Multi-metric Speech Evaluation

Jiatong Shi

Carnegie Mellon University, Langauge Technologies Institute

jiatongs@cs.cmu.edu

Content

- Motivation
- Toolkit Design: VERSA and VERSA-v2
- Efficient Unified Evaluation: Uni-VERSA
- Autoregressive Multi-metric Evaluation: ARECHO
- Future Scope

Motivation

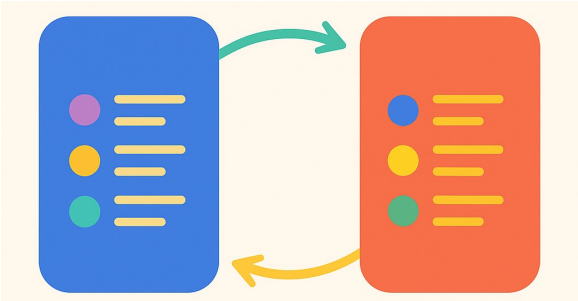
- Speech conveys far more than text

Motivation

- Human subjectiveness are naturally multi-factor
  - No single automatic measure tracks subjective MOS across all noise types, accents, or synthesis artifacts
  - A portfolio of metrics offers higher combined correlation and reliability

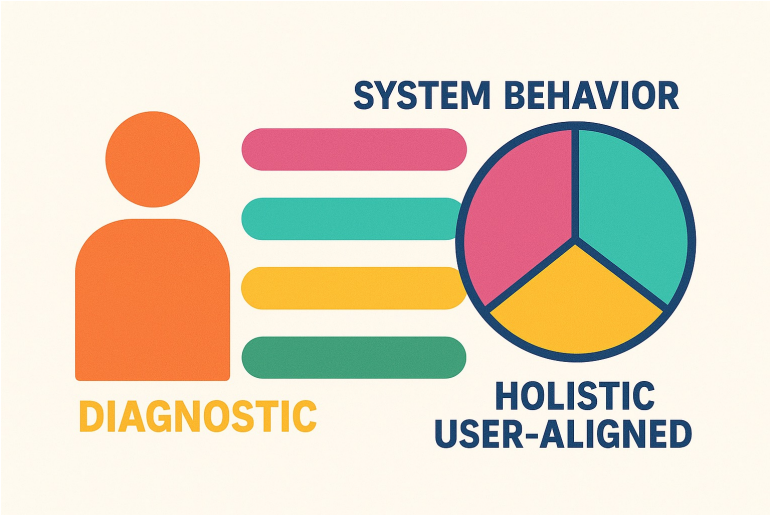
Motivation

- Feeds richer learning objectives
  - Many modern systems train with multi-objective or reinforcement signals derived from several metrics
  - Reliable evaluation on the same set of metrics closes the training-testing loop



Motivation

- Multi-metric mindset towards
  - A diagnostic, holistic, and user-aligned view of system behavior
  - Crucial for both cutting-edge research and real-world deployment



Roadmap

VERSA
Comprehensive Toolkit

Uni-VERSA
Fast-efficient multi-metric estimation

ARECHO
High-performance + interpretable multi-metric estimation

Future
Glance at the future

VERSA: Versatile Evaluation of Speech and Audio

Jiatong Shi[1], Hye-jin Shim[1], Jinchuan Tian[1], Siddhant Arora[1], Haibin Wu[2], Daruis Petermann[3], Jia Qi Yip[4], You Zhang[5], Yuxun Tang[6], Wangyou Zhang[7], Dareen Safar Alharthi[1], Yichen Huang[1], Koichi Saito[8], Jionghao Han[1], Yiwen Zhao[1], Chris Donahue[1], Shinji Watanabe[1]

[1]CMU, [2]Microsoft, [3]IU, [4]NTU, [5]UR, [6]RUC, [7]SJTU, [8]Sony AI

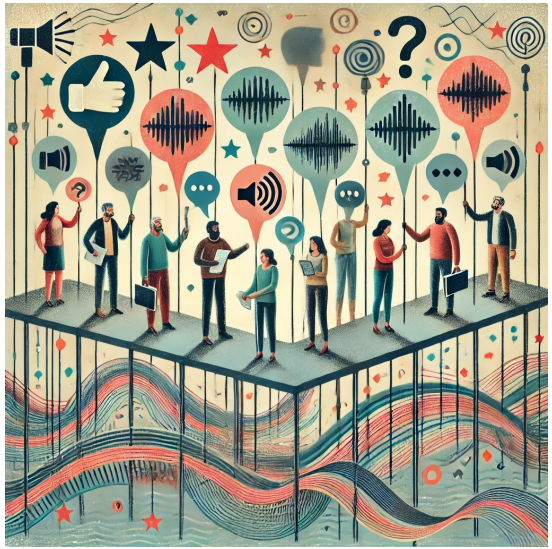
07/17/2025

VERSA: born from challenges

- Plenty of difficulties exist in evaluation for generated speech/audio quality

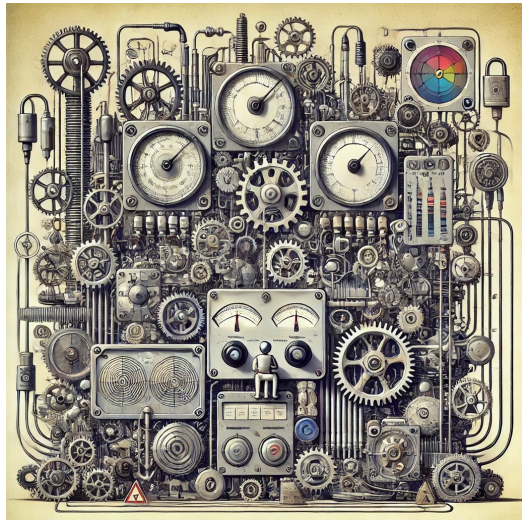
VERSA: born from challenges

- Plenty of difficulties exist in evaluation for generated speech/audio quality
  - Collecting subjective evaluations is not easy



VERSA: born from challenges

- Plenty of difficulties exist in evaluation for generated speech/audio quality
  - Collecting subjective evaluations is not easy
  - Implementing objective evaluations is not easy



VERSA: born from challenges

- Plenty of difficulties exist in evaluation for generated speech/audio quality:
  - Collecting subjective evaluations is not easy
  - Implementing objective evaluations is not easy
  - Increasing need to support multi-domain larger-scale evaluation with unified speech modeling

Introducing VERSA

- VERSA (Versatile Evaluation for Speech and Audio)
  - Targets a general interface for speech and audio evaluation
  - A collection of conventional/recent automatic quality evaluation metrics
  - Highly integration to toolkits / challenges

VERSA Framework

- Simple Command line Interface

```
python versa/bin/scorer.py \
    --score_config egs/general.yaml \
    --gt <ground truth audio list> \
    --pred <candidate audio list> \
    --output_file test_result
```

- Easy Configuration

```yaml
# Example Speaker Similarity YAML config

# speaker related metrics
# -- spk_similarity: speaker cosine similarity
#                    model tag can be any ESPnet-SPK huggingface repo at
#                    https://huggingface.co/espnet
- name: speaker
  model_tag: default
```

```yaml
# Example Speaker Similarity YAML config

# speaker related metrics
# -- spk_similarity: speaker cosine similarity
- name: speaker
  model_tag: espnet/voxcelebs12_ecapa_wavlm_joint
```

VERSA Framework

- Flexible I/O Interface
- Easy Manageable Configuration
- Strict Dependency Control
- Job-scheduling Support
- Continuous Integration Tests
- Efficient Resource Management (cache control)
- Community-driven Effort

```
+---versa
|   |launch.sh # launch examples with SLURM
|   \---egs    # example configs/running scripts
|   |   |general.yaml        # default metric list
|   |   |run_cpu.sh          # run CPU metrics
|   |   |run_gpu.sh          # run GPU metrics
|   |   \---separate_metrics # Config for individual metrics
|   |           mcd_f0.yaml
|   |           spk.yaml
|   |           ...
|   \---versa                # main library
|       |scorer_shared.py    # module register interface
|       \---bin              # binary interface
|               scorer.py    # scorer interface
|               aggregate_results.py  # aggregate corpus-level results
|       \---corpus_metrics   # corpus-level metrics
|               fad.py
|               clap_score.py
|               ...
|       \---utterance_metrics # utterance-level metrics
|               mcd_f0.py
|               spk.py
|               ..
|   \---tools                # external tools
|       install_emo2vec.sh
|       install_fadtk.sh
|       ...
```

Metrics in VERSA

- Up to **90 metrics** in speech and audio evaluation supported currently

- Community-driven efforts for future adaptation!

