Speech Understanding Assignment 1

Shaik Mohammed Touseef (B21ME061)
Shubham Mishra (B21ME065)

1 Introduction

Speech-to-Speech Translation (S2ST) involves converting spoken language from one language into spoken output in another.

2 Important Concept for this Problem Statement

Below is an overview of these concepts:

2.1 Vocoding and Vocoders

Vocoding is a technique that analyzes and synthesizes the human voice by decomposing speech signals into their fundamental components. A vocoder (voice encoder) achieves this by extracting features like pitch, tone, and formant frequencies, allowing for the manipulation and reconstruction of speech signals. In S2ST systems, vocoders are crucial for converting intermediate representations, such as spectrograms, back into audible speech.

HiFi-GAN is a state-of-the-art vocoder that utilizes Generative Adversarial Networks (GANs) to produce high-fidelity audio. It is known for its efficiency and ability to generate natural-sounding speech, making it a popular choice in modern speech synthesis applications.

arxiv.org

2.2 Isochrony Preservation

Isochrony refers to the rhythmic timing of speech, ensuring that syllables and words occur at regular intervals. In S2ST, preserving isochrony is vital to maintain the natural flow and timing of the original speech in the translated output. This involves aligning the translated speech to match the duration and rhythm of the source speech, which is essential for applications like dubbing in multimedia content. Models like **TransVIP** have been developed to address this by incorporating mechanisms to control isochrony during translation.

arxiv.org

2.3 Speech Quantization

Speech quantization involves transforming continuous speech features into discrete tokens, facilitating efficient processing and modeling in S2ST systems. This process is typically achieved through self-supervised learning (SSL) methods and results in two main types of tokens:

- **Semantic Tokens**: These capture the linguistic content of speech, focusing on the meaning conveyed by the speaker.
- **Acoustic Tokens**: These represent the acoustic properties of speech, such as intonation, stress, and rhythm.

By quantizing speech into these tokens, S2ST systems can more effectively model and translate spoken language, preserving both the semantic and acoustic characteristics of the original speech.

3 Importance in Real-World Applications

3.1 International Politics and Business

In **parliaments, multinational corporations, and diplomacy**, real-time negotiations require multilingual communication.

Example: The translation of **Palestinian President Mahmoud Abbas's speech at the UN General Assembly** was crucial for accurately conveying his political stance on Palestine's bid for non-member observer status. Source

3.2 Education

S2ST systems help people who speak languages that don't have a standard way of writing. These systems make it easier for students who speak different languages to learn and take classes.

Example: Taiwanese Hokkien, which lacks a widely accepted writing system, has seen research-driven S2ST development to assist in education. ACL Anthology

3.3 Healthcare and Disaster Response

When emergencies happen, language barriers can make it hard for medical teams and first responders to help people effectively. S2ST systems can quickly translate between languages to help save lives.

Example: The **COVID-19 pandemic** highlighted the need for multilingual communication in healthcare, particularly in hospitals treating diverse populations. ACL Anthology

4 State-of-the-Art Models in S2ST

4.1 Meta's SeamlessM4T

- **Approach**: Massively **multilingual and multimodal machine translation** for speech and text.
- **Strengths**: Supports **101+ languages**; maintains **emotions and speaker characteristics** during translation.
- **Results**: **26.1 ASR-BLEU** on FLEURS (eng–X), outperforming cascaded models.
- **Limitations**: Challenges in **proper noun recognition, accents, and gender bias**.

Paper

4.2 Google's Translathon-2 and Translathon-3

- **Approach**: Direct **S2ST with improved phoneme-based decoding**.
- **Strengths**: More **efficient decoding**, **bridging the performance gap** between direct and cascaded systems.
- **Limitations**: Still lags behind **two-stage cascaded systems** in **accuracy**.

Microsoft Research

4.3 TransVIP

- **Approach**: TransVIP employs a consecutive generation approach, simplifying the complex S2ST task into two sequential tasks while maintaining an end-to-end framework.
- **Results**: Matches the score of ASR-BLEU for cascaded systems while preserving speaker voice.
- **Limitations**: It is not a direct speech-to-speech model but a cascaded system.

