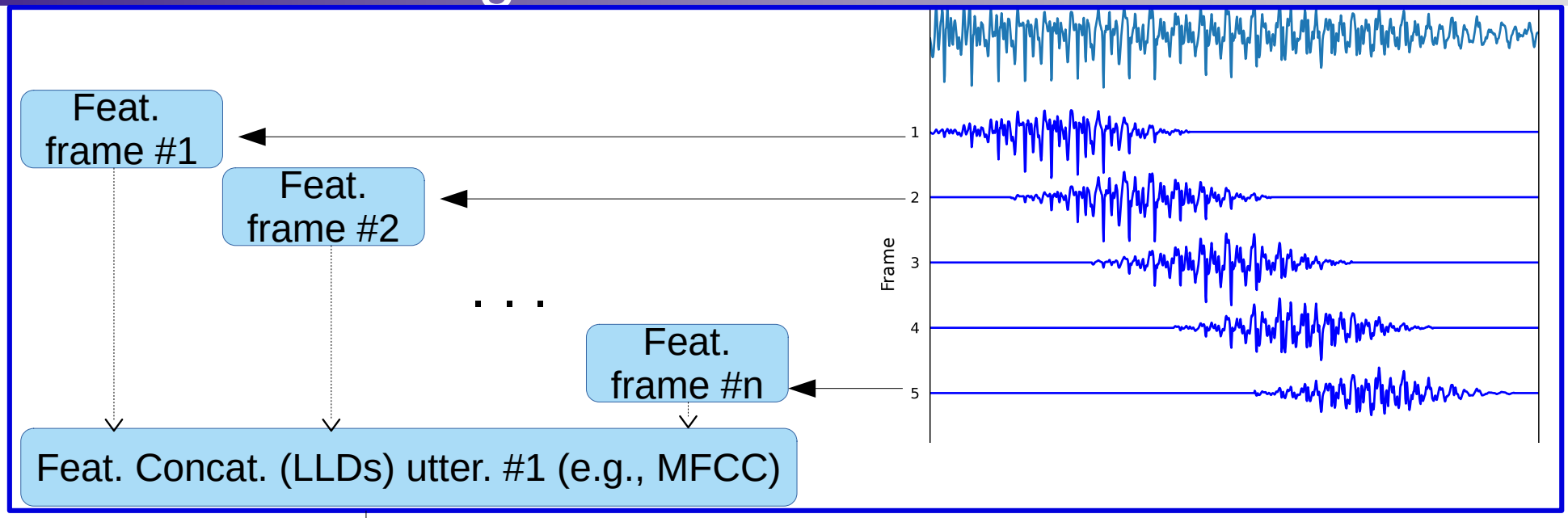
Filter bank



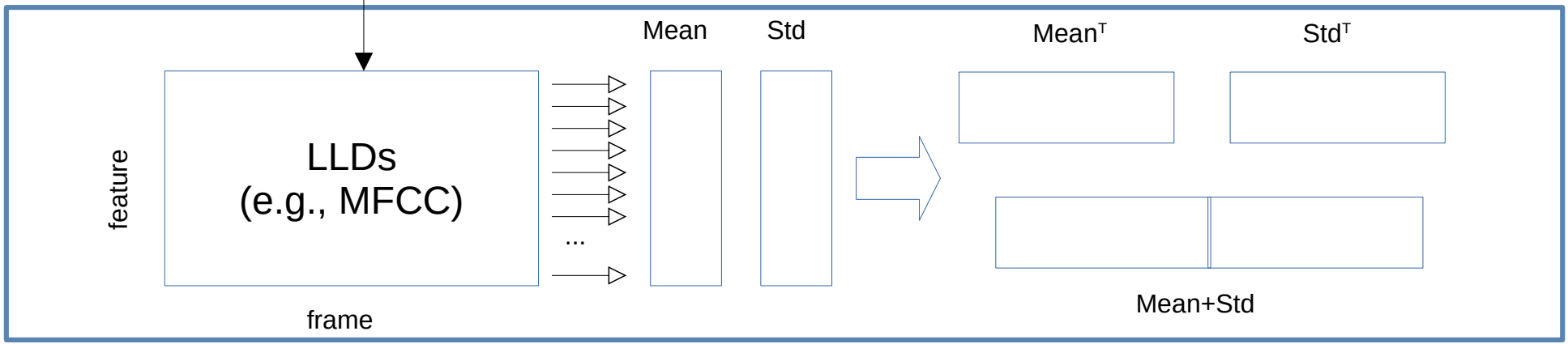
MFCC

Local features vs. global features

LLD



HSF



Results: LLD vs. HSF (IEMOCAP data)

