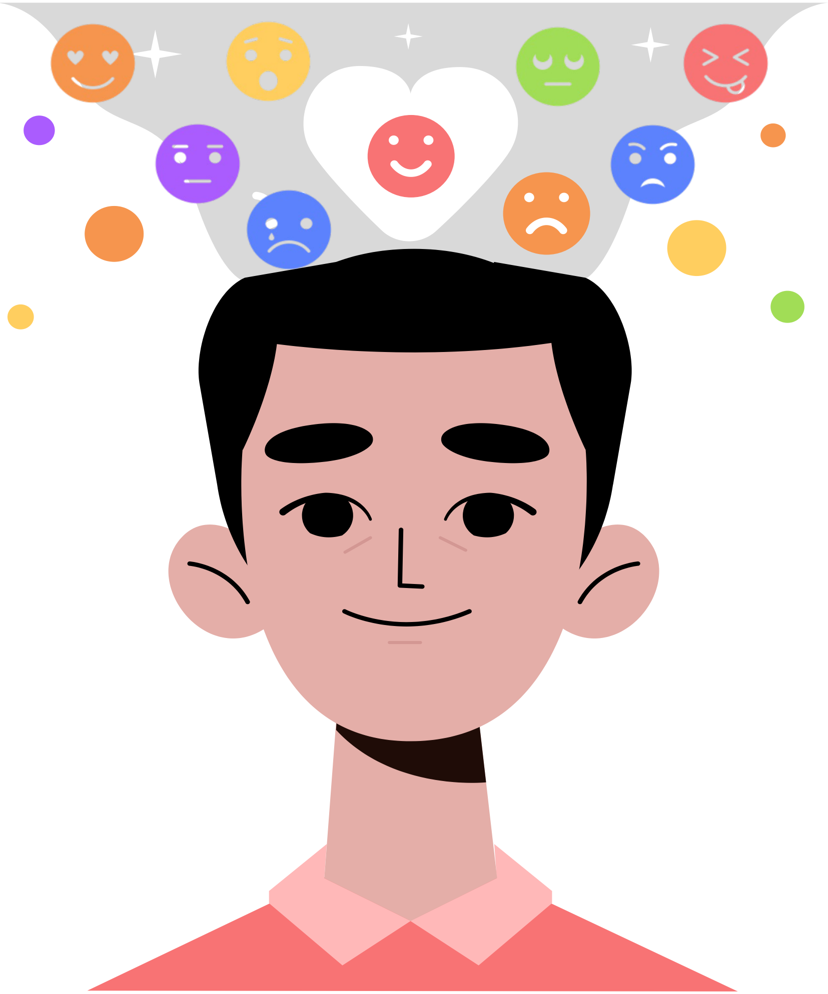
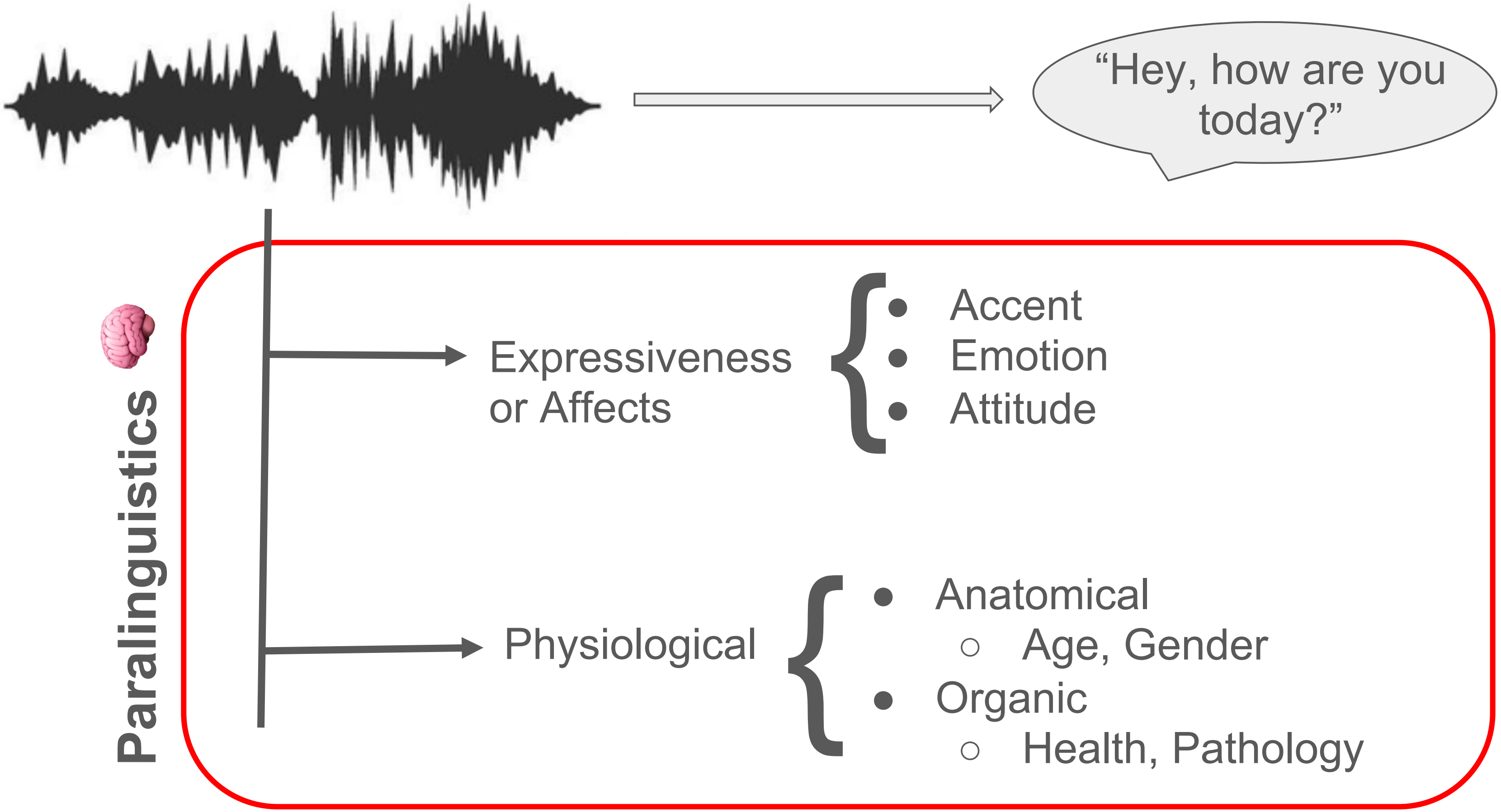
Computational Paralinguistics

Tilak Purohit

tilak.purohit@{idiap,epfl}.ch




"Hey, how are you today?"
Paralinguistics
Expressiveness or Affects
• Accent
• Emotion
• Attitude
Physiological
• Anatomical
○ Age, Gender
• Organic
○ Health, Pathology

Computational + Paralinguistics

Roughly means something is done by a computer and not by a human being

‘Paralinguistics’ means ‘alongside linguistics’ (from the Greek preposition παρα)

Term coined in 1950’s

Safe to claim that 30 years ago, neither the term ‘computational paralinguistics’ nor the field it denotes existed !

Paralinguistics: Going beyond linguistics

Paralinguistics deals with *traits (long-term events)* and *states(short-term events)*

- *Long-term traits:*
  - *Biological (age, gender)*
  - *Cultural (ethnicity, race [dialect])*
  - *Personality ('big-five' personality traits)*
- *Medium-term b/w traits and states:*
  - *sleepiness, intoxication (e.g., alcoholisation), health state (e.g. depression), mood.*
- *Short-term states:*
  - *emotion-related states or affects, such as stress, happy, excited, frustration, pain*

!! concerned with how you say something rather than what you say !!

