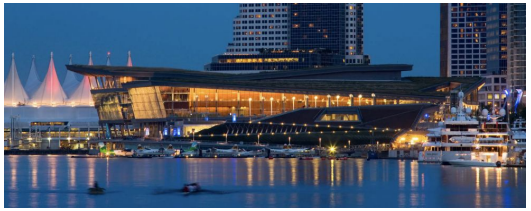
NEURAL INFORMATION PROCESSING SYSTEMS 2024



SSDM: Scalable Speech Dysfluency Modeling

Jiachen Lian, Xuanru Zhou, Zoe Ezzes, Jet Vonk, Brittany Morin, David Baquirin, Zachary Miller, Maria Luisa Gorno Tempini, Gopala Krishna Anumanchipalli

Speaker: Jiachen Lian

UC Berkekeley, UCSF

EECS

BAIR BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

UCSF University of California San Francisco

NIH National Institutes of Health

Background: Dysfluent speech

Which one is dysfluent speech?

Text: You wish to know all about my grandfather

Speech1		Closer [accented]
Speech2		[Uh,] you wish to know all about my grandfather
Speech3		[Uh,] you wish to know all about my [Uh], [my] grandfather
Speech4		[Uh,] [are you rich in all by] my grandfather [so fast]
Speech5		[Uh,] [Y-Y-] You [w-w-w-w]-ish to know all about my [gran-g-g-g-g-g-g]-grandfather [so fast]

Background: Dysfluent speech

Definition

We refer to dysfluent speech as any form of speech characterized by abnormal patterns such as pronunciation error/deficiency, repetition, prolongation, and irregular pauses, etc

[1] Lian, J. et al. "Unconstrained Dysfluency Modeling for Dysfluent Speech Transcription and Detection" 2023 ASRU

Background-Market Values

- Speech Therapy
    - Speech Disorders: Dyslexia, Aphasia, etc
    - Spoken Language Pathologist (SLP)

Market Value: 6.93B $ by 2023 https://www.fortunebusinessinsights.com/u-s-speech-therapy-market-105574

- Spoken Language Learning



Market Value: 337.2B $ by 2023 https://techreport.com/statistics/language-learning-market-statistics/

Background: Importance

- ➔ **17.9 million adults** and **1.4 percent of children** in the U.S. suffer from chronic communication and speech disorders [1]
- ➔ **1 in 14 kids** is at risk of developing language disabilities [2]
- ➔ Treatment not affordable by low-income families. [3]
- ➔ Not all hospitals have trained SLPs. Training SLP is expensive. [4]

[1] https://www.nidcd.nih.gov/health/statistics/quick-statistics-voice-speech-language
[2] https://www.nidcd.nih.gov/health/developmental-language-disorder
[3] https://www.trustedhealth.com/blog/speech-language-pathologist-licensure-guide
[4] https://dyslexichelp.org/why-doesnt-medical-insurance-cover-dyslexia/

Unfortunately, state-of-the-art speech AI models do not work

Aphasia Speech

What do we want from AI

- Reference Text:
  Giving those who observed him a pronounced feeling of the utmost respect.
- Ground Truth Transcription from SLP

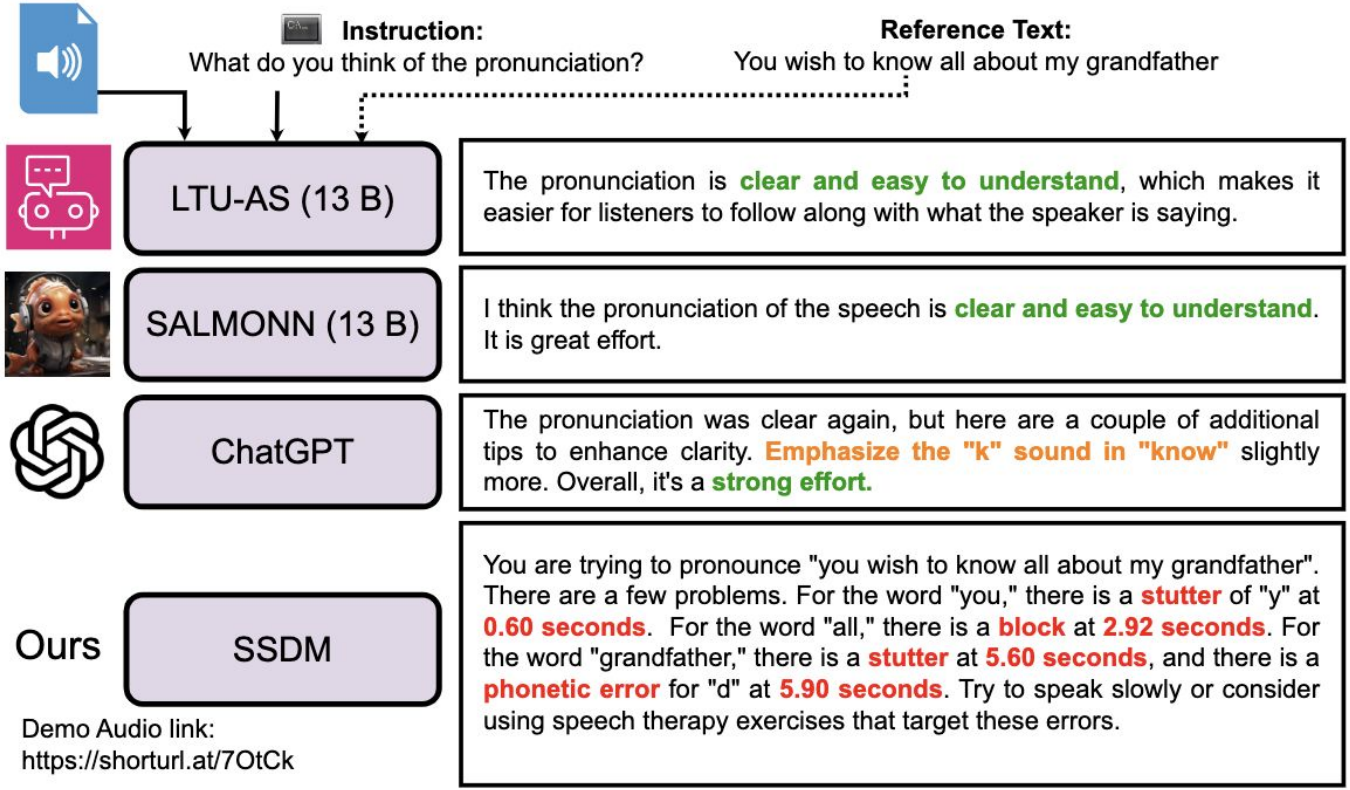
Segment 7: giving <th- th-> those who ob-server [intersegmental pause, addition of /er/] him a [pause] <por- por-> [very distorted] pronounced [distorted] feeling of the [distorted] the utmost respecks [sound substitution /ks/ for /t/]

