

2-P-30 音素事後確率と d -vector を用いたノンパラレル多対多VAE音声変換における学習データ量と d -vector 次元数に関する評価

☆中村 泰貴 (東京大学), 齋藤 佑樹 (東大院・情報理工), 西田 京介, 井島 勇祐 (NTT), 高道 慎之介 (東大院・情報理工)

1. 概要

■ ノンパラレル多対多VAE音声変換^[1]

- ✓ 音声変換：入力音声の言語情報を保持しつつ、**話者性を変換**
→ 音声変換モデルとして、Variational AutoEncoder (VAE)^[2]を利用
- ✓ 言語情報：学習済み音声認識モデルから得られる音素事後確率^[3]で保持
- ✓ 話者情報：学習済み話者認証モデルから得られる **d -vector**^[4] で表現
- ✓ 特徴：パラレルデータが不要 & 任意話者対での変換が可能

■ 疑問

- ? 最適な **d -vector の次元数** & モデル学習に用いる**話者数**
→ 高品質な音声変換を実現する適切な **d -vector 次元数** & **話者数**とは?

■ そこで今回は

大規模主観評価により、各モデルの学習に用いる**話者数と d -vector 次元数が最終的な変換音声の品質に与える影響を調査**

■ 結果として

- ? 最適な **d -vector の次元数**
○ 低次元であるほど良い
- ? モデル学習に用いる最適な**話者数**
○ 多いほどよい

2. ノンパラレル多対多VAE音声変換

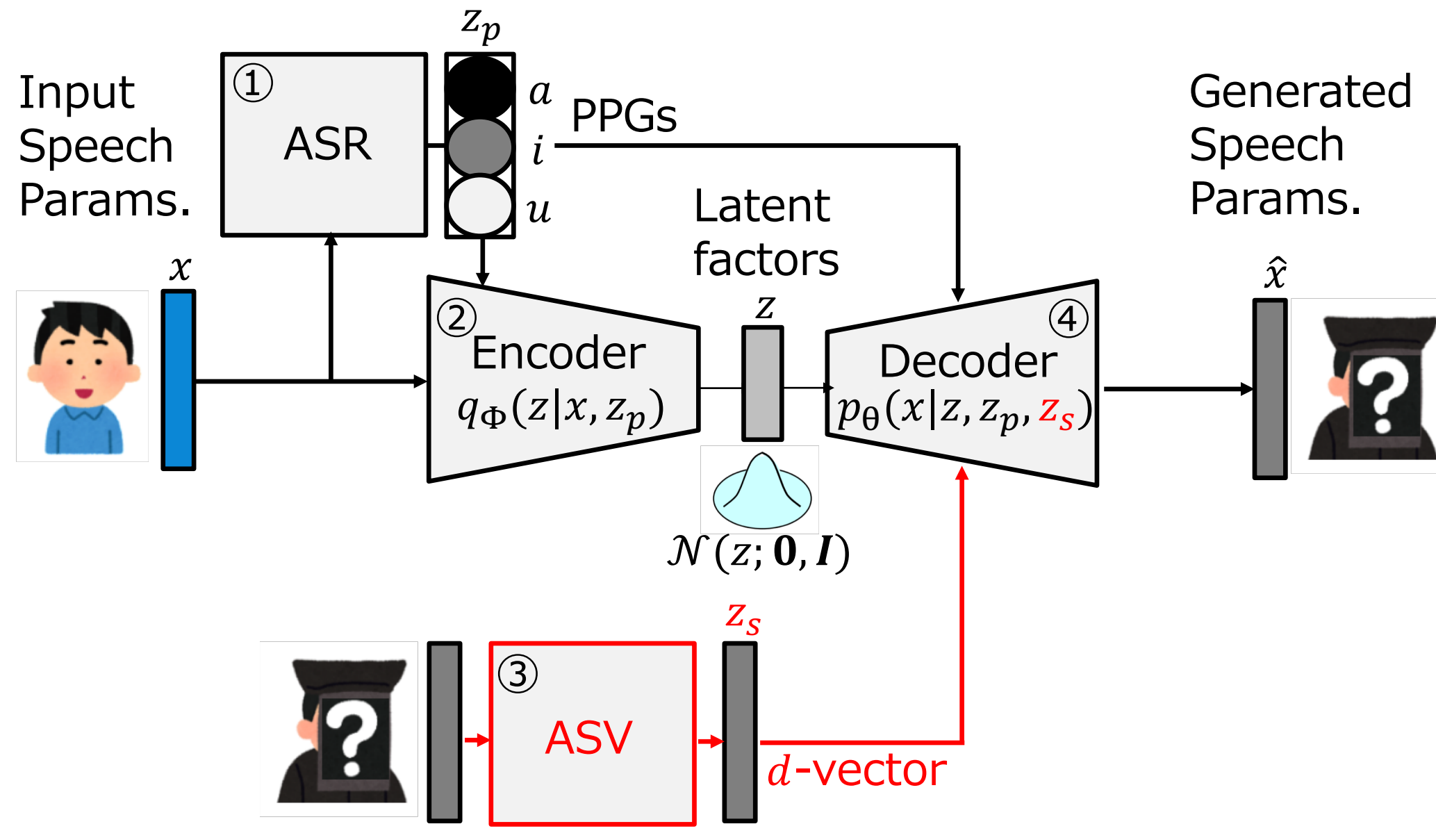
■ d -vector

- 事前学習された話者認証 (Automatic Speaker Verification: ASV) モデルの中間表現 (ボトルネック特徴量) として得られる
- 学習データに含まれない変換先話者 (未知話者) の**話者性を推定・表現**できる
→ 任意話者対での変換が可能になる
- ASVの**事前学習の話者数が多いほど**、 **d -vector がより多様な話者を表現可能**となることが期待できる
→ 変換音声の品質の向上が期待できる
- **d -vector の次元数を増やすことでASVの性能は改善**すると予想されるが、話者空間における話者の分布がより**疎**になることも考えられ、変換音声の品質向上に寄与するかは疑問である

■ 音素事後確率 (Phonetic PosteriorGram: PPG)

- 事前学習された音声認識 (Automatic Speech Recognition: ASR) モデルから得られる
- 入力音声の話者非依存な言語情報を推定・表現できる
→ 入力音声の音韻性を強く保持した、話者情報の変換が可能になる
