音声対話システムのテキスト音声合成における 声質変換とx-vector埋め込みを用いた感情制御方式の検討

小原 俊一 岡山大学 阿部研究室

OKAYAMA UNIVERSITY

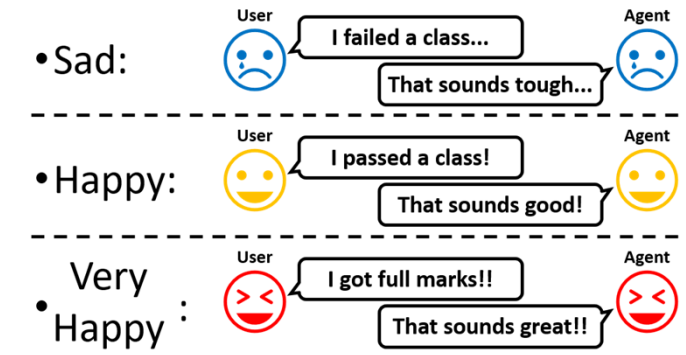
1. 研究背景・目的

音声対話システム（SDS）

- 音声を用いて機械と対話するシステム
  - ✓ 音声の**非言語情報**（感情・韻律など）が重要
  - ✓ **これを欠くとシステムの応答合成音声は無味乾燥に**

目的

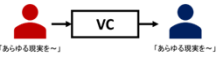
□ **感情とその強さをユーザに合わせた音声合成**



x-vector

- 話者認識モデルから抽出される話者埋め込みベクトル
- **話者性**に加えて，**感情情報**も内包
- **SDSでは，入力音声と合成音声で話者が異なる**

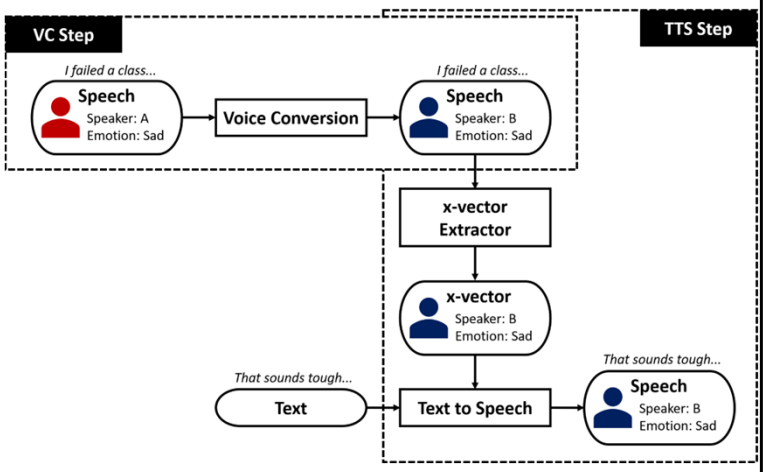
声質変換（Voice Conversion: VC）



提案方式

□ **VCで話者性を変更し，x-vectorで感情を付与するTTS**

2. 提案方式



VC Step

- 感情を保持したまま話者性をTTSモデルの話者へ変換

TTS Step

- 変換音声のx-vectorで感情表現を付与

3. データセット

使用データセット：STUDIES

- 講師（女性）と{男子，女子}生徒の模擬対話音声
- 発話ごとに感情ラベルが振られており，平静（Neutral）と喜び（Happy），悲しみ（Sad）を評価

4. 評価実験

VCアーキテクチャ：VITS

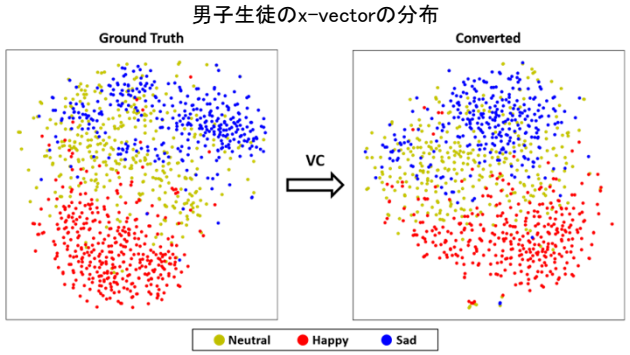
- 男子生徒と女子生徒，講師で学習，VCとして使用

TTSアーキテクチャ：X-VITS (ESPnet2)

- 講師の感情音声で学習

VC Stepに関する分析

- VC後も概ね感情が保持されている



感情の制御に関する感情認識テスト

- 変換音声のx-vectorで感情を制御可能

① OneHot-EmoID

Correct emotions	Subject-perceived emotions		
	Neutral	Happy	Sad
Neutral	**0.940**	0.000	0.060
Happy	0.120	**0.880**	0.000
Sad	0.110	0.000	**0.890**

② Teacher-GT

Subject-perceived emotions		
Neutral	Happy	Sad
0.970	0.020	0.010
0.000	**1.000**	0.000
0.080	0.000	**0.920**

③ M-student-Converted (proposed)

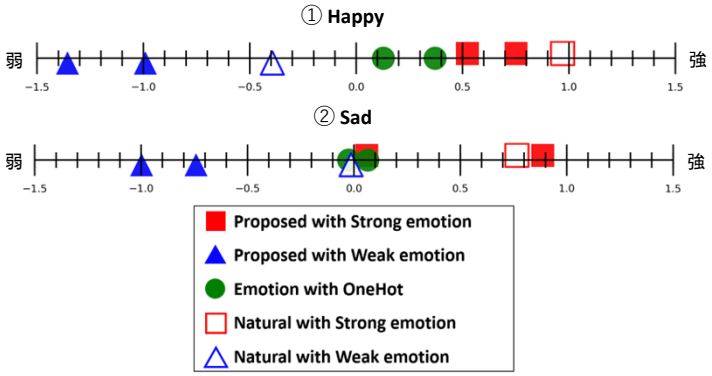
Correct emotions	Subject-perceived emotions		
	Neutral	Happy	Sad
Neutral	**0.860**	0.030	0.110
Happy	0.000	**1.000**	0.000
Sad	0.370	0.000	**0.630**

④ F-student-Converted (proposed)

Subject-perceived emotions		
Neutral	Happy	Sad
0.870	0.040	0.090
0.040	**0.960**	0.000
0.730	0.030	**0.240**

感情の強さに関する一対比較実験

- 喜びも悲しみも強さを表現可能



5. まとめと今後の課題

まとめ

- 声質変換とx-vector埋め込みで感情とその強さを合成音声に反映させる方式の検討
- 変換音声のx-vectorで感情とその強さを制御可能

今後の課題

- 提案方式を音声対話システムに用いた場合の評価
- ant-to-one VCやSERの事後確率で条件付ける方式の検討