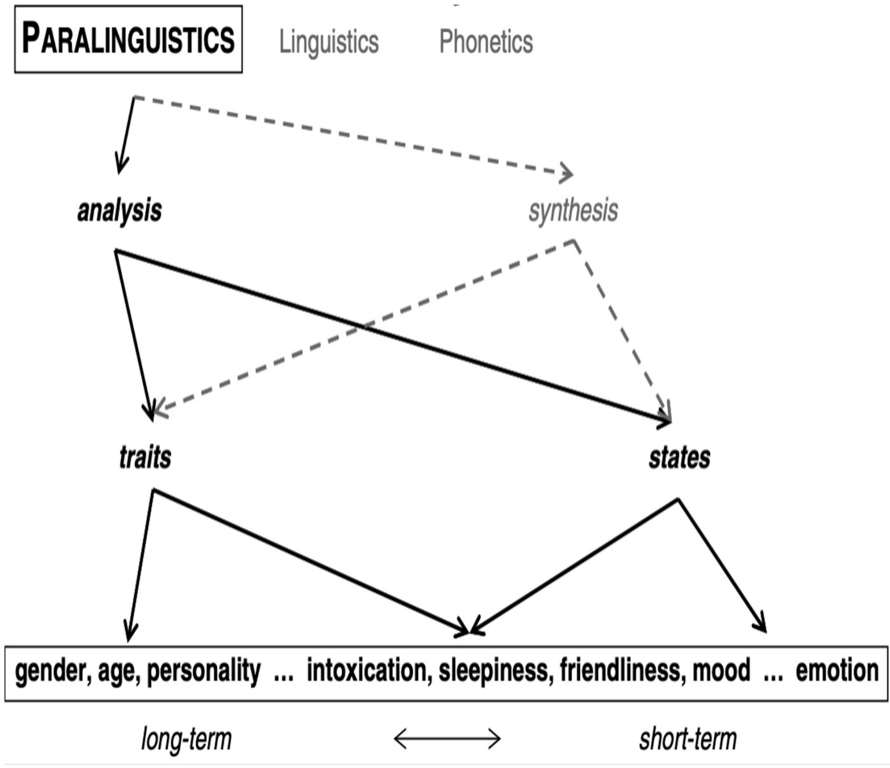


Image credit: computational paralinguistics book

Application areas

Understanding the user's states and traits can enhance the interactions between humans and human-computer interaction (HCI) interfaces.

- **Call Centers**
  - Quality of service
  - Coping with frustrated users
- **Education**
  - Detect attention & frustration
- **Observational practices**
  - Diagnosis and coaching
- **Healthcare**
  - Empathy detection in medical training
  - Assessment of therapist

Speech Analysis: 3 main speech organ groups

Lungs → Respiration, **Larynx** → Phonation and **Vocal Tract** → Articulation

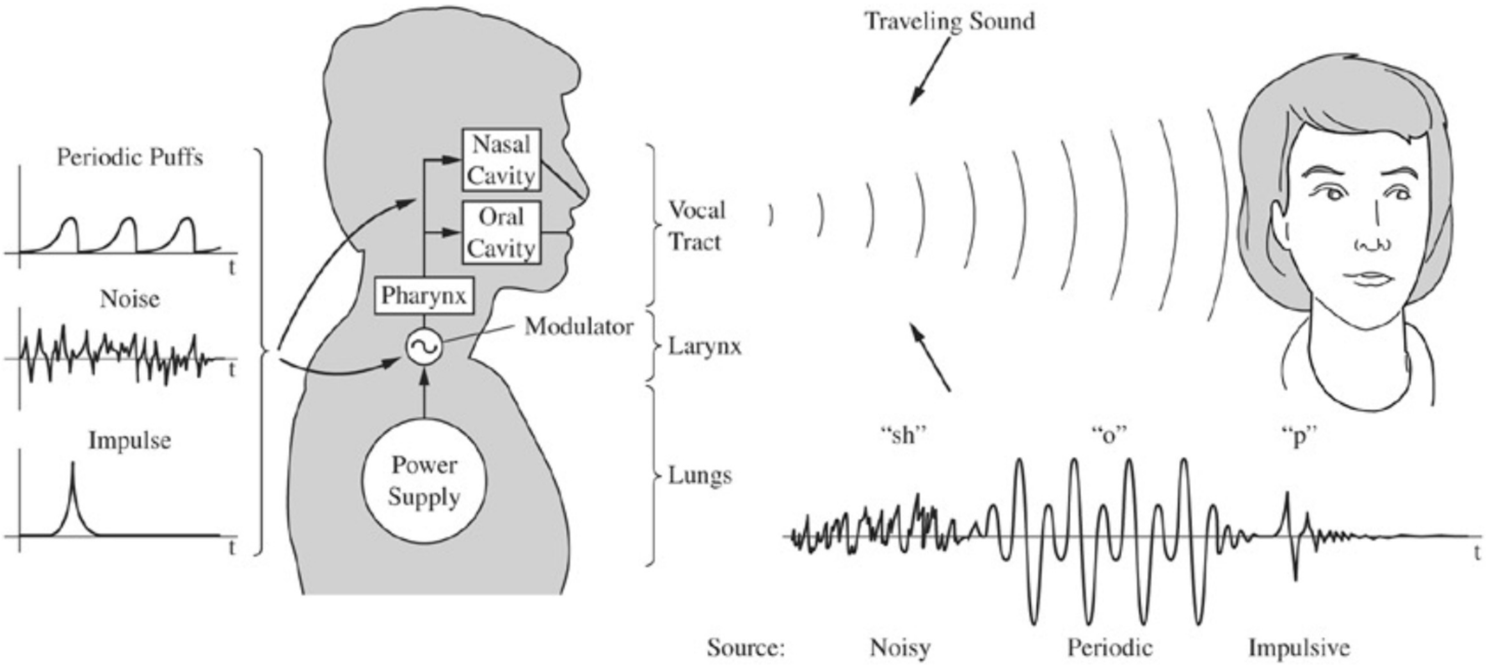
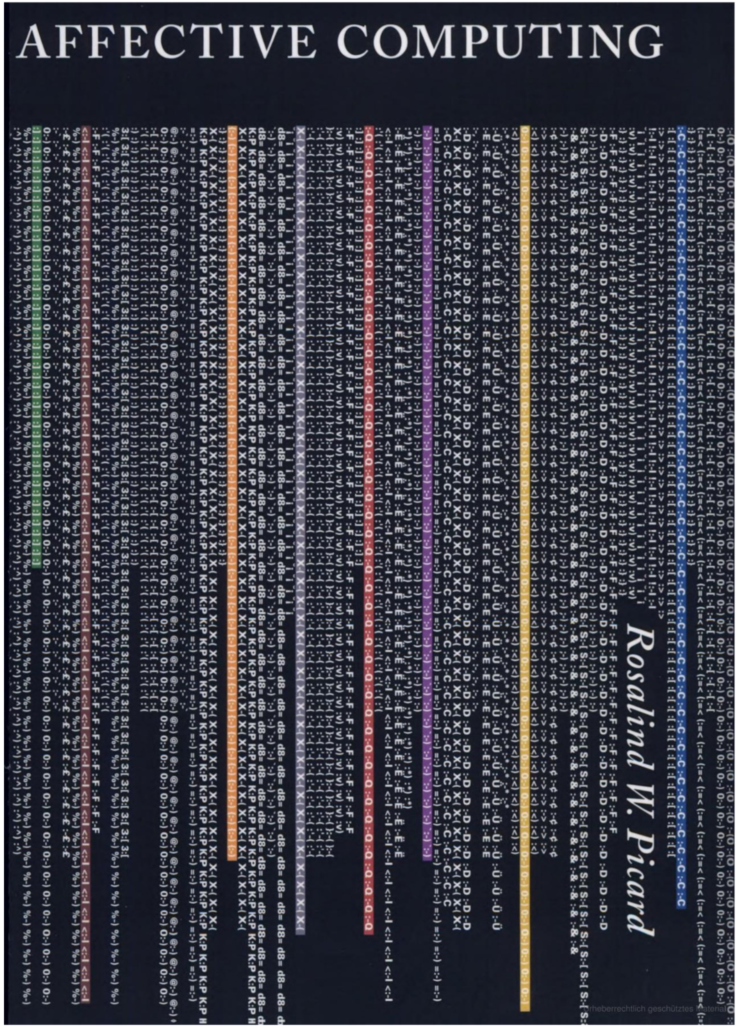


Image credit: USC SAIL

Speech Emotion Recognition

GOAL→ Design SER algorithms that replicate the **human perception process to infer emotions**.



