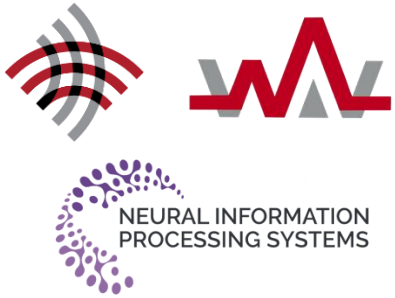


ARECHO: Autoregressive Evaluation via Chain-Based Hypothesis Optimization for Speech Multi-Metric Estimation

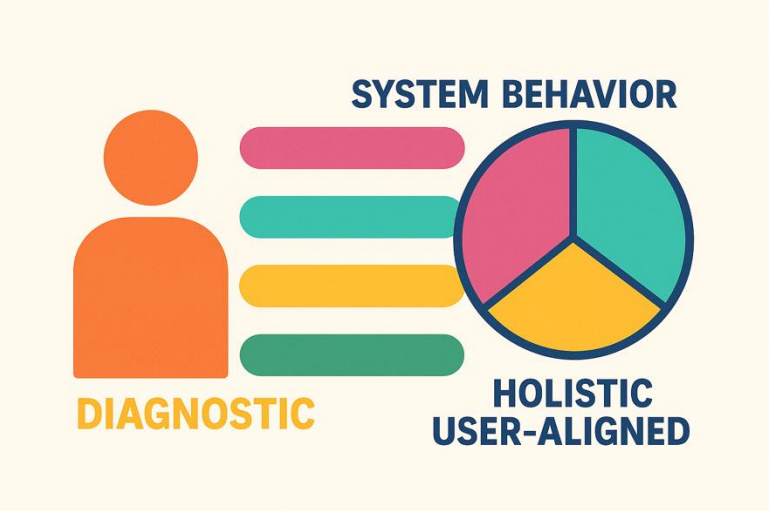
Jiatong Shi[1], Yifan Cheng[2], Bo-hao Su[1], Hye-jin Shim[1], Jinchuan Tian[1],

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

[1]CMU, [2]HUST

Motivation

- Multi-metric mindset towards
  - A diagnostic, holistic, and user-aligned view of speech generation system behavior



Introducing VERSA

VERSA: A Versatile Evaluation Toolkit for Speech, Audio, and Music

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

[1]Carnegie Mellon University, [2]Microsoft, [3]Indiana University, [4]Nanyang Technological University, [5]University of Rochester, [6]Renmin University of China, [7]Shanghai Jiaotong University, [8]Sony AI

- 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

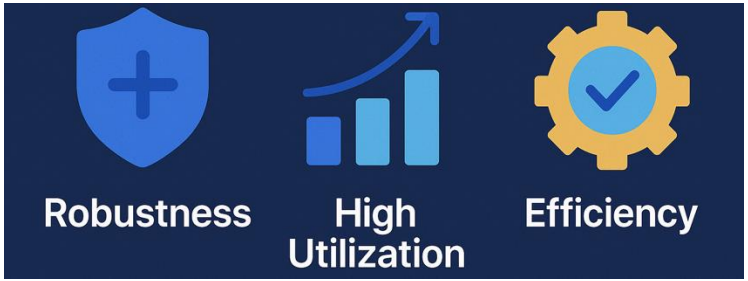
ESPnet Muskit URGENT CHALLENGE SINGING VOICE CONVERSION CHALLENGE

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



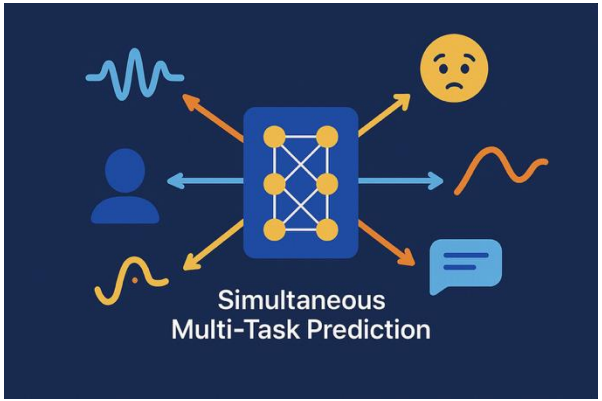
General Concept of Uni-VERSA

Uni-VERSA: Versatile Speech Assessment with a Unified Network

Jiatong Shi[1], *Hye-Jin Shim*[1], *Shinji Watanabe*[1]

