Genre-conditioned Acoustic Models for Automatic Lyrics Transcription of Polyphonic Music

Presented by Xiaoxue Gao

Xiaoxue Gao[1], Chitralekha Gupta[1] and Haizhou Li[2]

[1]National University of Singapore, Singapore

[2]The Chinese University of Hong Kong, Shenzhen, China

22-27 MAY 2022

ICASSP 2022

SINGAPORE

Outline

- Introduction
- Related work
- Proposed approach –Genre-conditioned acoustic model
- Experiments and results
- Conclusions

01 Introduction --Lyrics Transcription of Polyphonic Music

- Goal: Automatic lyrics transcription of polyphonic music (ALTP) aims to transcribe lyrics from a song that contains singing vocals mixed with background music.

Polyphonic music audio input

Automatic lyrics transcription of polyphonic music

Close your eyes give me your hands darling

Predicted Lyrics

01 Introduction --Lyrics Transcription of Polyphonic Music

- Applications:
  - Automatic generation of karaoke lyrical content
  - Music video subtitling
  - Query-by-singing
- Challenges:
  - Background music interference
  - Music genre discrepancy problem

Outline

- Introduction
- Related work
- Proposed approach –Genre-conditioned acoustic model
- Experiments and results
- Conclusions

02 Related work --Lyrics Transcription of Polyphonic Music

1) Extraction-transcription approach:

- Singing vocal extraction + transcription

Limitations:

- Feature mismatch between acoustic model training and testing
- Imperfect singing vocal extraction brings artifacts and distortions to the extracted singing

2) Music-aware approach:

- Make use of background music information in polyphonic music

Outline

- Introduction
- Related work
- Proposed approach –Genre-conditioned acoustic model
- Experiments and results
- Conclusions

03 Proposed approach

-- Genre-conditioned acoustic model

Motivation:

- Different genres exhibit significantly different levels of lyrics intelligibility in polyphonic music.
- Genres vary in their musical characteristics such as instrumental accompaniment, singing vocal loudness, syllable rate, and singing style.
- We believe that the predictable genre-class information might help an automated lyrics transcription system with lyrics intelligibility.

03 Proposed approach

-- Genre-conditioned acoustic model

Motivation:

- Different genres exhibit significantly different levels of lyric intelligibility in polyphonic music.
- Genres vary in their musical characteristics such as instrumental accompaniment, singing vocal loudness, syllable rate, and singing style.
- We believe that the predictable genre-class information might help an automated lyrics transcription system with lyrics intelligibility.

Genre broad classes: shared characteristics between music genres that affect lyrics intelligibility.

- **Hiphop**: contain rap with electronic music and higher syllable rate such as Rap, Hip Hop, and Rhythms & Blues. Rap shows lower lyrics intelligibility than Pop.
- **Metal**: loud and dense background music, such as Metal and Hard Rock. "Death metal" shows zero lyrics intelligibility.
- **Pop**: clear and louder vocals, such as Country, Jazz, Reggae etc.