Speech Emotion Recognition

Categorical attributes : **(Classification task)**

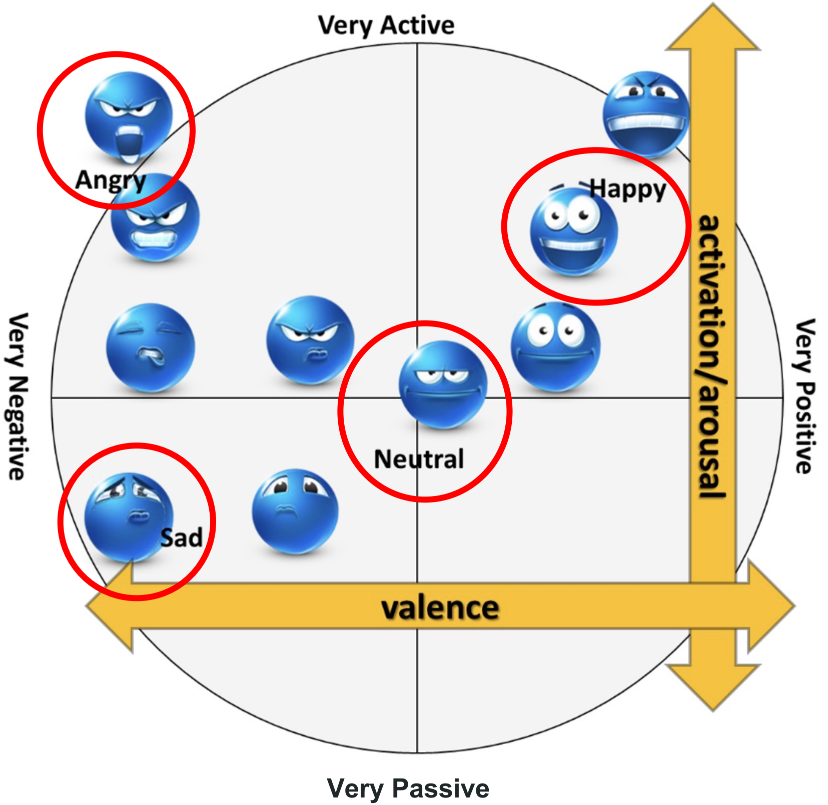
4 basic emotion categories namely:

Happy(😀) Angry(😡) Neutral(😐) Sad(😓)

Dimensional attributes: **(Regression task)**

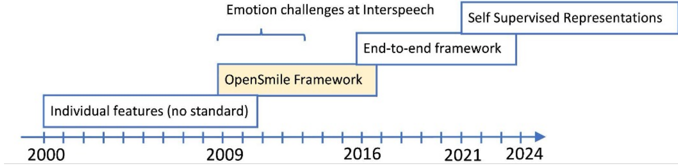
Valence (negative vs. positive)

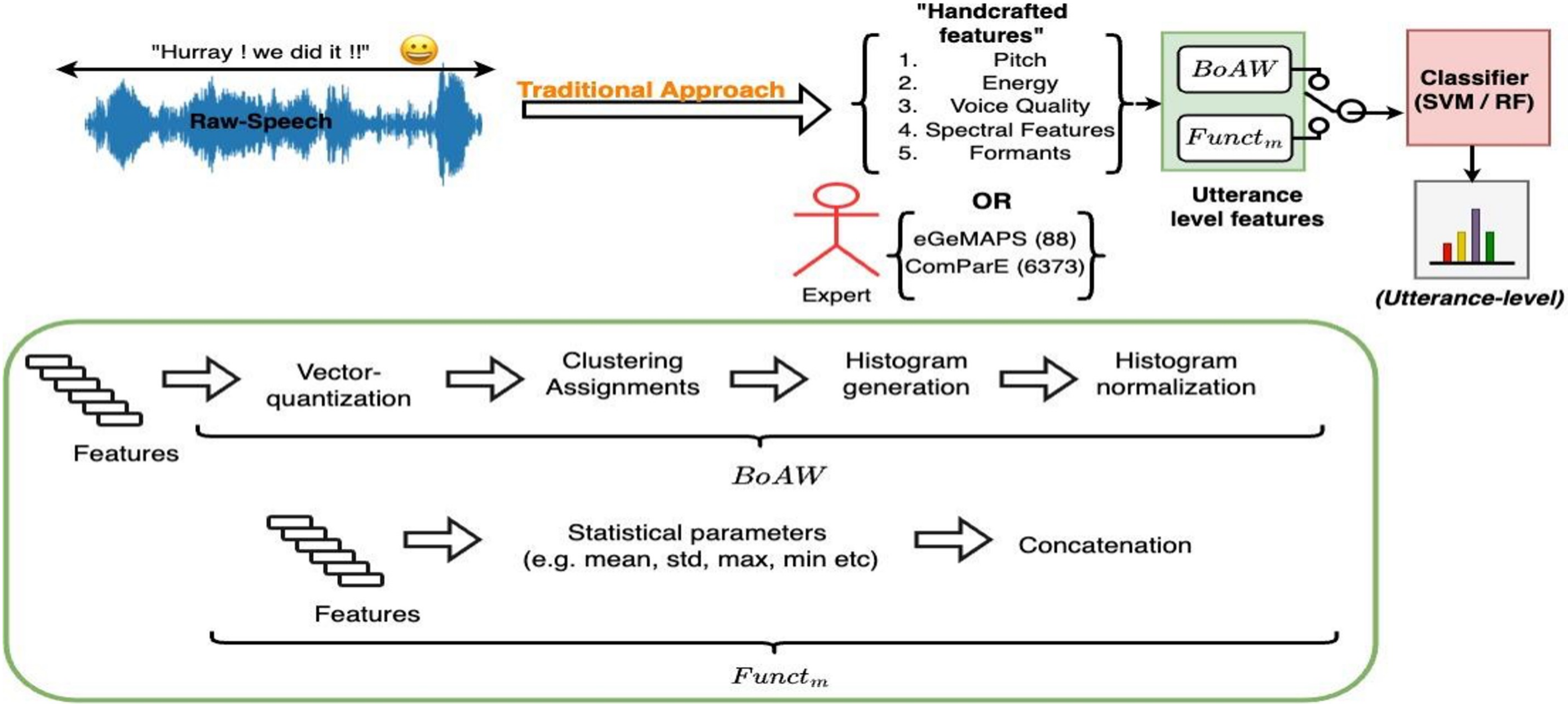
Arousal (calm vs. active)



Happy Angry Sad

RECAP: Using handcrafted features





Study design (SER)

Categorical attributes :

Corpus **IEMOCAP**, 4 basic emotion categories namely:

Happy(😀) Angry(😡) Neutral(😐) Sad(😓)

Happy Angry Sad

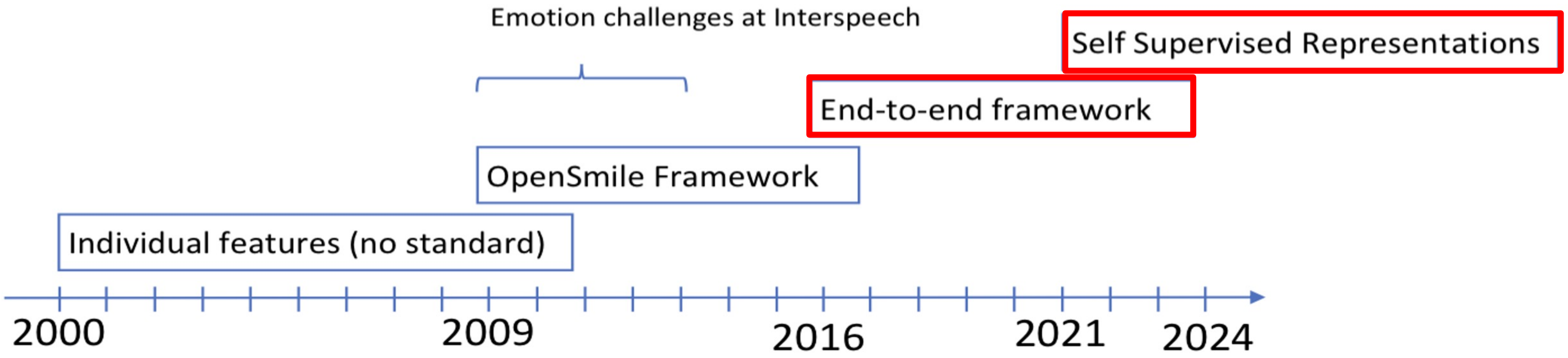
Protocol:

Conducted speaker-independent experiments by following **Leave-One-Speaker-Out (LOSpO)** methodology for training.