*code said “-ed” was missed but I couldn’t tell based on the audio

- Whisper Transcription (best ASR model in the world, from OpenAI [1])
  giving those who observe him a pronounced feeling of the utmost respects.

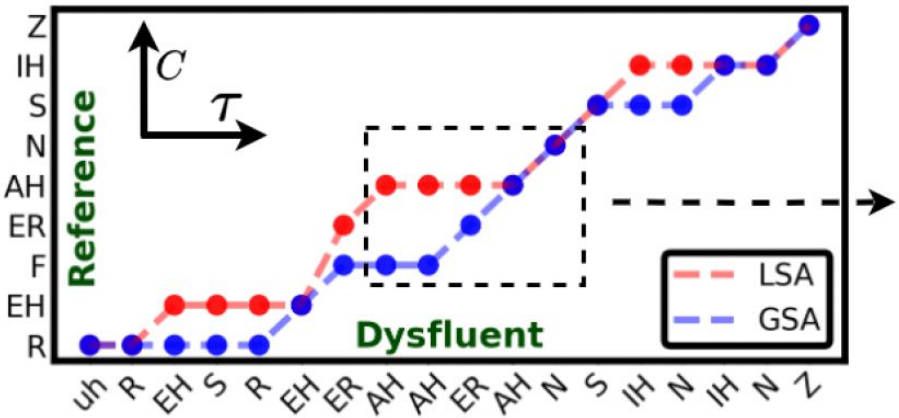
[1] Radford, Alec, et al. "Robust speech recognition via large-scale weak supervision." *International Conference on Machine Learning*. PMLR, 2023.

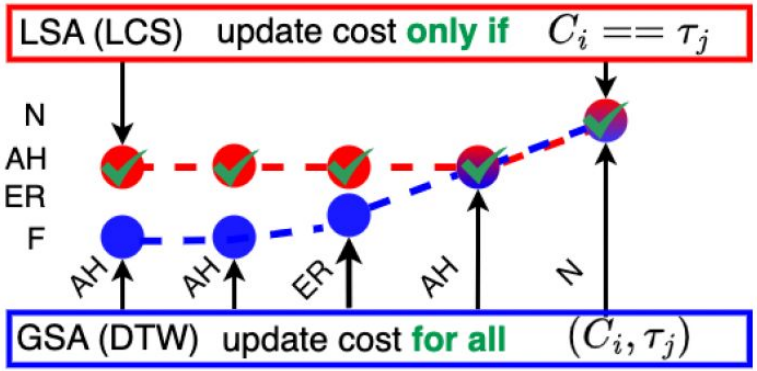
End-to-end Instruction Tuning



Towards Dysfluency Modeling

Revisit Monotonic Alignment



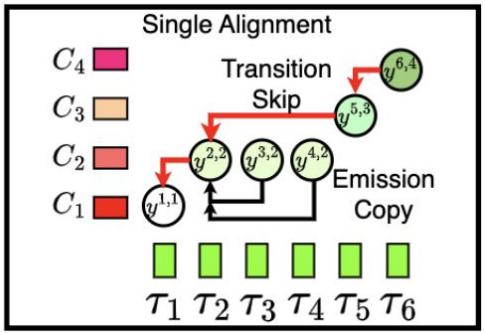


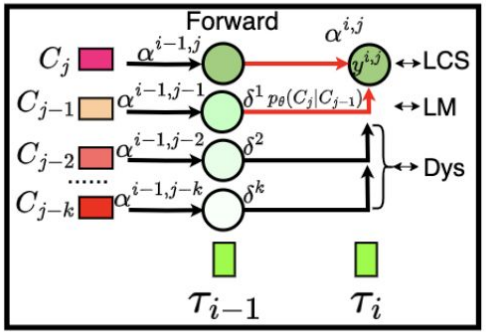
LSA: Local Sequence Aligner. E.g. LCS (Longest Common Subsequence)
GSA: Global Sequence Aligner. E.g. DTW (Dynamic Time Wraping)

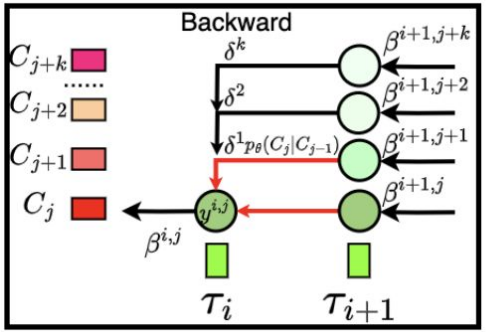
AH: (AH, AH, ER, AH)

Differentiable and Stochastic!

