Fast and Low-Latency End-to-End Speech Recognition and Translation(Abstract_要旨)

AUTHOR(S):

Inaguma, Hirofumi

（ 続紙 1 ）

京都大学	博士（情報学）	氏名	稲熊 寛文
論文題目	Fast and Low-Latency End-to-End Speech Recognition and Translation （高速・低遅延なEnd-to-End音声認識・翻訳）		

（論文内容の要旨）

Latency is an important factor for practicality in any real-time information processing system. The cause of latency includes various factors depending on the tasks, and achieving a better tradeoff of accuracy and latency is an open-ended research goal. This thesis focuses on latency in two fundamental speech-to-text generation tasks in real-world applications: streaming automatic speech recognition (ASR) and speech translation (ST). In streaming ASR, how quickly the system can display tentative recognition outputs while a speaker is talking impacts user experience. On the other hand, in ST, the system response time from the completion of speaking is important for utterance interpretation and smooth communication.

In the past years, advances in deep learning have enabled end-to-end training from speech to the target text sequence. Although it pushes up the limit of accuracy, an uncontrollable token emission latency is introduced in streaming end-to-end ASR models. Meanwhile, computational latency with incremental left-to-right decoding is not negligible for real-world applications of ST systems, even though end-to-end models are faster than traditional cascade models. Therefore, this thesis addresses two fundamental latency: emission latency in streaming end-to-end ASR and computational latency in end-to-end ST.

In Chapter 2, the conventional approaches of ASR and ST are formulated, and their problems are discussed.

In Chapter 3, we address the reduction of emission latency of streaming attention-based encoder-decoder (AED) ASR models. We propose alignment knowledge distillation from hybrid ASR systems and connectionist temporal classification (CTC) models. Unlike conventional knowledge distillation methods in an output probability space, the knowledge of token boundary positions is distilled in a latent alignment space. Because the teacher CTC model is trained jointly with the student AED model, the alignment knowledge distillation from CTC can be regarded as self-distillation, a purely end-to-end training framework. We formulate several training objective functions to synchronize the token boundaries of the student model to those of the teacher model. Moreover, we propose an alignment-free emission latency reduction method without any teacher model, which can be combined with the above distillation methods. We experimentally demonstrate that distillation from CTC achieves the best tradeoff between the recognition accuracy and emission latency. It also makes the model robust to long-form and noisy speech.

In Chapter 4, we address the reduction of display latency of streaming AED ASR models. To take advantage of efficient batched output-synchronous and low-latency input-synchronous search methods, we propose a block-synchronous beam search decoding. This guarantees a controllable display latency and achieves comparable accuracy to the label-synchronous decoding.

Moreover, to relieve the model of pre-segmentation with a separate voice activity detection (VAD) model, we propose a VAD-free streaming inference algorithm. It leverages probabilities from an auxiliary CTC layer to determine a suitable timing to reset the model states, and thus overcomes the vulnerability to long-form speech. We experimentally show that the proposed algorithm can recognize long-form speech stably for up to a few hours without any external VAD model. Moreover, it outperforms cascading VAD and ASR models in accuracy.

In Chapter 5, we address the reduction of computational latency in end-to-end ST models. To accelerate the decoding speed, a non-autoregressive (NAR) model, which generates multiple tokens in parallel, is incorporated. We approach this in two directions. The first approach is to adopt an NAR translation decoder instead of an autoregressive (AR) translation decoder that generates tokens incrementally. To make the best of both decoders, we propose a unified framework in which both NAR and AR decoders are jointly trained on the shared encoder, and the latter is used for rescoring the output from the former. This framework brings a large improvement of translation quality over the baseline NAR model only with a small additional cost for the rescoring. The second approach is to adopt a two-pass multi-decoder architecture that decomposes the overall ST task into ASR and MT sub-tasks by introducing an intermediate ASR decoder. We propose to condition the second-pass AR translation decoder on continuous representations from the first-pass NAR ASR decoder. Compared to a vanilla encoder-decoder model, this slows down the decoding speed, but the translation quality is improved by a large margin. On the other hand, this framework significantly improves the decoding speed compared to when using a first-pass AR ASR decoder without sacrificing the quality.

In Chapter 6, we address data sparseness in end-to-end ST models by making the most of bilingual ST corpora. We propose a simple yet effective framework for multilingual end-to-end ST, in which speech utterances in source languages are directly translated to the desired target languages with a universal sequence-to-sequence architecture. We experimentally show the effectiveness in one-to-many and many-to-many language directions. Moreover, to leverage the full potential of the source and target language information, we propose bidirectional sequence-level knowledge distillation (SeqKD), in which SeqKD from both source-to-target and target-to-source machine translation (MT) models is combined. The target-to-source MT model generates paraphrases of reference transcriptions via back-translation, and a bilingual end-to-end ST model is trained to predict the paraphrases as an auxiliary task with a shared decoder. Because paraphrasing is a monolingual translation task, bidirectional SeqKD can be regarded as pseudo multilingual training. We experimentally show that bidirectional SeqKD improves the translation quality of both AR and NAR models.

Chapter 7 concludes this thesis with a brief look at future work.

（続紙 ２ ）

（論文審査の結果の要旨）

音声認識や音声翻訳は、深層学習に基づくEnd-to-Endモデルの導入により大きな進歩を遂げている。その中で、入力発話全体をみるモデルが精度の点では優れているものの、出力の遅延が特に字幕付与などのリアルタイム応用では問題となる。本論文では、ストリーム型音声認識における個々のトークン（文字／単語）の出力遅延と音声翻訳における文全体の翻訳結果の計算遅延の削減に焦点をあてて行った研究成果をまとめたもので、主な成果は以下の通りである。

1. 注意機構に基づくエンコーダ・デコーダモデルをストリーム型の音声認識に適用する上で、出力遅延を削減する方法を複数提案した。これは、個々のトークンの時間境界情報を知識蒸留することにより、注意重みの計算の正則化に用いるものである。特に、知識蒸留の教師として、エンコーダを共有するConnectionist Temporal Classification (CTC)モデルを用いる手法では、これを含めた全体が１つのEnd-to-End系として最適化される。本手法により、認識精度と出力遅延の両面において最も優れた性能が実現されることを示した。

2. 音声認識デコーダにおける仮説探索において、出力トークンに同期した探索と入力フレームに同期した探索を組み合せて、最終的な出力遅延を制御するブロック同期型探索を提案した。さらに発話区間検出の前処理を必要としない完全なストリーム型の音声認識システムを実現した。

3. 音声認識と機械翻訳を密結合したEnd-to-End音声翻訳において計算遅延を削減するために、高速に並列計算可能な非自己回帰型(NAR)のモデルを活用する方法を複数提案した。１つは、NARモデルで複数の仮説を生成した後に、自己回帰型(AR)モデルにより再評価を行うもので、NARモデル単独よりも大幅に精度を改善し、ARモデル単独よりも大幅に高速な処理を実現した。もう１つは、NARモデルによる音声認識デコーダから機械翻訳デコーダに注意機構を張るもので、処理速度と翻訳精度の両立を実現した。

以上のように本論文は、音声認識と音声翻訳の両方において、アルゴリズムの改善を提案しながら、最先端の性能を示しているもので、学術上・実用上寄与するところが少なくない。よって、本論文は博士（情報学）の学位論文として価値あるものと認める。また、令和 3年 8月27日に論文とそれに関連した内容に関する口頭試問を行った結果、合格と認めた。なお、本論文のインターネットでの全文公開についても支障がないことを確認した。