Vol.2020-SLP-131 No.7
2020/2/13

国際会議既発表セッション

稲熊 寛文[1]　大淵 康成[2]　岡本 拓磨[3]　鈴木 貴仁[4]　中村 亮裕[4]

概要：情報処理学会 音声言語情報処理研究会では，原稿の執筆等をせずに気軽に研究紹介ができることを目指して，2019 年より「国際会議既発表セッション」という新たな取り組みを開始した．国内学会・研究会では未発表だが，最近の国際会議等で発表済み，投稿済みである論文を紹介する位置付けである．

1. はじめに

情報処理学会 音声言語情報処理研究会では，原稿の執筆等をせずに気軽に研究紹介ができることを目指し，2019 年より「国際会議既発表セッション」という新たな取り組みを開始した．国内学会・研究会では未発表だが，最近の国際会議等で発表済み，投稿済みである論文を紹介する位置付けである．

2020 年 2 月に開催する情報処理学会 音声言語情報処理研究会で紹介する論文は以下の通りである[*1]．

2. Multilingual End-to-End Speech Translation

著者

○ Hirofumi Inaguma (Kyoto University), Kevin Duh (JHU), Tatsuya Kawahara (Kyoto University), Shinji Watanabe (JHU)

概要

In this paper, we propose a simple yet effective framework for multilingual end-to-end speech translation (ST), in which speech utterances in source languages are directly translated to the desired target languages with a universal sequence-to-sequence architecture. While multilingual models have shown to be useful for automatic speech recognition (ASR) and machine translation (MT), this is the first time they are applied to the end-to-end ST problem. We show the effectiveness of multilingual end-to-end ST in two scenarios: one-to-many and many-to-many translations with publicly available data. We experimentally confirm that multilingual end-to-end ST models significantly outperform bilingual ones in both scenarios. The generalization of multilingual training is also evaluated in a transfer learning scenario to a very low-resource language pair.

会議名

ASRU2019

[1] 京都大学
[2] 東京工科大学
[3] 情報通信研究機構
[4] 静岡大学

[*1] 本稿の著者は各論文の紹介者で，五十音順である．

3. Personalized Quantification of Voice Attractiveness in Multidimensional Merit Space

著者
○大淵 康成（東京工科大）

概要
声の魅力度を自動推定することは，ビジネスやエンタテインメントなどの様々な場面で有益である．そのための機械学習的なアプローチとしては，学習データに魅力度スコアを付与し，音響特徴量から魅力度を推定するモデルを作るという方法が考えられる．しかし，どのような声に魅力を感じるかには個人差が大きく，単一のスコアによるモデル化では不十分である．本研究では，魅力度を多次元ベクトルとして定義し，聴取者の嗜好ベクトルとの内積で魅力度が決まるようなモデルを考える．複数音声を聴取して好悪の相対的なラベリングを行い，その結果を最もよく表現するマッピングを求める．更に，音響特徴量を用いた音声からの魅力度ベクトルの推定についても実験を行い，多次元化の有効性を示す．

会議名
Interspeech 2017

4. Tacotron-based acoustic model using phoneme alignment for practical neural text-to-speech systems

著者
○岡本 拓磨（NICT），戸田 智基（名大/NICT），志賀 芳則，河井 恒（NICT）

概要
Although sequence-to-sequence (seq2seq) models based on attention mechanism in neural text-to-speech (TTS) systems can jointly optimize duration and acoustic models, and achieve high-quality synthesis, these involve a risk that speech samples cannot be sometimes successfully synthesized due to the attention prediction errors. Therefore, these seq2seq models cannot be directly used in practical TTS systems. On the other hand, conventional pipeline models are broadly introduced in practical TTS systems since there are few crucial prediction errors in the duration model. To realize high-fidelity practical TTS systems without attention prediction errors, this paper investigates Tacotron-based acoustic models with phoneme alignment instead of attention mechanism. The phoneme durations are obtained from HMM-based forced alignment and a conventional bidirectional LSTM-based duration model is introduced. Then, a seq2seq model with forced alignment instead of attention is investigated and an alternative model with Tacotron decoder and phoneme duration is proposed. The results of experiments with full-context label input using WaveGlow vocoder indicate that the proposed model can realize a high-quality TTS system for Japanese with a real-time factor of 0.13 using a GPU without attention prediction errors.

会議名
ASRU 2019

5. Knowledge Distillation for Throat Microphone Speech Recognition

著者

○鈴木 貴仁（静岡大），緒方 淳（産総研），綱川 隆司，西田 昌史，西村 雅史（静岡大）

概要

皮膚振動を捉える咽喉マイクは外部雑音に頑健であり，高雑音環境下での音声認識に活用する研究が行われている．しかしながら，接話マイク等の空気振動を捉える通常のマイクとは異なった特性を持つ音が収録されるため，一般的な音声認識システムにそのまま咽喉マイク音声を入力しても高い認識精度は得られない．また，利用可能な咽喉マイクの音声データは限られているため，咽喉マイク音声のみで高精度な音響モデルを学習することも困難である．本研究では，咽喉マイクと接話マイクのパラレルデータ用いて知識蒸留法によって咽喉マイク用ハイブリッド方式音響モデルを学習する手法を提案する．咽喉マイク用 DNN（生徒モデル）は大量の接話マイク音声で学習したハイブリッド方式音響モデルの DNN（教師モデル）の出力を模倣するように学習する．さらに，生徒モデルの前段と後段をそれぞれ，咽喉マイク特徴量から接話マイクのボトルネック特徴量への特徴マッピング用 LSTM のパラメータと接話マイクのボトルネック特徴量から音素を識別するネットワークのパラメータで初期化することで認識精度の更なる改善を試みた．咽喉マイクで収録した新聞記事読み上げ音声を用いた評価の結果，咽喉マイク音声のみで学習したハイブリッド方式音響モデルと比較して提案手法は 9.8%の文字誤り率の削減を達成した．

会議名

Interspeech 2019

6. Estimation of Number of Chewing Strokes and Swallowing Events by Using LSTM-CTC and Throat Microphone

著者

○中村 亮裕・Muhammad Mehedi Billah・阿部 太樹（静岡大），齊藤 隆仁，池田 大造（NTT ドコモ），峰野 博史，西村 雅史（静岡大）

概要

健康維持の観点から重要とされている咀嚼から嚥下に至る食事行動を，中咽頭で収録された食事音を用いて，簡便にモニタリングできるシステムの検討を行っている．咀嚼と嚥下の回数の検出について，フレーム単位の正確なラベル（強ラベル）が付与された 600 個程度の少量データで学習した LSTM では十分な検出性能が得られていなかった．本研究では，正確な時間情報の無いラベル（弱ラベル）が付与された大量のデータを用いて LSTM-CTC による学習を行った．弱ラベルを強ラベルの約 5 倍用意すると従来手法と同等の性能となり，約 10 倍用意するとさらに誤りが 50%減少する結果が得られ，咀嚼や嚥下の検出を大幅に改善できる見通しを得られた．

会議名

GCCE 2019