LLD

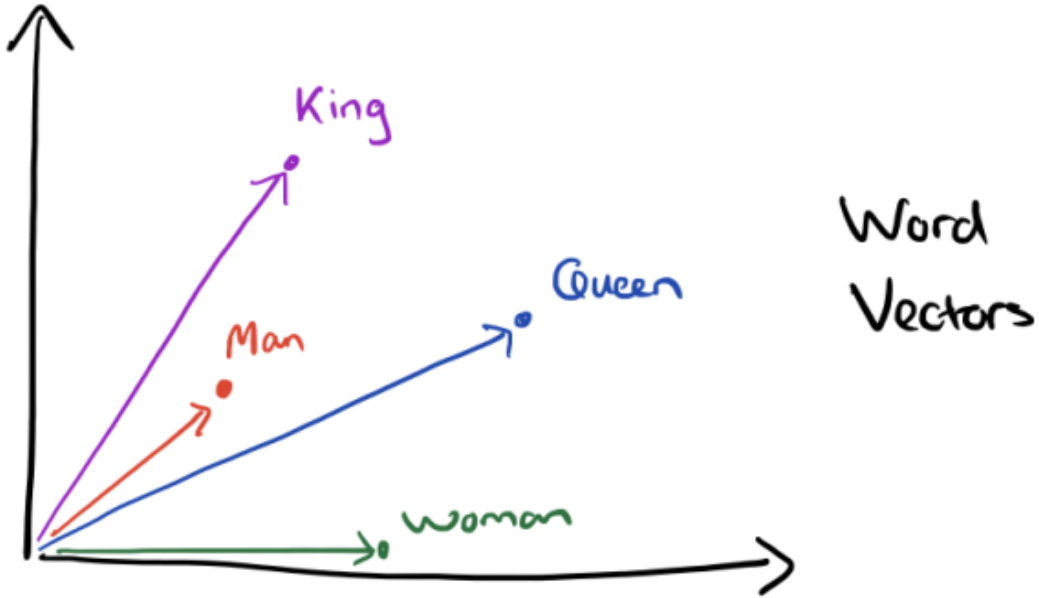
Feature	Dim	V	A	D	Mean
MFCC	(3414, 40)	0.148	0.488	0.419	0.352
Log mel	(3414, 128)	0.103	0.543	0.438	0.362
GeMAPS	(3409, 23)	0.164	0.527	0.454	0.382
pAA	(3412, 34)	0.130	0.513	0.419	0.354
pAA_D	(3412, 68)	0.145	0.526	0.439	0.370

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

Linguistic features

- Word embedding/word vectors: a numerical representation of word in a space of dimension (e.g. 300-dimensional vector)
- Some common used linguistic features:
  - Word embedding
  - word2vec
  - GloVe
  - FastText
  - BERT

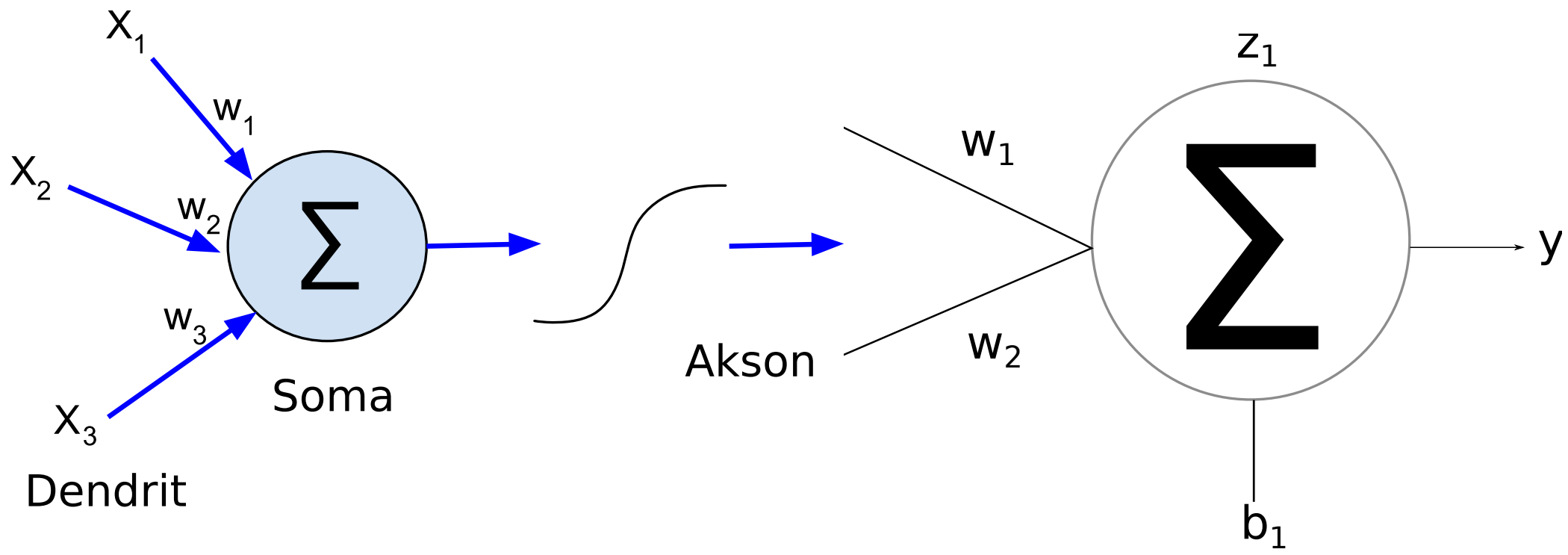


Outline

1. Introduction:
   Background, Applications, Companies
2. Emotion Recognition:
   Emotion models, Datasets, Evaluation metric, Tools
3. Acoustic and Linguistic Features
4. **Classifiers**
5. Fusion methods
6. Conclusions: Summary, Future research

DEMO

Classifier: Neural Network (NN)



$$z = b + \sum x_i w_i$$

$$y = f(z) = \frac{1}{1 + e^{-z}}$$

Classifier: CNN vs LSTM

Machine learning: small data, single/small layer, SVM, Random Forest, Naive Bayes, etc

