ISS-SP-003

# Zero-Shot 声質変換における雑音環境下音声から話者情報の獲得

坂本　瞭†　　竹内　太法††　　立蔵　洋介†
† 静岡大・院・総合科学技術研　　†† 静岡大・創造科学技術大学院

## 1. はじめに

声質変換は，入力話者による音声の音韻情報を保持したまま声質のみを対象とする話者に合わせて変換する技術である．この技術は話者変換や感情変換，発話支援など様々なタスクへの応用が期待されている．しかし，従来手法では目標話者のクリーンな音声が 5〜10 分程度必要であり，ユーザが目標話者ごとに大量のクリーン音声を集めることは容易ではない．したがって，雑音が重畳した少量の目標話者音声で変換可能なモデルが必要である．少量の目標話者音声で変換可能な手法として Zero-Shot 学習を用いた AUTOVC [1] が提案されている．そこで本稿では，AUTOVC における，話者埋め込みを獲得するモデルの雑音に対する頑健性を向上させるため，学習データに雑音を重畳させ，モデルの学習をおこない，雑音の頑健性について評価する．

## 2. AUTOVC

AUTOVC は音声から話者情報を表す話者埋め込みを生成する話者エンコーダ $Es$，音声から音声情報と韻律情報を表すコンテンツ埋め込みを生成するコンテンツエンコーダ $Ec$，話者埋め込みとコンテンツ埋め込みから音声を生成するデコーダ $D$ の 3 つの主要なモジュールから構成されている．元話者の音声データのスペクトログラムを$X_1$，目標話者の音声データのスペクトログラムを$X_2$と表す．変換された音声 $\hat{X}_{1\to2}$ は $X_1$のコンテンツ$C_1$，$X_2$の話者埋め込み $S_2$ を用いて次の式で与えられる．

$$\hat{X}_{1\to2} = D(C_1, S_2),\ C_1 = Ec(X_1),\ S_2 = Es(X_2) \qquad (1)$$

## 3. 雑音重畳音声を用いた性能評価

提案手法を図1に示す．$Es$を雑音に頑健なモデルにするため，初めにクリーンな音声のみを使用してモデルを学習し，次に学習したモデルに対して雑音重畳音声を用いて再学習した．音声は Voxceleb1 [2]と Librispeech [3] を組み合わせ，学習データとして 3525 話者，テストデータとして学習に含まれていない 4 話者（男女 2 名ずつ）を用いた．雑音として学習データには，白色雑音および DEMAND dataset [4] から PSTATION, NFIELD, DLIVING の 4 種類を用いた．テストデータには，STRAFFIC，PCAFETER，TMETRO の 3 種類を用いた．雑音重畳音声は，学習データおよびテストデータの各発話に対してランダムで選ばれた雑音を SNR 10〜30 dB で重畳して作成した．

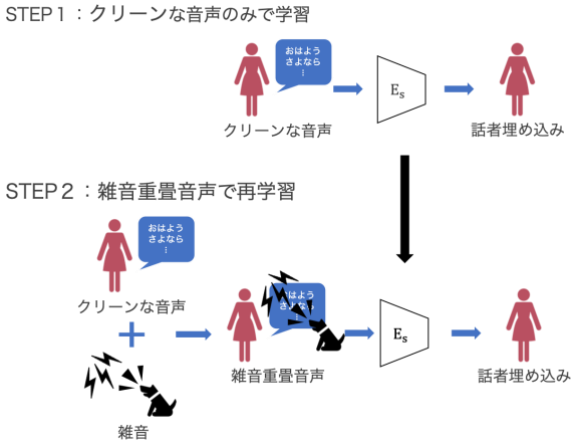


図1　提案手法の概要図

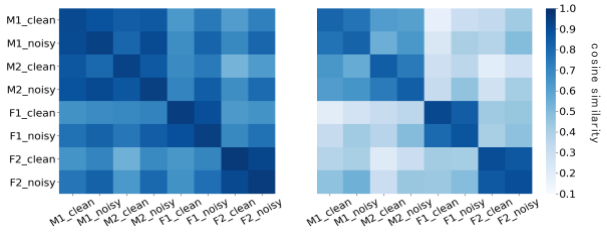


図2　クリーン音声のみで学習した場合（左）と雑音重畳音声で再学習した場合（右）のコサイン類似度の評価結果

学習した$Es$にテストデータからクリーンな音声，雑音重畳音声を入力し，得られた話者埋め込みと話者ごとのセントロイドとのコサイン類似度を用いて性能評価をおこなった．図2にコサイン類似度による評価結果を示す．結果より，雑音重畳音声を用いて再学習した場合，同一話者の類似度は高く，異なる話者間の類似度は低くなり話者ごとに大きく異なる話者埋め込みを出力していることが確認できた．またクリーン音声のみで学習した場合に比べ，雑音の有無にかかわらず，話者ごとに類似度が高くなっているため雑音の頑健性も向上していた．よって，クリーン音声で学習したモデルを混合音声で再学習した場合，雑音に頑健なモデルであることが示唆された．

## 参考文献

[1] K. Qian et al., *ICML*, pp. 5210-5219, 2019.
[2] A. Nagrani et al., *INTERSPEECH*, 2017.
[3] V. Panayotov et al., *ICASSP*, pp. 5206-5210, 2015.
[4] J. Thiemann et al., *Journal of ASA*, pp. 3591-3591, 2013.