Speech-to-Speech Translation for a Real-world Unwritten Language

Sougata Moi (M23MAC008)

Mitesh Kumar (M23MAC004)

Link: https://arxiv.org/pdf/2211.06474

The paper focuses on Speech-to-Speech Translation (S2ST) for unwritten languages, using English-to-Taiwanese Hokkien as a case study. The research addresses challenges such as the lack of standard writing systems and limited training data. It presents an end-to-end S2ST system incorporating data collection, modelling, and benchmarking, emphasizing self-supervised learning and text supervision techniques.

Dataset

The dataset used for training and evaluation consists of three main components:

- **Human-annotated data:** A supervised dataset created through bilingual speakers transcribing and translating English-Hokkien speech pairs. The dataset consists of 61.4 hours of speech data for Hokkien-to-English translation and 35 hours for English-to-Hokkien translation.
- **Weakly supervised data:** Pseudo-labeled data generated through automatic translation models and speech-to-text techniques. This includes 1.5k hours of English speech converted into Hokkien speech using machine translation, and 8k hours of Hokkien speech automatically translated into English.
- **Mined data:** Large-scale corpus obtained by mining parallel speech pairs from unlabeled multilingual speech datasets. This approach resulted in 8.1k hours of mined Hokkien-to-English data and 197 hours of English-to-Hokkien S2ST data.

For **evaluation**, the study introduces the **TAT-S2ST benchmark dataset**, which consists of manually annotated speech pairs for both English and Hokkien. The dataset includes:

- A **development set** containing 1.62 hours of English speech and 1.46 hours of Hokkien speech from 10 speakers.
- A **test set** containing 1.47 hours of English speech and 1.42 hours of Hokkien speech from 10 different speakers.
- Reference transcriptions in Tâi-lô (a romanization system for Hokkien) to enable automatic evaluation using speech recognition.

The dataset is designed to support rigorous benchmarking and facilitate future research in speech-to-speech translation for unwritten languages.

Model Architectures and Training

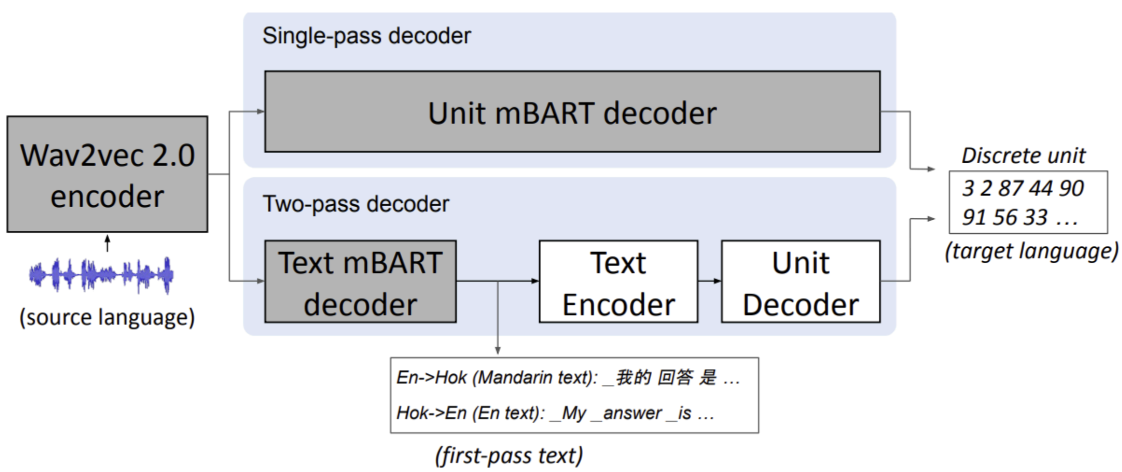

Figure 1: Model architecture of S2ST with single-pass and two-pass decoder. The blocks in shade illustrate the modules that are pre-trained. Text in italic is the training objective.

