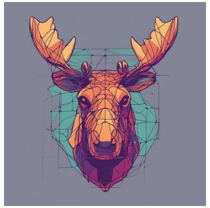
oplatek/moosenet-plda

MooseNet: Synthesized Speech Metric

MooseNet: A Trainable Metric for Synthesized Speech with a PLDA Module

Ondřej Plátek and Ondřej Dušek
{oplatek,odusek}@ufal.mff.cuni.cz



Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

Content

- MOS Prediction task & Data
- MooseNet NN Model
- PLDA Intro & Use
- Experiments
- Summary
- Q&A

Paper:
github.com/oplatek/moosenet-plda
arxiv.org/abs/2301.07087

Personal:
oplatek@ufal.mff.cuni.cz
linkedin.com/in/ondrejplatek

Task: Mean Opinion Scores Prediction

Task & Data: VoiceMOS Challenge

Task

- **Prediction** of speech utterance **score**
- Single score for utterance
- **Gold** score: **Mean** of annotators scores
- Large variance:
  - modelling annotator helps[2]
  - modelling data collection helps [2]
- Models based on SSL are SOTA
  - 2022: Wav2vec 2.0[3], HuBERT [4]

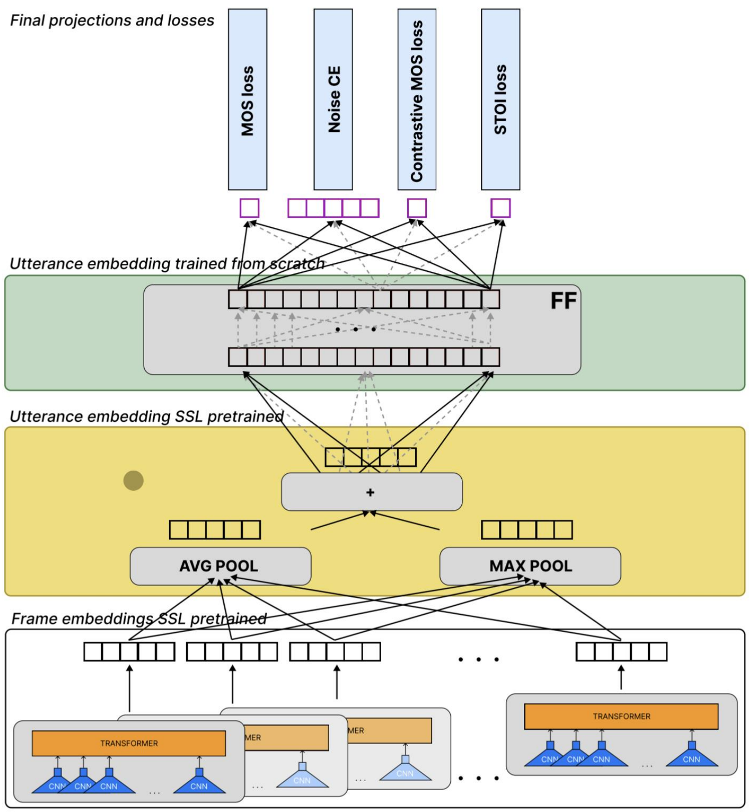
Data[1]

- **VoiceMOS'** two tracks: main & OOD
- Single isolated utterances
- Each rated by multiple annotators
- English and Chinese (OOD)
- Single *overall* score
- Main track from multiple datasets

1. W.-C. Huang, E. Cooper, Y. Tsao, H.-M. Wang, T. Toda, and J. Yamagishi, The VoiceMOS Challenge 2022.
2. W.-C. Huang, E. Cooper, J. Yamagishi, and T. Toda, LDNet: Unified Listener Dependent Modeling in MOS Prediction for Synthetic Speech
3. A. Baevski, H. Zhou, A. Mohamed, and M. Auli, Wav2Vec 2.0: A Framework for Self-Supervised Learning of Speech Representations.
4. W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units

