INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Acoustic Modeling for Automatic Lyrics-to-Audio Alignment

Chitralekha Gupta, Emre Yılmaz, Haizhou Li

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

Abstract

Automatic lyrics to polyphonic audio alignment is a challenging task not only because the vocals are corrupted by background music, but also there is a lack of annotated polyphonic corpus for effective acoustic modeling. In this work, we propose (1) using additional speech and music-informed features and (2) adapting the acoustic models trained on a large amount of solo singing vocals towards polyphonic music using a small amount of in-domain data. Incorporating additional information such as voicing and auditory features together with conventional acoustic features aims to bring robustness against the increased spectro-temporal variations in singing vocals. By adapting the acoustic model using a small amount of polyphonic audio data, we reduce the domain mismatch between training and testing data. We perform several alignment experiments and present an in-depth alignment error analysis on acoustic features, and model adaptation techniques. The results demonstrate that the proposed strategy provides a significant error reduction of word boundary alignment over comparable existing systems, especially on more challenging polyphonic data with long-duration musical interludes.

Index Terms: Lyrics-to-audio alignment, ASR, model adaptation, speech and music informed features

1. Introduction

The goal of an automatic lyrics-to-audio alignment algorithm is the time synchronization between the lyrics and the singing vocals with or without background music. It potentially enables various applications such as generating karaoke scrolling lyrics, music video subtitling, and music retrieval.

The task of lyrics-to-audio alignment is often seen as an extension of the speech-to-text alignment task. ASR systems have been used to force-align lyrics to singing vocals [1–5]. Singing voice, however, covers a much wider range of intrinsic variations than speech both in terms of timbre and fundamental frequencies [6]. One can reduce the mismatch between speech and singing signals by adapting the speech acoustic models with a small amount of singing data using maximum a posterior (MAP) or maximum likelihood linear regression (MLLR) [4,5]. Mesaros et al. [4] used 49 fragments of songs, 20-30 seconds long, along with their manual transcriptions to adapt Gaussian mixture model (GMM)-hidden Markov model (HMM) speech models for singing. These studies provide a direction for solving the problem of lyrics alignment in music, but they suffer from a lack of lyrics annotated data.

Kruspe [7] and Dzhambazov [8] presented systems for the lyrics alignment challenge in MIREX 2017. The acoustic models in [7] were trained using 6,000 songs from the Smule's public solo-singing karaoke dataset called Digital Archive of Mobile Performances (DAMP) [9]. This dataset is collected via a karaoke app, therefore has no consistent recording condition, contains out-of-vocabulary words, and incorrectly pronounced words because of unfamiliar lyrics [5]. Moreover, the dataset does not have lyrics time annotation.

Gupta et al. [5] designed a semi-supervised algorithm to automatically obtain weak line-level lyrics annotation of a subset of approximately 50 hours of solo-singing DAMP data. They adapted DNN-HMM speech acoustic models to singing voice with this data, that showed 36.32% word error rate (WER) in a free-decoding experiment on short solo-singing test phrases from the same dataset. In [10], these singing-adapted models were further enhanced to capture long duration vowels with a duration-based lexicon modification, that reduced the WER to 29.65%. However, acoustic models trained on solo-singing data result in a significant drop in performance when applied to singing vocals in the presence of background music[1]. Singing vocals are often highly correlated with the corresponding background music, resulting in overlapping frequency components [6]. The varied range of voice quality of artists combined with different types of musical instruments makes the problem of lyrics alignment highly challenging in polyphonic music.

To suppress the background accompaniment, some approaches have incorporated singing voice separation techniques as a pre-processing step [1, 4, 8, 11]. However, this step makes the system dependent on the performance of the singing voice separation algorithm, as the separation artifacts may make the words unrecognizable. Moreover, this requires a separate training setup for the singing voice separation system.

Recently, multiple researchers have explored data intensive approaches to lyrics-to-audio alignment. In MIREX 2018, Wang [12] presented a system that achieved a mean alignment error (AE) of 4.12 seconds on a standard test data for word alignment evaluation (Mauch's polyphonic dataset [2]). They used 7,300 annotated English songs from KKBOX Inc.'s music library to train GMM-HMM models. Stoller et al. [13] presented an end-to-end system based on the Wave-U-Net architecture that predicts character probabilities directly from raw audio. The system was trained on more than 44,000 songs with line-level lyrics annotations from the Spotify's music library. They achieved an impressive 0.35s mean AE on the Mauch's dataset. However, end-to-end systems require a large amount of annotated training polyphonic music data to perform well as seen in [13], while publicly available acoustic resources for polyphonic music are limited.

