Meta

Expressive Speech-to-Speech Translation

Min-Jae Hwang
Postdoctoral Researcher at Seamless, FAIR

BISH Bash event on Feb 15th, 2024

AI at Meta

Agenda

Introduction

SeamlessExpressive

Conclusion

Speech-to-Speech Translation (S2ST) System

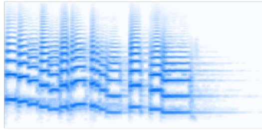
Concept

- Automatically converts speech signal in one language to speech signal in another language.
- Playing a crucial role in breaking down language barriers between different cultures in international conversation situations.

Direct S2ST with Discrete Units

Speech-to-unit translation (S2UT) model [Ann et al., 2021, Inaguma et al., 2023]

Mel-spectrogram



Language A

S2UT Model

Discretized semantic units

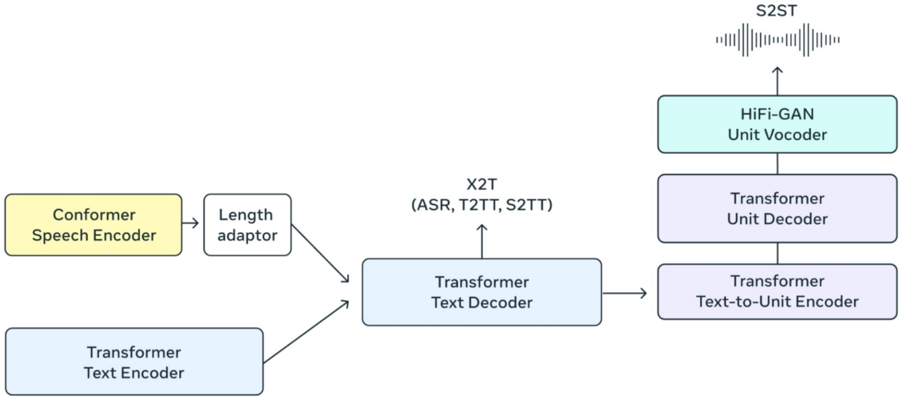
… 955 4 265 442 115 …

Language B

- Translate input speech into **discretized semantic units**, e.g., HuBERT [Hsu et al., 2021] and XLS-R [Babu et al., 2022]
  - Generate speech waveform using unit HiFi-GAN vocoder [Kong et al., 2020] from translated units
- Provide **high quality content translation performance**
  - Constraint model’s output space into semantic information
  - Leveraging advanced modeling techniques, e.g., model pretraining or data augmentation

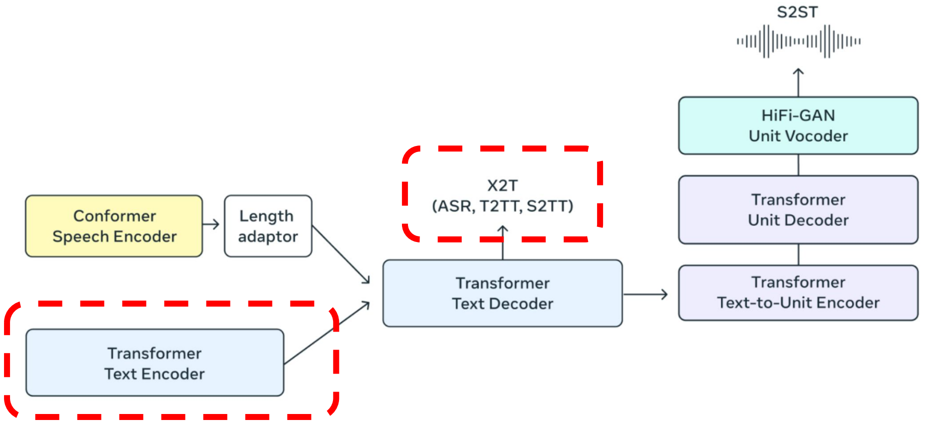
SeamlessM4T: Massively Multilingual & Multimodal Machine Translation Model

State-of-the-art S2UT model **[Seamless Communication, 2023]**



SeamlessM4T: Massively Multilingual & Multimodal Machine Translation Model

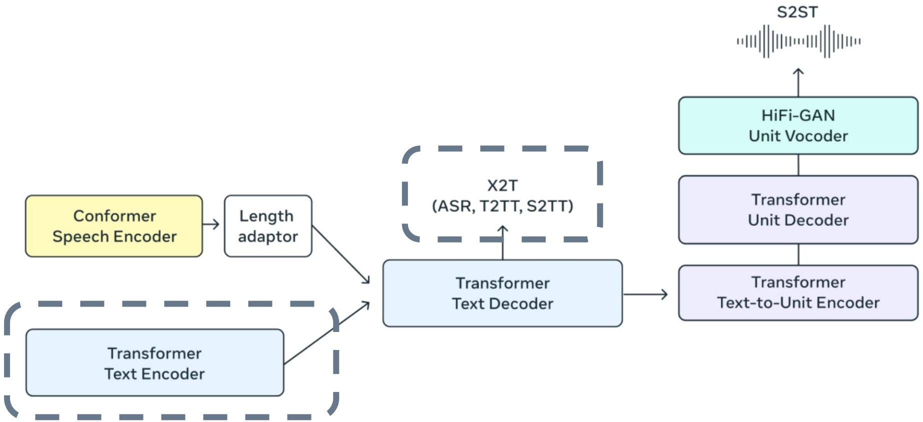
State-of-the-art S2UT model [Seamless Communication, 2023]



1. More modalities (Text input & output)

SeamlessM4T: Massively Multilingual & Multimodal Machine Translation Model

State-of-the-art S2UT model [Seamless Communication, 2023]



1. More modalities (Text input & output)

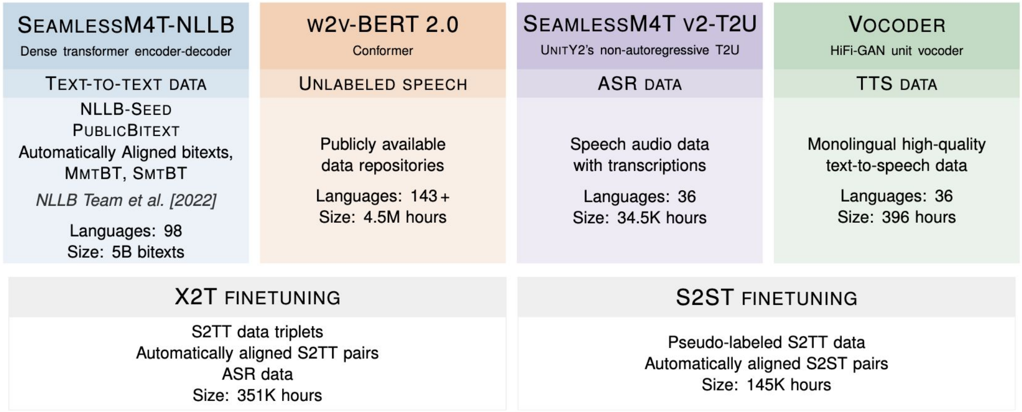
2. Pretrained components

SEAMLESSM4T-NLLB T2TT encoder-decoder	**w2V-BERT 2.0** Unsupervised speech pre-training	**T2U** Text-to-Unit encoder-decoder	**Vocoder** Speech resynthesis