MooseNet NN Training

Neural Network (NN) Architecture & PLDA integration



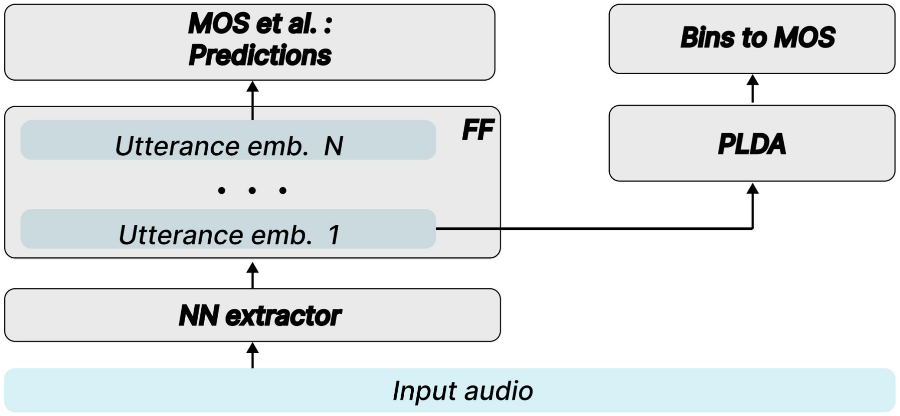


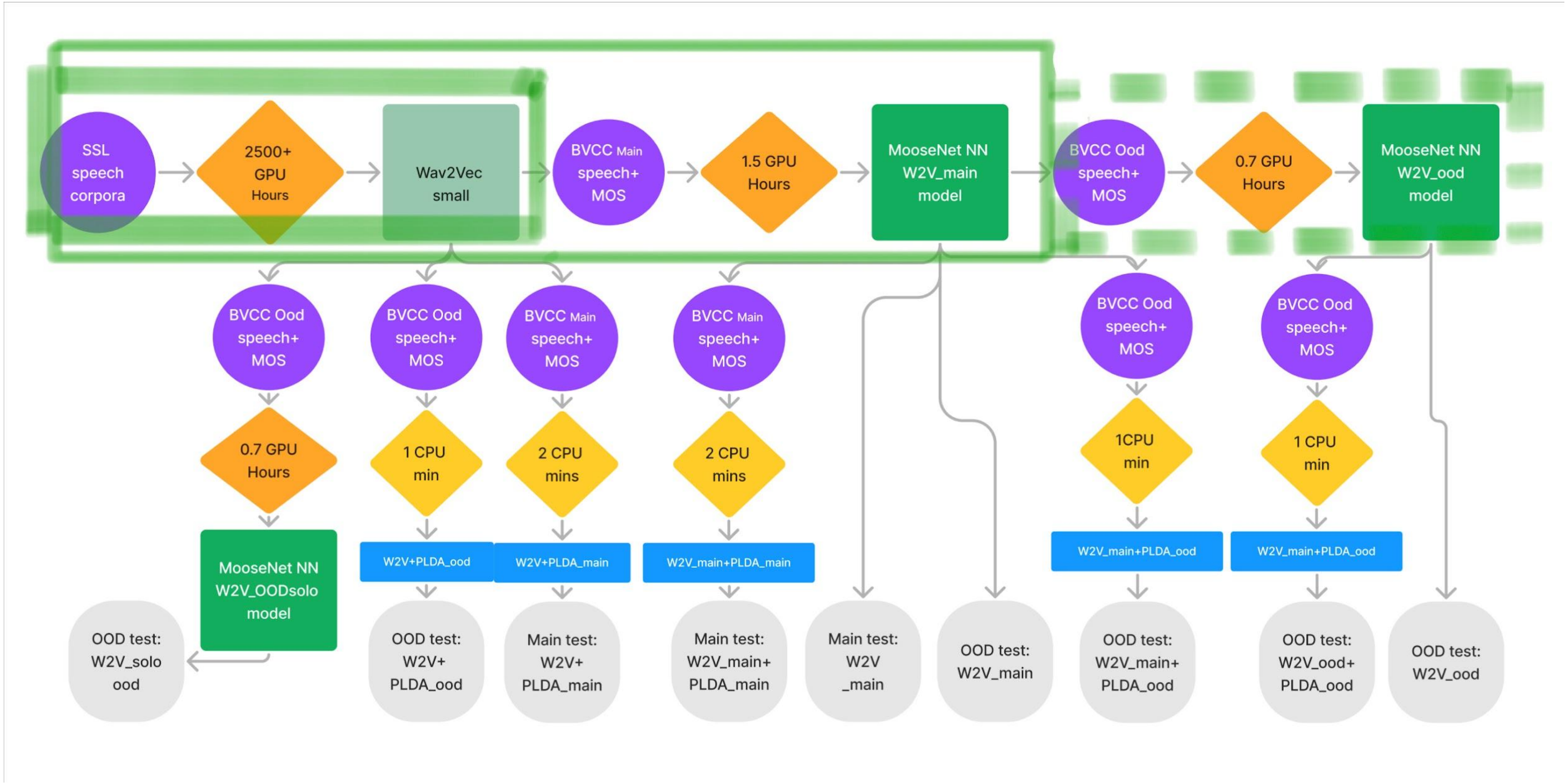
Figure 1: *PLDA can use any layer after global pooling as utterance level embedding as its features.*

Utterance embedding for PLDA

SSL Models: Wav2vec 2.0 & XLSR[1]

[1]A. Babu et al., "XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale."

