Low-resource Data Modelling for Speech and Audio: Perspectives from Statistics and Machine Learning

Sneha Das (sned@dtu.dk), Section for Statistics and Data Analysis

DSTS two-day meeting, May 2023

DTU Compute
Department of Applied Mathematics and Computer Science

Speech-technology is all around us!

Smart speakers

Speech synthesis

Automatic speech recognition

Voice activity detection

Voice biometrics

Voice conversion

Speech emotion recognition

Source: speech and audio Free text in full data from app.dimensions.ai (07.05.2023)

Outline

- Speech processing
  - Examples
  - Speech generation
  - Speech processing
- Existing challenges
- Low-resource methods
  - Resource constraints
  - Ex1: Modelling emotions from speech
  - Ex2: Automatic speech recognition
- Conclusions

WristAngel: Intervention and Research for OCD Treatment in Child and Adolescent Psychiatry

PIs of project WristAngel

- Line H. Clemmensen DTU Compute
- Nicole Nadine Lønfeldt Child and Adolescent Mental Health Center, Copenhagen
- Anne Katrine Pagsberg Faculty of Health, Department of Clinical Medicine, KU

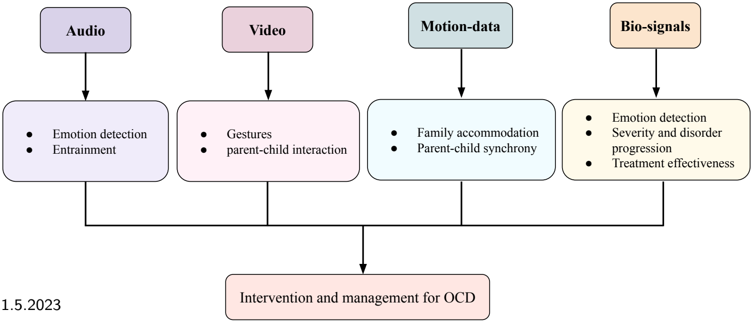
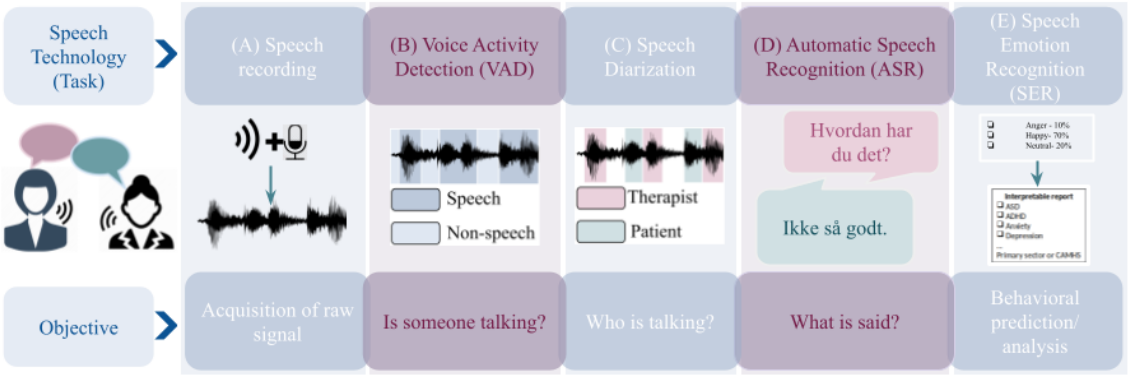


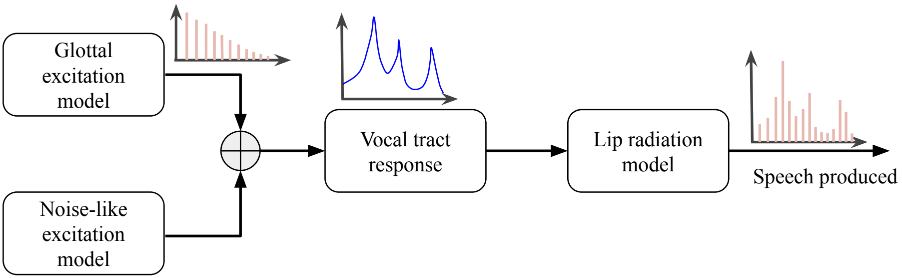
Table of Contents

Examples

Figure 1: Speech tasks

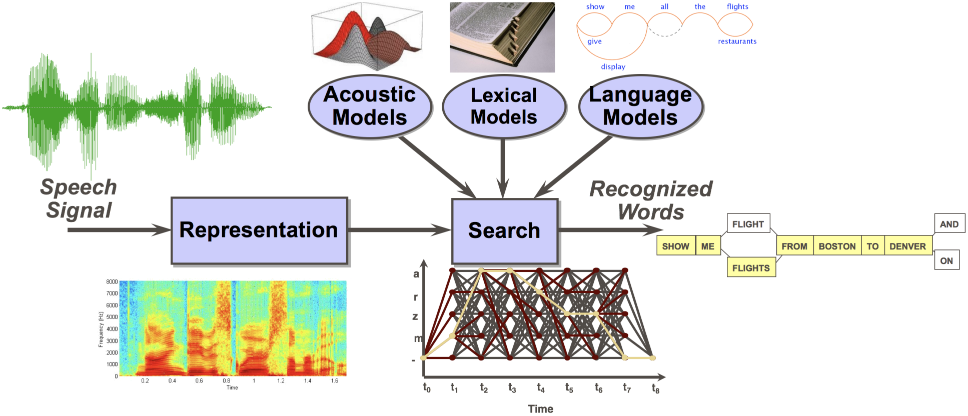


Speech processing: preliminaries



- Speech: looks like a time series data. Spectrogram, Quasi-stationary
- Speech generation: Glottal excitation shaped by vocal tract
- Multiple sources together $\rightarrow$ Speech signal[1]