SeamlessM4T: Massively Multilingual & Multimodal Machine Translation Model

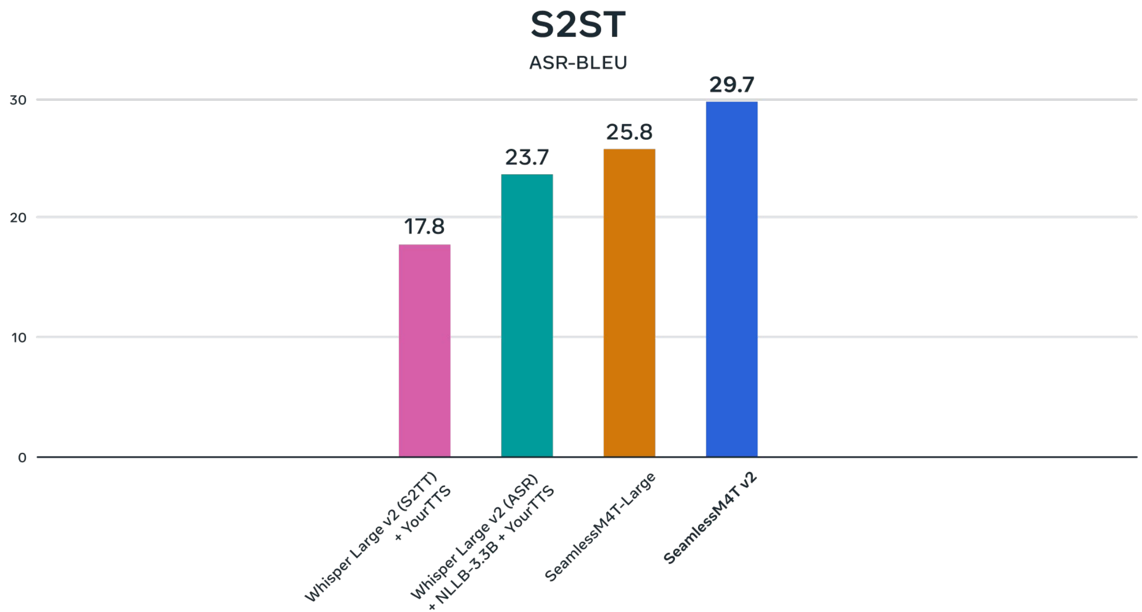
State-of-the-art S2UT model **[Seamless Communication, 2023]**

3. Massively multilingual training data



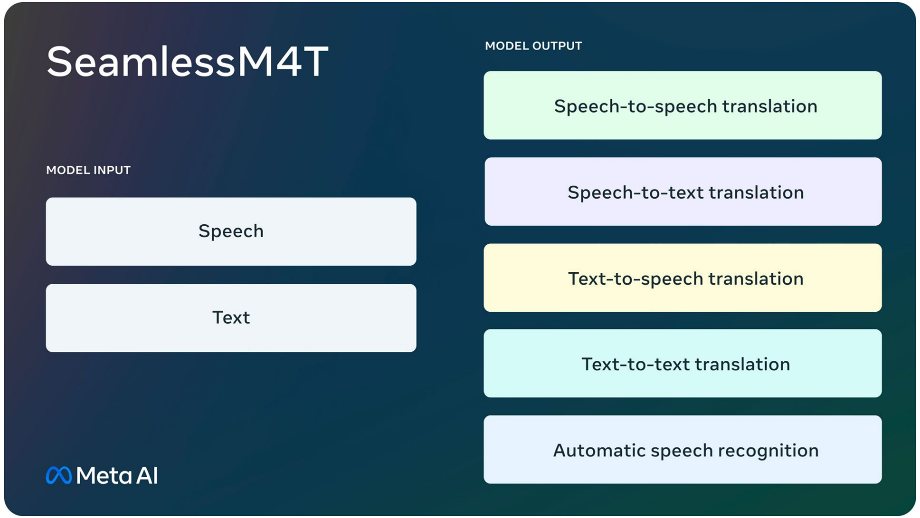
SeamlessM4T: Massively Multilingual & Multimodal Machine Translation Model

State-of-the-art S2UT model [Seamless Communication, 2023]



SeamlessM4T: Massively Multilingual & Multimodal Machine Translation Model

State-of-the-art S2UT model [Seamless Communication, 2023]

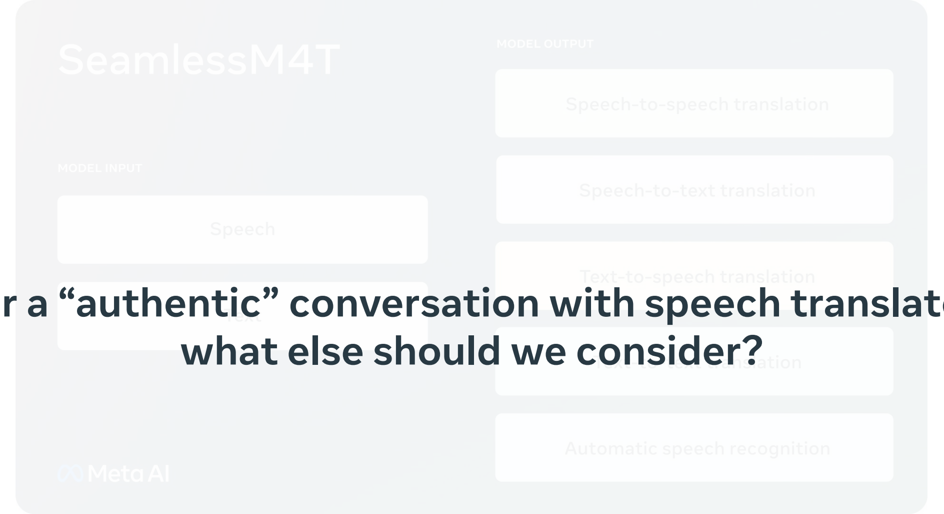


Inputs
101+ language speech
96 language text

Outputs
36 language speech
96 language text

SeamlessM4T: Massively Multilingual & Multimodal Machine Translation Model

State-of-the-art S2UT model [Seamless Communication, 2023]

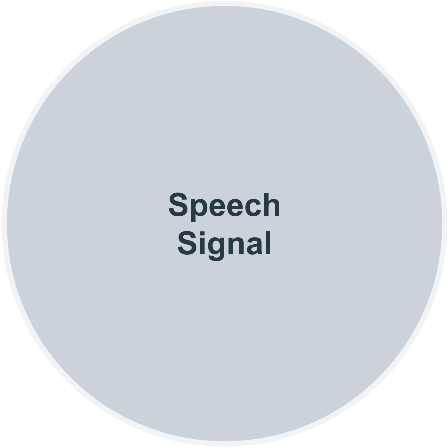


For a "authentic" conversation with speech translator, what else should we consider?

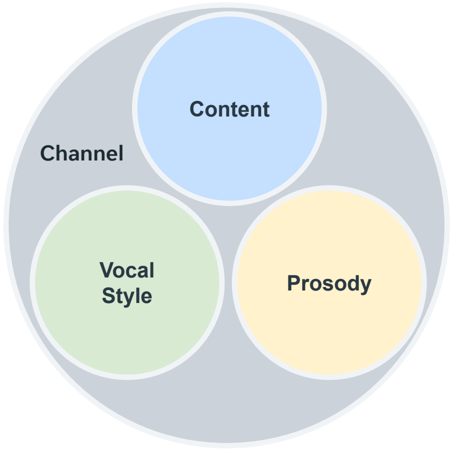
Inputs
101+ language speech
96 language text

Outputs
36 language speech
96 language text

Disentangled Property of Speech

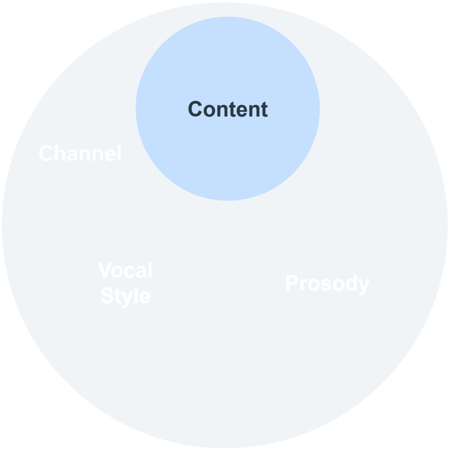


Disentangled Property of Speech



- **Content information**
  - Related to linguistic property
  - E.g., Semantic meaning, language identity, etc.
- **Vocal style information**
  - Related to vocal style
  - E.g., Voice color or speaking style
- **Prosody information**
  - Related to intonation, accent, rhythm, emotion, etc.
- **Channel information**
  - Information other than speech
  - E.g., Background noise or reverberation, etc.