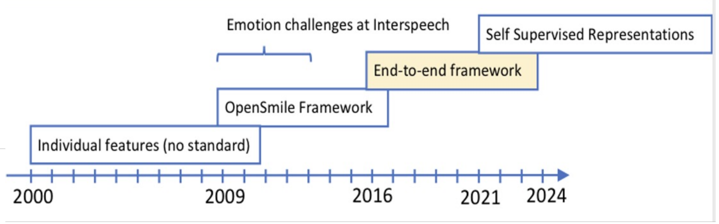
Evaluation Matrices:

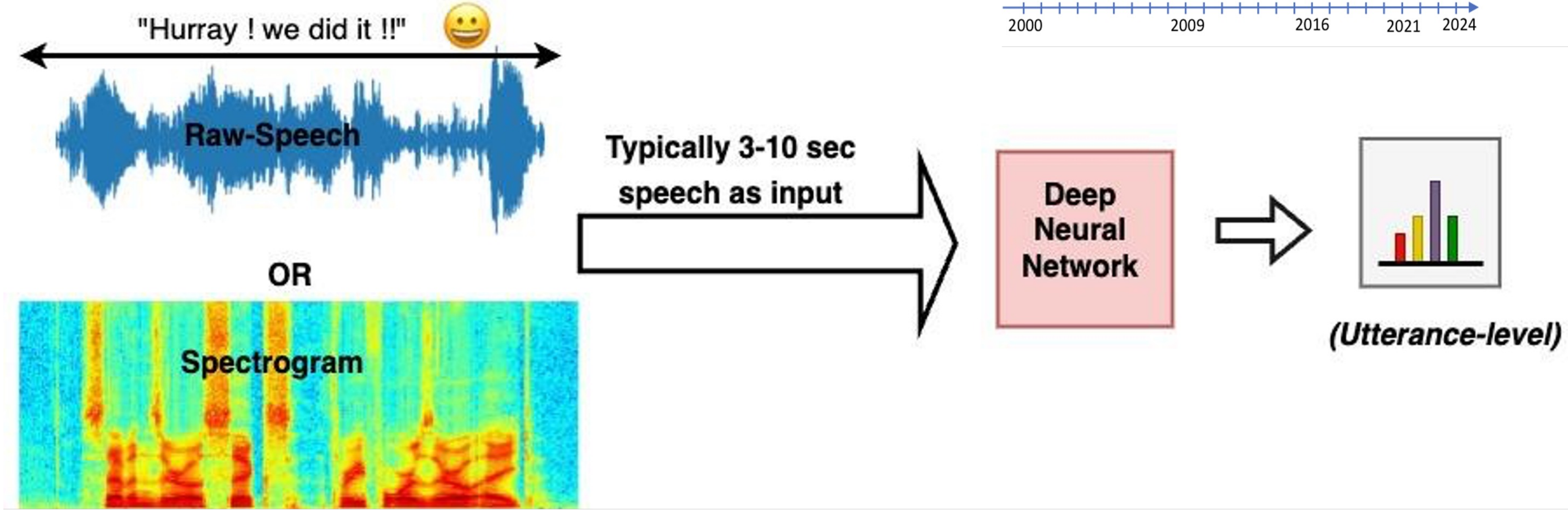
Performance measurement : **Unweighted Average Recall (UAR)**.

Moving on to DL based methods



Goal: Learn features from data





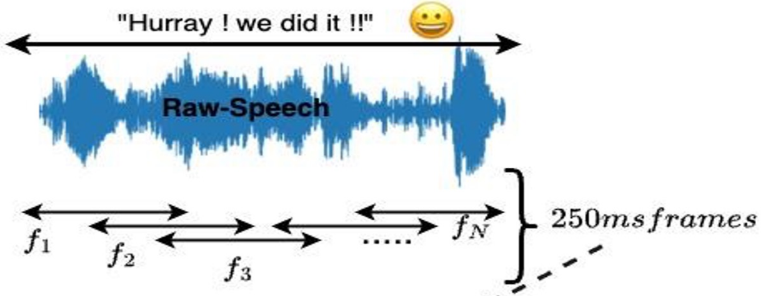
M. Neumann and T. Vu, "Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech," in Proc. of Interspeech, 2017.

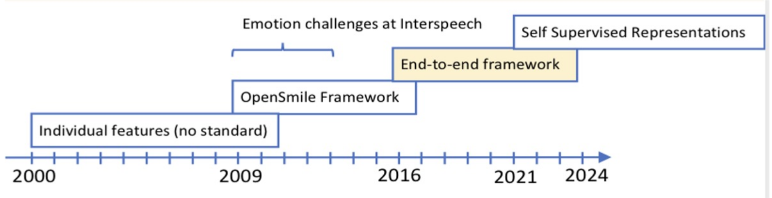
J.L. Li et al., "A waveform-feature dual branch acoustic embedding network for emotion recognition," Frontiers in Computer Science, 2020.

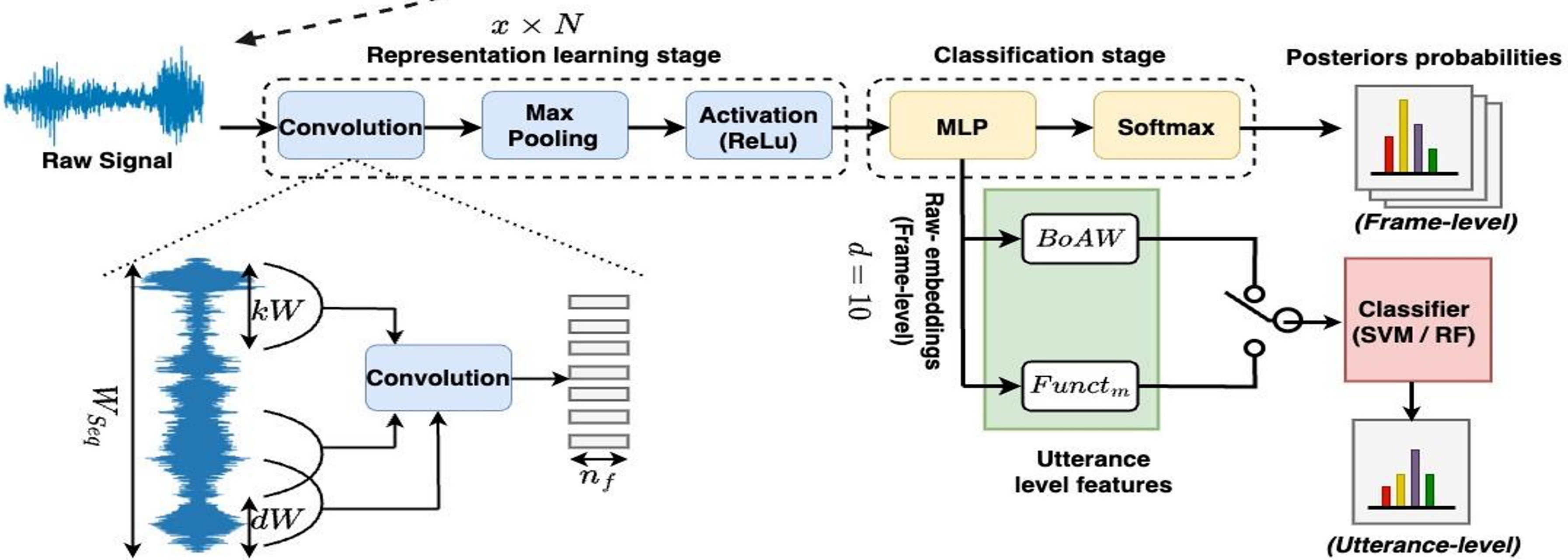
P. Kumawat and A. Routray, "Applying TDNN Architectures for Analyzing Duration Dependencies on Speech Emotion Recognition," in Proc. of Interspeech, 2021.