[1] S. Das, Robust and Efficient Methods for Distributed Speech Processing - Perspectives on Coding, Enhancement and Privacy, PhD Dissertation (2021)



Source: https://www.csail.mit.edu/research/automatic-speech-recognition

Table of Contents

Large (language) models → ChatGPT

Explain causality with math

Causality can be expressed mathematically through the use of statistical models. One common approach is to use regression analysis, which allows us to estimate the effect of one variable (the independent variable or cause) on another variable (the dependent variable or effect), while controlling for other factors that may be related to both variables.

In a simple linear regression model, we can express the relationship between a dependent variable Y and an independent variable X as:

$$Y = \beta 0 + \beta 1X + \varepsilon$$

where β0 is the intercept, β1 is the slope coefficient that represents the effect of X on Y, and ε is the error term that captures the random variation in Y that is not explained by X.

To test for causality, we need to establish that the observed relationship between X and Y is not due to other factors that may be influencing both variables. One way to do this is through a randomized controlled trial (RCT), where participants are randomly assigned to different levels of X, and the effect on Y is measured. In this case, we can estimate the causal effect of X on Y by comparing the average values of Y in the different treatment groups.

Issues

Low-resource learning

- Data
- Labels
- Computation

Evolving data streams

- Eg: Biosignals

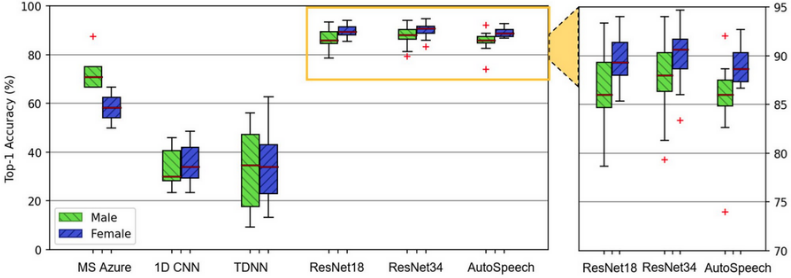
Variability within subjects

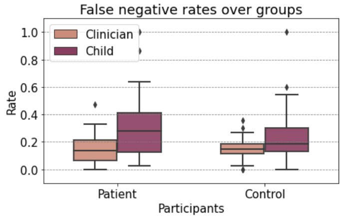
- Eg: Clinical data

Biases

- Language
- Cultural
- Dialects
- Atypical

Biases





[Chen, et.al, 2020]

[Das, et.al, 2021]

Table of Contents

Low-resource machine learning

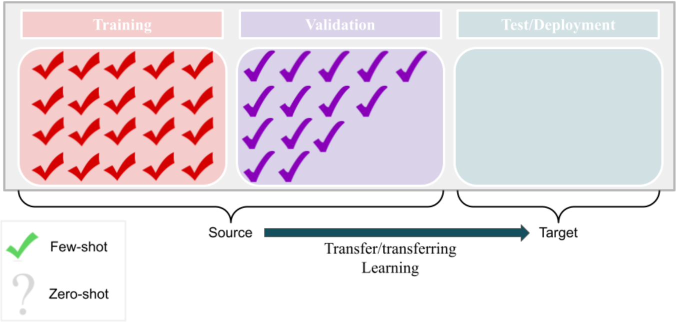
- Transfer-learning, Few-shot and N-shot learning, Zero-shot ...

Low-resource machine learning

- Transfer-learning, Few-shot and N-shot learning, Zero-shot ...
- What do these words mean?

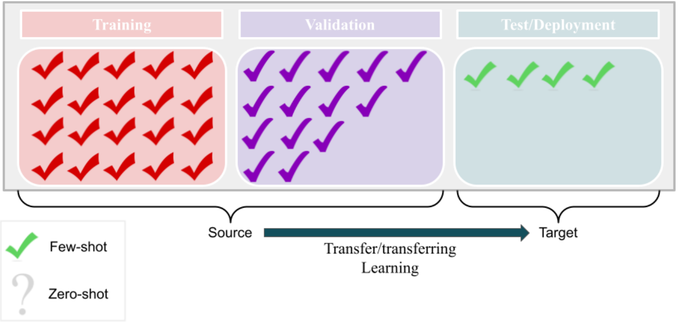
Low-resource machine learning

- Transfer-learning, Few-shot and N-shot learning, Zero-shot ...
- What do these words mean?



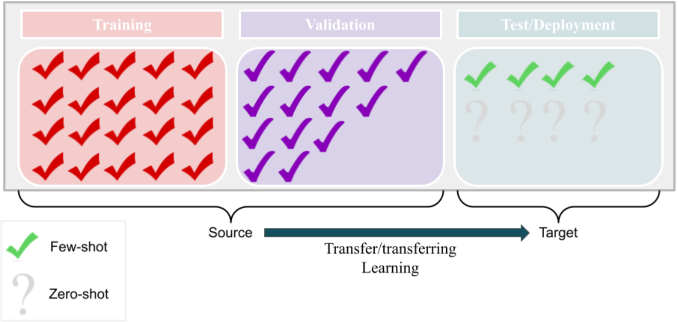
Low-resource machine learning

- Transfer-learning, Few-shot and N-shot learning, Zero-shot ...
- What do these words mean?