Differentiable and Stochastic Local Sequence Aligner





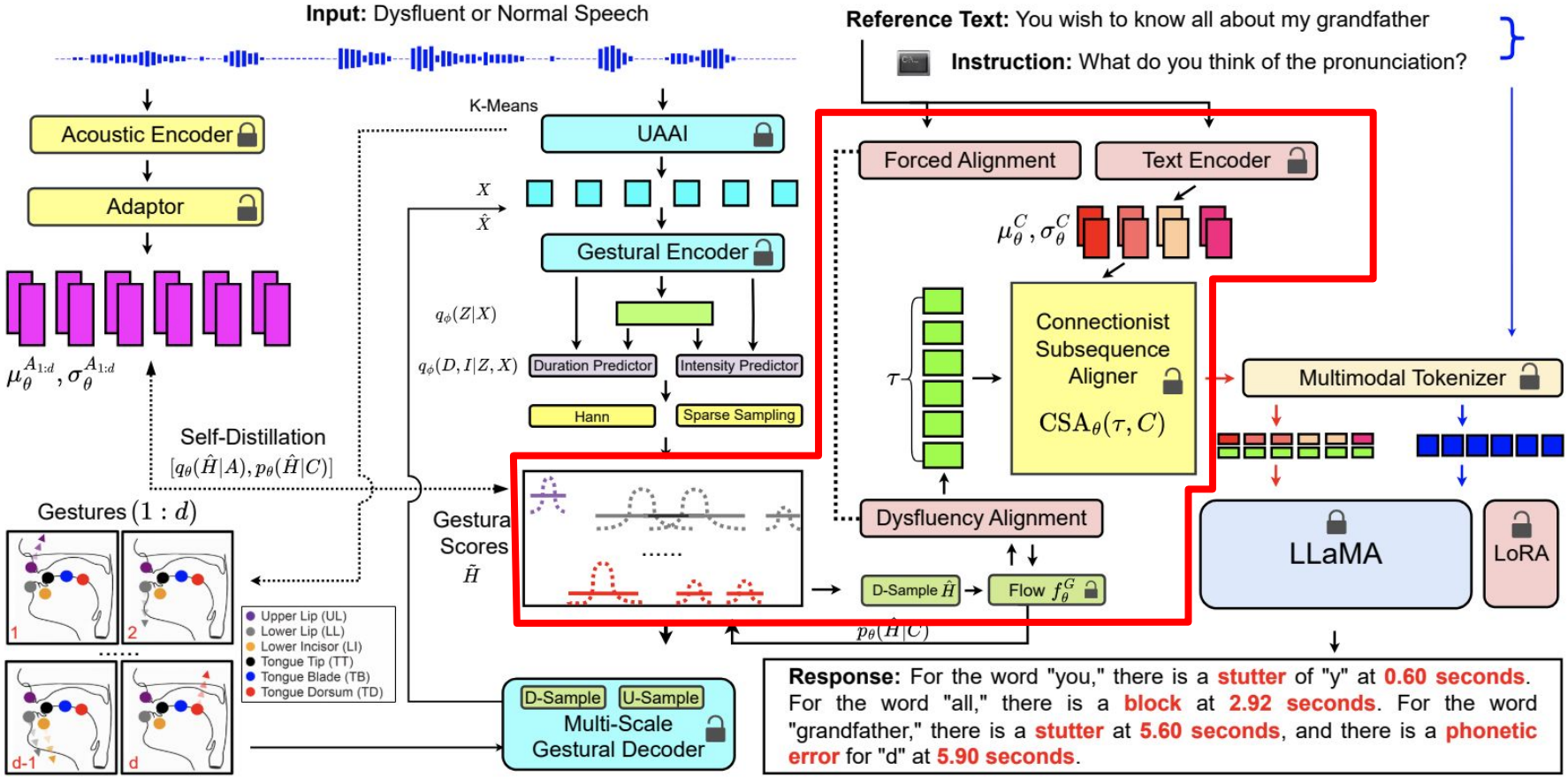


$$\max_{\theta} \mathop{\mathbb{E}}_{C,\tau} \sum_{j=1}^{N} p_\theta(\gamma_j^{\mathrm{LSA}}(C)|C,\tau) = \max_{\theta} \mathop{\mathbb{E}}_{C,\tau} \sum_{j=1}^{N} p_\theta(\Gamma_j^{\mathrm{LSA}}(\tau)|C,\tau) \approx \max_{\theta} \mathop{\mathbb{E}}_{C\tau} \sum_{j=1}^{N} p_\theta(\Gamma_j^{'}(\tau)|\tau)$$

$$\mathcal{L}_{\mathrm{CSA}} = -\mathop{\mathbb{E}}_{C,\tau} \sum_{j=1}^{N} p_\theta(\Gamma_j^{'}(\tau)|\tau) = -\sum_{i=1}^{t''} \sum_{j=1}^{L} \frac{\alpha_\theta^{i,j} \beta_\theta^{i,j}}{y^{i,j}} \qquad y^{i,j} = p_\theta(C_j|\tau_i) \approx \frac{\exp^{\tau_i \cdot C_j^S}}{(\sum_{k=1}^{L} \exp^{\tau_i \cdot C_k^S})}$$

$$\alpha_\theta^{i,j} = \alpha_\theta^{i-1,j} + \sum_{k=1}^{j} \delta^k \alpha_\theta^{i-1,j-k} \cdot y^{i,j} \cdot \left(p_\theta(C_{j-1}^S|C_j^S) \cdot \mathbf{1}_{\{k=1\}} + \mathbf{1}_{\{k\neq 1\}}\right)$$

$$\beta_\theta^{i,j} = \beta_\theta^{i+1,j} + \sum_{k=1}^{t'-j} \delta^k \beta_\theta^{i+1,j+k} \cdot y^{i+1,j+k} \cdot \left(p_\theta(C_j^S|C_{j+1}^S) \cdot \mathbf{1}_{\{k=1\}} + \mathbf{1}_{\{k\neq 1\}}\right)$$


Input: Dysfluent or Normal Speech
Reference Text: You wish to know all about my grandfather
Instruction: What do you think of the pronunciation?
Acoustic Encoder
Adaptor
$\mu_\theta^{A_{1:d}}, \sigma_\theta^{A_{1:d}}$
K-Means
UAAI
X
$\hat{X}$
Gestural Encoder
$q_\phi(Z|X)$
$q_\phi(D, I|Z, X)$
Duration Predictor
Intensity Predictor
Hann
Sparse Sampling
Self-Distillation
$[q_\theta(\hat{H}|A), p_\theta(\hat{H}|C)]$
Gestures $(1:d)$
Gestura
Scores
$\tilde{H}$
Upper Lip (UL)
Lower Lip (LL)
Lower Incisor (LI)
Tongue Tip (TT)
Tongue Blade (TB)
Tongue Dorsum (TD)
D-Sample
U-Sample
Multi-Scale
Gestural Decoder
Forced Alignment
Text Encoder
$\mu_\theta^C, \sigma_\theta^C$
τ
Connectionist
Subsequence
Aligner
$\mathrm{CSA}_\theta(\tau, C)$
Dysfluency Alignment
D-Sample $\hat{H}$
Flow f_θ^G
$p_\theta(\hat{H}|C)$
Multimodal Tokenizer
LLaMA
LoRA
Response: For the word "you," there is a stutter of "y" at 0.60 seconds. For the word "all," there is a block at 2.92 seconds. For the word "grandfather," there is a stutter at 5.60 seconds, and there is a phonetic error for "d" at 5.90 seconds.