Metrics in VERSA

Dependency	Example Evaluation Metrics
Dependent	MCD, F0-RMSE, F0-CORR, SI-SNR, CI-SDR, PESQ, STOI, WARPQ, VISQOL, SpeechDiscreteBLEU, SpeechDiscreteDistance, SpeechBERT, etc.
Independent	DNSMOS, UTMOS, PLCMOS, SingMOS, Torch-squim (PESQ, STOI, SI-SNR), Sheet-SSQA, PAM, SpoofS, SWR, etc.
Non-matching	CER/WER (ESPnet/ESPnet-OWSM, Whisper),, CLAP-score, Log-WMSE, EMO-SIM, APA, LLR, NOMAD, LLR, etc.
Distributional	FAD, KID, Coverage, Density, KLD

Check full list and details about metrics in the paper!

Comparison to existing open-source toolkits

Toolkit	Github Repo	Evaluation-Focus	Supported Metrics
ESPnet	espnet/espnet	❌	16
Amphion	open-mmlab/Amphion	❌	15
Stable-Audio-Metrics	Stability-AI/stable-audio-metrics	✅	3
AudioLDM-Eval	haoheliu/audioldm_eval	✅	9
VERSA	wavlab-speech/versa	✅	

Next VERSA-V2 targets double (130)!

Compared to other open-source toolkits, VERSA is dedicated to speech and audio evaluation with **extensive integration** of both conventional and recent evaluation metrics.

(Check our paper for a full comparison!)

Links and Demos

Paper

GitHub

Colab Demo

Gradio Demo

https://github.com/wavlab-speech/versa.git

Uni-VERSA: Versatile Speech Assessment with a Unified Network

Jiatong Shi, Hye-jin Shim, Shinji Watanabe

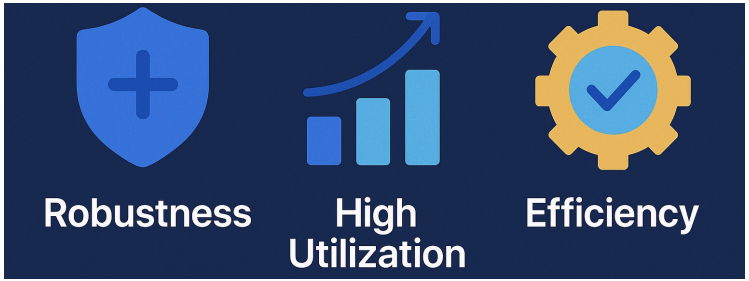
CMU

Further Step from VERSA

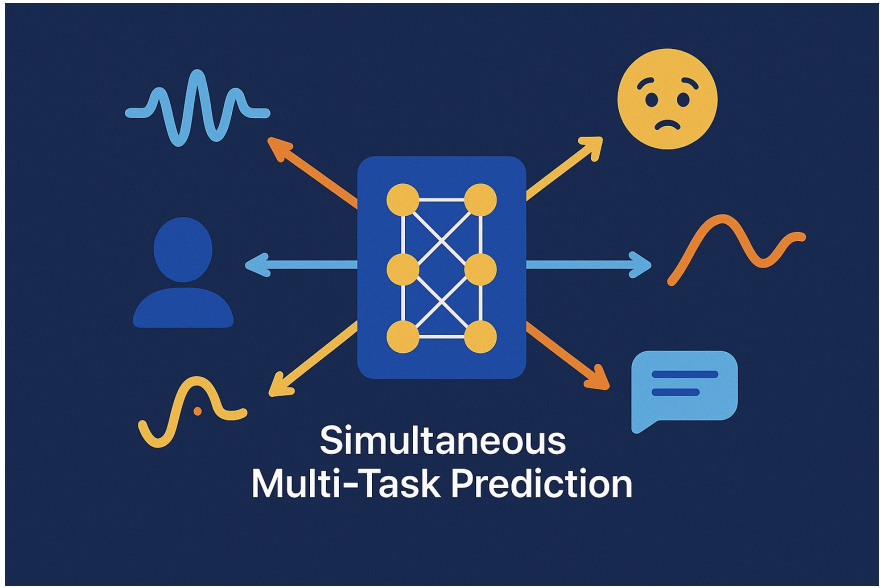
- Can we expand further from VERSA, with a single-unified model?

- What are the potential benefits?

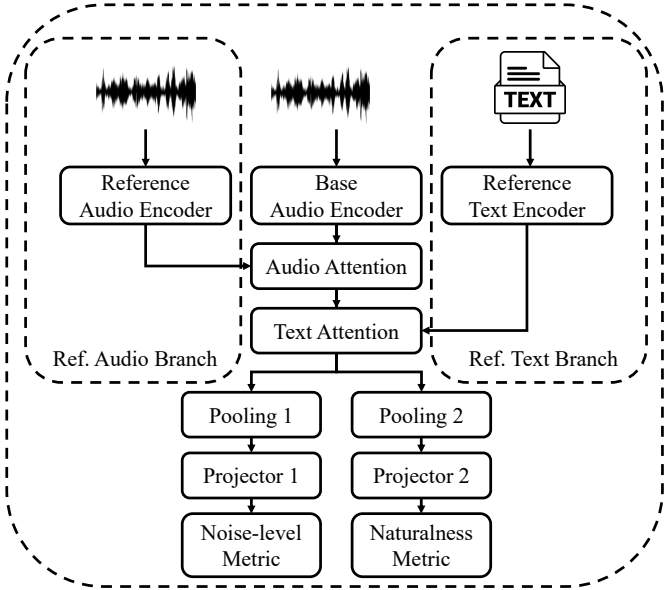
  - Enhanced robustness to various conditions
  - Better utilization of the information
  - Elevated efficiency of inference



General Concept of Uni-VERSA



The Development of Uni-VERSA



Conceptual Diagram of the Problem

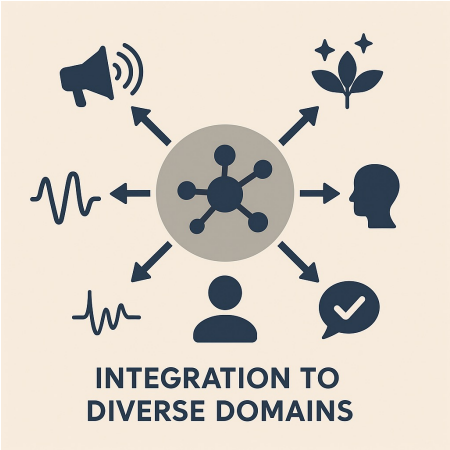
An example of predicting different aspects of speech assessment estimation **simultaneously**

Practical Challenges

- Integration to Diverse Domains
  - Wide range focuses from noise-level, naturalness, prosody information, speaker information, and intelligibility