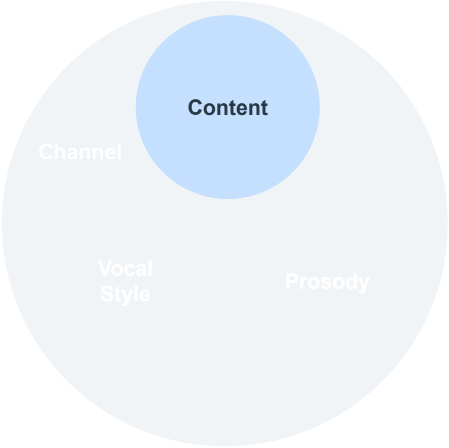
Disentangled Property of Speech



- **Content information**
    - Related to linguistic property
    - E.g., Semantic meaning, language identity, etc.

Target of conventional S2ST models

Disentangled Property of Speech

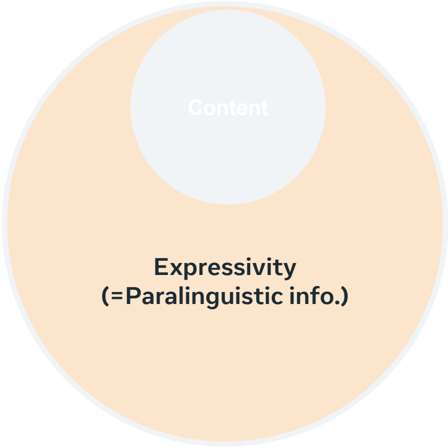


- Content information
  - Related to linguistic property
  - E.g., Semantic meaning, language identity, etc.

Target of conventional S2ST models

Conventional S2ST models ignores paralinguistic information.
So they generate monotone translated speech.

Disentangled Property of Speech



- Content information
  - Related to linguistic property
  - E.g., Semantic meaning, language identity, etc.

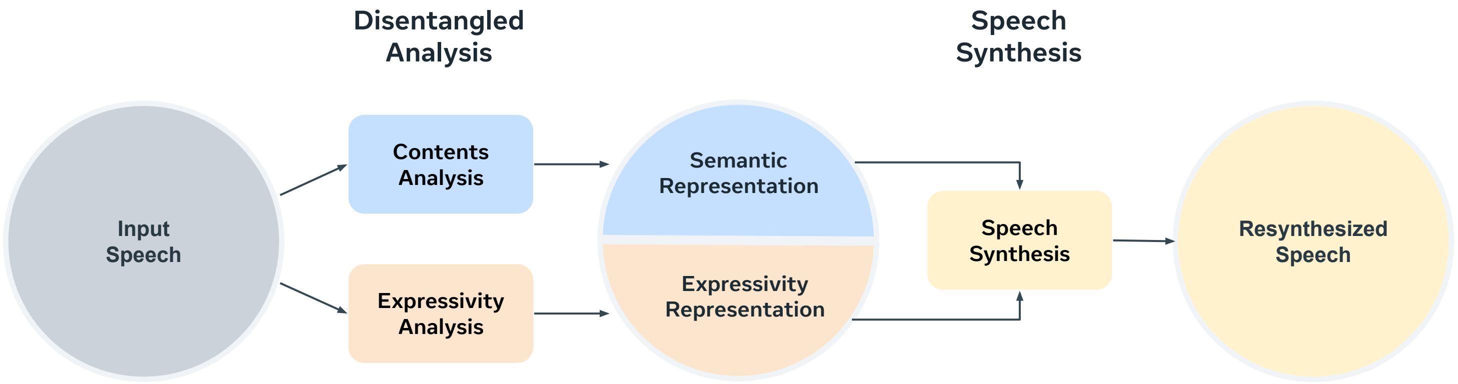
Target of conventional S2ST models

Conventional S2ST models ignores paralinguistic information.
So they generate monotone translated speech.

For the naturalistic conversation, paralinguistic (or expressivity) information also should be conveyed to listener!

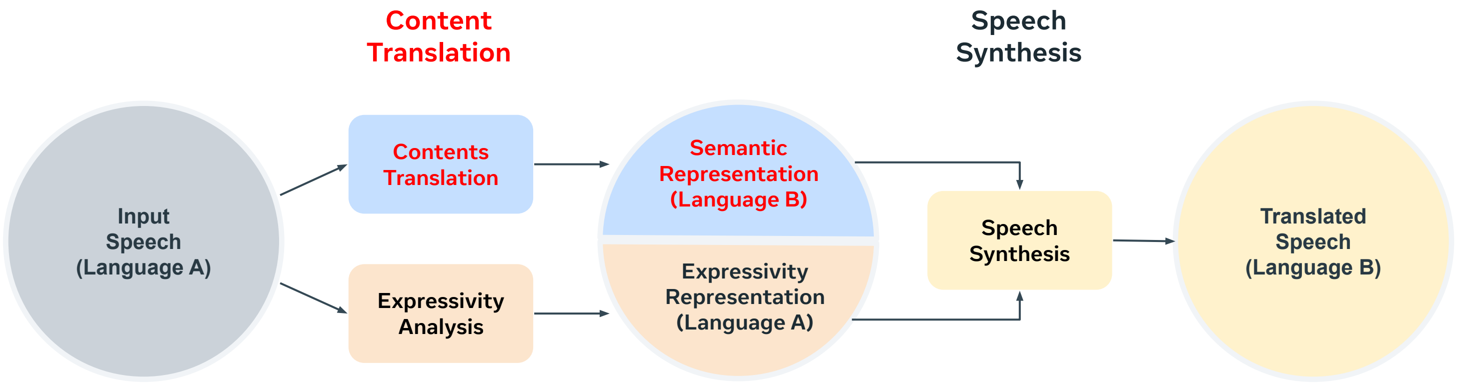
Idea for expressivity-preserved S2ST

Analysis-synthesis framework for **disentangled representation** of speech components



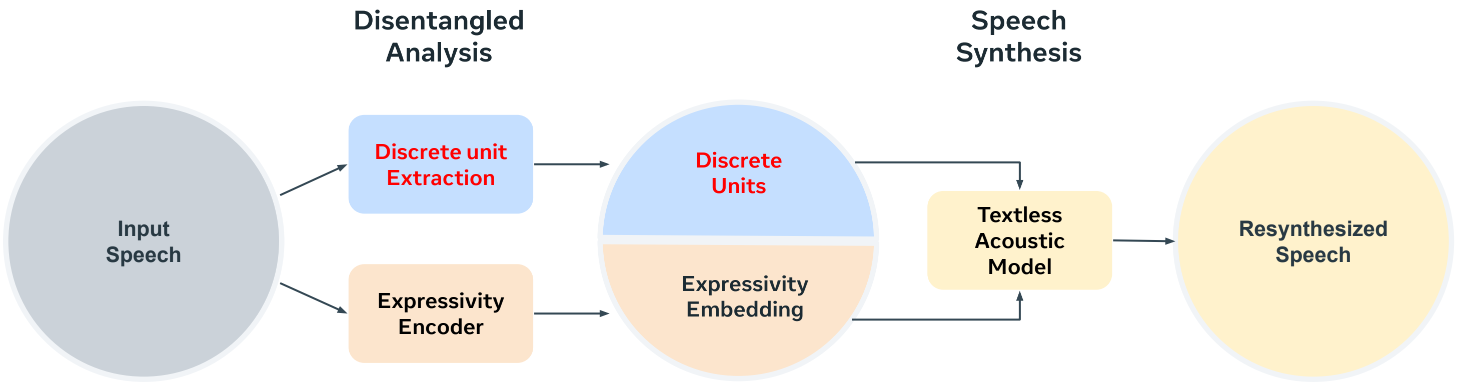
Idea for expressivity-preserved S2ST

Replace semantic representation with the one of target language for expressive S2ST