Deep learning: bigger data (> 5000 samples), deeper layer (>2), CNN, LSTM, etc

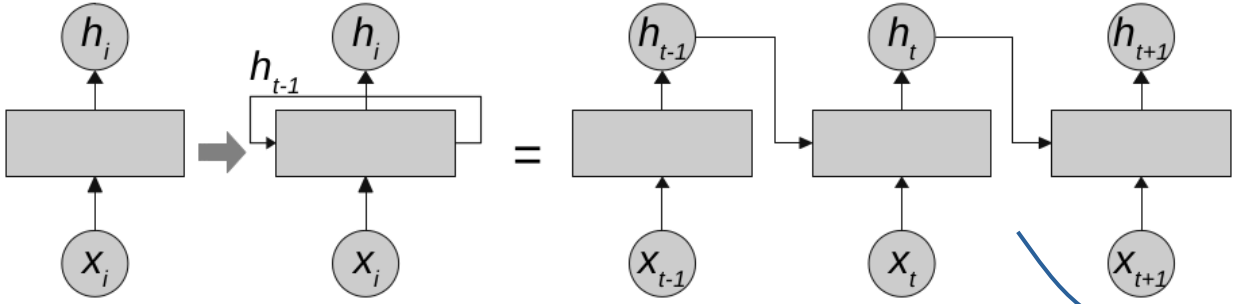
scikit-learn

Keras/Tensorflow

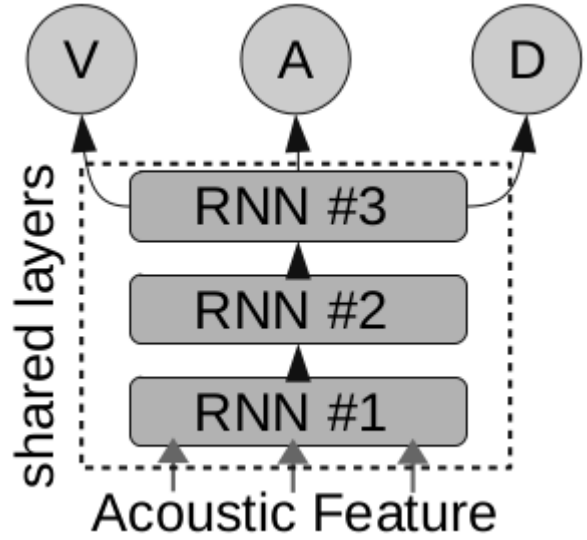
CNN vs. RNN: What are they and how do they differ?

	Convolutional neural network (CNN)	Recurrent neural network (RNN)
ARCHITECTURE	Feed-forward neural networks using filters and pooling	Recurring network that feeds the results back into the network
INPUT/OUTPUT	The size of the input and the resulting output are fixed (i.e., receives images of fixed size and outputs them to the appropriate category along with the confidence level of its prediction)	The size of the input and the resulting output may vary (i.e., receives different text and output translations—the resulting sentences can have more or fewer words)
IDEAL USAGE SCENARIO	Spatial data (such as images)	Temporal/sequential data (such as text or video)
USE CASES	Image recognition and classification, face detection, medical analysis, drug discovery and image analysis	Text translation, natural language processing, language translation, entity extraction, conversational intelligence, sentiment analysis, speech analysis

