Ensemble of multi-stream diffusion model enhances speech

08 多ストリーム拡散モデルのアンサンブル推論による音声強調

騒がしい環境で録音音声をクリーンな音声に変換

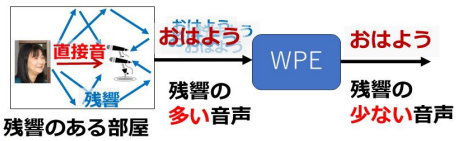
どんな研究	近年、生成AI技術により高品質な音声生成が可能になってきています。本研究では、生成AIを活用して録音音声から雑音や残響を高精度に除去する音声強調手法を紹介します。従来のNTT最先端音声強調技術と生成AI技術を統合し、格段に高品質な音声強調を実現します。
どこが凄い	生成AI（拡散モデル）による音声生成に、従来の最先端技術に基づく複数の強調音声を統合的に組み込む方法を考案し、格段に性能を向上しました。また、拡散モデルの複数回の出力を平均することが、音声強調の性能改善に極めて効果的であることを世界で初めて示しました。
めざす未来	騒がしい環境で、目的の音声だけをスタジオで収録したような高品質な音声収録が可能になります。日常環境の中で、高品質な音声コンテンツやデータが収集できたり、より快適な遠隔会議が実現できるなど、様々な音声アプリケーションの利便性が大幅に向上することが期待できます。

目的：生成AI技術を用いて従来の音声強調処理を性能向上させたい

～ 従来の最先端音声強調技術 ～

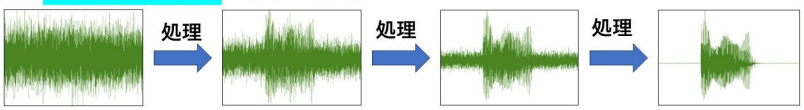
■ 目的音声抽出技術 SpeakerBeam
- 聞きたい人の声を取り出す

■ 残響抑圧技術 WPE
- 音声に含まれる残響を抑圧



～ 生成AI技術 ～

■ 拡散モデルによる高品質な音声の生成
- 雑音から処理を繰り返して、高品質な音声に近づける



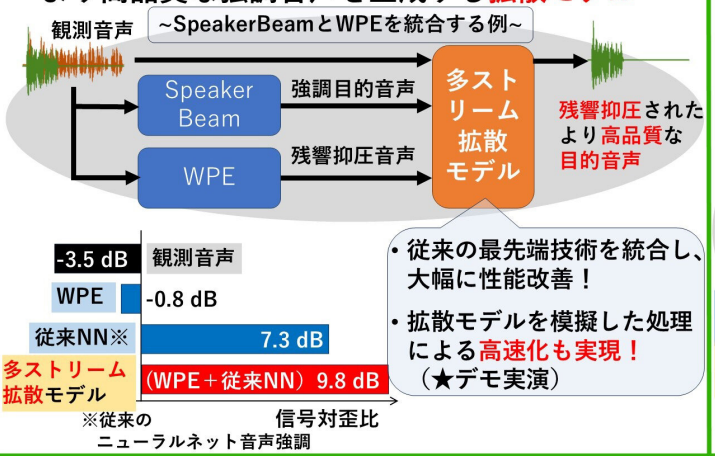
■ 音声強調への応用（従来技術）
- 観測音声から処理を繰り返して、雑音・残響に埋もれた音声を復元し、強調音声を得る

★課題1 拡散モデルによる音声強調の従来技術は性能が不十分☹
★課題2 従来の最先端技術（SpeakerBeam、WPEなど）との組み合わせ方が未知☹

①多ストリーム拡散モデル

★課題1、2を解決☺

■ 複数の強調音声（多ストリーム）を統合して、より高品質な強調音声を生成する拡散モデル

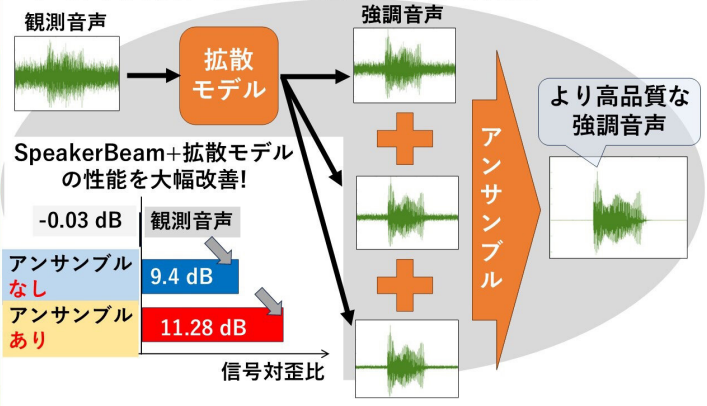


- 従来の最先端技術を統合し、大幅に性能改善！
- 拡散モデルを模擬した処理による高速化も実現！（★デモ実演）

②拡散モデルのアンサンブル推論

★課題1を解決☺

■ 拡散モデルは強調音声を確率的変動込みで生成
⇨複数回生成した強調音声のアンサンブル（平均）により大幅に性能向上することを発見



関連文献

[1] N. Kamo, M. Delcroix, T. Nakatani, “Target speech extraction with conditional diffusion model,” in *Proc. INTERSPEECH*, pp. 176-180, 2023.

[2] T. Nakatani, N. Kamo, M. Delcroix, S. Araki, “Multi-stream diffusion model for probabilistic integration of model-based and data-driven speech enhancement,” in *Proc. IWAENC*, pp. 65-69, 2024.

連絡先

加茂 直之（Naoyuki Kamo）メディア情報研究部 信号処理研究グループ