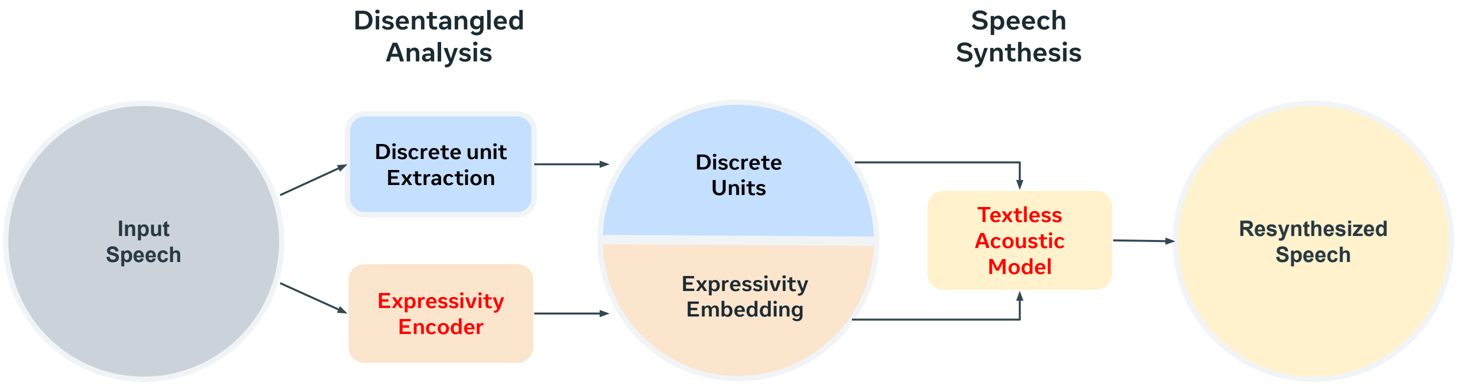
Idea for expressivity-preserved S2ST (detailed)

We know that the **discrete units** are efficient way to represent semantic information



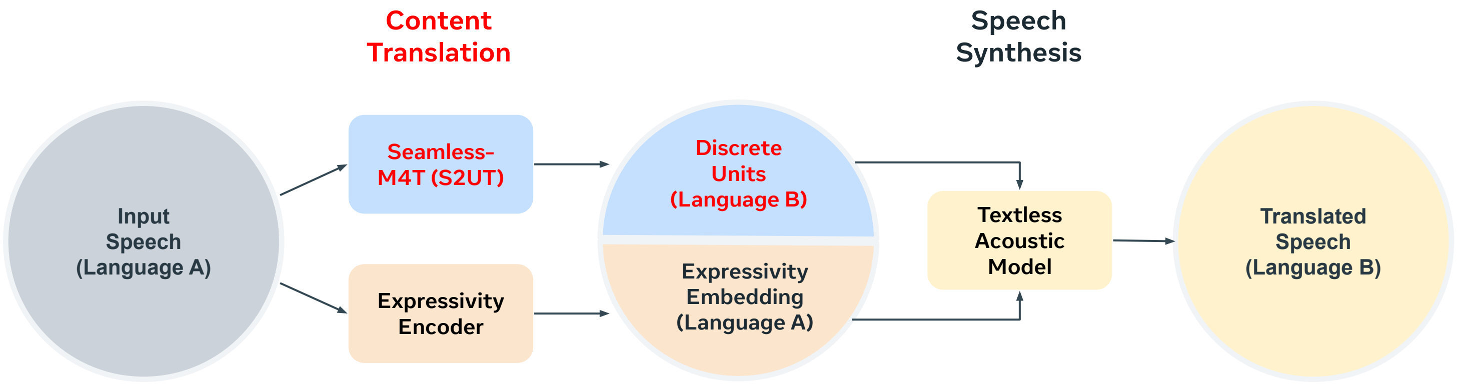
Idea for expressivity-preserved S2ST (detailed)

Fix unit extractor, then **train expressivity encoder** and **acoustic model**

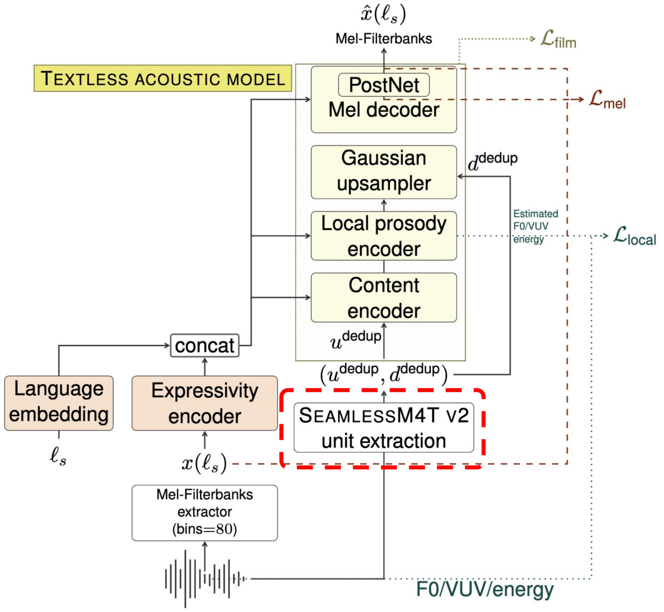


Idea for expressivity-preserved S2ST (detailed)

For S2ST, synthesize speech from **translated units (target language)** and **expressivity embedding (source language)**



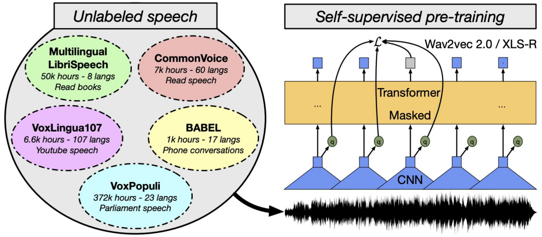
PRETSSEL: **P**aralinguistic **RE**presentation-based **T**extle**SS** acoustic mod**EL**



[PRETSSEL]

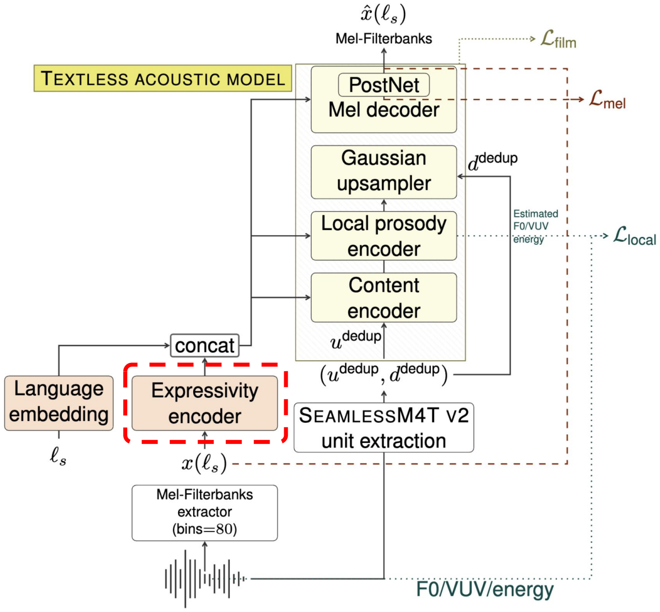
Discrete unit extractor

- Encode linguistic information of input speech
  - Input : Speech waveform
  - Output : Discretized XLS-R 10K units [Babu et al., 2022]
- Pretrained XLS-R model followed by K-means clustering
  - Align with SeamlessM4T unit extractor