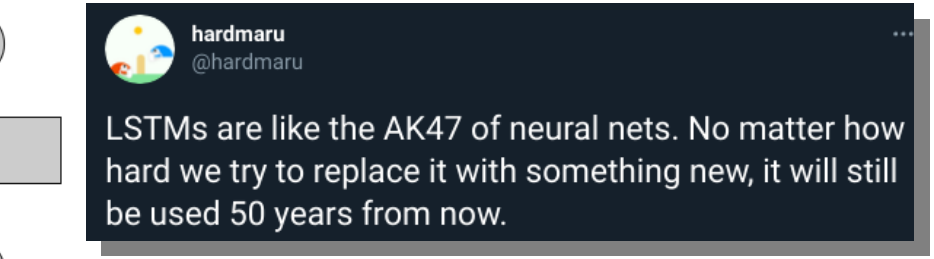
RNN LSTM

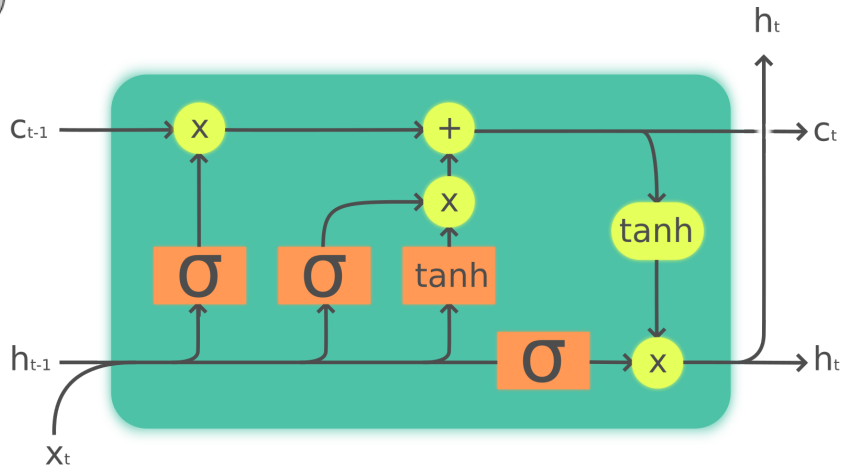


NN RNN Unrolled RNN



Simple RNN for dimensional SER

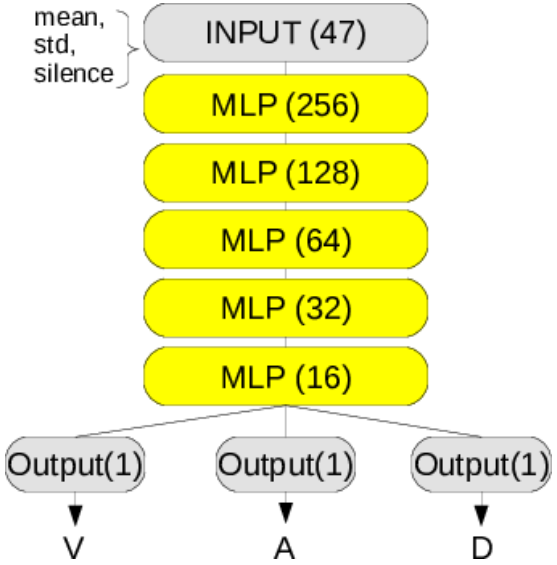




Legend: Layer Pointwize op Copy

More details: https://d2l.ai/

Deep Learning



LSTM Model	CNN Model
LSTM (256)	CNN (256,3)
LSTM (128)	CNN (128,12)
LSTM (64)	CNN (64,12)
LSTM (32)	CNN (32,12)
LSTM (16)	CNN (16,12)
Flatten()	Flatten()
#: 575,505	#:542,181

InputLayer (Acoustic features)	input:	(None, 3409, 23)
	output:	(None, 3409, **23**)
BatchNormalization	input:	(None, 3409, 23)
	output:	(None, 3409, **23**)
1 Dense / CuDNNLSTM 3	input:	(None, 3409, 23)
	output:	(None, 3409, **64**)
Dense / CuDNNLSTM	input:	(None, 3409, 64)
	output:	(None, 3409, **64**)
Dense / CuDNNLSTM	input:	(None, 3409, 64)
	output:	(None, 3409, **64**)
Flatten	input:	(None, 3409, 64)
	output:	(None, 218176)
Dropout (0.4)	input:	(None, 218176)
	output:	(None, 218176)

2 CuDNNGRU / CuDNNGRU / CuDNNGRU

batch size: 64
Trainable params: 743,921

Evaluation metrics: MSE, MAE, MAPE, **CCC**

Dense V	input:	(None, 218176)
	output:	(None, **1**)
Dense A	input:	(None, 218176)
	output:	(None, 1)
Dense D	input:	(None, 218176)
	output:	(None, **1**)

MLP vs. LSTM vs. GRU vs. CNN

Classifier	Song		Speech	
	Accuracy	UAR	Accuracy	UAR
FC/MLP	0.794	0.804	0.729	0.755
LSTM	0.820	0.813	0.785	0.781
GRU	0.812	0.844	0.785	0.764
Conv1D	0.743	0.806	0.687	0.690

[1] B. T. Atmaja and M. Akagi, “On The Differences Between Song and Speech Emotion Recognition: Effect of Feature Sets, Feature Types, and Classifiers,” in 2020 IEEE REGION 10 CONFERENCE (TENCON), 2020, pp. 968–972.

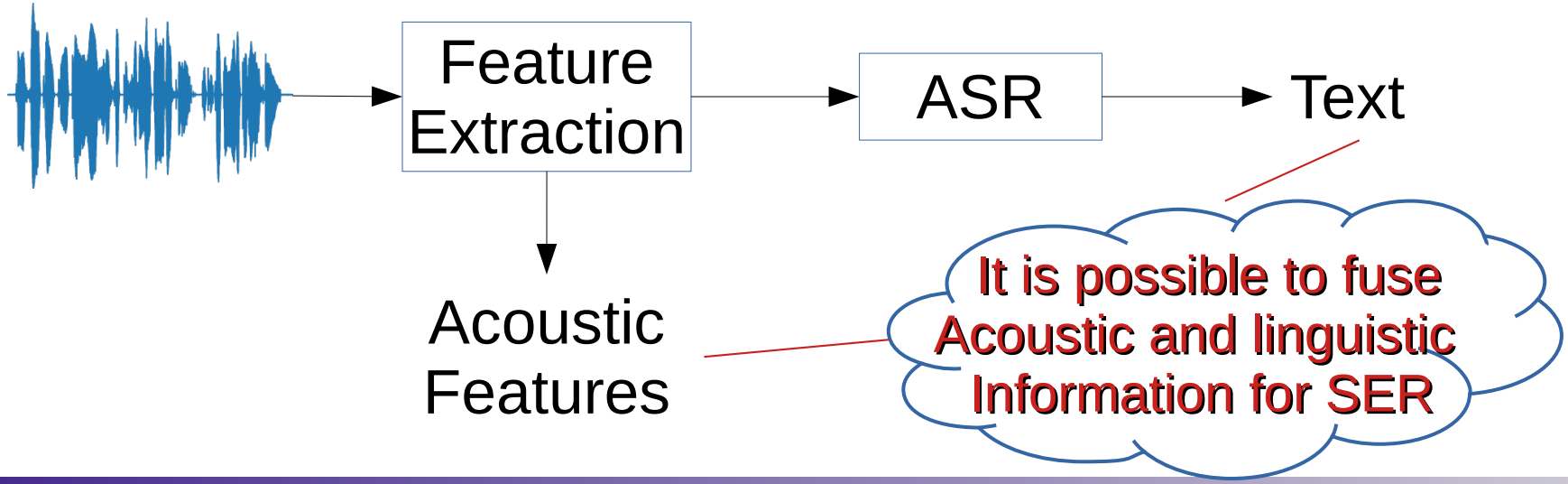
Outline

1. Introduction:
   Background, Applications, Companies
2. Emotion Recognition:
   Emotion models, Datasets, Evaluation metric, Tools
3. Acoustic and Linguistic Features
4. Classifiers
5. **Fusion methods**
6. Conclusions: Summary, Future research