03 Proposed approach

-- Genre-conditioned acoustic model

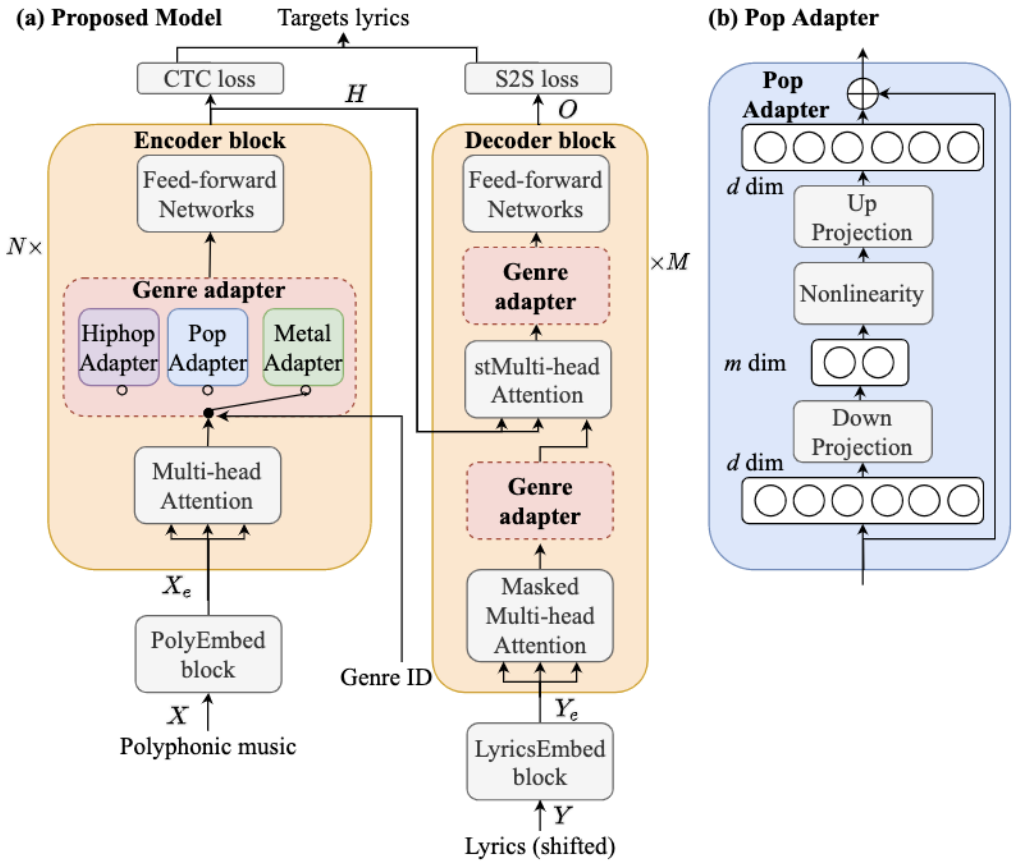
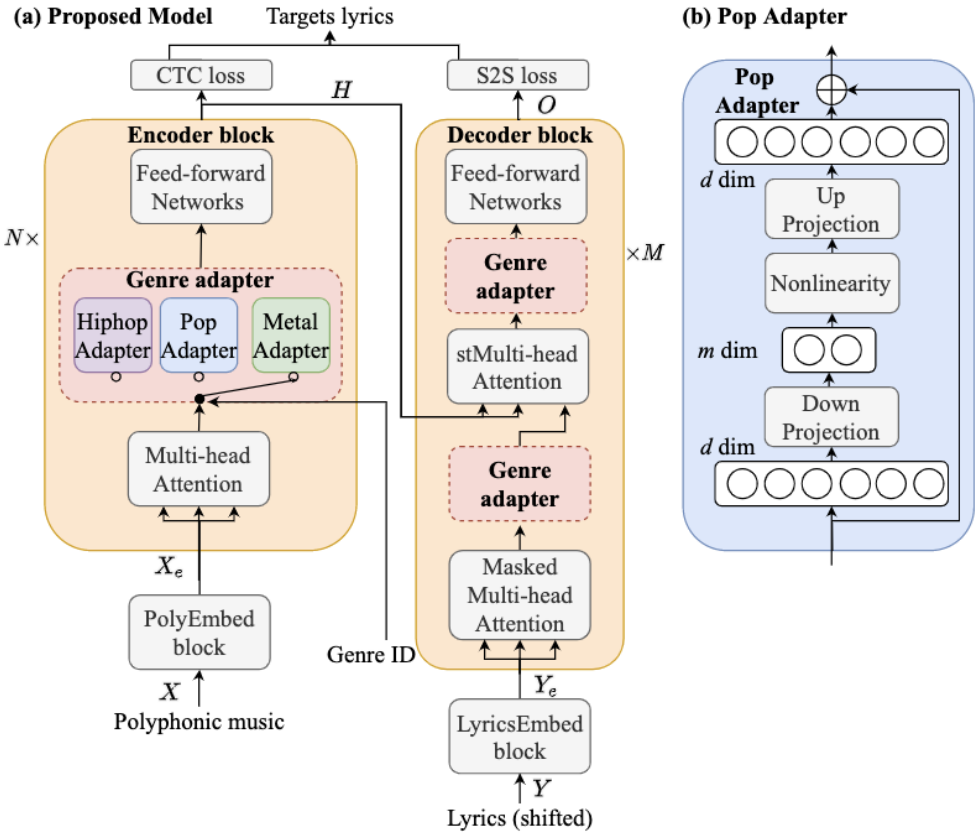


Fig. 1: An overview architectures of (a) the proposed genre-conditioned lyric transcriber with a Transformer architecture; and (b) the pop adapter, which has the same design for metal and hiphop adapters.