What is the smallest acoustic unit/segment in speech that contains emotion discriminative information?

Can emotion discriminative information be effectively learned/modeled from short segment of speech (of duration around 250 ms)?

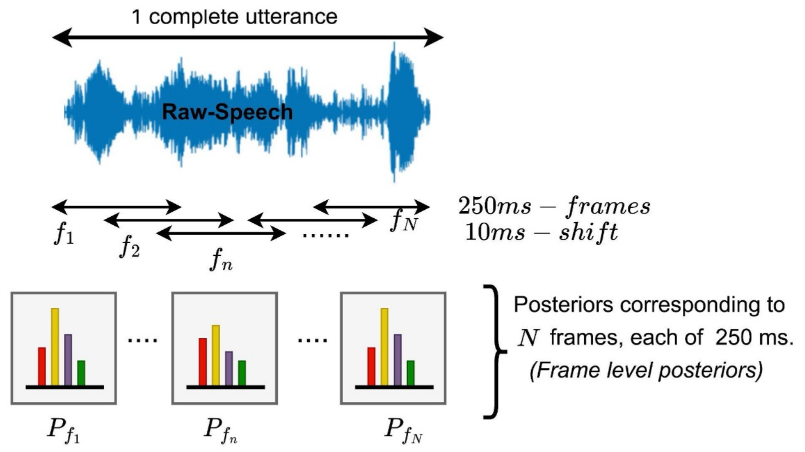






T. Purohit et al, “Towards Learning Emotion Information from Short Segments of Speech”. In Proc. of ICASSP, 2023, Rhodes island, Greece.

Performance



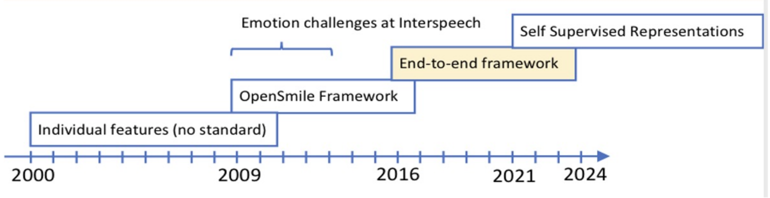


Table 2. Performance of previously reported systems measured in terms of UAR and Weighted Accuracy (WA); Utterance level (UL)

Method (Feature) – Duration	Metric	%
Att. CNN (logMel) – 7.5s [9]	WA	56.1
DBN-ivector (MFCC) – UL [13]	WA	57.2
CNN+LSTM (raw aud.) – 6s [14]	UAR	52.8
TDNN (MFCC) – 4s [15]	UAR	58.6

Systems	Classifier	UAR
Baseline systems - Speaker Independent		
$\text{ComParE}_{LLD \times F}$	SVM	56.57
$\text{BoAW}(\text{ComParE}_{LLD})$	SVM	56.63
Proposed systems - Speaker Independent		
Raw-CNN	Softmax	57.4
$\text{Funct}_{m,sd,sk,k}$(S-EMBEDDINGS)	SVM	56.7

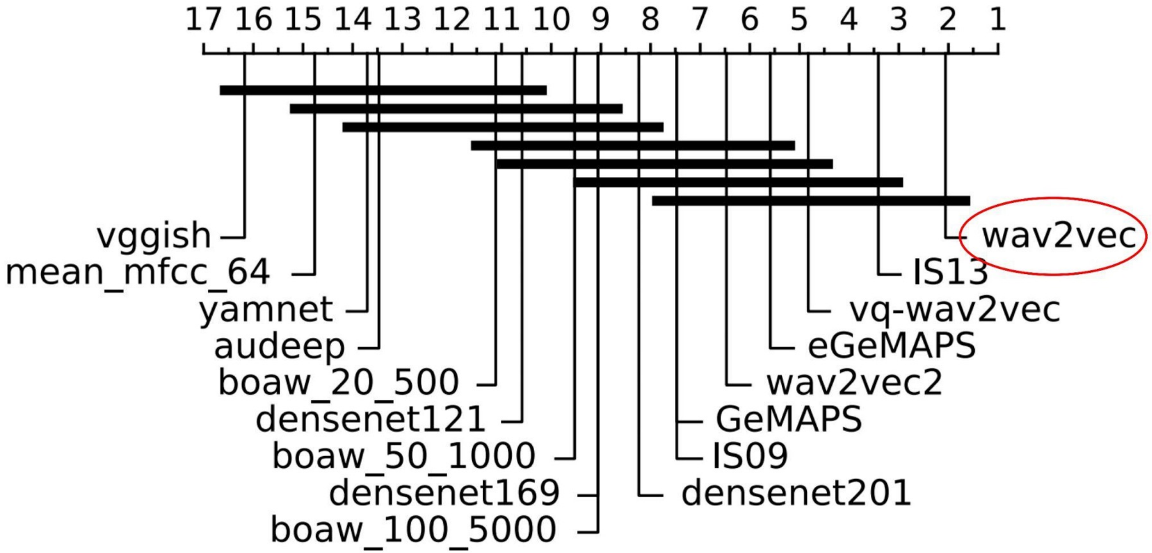
Takeaway:
End-to-End modelling system can capture emotion discrimination information from short speech-segments

Different Acoustic feature & Neural Rep. Evaluation

17 different SER corpus and 17 different representations were evaluated by Keesing et al.

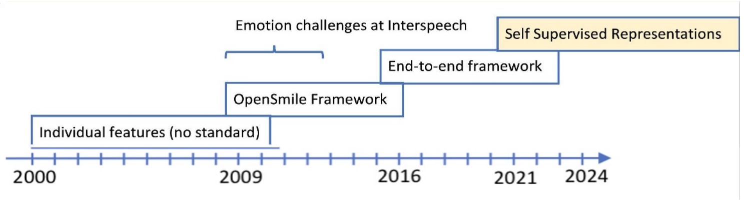
Observation:

Self-superpervised representation achieved the best average performance.



A. Keesing et al, "Acoustic Features and Neural Representations for Categorical Emotion Recognition from Speech". In Proc. of INTERSPEECH 2021, Brno, Czechia.

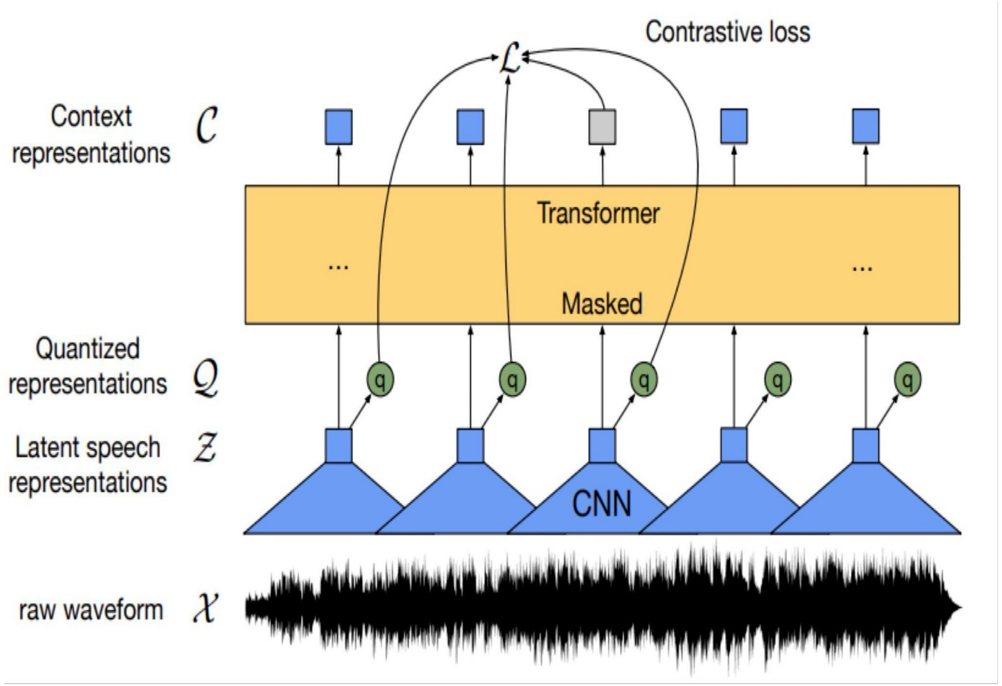
Self Supervised Representations (SSLs)



- Trained using 1000 hrs of unlabelled speech data in a self supervised fashion.
- Model **learns some intrinsic properties** of the data.
- Four major speech SSL models or Speech Foundation Models (SFMs):

➔Wav2vec2.0 ➔ HuBERT

➔ Hubert ➔ WavLM



A Baevski et al, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations". In Proc. of Neurips 2020, (Virtual).