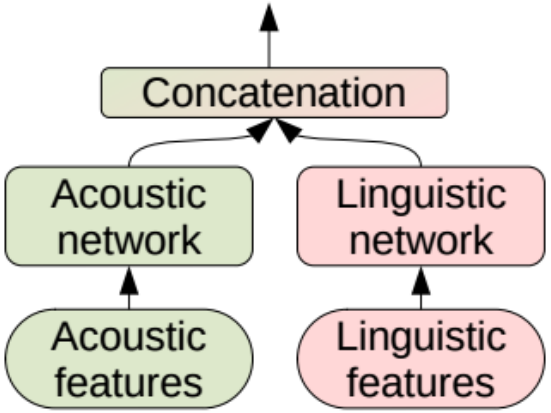
DEMO

Why fusing acoustic and linguistic?

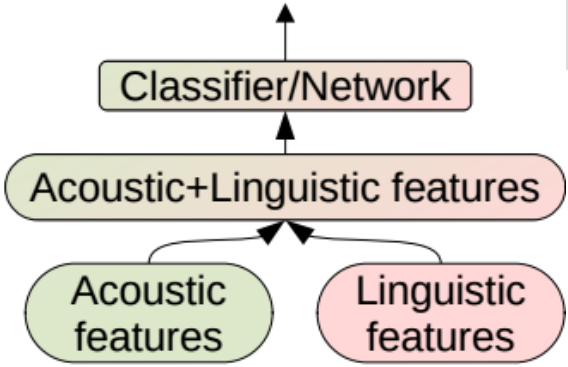
- Speech can be transcribed into text using **A**utomatic **S**peech **R**ecognition (ASR)
- Linguistic information can be extracted from transcription
- Human communicate emotion through speech and language (Kotz et al., 2011)
- More data tends to be more effective (Halevy et al., 2009)



