音声・音響信号処理における世界最大規模の国際学会 「INTERSPEECH 2023」にて、4本の論文が採択

2023.06.30　技術情報

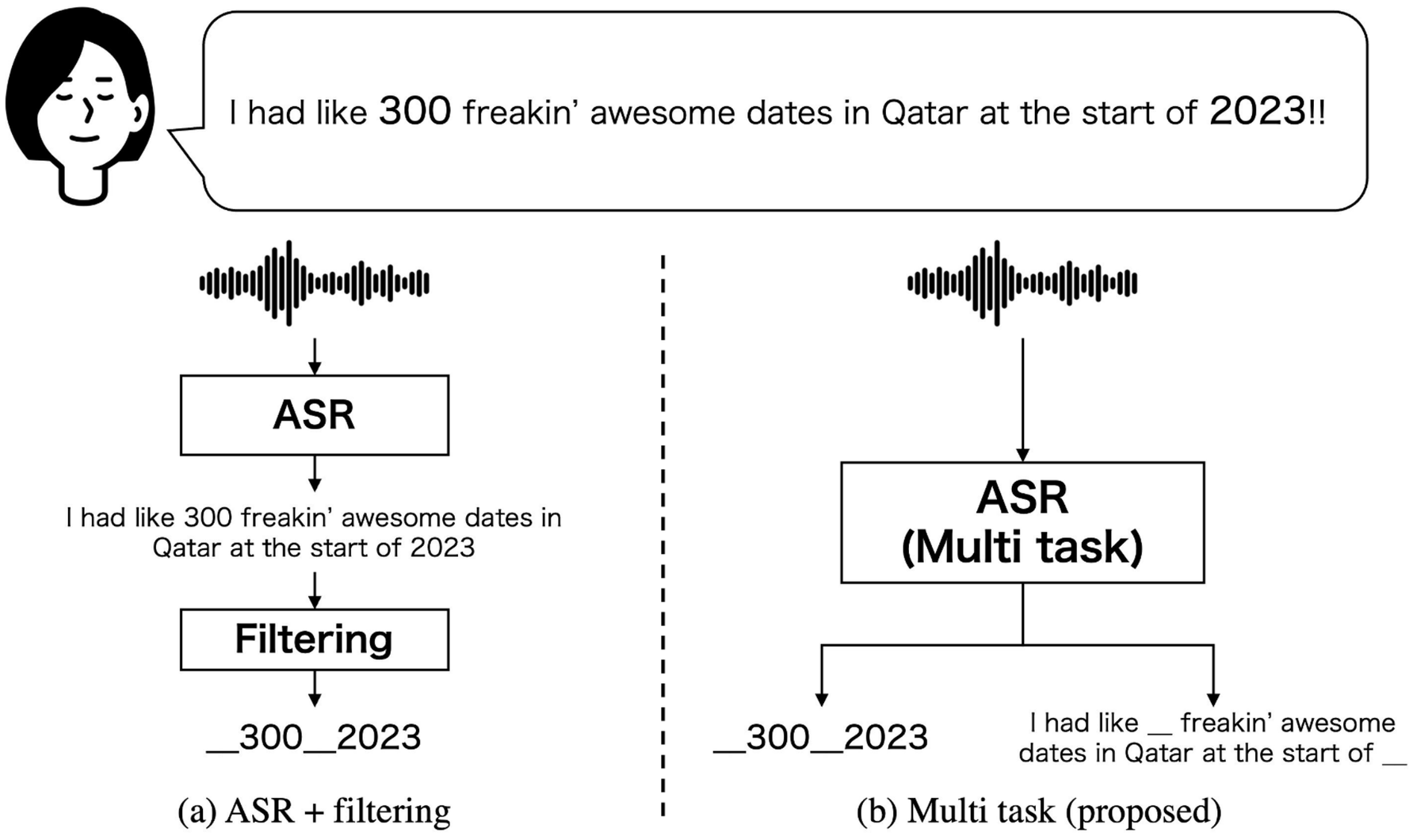
音声認識や環境音認識における研究成果が評価

LINE株式会社（以下、LINE）による論文4本が、2023年8月20日から24日にかけて開催される、音声・音響信号処理における世界最大規模の国際学会「INTERSPEECH 2023」（ダブリン、アイルランド）にて採択されました。