[XLS-R model overview]

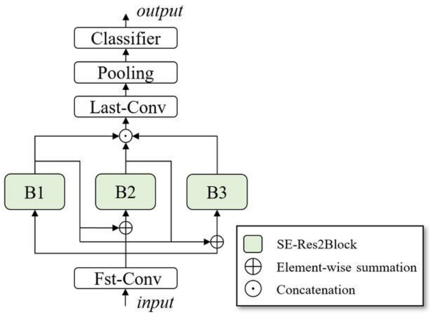
PRETSSEL: Paralinguistic REpresentation-based TextleSS acoustic modEL



[PRETSSEL]

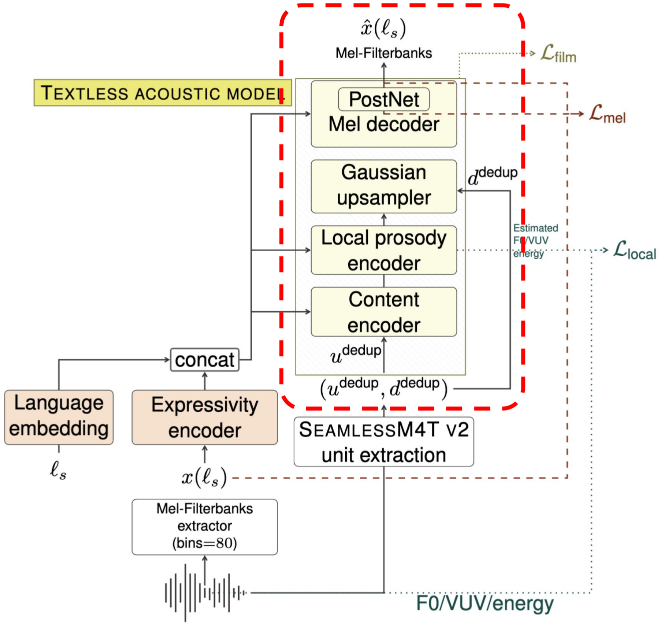
Expressivity encoder

- Encode paralinguistic information of input speech
  - Input : Mel-filterbank features
  - Output : Global expressivity embedding vector
- Modified ECAPA-TDNN architecture [Desplanques et al., 2020]
  - Replace batch norm. layer with layer norm. layer



[ECAPA-TDNN architecture]

PRETSSEL: **P**aralinguistic **RE**presentation-based **T**extle**SS** acoustic mod**EL**

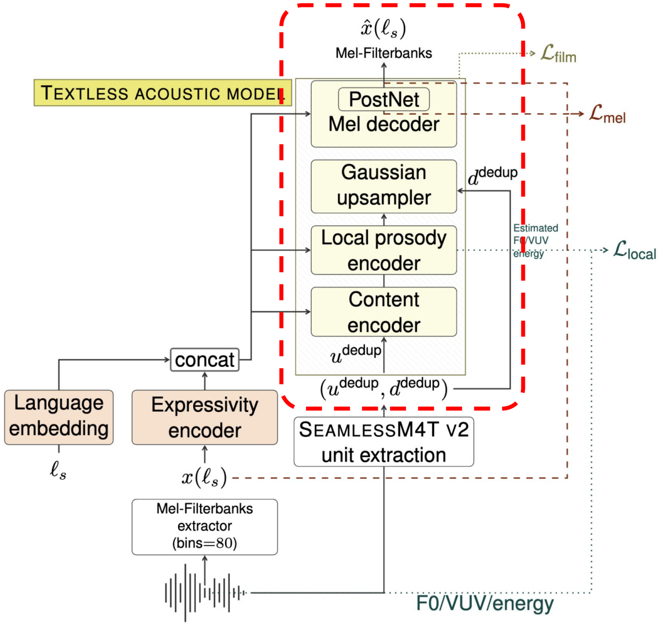


[PRETSSEL]

Textless acoustic model

- Synthesize speech from disentangled representations
  - Input : (1) XLS-R 10K units, (2) Expressivity embedding
    Linguistic ***Paralinguistic***
  - Output : Mel-filterbank features
- Modified FastSpeech2 architecture [Ren et. al., 2021]
  - Contents encoder
    - Encode unit representations
    - Feed-forward Transformer (FFT) blocks
  - Local prosody encoder
    - Predict and embed F0 and energy to encoder output
  - Mel-decoder
    - Predict output Mel-filterbank features
    - Feed-forward Transformer (FFT) blocks

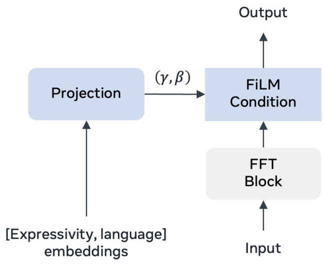
PRETSSEL: Paralinguistic REpresentation-based TextleSS acoustic modEL



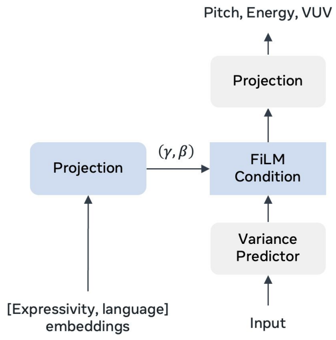
[PRETSSEL]

Textless acoustic model (cont.)

1. **FiLM conditioning layer** for better expressivity conditioning
   - Formula [Oreshkin et al., 2018]
     - $\text{FiLM}(x, c) = (\gamma + 1) \cdot x + \beta$
       $\gamma = \text{proj}(c) \cdot \theta_\gamma$
       $\beta = \text{proj}(c) \cdot \theta_\beta$
   - Apply FiLM to FFT blocks and prosody predictors
     - Use **Expressivity and language embeddings** as condition

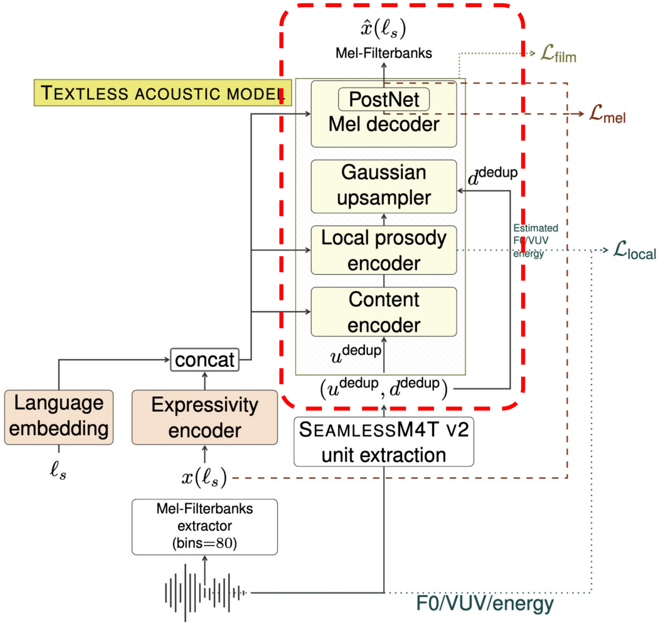


[FFT block with FiLM]



[Prosody predictor with FiLM]

PRETSSEL: **P**aralinguistic **RE**presentation-based **T**extle**SS** acoustic mod**EL**



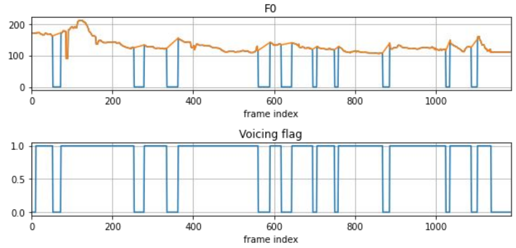
[PRETSSEL]

Textless acoustic model (cont.)