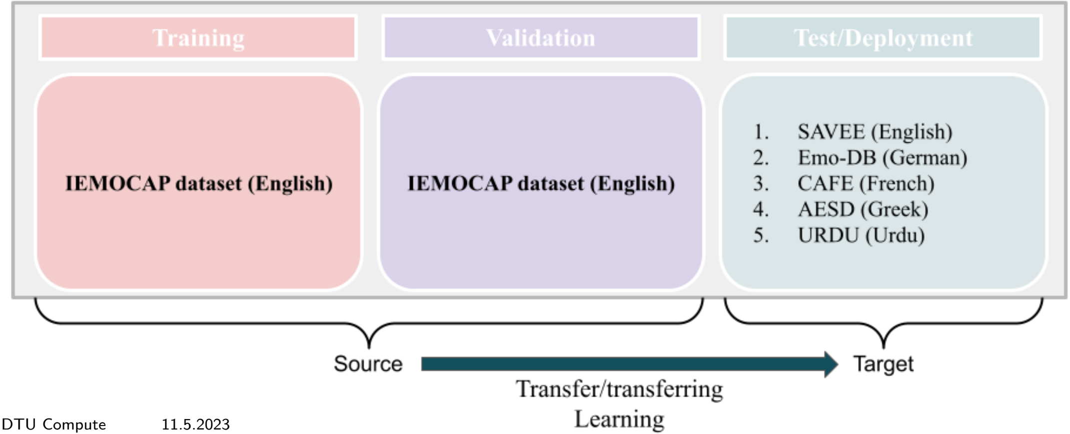
Low-resource machine learning

- Transfer-learning, Few-shot and N-shot learning, Zero-shot ...
- What do these words mean?



In this talk...

[Clemmensen, L, et al. JMIR Research Protocols 2022]



Audio-features $\rightarrow$ (Simple!) Emotion-recognition

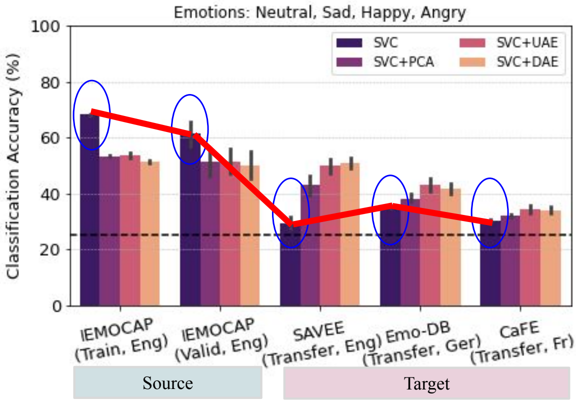
- Input-features: descriptive features of speech features (f_0, tonality, intonation, etc)

Audio-features $\rightarrow$ (Simple!) Emotion-recognition

- Input-features: descriptive features of speech features (f_0, tonality, intonation, etc)
- Input-features $R^{88\times1}$ $\rightarrow$ Support vector machine (SVM) [Das, S, et al. 2021]

Audio-features → (Simple!) Emotion-recognition

- Input-features: descriptive features of speech features (f_0, tonality, intonation, etc)
- Input-features $R^{88 \times 1}$ → Support vector machine (SVM) [Das, S, et al. 2021]



Audio-features → Feature-embedding → Emotion-recognition

- Learning *emotion-relevant* representations of speech!

Audio-features → Feature-embedding → Emotion-recognition

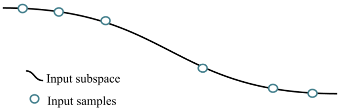
- Learning *emotion-relevant* representations of speech!
- (Denoising) Autoencoder, DAE [Lu, Xugang, et al. 2013]

Audio-features $\rightarrow$ Feature-embedding $\rightarrow$ Emotion-recognition

- Learning *emotion-relevant* representations of speech!
- (Denoising) Autoencoder, DAE [Lu, Xugang, et al. 2013]
- Learns the subspace where the noise free input exists (Type of regularization)

Denoising autoencoder

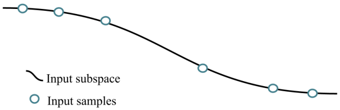
$$\mathbf{x} \in \mathcal{R}^{88\times 1}$$



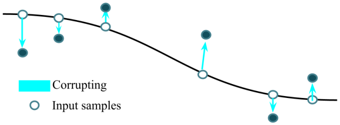
[Das, S, et al. NLDL 2022.]

Denoising autoencoder

$$\mathbf{x} \in \mathcal{R}^{88\times 1}$$



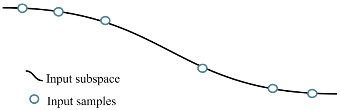
$$\mathbf{x_n} = \mathbf{x} + \mathcal{N}(\mathbf{0}, \sigma_\mathbf{n})$$



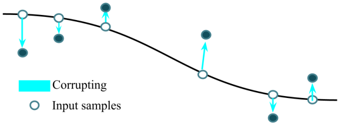
[Das, S, et al. NLDL 2022.]