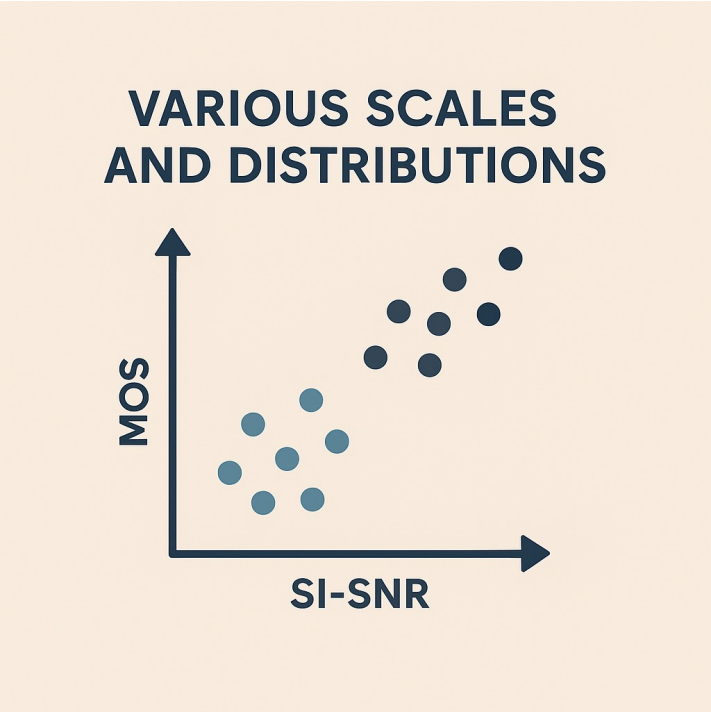
- Various Scales and Distributions

- Semi-supervised Learning



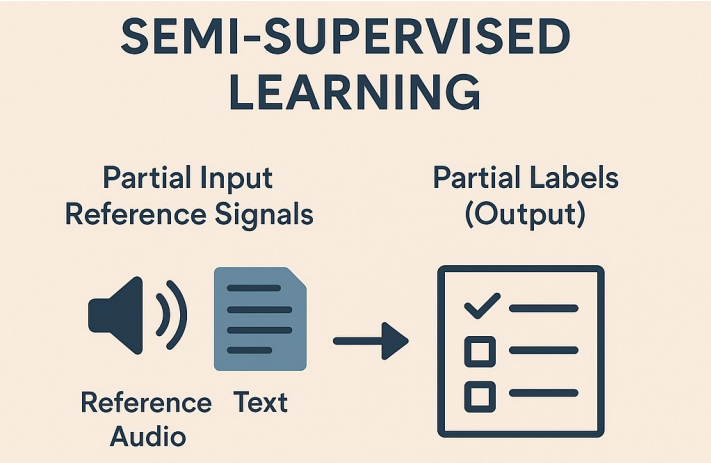
Practical Challenges

- Integration to Diverse Domains
- Various Scales and Distributions
  - SI-SNR vs. MOS
- Semi-supervised Learning



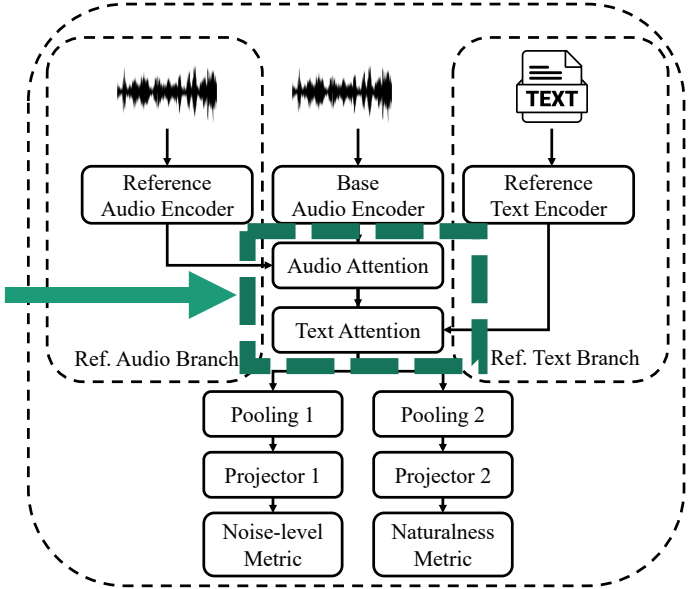
Practical Challenges

- Integration to Diverse Domains
- Various Scales and Distributions
- Semi-supervised Learning
  - Partial Reference Signals
    - Reference Audio, Reference Text
  - Partial Labels



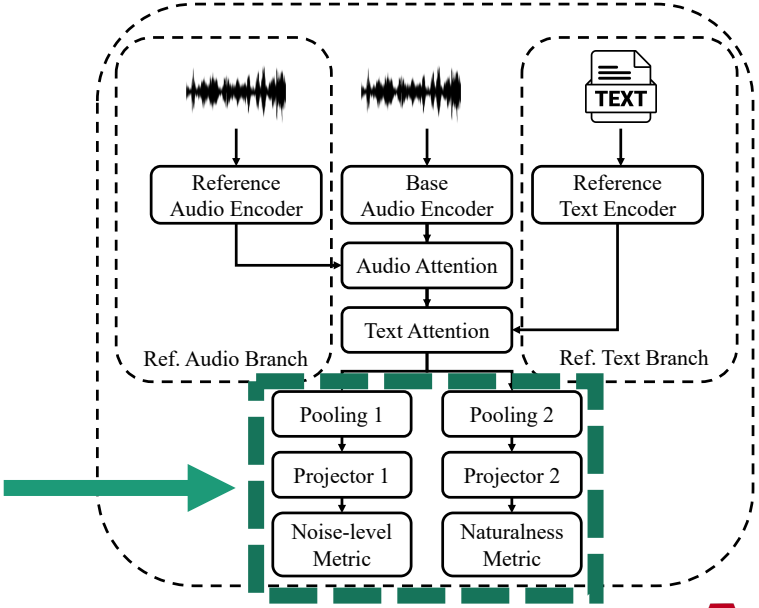
Our Initial Trial

- **Use cross-attention to add reference information**
- Use Individual predictor for each metric
- The simplest solution for the start



Our Initial Trial

- Use cross-attention to add reference information
- **Use Individual predictor for each metric**
- The simplest solution for the start



Data Curation

- A curation from URGENT challenge 2024
- Train-dev-test: 249-14-30 hrs
  - 50/50: data with / without reference
- 5 domains -> 11 metrics
  - Utilize expert models from VERSA for the prediction targets
- Evaluation:
  - Correlation coefficient (Pearson, Spearman)

Reference speech / reference text

Domain	Metric	Range	Reference Type
Noise level	SI-SNR	[-inf, inf]	Signal
	PESQ	[1, 4.5]	Signal
	DNSMOS	[1, 5]	/
Prosody	F0-CORR	[-1, 1]	Signal
Naturalness	MOS	[1, 5]	/
	UTMOS	[1, 5]	/
	SHEET-base	[1, 5]	/
Intelligibility	WER	[0, inf]	Text
	STOI	[0, 1]	Signal
	SBERT	[0, 1]	Signal
Speaker characteristics	SPK-SIM	[-1, 1]	Signal

Experimental Setup

- Main Investigation
  - The effect of audio encoder
  - The effect of reference signals
  - The effect of joint prediction

Experimental Results (Audio Encoder)

Pearson / Spearman Correlation Coefficient for Evaluation

Feature	Noise-Level			Prosody	Naturalness			Intelligibility			Speaker	Avg.
	SI-SNR	PESQ	DNSMOS	F0-CORR	MOS	UTMOS	SHEET-base	WER	STOI	SBERT	SPK-SIM	
Fbank	.69 / .73	.78 / .80	.90 / .89	.56 / .51	.72 / .72	.92 / .91	.85 / .85	.57 / .53	.79 / .74	.80 / .73	.75 / **.76**	.76 / .73
HuBERT	.84 / .83	.80 / .81	.90 / **.89**	.61 / .56	.78 / .77	.97 / .95	.93 / .93	.74 / .71	.92 / .89	.94 / .89	.76 / .71	.83 / .80
MR-HuBERT	.85 / .85	**.83** / **.85**	.90 / .88	**.65** / .60	**.78** / **.78**	**.97** / **.95**	**.94** / **.94**	.74 / .74	.91 / .89	.94 / .90	**.79** / .75	.84 / .82
WavLM	**.85** / **.85**	.82 / .84	**.90** / .88	.64 / **.61**	.77 / .77	.97 / .94	.93 / .94	**.79** / **.79**	**.93** / **.89**	**.95** / **.91**	.79 / .74	**.84** / **.82**

Four Encoder:

- Fbank: Mel-filterbank spectrogram
- HuBERT: Hidden unit BERT pre-trained audio encoder (Hsu et al. 2021)
- MR-HuBERT: Multi-resolution HuBERT (Shi et al. 2024)
- WavLM: Enhanced HuBERT with noise/multi-speaker augmentation (Chen et al. 2022)

Experimental Results (Audio Encoder)

Pearson / Spearman Correlation Coefficient for Evaluation

Feature	Noise-Level			Prosody	Naturalness			Intelligibility			Speaker	Avg.
	SI-SNR	PESQ	DNSMOS	F0-CORR	MOS	UTMOS	SHEET-base	WER	STOI	SBERT	SPK-SIM	
Fbank	.69 / .73	.78 / .80	.90 / .89	.56 / .51	.72 / .72	.92 / .91	.85 / .85	.57 / .53	.79 / .74	.80 / .73	.75 / **.76**	.76 / .73
HuBERT	.84 / .83	.80 / .81	.90 / **.89**	.61 / .56	.78 / .77	.97 / .95	.93 / .93	.74 / .71	.92 / .89	.94 / .89	.76 / .71	.83 / .80
MR-HuBERT	.85 / .85	**.83** / **.85**	.90 / .88	**.65** / .60	**.78** / **.78**	**.97** / **.95**	**.94** / **.94**	.74 / .74	.91 / .89	.94 / .90	**.79** / .75	.84 / .82
WavLM	**.85** / **.85**	.82 / .84	**.90** / .88	.64 / **.61**	.77 / .77	.97 / .94	.93 / .94	**.79** / **.79**	**.93** / **.89**	**.95** / **.91**	.79 / .74	**.84** / **.82**