Neural Gestural Scores are Scalable Speech Representations

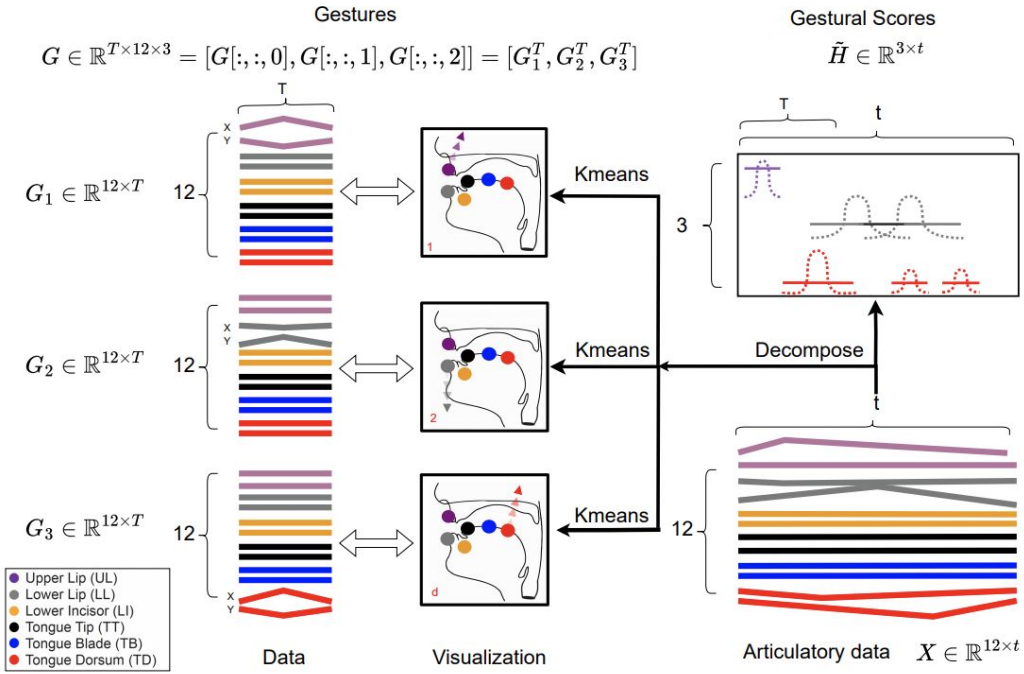
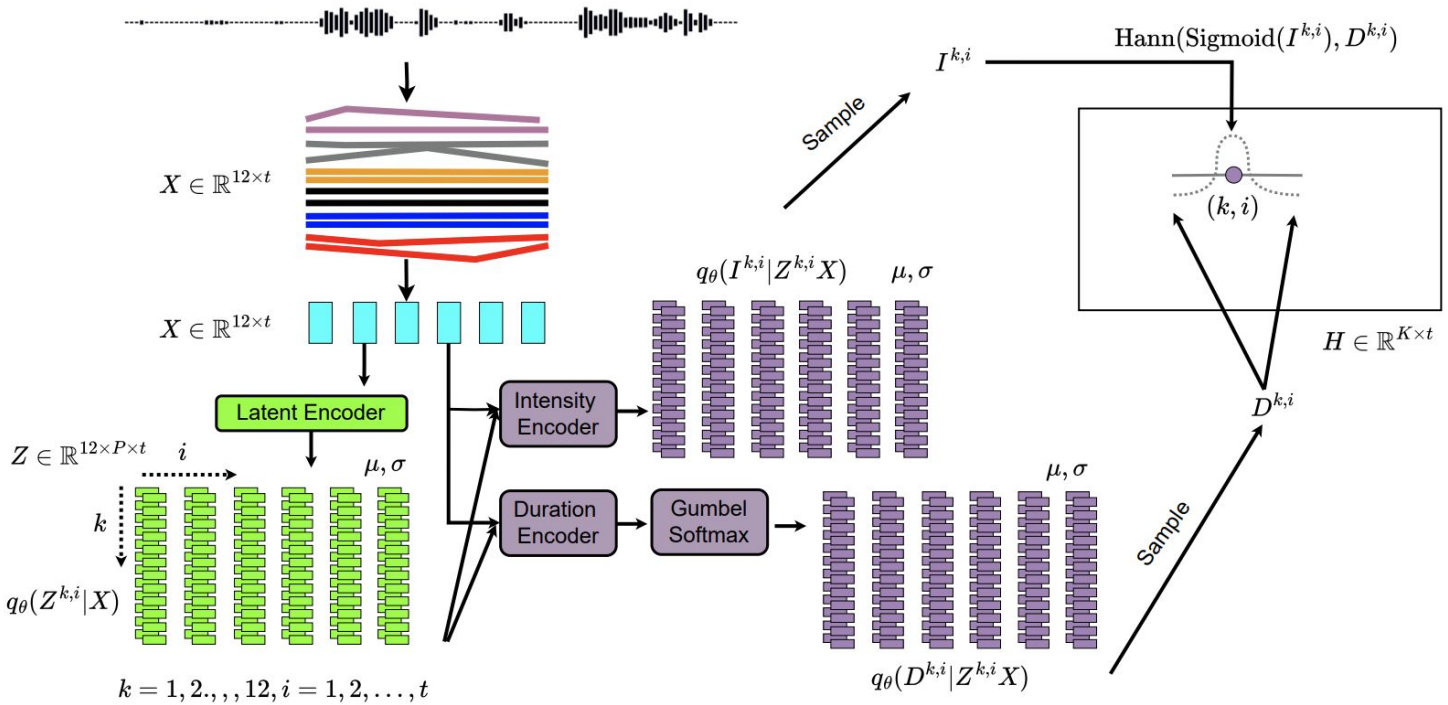
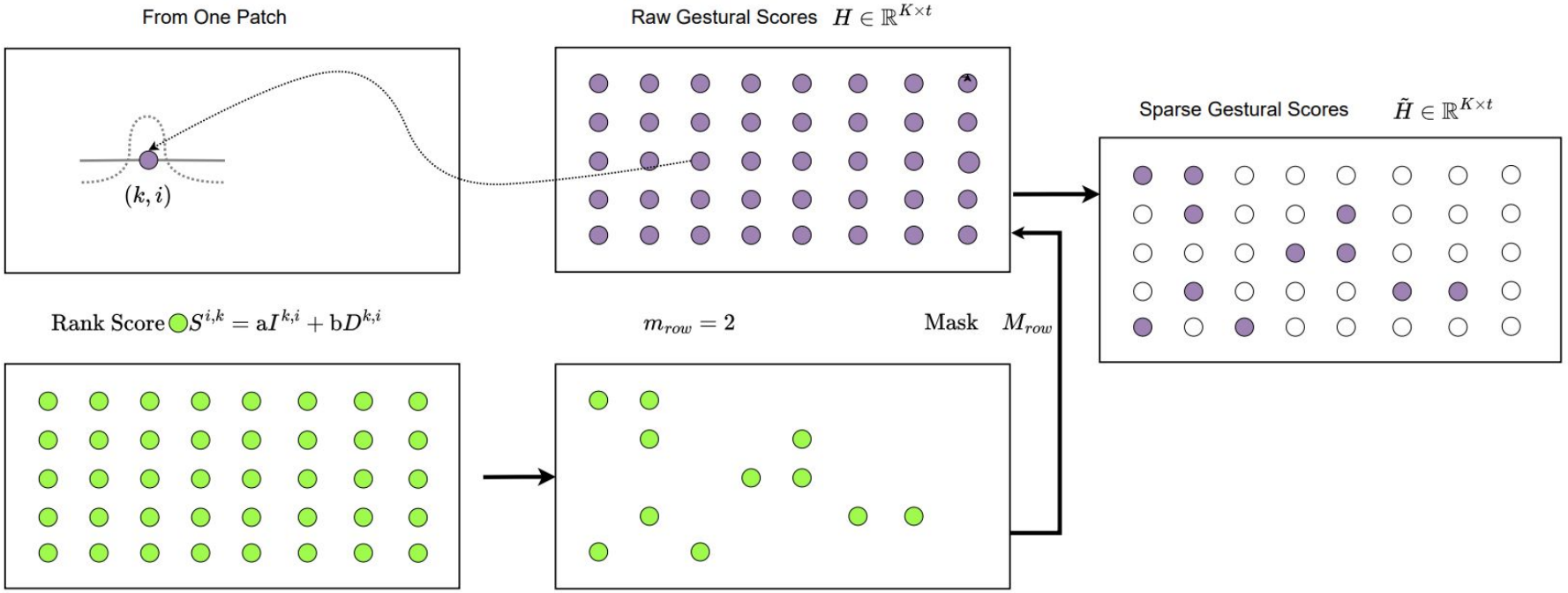