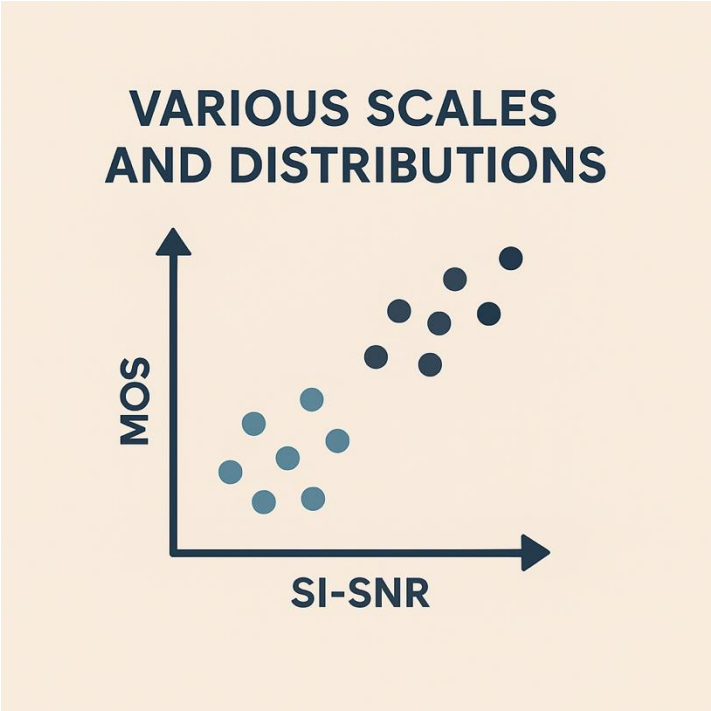
[1]Language Technologies Institute, Carnegie Mellon University, U.S.A.

jiatongs@cs.cmu.edu



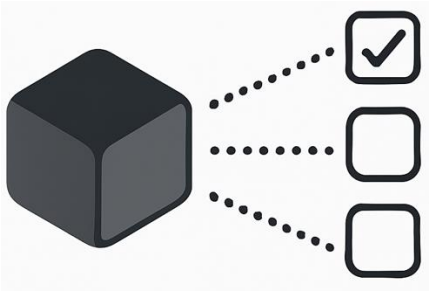
However, challenges remains

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



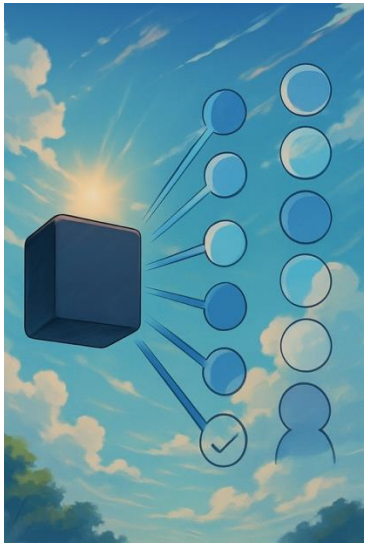
However, challenges remains

- Diverse Scale Issues
- **Limited Data Availability**
  - **We can not have a full set of metrics for all the data**
- 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 metric labels**



ARECHO: Autoregressive Evaluation

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

Experiments

- Dataset combination (87 metrics included)

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

Training:
- Uni-VERSA (baseline);
- Uni-VERSA-T (baseline + tokenization);
- ARECHO (proposed)