Denoising autoencoder

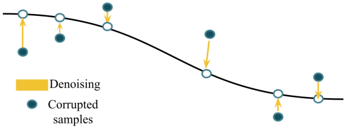
$$\mathbf{x} \in \mathcal{R}^{88\times 1}$$



$$\mathbf{x_n} = \mathbf{x} + \mathcal{N}(\mathbf{0}, \sigma_\mathbf{n})$$

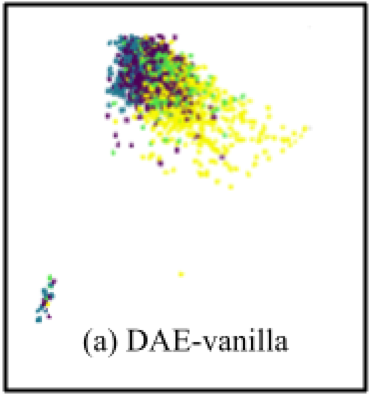


$$\arg\min_{f_\theta, g_\phi} \mathcal{L}_{\text{rec}} = \mathbb{E}\|\mathbf{x} - g_\phi(f_\theta(\mathbf{x_n}))\|_2^2$$



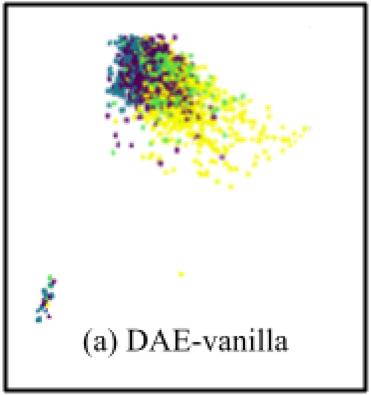
[Das, S, et al. NLDL 2022.]

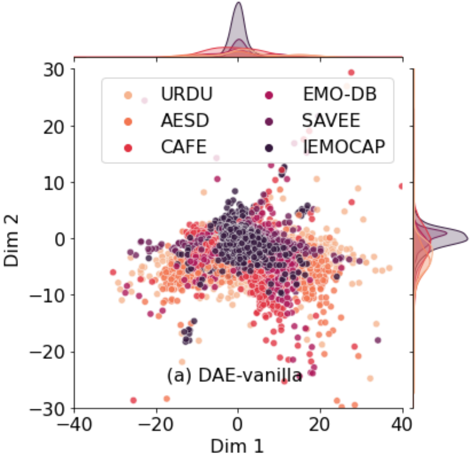
Latent representation of the model



[Das, S, et al. ICASSP 2022.]

Latent representation of the model





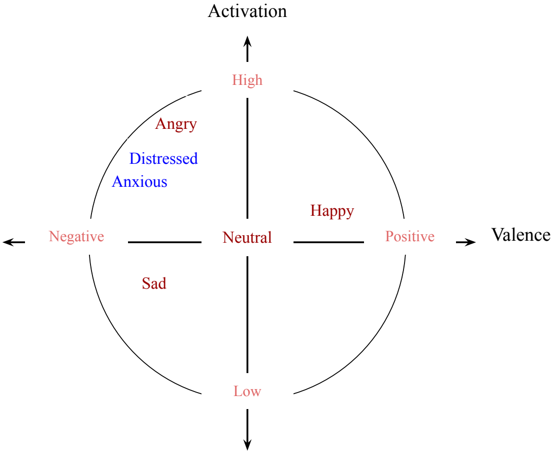
[Das, S, et al. ICASSP 2022.]

Discrete point-estimates → Continuous densities

- DAE: Generated latent space is discontinuous → no meaning in the gaps of the space.

Discrete point-estimates → Continuous densities

- DAE: Generated latent space is discontinuous → no meaning in the gaps of the space.
- Emotions are not discrete!

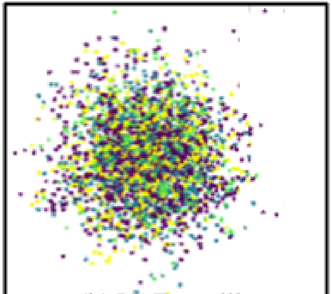


Variational autoencoder (VAE)

The loss function:

$$\begin{aligned}\text{arg}\min_{\theta,\phi} \quad \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{KL}} &= -\mathbb{E}_{\mathbf{z}\sim q_\theta(\mathbf{z}|\mathbf{x})} \log p_\phi(\mathbf{x}|\mathbf{z}) \\ &\quad + D_{KL}(q_\theta(\mathbf{z}|\mathbf{x})||p(\mathbf{z})),\end{aligned} \tag{1}$$

Posterior collapse (VAE) and KL-annealing



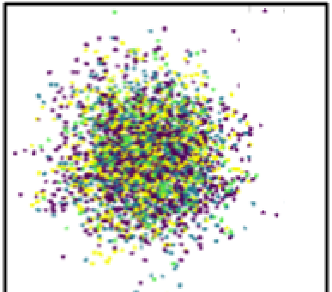
(b) VAE-vanilla

Posterior collapse (VAE) and KL-annealing

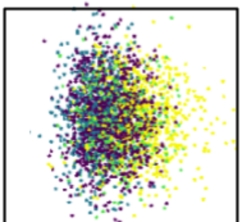
$$\arg\min_{\theta,\phi} \quad \mathcal{L}_{\text{rec}} + \beta\mathcal{L}_{\text{KL}} = -\mathbb{E}_{\mathbf{z}\sim q_\theta(\mathbf{z}|\mathbf{x})} \log p_\phi(\mathbf{x}|\mathbf{z}) + \beta_e D_{KL}(q_\theta(\mathbf{z}|\mathbf{x})||p(\mathbf{z})), \tag{2}$$

where the standard formulation of β_e:

$$\beta_e = \begin{cases} f(\tau) = \frac{0.25}{R}\tau, & \tau \leq R \\ 0.25, & \tau > R \end{cases} \quad \text{where} \quad \tau = \frac{\text{mod}(e-1, \frac{T}{M})}{\frac{T}{M}}, \tag{3}$$