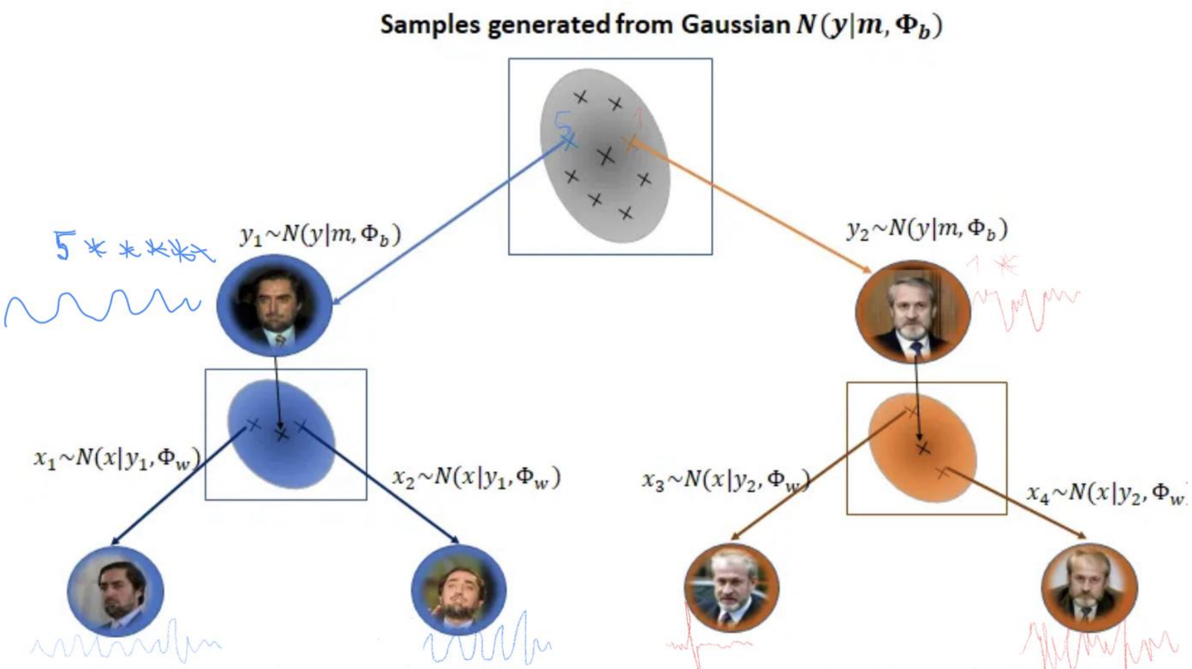
MooseNet Training



Wav2vec + W2V_main + W2V_ood

Probabilistic Linear Discriminant Analysis (PLDA)

PLDA Generative Model for Audio Quality Classes



Source: Images of people are from LFW dataset, representation by author. Plots are for illustration.

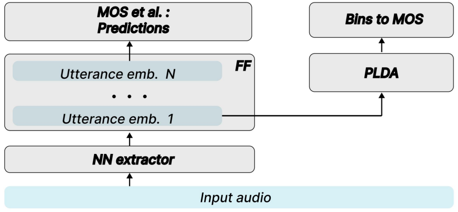


Figure 1: *PLDA can use any layer after global pooling as utterance level embedding as its features.*

PLDA **generative** model

y ~ distribution models **different** classes

x ~ distribution models **similarity** for the class y_i

Image is from https://towardsdatascience.com/probabilistic-linear-discriminant-analysis-plda-explained-253b5effb96

