[1-P-18] 主観的話者間類似度に基づくDNN話者埋め込みを用いた多数話者DNN音声合成の実験的評価

◎ 齋藤 佑樹　高道 慎之介　猿渡 洋　(東大院・情報理工)

1. 本発表の概要

主観的話者間類似度に基づく DNN 話者埋め込みを, Variational AutoEncoder (VAE) を用いた多数話者音声生成モデリングに適用し, 合成音声の品質を主観評価

結果: (1) 話者認識に基づく DNN 話者埋め込みと比べて, 合成音声の品質改善 & (2) Gauss カーネルを用いた話者埋め込み学習により, 合成音声の品質劣化

研究プロジェクト Web ページ

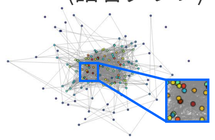
2. 従来手法

主観的話者間類似度に基づく DNN 話者埋め込み[1]

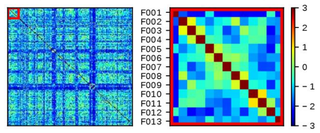
話者間類似度の大規模な主観スコアリングの結果を用いた DNN 話者埋め込みの学習法

主観スコアリングのインストラクション (4,000 名以上が参加)

類似度スコアのグラフ表現 (話者グラフ)

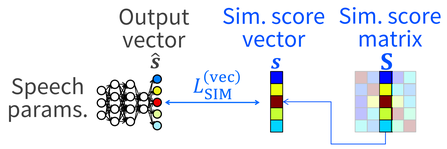


類似度スコアの行列表現 (類似度スコア行列 $\mathbf{S}$)



類似度スコアベクトル埋め込み

音声特徴量から行列 $\mathbf{S}$ のベクトルを予測

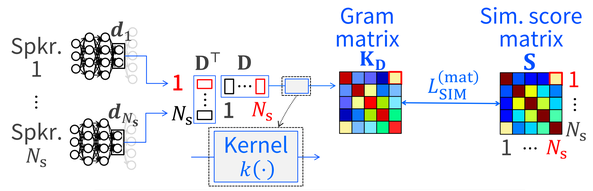


$$L_{\mathrm{SIM}}^{(\mathrm{vec})}(\boldsymbol{s},\hat{\boldsymbol{s}}) = \frac{1}{N_{\mathrm{S}}}(\hat{\boldsymbol{s}}-\boldsymbol{s})^{\top}(\hat{\boldsymbol{s}}-\boldsymbol{s})$$

N_{S}: 主観スコアリングに用いた話者数

類似度スコア行列埋め込み

話者埋め込みの Gram 行列と行列 $\mathbf{S}$ の差を最小化



$$L_{\mathrm{SIM}}^{(\mathrm{mat})}(\mathbf{D},\mathbf{S}) = \frac{1}{Z_{\mathrm{S}}}\left\|\tilde{\mathbf{K}}_{\mathbf{D}}-\tilde{\mathbf{S}}\right\|_F^2$$

Z_{S}: 話者数に依存する正規化係数
$\tilde{\mathbf{K}}_{\mathbf{D}}, \tilde{\mathbf{S}}$: $\mathbf{K}_{\mathbf{D}}, \mathbf{S}$ の対角成分を除外した行列

VAE[2] を用いた多数話者音声生成モデリング[3]

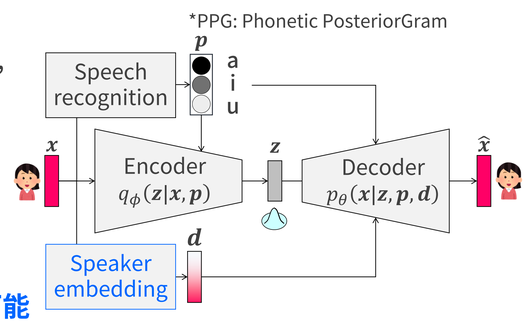
学習済み音声認識・ 話者埋め込み DNN の導入により, 音韻性・話者性を分離して音声特徴量 $\boldsymbol{x}$ をモデル化

Encoder $q_\phi(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{p})$:
$\boldsymbol{x}$ と PPG*[4] $\boldsymbol{p}$ から潜在変数 $\boldsymbol{z}$ を抽出

Decoder $p_\theta(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{p},\boldsymbol{d})$:
$\boldsymbol{z}$ と $\boldsymbol{p}$ と話者埋め込み $\boldsymbol{d}$ から合成音声特徴量 $\hat{\boldsymbol{x}}$ を生成

学習に使われなかった話者の合成音声特徴量も生成可能

*PPG: Phonetic PosteriorGram



Research questions

Q1: 主観的話者間類似度を考慮した話者埋め込みは, 合成音声の品質を改善するか？

本発表では, **話者認識に基づく手法**[5] と **主観的話者類似度に基づく手法** を比較

Q2: 類似度スコア行列埋め込みの学習に適したカーネル関数は何か？

本発表では, **sigmoid カーネル** と **Gauss カーネル** を比較

3. 実験的評価

話者埋め込み・多数話者音声合成共通の条件

データセット	JNAS[6] 女性話者 153 名, 16 kHz サンプリング 評価話者: F001—F013 の 13 名 学習話者: 上記 13 名以外の 140 名
類似度スコア行列	先行研究[1]と同様: −3 (似てない) 〜 +3 (似ている)
音声分析・合成	STRAIGHT ボコーダ[7]
DNN の入力特徴量	1—39 次メルケプと動的特徴量 (Δ)
DNN の最適化	学習率を 0.01 とした AdaGrad[8]
比較手法	(1) d-vec. : 話者認識に基づく埋め込み[5] (2) Sim. (vec): 類似度スコアベクトル埋め込み (3) Sim. (mat): 類似度スコア行列埋め込み (4) Sim. (mat-re): 同上 (類似話者対のみを考慮)