- ASRの**事前学習の話者数が多いほど**、**話者非依存な特徴量**となることが期待できる
→ 変換音声の品質の向上が期待できる

■ 音声変換の流れ



- ① 入力音声の発話内容を z_p によって保持
- ② 入力された音声パラメータと z_p から潜在変数 z を抽出
- ③ 学習データに含まれない**未知話者の話者表現**を z_s により推定
- ④ z , z_s , z_p を用いて音声パラメータを復元

■ 最適化される目的関数 (変分下限)

$$-D_{KL}(q_{\phi}(z|x, z_p)||p_{\theta}(z)) + \mathbb{E}_{q_{\phi}(z|x, z_p)}[\log p_{\theta}(x|z, z_p, z_s)]$$

↓
zに対する制約

↓
xの二乗再構成誤差

3. 実験的評価

実験条件

音声コーパス	■ 音声認識/話者認証/VAE音声変換モデル学習用 {50spk (男女25名ずつ), 130spk, 260spk} ■ VAE音声変換モデル評価用 3名 (男2名, 女1名) による425発話を含み、内25発話を評価に用いる
音声パラメータ	■ スペクトル特徴量 → 40次元のメルケプストラム (0次は変換せず) ■ 音源特徴量 → 対数F0, U/V, 10帯域の非周期成分
変換話者対	■ {m2m (男性話者から男性話者への変換), m2f (男性話者から女性話者への変換) }
d -vector次元数	■ {8d (8次元), 16d, 32d } ※ 評価データ以外の200発話により推定する
評価方法	■ 変換音声の自然性 → プリファレンスABテスト ■ 変換音声の話者類似性 → プリファレンスXABテスト ※ 各評価の受聴者数は25人 ※ 1人あたり10サンプル

d -vector の次元数・学習話者数に関する評価結果

※ 原稿での値に一部誤りあり

■ d -vector 次元数に対する自然性

spk		8d vs. 16d	16d vs. 32d	8d vs. 32d
50	m2m	0.548 - 0.452	0.516 - 0.484	0.596 - 0.404
	m2f	0.504 - 0.496	0.536 - 0.464	0.536 - 0.464
130	m2m	0.472 - 0.528	0.580 - 0.420	0.568 - 0.432
	m2f	0.504 - 0.496	0.584 - 0.416	0.548 - 0.452
260	m2m	0.516 - 0.484	0.556 - 0.444	0.568 - 0.432
	m2f	0.564 - 0.436	0.492 - 0.508	0.640 - 0.360