Audio Quality Classes - Binning MOS scores

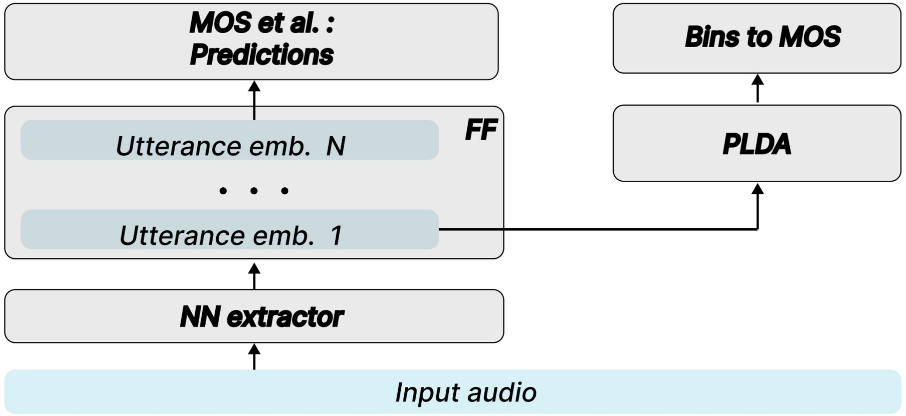
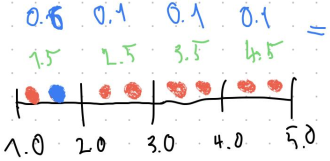


Figure 1: *PLDA can use any layer after global pooling as utterance level embedding as its features.*