話者埋め込みの条件

話者埋め込みの次元	8
Sim. (mat) のカーネル関数	Sigmoid: $k(\boldsymbol{d}_i,\boldsymbol{d}_j) = \tanh(\boldsymbol{d}_i^{\top}\boldsymbol{d}_j)$ Gauss: $k(\boldsymbol{d}_i,\boldsymbol{d}_j) = \exp\left(-\left\|\boldsymbol{d}_i-\boldsymbol{d}_j\right\|^2\right)$
話者埋め込みの推定	当該話者による発話の有声区間の平均

多数話者音声合成の条件

PPG / 潜在変数の次元	43 / 64
VAEの学習基準	音声特徴量の対数尤度の変分下限[3]
F_0, 非周期成分, 0次メルケプ	入力音声のものをそのまま使用
音声パラメータ生成	最尤パラメータ生成[9]

合成音声の品質に関する主観評価 (スコアの詳細は原稿参照)

A ＞ B: 手法 A のスコア ＞ 手法 B のスコア ($p < 0.05$) となった評価話者数
A ＜ B: 手法 A のスコア ＜ 手法 B のスコア ($p < 0.05$) となった評価話者数
A ≒ B: 手法間のスコアに有意差がなかった評価話者数

(1) 話者埋め込み学習法の比較 (計 1,950 名が評価)

プリファレンスABテスト (自然性)

A vs. B	A ＞ B	A ＜ B	A ≒ B
d-vec. vs. Sim. (vec)	0	12	1
d-vec. vs. Sim. (mat)	0	7	6
d-vec. vs. Sim. (mat-re)	0	8	5

プリファレンスXABテスト (話者類似性)

A vs. B	A ＞ B	A ＜ B	A ≒ B
d-vec. vs. Sim. (vec)	0	10	3
d-vec. vs. Sim. (mat)	4	4	5
d-vec. vs. Sim. (mat-re)	4	2	7

結果: Sim. (vec) は, ほぼ全評価話者で自然性・話者類似性を有意に改善
Sim. (mat) は, 一部の評価話者で話者類似性を有意に劣化 → 学習データへの依存性大？

(2) Sim. (mat) のカーネル関数の比較 (計 1,300 名が評価)

プリファレンスABテスト (自然性)

A vs. B	A ＞ B	A ＜ B	A ≒ B
Sigmoid vs. Gauss	11	0	2

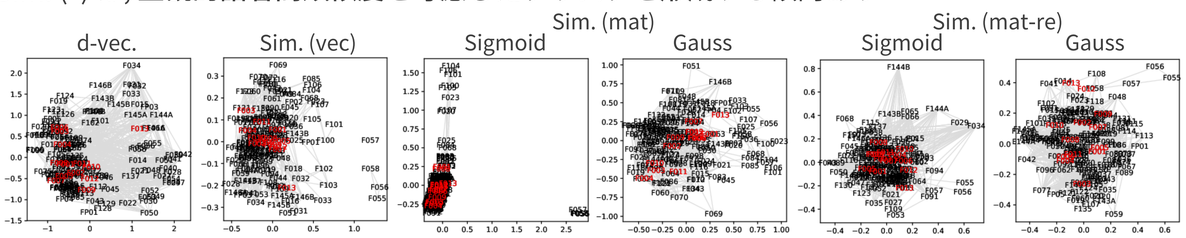
プリファレンスXABテスト (話者類似性)

A vs. B	A ＞ B	A ＜ B	A ≒ B
Sigmoid vs. Gauss	13	0	0

結果: Gauss カーネルを用いた Sim. (mat) は, ほぼ全評価話者で自然性・話者類似性を有意に劣化
Sim. (mat-re) に関しては, 一部の評価話者で有意な改善 (詳細は原稿参照)

話者埋め込みの 2 次元 PCA プロット

Sim. (*) は, 主観的話者間類似度を考慮したクラスタを形成する傾向あり



[参考文献] [1] 齋藤 他, *音講論 (春)*, 2019. [2] Kingma et al., *arXiv:1312.6114*, 2013. [3] Saito et al., *Proc. ICASSP*, 2018. [4] Sun et al., *Proc. ICME*, 2016. [5] Variani et al., *Proc. ICASSP*, 2014. [6] Itou et al., *Journal of the ASJ (E)*, 1999. [7] Kawahara et al., *Speech Communication*, 1999. [8] Duchi et al., *Journal of Machine Learning Research*, 2011. [9] Tokuda et al., *Proc. ICASSP*, 2000.

[謝辞] 本研究の一部は，総務省の委託「知覚モデルに基づくストレスフリーなリアルタイム広帯域音声変換の研究」, セコム科学技術支援財団および JSPS科研費18J22090, 17H06101の助成を受け実施した．