4.4 Textless Speech Translation

- **Approach**: Methods like MSLM-S2ST use a single decoder-only autoregressive model for semantic translation and acoustic generation.
- **Results**: Scored **24.78 ASR-BLEU** on Spanish-English, lagging behind Enc-Dec models by 2 points.
- **Limitations**: ASR-BLEU gaps indicate weaker translation accuracy.

5 Metrics used in various research studies

5.1 ASR-BLEU (Automatic Quality)

What it measures: BLEU (Bilingual Evaluation Understudy) evaluates machine-translated text quality by comparing it to reference translations. In this context, it measures the quality of transcriptions generated by Automatic Speech Recognition (ASR) systems.

BLEU calculates the precision of n-grams in the generated text compared to the reference text and includes a brevity penalty:

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{1}$$

where BP is the brevity penalty, w_n is the weight for n-grams, and p_n is the precision of n-grams.

5.2 ASR-chrF (Automatic Bias)

What it measures: chrF (Character F-score) evaluates text quality by comparing character n-grams between generated and reference texts. It measures bias in transcriptions.

chrF is calculated as the harmonic mean of precision and recall for character n-grams:

$$chrF = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{2}$$

where precision is the ratio of correctly predicted character n-grams to the total predicted, and recall is the ratio of correctly predicted character n-grams to the total in the reference.

5.3 Blaser 2.0 (Automatic Quality and Model-based Bias)

What it measures: Blaser 2.0 is a composite metric evaluating both quality and bias in generated text. It may combine aspects of BLEU, chrF, and other metrics for a comprehensive assessment.

5.4 XSTS (Human Quality)

What it measures: XSTS (Cross-lingual Semantic Textual Similarity) evaluates semantic similarity between generated and reference texts, often assessed by human evaluators. XSTS is typically computed based on human-rated semantic similarity scores, usually on a scale (e.g., 1 to 5).

5.5 MOS (Human Naturalness)

What it measures: MOS (Mean Opinion Score) is a subjective metric used to evaluate synthesized speech naturalness, rated by human listeners.

MOS is computed as the average of human ratings, typically on a scale from 1 (poor) to 5 (excellent):

$$MOS = \frac{1}{N} \sum_{i=1}^{N} R_i \tag{3}$$

where R_i is the rating from the i-th evaluator, and N is the total number of evaluators.

5.6 ASR-ETOX (Automatic Toxicity)

What it measures: ETOX evaluates text toxicity by detecting harmful or inappropriate language. ETOX typically involves a classifier scoring the text for toxic language. The exact formula depends on the classifier used, but it generally outputs a probability or toxicity score.

6 Challenges and Future Directions

6.1 Open Problems

- **Low-Resource Language Support**: Many languages have **little or no training data**.
- **Accent Recognition**: Current models struggle with **regional accents and dialects**. Some research addresses this via speech normalization.
- **Speech Quality Preservation**: Maintaining **voice, tone, and emotional expressiveness** remains challenging.

6.2 Future Opportunities

- **Data Augmentation for Low-Resource Languages**: Techniques like **transfer learning** can improve translations.
- **Improved Evaluation Metrics**: Better tools are needed to assess **emotions, fluency, and real-time performance**.
- **Self-Supervised Learning**: Reducing reliance on **large labeled datasets**.

7 Spectrograms and Windowing Techniques

Here, the results and such of the second part are compiled. Link to GitHub.

7.1 Windowing analysis

With respect to different windowing techniques (keeping window size and hop size constant):