(b) VAE-vanilla

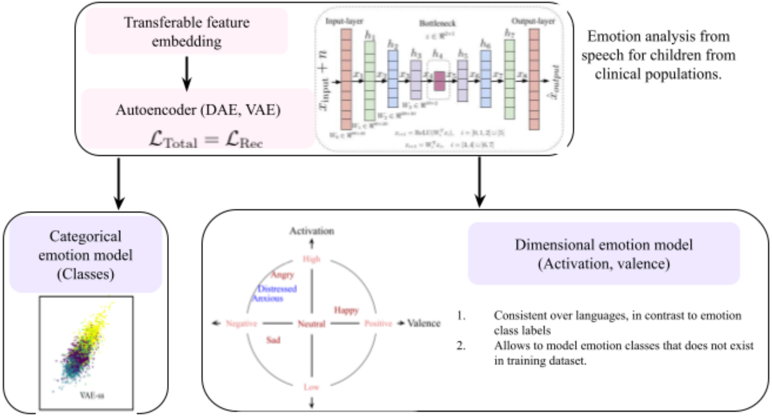


(c) VAE-annealing

[Das, S, et al. ICASSP 2022.]

Transferability: What variable to condition on?

Emotion class (discrete) or dimensional model (continuous)?

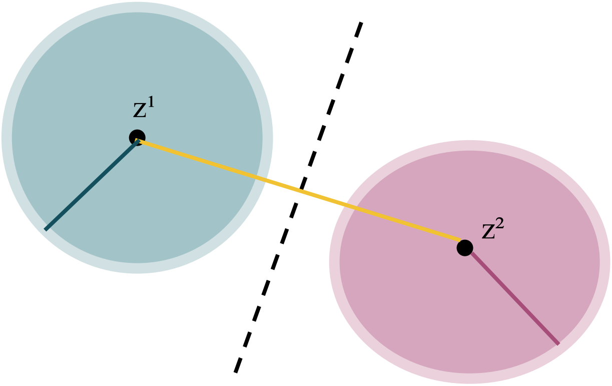


Emotion-class constrained VAE

Centre Loss [Das, S, et al. ICASSP 2022.]:

$$\arg\min_{\theta,\phi} \quad \mathcal{L}_{\text{rec}} + \beta_e \mathcal{L}_{\text{KL}} + \gamma \mathcal{L}_{\text{clus}},$$

$$\mathcal{L}_{\text{clus}} = \frac{D_{\text{intra}}}{D_{\text{inter}}} = \frac{\sum_{k=1}^{K} \sum_{\forall i \in k} D(\mathbf{z_i}, \overline{\mathbf{z}}^{\mathbf{k}})}{\sum_{k=1}^{K-1} \sum_{j=k+1}^{K} D(\overline{\mathbf{z}}^{\mathbf{k}}, \overline{\mathbf{z}}^{\mathbf{j}})}, \tag{4}$$



Dimensional-model constrained VAE

- Metric learning: models learning based on similarity and dissimilarity.

Dimensional-model constrained VAE

- Metric learning: models learning based on similarity and dissimilarity.
- Contrastive, centre-loss, triplet-loss

Dimensional-model constrained VAE

- Metric learning: models learning based on similarity and dissimilarity.
- Contrastive, centre-loss, triplet-loss
- Problem: No loss function to learn continuous contrasts.

VAE with metric-loss

- We came up with one: Continuous metric loss.

$$\arg\min_{f_\theta, g_\phi} \quad \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{KL}} + \mathcal{L}_{\text{met}} = \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{KL}} + \mathcal{L}_{\text{res}} + \mathcal{L}_{\text{sl}}, \tag{5}$$

$$\mathcal{L}_{\text{res}} = \mathbb{E}\|\mathbf{z_d} - \hat{\mathbf{z}}_\mathbf{d}\|_2^2, \quad \hat{\mathbf{z}}_\mathbf{d} = p\mathbf{l_d}, \quad \mathbf{l_d} = d(l_i, l_{i+1}) \tag{6}$$

$$p = (\mathbf{l_d}^T \mathbf{l_d})^{-1} \mathbf{l_d}^T \mathbf{z_d} \tag{7}$$

$$\mathcal{L}_{\text{sl}} = \left\| \frac{\hat{\mathbf{z}}_\mathbf{d}(a_1) - \hat{\mathbf{z}}_\mathbf{d}(a_2)}{\mathbf{l_d}(a_1) - \mathbf{l_d}(a_2)} - 1 \right\|_2, \tag{8}$$