Fou

-

-

-

-

)

hen et al. 2022)

The effect of audio encoder is **significant**

->

Fbank << HuBERT < MR-HuBERT ~ WavLM

Experimental Results (Reference Signal)

Pearson / Spearman Correlation Coefficient for Evaluation

Model	Noise-Level			Prosody	Naturalness			Intelligibility			Speaker	Avg.
	SI-SNR	PESQ	DNSMOS	F0-CORR	MOS	UTMOS	SHEET-base	WER	STOI	SBERT	SPK-SIM	
No Ref.	.84 / .84	**.85 / .86**	.90 / .89	.64 / .60	.77 / .78	.97 / .94	**.94 / .94**	.79 / .78	.90 / .86	.93 / .90	.79 / .75	.85 / .82
Audio Ref.	.84 / .82	.82 / .83	.90 / .89	**.66 / .62**	.78 / .78	**.97 / .95**	.94 / .94	.78 / .76	.92 / .89	.95 / .92	**.81** / .77	.85 / .82
Text Ref.	.84 / .83	.84 / .86	**.90 / .89**	.64 / .57	**.79 / .79**	.97 / .94	.93 / .93	.78 / .76	.92 / **.90**	**.96 / .93**	.80 / **.78**	**.85 / .83**
Full Ref.	**.85 / .85**	.82 / .84	.90 / .88	.64 / .61	.77 / .77	.97 / .94	.93 / .94	**.79 / .79**	**.93** / .89	.95 / .91	.79 / .74	.84 / .82

Four Models:

- No reference: model without reference
- Audio reference: model with reference audio
- Text reference: model with reference text
- Full reference: model with both reference signals

Experimental Results (Reference Signal)

Pearson / Spearman Correlation Coefficient for Evaluation

Model	Noise-Level			Prosody	Naturalness			Intelligibility			Speaker	Avg.
	SI-SNR	PESQ	DNSMOS	F0-CORR	MOS	UTMOS	SHEET-base	WER	STOI	SBERT	SPK-SIM	
No Ref.	.84 / .84	**.85 / .86**	.90 / .89	.64 / .60	.77 / .78	.97 / .94	**.94 / .94**	.79 / .78	.90 / .86	.93 / .90	.79 / .75	.85 / .82
Audio Ref.	.84 / .82	.82 / .83	.90 / .89	**.66 / .62**	.78 / .78	**.97 / .95**	.94 / .94	.78 / .76	.92 / .89	.95 / .92	**.81** / .77	.85 / .82
Text Ref.	.84 / .83	.84 / .86	**.90 / .89**	.64 / .57	**.79 / .79**	.97 / .94	.93 / .93	.78 / .76	.92 / **.90**	**.96 / .93**	.80 / **.78**	**.85 / .83**
Full Ref.	**.85 / .85**	.82 / .84	.90 / .88	.64 / .61	.77 / .77	.97 / .94	.93 / .94	**.79 / .79**	**.93** / .89	.95 / .91	.79 / .74	.84 / .82

Notable Findings:

- Models without certain references **sometimes outperform** those with more references
- Examples:
  - Model using reference text: Lower WER prediction correlations
  - Model without audio references: Best PESQ prediction results

Experimental Results (Reference Signal)

Pearson / Spearman Correlation Coefficient for Evaluation

Model	Noise-Level			Prosody	Naturalness			Intelligibility			Speaker	Avg.
	SI-SNR	PESQ	DNSMOS	F0-CORR	MOS	UTMOS	SHEET-base	WER	STOI	SBERT	SPK-SIM	
No Ref.	.84 / .84	**.85 / .86**	.90 / .89	.64 / .60	.77 / .78	.97 / .94	**.94 / .94**	.79 / .78	.90 / .86	.93 / .90	.79 / .75	.85 / .82
Audio Ref.	.84 / .82	.82 / .83	.90 / .89	**.66 / .62**	.78 / .78	**.97 / .95**	.94 / .94	.78 / .76	.92 / .89	.95 / .92	**.81** / .77	.85 / .82
Text Ref.	.84 / .83	.84 / .86	**.90 / .89**	.64 / .57	**.79 / .79**	.97 / .94	.93 / .93	.78 / .76	.92 / **.90**	**.96 / .93**	.80 / **.78**	**.85 / .83**
Full Ref.	**.85 / .85**	.82 / .84	.90 / .88	.64 / .61	.77 / .77	.97 / .94	.93 / .94	**.79 / .79**	**.93** / .89	.95 / .91	.79 / .74	.84 / .82

N

-

-

th more references

Potential cause: Uni-VERSA model **overfitting** when given redundant reference signals

Further verifications: the PESQ for training set with audio references are much better than those without.

Experimental Results (Joint Prediction)

Pearson / Spearman Correlation Coefficient for MOS Evaluation

Feature	Pred. Type	LCC	SRCC
Fbank	Single	.53	.55
	Joint	.72	.72
WavLM	Single	**.80**	**.80**
	Joint	.77	.77

With weak features (Fbank), joint prediction show benefits

With WavLM, joint prediction may not be helpful to performance

Do we still need multi-metric prediction?

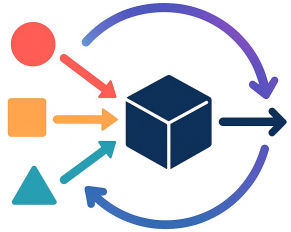
Benefits from Uni-VERSA

- Efficiency
  - Using Uni-VERSA can enjoy up to **109x** speed up (same batch size)
    - With only the 11 metrics included
    - Why?

Benefits from Uni-VERSA

- Efficiency
  - Using Uni-VERSA can enjoy up to **109x** speed up
    - With only the 11 metrics included
    - Why?
      - Full GPU inference
      - Reduced shared forward from audio encoder representation
      - Parallel prediction

Benefits from Uni-VERSA



- Multi-metric model as a loss term for generation tasks
  - Convert non-differentiable metrics into backward-friendly objectives
  - Enjoy efficient loss computation (multi-metric with a single model)
  - Reduce the need of real-world data
- E.g., Improving Speech Enhancement with Multi-Metric Supervision from Learned Quality Assessment (Wei et al. 2025)

Links and Demos

Paper

Colab Demo

Huggingface Pre-trained Models

Code

ARECHO: Autoregressive Evaluation via Chain-Based Hypothesis Optimization for Speech Multi-Metric Estimation

Jiatong Shi[1], Yifan Cheng[2], Bo-hao Su[1], Hye-jin Shim[1], Jinchuan Tian[1],

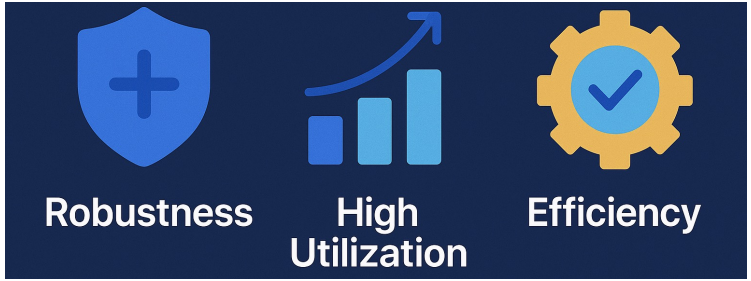
Samuele Cornell[1], Yiwen Zhao[1], Siddhant Arora[1], Shinji Watanabe[1]

