Breaking Language Barriers: Speech-to-Speech Translation for Unwritten Languages

Mitesh Kumar, M23MAC004
Sougata Moi, M23MAC008

Introduction and Application

- **Why focus on unwritten languages?**
  - 40% of the world's languages lack standardized writing systems, making text-based translation methods impractical.
  - Examples include many endangered languages and oral dialects like Taiwanese Hokkien.
- **Challenges faced by traditional S2ST systems:**
  - Heavy reliance on text-based intermediate steps (e.g., ASR, MT, and TTS).
  - Lack of parallel text corpora for training models.
  - Difficulty in capturing **tone, accent, and speaker characteristics** without effective representations.
- **The Hokkien Case Study:**
  - **Taiwanese Hokkien** is spoken by over 70% of Taiwan's population (~15.8 million).
  - Lacks a widely adopted writing system; tonal language with complex **tone sandhi** (tone changes based on context).
  - Existing approaches using synthetic data have limited success due to **low-resource nature** of this language pair.

Proposed Solution and Real-World Importance

Proposed Solution:

- Instead of text, use **discrete unit-based representations** that encode linguistic and **non-linguistic speech features**.
- Bypasses the need for parallel text by translating speech directly into speech.

Real-world importance:

- Enables communication in **multilingual, low-resource environments** (e.g., rural and oral-first communities).
- Preserves oral languages by integrating them into digital ecosystems, promoting **inclusivity and access**.

Speech-to-Speech Translation For A Real-world Unwritten Language

Dataset

Training Data:

- **Human Annotated Data:**
  - Hokkien→English: 61.4 hours total
  - English→Hokkien: 35 hours English + 51 hours Hokkien
  - Sources: Hokkien dramas, TAT dataset, MuST-C
- **Weakly Supervised Data:**
  - English→Hokkien: 1.5k hours from Librispeech and TED-LIUM3
  - Hokkien→English: 8k hours from Hokkien dramas
- **Mined Data:**
  - Hokkien→English S2T: 8.1k hours
  - English↔Hokkien S2ST: 197 hours

Test Data:

- 1.47 hours of English speech – 10 speakers (5 male, 5 female) for English
- 1.42 hours of Hokkien speech – 4 speakers (2 male, 2 female) for Hokkien

Model Architecture

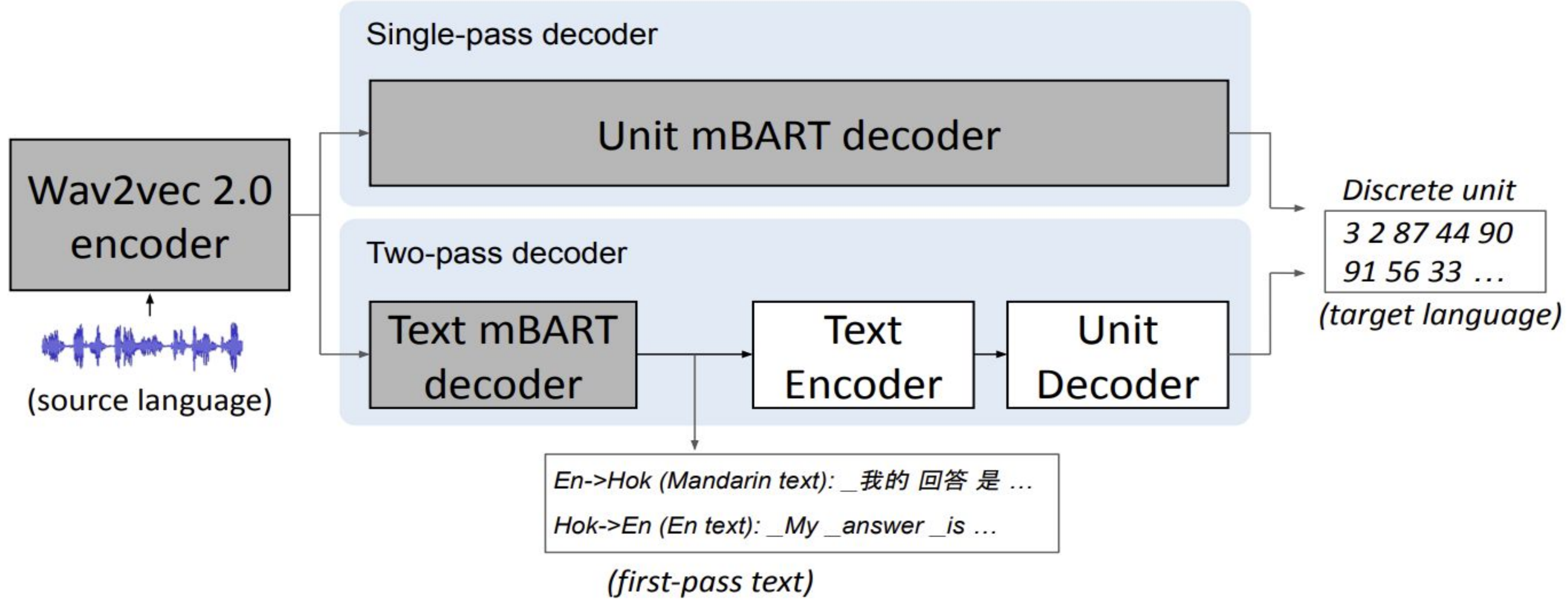

Figure 1: Model architecture of S2ST with single-pass and two-pass decoder. The blocks in shade illustrate the modules that are pre-trained. Text in italic is the training objective.