03 Proposed approach

-- Genre-conditioned acoustic model

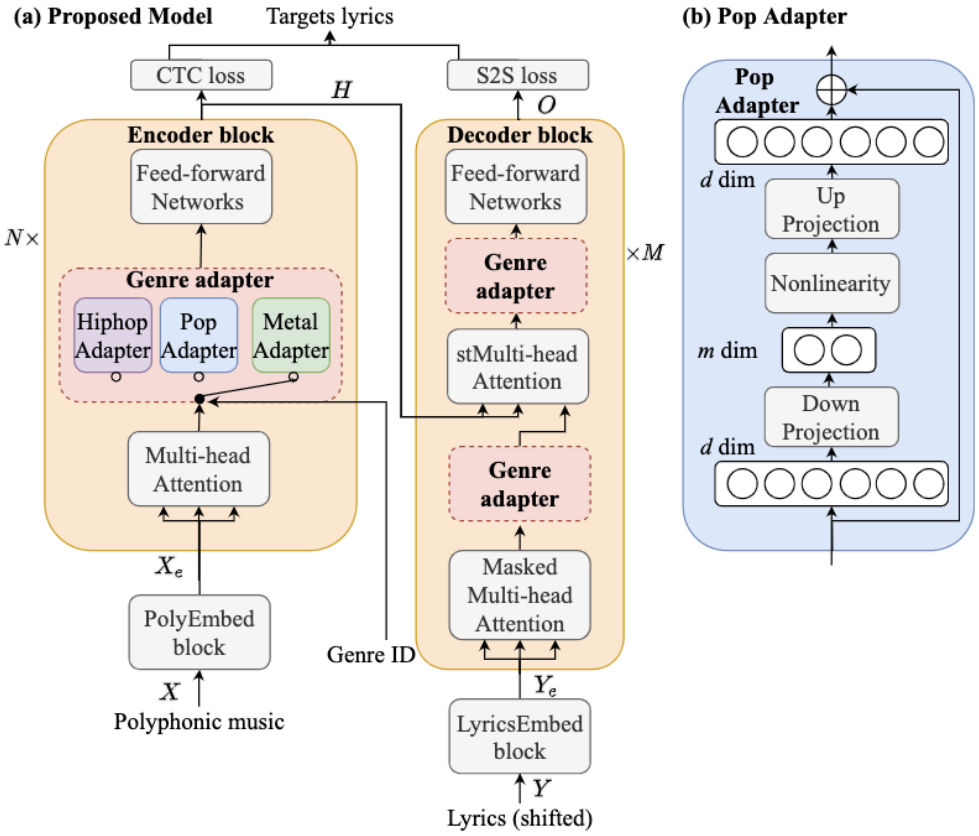


Goal: base model provides holistic polyphonic representation with information that is common across the different genres, while the genre adapters capture the genre-specific characteristics.

Fig. 1: An overview architectures of (a) the proposed genre-conditioned lyric transcriber with a Transformer architecture; and (b) the pop adapter, which has the same design for metal and hiphop adapters.

03 Proposed approach

-- Genre-conditioned acoustic model



Goal: base model provides holistic polyphonic representation with information that is common across the different genres, while the genre adapters capture the genre-specific characteristics.