VAE with metric-loss

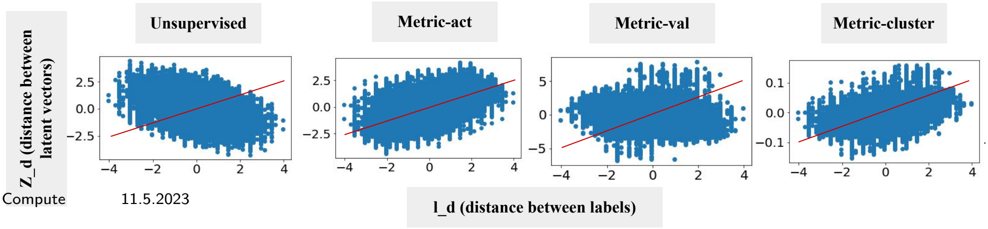
- We came up with one: Continuous metric loss.
- Minimize slope and residual.

$$\arg\min_{f_\theta, g_\phi} \quad \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{KL}} + \mathcal{L}_{\text{met}} = \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{KL}} + \mathcal{L}_{\text{res}} + \mathcal{L}_{\text{sl}}, \tag{5}$$

$$\mathcal{L}_{\text{res}} = \mathbb{E}\|\mathbf{z_d} - \hat{\mathbf{z}}_\mathbf{d}\|_2^2, \quad \hat{\mathbf{z}}_\mathbf{d} = p\mathbf{l_d}, \quad \mathbf{l_d} = d(l_i, l_{i+1}) \tag{6}$$

$$p = (\mathbf{l_d}^T \mathbf{l_d})^{-1} \mathbf{l_d}^T \mathbf{z_d} \tag{7}$$

$$\mathcal{L}_{\text{sl}} = \left\| \frac{\hat{\mathbf{z}}_\mathbf{d}(a_1) - \hat{\mathbf{z}}_\mathbf{d}(a_2)}{\mathbf{l_d}(a_1) - \mathbf{l_d}(a_2)} - 1 \right\|_2, \tag{8}$$

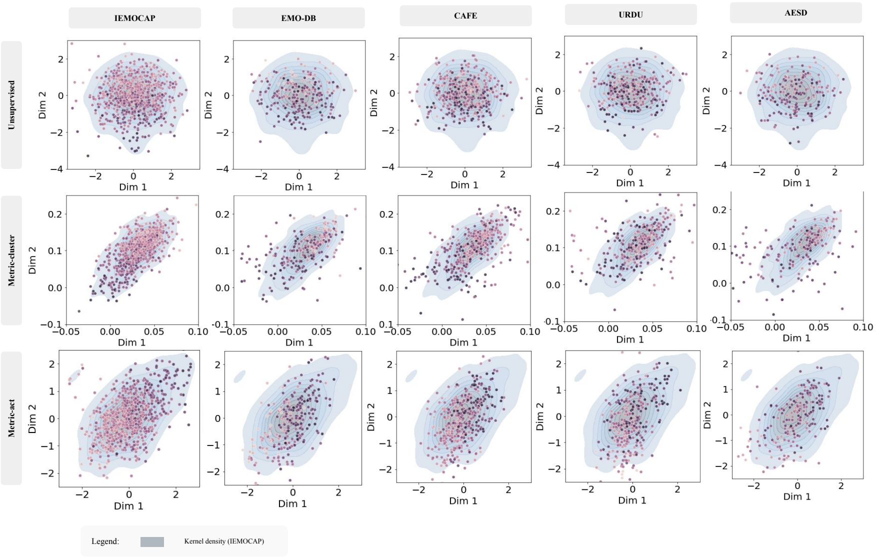


Some results

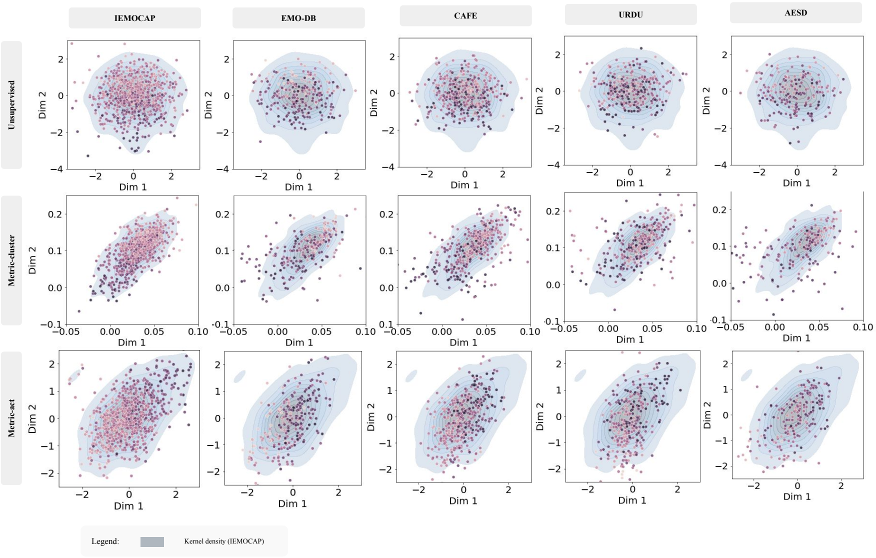
Evaluation criteria: Rank-order correlation, classification-accuracy [Das, S, et al. ISCA SPSC symposium 2022.]

Table: Spearman's rank order correlation for the validation and transfer datasets aggregated over all model runs with different folds and random initial seeds. Higher correlation implies a larger correspondence to the ground truth labels (activation).