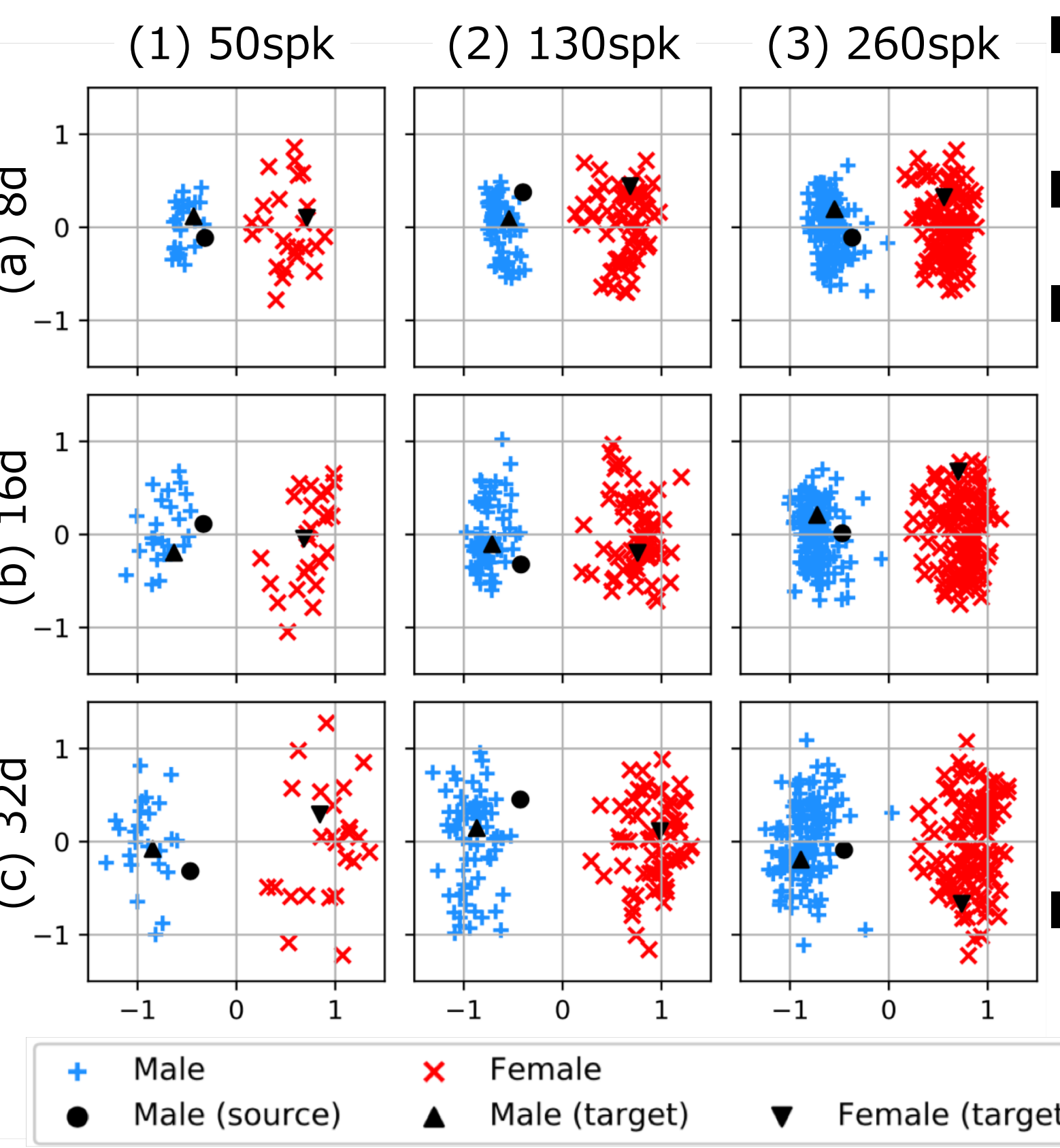
- **d -vector 次元数を減らすほど**
☆ 多くの場合で**自然性が向上**

■ 学習話者数に対する自然性

	50spk vs. 130spk	130spk vs. 260spk	50spk vs. 260spk
m2m	0.496 - 0.504	0.472 - 0.528	0.400 - 0.600
m2f	0.488 - 0.512	0.476 - 0.524	0.380 - 0.620

- **学習に用いる話者数を増やすほど**
☆ 全パターンで**自然性が向上**

d -vector の可視化結果



- 各話者の d -vector を主成分分析により圧縮・可視化した
- 各点は、それぞれの話者に対応する
- **学習話者数を増やすと**
☆ 話者空間での話者の分布がより**密**になった
☆ 未知話者に対し、頑健性のある学習が可能になった
→ 音声品質の向上に繋がったと考えられる
- **d -vector の次元数を増やすと**
☆ 話者空間での話者の分布がより**疎**になっている可能性がある
→ 音声品質の劣化に繋がったと考えられる

