Lyrics-to-Audio Alignment with Music-aware Acoustic Models

Chitralekha Gupta Emre Yılmaz Haizhou Li

Dept. of Electrical and Computer Engineering, National University of Singapore, Singapore,
Emails: {chitralekha,emre,haizhou.li}@nus.edu.sg

Introduction

- **Goal:** To automatically align the lyrical content with mixed singing audio (singing voice+musical accompaniment) at a word level.
- Automatic lyrics alignment in polyphonic music are challenging tasks because the singing vocals are corrupted by the background music.

- We propose music-aware acoustic models that learn music genre-specific characteristics to train polyphonic acoustic models [1].
- With such genre-based approach, we explicitly model the music without removing it during acoustic modeling.

Genre-informed acoustic modeling

- The characteristics of genres that affects lyric intelligibility are **relative volume of the singing vocals compared to the background accompaniment**, **syllable rate**, and **the kinds of the instrumental accompaniments**.
- Genre-informed acoustic modelling of phones and non-vocal sections would capture the combined effect of background music and singing vocals, depending on the genre.
- We train 3 different types of acoustic models corresponding to the **three genre broadclasses** (Table 1), for (a) **genre-informed "silence" or non-vocal models**, and (b) **genre-informed phone models**.

Table 1: Genre broadclasses grouping

Genre Broadclasses	Characteristics	Genres
hiphop	rap, electronic music	Rap, Hip Hop, R&B
metal	loud and many background accompaniments, a mix of percussive instruments, amplified distortion, vocals not very loud, rock, psychedelic	Metal, Hard Rock, Electro, Alternative, Dance, Disco, Rock, Indie
pop	vocals louder than the background accompaniments, guitar, piano, saxophone, percussive instruments	Country, Pop, Jazz, Soul, Reggae, Blues, Classical

Framework Overview

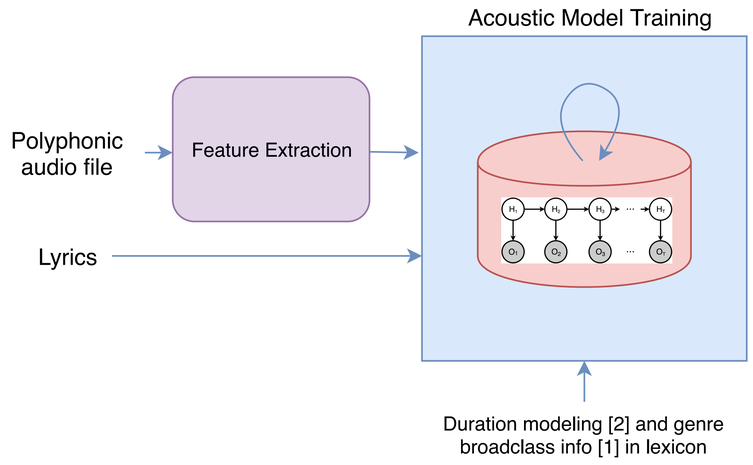

Figure 1: Schematic diagram of genre-informed acoustic modeling.

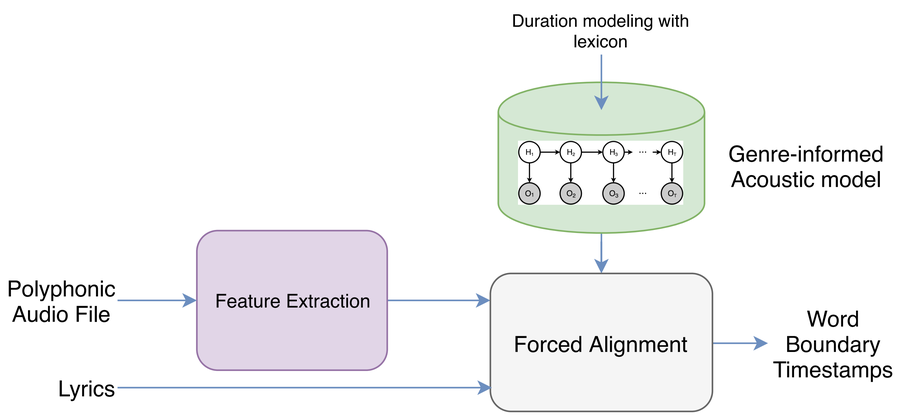

Figure 2: Schematic diagram of lyrics-to-audio alignment at run-time.