Single-Pass Decoding (S2UT)

- **Overview:** The single-pass decoding model directly translates source speech into target speech without using intermediate text representations. It uses discrete units extracted through clustering techniques on speech representations.
- **Architecture Components:**
  - **wav2vec 2.0 encoder:**
    - Pre-trained on large amounts of unlabeled speech data.
    - Extracts 80-dimensional log-mel filterbank features from input speech.
  - **unit mBART decoder:**
    - Converts encoded speech into target discrete units using cross-entropy loss.
    - The decoder is pre-trained on a multilingual dataset using mBART with discrete unit targets.
- **Training Process:**
  - The encoder-decoder system is trained end-to-end using a combination of human-annotated, weakly supervised, and mined data. The model optimizes for discrete unit prediction using cross-entropy loss.
- **Limitations:**
  - Lacks intermediate supervision from text, making it prone to errors in low-resource conditions.
  - Struggles with complex tonal changes and speaker variability.

Two-Pass Decoding (UnitY)

- **Overview:** The two-pass decoding architecture addresses the limitations of the single-pass model by introducing an intermediate text prediction stage. This additional supervision helps improve generalization and translation accuracy, particularly for tonal languages like Hokkien.

- **Architecture Components:**
  - **wav2vec 2.0 encoder:**
    - Same as in the single-pass model, encoding input speech into meaningful representations.
  - **Text mBART decoder:**
    - Predicts intermediate text sequences (Mandarin) from encoded speech representations.
    - Pre-trained using large bilingual corpora of Mandarin and English.
  - **Text encoder:**
    - Two randomly initialized Transformer layers refine the intermediate text output.
  - **Unit decoder:**
    - Converts intermediate text into target discrete units for the final speech synthesis.
  - **HiFi-GAN vocoder:**
    - Converts discrete unit sequences into the final speech waveform.
- **Training Process:**
  - The model is trained using a combination of human-annotated and weakly supervised data. The use of intermediate text helps bridge the gap between low-resource Hokkien and high-resource Mandarin.

strengths and Limitations

Strengths:

1. **Direct Speech-to-Speech Translation:**
   - Avoids the dependence on text-based intermediate representations, making it suitable for unwritten languages.

2. **Comprehensive Data Strategy:**
    ○ Combines **human-annotated, weakly supervised, and mined data** to overcome low-resource constraints.
3. **Intermediate Language Supervision:**
    ○ Leverages Mandarin as a high-resource language, which significantly boosts performance.
4. **Innovative Architecture:**
    ○ Demonstrates the effectiveness of **discrete units** as an intermediate representation in translation systems.

Limitations:

1. **Data Scarcity:**
    ○ The system still depends on collecting large amounts of data, which can be challenging for very rare languages.
2. **Noise in Mined Data:**
    ○ Automatically mined data often contain inaccuracies, affecting the model's robustness.
3. **Domain Generalization:**
    ○ The system's performance may degrade when used in **domains different from the training data**.
4. **Latency Issues:**
    ○ Real-time deployment of the system may require additional optimizations to reduce inference time.

Evaluation and Results

Evaluation Methodology:

The model was evaluated using **ASR-BLEU (Automatic Speech Recognition BLEU)**, which measures how well the transcribed output matches the reference text. BLEU scores provide a proxy for evaluating translation accuracy in the absence of text-based training.

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

Table 4: Results of En→Hokkien models trained with mined En↔Hokkien S2ST data. We report dev / test ASR-BLEU on TAT-S2ST dataset.

		Training data			ASR-BLEU	
ID	Model	Human (35-hr)	Weakly (1.5k-hr)	Mined (197-hr)	Dev	Test
3	Single-pass decoding	✓	✗	✗	0.1	0.1
8		✓	✗	✓	0.1	0.1
4		✓	✓	✗	6.6	6.0
9		✓	✓	✓	6.7	6.0
5	Two-pass (UnitY)	✓	✗	✗	0.9	0.4
10		✓	✗	✓	5.7	4.9
6		✓	✓	✗	7.8	7.3
11		✓	✓	✓	8.0	7.5

