LINE、音声処理における世界最大規模の国際学会 「INTERSPEECH 2022」にて12本の論文採択

2022.07.05　技術情報

音声認識技術、音声合成技術で研究成果が多数評価され、論文採択数は昨年の2倍に

LINE株式会社（本社：東京都新宿区、代表取締役社長：出澤 剛）は、音声処理における世界最大規模の国際学会「INTERSPEECH 2022」にて、12本の論文が採択されましたので、お知らせいたします。

「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す。今年で23回目の開催となり、採択された論文は9月18日から22日にかけて開催される「INTERSPEECH 2022」（韓国、仁川）にて発表されます。

今回、LINEとしては昨年の6本から、大幅に採択論文数を伸ばす結果となりました。なお12本のうち7本がLINE主著、5本がNAVERや大学など他グループとの共著となります。

■LINEが研究開発に注力してきた音声処理技術

LINEではAI事業を戦略事業の一つとして位置付け、AIテクノロジーブランド「LINE CLOVA」をはじめ、AI技術を活用した新たなサービスの創出を進めております。また、これまで世の中になかった新たな技術を生み出し事業の発展を加速することを目的に、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、言語処理、画像処理などのメディア処理分野において、機械学習を技術の軸とした研究開発をすすめており、各チームが事業や担当領域を超えて連携しながら「研究 > 開発 > 事業化」のサイクルをスピードアップすることを目指しています。

さらに音声処理分野においては、様々なサービスに展開している音声認識・音声合成技術を中心に、これまでに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まいりました。例えば、質の高い音声を高速に合成することのできるParallel Wave GAN*1や、高速な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などの最先端の技術を開発してきたほか、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。

■Self-Conditioned CTCの拡張方式、平静音声から感情音声合成モデルを構築する方式など、12本の論文が採択

今回の「INTERSPEECH 2022」では、音声認識技術、音声合成技術の両面で、研究成果が多数評価されました。

まず音声認識の研究としては、上述のSelf-Conditioned CTCを拡張した、2件の論文 [1][2] が採択されました。[1] では、認識モデルの学習時に、中間層での予測結果に対して意図的に誤りを加えることによって、モデルの頑健性が高められることを示しました。[2] では、中間層での予測結果を外部言語モデルと仮説探索を使って改善させることで、最終的な認識精度を向上できることを示しました。また、音源分離や多チャンネル信号処理を用いた研究としては、複数のマイクを使って音源分離を行う手法の一つであるT-ISSIにニューラルネットワークを使った音源モデルを導入し、残響除去と音源分離を同時に実現するよう学習することで、高い音声認識精度が得られることを示した論文 [3] が採択されました。さらに、ニューラルネットワークを用いた多チャンネル音源分離技術において、モデル学習の教師として従来必要であったクリーンな音声を利用せず、古典的な信号処理手法から得られる音源の空間的な情報を手掛かりに、高精度なモデルが学習可能であることを示した論文 [4] が採択されました。

また、音声合成の研究としては、少量の平静音声を収録した話者の感情音声合成モデルを構築する論文 [5] が採択されました。提案手法では、平静音声を有する目標話者と、感情音声を有する他話者との間で声質変換を行い、目標話者の擬似感情音声を生成し、感情音声合成モデルの構築に活用しています。また、自然な韻律を実現することを目指して、日本語アクセント予測に取り組んだ論文 [6] が採択されました。日本語アクセントは、アクセントの変化が起きる区間を指す「アクセント句」と、句内でアクセントの変化が起きている箇所を指す「アクセント核」という、2つの要素によって構成されています。提案手法では、multi-task learningによってこれらを同時に予測することで、独立して「アクセント句」と「アクセント核」を予測していた従来手法よりも、精度面と合成音の自然生面において、大幅な向上が確認されました。さらに、学習データの録音環境と言語情報・話者情報を分離するための正則化の導入により、雑音や残響を含む音声からクリーンな音声合成モデルを学習する手法を提案した論文 [7] が採択されました。

*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擬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