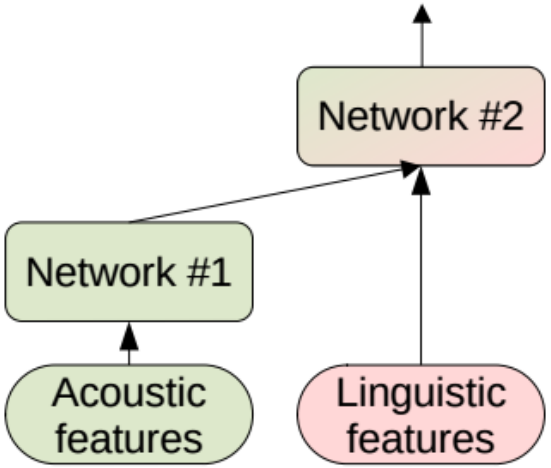
Fusion methods



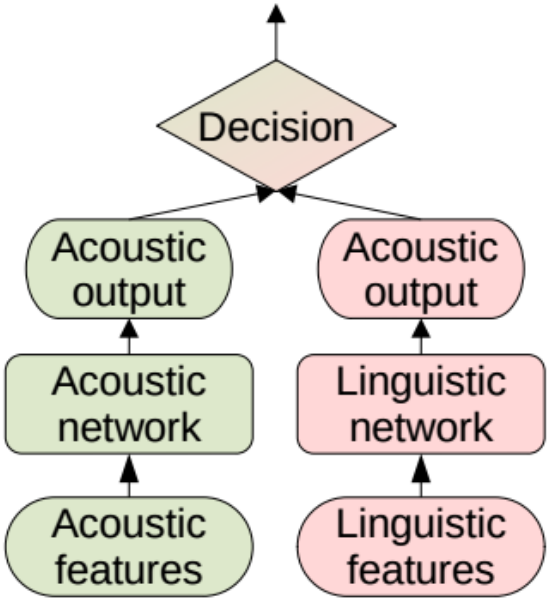
(a) Network/model concatenation



(b) Feature concatenation

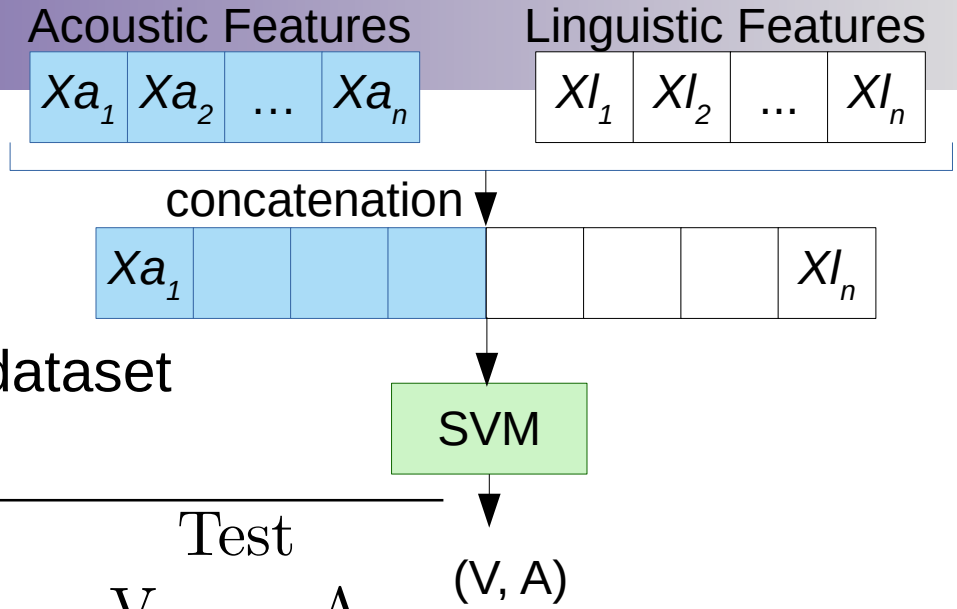


(c) Hierarchical fusion



(d) Decision-level fusion

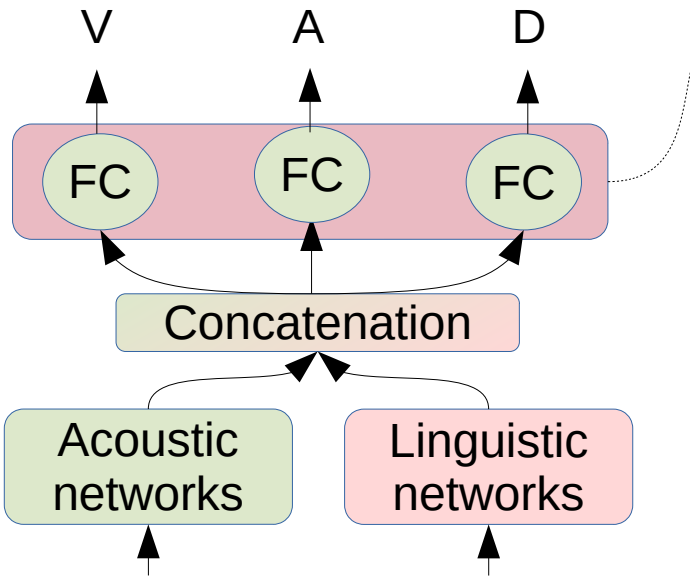
Feature concatenation



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

Network concatenation with multitask learning (MTL)



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:

$$\mathrm{CCCL}_{tot} = \alpha\, CCCL_V + \beta\, CCCL_A + (1 - \alpha - \beta)\, CCCL_D$$

Total loss function with 3 parameters:

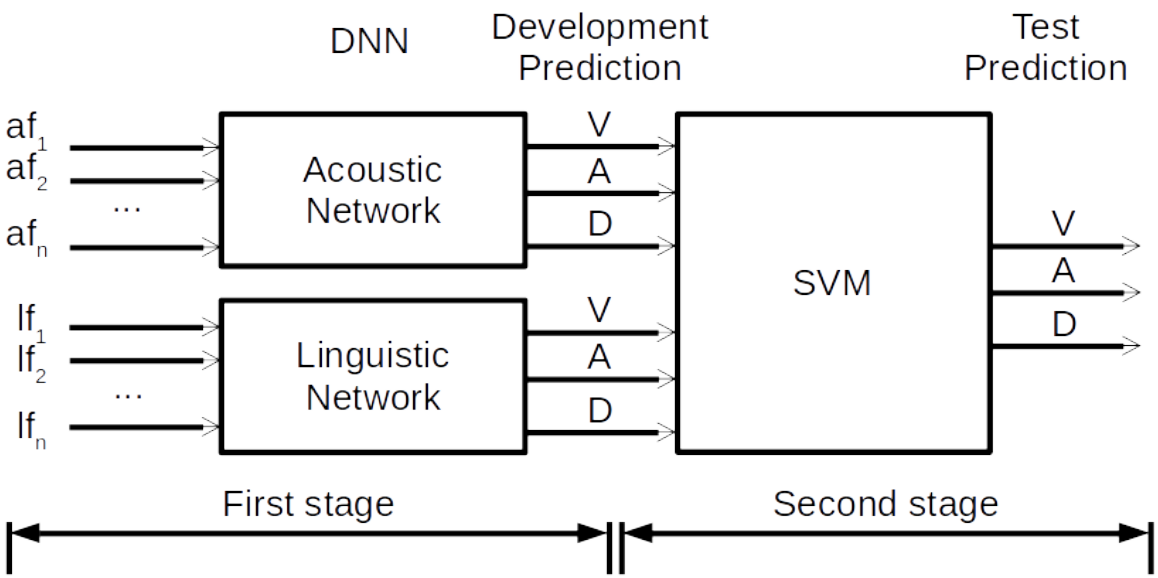
$$CCCL_{tot} = \alpha\, CCCL_V + \beta\, CCCL_A + \gamma\, CCCL_D$$

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

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')

Outline

1. Introduction:
   Background, Applications, Companies
2. Emotion Recognition:
   Emotion models, Datasets, Evaluation metric, Tools
3. Acoustic and Linguistic Features
4. Classifiers
5. Fusion methods
6. **Conclusions: Summary, Future research**

DEMO

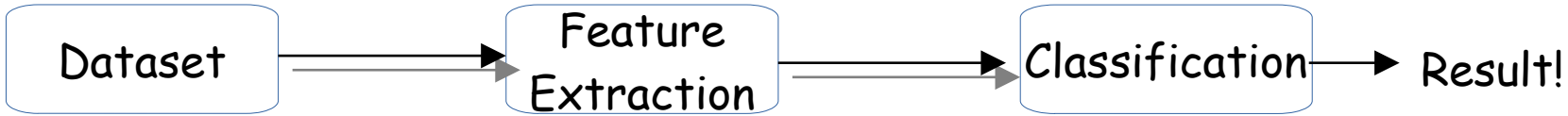
Summary

1) Emotion can be recognized from speech
2) Speech emotion recognition (SER) can be applied for many applications like call center and voice assistant
3) The research of of SER is now being implemented for commercial; however *more research is needed particularly for Indonesian language* (aside from the universality of SER)
4) SER can be combined with other speech tasks: *ASR, speaker recognition, age recognition, gender recognition.*