Method	IEMOCAP ($\mu \pm \sigma$)		EMO-DB ($\mu \pm \sigma$)		CAFE ($\mu \pm \sigma$)		URDU ($\mu \pm \sigma$)		AESD ($\mu \pm \sigma$)	
	Transfer	Supervised	Transfer	Supervised	Transfer	Supervised	Transfer	Supervised	Transfer	Supervised
Unsupervised	0.26 ± 0.17	0.26 ± 0.17	0.31 ± 0.22	0.31 ± 0.22	0.24 ± 0.14	0.24 ± 0.14	0.12 ± 0.1	0.1 ± 0.07	0.18 ± 0.11	0.16 ± 0.09
Metric-cluster	0.19 ± 0.14	0.19 ± 0.14	0.23 ± 0.16	0.28 ± 0.19	0.12 ± 0.08	0.07 ± 0.04	0.07 ± 0.06	0.09 ± 0.07	0.12 ± 0.06	0.11 ± 0.05
Metric-act	$\mathbf{0.76 \pm 0.05}$	$\mathbf{0.76 \pm 0.05}$	$\mathbf{0.53 \pm 0.08}$	$\mathbf{0.61 \pm 0.04}$	$\mathbf{0.35 \pm 0.04}$	$\mathbf{0.39 \pm 0.03}$	$\mathbf{0.38 \pm 0.05}$	$\mathbf{0.39 \pm 0.05}$	$\mathbf{0.31 \pm 0.01}$	$\mathbf{0.31 \pm 0.01}$
Metric-val	0.29 ± 0.11	0.29 ± 0.11	-0.05 ± 0.03	0.27 ± 0.24	0.31 ± 0.09	0.32 ± 0.1	0.03 ± 0.08	0.07 ± 0.1	0.01 ± 0.05	0.14 ± 0.1



- Scatter range and orientation wrt KDE: Metric-act $\rightarrow$ Unsupervised.



- Scatter range and orientation wrt KDE: Metric-act → Unsupervised.
- Lower correlation for CAFE, URDU, AESD → Different language family (Needs dedicated investigation).

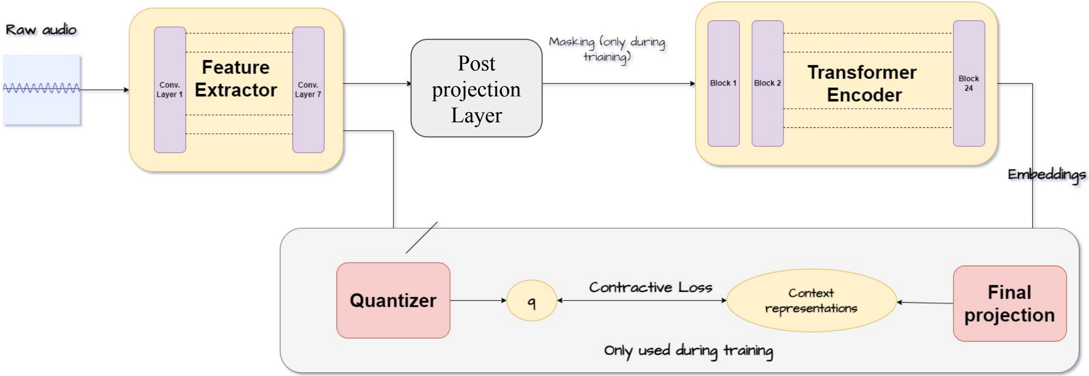
Automatic Speech Recognition and Transcriptions

- Clinical documentation
- Screening, diagnosis, management.

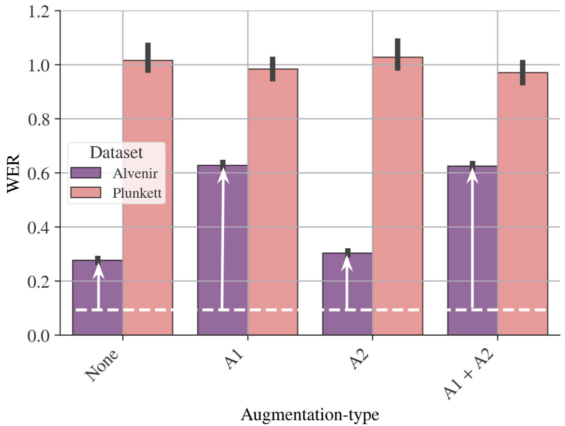
Automatic Speech Recognition and Transcriptions

1. State-of-the-art Models $\rightarrow$ English + Adults
2. State-of-the-model for Danish $\rightarrow$ Alvenir
3. Challenges:
    - Transcribe speech from children in Danish
    - Clinical conversations between clinician and child.
    - Do we have data?

Baseline and Wav2vec Model



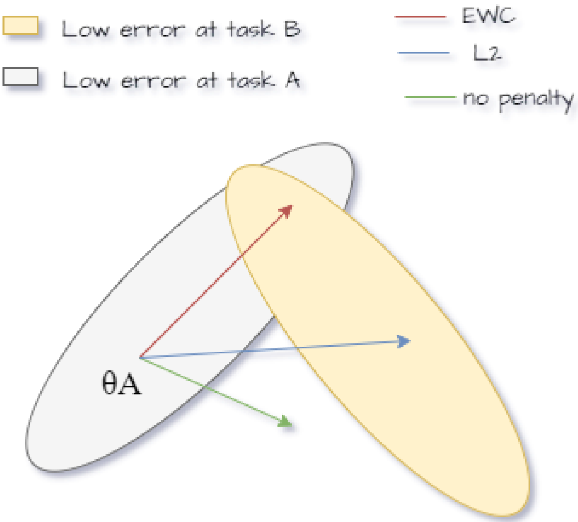
Fine-tuning model using children's dataset