$$\sum_{i=1}^{Nbins} BinCenterScore_i * P(i|x)$$

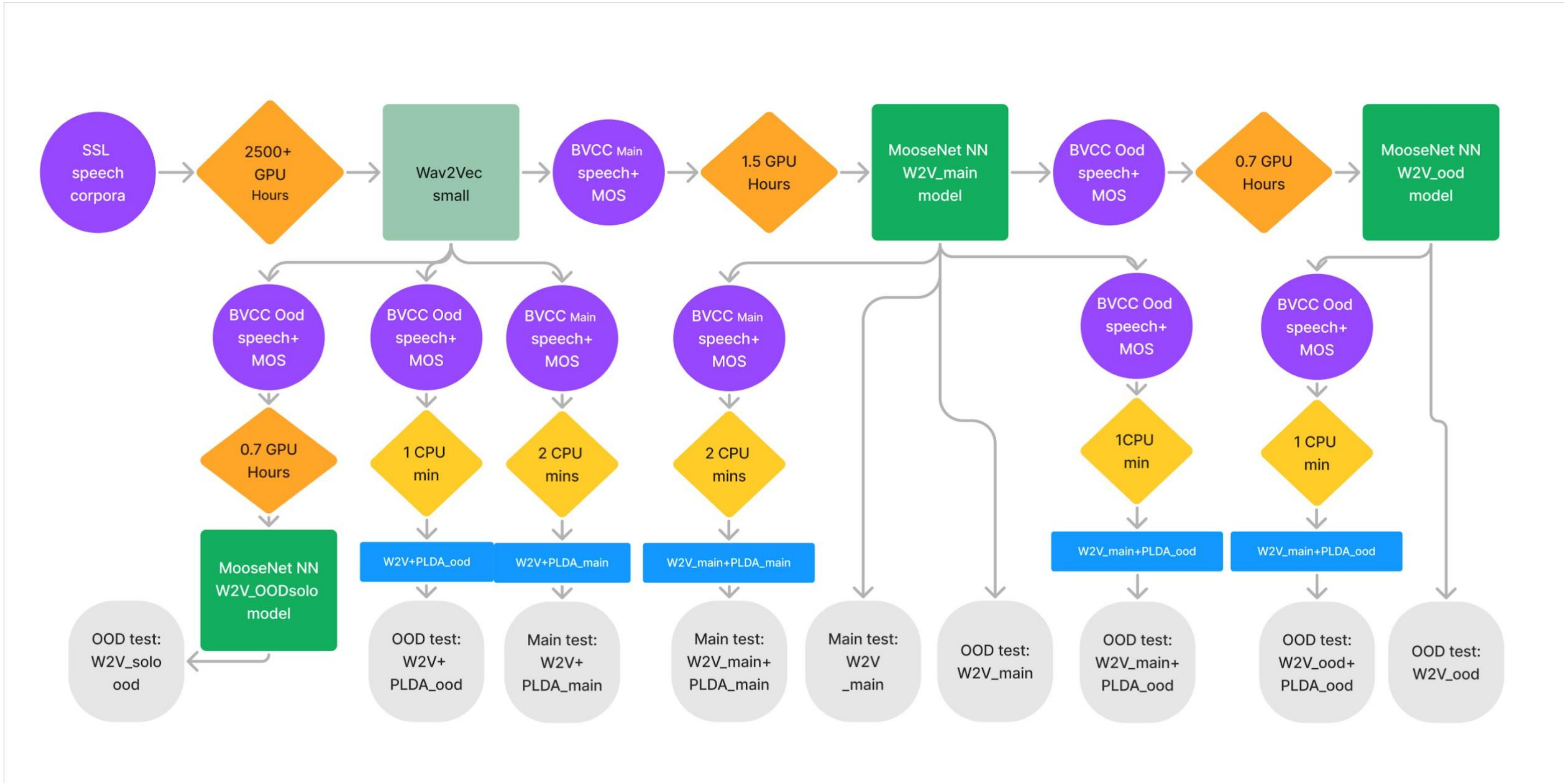


= 1.5*0.6 + 2.5*0.1 + 3.5*0.1 + 4.5*0.1 = 1.95

- PLDA needs classification.
- VoiceMOS main training set contains **only 33 unique scores** for 4974 utterances :)
- PLDA requires representant for each bin.
- Specify number of bins -> boundaries set to have equal number of samples.
- **Posterior Probabilities used as weights.**

Experiments & Results

Experiments Overview



NN(green) PLDA(blue) ckpts

Baselines and RQ1 MooseNet NN on Main Track

Main test system-level:	MSE	SRCC
LDNet baseline	0.178	0.873
SSL-Baseline (B01)	0.148	0.921
W2V_main w/o contrast	0.149±0.033	0.922±0.007
W2V_main w/o augmnt.	**0.137**±0.047	0.922±0.005
W2V_main w/o STOI	0.140±0.033	0.922±0.007
W2V_main_logCosh/Gauss	0.159±0.035	0.922±0.006
W2V_main	0.142±0.032	**0.923**±0.006

- Faster convergence but no significant quality improvement.
- STOI - multi task training with STOI regression computed for original and degraded audio.
- Contrastive Loss [1].
- We first used LogCosh loss[2] and then Gauss loss[3].
- We used dynamic volume and tempo augmentation.

[1]T. Saeki, D. Xin, W. Nakata, T. Koriyama, S. Takamichi, and H. Saruwatari, UTMOS: UTokyo-SaruLab System for VoiceMOS Challenge 2022.

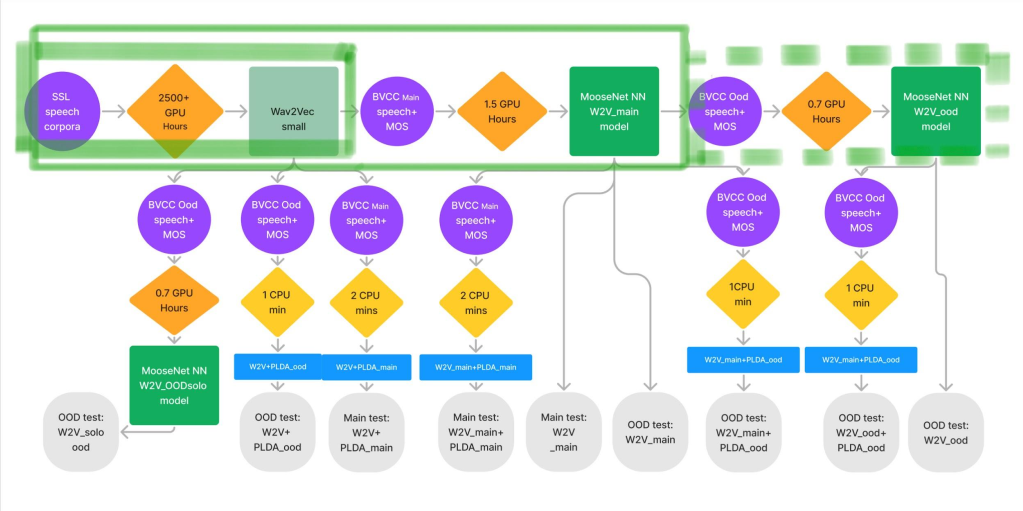
[2]LearningResve A. Saleh,A.K.Md. Ehsanes Saleh, Statistical Properties of the log-cosh Loss Function Used in Machine Learning

[3] Nix, D. A. and Weigend, A. S., “Estimating the mean and variance of the target probability distribution”

RQ2 MooseNet Trained on Main. Evaluated on OOD Track.

