DO PROSODY TRANSFER MODELS TRANSFER PROSODY?

- ICASSP 2023
- CSTR, University of Edinburgh
- Paper

Overview

- 현재 Prosody Transfer
  - 훈련에서는 Text-reference speech 를 같은 source speech에서 선정
  - Inference 는 text와 reference speech를 다른 source에서 선정
  - 따라서, 훈련과 inference 간의 Inconsistency가 있음

- 제안하는 것
  - 훈련에서도 text와 reference speech를 다른 Source 에서 선정해보자
    - Reference speech를 선정하는 2가지 방법 제안
      - Text-based
      - F0-based
  - 결과적으로 기존 방법보다 성능이 좋지 않았음
  - 이는 현재 Prosody Transfer model들이 transferable 한 prosody representation을 학습하는 게 아니라, reference speech 와 주어진 text (즉, reference speech의 Text) 모두에 Highly dependent한 utterance-level representation을 학습한다고 볼 수 있음

Reference Speech 선정 방법

- **Text-based method**
  - 같은 text를 target speaker와는 다른 화자가 읽은 것을 reference speech로 선정
  - 이렇게 선정한 reference speech의 prosody가 target speech와 완전 동일하지는 않지만, informative하다고 가정

- **F0-based method**
  - DTW(Dynamic Time warping)를 이용하여 각 utterance들을 align 한 다음, F0 similarity 를 계산
  - 가장 F0 Similarity가 높은 음성을 reference speech로 선정

Backbone – Daft-Exprt

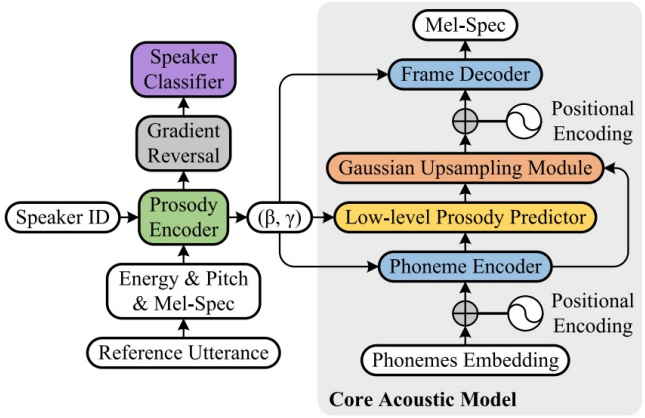

Figure 1: *Daft-Exprt architecture*

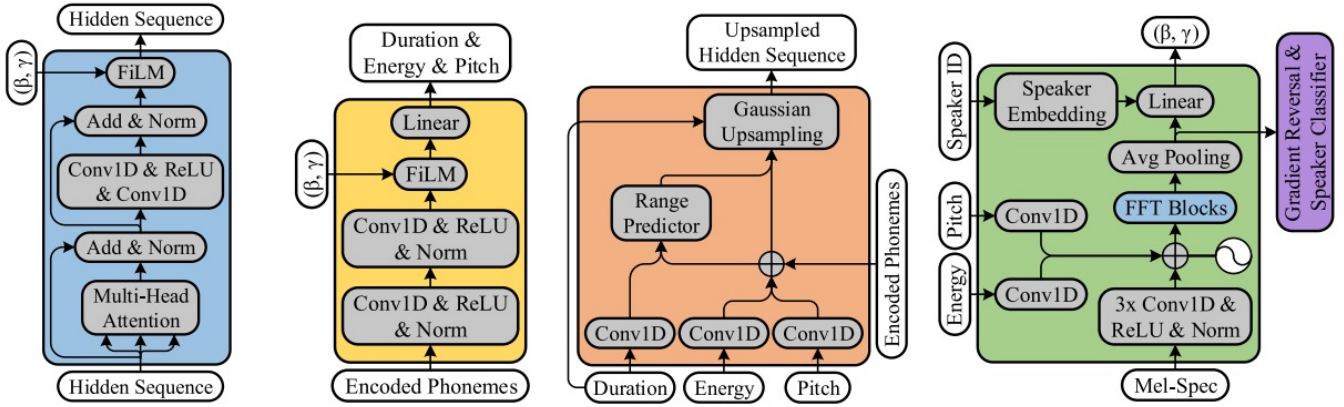

Figure 2: *Daft-Exprt modules*

- ***Daft-Exprt: Cross-Speaker Prosody Transfer on Any Text for Expressive Speech Synthesis***
  - **InterSpeech 2022 / Ubisoft La Forge, Montreal, Canada**
- **Reference speech의 mel-spectrogram 뿐만 아니라 energy, pitch 와 같은 feature 도 활용**

Experiments

- 훈련 데이터 : **Parallel audiobook corpus**
  - **From LibriVox**
  - **Parallel : every sentence is ready by multiple speakers**

- **4가지 reference 선정 방식 비교 (in training phase)**
  - **Daft-Exprt : target speech를 reference speech 로 이용 (기존 방식)**
  - **Shuffle : random 으로 reference speech 선정하여 훈련**
  - **Text-based**
  - **F0-based**

Result

- **Naturalness**

Table 1. MOS results for both real and synthesized samples. Target-speakers are chosen at random.