A bit of detail on Speech Foundation Models (SFMs)

Wav2vec2.0

WavLM

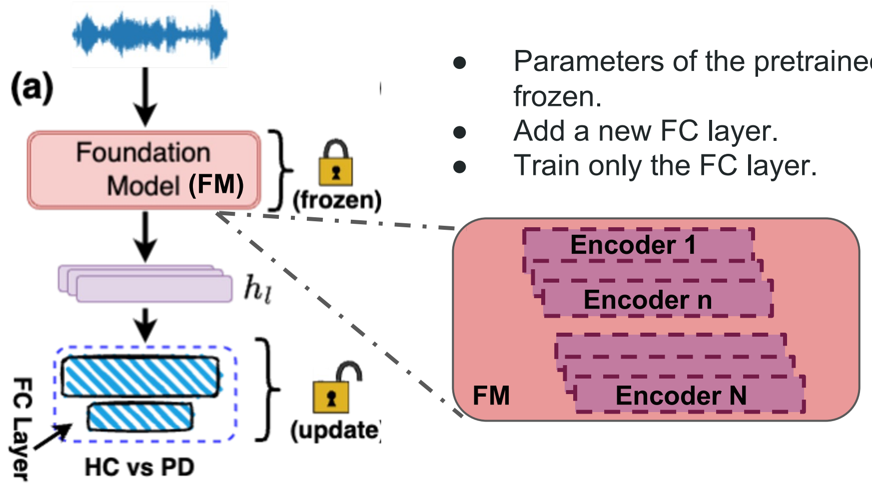
HuBERT

Whisper

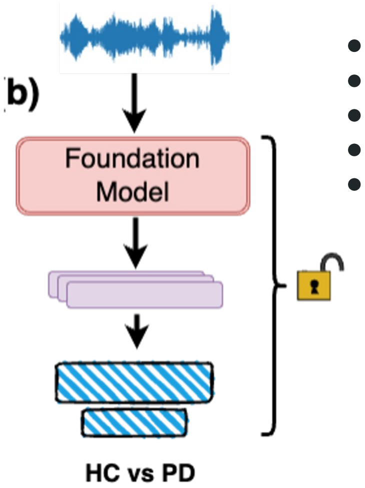
		BASE	LARGE	X-LARGE
CNN Encoder	strides	5, 2, 2, 2, 2, 2, 2		
	kernel width	10, 3, 3, 3, 3, 2, 2		
	channel	512		
Transformer	layer	12	24	48
	embedding dim.	768	1024	1280
	inner FFN dim.	3072	4096	5120
	layerdrop prob	0.05	0	0
	attention heads	8	16	16
Projection	dim.	256	768	1024
Num. of Params		95M	317M	964M

Model architecture summary for BASE, LARGE, and X-LARGE

How to use these models



- Parameters of the pretrained FMs frozen.
- Add a new FC layer.
- Train only the FC layer.

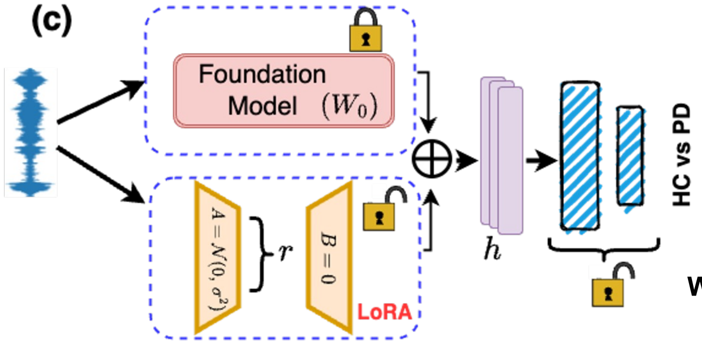


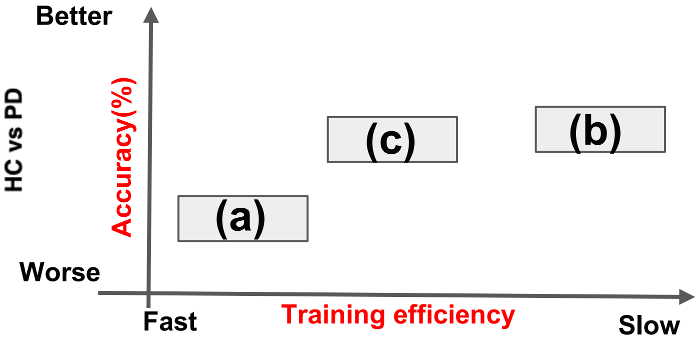
- Unfreeze the parameters of FMs
- Add a new FC layer.
- Train everything together.
- *Provides defacto initialization*
- "Gold standard" for optimizing performance.

- PEFT
- There exists a low dimension reparameterization that is as effective for fine-tuning as the full parameter space.

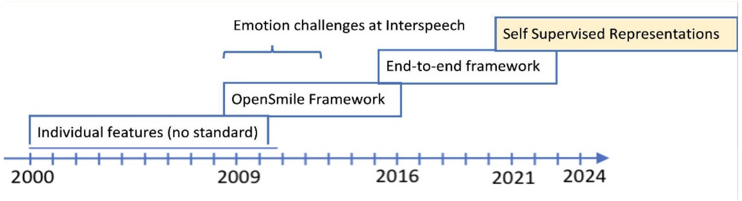
Armen Aghajanyan, et.al, "Intrinsic Dimensionality Explains the Effectiveness of Language Model Fine-Tuning."

- Save only task specific parameters





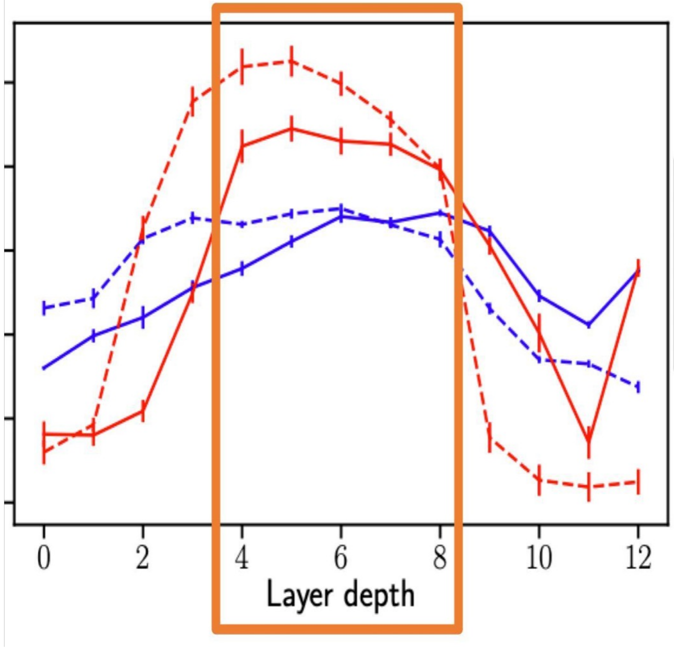
Layer-depth for SER



Analysed layers that contributes towards emotion recognition task.