OOD test system-level:	MSE	SRCC
LDNet baseline	0.091	0.934
SSL-Baseline (B01)	0.099	0.975
W2V_main	2.657±0.399	0.710±0.040

- Poor performance.
- Absolute values are nonsense.
- Still some correlation.



Results: Zero-shot generalization?

RQ3 PLDA benefits from fine-tuned MooseNet NN?

OOD test system-level:	MSE	SRCC
LDNet baseline	0.091	0.934
SSL-Baseline (B01)	0.099	0.975
W2V_main	2.657±0.399	0.710±0.040
XLSR_main	2.630±0.301	0.748±0.041
W2V_main+PLDA_ood	**0.190**±0.061	0.860±0.042
XLSR_main+PLDA_ood	0.197±0.051	**0.866**±0.039
W2V_ood	0.263±0.128	0.955±0.013
XLSR_ood	**0.058**±0.011	0.942±0.007
W2V_ood+PLDA_ood	0.063±0.008	**0.956**±0.011
XLSR_ood+PLDA_ood	0.062±0.008	0.945± 0.004
W2V_solo-ood	0.265±0.144	0.927±0.023
W2V+PLDA_ood	**0.057**±0.009	**0.955**±0.001
XLSR+PLDA_ood	0.145±0.012	0.886±0.018

- Is more fine-tuning beneficial to PLDA?
- Yes it is :)
- Note also that PLDA improves the fine-tuned models W2v_ood and XLSR_ood :)

Results: PLDA is consistent

RQ4 Can PLDA Be Used without SSL Model Fine-tuning?

Main test system-level:	MSE	SRCC
LDNet baseline	0.178	0.873
SSL-Baseline (B01)	0.148	0.921
W2V_main w/o contrast	0.149±0.033	0.922±0.007
W2V_main w/o augmnt.	**0.137**±0.047	0.922±0.005
W2V_main w/o STOI	0.140±0.033	0.922±0.007
W2V_main_logCosh/Gauss	0.159±0.035	0.922±0.006
W2V_main	0.142±0.032	**0.923**±0.006
W2V_main 50% train	**0.150**±0.044	**0.924**±0.006
W2V_main 5% train	0.307±0.176	0.884±0.006
W2V_main 136 train	0.289±0.072	0.853±0.006
XSLR_main	0.117±0.035	0.929±0.007
W2V_main+PLDA_main	0.105±0.009	0.922±0.006
XSLR_main+PLDA_main	**0.101**±0.010	**0.929**±0.005
W2V+PLDA_main	0.167±0.000	**0.867**±0.000
XLSR+PLDA_main	**0.076**±0.326	0.804±0.109

OOD test system-level:	MSE	SRCC
LDNet baseline	0.091	0.934
SSL-Baseline (B01)	0.099	0.975
W2V_main	2.657±0.399	0.710±0.040
XLSR_main	2.630±0.301	0.748±0.041
W2V_main+PLDA_ood	**0.190**±0.061	0.860±0.042
XLSR_main+PLDA_ood	0.197±0.051	**0.866**±0.039
W2V_ood	0.263±0.128	0.955±0.013
XLSR_ood	**0.058**±0.011	0.942±0.007
W2V_ood+PLDA_ood	0.063±0.008	**0.956**±0.011
XLSR_ood+PLDA_ood	0.062±0.008	0.945± 0.004
W2V_solo-ood	0.265±0.144	0.927±0.023
W2V+PLDA_ood	**0.057**±0.009	**0.955**±0.001
XLSR+PLDA_ood	0.145±0.012	0.886±0.018

- On Small OOD dataset PLDA performs the best.
- Interestingly, fine-tuning MooseNet NN on Main+OOD does not help much.
- Future work: Maybe the fine-tuning first on the main track is not beneficial for better discriminative features.
- On larger datasets the MooseNet NN performance is better - will discuss on next slide.

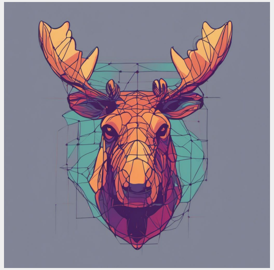
Results: PLDA is consistent

RQ5 How is the NN and PLDA Data Hungry?

