PDAugment: Data Augmentation by Pitch and Duration Adjustment for Automatic Lyrics Transcription

[1]Chen Zhang , [1]Jiaxing Yu, [1]Luchin Chang, [2]Xu Tan, [3]Jiawei Chen, [2]Tao Qin, [1]Kejun Zhang

[1]Department of Computer Science, Zhejiang University, [2]Microsoft Research Asia, [3]Soutch China University of Technology

Motivation & Background

- Automatic Lyrics Transcription (ALT) is widely used in many MIR related tasks and applications;
- ALT system requires a large amount of paired singing voice data;
- Collecting enough singing voice data costs a lot, while collecting speech data is easy;
- PDAugment aims to leverage natural speech when training ALT.

Natural speech v.s. Singing voice:

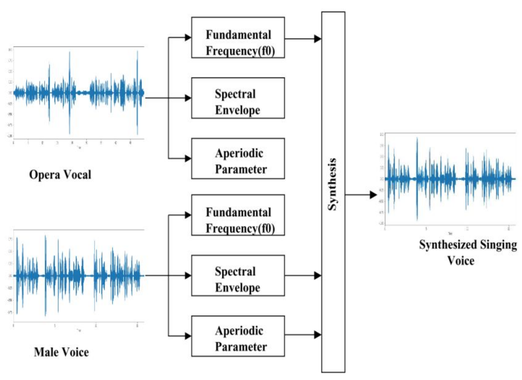
- Pitch (semitone): Pitch Range, Pitch Stability
- Duration (s): Duration Range, Duration Variance

Property	*Speech*	*Singing Voice*
Pitch Range (semitone)	12.71	14.61
Pitch Stability	0.93	0.84
Duration Range (s)	0.44	2.40
Duration Variance	0.01	0.11

Table 1: The differences of acoustic properties between speech and singing voice.

Voice to Singing Style Transfer:

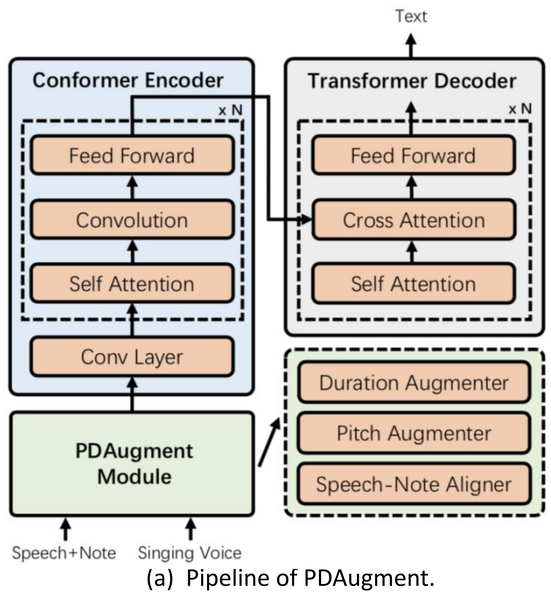
- F0 contour: from opera vocal
- SP & AP: from speech



(a) the proposed V2S system in their paper.

PDAugment

- **Overall Pipeline**:
  - Conformer encoder + Transformer decoder;
  - PDAugment Module:
    - speech-note aligner
    - pitch adjustment
    - duration adjustment
  - Input:
    - speech + note
    - singing voice



(a) Pipeline of PDAugment.

Pitch adjustment

Figure 2: The change of F0 contour before and after pitch augmenter. The content of this example is "opening his door".

Duration adjustment

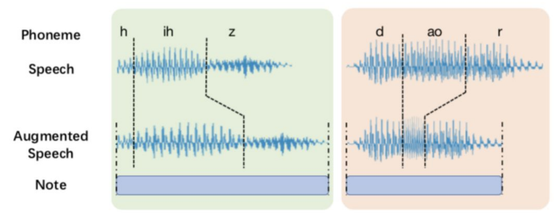


Figure 3: The change of duration before and after duration augmenter. The content of this example is "his door". The left block shows the case of lengthening the duration and the right shows the case of shortening the duration.

Datasets

- **Speech dataset**:
  - LibriSpeech: 960 hours from 2339 speakers.
- **Singing voice dataset**:
  - DSing30: 4K recordings from 3205 users;
  - DALI: 1.2K songs (70 hours).
- **Music score dataset**:
  - Pop music subset of FreeMIDI: 4K midi files.
- **Lyric dataset**:
  - crawl from the web (mojim.com): 17M sentences of pop English song lyrics.

Baselines & Evaluation Metrics

Baselines:

- Naive ALT: trained with only singing voice datasets.
- ASR augmented: add natural speech dataset directly.
- Previous SOTA:
  - DSing30: Dilated convolutional neural networks with self-attention;
  - DALI: V2S style transfer.

Evaluation Metric:

Similar to ASR, use **WER** (word error rate) as the metric.

Experiments

Main Results

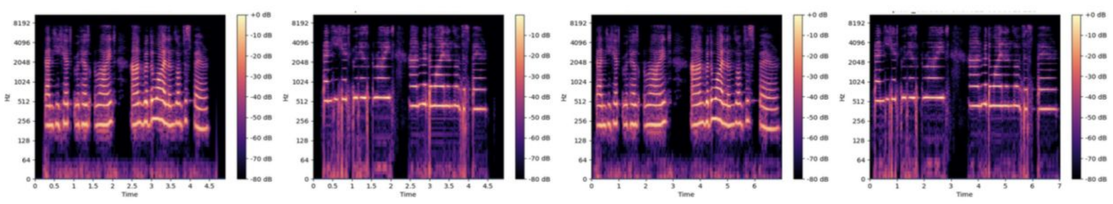
Method	*DSing30 Dev*	*DSing30 Test*
Naive ALT	28.2	27.4
ASR Augmented	20.8	20.5
Previous SOTA [13]	17.7	15.7
PDAugment	**10.1**	**9.8**

Table 2: The WERs (%) of DSing30 dataset.

Method	*Dali Dev*	*Dali Test*
Naive ALT	80.9	86.3
ASR Augmented	75.5	75.7
Previous SOTA [11]	75.2	78.9
PDAugment	**53.4**	**54.0**

Table 3: The WERs (%) of Dali corpus.

Visualization of augmented speech



(a) Original speech sample. (b) After Pitch Augmenter. (c) After Duration Augmenter. (d) After PDAugment.

Figure 4: Spectrograms of speech example after pitch augmenter or/and duration augmenter.

Method Analyses

- Augmentation types: by enable pitch adjustment and duration adjustment separately.
- Random augmentation: adjusts the pitch and duration of speech randomly. (PDAugment adjusts according to existing music score information.)

Setting	*DSing30 Dev*	*DSing30 Test*
PDAugment	**10.1**	**9.8**
- Pitch Augmenter	13.6	13.4
- Duration Augmenter	13.8	13.8
- Pitch & Duration Augmenters	20.8	20.5
Random Aug	17.9	17.6

Table 4: The WERs (%) of different types of augmentation. *Without Aug* means training the ALT model on DSing30 and the original LibriSpeech data.

Summary

- **In this paper**,
  - we developed PDAugment, a data augmentation method by **adjusting pitch and duration**, to make better use of natural speech data for ALT training.
- **PDAugment contains**
  - a **speech-note aligner** to align the speech with note;
  - two **augmenters** to adjust pitch and duration respectively.
- **In the future, we are planning to**:
  - narrow the gap between natural speech and singing voice from more aspects;
  - add some musical-specific constraints in the decoding stage.