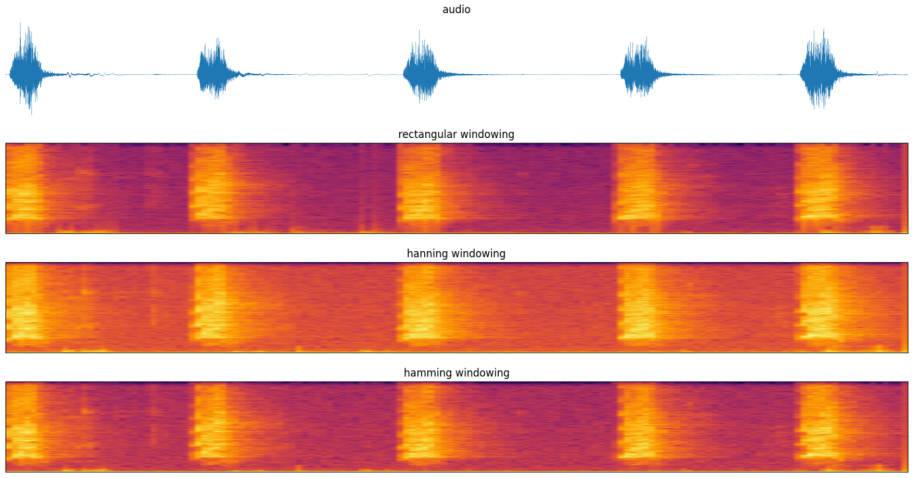

Figure 1: Different windowing techniques on an audio

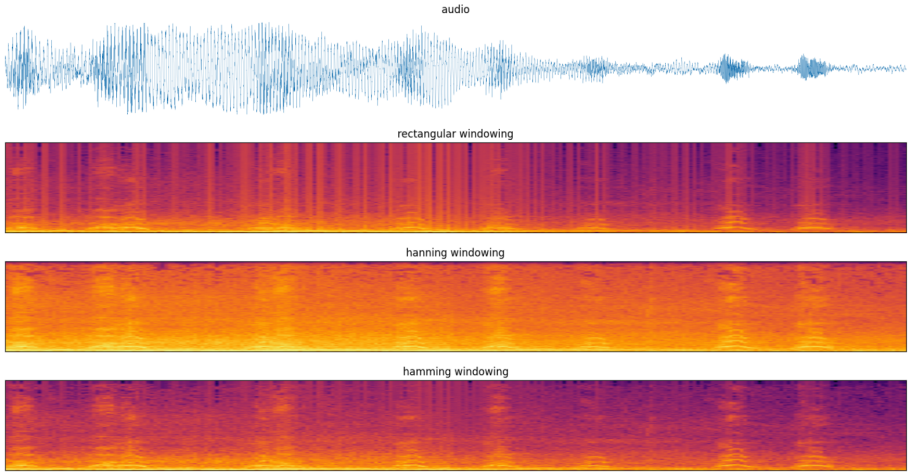

Figure 2: Enter Caption

As can be seen, hamming windowing leads to a balance between resolution and leaky suppression, while hanning windowing leads to more of the suppression, and rectangular windowing to that of the resolution. An interactive version and more analysis can be seen in the notebook on Github.

7.2 Music analysis

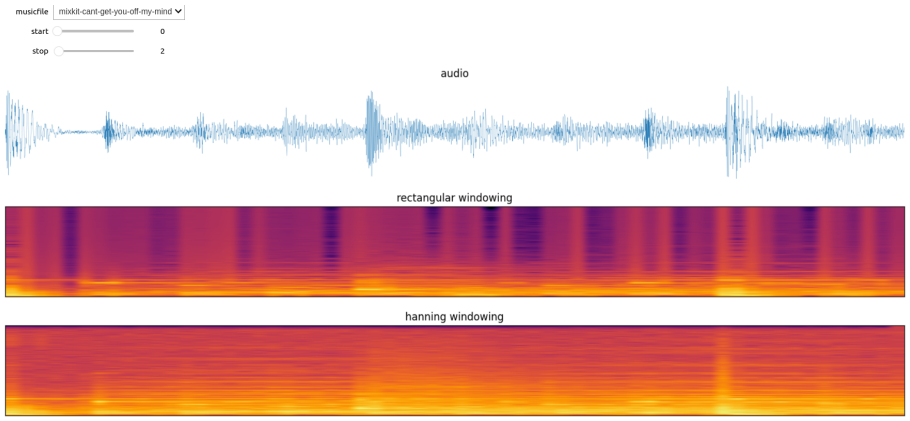

Figure 3: Music spectrogram

Again, interactive version on Github.