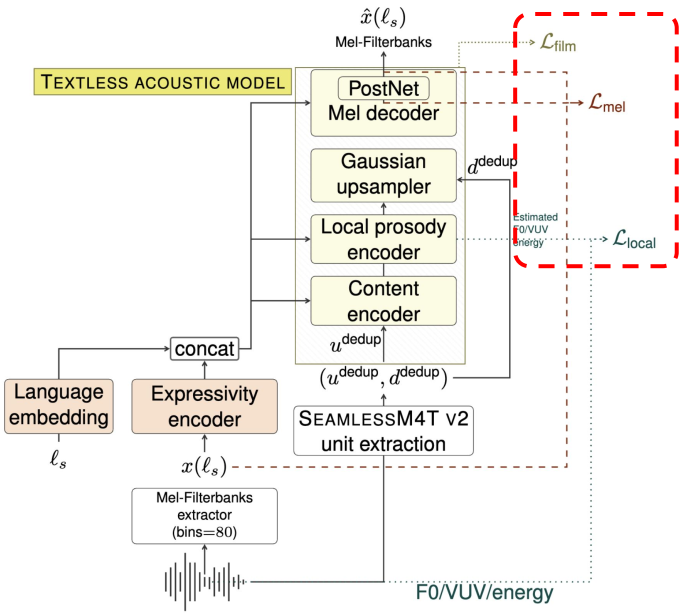
2. Duration modeling
   - **Obtain unit duration from external S2UT model**
   - Use **Gaussian upsampler** [Shen et al. 2020]

$$c_i = \frac{d_i}{2} + \sum_{j=1}^{i-1} d_j, \quad w_{ti} = \frac{\mathcal{N}\left(t; c_i, \sigma_i^2\right)}{\sum_{j=1}^{N} \mathcal{N}\left(t; c_j, \sigma_j^2\right)}, \quad \boldsymbol{u}_t = \sum_{i=1}^{N} w_{ti} \boldsymbol{h}_i.$$

3. Individual F0 and voicing flag (VUV) modeling
   - Interpolate F0 to **obtain continuous F0 and VUV,**
   - Predict them separately



PRETSSEL: **P**aralinguistic **RE**presentation-based **T**extle**SS** acoustic mod**EL**



[PRETSSEL]

Training criteria

$$\mathcal{L}_{total} = \mathcal{L}_{mel} + \lambda_l \cdot \mathcal{L}_{local} + \lambda_f \cdot \mathcal{L}_{film},$$

- Mel-reconstruction loss

  $$\mathcal{L}_{mel} = \mathcal{L}_1(\hat{y}_{before}, y) + \mathcal{L}_2(\hat{y}_{before}, y) + \mathcal{L}_1(\hat{y}_{after}, y) + \mathcal{L}_2(\hat{y}_{after}, y),$$

  - L1 and L2 losses of before and after PostNet

- Local prosody prediction loss

  $$\mathcal{L}_{local} = \mathcal{L}_2(\hat{p}, p) + \mathrm{BCE}(\hat{u}, u) + \mathcal{L}_2(\hat{e}, e),$$

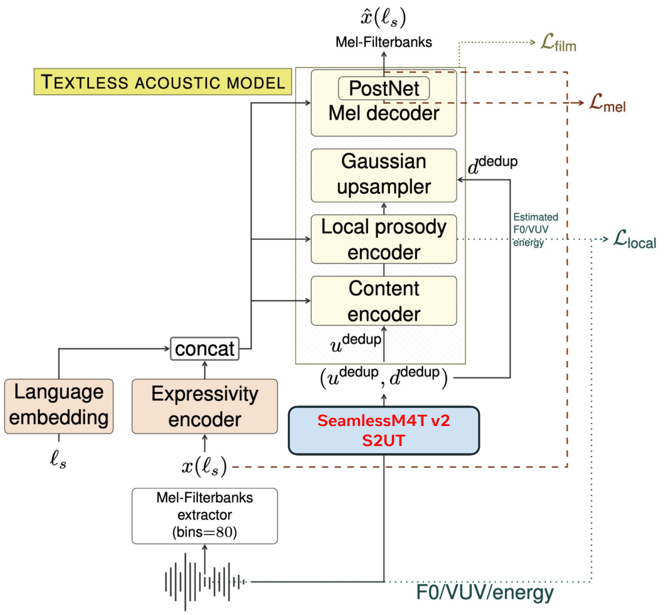
  - L2 losses for pitch and energy
  - Binary cross entropy loss for VUV

- FiLM regularization loss

  $$\mathcal{L}_{film} = \sum_{\theta_\gamma, \theta_\beta} (\theta_\gamma^2 + \theta_\beta^2),$$

  - L2 regularization for FiLM parameters

PRETSSEL: **P**aralinguistic **RE**presentation-based **T**extle**SS** acoustic mod**EL**

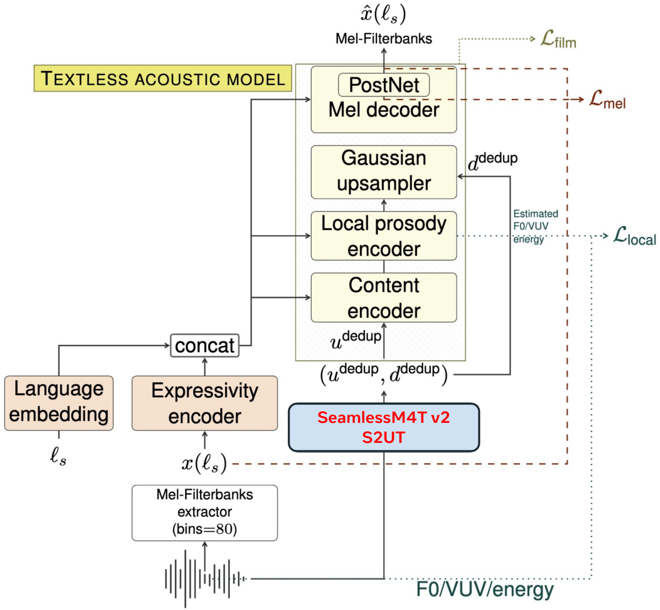


[PRETSSEL]

Expressive S2ST inference

1. Extract expressivity embedding at source language
   - Execute expressivity encoder
2. Obtain XLS-R 10K units at target language
   - Execute SeamlessM4T S2UT model
3. Generate Mel-filterbank features at target language
   - Execute textless acoustic model
4. Generate speech waveform from Mel-filterbank features
   - Execute HiFi-GAN vocoder [Kong et al., 2020]

PRETSSEL: **P**aralinguistic **RE**presentation-based **T**extle**SS** acoustic mod**EL**

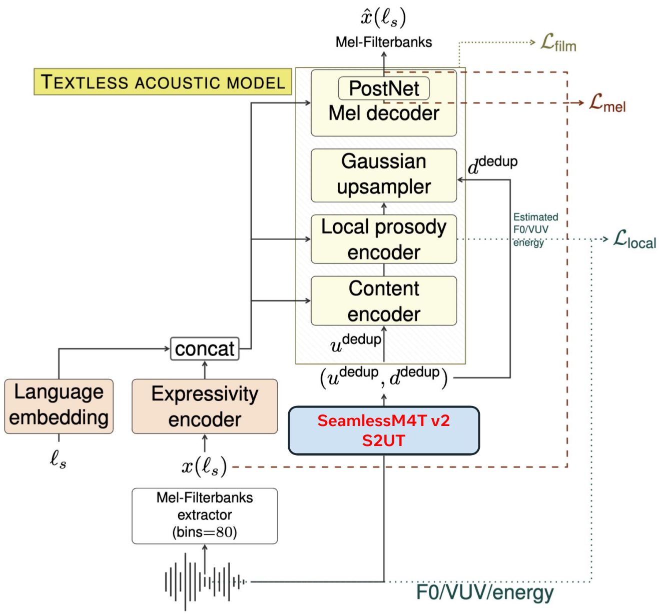


[PRETSSEL]

