Semi-supervised end-to-end training with text-to-speech

16 少量の入出力ペアから高精度に音声認識を学習

～音声合成を活用した半教師ありEnd-to-End学習～

どんな研究

音声認識は音声を書き起こした文字列へ変換する仕組みです。音声認識モデルの学習用に人手で用意する対応付いた音声と文字列のペアデータが少ない場合、高精度なモデルの実現は困難でした。この研究ではペアではない音声のみ・文字列のみのデータも活用できる学習方法を提案します。

どこが凄い

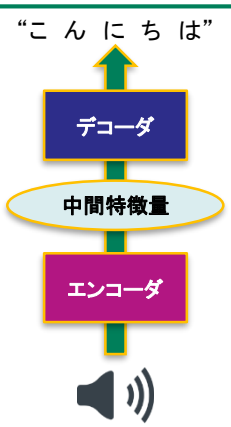
音声認識モデルと音声合成モデルの構造が似ていることに注目し、二つを組み合わせて音声・テキストのみで学んだ特徴量と音声認識の特徴量が近づく≒音声認識に活かせる半教師あり学習を実現しました。実験では少量のペアデータのみで学習する場合と比べて、文字誤り率を半分に削減しました。

めざす未来

既存の方法よりも少ない音声と文字列の学習用ペアデータで高精度な音声認識モデルを学習できます。将来的には、マイナーな言語や大量に準備しにくい音声（子供など）といったペアデータがほとんど得られない環境の音声認識など、より挑戦的な場面で活用できる技術の実現を目指します。

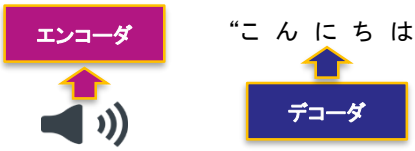
音声認識の枠組み

- 音声と文字列の学習用ペアを準備
- 音声を受け取るエンコーダと文字列を出力するデコーダを学習
- 未知の入力音声から正しく文字列が出力できるか評価



音声認識の課題

- エンコーダ・デコーダ型モデルの学習は、沢山の音声と書き起こし文字列のペアデータが必要
- 大量のペアデータの準備にはコストと時間がかかってしまう
- もし音声だけでエンコーダを、文字列だけでデコーダを学習できればデータの準備が簡単に

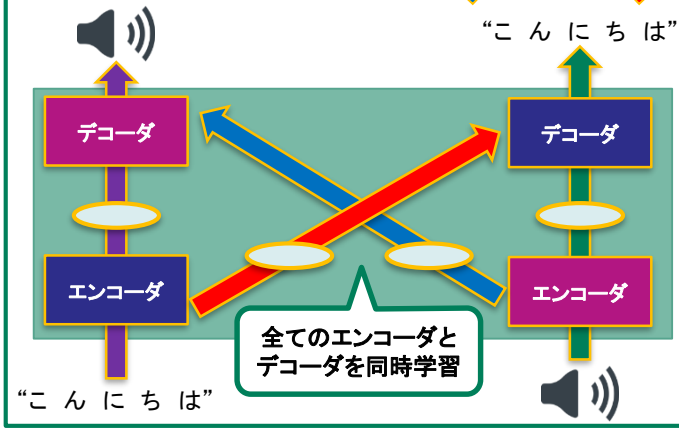


半教師あり学習への拡張

ポイント1：音声認識と音声合成を組み合わせる

・ペアデータが必要なタスク（音声認識↑、音声合成↑）

・ペアデータが不要なタスク（音声復元↖、文字列復元↗）



学習タスク				
音声認識	文字列復元	音声復元	音声合成	音声認識エラー率
✔				15.0%
✔	✔			9.0%
✔		✔		8.7%
✔	✔	✔	✔	8.4%

ポイント2：それぞれの中間特徴量 が似るように学習

関連文献

[1] S. Karita, S. Watanabe, T. Iwata, A. Ogawa, M. Delcroix, “Semi-supervised end-to-end speech recognition,” in *Proc. Interspeech*, 2018.

[2] S. Karita, S. Watanabe, T. Iwata, M. Delcroix, A. Ogawa, T. Nakatani, “Semi-supervised end-to-end speech recognition using text-to-speech and autoencoders,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2019.

連 絡 先

苅田 成樹 (Shigeki Karita) メデイア情報研究部 信号処理研究グループ

Email: cs-liaison-ml at hco.ntt.co.jp

Innovative R&D by NTT

オープンハウス 2019