Voice-Changing Translator

Jinglan Yang (121090699)
School of Data Science
Chinese University of Hong Kong, Shenzhen
121090699@link.cuhk.edu.cn

Abstract

This report proposes a Voice-Changing Translator, which could translate the content spoken by the user into the target language, and then speak the translated content in the user's voice. This voice changing translator aims to help people communicate and learn foreign language pronunciation.

1 Introduction

Speech can be broken down into three factors: spoken language content, speaker timbre and speech prosody. Everyone's timbre and prosody are very different. With familiar people, we can identify who is speaking even just by hearing their voice. From the voice, we can infer a person's age, gender, personality, where they are from, etc. However, the number of languages a person can speak are limited. If a person goes to a country or region where he or she does not understand the local language, it will be difficult for him or her to communicate with the locals. Currently there are only translators on the market that realizing text-to-text, speech-to-text, or text-to-speech. This report proposes a voice-changing translator: the user uses his or her own voice and language to speak what he or she wants to say to the voice-changing translator. The voice-changing translator can use the user's voice to speak in the target language, thereby helping people communicate better.

There are two types of target users of the voice-changing translator: first, for users who travel or work in countries and regions where he or she does not understand local language, the voice-changing translator can help these users communicate with local residents and reduce the user's burden of learning the local language; And, because the voice-changing translator can imitate the user's voice, it can not only help convey spoken content, but also imitate the timbre and prosody of the user's voice. Second, help people learning foreign languages. Learning a foreign language requires not only learning grammar, but also learning pronunciation. For example, if a user wants to learn Scottish-accented English, the voice-changing translator can imitate the sound of the user speaking Scottish-accented English. When the user hears the Scottish-accented English voice spoken by "himself" or "herself", the user will be able to learn how to pronounce the Scottish-accented English faster, gaining a deeper understanding and learn foreign language pronunciation more efficiently.

The voice-changing translator proposed in this paper can be applied to translation software and foreign language learning software.

2 Related Work

Previous research has focused on developing robust speech recognition and machine translation systems. Transformer is the state-of-the-art model in recent machine translation evaluations. Deep transformer (Wang, etc; 2019), weighted transformer network (Ahmed, etc; 2017), and multimodal transformer (Yao, etc; 2020) have been proposed to improve the performance. Studies in voice conversion and synthesis have explored methods for modifying speaker characteristics while preserving speech content. Techniques such as Adaptive Wavenet Vocoder for Residual Compensation in GAN-Based Voice Conversion (Sisman, etc; 2018), Auto-Encoder VC like PitchNet (Deng, etc; 2018) have demonstrated promising results in transforming voices across languages and accents. The

voice-changing translator proposed in this paper combines speech recognition, machine translation, and voice conversion technologies, realizing user's voice multi-language speech to speech conversion.

3 Approach

The function implementation of the voice-changing translator consists of five parts: speech-to-text, text translation, speech feature extraction, text-to-speech, and speech synthesis.

3.1 Speech-to-Text

Given a voice recording, use automatic speech recognition technology (ASR) to map the waveform to the correct text. First, the input waveform is converted into an acoustic wave sequence feature vector. Each vector represents an information signal within a small-time window; it needs to undergo processes such as sampling, quantization, windowing, and discrete Fourier transform. Then, use encoder-decoder Model (such as RNN or Transformer) or Connectionist Temporal Classification (CTC) to convert the digital signal into text. However, Noise, accents, and non-standard dialects will cause model not map Speech to the correct text; possible solutions are to obtain high-quality speech (less noise interference) or to denoise the language and expand the corpus (including Accents and non-standard dialects) enable non-standard languages to have corresponding texts.

3.2 Text Translation

Use machine translation to translate text from one language to another. First, use subword tokenization algorithms (select algorithms such as BPE algorithm, wordpiece algorithm, and unigram algorithm, based on language type) to build vocabulary. Then, a parallel corpus is constructed as training data. Use the constructed training dataset to train the encoder-decoder Transformer or sequence-to-sequence model. Finally, use the trained model to translate our target text. However, MT may encounter some problems: Word Order Typology (requiring the system to do huge structural reorderings as it generates the output), Lexical Divergences, Morphological Typology (use BPE models to handle languages with rich morphology), Referential density.

3.3 Speech Feature Extraction

Extract voice features from the user's voice, such as pitch, speaking speed, voice quality, etc. This is done using the pre-trained speech feature extraction model.

3.4 Text-to-Speech

Use text-to-speech (TTS) to map text to waveforms. A common architecture is encoder-decoder Architecture (LSTMs or Transformers). The TTS task consists of two parts: first, mapping from strings of letters to mel spectrograph using encoder-decoder model. Second, mapping from mel spectrograms to waveforms. First, we need to normalize the text, then use encoder-decoder with attention for Spectrogram prediction, and finally, use vocoder to map Spectrogram to wavefiles.

3.5 Speech Synthesis

The generated speech of the target language is combined with the user's speech characteristics and is synthesized using sound synthesis algorithm to produce the final voice changing effect. Finally, output the synthesized speech.

4 Anticipated Results

The proposed voice-changing translator allows users to speak in their own language and voice, while the system translates and reproduces the content into the desired target language. This approach not only facilitates communication, but also preserves the unique characteristics and emotional nuances conveyed by the user's voice. It solves the problem that existing translators can only reproduce the text content in the language and lose the timbre, prosody, and emotion in the speech.

References

Ahmed, K., Keskar, N. S., Socher, R. (2017). Weighted transformer network for machine translation. arXiv preprint arXiv:1711.02132.

Dan Jurafsky, James H. Martin. 2024. Speech and Language Processing (3rd ed. draft). https://web.stanford.edu/ jurafsky/slp3/

Deng, C., Yu, C., Lu, H., Weng, C., Yu, D. (2020, May). Pitchnet: Unsupervised singing voice conversion with pitch adversarial network. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 7749-7753). IEEE.

Sisman, B., Zhang, M., Sakti, S., Li, H., Nakamura, S. (2018, December). Adaptive wavenet vocoder for residual compensation in gan-based voice conversion. In 2018 IEEE Spoken Language Technology Workshop (SLT) (pp. 282-289). IEEE.

Wang, Q., Li, B., Xiao, T., Zhu, J., Li, C., Wong, D. F., Chao, L. S. (2019). Learning deep transformer models for machine translation. arXiv preprint arXiv:1906.01787.

Yao, S., Wan, X. (2020, July). Multimodal transformer for multimodal machine translation. In Proceedings of the 58th annual meeting of the association for computational linguistics (pp. 4346-4350).