[1]CMU, [2]HUST

07/17/2025

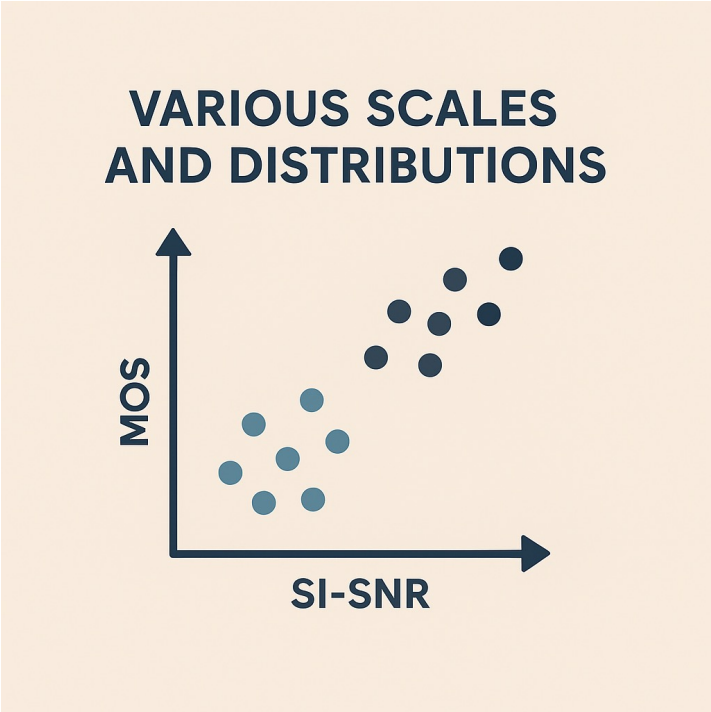
Successes in Uni-VERSA

- Robustness to various conditions
- Better utilization of the information
- Elevated efficiency of inference



However, challenges remains

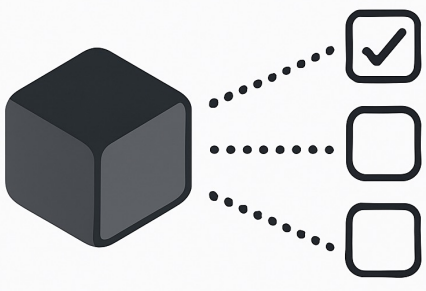
- **Diverse Scale Issues**
- Limited Data Availability
- Dependency Modeling with Flexible Control



However, challenges remains

- Diverse Scale Issues

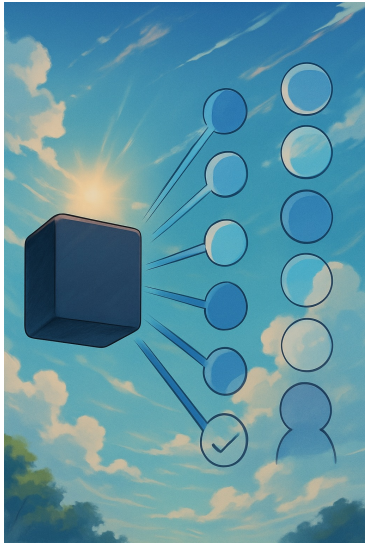
- **Limited Data Availability**
  - **In Uni-VERSA, we use a metric-balanced set**
  - **Reality is not the case though**



- Dependency Modeling with Flexible Control

However, challenges remains

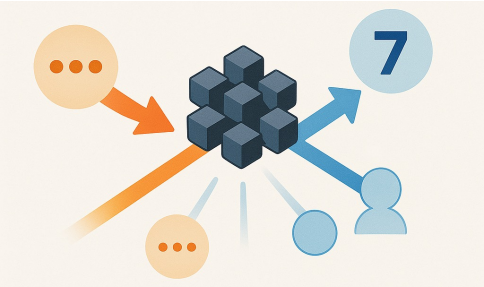
- Diverse Scale Issues
- Limited Data Availability
- **Dependency Modeling with Flexible Control**
  - **Difficulty in leveraging dependency benefits**
  - **Inefficient modeling with incomplete labels**



ARECHO: Autoregressive Evaluation

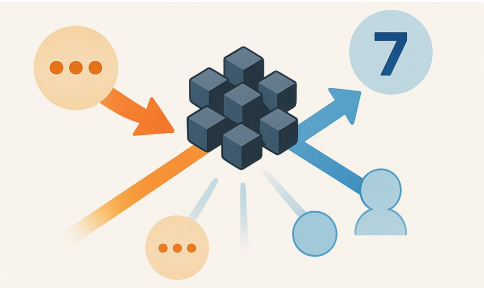
- Tokenizing Everything
- Dynamic Classifier Chain
- Two-step Confidence-oriented Decoding

Everything in Tokenization



- Categorical metrics -> direct tokenization
- Numerical metrics -> quantile-based partitioning
  - Flexibly control the distribution of token bins with different distribution percentiles

Everything in Tokenization

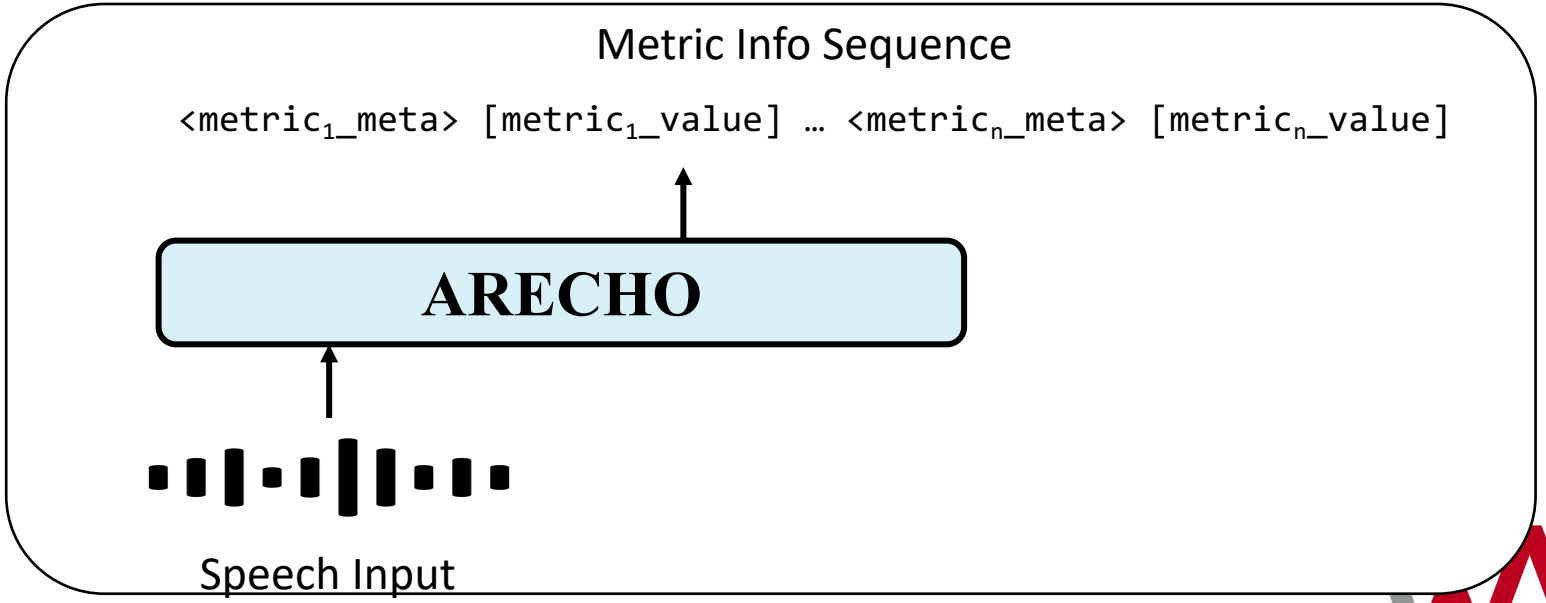


- Categorical metrics -> direct tokenization
- Numerical metrics -> quantile-based partitioning
  - Flexibly control the distribution of token bins with different distribution percentiles

A direct fix to:
- Diverse metric scales / types

Dynamic Classifier Chain

- Main concept is as simple as an autoregressive network:



Metric Sequence

- Motivating Example
  - <Gender> [Male] <Emotion> [Happy] <MOS> [3.78]

- During Training, we randomly shuffle the metric order
  - <Gender> [Male] <Emotion> [Happy] <MOS> [3.78]
  - <Emotion> [Happy] <Gender> [Male] <MOS> [3.78]
  - <MOS> [3.78] <Gender> [Male] <Emotion> [Happy]
  - …