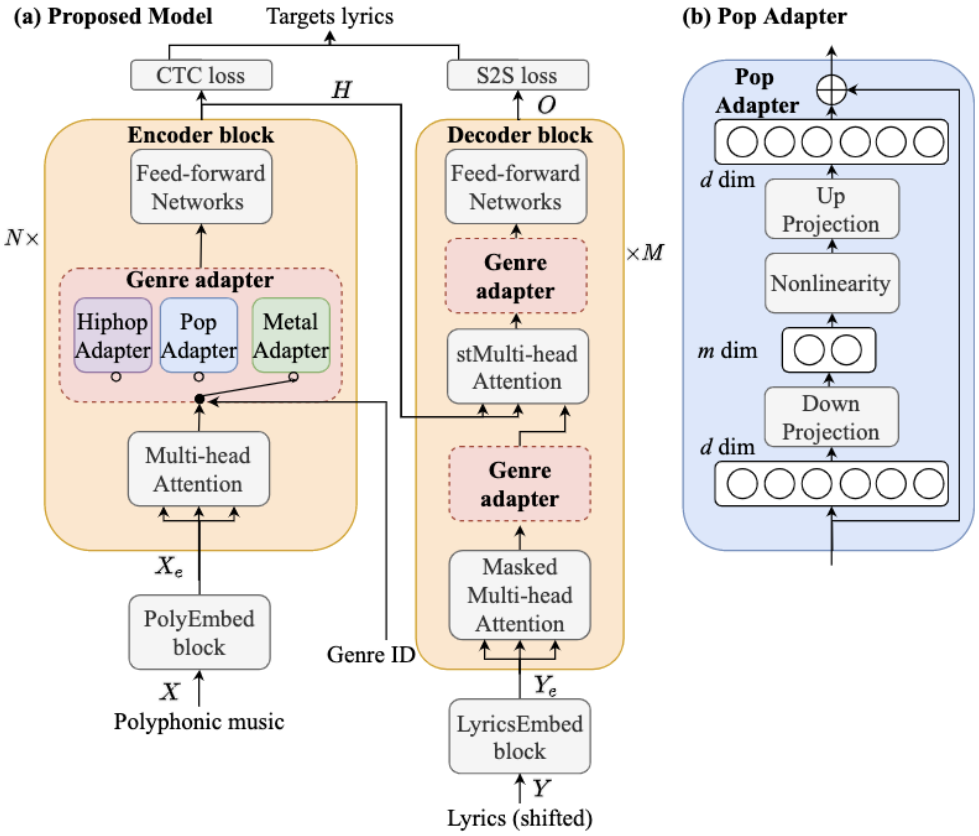
Genre-conditioned Encoder and decoder:

$$\mathbf{X}_e = \text{PolyEmbed}(\mathbf{X}),$$
$$\mathbf{H} = \text{GenreEncoder}(\mathbf{X}_e),$$
$$\mathbf{Y}_e = \text{LyricsEmbed}(\mathbf{Y}),$$
$$\mathbf{O} = \text{GenreDecoder}(\mathbf{H}, \mathbf{Y}_e)$$

Fig. 1: An overview architectures of (a) the proposed genre-conditioned lyric transcriber with a Transformer architecture; and (b) the pop adapter, which has the same design for metal and hiphop adapters.

03 Proposed approach

-- Genre-conditioned acoustic model



Goal: base model provides holistic polyphonic representation with information that is common across the different genres, while the genre adapters capture the genre-specific characteristics.

Genre-conditioned Encoder and decoder:

$$\mathbf{X}_e = \text{PolyEmbed}(\mathbf{X}),$$
$$\mathbf{H} = \text{GenreEncoder}(\mathbf{X}_e),$$
$$\mathbf{Y}_e = \text{LyricsEmbed}(\mathbf{Y}),$$
$$\mathbf{O} = \text{GenreDecoder}(\mathbf{H}, \mathbf{Y}_e)$$

$$\mathcal{L}_{\text{Genre-con}} = \alpha\mathcal{L}^{\text{CTC}} + (1-\alpha)\mathcal{L}^{\text{S2S}},$$
$$\mathcal{L}^{\text{CTC}} = \text{Loss}_{\text{CTC}}(\mathbf{G}_{ctc}, \mathbf{R}),$$
$$\mathcal{L}^{\text{S2S}} = \text{Loss}_{\text{S2S}}(\mathbf{G}_{s2s}, \mathbf{R})$$

Fig. 1: An overview architectures of (a) the proposed genre-conditioned lyric transcriber with a Transformer architecture; and (b) the pop adapter, which has the same design for metal and hiphop adapters.

Outline

- Introduction
- Related work
- Proposed approach –Genre-conditioned acoustic model
- Experiments and results
- Conclusions

04 Experiments and results

Database:

- Training data: 4430 songs
- Validation data: 170 songs
- Testing data: Hansen (10 songs), Jamendo (20 songs) and Mauch (20 songs)

Experimental setup:

- ESPnet [37] with pytorch backend, 83-dim fbank feature
- **Transformer**: 12 encoder blocks, 6 decoder blocks, 8 heads in MHA
- **Genre-adapter**: the down-projection and up-projection layers are linear layers with d = 512 and m = 256, along with a ReLU non-linearity function.

[37] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. Enrique Yalta Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, “Espnet: End-to-end speech processing toolkit,” in INTERSPEECH, 2018, pp. 2207–2211. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-1456

04 Experiments and results

- Genre MHA: that does not have the first decoder adapter which only takes lyrical features as input
- Genre MHA + Mask MHA: has all the three adapters
- Genre MHA Ablation: one common adapter with pop, hiphop and metal parameters shared.