Main test system-level:	MSE	SRCC
LDNet baseline	0.178	0.873
SSL-Baseline (B01)	0.148	0.921
W2V_main w/o contrast	0.149±0.033	0.922±0.007
W2V_main w/o augmnt.	**0.137**±0.047	0.922±0.005
W2V_main w/o STOI	0.140±0.033	0.922±0.007
W2V_main_logCosh/Gauss	0.159±0.035	0.922±0.006
W2V_main	0.142±0.032	**0.923**±0.006
W2V_main 50% train	**0.150**±0.044	**0.924**±0.006
W2V_main 5% train	0.307±0.176	0.884±0.006
W2V_main 136 train	0.289±0.072	0.853±0.006
XSLR_main	0.117±0.035	0.929±0.007
W2V_main+PLDA_main	0.105±0.009	0.922±0.006
XSLR_main+PLDA_main	**0.101**±0.010	**0.929**±0.005
W2V+PLDA_main	0.167±0.000	**0.867**±0.000
XLSR+PLDA_main	**0.076**±0.326	0.804±0.109
OOD test system-level:	**MSE**	**SRCC**
LDNet baseline	0.091	0.934
SSL-Baseline (B01)	0.099	0.975
W2V_main	2.657±0.399	0.710±0.040
XLSR_main	2.630±0.301	0.748±0.041
W2V_main+PLDA_ood	**0.190**±0.061	0.860±0.042
XLSR_main+PLDA_ood	0.197±0.051	**0.866**±0.039
W2V_ood	0.263±0.128	0.955±0.013
XLSR_ood	**0.058**±0.011	0.942±0.007
W2V_ood+PLDA_ood	0.063±0.008	**0.956**±0.011
XLSR_ood+PLDA_ood	0.062±0.008	0.945± 0.004
W2V_solo-ood	0.265±0.144	0.927±0.023
W2V+PLDA_ood	**0.057**±0.009	**0.955**±0.001
XLSR+PLDA_ood	0.145±0.012	0.886±0.018

- Surprisingly the MooseNet NN improves quite quickly - See the ablation study on main track.
- The experiments on OOD track shows that PLDA outperforms pure NN MooseNet
- However, we compared PLDA trained on full set with 5% of the main track data which is enough for MooseNet NN to beat SRCC ranking.
- 50% of the main train set beets the PLDA which used 100% of the data in both MSE and SRCC
- **In general, PLDA excels in adjusting the scale but the NN feature are already very discriminative and can be easily further fine-tuned.**

Results: PLDA is consistent



MooseNet: SSL Model Finetuning + PLDA

Summary

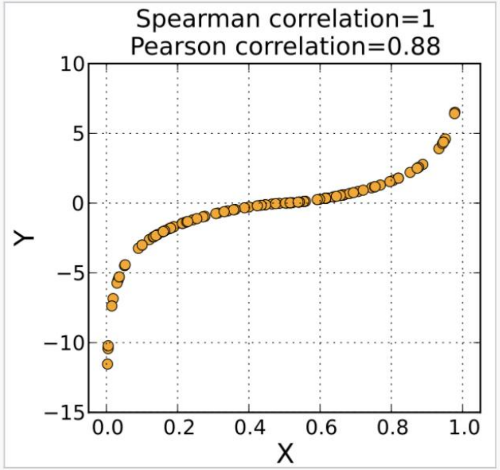
- MooseNet NN: Predicts MOS using regression on top of SSL model.
- MooseNet + PLDA:
- PLDA: classification into numerical labels which are weighted
  - PLDA clusters input features: Each cluster has a MOS label.
  - Posterior probabilities are used as weights for numerical labels.
- PLDA shines for small datasets.
- PLDA cannot improve SSL embeddings to be more discriminative — ranking is not improved but improves scale.

https://github.com/oplatek/moosenet-plda

{oplatek,odusek}@ufal.mff.cuni.cz

Expected QA

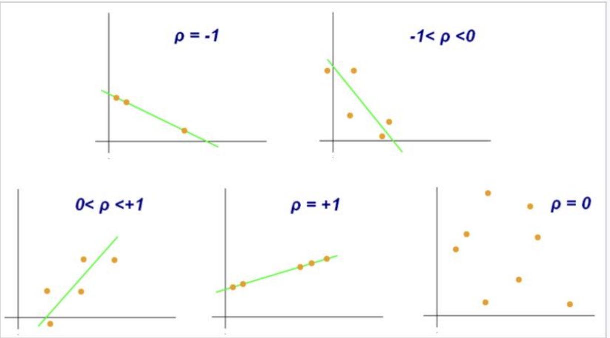
SRCC metric - Spearman Correlation Coefficient