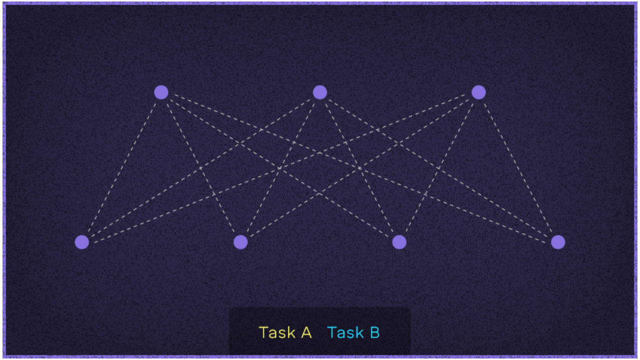
- Testing on Alvenir + Plunkett
- Catastrophic forgetting → Not acceptable (!)

How to avoid Catastrophic forgetting?[J. Kirkpatrick, et.al, 2017]

- Elastic weight consolidation: $L(\theta) = L_B(\theta) + \sum_i \frac{\lambda}{2} F_i(\theta_i - \theta^*_{A,i})^2$

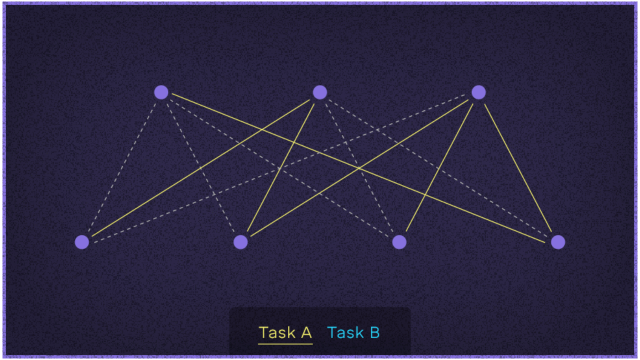


Elastic weight consolidation



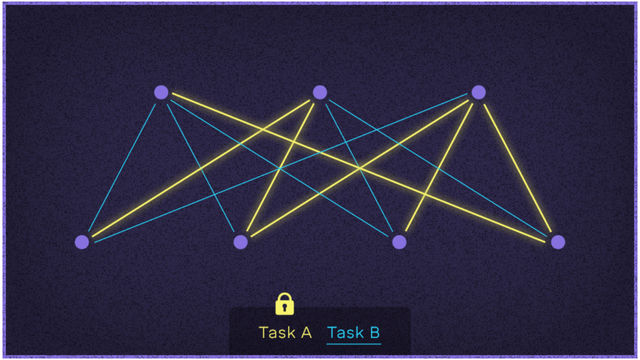
source: https://www.deepmind.com/blog/enabling-continual-learning-in-neural-networks

Elastic weight consolidation



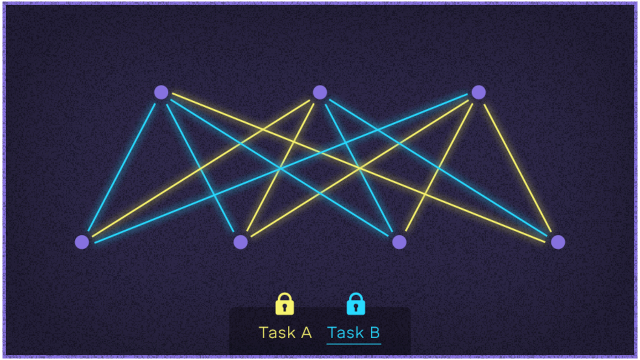
source: https://www.deepmind.com/blog/enabling-continual-learning-in-neural-networks

Elastic weight consolidation



source: https://www.deepmind.com/blog/enabling-continual-learning-in-neural-networks

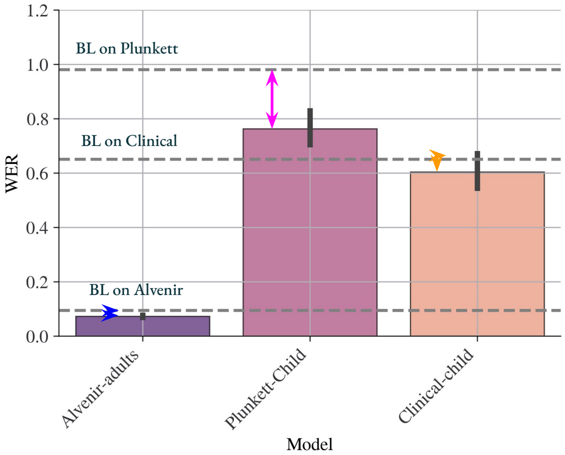
Elastic weight consolidation



source: https://www.deepmind.com/blog/enabling-continual-learning-in-neural-networks

Results

Performance of the best model[1]



[1] Garofalaki. M, Speech and natural language processing for clinical in-the-wild data 2023.

Table of Contents

Summary

- As models getting larger (hungrier for data!), so is the need to devise (smarter!) methods.
- Carefully devise loss-functions.
- Need to re-visit how we evaluate ML/DL models.

Thankyou!

Email: sned@dtu.dk; @dassneh