Figure 6: Gestures, Gestural Scores, Raw Data Visualization

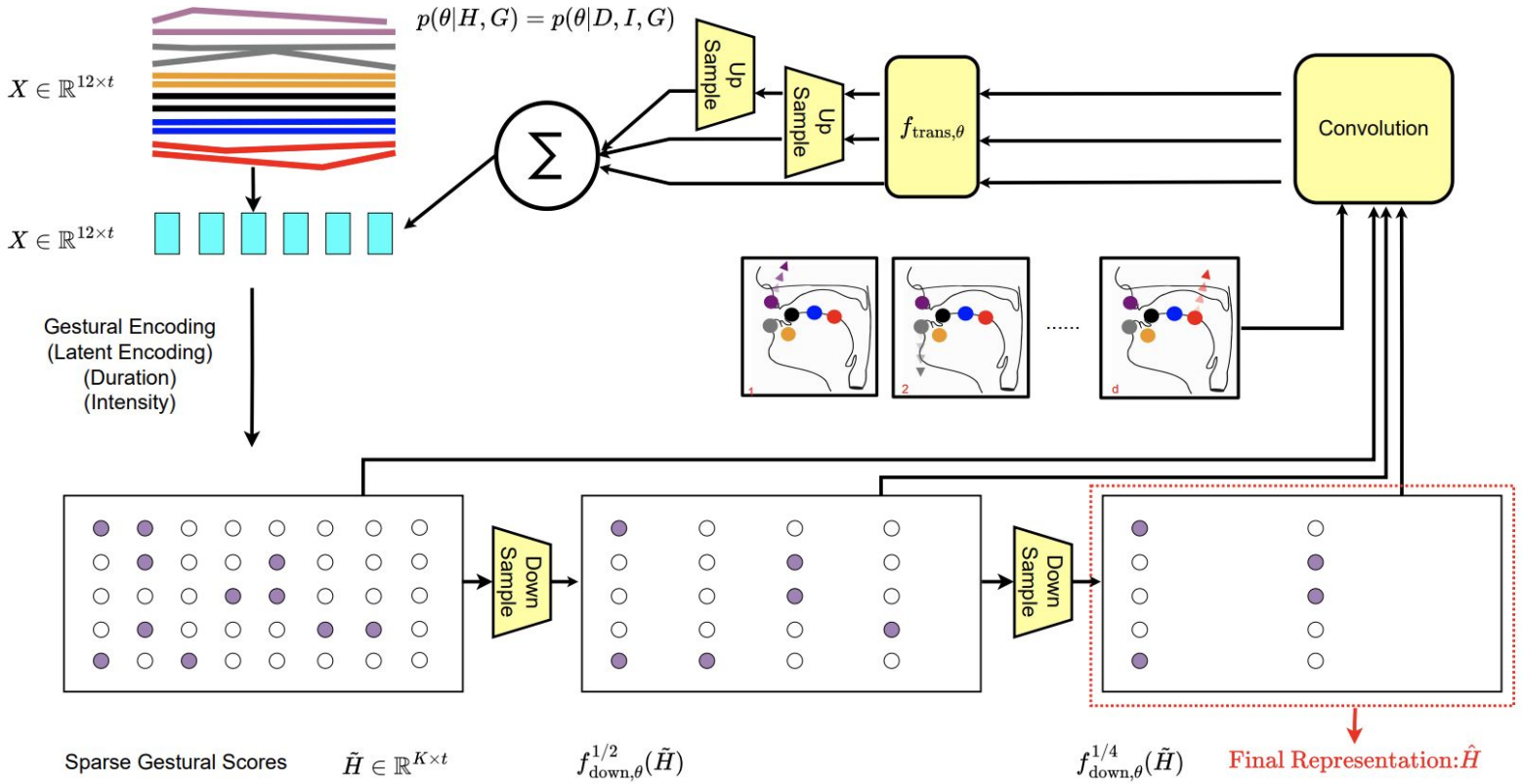
Duration and Intensity Modeling

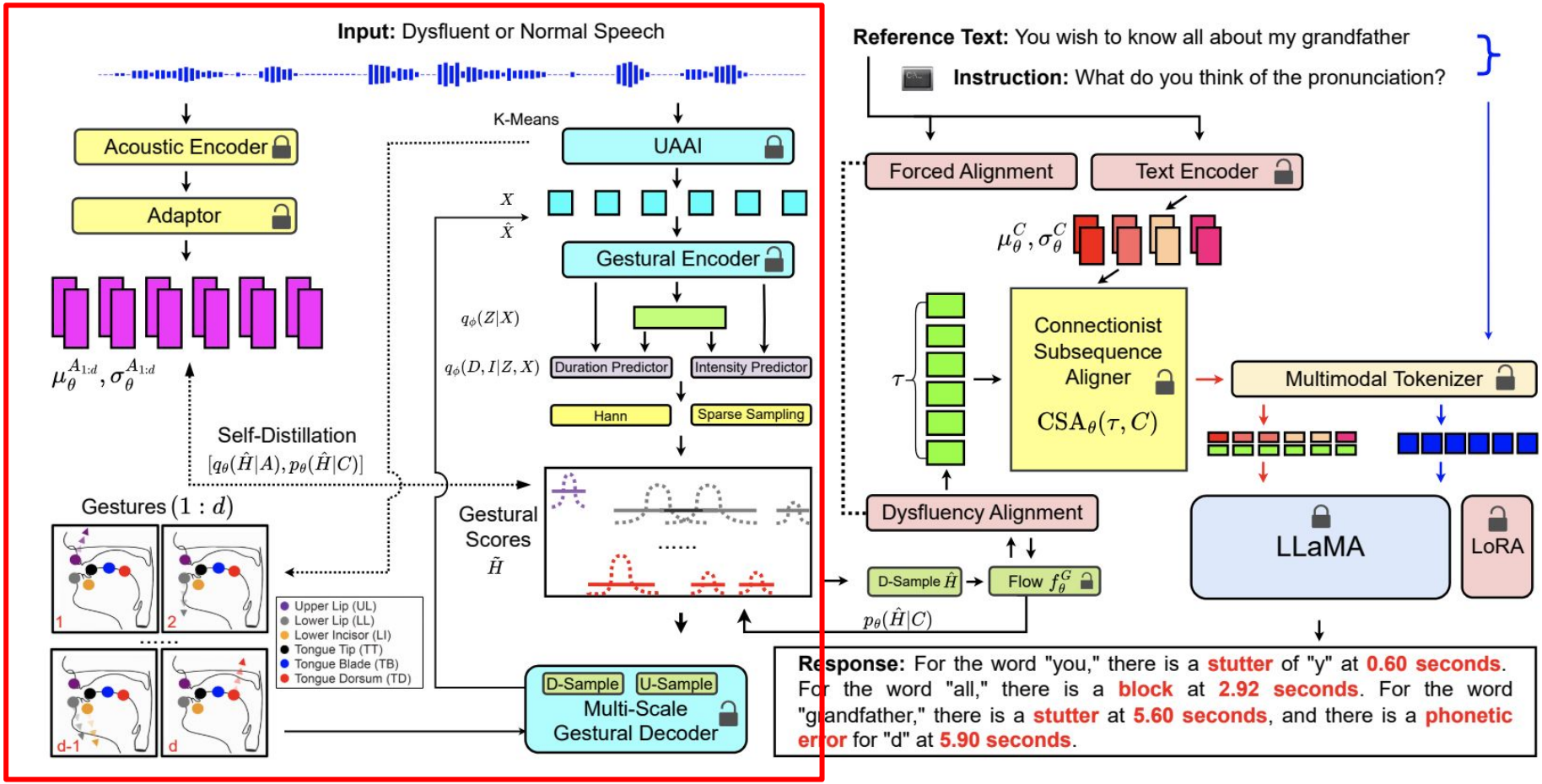


Sparse Sampling

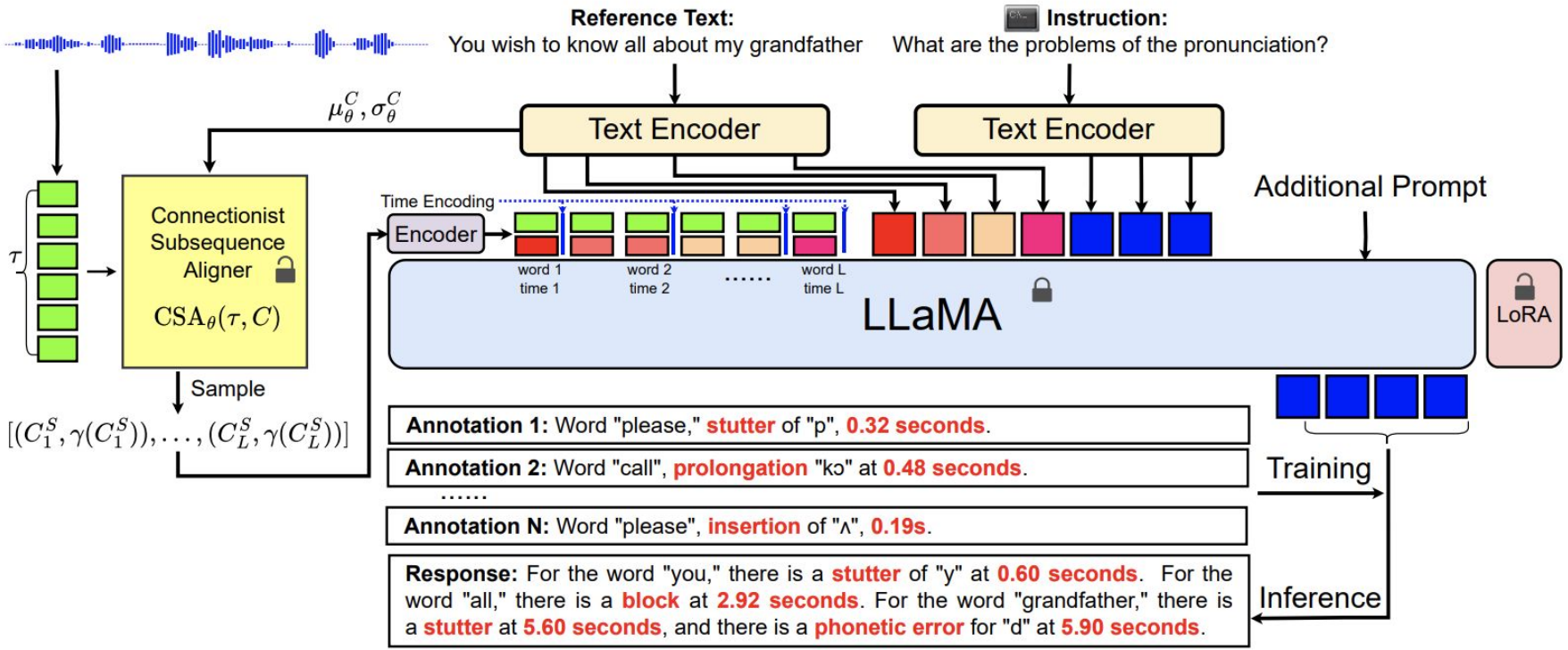


Reconstruction




Input: Dysfluent or Normal Speech
Acoustic Encoder
Adaptor
$\mu_\theta^{A_{1:d}}, \sigma_\theta^{A_{1:d}}$
K-Means
UAAI
X
$\hat{X}$
Gestural Encoder
$q_\phi(Z|X)$
$q_\phi(D, I|Z, X)$
Duration Predictor
Intensity Predictor
Hann
Sparse Sampling
Self-Distillation
$[q_\theta(\hat{H}|A), p_\theta(\hat{H}|C)]$
Gestures $(1:d)$
1
2
d-1
d
Upper Lip (UL)
Lower Lip (LL)
Lower Incisor (LI)