Model Architecture : Discrete Unit to Speech

Overview of the Model:

The S2ST system consists of two core components:

1. **Sequence-to-sequence S2UT model:** Converts input source speech into target discrete units.
2. **HiFi-GAN Vocoder:** Converts discrete unit sequences into target Speech waveforms.

Single-Pass Decoding (S2UT):

Key Components:

- **wav2vec 2.0 encoder:**
  - Extracts 80-dim log-mel filterbank features and transforms input speech into meaningful representations.
  - Pre-trained on large speech datasets (e.g., LibriSpeech) using self-supervised learning.
- **unit mBART decoder:**
  - Converts the encoded speech representations into sequences of discrete units.
  - Pre-trained using cross-entropy loss on unit sequences extracted via HuBERT clustering.

Single-Pass Decoding (S2UT):

How Single-Pass Decoding Works:

- Source speech is passed through the **wav2vec 2.0 encoder**, generating latent speech representations.
- These representations are directly decoded by the **unit mBART decoder** into discrete units that represent the target language speech.
- The output discrete units are passed through a **HiFi-GAN vocoder** to generate the target waveform.

Two-Pass Decoding (UnitY):

Key Components:

- **wav2vec 2.0 encoder:** Same as in the single-pass model, responsible for encoding input speech into latent representations.
- **Text mBART decoder:**
    - Pre-trained on large corpora of Mandarin and English text.
    - Generates intermediate text representations based on the source speech.
- **Text Encoder:** Processes the intermediate text into a format compatible with the unit decoder.
- **Unit Decoder:** Converts the intermediate text representations into discrete units of target language speech.
- **HiFi-GAN vocoder:** Converts the discrete units into the target speech waveform.

Two-Pass Decoding (UnitY):

How Two-Pass Decoding Works:

- The input speech is processed by the **wav2vec 2.0 encoder** to generate latent speech representations.
- These representations are fed into the **text mBART decoder**, which predicts an intermediate text
- The **text encoder** processes the predicted text, and the **unit decoder** converts it into a sequence of discrete units.
- The discrete units are then passed through the **HiFi-GAN vocoder** to generate the final target waveform.

Evaluation

Table 3: Dev / test ASR-BLEU on TAT-S2ST dataset. (*: synthetic Hokkien speech is generated by applying unit vocoder on the normalized units extracted from the ground truth Hokkien speech in TAT-S2ST, while synthetic En speech is generated by applying En T2U followed by the unit vocoder on the ground truth En text.)

		En→Hokkien				Hokkien→En			
		Training data		ASR-BLEU		Training data		ASR-BLEU	
ID	Model	Human (35-hr)	Weakly (1.5k-hr)	Dev	Test	Human (61.4-hr)	Weakly (8k-hr)	Dev	Test
Cascaded systems:									
1	Three-stage	✓	✓	7.5	6.8	✓	✓	9.9	8.8
2	Two-stage	✓	✓	7.1	6.6	✓	✓	12.5	10.5
Single-stage S2UT systems:									
3	Single-pass decoding	✓	✗	0.1	0.1	✓	✗	0.1	0.1
4	Single-pass decoding	✓	✓	6.6	6.0	✓	✓	8.8	8.1
5	Two-pass decoding (UnitY)	✓	✗	0.9	0.4	✓	✗	4.2	3.8
6	Two-pass decoding (UnitY)	✓	✓	**7.8**	**7.3**	✓	✓	**13.6**	**12.5**
7	Synthetic target*	✗	✗	55.5	53.4	✗	✗	76.2	78.5

Conclusion & Future Scope

Conclusion:

- This study developed the **first English ↔ Hokkien S2ST system** targeting an unwritten language.
- Demonstrated the **effectiveness of combining human-annotated, weakly supervised, and mined data** in low-resource settings.
- Showed that **two-pass decoding leveraging high-resource intermediate languages** (like Mandarin) significantly improves translation accuracy.
- Highlighted the potential of **discrete unit-based approaches** for preserving and translating oral languages.

Conclusion & Future Scope

Future Scope:

- **Support for diverse languages:** Expand the model's applicability to other unwritten and endangered languages and we are still using a high resource language as reference, we need to reduce the importance of the reference language in the whole training process.
- **Real-time implementation:** Work towards building efficient S2ST models suitable for real-time applications.
- **Robust domain adaptation:** Ensure models perform consistently across varied real-world conditions and domains.

Thank You