Future research direction

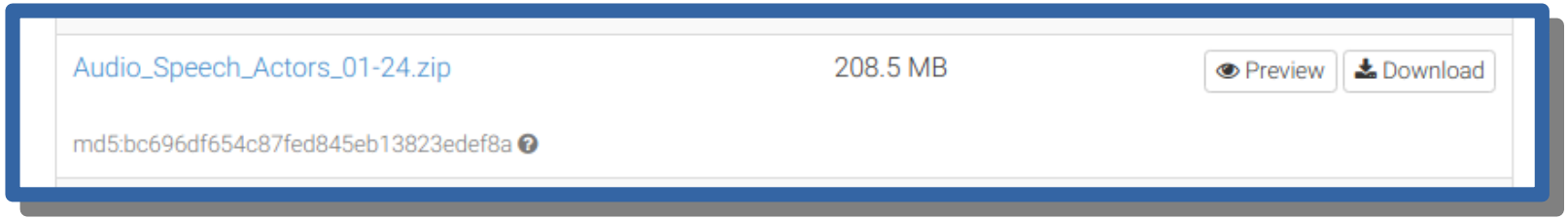
- Accelerating high-level feature extraction for speech emotion recognition
- A method to calculate silent pause features that discriminate significantly among removing, keeping, and utilizing silence.
- The contribution of fusing LLD and HSF compared to individual region of analysis, the thresh-hold, and its complexity
- Bimodal acoustic-linguistic emotion recognition by two spaces resultant
- Fine-tuned BERT on acoustic-linguistic dimensional SER
- Fully lexical controlled vs. lexical uncontrolled emotion recognition
- Bottleneck between acoustic and linguistic processing
- Concurrent speech and emotion *(and others)* recognition
- Model generalization

- **DEMO:**
  - http://bagustris.blogspot.com/2019/06/implementasi-pengenalan-emosi-dari.html
  - paper: https://ieeexplore.ieee.org/document/9293852
  - github: https://github.com/bagustris/ravdess_song_speech
  - Requirement:
    - Keras == 2.3.1
    - Tensorflow == 1.15.5
    - Librosa == 0.8.0
  - Flow:



Demo #1: Download the dataset

https://zenodo.org/record/1188976#.YDylMllfih5



Unzip the zip file

Demo #2: Run feature extraction code

```
>>> run feature_extraction.py
```

Code link: https://gist.github.com/bagustris/3c73a40c8e736908db775a532d195205

Demo #3: Run classification code

```
>>> run classification.py
```

Code link: https://gist.github.com/bagustris/8c8597ae6cfc90007c2cc4379979c0c8

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

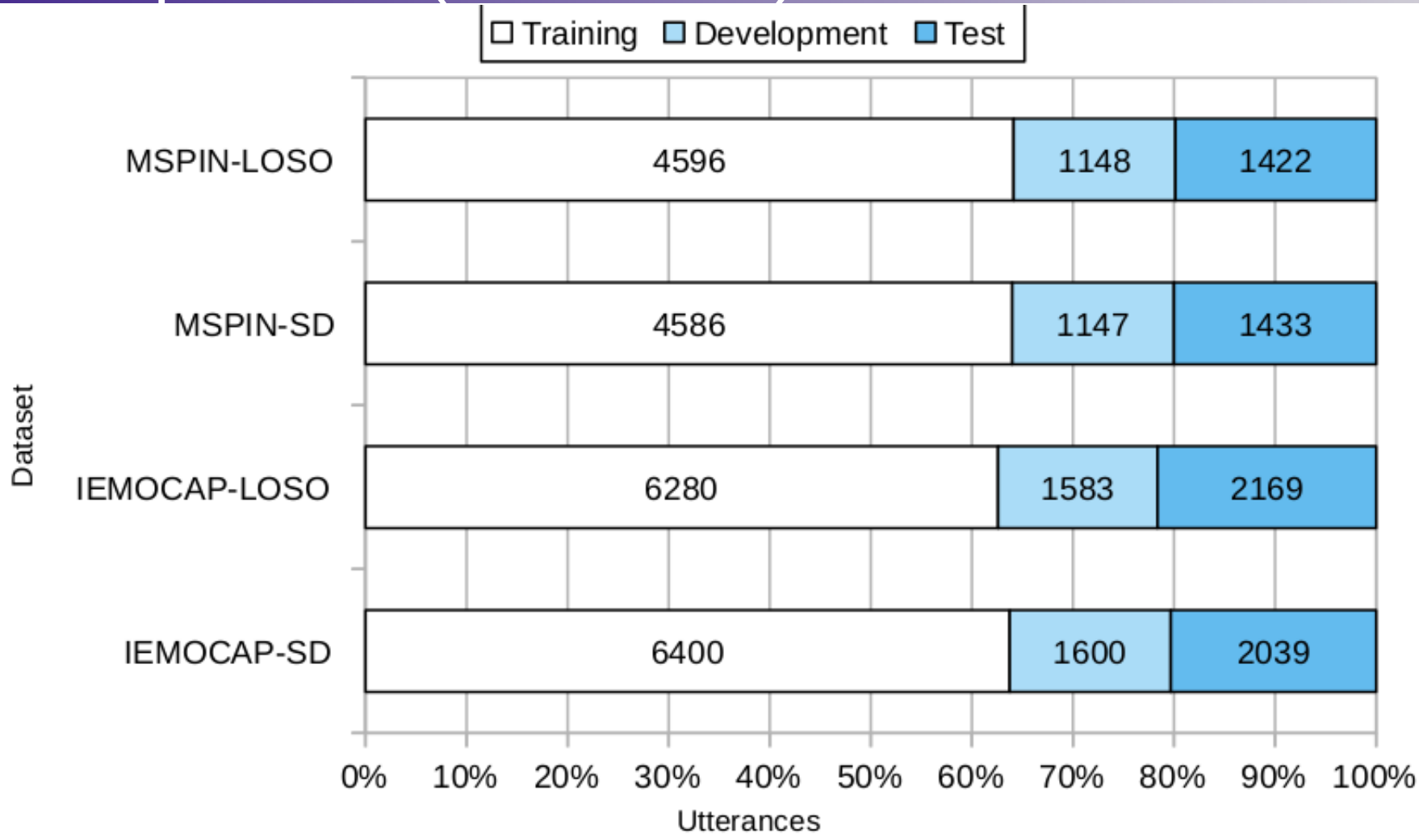
List of abbreviation

- 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)

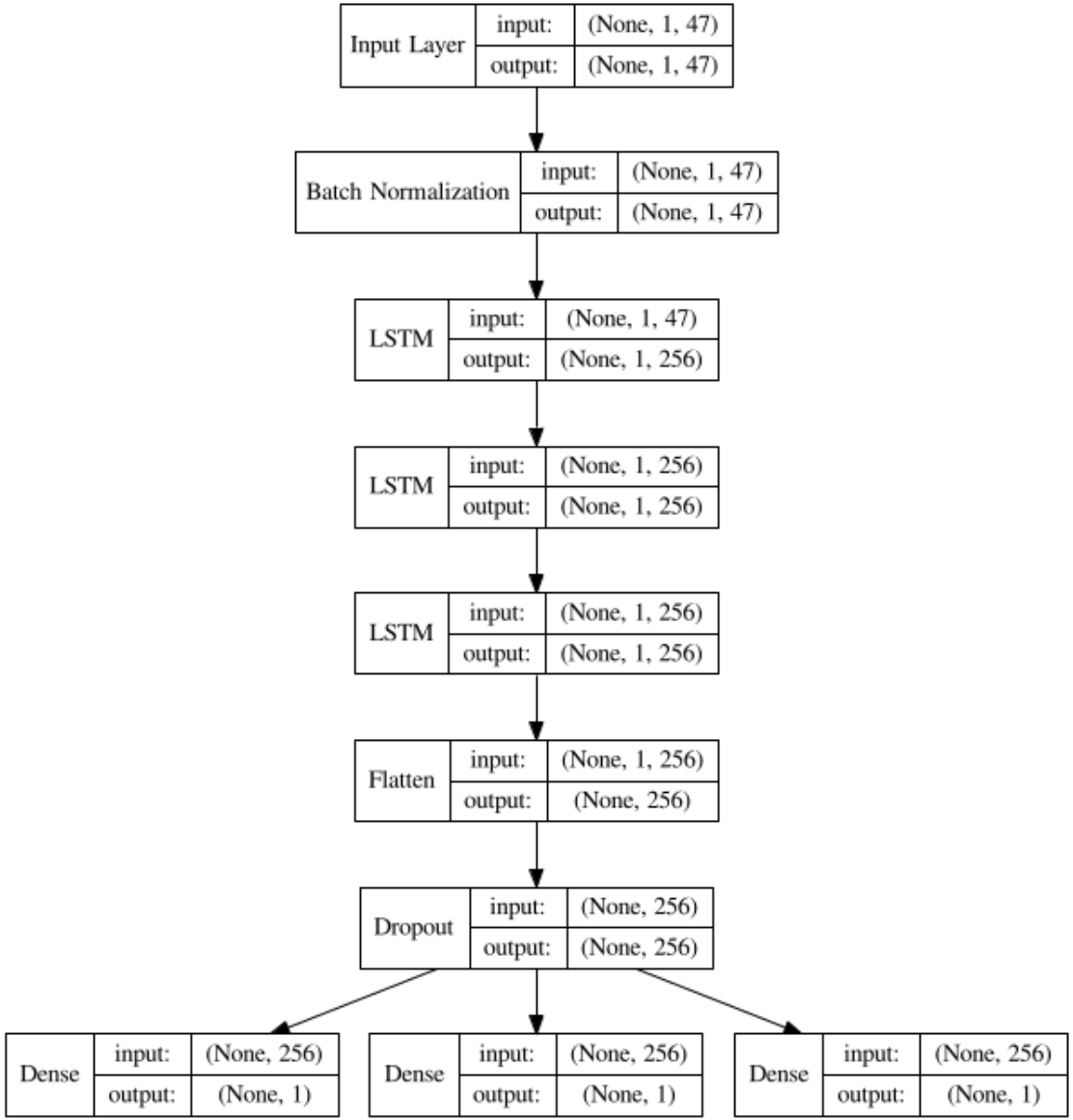
- SD: speaker dependent
- LOSO: leave one session out, SI: speaker independent
- WER: word error rate
- pAA: pyAudioAnalysis
- pAA_D: pyAudioanalysis with their deltas
- MTL: multi-task learning
- af: acoustic feature
- lf: linguistic feature
- WE: word embedding
- Std: standard deviation

Dataset partition (late fusion)



DNN model (Acoustic)

* used in two-stage processing



DNN model (Linguistic)

* used in two-stage processing