音声・音響信号処理における世界最大の国際学会 「ICASSP 2023」にて、8本の論文が採択

2023.04.14　技術情報

音声認識、音声合成の研究成果が評価され、主著の採択論文数は昨年から倍増

LINE株式会社（以下、LINE）による論文8本が、6月4日から10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「ICASSP 2023」（ロドス島、ギリシャ）にて採択されました。「ICASSP」(International Conference on Acoustics, Speech, and Signal Processing)は、米国電気電子学会の中で最も長い歴史を持つ信号処理学会である「IEEE Signal Processing Society」が主催する国際学会で、今年で48回目の開催となります。

8本の論文のうち6本はLINEの主著、2本は他グループとの共著となり、いずれも開催期間中に発表されます。これにより、LINE主著の採択論文数は、昨年の3本から倍増の成果となりました。

■感情音声合成や音源分離等において、より自然な音声合成を実現する提案に評価

論文[4] では、感情音声合成におけるテキストから音声波形への変換プロセスでの、音声のピッチ情報（声の高さ）を利用したEnd-to-Endモデルに関する提案が採択されました。変換プロセスを単一モデルで行うEnd-to-Endモデルでは質の高い音声を生成することが可能ですが、従来のモデルでは、より豊かな表現が必要となる感情音声合成において、自然な音声を合成することが難しいケースが多く見られました。

提案法では、感情音声合成の際により重要となるピッチ情報を、陽にモデル化しました。これによって生成音声におけるピッチ情報をより正確に表現することが可能となり、従来法では生成が難しいとされていた、ピッチが極端に高い・低い等の発話についても、より自然かつ安定した生成結果が得られることを示しました。（図1）

また論文[5] では、複数話者が混在した音声を分離する音源分離において、画像生成にも活用される拡散モデルを用いた方式が採択されました。（図2）機械学習を利用する従来の音源分離では、教師データにおける音声の分離度を最大化する識別モデルを利用することが主流でしたが、高い分離度の音声であっても、人間が聞いた場合には不自然に聞こえるケースが散見されました。

提案法は、画像生成にも活用される生成モデルの一つである拡散モデルを音源分離に活用することで、自然な音声の生成を実現しました。拡散モデルの活用によって分離音の歪みが少なくなり、人間の知覚に基づく音声品質評価指標(DNSMOS)において、従来法を上回ることを示しました。

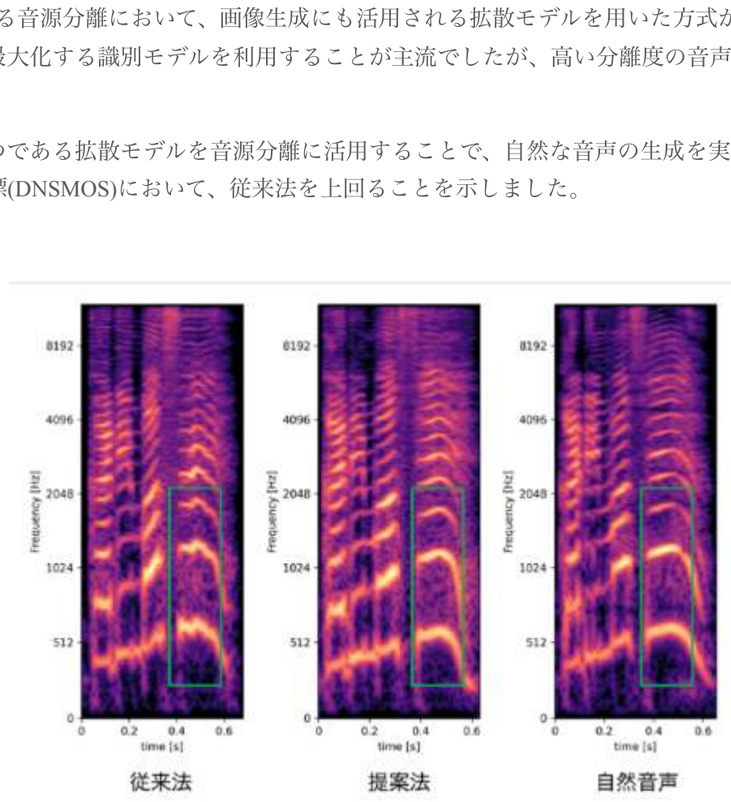

図1: 生成音声と肉声音声のメルスペクトルグラム

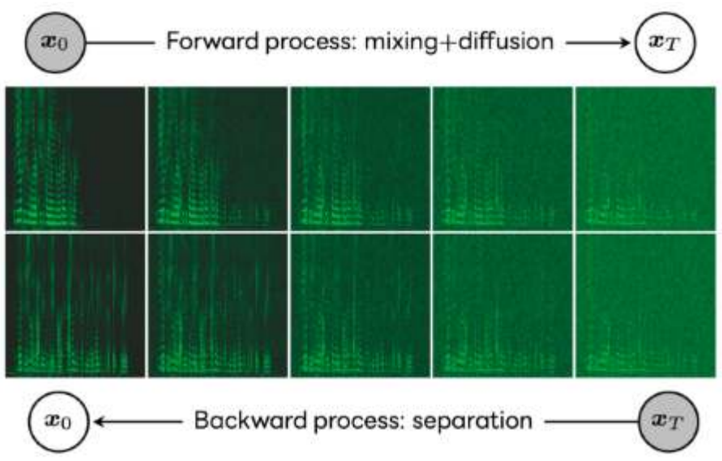

図2: Diffusion混合過程

採択された論文

1. R. Yamamoto et al., "NNSVS: NEURAL NETWORK BASED SINGING VOICE SYNTHESIS TOOLKIT"
2. R. Yoneyama et al., "Non-parallel High-Quality Audio Super Resolution with Domain Adaptation and Resampling CycleGANs"
3. M. Kawamura et al., "LIGHTWEIGHT AND HIGH-FIDELITY END-TO-END TEXT-TO-SPEECH WITH MULTI-BAND GENERATION AND INVERSE SHORT-TIME FOURIER TRANSFORM"
4. Y. Shirahata et al., "Period VITS: Variational inference with explicit pitch modeling for End-to-End emotional speech synthesis"
5. R. Scheibler et al., "DIFFUSION-BASED GENERATIVE SPEECH SOURCE SEPARATION"
6. Y. Fujita et al., "Neural Diarization with Non-autoregressive Intermediate Attractors"
7. T. Kawamura et al., "Effectiveness of Inter- and Intra-Subarray Spatial Features for Acoustic Scene Classification"
8. H. Zhao, et al., "Conversation-oriented ASR with multi-look-ahead CBS architecture"

※1-6がLINE主著、7が東京都立大学との共著、8が早稲田大学との共著の論文となります。

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まいりました。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では、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

2023年4月19日 18:20

採択された論文[3]のタイトルについて、「FAST AND HIGH-FIDELITY END-TO-END TEXT-TO-SPEECH FOR EDGE DEVICES WITH MULTI-BAND GENERATION AND INVERSE SHORT-TIME FOURIER TRANSFORM」より、「LIGHTWEIGHT AND HIGH-FIDELITY END-TO-END TEXT-TO-SPEECH WITH MULTI-BAND GENERATION AND INVERSE SHORT-TIME FOURIER TRANSFORM」へと修正