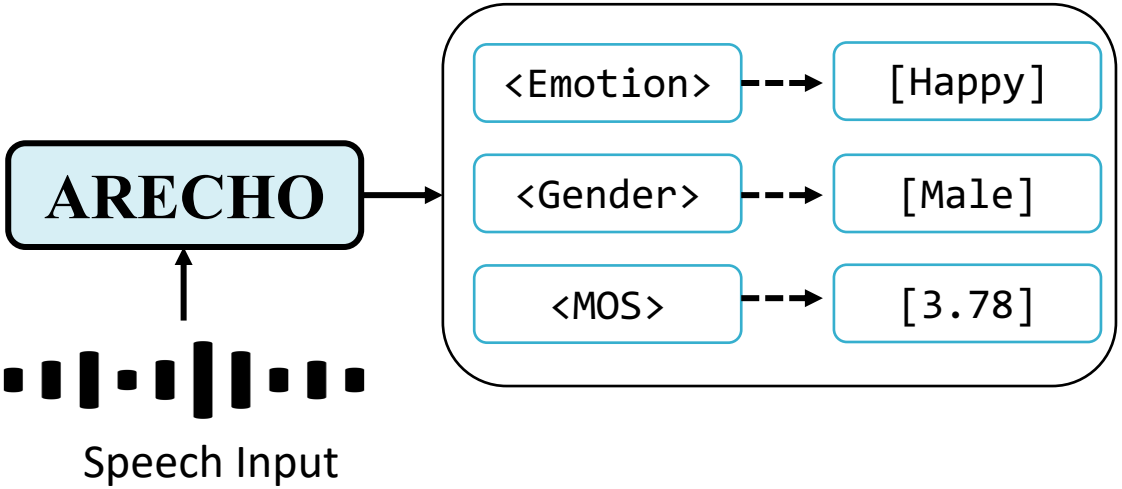
Metric Sequence

Key features:

- **Dependency control:** prior metric predictions are part of the sequence history
- **Metric identification:** Metadata token specify which metric to predict, even when value formats overlap
- **Flexible Inference:** beginning with a metadata token, generate the corresponding value token -> arbitrary queries enable dynamic/chain-style reasoning
- **Direct fix to limited labels:** simply omits the unavailable labels (no need for masking / imputation)

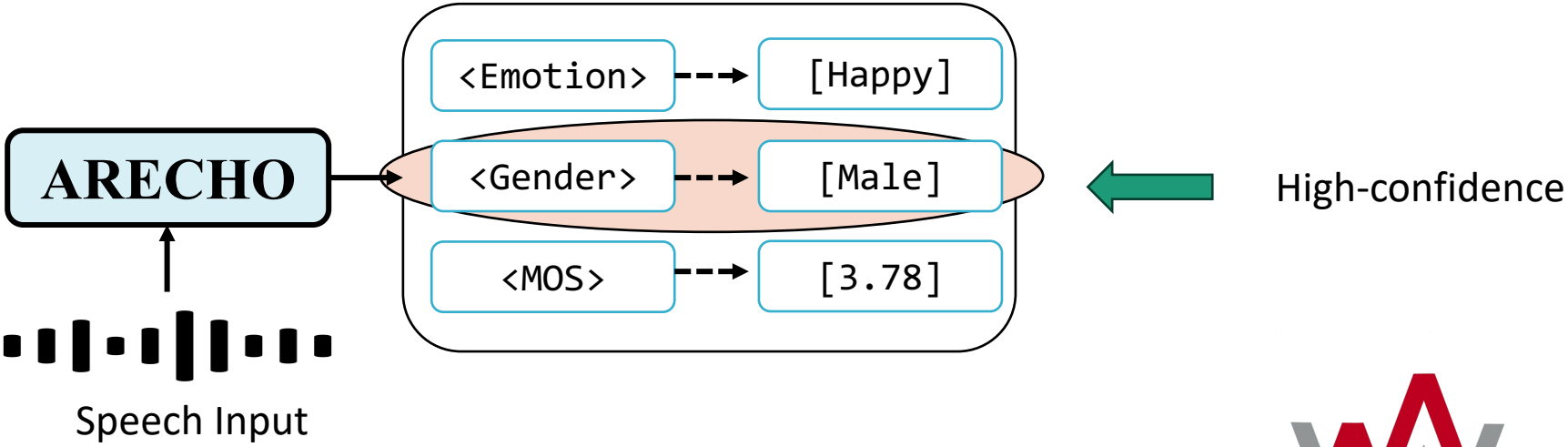
Two-step Confidence-oriented Decoding

- Make decoding decision every two steps



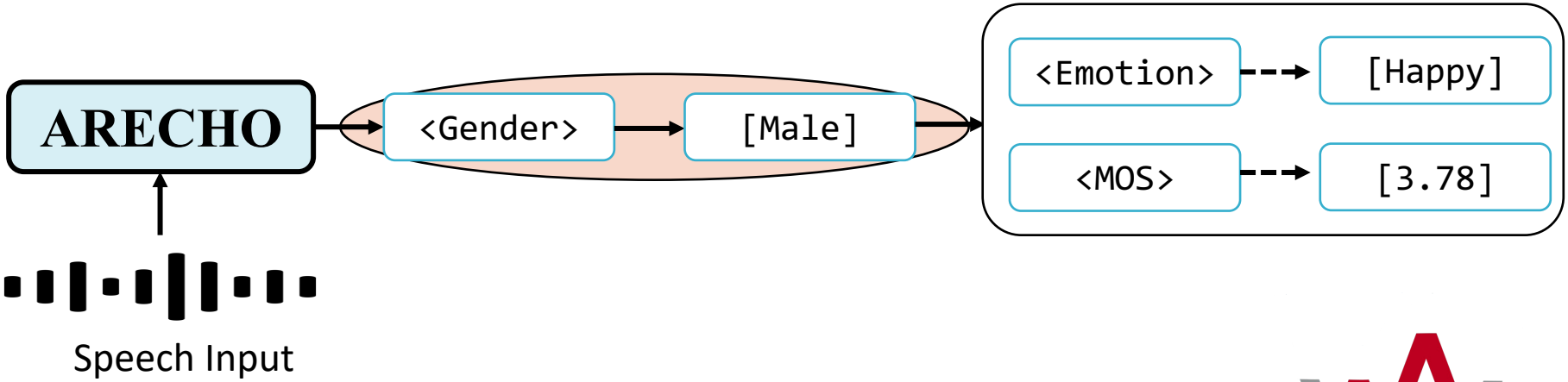
Two-step Confidence-oriented Decoding

- Make decoding decision every two steps



Two-step Confidence-oriented Decoding

- Make decoding decision every two steps



Experiments

- Dataset combination
  - Basic speech data -> sampled from OWSM (Peng et al. 2023)
  - Corrupted speech data (simulation) -> simulated based on URGENT challenge (Zhang et al. 2024)
  - Enhanced speech data -> samples from the URGENT challenge submission (Zhang et al. 2024)
  - Synthesized speech data -> synthesized data from NISQA and VoiceMOS (Mittag et al. 2021; Huang et al. 2022)

Experiments

- Dataset combination
  - Basic speech data -> sampled from OWSM (Peng et al. 2023)
  - ge
  - (
  - Synthesized speech data -> synthesized data from NISQA and VoiceMOS (Mittag et al. 2021; Huang et al. 2022)

Two combination:
- Base (a 300h domain-balanced training set)
- **Scale (a 2kh larger-scale training set)**

Covered metrics in ARECHO

- 47 independent metrics
- 25 dependent metrics
- 7 non-matching metrics
- 8 ground truth metrics from data annotations

- → Sum up to 87 metrics (65 numerical + 22 categorical)

Main Results

MSE: Mean square error;
LCC: Linear correlation coefficient
KTAU: Ktau ranking-based correlation
Acc: Accuracy
F1: F1 measure

Data	Domain	Model	Token	Chain	Regression Metrics MSE (↓)	LCC (↑)	KTAU (↑)	Classification Metrics Acc (↑)	F1 (↑)
Scale	Dev.	UniVERSA	✗	✗	116.01	0.89	0.37	0.73	0.49
		UniVERSA-T	✓	✗	**27.98**	**0.86**	0.37	0.74	**0.52**
		ARECHO	✓	✓	29.61	0.85	**0.37**	**0.75**	**0.52**
	Enhanced	UniVERSA	✗	✗	43.05	**0.84**	0.34	0.72	0.47
		UniVERSA-T	✓	✗	69.94	0.80	0.36	0.74	0.50
		ARECHO	✓	✓	**32.63**	0.83	**0.37**	**0.75**	**0.53**
	Corrupted	UniVERSA	✗	✗	151.97	**0.88**	0.37	0.75	0.54
		UniVERSA-T	✓	✗	39.80	0.77	0.37	0.76	0.54
		ARECHO	✓	✓	**34.37**	0.84	**0.38**	**0.77**	**0.56**
	Synthesized	UniVERSA	✗	✗	**6.46**	0.84	0.32	0.71	0.47
		UniVERSA-T	✓	✗	8.23	0.84	0.34	0.73	0.49
		ARECHO	✓	✓	8.63	**0.85**	**0.36**	**0.75**	**0.54**
	Avg. Test	UniVERSA	✗	✗	67.16	**0.85**	0.35	0.73	0.50
		UniVERSA-T	✓	✗	39.32	0.80	0.36	0.74	0.51
		ARECHO	✓	✓	**25.21**	0.84	**0.37**	**0.76**	**0.54**