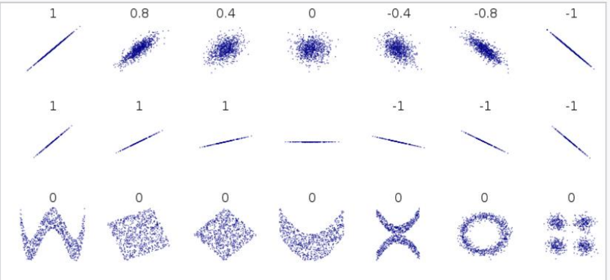
A Spearman correlation of 1 results when the two variables being compared are monotonically related, even if their relationship is not linear. This means that all data points with greater *x* values than that of a given data point will have greater *y* values as well. In contrast, this does not give a perfect Pearson correlation.

Table 3: *Linear correlation coefficients between system-level metrics, using the main track results.*

	MSE	LCC	SRCC	KTAU
MSE	1.00	-.875	-.862	-.870
LCC	-	1.00	.997	.994
SRCC	-	-	1.00	.994
KTAU	-	-	-	1.00



Examples of scatter diagrams with different values of correlation coefficient (ρ)



Several sets of (x, y) points, with the correlation coefficient of x and y for each set. Note that the correlation reflects the strength and direction of a linear relationship (top row), but not the slope of that relationship (middle), nor many aspects of nonlinear relationships (bottom). N.B.: the figure in the center has a slope of 0 but in that case the correlation coefficient is undefined because the variance of Y is zero.

- We use only SRCC and MSE.
- SRCC, KTAU and LCC correlate highly on VoiceMOS dataset.
- Pearson depends on scale, Spearman does not.
- Spearman evaluate ranking
- MSE evaluates absolute values.
- MSE and SRCC are complementary.

Table 3 is from W.-C. Huang, E. Cooper, Y. Tsao, H.-M. Wang, T. Toda, and J. Yamagishi, The VoiceMOS Challenge 2022.

The other two pictures are from https://en.m.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient

All Results and Experiments

Main test system-level:	MSE	SRCC
LDNet baseline	0.178	0.873
SSL-Baseline (B01)	0.148	0.921
W2V_main w/o contrast	0.149±0.033	0.922±0.007
W2V_main w/o augmnt.	**0.137**±0.047	0.922±0.005
W2V_main w/o STOI	0.140±0.033	0.922±0.007
W2V_main_logCosh/Gauss	0.159±0.035	0.922±0.006
W2V_main	0.142±0.032	**0.923**±0.006
W2V_main 50% train	**0.150**±0.044	**0.924**±0.006
W2V_main 5% train	0.307±0.176	0.884±0.006
W2V_main 136 train	0.289±0.072	0.853±0.006
XSLR_main	0.117±0.035	0.929±0.007
W2V_main+PLDA_main	0.105±0.009	0.922±0.006
XSLR_main+PLDA_main	**0.101**±0.010	**0.929**±0.005
W2V+PLDA_main	0.167±0.000	**0.867**±0.000
XLSR+PLDA_main	**0.076**±0.326	0.804±0.109
OOD test system-level:	**MSE**	**SRCC**
LDNet baseline	0.091	0.934
SSL-Baseline (B01)	0.099	0.975
W2V_main	2.657±0.399	0.710±0.040
XLSR_main	2.630±0.301	0.748±0.041
W2V_main+PLDA_ood	**0.190**±0.061	0.860±0.042
XLSR_main+PLDA_ood	0.197±0.051	**0.866**±0.039
W2V_ood	0.263±0.128	0.955±0.013
XLSR_ood	**0.058**±0.011	0.942±0.007
W2V_ood+PLDA_ood	0.063±0.008	**0.956**±0.011
XLSR_ood+PLDA_ood	0.062±0.008	0.945± 0.004
W2V_solo-ood	0.265±0.144	0.927±0.023
W2V+PLDA_ood	**0.057**±0.009	**0.955**±0.001
XLSR+PLDA_ood	0.145±0.012	0.886±0.018