Tongue Tip (TT)
Tongue Blade (TB)
Tongue Dorsum (TD)
Gestural Scores $\tilde{H}$
D-Sample
U-Sample
Multi-Scale Gestural Decoder
Reference Text: You wish to know all about my grandfather
Instruction: What do you think of the pronunciation?
Forced Alignment
Text Encoder
$\mu_\theta^C, \sigma_\theta^C$
τ
Connectionist Subsequence Aligner
$\mathrm{CSA}_\theta(\tau, C)$
Multimodal Tokenizer
Dysfluency Alignment
D-Sample $\hat{H}$
Flow f_θ^G
$p_\theta(\hat{H}|C)$
LLaMA
LoRA
Response: For the word "you," there is a stutter of "y" at 0.60 seconds. For the word "all," there is a block at 2.92 seconds. For the word "grandfather," there is a stutter at 5.60 seconds, and there is a phonetic error for "d" at 5.90 seconds.

Language Models



Where is annotated training data?

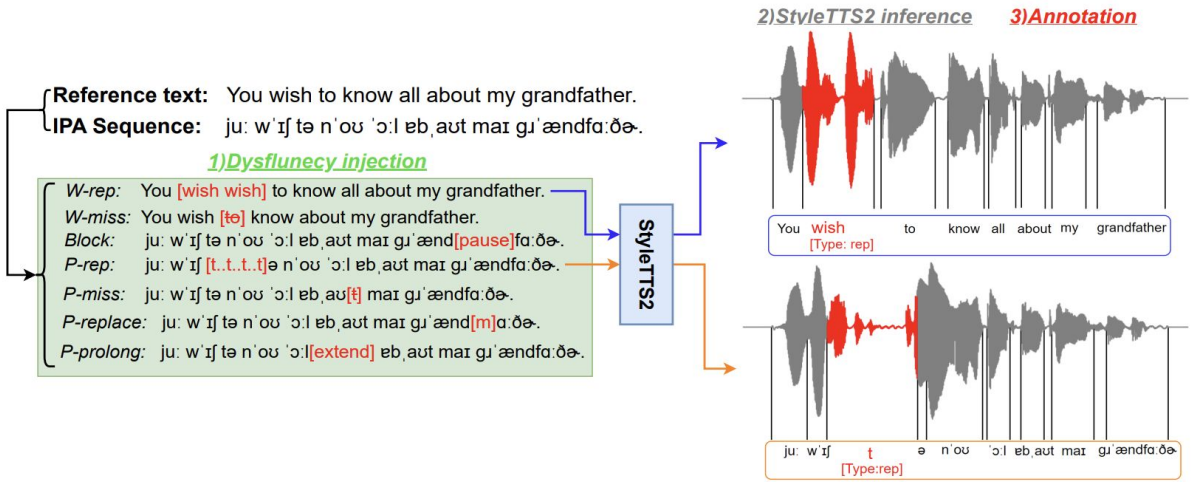


Table 4: Types of Dysfluency Data in VCTK++ and Libri-Dys

Dysfluency	# Samples VCTK++ [1]	Percentage VCTK++	# Samples Libri-Dys	Percentage Libri-Dys
Prolongation	43738	33.28	288795	13.24
Block	43959	33.45	345853	15.97
Replacement	0	0	295082	13.63
Repetition (Phoneme)	43738	33.28	340916	15.75
Repetition (Word)	0	0	301834	13.94
Missing (Phoneme)	0	0	296076	13.68
Missing (Word)	0	0	296303	13.69
Total Hours of Audio	130.66		3983.44	