S2ST samples

	Source Speech	SeamlessM4T V2	SeamlessM4T V2 + PRETSSEL
Happy			
Sad			
Enunciated			

PRETSSEL: **P**aralinguistic **RE**presentation-based **T**extle**SS** acoustic mod**EL**



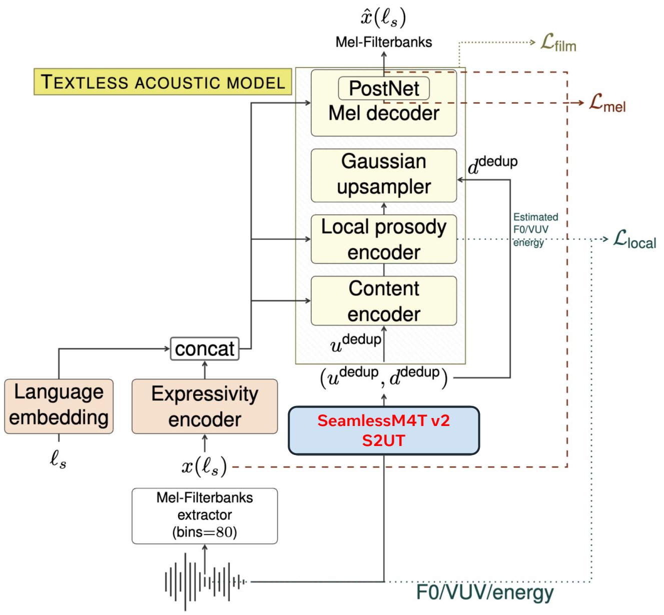
[PRETSSEL]

S2ST samples

	Source Speech	SeamlessM4T V2	SeamlessM4T V2 + PRETSSEL
Happy			
Sad			
Enunciated			

- PRETSEEL can clearly transfer source speech’s utterance-level expressivity!
    - e.g., vocal style or global emotion!

PRETSSEL: **P**aralinguistic **RE**presentation-based Textle**SS** acoustic mod**EL**



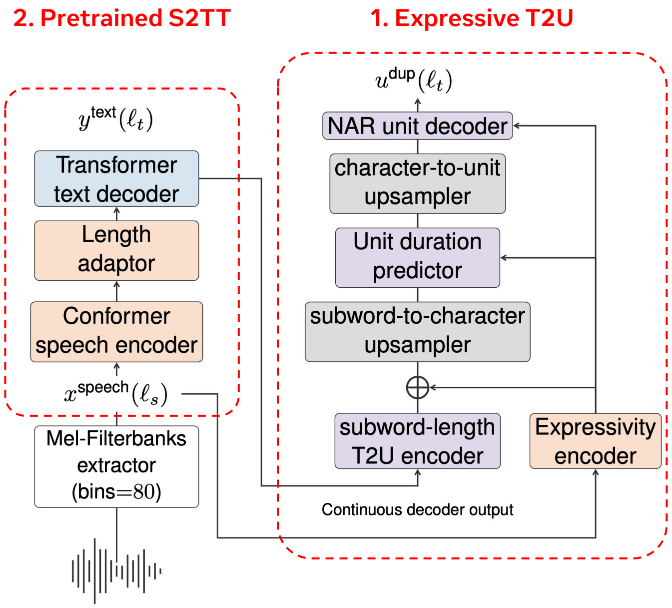
[PRETSSEL]

S2ST samples

	Source Speech	SeamlessM4T V2	SeamlessM4T V2 + PRETSSEL
Happy			
Sad			
Enunciated			

- PRETSEEL can clearly transfer source speech's utterance-level expressivity!
  - e.g., vocal style or global emotion!
- However, phrase-level expressivity are still missed in the output speech…
  - e.g., Rhythm, pause

Prosody UnitY2: Expressivity-aware S2UT model



[Prosody UnitY2]

Expressivity-variant of SeamlessM4T V2

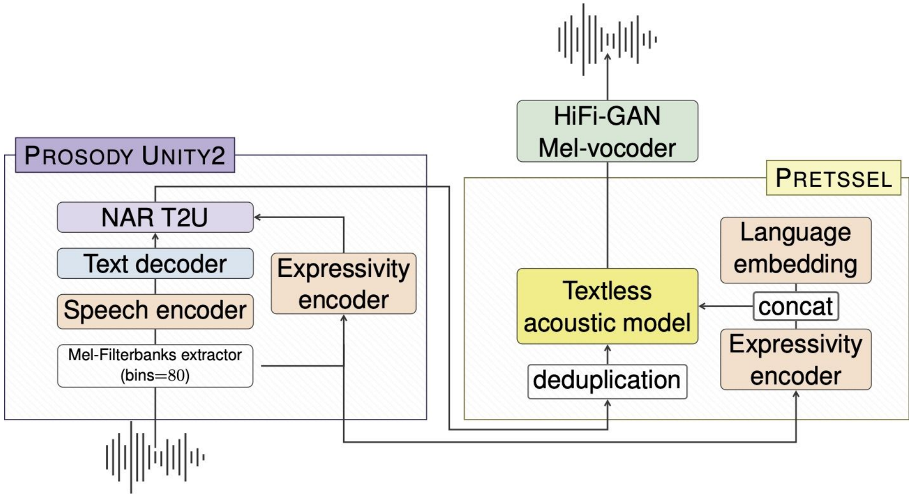
1. Expressive text-to-unit (T2U) architecture
   - Inject expressivity embedding to various positions of T2U model
     - i. Output of subword-length T2U encoder
     - ii. Unit-duration predictor using FiLM layer
     - iii. Unit decoder using FiLM layer

   ➡ ***Can transfer expressivity information of input speech!***

2. Fine-tuning from pretrained S2TT module
   - Initialize S2TT model from SeamlessM4T V2's S2TT model
   - Fine-tune entire S2UT modele using expressive S2ST data

   ➡ ***Can be efficiently trained!***

SeamlessExpressive: Expressivity-preserved Speech-to-Speech Translation



[SeamlessExpressive]

Performance Evaluation - Metrics

- **ASR-BLEU.** Content translation quality
- **V-SIM.** Vocal style preservation performance
- **AutoPCP.** Utterance-level prosody preservation performance
- **Rhythm.** Phrase-level prosody preservation performance
  - **Speech rate.** Spearman correlation of speech rates between two speeches
  - **Pause.** Pause alignment score

Performance Evaluation - Results

- **Eng to [Spa, Deu, Fra] translation**

Model	ASR-BLEU↑	V-Sim↑	AutoPCP↑	Speech rate↑	Pause↑
SeamlessM4T v2	38.82	0.05	2.31	0.13	0.14
SeamlessM4T v2 + PRETSSEL	38.59	0.27	2.87	0.15	0.16
SeamlessExpressive	40.18	0.28	3.19	0.64	0.39

- **[Spa, Deu, Fra] to Eng translation**

Model	ASR-BLEU↑	V-Sim↑	AutoPCP↑	Speech rate↑	Pause↑
SeamlessM4T v2	25.32	0.06	2.36	0.06	0.14
SeamlessM4T v2 + PRETSSEL	24.75	0.33	2.76	0.09	0.14
SeamlessExpressive	33.82	0.33	2.92	0.59	0.36

Performance Evaluation - Results

- **Eng to [Spa, Deu, Fra] translation**

Model	ASR-BLEU↑	V-Sim↑	AutoPCP↑	Speech rate↑	Pause↑
SeamlessM4T v2	38.82	0.05	2.31	0.13	0.14
SeamlessM4T v2 + PRETSSEL	38.59	**0.27**	**2.87**	0.15	0.16
SeamlessExpressive	40.18	0.28	3.19	0.64	0.39