Effect of tokenization
-> Consistent improvements on classification metrics due to mitigated scale differences

				Chain	Regression Metrics			Classification Metrics	
					MSE (↓)	LCC (↑)	KTAU (↑)	Acc (↑)	F1 (↑)
				✗	116.01	0.89	0.37	0.73	0.49
				✗	**27.98**	**0.86**	0.37	0.74	**0.52**
				✓	29.61	0.85	**0.37**	**0.75**	**0.52**
	Enhanced	UniVERSA	✗	✗	43.05	**0.84**	0.34	0.72	0.47
		UniVERSA-T	✓	✗	69.94	0.80	0.36	0.74	0.50
		ARECHO	✓	✓	**32.63**	0.83	**0.37**	**0.75**	**0.53**
Scale	Corrupted	UniVERSA	✗	✗	151.97	**0.88**	0.37	0.75	0.54
		UniVERSA-T	✓	✗	39.80	0.77	0.37	0.76	0.54
		ARECHO	✓	✓	**34.37**	0.84	**0.38**	**0.77**	**0.56**
	Synthesized	UniVERSA	✗	✗	**6.46**	0.84	0.32	0.71	0.47
		UniVERSA-T	✓	✗	8.23	0.84	0.34	0.73	0.49
		ARECHO	✓	✓	8.63	**0.85**	**0.36**	**0.75**	**0.54**
	Avg. Test	UniVERSA	✗	✗	67.16	**0.85**	0.35	0.73	0.50
		UniVERSA-T	✓	✗	39.32	0.80	0.36	0.74	0.51
		ARECHO	✓	✓	**25.21**	0.84	**0.37**	**0.76**	**0.54**

Main Results

Effect of AR Modeling
-> Consistent improvements with AR modeling

Data	Domain	Model	Token	Chain	Regression Metrics MSE (↓)	LCC (↑)	KTAU (↑)	Classification Metrics Acc (↑)	F1 (↑)
Scale	Dev.	UniVERSA	✗	✗	116.01	0.89	0.37	0.73	0.49
		UniVERSA-T	✓	✗	**27.98**	**0.86**	0.37	0.74	**0.52**
		ARECHO	✓	✓	29.61	0.85	**0.37**	**0.75**	**0.52**
	Enhanced	UniVERSA	✗	✗	43.05	**0.84**	0.34	0.72	0.47
		UniVERSA-T	✓	✗	69.94	0.80	0.36	0.74	0.50
		ARECHO	✓	✓	**32.63**	0.83	**0.37**	**0.75**	**0.53**
	Corrupted	UniVERSA	✗	✗	151.97	**0.88**	0.37	0.75	0.54
		UniVERSA-T	✓	✗	39.80	0.77	0.37	0.76	0.54
		ARECHO	✓	✓	**34.37**	0.84	**0.38**	**0.77**	**0.56**
	Synthesized	UniVERSA	✗	✗	**6.46**	0.84	0.32	0.71	0.47
		UniVERSA-T	✓	✗	8.23	0.84	0.34	0.73	0.49
		ARECHO	✓	✓	8.63	**0.85**	**0.36**	**0.75**	**0.54**
	Avg. Test	UniVERSA	✗	✗	67.16	**0.85**	0.35	0.73	0.50
		UniVERSA-T	✓	✗	39.32	0.80	0.36	0.74	0.51
		ARECHO	✓	✓	**25.21**	0.84	**0.37**	**0.76**	**0.54**

Main Results

Metric-wise fine-grained control may help

Data	Domain	Model	Token	Chain	Regression Metrics MSE (↓)	LCC (↑)	KTAU (↑)	Classification Metrics Acc (↑)	F1 (↑)
Scale	Dev.	UniVERSA	✗	✗	116.01	0.89	0.37	0.73	0.49
		UniVERSA-T	✓	✗	**27.98**	**0.86**	0.37	0.74	**0.52**
		ARECHO	✓	✓	29.61	0.85	**0.37**	**0.75**	**0.52**
	Enhanced	UniVERSA	✗	✗	43.05	**0.84**	0.34	0.72	0.47
		UniVERSA-T	✓	✗	69.94	0.80	0.36	0.74	0.50
		ARECHO	✓	✓	**32.63**	0.83	**0.37**	**0.75**	**0.53**
	Corrupted	UniVERSA	✗	✗	151.97	**0.88**	0.37	0.75	0.54
		UniVERSA-T	✓	✗	39.80	0.77	0.37	0.76	0.54
		ARECHO	✓	✓	**34.37**	0.84	**0.38**	**0.77**	**0.56**
	Synthesized	UniVERSA	✗	✗	**6.46**	0.84	0.32	0.71	0.47
		UniVERSA-T	✓	✗	8.23	0.84	0.34	0.73	0.49
		ARECHO	✓	✓	8.63	**0.85**	**0.36**	**0.75**	**0.54**
	Avg. Test	UniVERSA	✗	✗	67.16	**0.85**	0.35	0.73	0.50
		UniVERSA-T	✓	✗	39.32	0.80	0.36	0.74	0.51
		ARECHO	✓	✓	**25.21**	0.84	**0.37**	**0.76**	**0.54**

Dependency Modeling

Top 5 – Bottom 5 metrics ranked by average position (Avg. Pos.) across three test sets

An option to show **dependency reasoning** in ARECHO

Test Set	Rank	Metric Name	Avg. Pos.
Enhanced	Top-1	Q-SpeakerGender	16.50
	Top-2	Q-SpeechImpairment	20.35
	Top-3	Q-SpeechStyle	21.47
	Btm-3	SNR Simulation	163.52
	Btm-2	NISQA Real MOS	167.91
	Btm-1	VoiceMOS Real MOS	171.58
Corrupted	Top-1	RIR Room Size	1.82
	Top-2	Q-SpeechImpairment	12.65
	Top-3	Q-SpeechDelivery	13.15
	Btm-3	CER	167.26
	Btm-2	NISQA Real MOS	170.62
	Btm-1	VoiceMOS Real MOS	171.38
Synthesized	Top-1	Q-Background	12.09
	Top-2	NISQA Coloration	27.51
	Top-3	Q-Purpose	27.95
	Btm-3	C_{bak}	154.75
	Btm-2	SNR Simulation	158.64
	Btm-1	CER	161.61

Dependency Modeling

Metrics related to

- Structured annotations
- Acoustic scene characteristics

Appear early in prediction sequence

Test Set	Rank	Metric Name	Avg. Pos.
Enhanced	Top-1	Q-SpeakerGender	16.50
	Top-2	Q-SpeechImpairment	20.35
	Top-3	Q-SpeechStyle	21.47
	Btm-3	SNR Simulation	163.52
	Btm-2	NISQA Real MOS	167.91
	Btm-1	VoiceMOS Real MOS	171.58
Corrupted	Top-1	RIR Room Size	1.82
	Top-2	Q-SpeechImpairment	12.65
	Top-3	Q-SpeechDelivery	13.15
	Btm-3	CER	167.26
	Btm-2	NISQA Real MOS	170.62
	Btm-1	VoiceMOS Real MOS	171.38
Synthesized	Top-1	Q-Background	12.09
	Top-2	NISQA Coloration	27.51
	Top-3	Q-Purpose	27.95
	Btm-3	C_{bak}	154.75
	Btm-2	SNR Simulation	158.64
	Btm-1	CER	161.61

Dependency Modeling

Depending on the context, some metrics are **inherently harder** to predict (e.g., SNR simulation for enhanced and synthesized test sets)

Test Set	Rank	Metric Name	Avg. Pos.
Enhanced	Top-1	Q-SpeakerGender	16.50
	Top-2	Q-SpeechImpairment	20.35
	Top-3	Q-SpeechStyle	21.47
	Btm-3	SNR Simulation	163.52
	Btm-2	NISQA Real MOS	167.91
	Btm-1	VoiceMOS Real MOS	171.58
Corrupted	Top-1	RIR Room Size	1.82
	Top-2	Q-SpeechImpairment	12.65
	Top-3	Q-SpeechDelivery	13.15
	Btm-3	CER	167.26
	Btm-2	NISQA Real MOS	170.62
	Btm-1	VoiceMOS Real MOS	171.38
Synthesized	Top-1	Q-Background	12.09
	Top-2	NISQA Coloration	27.51
	Top-3	Q-Purpose	27.95
	Btm-3	C_{bak}	154.75
	Btm-2	SNR Simulation	158.64
	Btm-1	CER	161.61