Analysis of Results

- The **ASR-BLEU scores** help measure how well the generated speech matches the reference speech.
- Training with only human-annotated data resulted in low BLEU scores, showing that pre-training alone is not enough.
- Adding weakly supervised data significantly improved the scores, with gains between **5.9 and 8.7 BLEU points**.
- The **UnitY model** performed better than the **single-pass S2UT model** in both directions.
    - In **English-to-Hokkien**, UnitY was **1.3 BLEU points higher**, proving that using related text (Mandarin) helps.
    - In **Hokkien-to-English**, UnitY outperformed S2UT by **4.4 BLEU points**, likely due to more available training data.
- Cascaded baseline models (multi-step translations) were tested:
    - The **two-stage system** was better or similar to the **three-stage system**.
    - The **best-performing one-stage system (UnitY)** surpassed the two-stage system by **0.7 BLEU in English-to-Hokkien** and **4.4 BLEU in Hokkien-to-English**.
    - This highlights the benefits of training the translation and speech generation steps together.
- **Leveraging Mined En↔Hokkien S2ST Data (En→Hokkien Direction)**
    - UnitY model trained using **Hokkien→Zh S2T** for pseudo-labeled Mandarin text as an auxiliary task.
    - Single-pass decoding **S2UT model** has low BLEU score (row 8).
    - **UnitY model improves by 4.5 BLEU** with 197-hr mined S2ST data (row 5 vs. 10).
    - Noisy pseudo-labeled Mandarin text **still benefits training**.
    - Combining with **weakly supervised data does not show significant gain** (row 4 vs. 9, 6 vs. 11).
    - Mined data is **only 13%** of total weakly supervised data, limiting its impact.

- **Effect of Mined Hokkien→En S2T Data Converted to S2ST**
  - Mined **Hokkien→En S2T** data converted using **En T2U model** for UnitY training.
  - **4.7k-hr mined data** (t = 1.065) improves model by **3.6 BLEU** with only human annotated data.
  - **8.1k-hr mined data** (t = 1.06) provides only **0.9 BLEU gain**.
  - **7.8 BLEU gap** exists between mined data model and UnitY model trained with **human annotated + 8k-hr weakly supervised data**.
  - **Both weakly supervised and mined data come from Hokkien dramas dataset**, highlighting pseudo-labeling's effectiveness.
  - **Mined data quality limits improvements**, but combining all three data types is still beneficial.
  - Adding **8.1k-hr mined data** to human annotated + weakly supervised data yields **1.6 BLEU gain**.

Future Opportunities:

1. **Support for Diverse Languages:**
   - Expand the model's applicability to other unwritten and endangered languages, addressing the need to reduce dependence on high-resource reference languages.
2. **Real-Time Applications:**
   - Developing optimized models for **low-latency, real-time S2ST applications**.
3. **Domain Adaptation Techniques:**
   - Incorporating domain-specific fine-tuning to improve performance in various use cases (e.g., healthcare, emergency services).
4. **Contextual and Stylistic Adaptation:**
   - Improving the system's ability to capture **intonation, speaker style, and**
   - **contextual meaning** beyond word-for-word accuracy.

Citation

1. A. Lee et al., "Direct speech-to-speech translation with discrete units," *arXiv preprint arXiv:2107.05604*, 2021. [Online]. Available: https://arxiv.org/abs/2107.05604
2. C. Zhang et al., "UWSpeech: Speech to Speech Translation for Unwritten Languages," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 16, pp.

14319–14327, 2021. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/17684
3. A. Tjandra, S. Sakti, and S. Nakamura, "Speech-to-speech translation between untranscribed unknown languages," *arXiv preprint arXiv:1910.00795*, 2019. [Online]. Available: https://arxiv.org/abs/1910.00795
4. P.-J. Chen et al., "Speech-to-Speech Translation For A Real-world Unwritten Language," *arXiv preprint arXiv:2211.06474*, 2022. [Online]. Available: https://arxiv.org/abs/2211.06474
5. A. Lee et al., "Direct Speech-to-Speech Translation with Discrete Units," in *Proc. Interspeech 2021*, 2021, pp. 3451–3455. [Online]. Available: https://www.isca-speech.org/archive/interspeech_2021/lee21c_interspeech.html