Model	MOS	
Ground Truth	4.2 ± 0.1	
Ground Truth + HiFi-GAN	3.7 ± 0.1	
	Same-text	**Different-text**
shuffle	2.8 ± 0.2	2.9 ± 0.2
text-based	2.6 ± 0.2	2.4 ± 0.2
F_0-based	2.9 ± 0.2	2.6 ± 0.2
Daft-Exprt	3.2 ± 0.2	2.4 ± 0.2

Same-text / Different-text

4가지 refernce speech 선정 방식은 훈련 시의 reference speech 선정 방식

Same-text / Different-text 는 inference 시 reference speech 선정 방식을 의미함

Same-text는 target speech 와 같은 text를 읽은 음성
Different-text는 target speech 와 다른 text를 읽은 음성

Target speaker는 Random으로 선정

Daft-Exprt는 훈련과 다른 방식으로 Inference 했을 때 (different-text) 성능이 많이 떨어짐

Shuffle, text-based, F0-based 방식은 same-text/different-text 방식에 따른 성능 차이가 미미함

기존 방식으로 훈련했을 때, same-text setting 에서만 성능이 유의미하게 좋음

Result

- **Prosodic Similarity, Speaker Preservation**

Table 2. PT MUSHRA-like scores and target-speaker classification accuracy. Target-speakers are randomly sampled.

Model	MUSHRA-like		Speaker
	Same-text	Different-text	classif.
shuffle	39.0 ± 4.2	25.5 ± 3.1	91.5%
text-based	38.7 ± 4.4	30.4 ± 3.2	88.4%
F_0-based	42.9 ± 4.5	28.4 ± 3.1	91.0%
Daft-Exprt.	61.5 ± 4.8	49.3 ± 4.1	46.1%

Speaker Classification Accuracy
: AXY test

평가자들이 듣고, 합성음 A가
Target speaker X에 가깝다고 생각하면 1점
Reference speaker Y에 가깝다고 생각하면 0점

⇒ Subjective evaluation

MUSHRA-like: prosodic similarity 를 측정, speaker classification 은 speaker preservation을 측정

Natualness와 마찬가지로 prosodic similarity, speaker preservation에서도 기존 방식(Daft-Exprt)로 훈련하고 추론했을 때 가장 성능이 좋음

특히, 다른 reference speech로 훈련했을 때(shuffle, text-basec, F0-based)도 same-text setting에서 유사성 성능이 더 좋았음

특히, Daft-Exprt는 speaker classification 성능이 많이 떨어짐 => speaker-leakage

Result

- **F0 Similarity (Objective)**

Table 3. A normalized DTW-based metric indicates how well F_0 contours align with the reference F_0 contour (lower is better, 0 indicating perfect alignment, 1 indicates worst alignment). Mean absolute F0 error indicates how well each model preserves the target-speaker identity.

	F_0 DTW error		Mean F_0 target error	
	same spkr.	diff spkr.	same spkr.	diff spkr.
shuffle	0.60	0.90	19.9 Hz	20.6 Hz
text-based	0.60	0.95	16.9 Hz	18.2 Hz
F_0-based	0.50	0.85	25.7 Hz	20.9 Hz
Daft-Exprt	0.35	0.45	25.4 Hz	43.5 Hz

target speaker와 비교

DTW error
F0 contour를 DTW로 align 했을 때 error
(낮을 수록 더 좋음)

MAE on F0 target

모든 훈련 방식에 대해 target speaker와 같은 speaker의 음성을 reference 로 줬을 때 성능이 더 좋음

Result

- **Reference speech 선정 방식 비교**

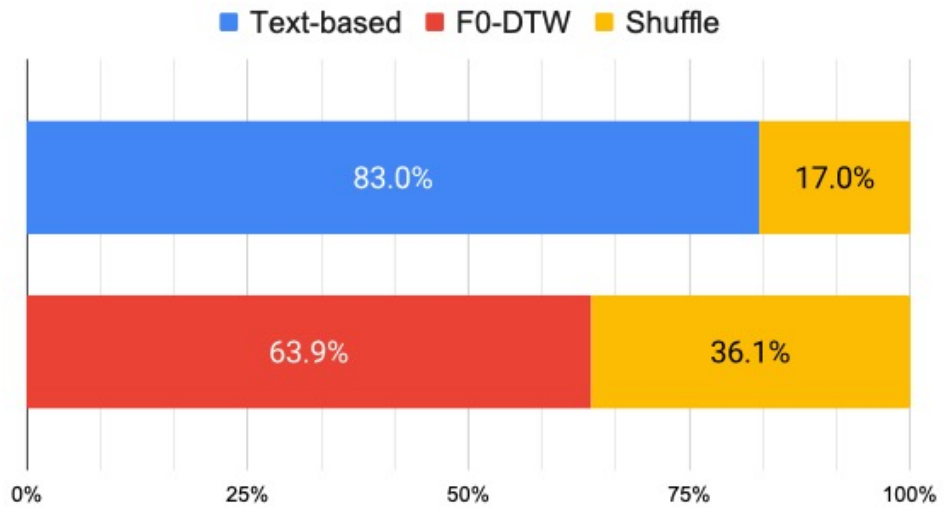

Fig. 2. Our evaluation task shows that evaluators prefer the two proposed methods over randomly sampled utterances.

앞에서 성능이 **shuffle**이랑 비슷하게 나와서 더 실험해 봄

A/B preference test했을 때,
Text-based, F0-based 모두 **shuffle** 보다 좋았음

Different-text setting에서 평가

Conclusion

- Target speech 와 다른 reference speech를 사용해서 훈련하면 성능이 떨어짐
- Prosody Transfer 방식이 실제로 prosody를 transfer한다면, target speech와는 다른 reference speech를 이용해서 훈련해서도 성능이 잘 나와야 하지만, 그렇지 않았음
- Prosody Transfer 방식은 transferable representation 보다는 reference speech의 speaker와 content 에 high dependent한 정보를 encode한다고 볼 수 있음
  - Same-text, Same-speaker setting에서 가장 성능이 좋았음
  - Prosodically similar한 reference speech를 이용해서 훈련하는 방식도 실험 해봤지만 그 성능은 좋지 않았음
- 결론적으로 prosody tranfer가 사용하는 방식은 transferable 한 prosodic information을 encode한다고 볼 수 없음