採択された論文

1. Yu Nakagome, Tatsuya Komatsu, Yusuke Fujita, Shuta Ichimura, Yusuke Kida, "InterAug: Augmenting Noisy Intermediate Predictions for CTC-based ASR"

2. Tatsuya Komatsu, Yusuke Fujita, Jaesong Lee, Lukas Lee, Shinji Watanabe, Yusuke Kida, "Better Intermediates Improve CTC Inference"

3. Kohei Saijo, Robin Scheibler, "Independence-based Joint Dereverberation and Separation with Neural Source Model".

4. Kohei Saijo, Robin Scheibler, "Spatial Loss for Unsupervised Multi-channel Source Separation"

5. Ryo Terashima, Ryuichi Yamamoto, Eunwoo Song, Yuma Shirahata, Hyun-Wook Yoon, Jae-Min Kim, Kentaro Tachibana, "Cross-Speaker Emotion Transfer for Low-Resource Text-to-Speech Using Non-Parallel Voice Conversion with Pitch-Shift Data Augmentation"

6. Byeongseon Park, Ryuichi Yamamoto, Kentaro Tachibana, "A Unified Accent Estimation Method Based on Multi-Task Learning for Japanese Text-to-Speech"

7. Takaaki Saeki, Kentaro Tachibana, Ryuichi Yamamoto, "DRSpeech: Degradation-Robust Text-to-Speech Synthesis with Frame-Level and Utterance-Level Acoustic Representation Learning"

8. Hyunwook Yoon, Ohsung Kwon, Hoyeon Lee, Ryuichi Yamamoto, Eunwoo Song, Jae-Min Kim, and Min-Jae Hwang, "Language Model-Based Emotion Prediction Methods for Emotional Speech Synthesis Systems"

9. Eunwoo Song, Ryuichi Yamamoto, Ohsung Kwon, Chan-Ho Song, Min-Jae Hwang, Suhyeon Oh, Hyun-Wook Yoon, Jin-Seob Kim, Jae-Min Kim, "TTS-by-TTS 2: Data-selective Augmentation for Neural Speech Synthesis Using Ranking Support Vector Machine with Variational Autoencoder"

10. Yuki Saito, Yuto Nishimura, Shinnosuke Takamichi, Kentaro Tachibana, and Hiroshi Saruwatari, "STUDIES: Corpus of Japanese Empathetic Dialogue Speech Towards Friendly Voice Agent"

11. Yuto Nishimura, Yuki Saito, Shinnosuke Takamichi, Kentaro Tachibana, and Hiroshi Saruwatari, "Acoustic Modeling for End-to-End Empathetic Dialogue Speech Synthesis Using Linguistic and Prosodic Contexts of Dialogue History"

12. Yen-Ju Lu, Xuankai Chang, Chenda Li, Wangyou Zhang, Samuele Cornell, Zhaoheng Ni, Yoshiki Masuyama, Brian Yan, Robin Scheibler, Zhong-Qiu Wang, Yu Tsao, Yanmin Qian, Shinji Watanabe, "ESPnet-SE++: Speech Enhancement for Robust Speech Recognition, Translation, and Understanding"

※1-7がLINE主著、8-12がLINEと他グループ共著の論文となります。

【LINEが提供するAIテクノロジーブランド「LINE CLOVA」】

LINEが提供するAIテクノロジーブランド「LINE CLOVA」は、さまざまなAI技術やサービスを通して生活やビジネスに潜む煩わしさを解消し、社会機能や生活の改善によって便利で豊かな世界をもたらすことを目指しています。現在、音声に関連する技術として「CLOVA Speech（音声認識）」「CLOVA Voice（音声合成）」、そしてそれらの技術を組み合わせたソリューションの提供も行っています。これまでに、AIによる自然な対話応答によりユーザーの目的を実現する電話応対AIサービス「LINE AiCall」や、会議やインタビューなどの自由発話を高い精度で認識し、記録・管理するAI音声認識アプリ「CLOVA Note」などをリリースしました。

今後も、AI技術に関連する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。