「INTERSPEECH」はInternational Speech Communication Association（ISCA）が主催する、音声処理における世界最大規模の国際会議で、今年で24回目の開催となります。4本の論文のうち2本はLINEの主著、2本は東京大学との共著で、いずれも開催期間中に発表されます。

■音声認識や環境音認識において、精度向上を可能とする提案が評価

LINE主著の論文[1] では、音声認識とキーワード検出を組み合わせたシステムにおいて、キーワードの検出精度が向上するように音声認識モデルをチューニングする新たな提案が採択されました。従来法では、キーワードに特化した音声認識モデルのチューニングは困難で、音声認識に誤りがあるとキーワード検出に失敗していました。提案法では、音声認識の学習データをキーワード系列と非キーワード系列に分解することにより、音声認識とキーワード検出の同時チューニングを可能にします（図1）。同時チューニングによってキーワードに対する音声認識の誤りが少なくなり、音声からカタカナ名詞や数字列等のキーワードを検出する課題において、従来法を上回る検出精度を達成しました。



(a) 音声認識の後にキーワード検出 (b) 音声認識とキーワード検出を同時チューニング
図1:キーワード検出方法の比較

同じくLINE主著の論文[2] では、音声や環境音を認識するシステムにおいて、移動する複数の音源に追従できる新たな方式を提案しました。従来法の多くは、音源が短時間では移動しないことを仮定したモデルが用いられるため、移動する人の声や足音などが認識しづらくなる問題がありました。提案法では、音源を分離するモデルの内部に、移動に追従する注意機構を備えることにより、移動する音源を精度良く抽出できます（図2）。特に、移動する音源が複数になった場面での効果が高く、複数人による会話の音声認識と環境音の識別において、移動を考慮しない従来法を上回る精度を達成しました。