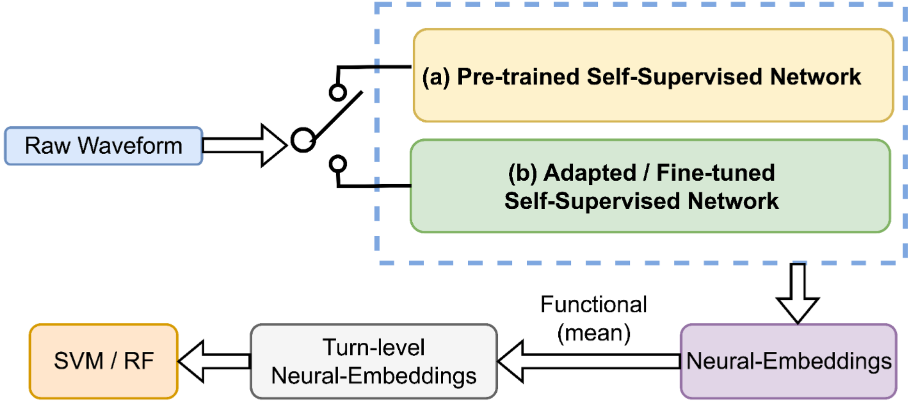
SSL better than Spectrograms.

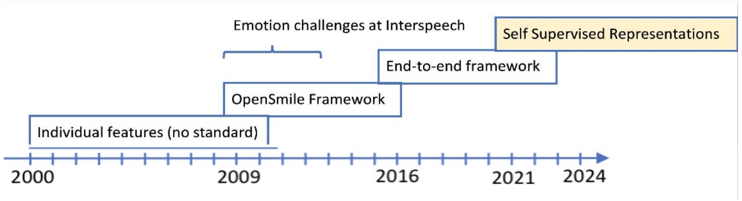
Fine-tuning needed ?

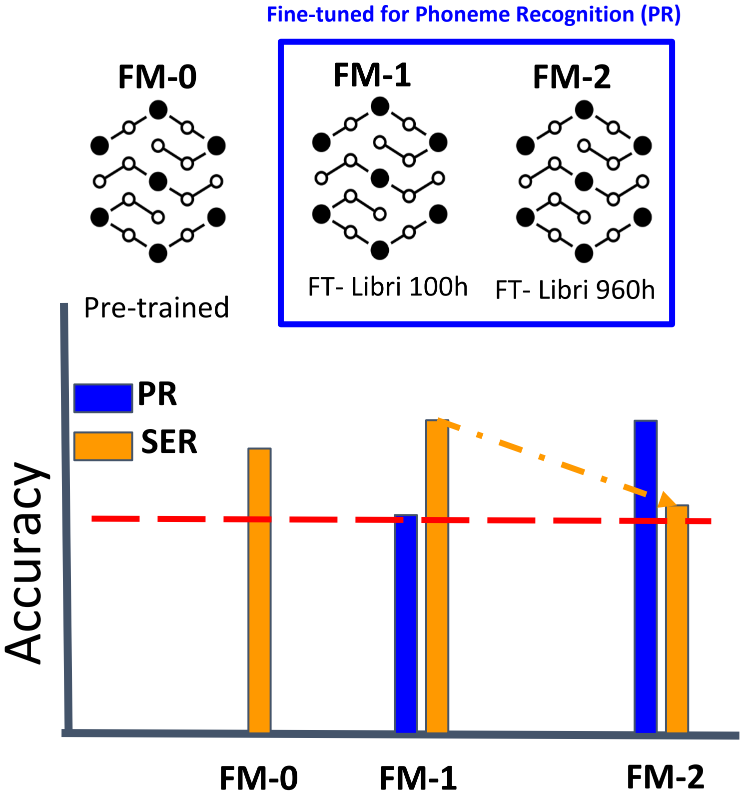


L. Pepino et al, *"Emotion Recognition from Speech Using wav2vec 2.0 Embeddings". In Proc. of INTERSPEECH 2021, Brno, Czechia.*

Fine-tuning for Auxiliary task



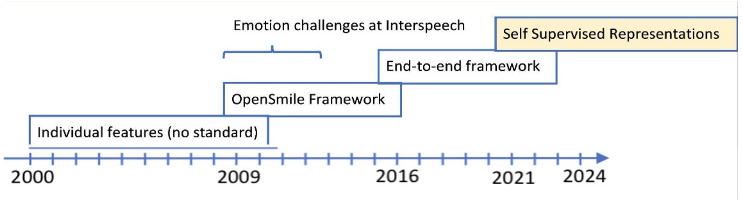




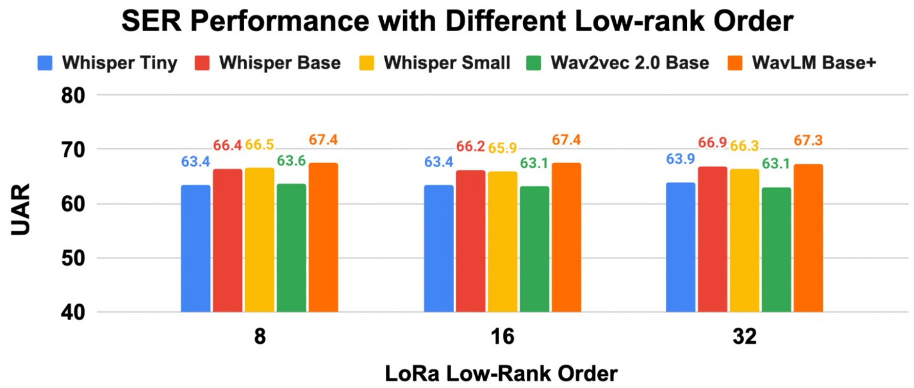
- Phonetic embeddings yield improved SER performance compared to Handcrafted features.
- SER inverse relation with ASR.

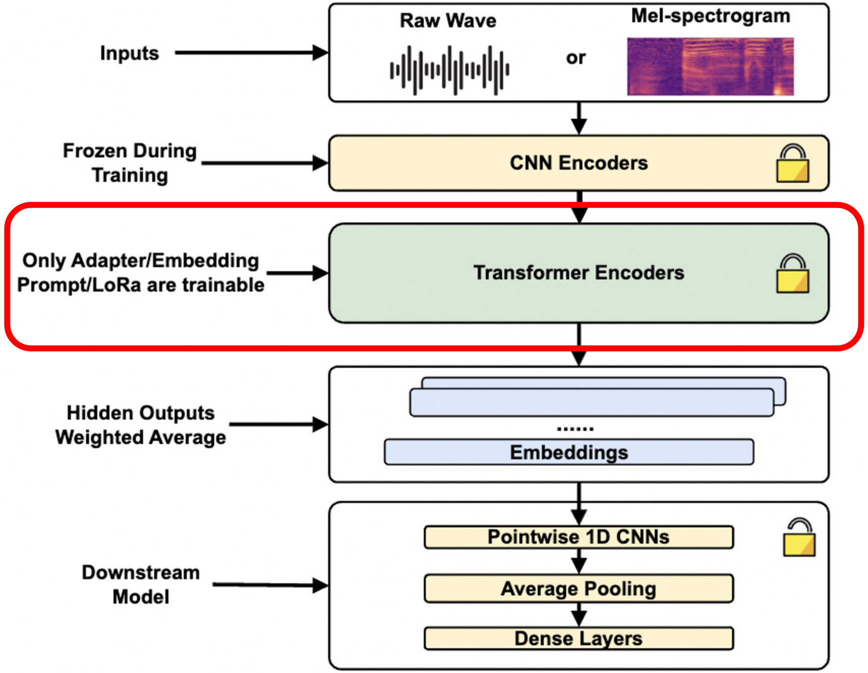
T. Purohit et al, "Implicit phonetic information modeling for speech emotion recognition". In Proc. of INTERSPEECH 2023, Dublin, Ireland.

Parameter efficient tuning for SER



- Used PEFT on transformer representation model for SER
- Utilized low rank approximation (LoRA).
- Best performance with reduced parameters.