Experimental Setup

- The ASR architecture consists of a factorized time-delay neural network (TDNN-F) model with 2 additional convolutional layers and a rank reduction layer; trained according to the standard Kaldi recipe.
- An augmented version of the training data is created by reducing (x0.9) and increasing (x1.1) the speed of each utterance.
- The acoustic model is trained using 40-dimensional MFCCs as acoustic features.
- A duration-based modified pronunciation lexicon is employed [2].
- For the genre-informed phone modeling, we label the phone units in the phonetic lexicon with genre broadclass labels.
- The alignment system chooses the best fitting phone models among all genres during the forced alignment, to prevent the additional requirement of genre information for songs in the testing phase

Training Dataset

Table 2: Training dataset description.

Name	Content	Lyrics Ground-Truth	Genre distribution
DALI [3]	3,913 songs	line-level boundaries, 180,033 lines	hiphop:119, metal:1,576, pop:2,218

Comparison with existing literature

Table 3: Comparison of lyrics alignment (mean absolute word alignment error (seconds)) performance with existing literature.

Test Datasets	MIREX 2017		MIREX 2018	ICASSP 2019		Interspeech2019	
	AK [4]	GD [5]	CW [6]	DS [7]	CG [8]	CG [9]	Ours [1]
Mauch	9.03	11.64	4.13	0.35	6.34	1.93	**0.19**
Hansen	7.34	10.57	2.07	-	1.39	0.93	**0.10**
Jamendo	-	-	-	0.82	-	-	**0.22**

Summary

- Lyrics alignment shows an improvement in performance with genre-informed silence + phone models over those with no genre info [1].
- The mean absolute word alignment error is less than 220 ms across three test datasets – Hansen, Mauch, and Jamendo.
- Our proposed strategies show a way to induce music knowledge in ASR to address the problem of lyrics alignment in polyphonic audio.

References

[1] C. Gupta, E. Yılmaz, H. Li, "Automatic Lyrics Alignment and Transcription in Polyphonic Music: Does Background Music Help?," *arXiv preprint:1909.10200v2 [eess.AS]*, 2019 (submitted to ICASSP 2020).

[2] C. Gupta, H. Li, and Y. Wang, "Automatic Pronunciation Evaluation of Singing," *in Interspeech*, 2018.

[3] G. Meseguer-Brocal, A. Cohen-Hadria, and G. Peeters, "Dali: A large dataset of synchronized audio, lyrics and notes, automatically created using teacher-student machine learning paradigm," *in Proc. ISMIR*, 2018.

[4] A. M. Kruspe, "Bootstrapping a system for phoneme recognition and keyword spotting in unaccompanied singing," *in Proc. ISMIR*, 2016.

[5] G. Dzhambazov and X. Serra, "Modeling of phoneme durations for alignment between polyphonic audio and lyrics," *In 12th Sound and Music Computing Conference*, pp. 281—286, 2015.

[6] C.C. Wang, "Lyrics-to-audio alignment for instrument accompanied singings," *in MIREX 2018*.

[7] D. Stoller, S. Durand, and S. Ewert, "End-to-end Lyrics Alignment for Polyphonic Music Using an Audio-to-character Recognition Model," *in Proc. ICASSP*, 2019.

[8] C. Gupta*, B. Sharma*, H. Li, and Y. Wang, "Automatic lyrics-to-audio alignment on polyphonic music using singing-adapted acoustic models," *in Proc. ICASSP*, 2019. (*equal contributors)

[9] C. Gupta, E. Yılmaz, and H. Li,, "Acoustic modeling for automatic lyrics-to-audio alignment," *in Proc. Interspeech*, 2019.