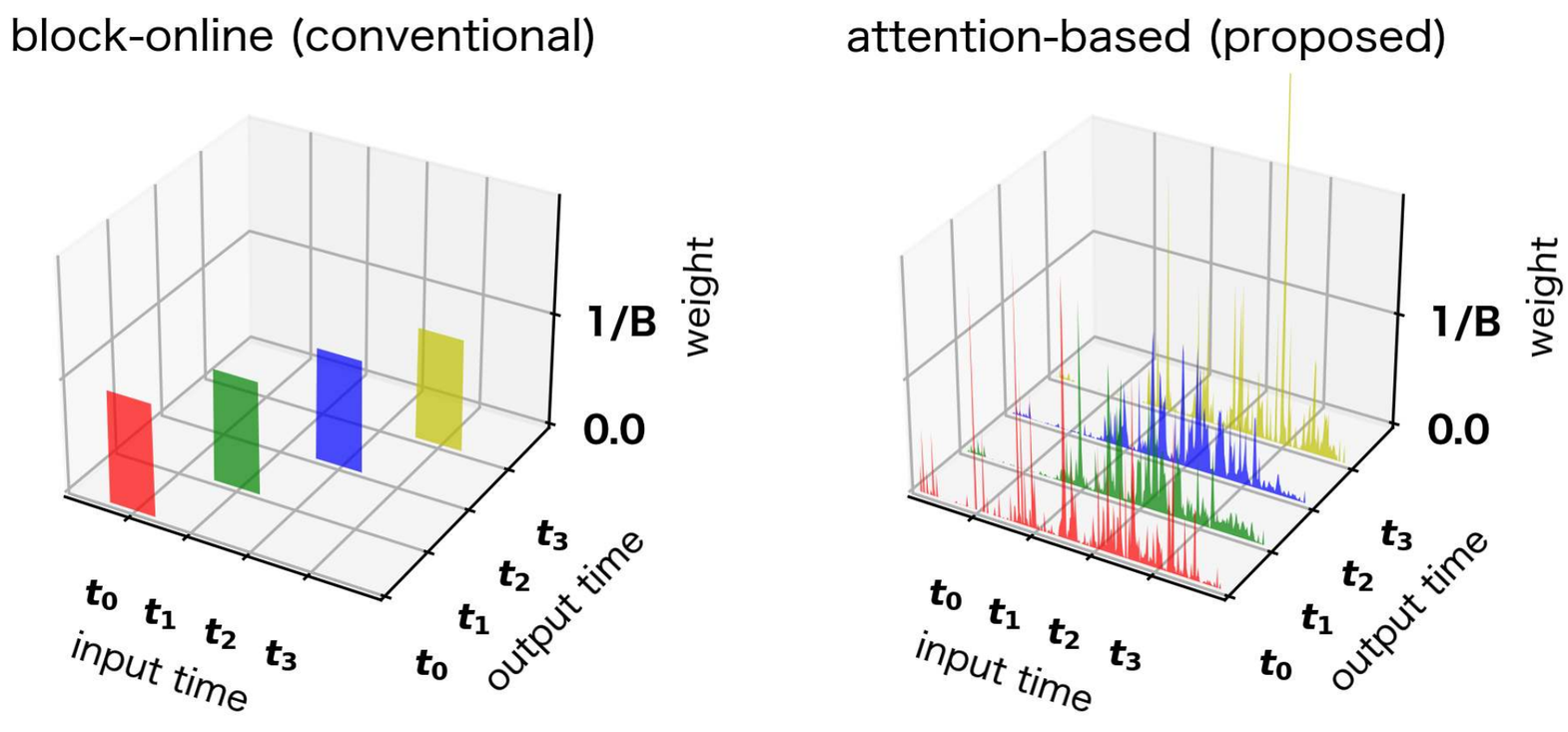


図2: 移動に追従する注意機構の比較

採択された論文

1."Target Vocabulary Recognition Based on Multi-Task Learning with Decomposed Teacher Sequences", Aoi Ito, Tatsuya Komatsu, Yusuke Fujita, Yusuke Kida

2."Multi-channel separation of dynamic speech and sound events", Takuya Fujimura, Robin Scheibler

3."CALLS: Japanese Empathetic Dialogue Speech Corpus of Complaint Handling and Attentive Listening in Customer Center", Yuki Saito, Eiji Iimori, Shinnosuke Takamichi, Kentaro Tachibana, and Hiroshi Saruwatari

4."ChatGPT-EDSS: Empathetic Dialogue Speech Synthesis Trained from ChatGPT-derived ContextWord Embeddings", Yuki Saito, Shinnosuke Takamichi, Eiji Iimori, Kentaro Tachibana, and Hiroshi Saruwatari

※1-2がLINE主著、3-4が東京大学との共著の論文です。

LINEではAI技術を活用した新たなサービスの創出を進めるとともに、AI技術そのものの研究開発活動にも注力しています。特に音声処理分野においては、音声認識・音声合成技術を中心に、これまで数々のトップカンファレンスにてインパクトのある研究成果を発表してきました。例えば、質の高い音声を高速で合成することができるParallel WaveGAN*1や、高速の音声認識を実現する手法である非自己回帰型音声認識*2モデルの中でも最も高い精度を示したSelf-Conditioned CTC*3といった最先端技術を開発してきました。また環境音分析では、国際的なコンペティションであるDCASE2020にて世界１位を獲得しています。

LINEは、今後もAI技術に関連した基礎研究を積極的に推進することで、既存サービスの品質向上や、新たな機能・サービスの創出に努めてまいります。

*1 Parallel WaveGAN (PWG)：機械学習の生成モデルのひとつであり2つのニューラルネットワークを用いて学習を行って入力されたデータから新しい疑似データを生成する「敵対的生成ネットワーク（Generative Adversarial Network / GAN）」を用いた非自己回帰型音声生成モデル
*2 非自己回帰型音声認識：過去に生成したテキストに依存せずに、各時点の音声を認識する手法
*3 Self-Conditioned CTC：End-to-End型の音声認識モデルの一種であり、ニューラルネットワークの中間層で予測したテキストを参照して最終的な予測を行う手法