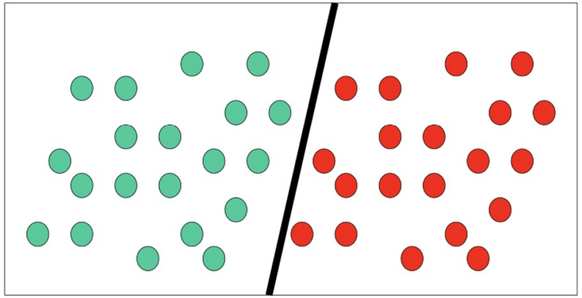


T. Feng et al, "PEFT-SER: On the Use of Parameter Efficient Transfer Learning Approaches For Speech Emotion Recognition Using Pre-trained Speech Models". In Proc. of ACII 2023, Cambridge, MA, USA.

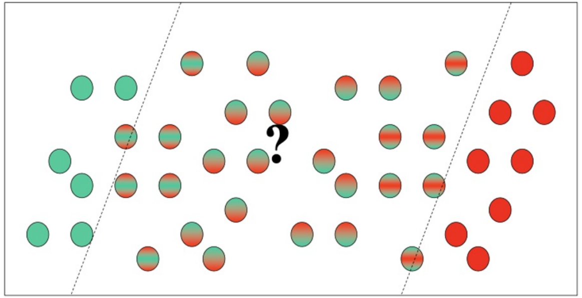
Challenges in the SER community

- Emotions are fuzzy in nature, annotation becomes challenging.
- Acted vs real emotions.
- Lack of Naturalistic databases.
- Low resource data.
- Domain adaptation: train on language-1 test on language-2. Does language matter?
- Cross cultural generalization.
- Privacy issue.

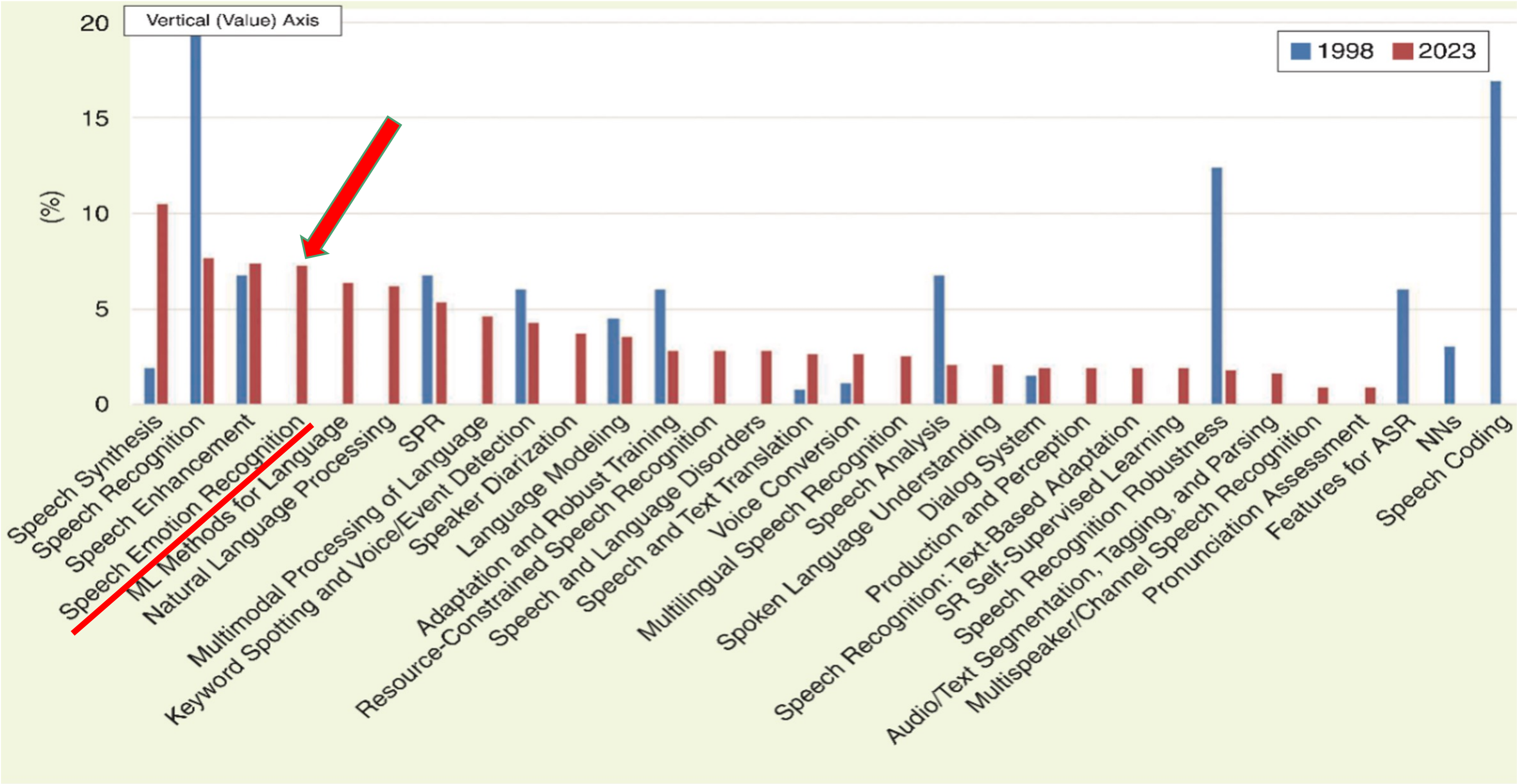
Conventional machine learning problem



Emotion recognition

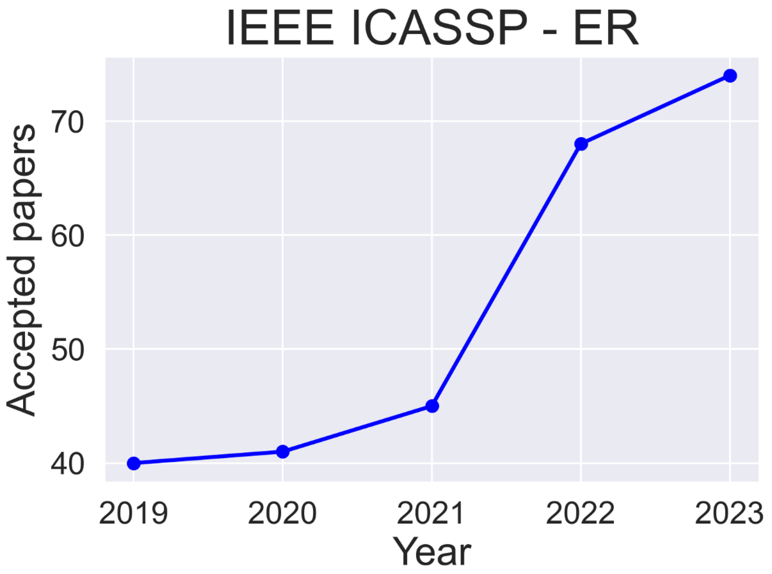


Looking on the bright side..



*Yu, Dong, et al. "Twenty-Five Years of Evolution in Speech and Language Processing." *IEEE Signal Processing Magazine* 40.5 (2023): 27-39.

Stats..



Interspeech 2024 (Kos Island, Greece)
57 accepted papers ; several sessions

Google Scholar

Top publications

Categories › Engineering & Computer Science › Signal Processing

	Publication	h5-index	h5-median
1.	IEEE Transactions on Image Processing	150	202
2.	IEEE Transactions on Wireless Communications	139	205
3.	IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)	129	195
4.	Conference of the International Speech Communication Association (INTERSPEECH)	111	171
5.	IEEE Wireless Communications Letters	97	142
6.	IEEE Transactions on Circuits and Systems for Video Technology	94	131
7.	IEEE Transactions on Signal Processing	93	147
8.	IEEE Journal of Selected Topics in Signal Processing	75	124
9.	IEEE/ACM Transactions on Audio, Speech, and Language Processing	74	124
10.	IEEE Signal Processing Magazine	71	147
11.	Signal Processing	69	112