- **[Spa, Deu, Fra] to Eng translation**

Model	ASR-BLEU↑	V-Sim↑	AutoPCP↑	Speech rate↑	Pause↑
SeamlessM4T v2	25.32	0.06	2.36	0.06	0.14
SeamlessM4T v2 + PRETSSEL	24.75	**0.33**	**2.76**	0.09	0.14
SeamlessExpressive	33.82	0.33	2.92	0.59	0.36

1. **Use of PRETSSEL dramatically improved utterance level expressivity preservation performance.**

Performance Evaluation - Results

- **Eng to [Spa, Deu, Fra] translation**

Model	ASR-BLEU↑	V-Sim↑	AutoPCP↑	Speech rate↑	Pause↑
SeamlessM4T v2	38.82	0.05	2.31	0.13	0.14
SeamlessM4T v2 + PRETSSEL	38.59	0.27	2.87	0.15	0.16
SeamlessExpressive	40.18	0.28	3.19	**0.64**	**0.39**

- **[Spa, Deu, Fra] to Eng translation**

Model	ASR-BLEU↑	V-Sim↑	AutoPCP↑	Speech rate↑	Pause↑
SeamlessM4T v2	25.32	0.06	2.36	0.06	0.14
SeamlessM4T v2 + PRETSSEL	24.75	0.33	2.76	0.09	0.14
SeamlessExpressive	33.82	0.33	2.92	**0.59**	**0.36**

1. Use of PRETSSEL dramatically improved utterance level expressivity preservation performance.
2. **Prosody UnitY2 dramatically improved phrase-level expressivity preservation performance.**

Performance Evaluation - Results

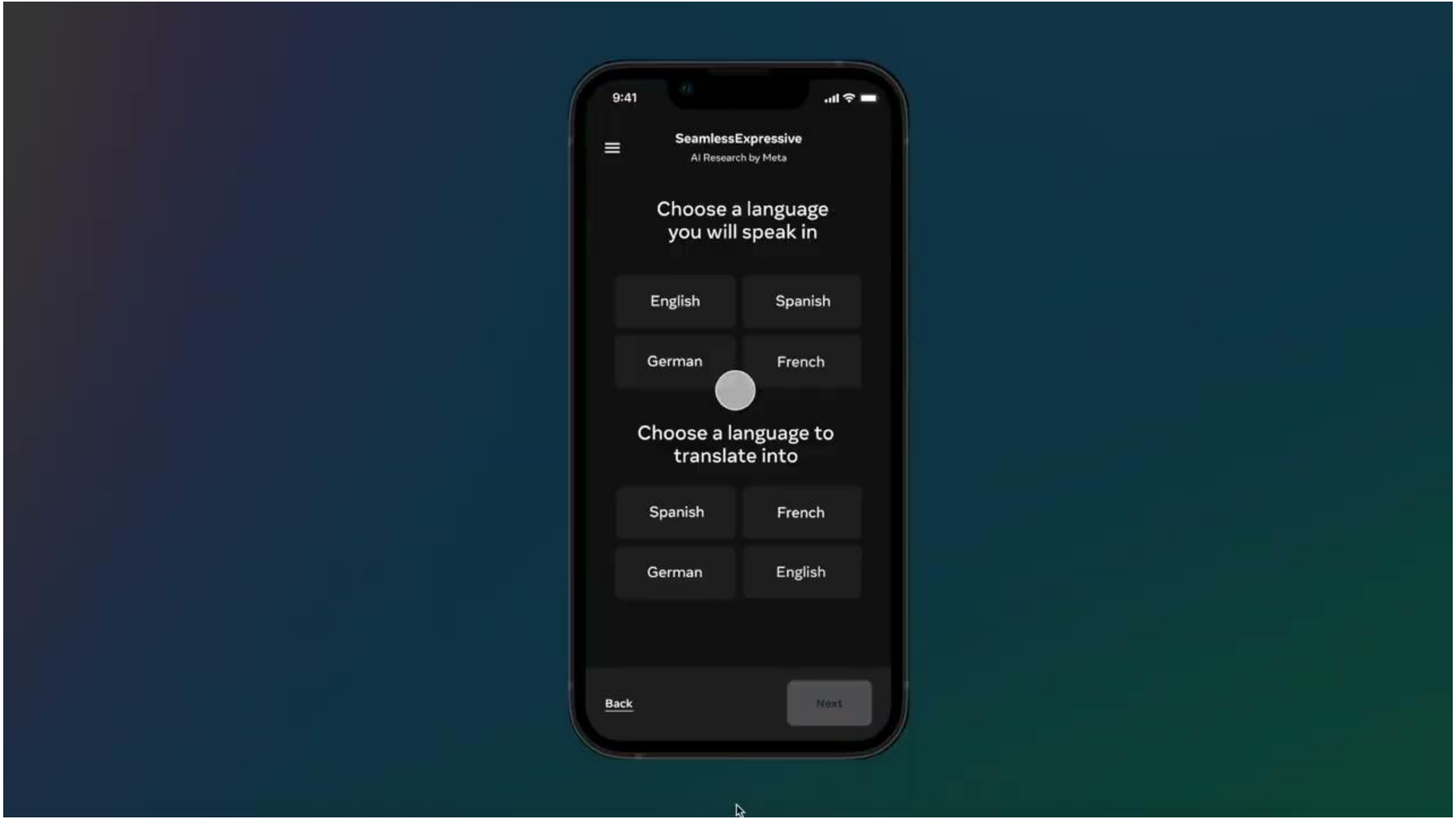
- **Eng to [Spa, Deu, Fra] translation**

Model	ASR-BLEU↑	V-Sim↑	AutoPCP↑	Speech rate↑	Pause↑
SeamlessM4T v2	38.82	0.05	2.31	0.13	0.14
SeamlessM4T v2 + PRETSSEL	38.59	0.27	2.87	0.15	0.16
SeamlessExpressive	**40.18**	**0.28**	**3.19**	**0.64**	**0.39**

- **[Spa, Deu, Fra] to Eng translation**

Model	ASR-BLEU↑	V-Sim↑	AutoPCP↑	Speech rate↑	Pause↑
SeamlessM4T v2	25.32	0.06	2.36	0.06	0.14
SeamlessM4T v2 + PRETSSEL	24.75	0.33	2.76	0.09	0.14
SeamlessExpressive	**33.82**	**0.33**	**2.92**	**0.59**	**0.36**

1. Use of PRETSSEL dramatically improved utterance level expressivity preservation performance.
2. Prosody UnitY2 dramatically improved phrase-level expressivity preservation performance.
3. **With PRETSSEL and Prosody UnitY2, SeamlessExpressive achieved best performances for all metrics.**


9:41
SeamlessExpressive
AI Research by Meta
Choose a language you will speak in
English
Spanish
German
French
Choose a language to translate into
Spanish
French
German
English
Back
Next

SeamlessExpressive Demo

References

[Lee et al., 2021] Ann Lee, et al. “Direct speech-to-speech translation with discrete units,” *in Proc. ACL*, 2021

[Inaguma et al., 2023] Hirofumi Inaguma et al. “Unity: Two-pass direct speech-to-speech translation with discrete units.,” *arXiv*, 2023

[Seamless Communication, 2023] Seamless Communication, “SeamlessM4T-Massively Multilingual & Multimodal Machine Translation.,” *arXiv*, 2023

[Babu et al., 2022] Arun Babu et al., “XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale,” *in Proc. Interspeech*, 2022

[Desplanques et al., 2020] Brecht Desplanques et al., “ECAPA-TDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification,” *in Proc. Interspeech*, 2020

[Ren et. al., 2021] Yi Ren et al., “FastSpeech 2: Fast and high-quality end-to-end text-to-speech,” *In Proc. ICLR*, 2021

[Oreshkin et al., 2018] Boris N. Oreshkin et al., “Tadam: Task dependent adaptive metric for improved few-shot learning.” *In Proc. NIPS*, 2018

[Kong et al., 2020] Jungil Kong et al., “HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis,” in Proc. *NeurIPS*, 2020

Meta