Table 2. The genre distribution for polyphonic music Dataset.

Statistics	Metal	Pop	Hiphop
Percentage in Poly-train	35%	59%	6%
Percentage in Poly-dev	48%	49%	3%
Percentage in Poly-test	34%	56%	10%

Table 3. Comparison between the proposed genre-adapter solutions and other existing competitive solutions to lyrics transcription (WER%) of polyphonic music.

Whole songs test	Hansen	Jamendo	Mauch
DS [10]	-	77.80	70.90
RB1 [14]	83.43	86.70	84.98
DDA2 [39]	74.81	72.15	75.39
DDA3 [39]	77.36	73.09	80.66
CG [31]	-	59.60	44.00
GGL2 [40]	48.11	61.22	45.35
GGL1 [40]	45.87	56.76	43.76
Line-level test	**Metal**	**Pop**	**Hiphop**
GGL1 [40]	59.70	37.07	57.08
Base model [23]	50.04	36.52	**51.19**
Genre MHA	**48.17**	**33.34**	52.32
Genre MHA Ablation	48.05	33.41	55.42
Genre MHA+MaskMHA	48.22	33.86	51.55

04 Experiments and results

- Genre MHA: that does not have the first decoder adapter which only takes lyrical features as input

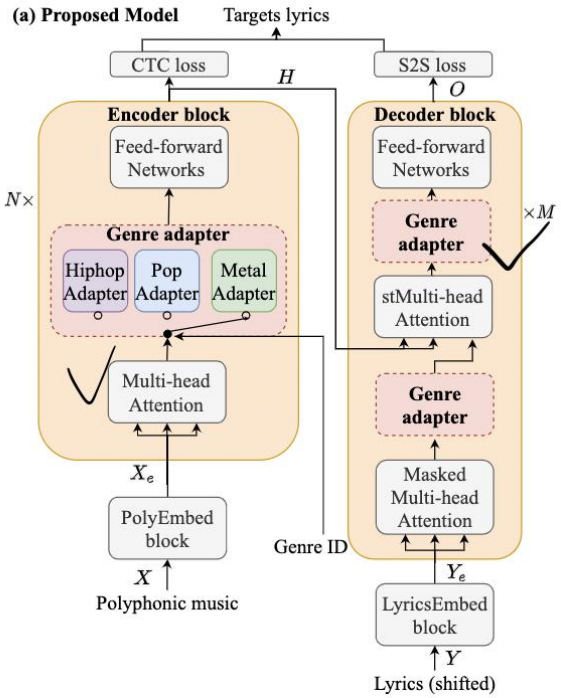


Table 2. The genre distribution for polyphonic music Dataset.

Statistics	Metal	Pop	Hiphop
Percentage in Poly-train	35%	59%	6%
Percentage in Poly-dev	48%	49%	3%
Percentage in Poly-test	34%	56%	10%

Table 3. Comparison between the proposed genre-adapter solutions and other existing competitive solutions to lyrics transcription (WER%) of polyphonic music.

Whole songs test	Hansen	Jamendo	Mauch
DS [10]	-	77.80	70.90
RB1 [14]	83.43	86.70	84.98
DDA2 [39]	74.81	72.15	75.39
DDA3 [39]	77.36	73.09	80.66
CG [31]	-	59.60	44.00
GGL2 [40]	48.11	61.22	45.35
GGL1 [40]	45.87	56.76	43.76
Line-level test	**Metal**	**Pop**	**Hiphop**
GGL1 [40]	59.70	37.07	57.08
Base model [23]	50.04	36.52	**51.19**
Genre MHA	**48.17**	**33.34**	52.32
Genre MHA Ablation	48.05	33.41	55.42
Genre MHA+MaskMHA	48.22	33.86	51.55

Outline

- Introduction
- Related work
- Proposed approach –Genre-conditioned acoustic model
- Experiments and results
- Conclusions

05 Conclusion

- Genre-conditioned Automatic Lyrics Transcription of Polyphonic Music

  - The proposed genre adapters for lyrics-genre pairs in polyphonic music provide genre-related knowledge to help with music interference problem.

  - Integrating genre-adapters with pre-trained models shows the flexibility of using adapters to explore different kinds of music data for the development of lyrics transcription system for polyphonic music.

THANK YOU!

Q&A

Further question please contact:

Xiaoxue Gao
xiaoxue.gao@u.nus.edu