Dependency Modeling

Q-Gender	Q-SpeechImpariment	Q-SpeakingStyle	Q-EnvQuality	Q-PitchRange	Q-VocComplexity	Q-VolumeLevel	RealLanguage	Q-ContentRegister	SRMR
SpoofS	NISQA-NOI	Q-Emotion	AA-PC	Q-Background	AA-PQ	Q-ChannelType	LID	Q-Clarity	SE-CI-SDR
DNSMOSP.835	SWR/SCR	Q-Purpose	WER	Q-VoiceType	SingMOS	SE-SI-SNR	Q-Lang	Q-SpeechRate	SCOREQ
NISQA-COL	Q-EmoVocalization	NISQA-LOUD	Q-SpeakerCount	Q-Age	PAM	UTMOS	AA-CE	NISQA-MOS	DNSMOSP.808
SSQA	PLCMOS	Q-Pitch	AA-CU	CER	SE-SDR	UTMOSv2	NISQA-DIS	CI-SDR	D-Distance
STOI	SE-SAR	SDR	SI-SNR	MCD	D-BERT	F0Corr	F0RMSE	SPK-SIM	D-BLEU
PESQ	URGENT MOS	SAR	CD	WSS	LLR	EMO-SIM	NCM	Covl	Reference Text Length
VISQOL	Csig	CSII-MID	Cbak	CSII-HIGH	SCOREQ w. Ref.	ASR-Mismatch	CSII-LOW	NOMAD	RIR Room Size
Noresqa	FWSEGSNR	RT60	Predicted Text Length	SNR Simulation	NISQA Real MOS	VoiceMOS Real MOS			

Check full metric order in the paper!

Future works

- Scaling!
- ARECHO as a reward model
- ARECHO + Uni-VERSA
- From low-level information to high-level understanding
- From fixed metric set to natural language
    - How to combine the information
    - How to activate the usage

Acknowledgements

- Thanks to all the collaborators to these projects, including

  Jinchuan Tian (CMU), Yifan Cheng (HUST), Bo-Hao Su (CMU), Samule Cornell (CMU), Yihan Wu (RUC), Jia Qi Yip (NTU - Singapore), You Zhang (Dolby), Wangyou Zhang (SJTU), Darius Petermann (IU), Yichen Huang (CMU), William Chen (CMU), Yuning Wu (RUC), Yuxun Tang (RUC), Dareen Alharhi (CMU), Yiwen Zhao (CMU), Jionghao Han (CMU), Wenhao Feng (CMU), Tejes Srivastava (Uchicago), Haibin Wu (Microsoft), Hye-Jin Shim (CMU), Chris Donahue (CMU), Qin Jin (RUC), Shinji Watanabe (CMU)

Icons and images are either from flaticon.com or generated from GPT-4o.

Thank you for listening!

TABLE V

COMPARISON OF SUPERVISION STRATEGIES USING DIFFERENT LOSS CONFIGURATIONS ON DNS20 TEST SETS, REPORTED AS (SYNTHETIC / SYNTHETIC REVERB). LPS DENOTES LEVENSHTEIN PHONEME SIMILARITY. RANK DENOTES THE OVERALL RANKING SCORE NORMALIZED TO [0, 1]. BEST RESULTS ARE IN **BOLD**, BASED ON FULL-PRECISION VALUES.

#	SQA Model	$\lambda_{spec}/\lambda_{score}/\lambda_{feat}$	SDR ↑	PESQ ↑	LSD ↓	MCD ↓	LPS ↓	ESTOI ↑	SBERT ↑	SpkSim ↑
1	Noisy		9.09 / 9.16	1.58 / 1.82	3.13 / 2.05	5.63 / 5.02	0.90 / 0.67	0.81 / **0.78**	0.88 / 0.86	**0.95** / **0.88**
2	None	1 / 0 / 0	18.21 / 10.97	2.86 / 1.99	2.49 / 2.37	2.03 / 3.99	0.96 / 0.83	0.93 / 0.77	0.93 / 0.88	0.93 / 0.84
3	Uni-VERSA	1 / 1 / 1	18.29 / 10.75	2.91 / 1.88	2.45 / 2.33	**1.98** / 3.99	0.96 / 0.83	0.93 / 0.76	0.93 / 0.87	0.93 / 0.83
4	Uni-VERSA-Ext (proposed)	1 / 1 / 1	18.27 / 10.72	2.91 / 1.90	2.41 / 2.32	1.99 / 3.94	0.96 / 0.83	0.93 / 0.77	0.93 / 0.88	0.93 / 0.83
5		1 / 1 / 0	**18.30** / 10.74	**2.91** / 1.88	2.48 / 2.35	1.99 / 3.99	0.96 / **0.83**	0.93 / 0.76	0.93 / 0.87	0.93 / 0.83
6		1 / 0 / 1	18.25 / 10.80	2.91 / 1.91	**2.41** / **2.31**	1.99 / **3.91**	0.96 / 0.83	0.93 / 0.77	0.93 / 0.88	0.93 / 0.84
7		0 / 1 / 1	17.93 / **11.15**	2.90 / **1.95**	2.50 / 2.46	2.20 / 3.98	**0.96** / 0.83	**0.93** / 0.77	**0.93** / **0.88**	0.92 / 0.82
8		0 / 1 / 0	1.56 / -1.43	1.04 / 1.03	5.71 / 5.84	7.64 / 9.51	0.29 / 0.15	0.33 / 0.18	0.59 / 0.52	0.19 / 0.08
9		0 / 0 / 1	17.94 / 11.12	2.90 / 1.95	2.49 / 2.46	2.19 / 3.99	0.96 / 0.83	0.93 / 0.77	0.93 / 0.88	0.92 / 0.83
		(continued)	CER ↓	UTMOS ↑	SIGMOS.OVRL ↑	SCOREQ ↑	DNSMOS ↑	Distill-MOS ↑	NISQA ↑	**Rank [22]** ↓
1	Noisy		**0.25** / 0.26	2.37 / 1.30	2.32 / 2.07	2.79 / 2.10	2.48 / 1.39	3.06 / 2.32	2.58 / 1.66	0.60 / 0.48
2	None	1 / 0 / 0	0.34 / 0.27	3.60 / 1.38	3.06 / 2.42	3.98 / 2.53	3.25 / 2.59	4.07 / 2.74	4.25 / 2.36	0.42 / 0.40
3	Uni-VERSA	1 / 1 / 1	0.32 / 0.26	3.61 / 1.38	3.11 / 2.46	4.03 / 2.51	3.24 / 2.52	4.09 / 2.65	4.26 / 2.34	0.35 / 0.48
4	Uni-VERSA-Ext (proposed)	1 / 1 / 1	0.31 / 0.25	3.69 / 1.39	3.00 / 2.50	4.09 / 2.69	3.28 / 2.61	4.29 / 2.67	4.49 / 2.72	**0.33** / 0.42
5		1 / 1 / 0	0.33 / 0.25	3.61 / 1.38	3.12 / 2.47	4.03 / 2.50	3.25 / 2.52	4.08 / 2.62	4.29 / 2.36	0.38 / 0.51
6		1 / 0 / 1	0.32 / 0.22	3.69 / 1.39	3.01 / 2.52	4.09 / 2.69	3.26 / 2.58	4.22 / 2.64	4.47 / 2.71	0.34 / 0.38
7		0 / 1 / 1	0.31 / 0.24	3.84 / 1.42	**3.14** / 2.58	**4.21** / **2.85**	**3.33** / **2.77**	**4.46** / **2.76**	**4.59** / **2.94**	0.39 / **0.38**
8		0 / 1 / 0	0.77 / 1.05	2.34 / **1.64**	2.47 / 2.17	3.07 / 2.66	2.75 / 2.35	2.77 / 2.24	2.75 / 1.93	0.77 / 0.76
9		0 / 0 / 1	0.32 / **0.23**	**3.84** / 1.42	3.13 / **2.59**	4.21 / 2.84	3.33 / 2.76	4.41 / 2.69	4.56 / 2.92	0.39 / 0.38