Dysfluent Scalability Evaluation:

Method	Eval Data	F1 (%, ↑)	dPER (%, ↓)	F1 (%, ↑)	dPER (%, ↓)	F1 (%, ↑)	dPER (%, ↓)	F1 (%, ↑)	dPER (%, ↓)	F1 (%, ↑)	dPER (%, ↓)	SF1 (%, ↑)	SF2 (%, ↓)
Training Data		*VCTK++*		*LibriTTS (100%)*		*Libri-Dys (30%)*		*Libri-Dys (60%)*		*Libri-Dys (100%)*			
HuBERT [108]	VCTK++	90.5	40.3	90.0	40.0	89.8	41.2	91.0	40.2	89.9	41.2	0.15	-0.1
	Libri-Dys	86.2	50.3	88.2	47.4	87.2	42.3	87.2	43.4	87.8	42.9	0.18	0.29
H-UDM [2]	VCTK++	91.2	39.8	91.0	38.8	90.7	39.0	91.3	39.9	90.9	40.2	0.12	0.45
	Libri-Dys	88.1	44.5	88.9	45.6	88.0	43.3	88.5	43.3	88.9	43.0	0.32	-0.09
GS-only	VCTK++	88.1	41.9	88.1	42.2	88.3	41.9	88.9	41.9	89.4	40.7	0.39	-0.36
	Libri-Dys	84.7	44.5	85.0	43.3	85.5	43.0	85.7	42.2	86.5	41.5	0.32	-0.53
GS w/o dist	VCTK++	91.4	39.0	91.6	38.5	91.5	38.8	92.0	37.2	92.6	37.1	0.38	-0.67
	Libri-Dys	88.0	42.4	88.3	41.9	88.7	41.0	88.9	39.4	90.0	39.0	0.11	**-0.76**
GS w/ dist	VCTK++	**91.5**	**39.0**	**91.7**	**38.3**	**91.7**	**38.6**	**92.1**	**37.0**	**93.0**	**37.0**	0.43	-0.64
	Libri-Dys	**88.2**	**40.9**	**88.9**	**40.9**	**89.0**	**40.8**	**89.2**	**39.0**	**90.8**	**39.0**	**0.56**	-0.72

Table 1: Scalable Dysfluent Phonetic Transcription Evaluation

Method	Eval Data	F1 (%, ↑)	MS (%, ↑)	F1 (%, ↑)	MS (%, ↑)	F1 (%, ↑)	MS (%, ↑)	F1 (%, ↑)	MS (%, ↑)	F1 (%, ↑)	MS (%, ↑)	SF1 (%, ↑)	SF2 (%, ↑)
Training Data		*VCTK++*		*LibriTTS (100%)*		*Libri-Dys (30%)*		*Libri-Dys (60%)*		*Libri-Dys (100%)*			
H-UDM [2]	VCTK++	78.3	60.7	82.5	63.9	84.3	66.1	84.2	65.3	84.1	65.2	-0.07	-0.35
	Libri-Dys	74.8	63.9	75.0	62.9	77.2	60.1	75.0	62.3	75.9	61.1	-0.61	0.64
SSDM	VCTK++	84.8	64.3	87.8	68.2	88.5	69.7	89.0	69.9	89.2	70.2	0.26	0.17
	Libri-Dys	78.9	68.3	79.0	69.4	79.3	69.8	80.6	69.9	81.4	70.4	0.76	0.19
w/o LLaMA	VCTK++	84.5	64.0	86.9	68.0	88.4	69.7	88.7	69.8	88.9	69.9	0.18	0.07
	Libri-Dys	78.2	68.1	78.3	69.0	78.8	69.2	79.6	69.3	80.7	70.0	0.65	0.25
w/ DTW	VCTK++	80.3	60.9	83.5	65.9	84.2	66.2	85.0	66.6	85.2	67.2	0.38	0.34
	Libri-Dys	75.9	65.6	76.3	67.4	76.7	67.5	77.9	68.2	78.0	68.4	0.51	0.32
w/o GS	VCTK++	84.3	64.1	86.9	65.0	87.4	66.2	87.1	66.3	87.2	66.5	-0.09	0.1
	Libri-Dys	76.9	66.1	77.0	66.4	77.7	67.8	78.6	68.1	78.8	68.4	0.42	0.21
w/ Curri	VCTK++	**85.6**	**65.1**	**87.1**	**68.5**	**88.8**	**69.9**	**89.2**	**70.2**	**90.0**	**71.9**	0.4	**0.63**
	Libri-Dys	**79.2**	**68.4**	**79.4**	**69.5**	**79.4**	**69.9**	**81.0**	**70.5**	**81.6**	**71.0**	**0.82**	0.39

Table 2: Scalable Dysfluent Detection Evaluation (Simulation)

State-of-the-art Comparison

Eval Data	LTU-AS-13B [24]		LTU-AS-13B-FT		SALMONN-13B [27]		SALMONN-13B-FT		ChatGPT [110]		SSDM		SSDM w/ Curri	
	F1(%, ↑)	MS(%, ↑)	F1(%, ↑)	MS(%, ↑)	F1(%, ↑)	MS(%, ↑)	F1(%, ↑)	MS(%, ↑)	F1(%, ↑)	MS(%, ↑)	F1(%, ↑)	MS(%, ↑)	F1(%, ↑)	MS(%, ↑)
VCTK++	7.2	0	12.2	1.7	7.3	0	14.2	0.5	25.3	0	89.2	70.2	**90.0**	**71.9**
Libri-Dys	8.9	0	9.7	1.7	7.7	0	11.0	2.5	18.3	0	81.4	70.4	**81.6**	**71.0**
nfvPPA	0	0	2.4	0	0	0	1.8	0	5.6	0	69.2	54.2	**69.9**	**55.0**

Table 3: Detection results from state-of-the-art models.

Conclusions

- Unified Business/Research Platform for language learning/Disordered Speech Diagnosis
- End-to-end and flexible deep speech understanding framework
- Efficient Scaling