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			Classification Metrics	
					MSE (↓)	LCC (↑)	KTAU (↑)	Acc (↑)	F1 (↑)
Base	Dev.	UniVERSA	✗	✗	160.06	0.69	0.53	0.68	0.42
		UniVERSA-T	✓	✗	40.95	0.78	0.68	0.70	0.46
		ARECHO	✓	✓	**25.73**	**0.86**	**0.72**	**0.71**	**0.51**
	Enhanced	UniVERSA	✗	✗	61.54	0.71	0.54	0.69	0.43
		UniVERSA-T	✓	✗	27.34	0.81	0.68	0.70	0.47
		ARECHO	✓	✓	**20.58**	**0.84**	**0.69**	**0.72**	**0.51**
	Corrupted	UniVERSA	✗	✗	170.65	0.61	0.48	0.70	0.46
		UniVERSA-T	✓	✗	77.72	0.74	0.67	0.71	0.50
		ARECHO	✓	✓	**44.22**	**0.82**	**0.70**	**0.72**	**0.55**
	Synthesized	UniVERSA	✗	✗	58.79	0.76	0.54	0.69	0.45
		UniVERSA-T	✓	✗	8.10	0.84	0.68	0.72	0.50
		ARECHO	✓	✓	**4.99**	**0.91**	**0.78**	**0.79**	**0.65**
	Avg. Test	UniVERSA	✗	✗	96.99	0.69	0.52	0.69	0.45
		UniVERSA-T	✓	✗	37.72	0.79	0.68	0.71	0.49
		ARECHO	✓	✓	**23.26**	**0.86**	**0.72**	**0.74**	**0.57**

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

			en	Chain	Regression Metrics MSE (↓)	LCC (↑)	KTAU (↑)	Classification Metrics Acc (↑)	F1 (↑)
Base	Enhanced	UniVERSA	✗	✗	160.06	0.69	0.53	0.68	0.42
		UniVERSA-T	✓	✗	40.95	0.78	0.68	0.70	0.46
		ARECHO	✓	✓	**25.73**	**0.86**	**0.72**	**0.71**	**0.51**
	Corrupted	UniVERSA	✗	✗	61.54	0.71	0.54	0.69	0.43
		UniVERSA-T	✓	✗	27.34	0.81	0.68	0.70	0.47
		ARECHO	✓	✓	**20.58**	**0.84**	**0.69**	**0.72**	**0.51**
	Synthesized	UniVERSA	✗	✗	170.65	0.61	0.48	0.70	0.46
		UniVERSA-T	✓	✗	77.72	0.74	0.67	0.71	0.50
		ARECHO	✓	✓	**44.22**	**0.82**	**0.70**	**0.72**	**0.55**
	Avg. Test	UniVERSA	✗	✗	58.79	0.76	0.54	0.69	0.45
		UniVERSA-T	✓	✗	8.10	0.84	0.68	0.72	0.50
		ARECHO	✓	✓	**4.99**	**0.91**	**0.78**	**0.79**	**0.65**
		UniVERSA	✗	✗	96.99	0.69	0.52	0.69	0.45
		UniVERSA-T	✓	✗	37.72	0.79	0.68	0.71	0.49
		ARECHO	✓	✓	**23.26**	**0.86**	**0.72**	**0.74**	**0.57**

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 (↑)
Base	Dev.	UniVERSA	✗	✗	160.06	0.69	0.53	0.68	0.42
		UniVERSA-T	✓	✗	40.95	0.78	0.68	0.70	0.46
		ARECHO	✓	✓	**25.73**	**0.86**	**0.72**	**0.71**	**0.51**
	Enhanced	UniVERSA	✗	✗	61.54	0.71	0.54	0.69	0.43
		UniVERSA-T	✓	✗	27.34	0.81	0.68	0.70	0.47
		ARECHO	✓	✓	**20.58**	**0.84**	**0.69**	**0.72**	**0.51**
	Corrupted	UniVERSA	✗	✗	170.65	0.61	0.48	0.70	0.46
		UniVERSA-T	✓	✗	77.72	0.74	0.67	0.71	0.50
		ARECHO	✓	✓	**44.22**	**0.82**	**0.70**	**0.72**	**0.55**
	Synthesized	UniVERSA	✗	✗	58.79	0.76	0.54	0.69	0.45
		UniVERSA-T	✓	✗	8.10	0.84	0.68	0.72	0.50
		ARECHO	✓	✓	**4.99**	**0.91**	**0.78**	**0.79**	**0.65**
	Avg. Test	UniVERSA	✗	✗	96.99	0.69	0.52	0.69	0.45
		UniVERSA-T	✓	✗	37.72	0.79	0.68	0.71	0.49
		ARECHO	✓	✓	**23.26**	**0.86**	**0.72**	**0.74**	**0.57**

Dependency Modeling

Top 3 – Bottom 3 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

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
- Same concept in pre-training/mid-training (meta data permutation)
- From fixed metric set to natural language (connecting to LLM)
    - 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!