In this study, we explore the use of additional speech and music-informed features, along with the standard acoustic features during the acoustic model training for singing voice. In addition, we adapt an acoustic model trained on a large amount of solo singing vocals using a limited amount of annotated polyphonic data to reduce the domain mismatch. The aim is to investigate the performance of content-informed features and adaptation methods in capturing the spectro-temporal characteristics of singing voice in polyphonic music.

[1] https://www.music-ir.org/mirex/wiki/2017:Automatic_Lyrics-to-Audio_Alignment_Results

http://dx.doi.org/10.21437/Interspeech.2019-1520

2. Speech and music-informed features

Speech and singing have many similarities because they share the underlying physiological mechanisms for production, such as articulatory movements in vocal production [14, 15]. Modern ASR systems use conventional acoustic features such as mel-scaled cepstral coefficients (MFCC) to capture the phonetic aspects in conjunction with speaker representations such as i-vectors [16] to capture speaker information. These features have been widely used for various MIR tasks such as genre classification, artist, and song identification [17–19]. However, the acoustic characteristics of singing and speech also differ in many ways, such as pitch range, vibrato, and phoneme duration [20, 21]. Moreover, the presence of different kinds of musical accompaniments, along with singing vocals, constitute additional frequency components in the music signal, that may render the lyrics unrecognizable [6]. We hypothesize that including additional speech and music informed low-level descriptors for acoustic modeling of sung lyrics will result in improved lyrics-to-audio alignment. Low-level descriptors provide discriminatory information about the temporal variations of the background music and the transitions between sung phonemes and notes, in addition to the timbral information provided by the conventional MFCC features.

The open-source feature extractor called OpenSMILE (or Open Speech and Music Interpretation by Large-space Extraction) [22] unites feature extraction algorithms from the speech processing and the MIR communities. It provides various audio low-level descriptors (LLD) that have been widely used for emotion recognition in speech [23], as well as for summarization [24], mood classification [25], and singing quality assessment [26] in music. In this work, we have divided these features into five feature groups, namely *voicing (V), energy (E), auditory (A), spectral (S), and chroma (C)*, as described in Table 1.

As indicated in early studies in speech-music discrimination [27, 28], the distribution of the first differential of pitch in singing voice shows a high concentration around zero delta pitch corresponding to steady notes. A similar behavior is observed for the delta amplitude as well. Also, large changes in pitch are observed in singing corresponding to transition between notes. These aspects are covered by the *voicing* and *energy* feature groups.

Singing vocals in presence of background music or chorus is similar to speech in the presence of noise. Relative spectra (RASTA) [29] is a filtered representation of an audio signal that is robust to additive and convolutional noise. It essentially suppresses the spectral components that change more quickly or slowly compared to the typical range of speaking rate. Therefore, the *auditory* feature group is expected to be robust to background music and chorus.

Spectral group of features represent the "musical surface" which denote the characteristics of music related to texture, timbre and instrumentation, as coined by Tzanetakis et al. [30]. The statistics of the distribution of various spectral descriptors such as spectral centroid, flux, energy over time represent the musical surface for pattern recognition purposes.

Chroma features have been used previously for tasks such as cover song identification, and music audio classification [31]. These features consist of a 12-element vector with each dimension representing the intensity associated with a particular musical semitone. While spectral features such as MFCCs represent the timbral characteristics, chroma features reflect the harmonic and melodic content of the music signal, and are shown to provide information independent of the spectral features [31].

Table 1: *Description of 5 acoustic feature groups.*

Group ID	Feature Group	Description	#LLDs
A	Auditory	RASTA-style auditory spectrum bands 1-26 (0-8 kHz)	26 + deltas
E	Energy	Sum of auditory spectrum (loudness), sum of RASTA-style auditory spectrum, RMS energy, zero crossing rate	4 + deltas
C	Chroma	Intensities in 12 musical semitones	12
S	Spectral	Spectral energy 250-650 Hz, 1 k-4kHz Spectral Roll Off Point 0.25, 0.50, 0.75, 0.90 Spectral Flux, Entropy, Variance, Skewness, Kurtosis, Slope, Psychoacoustic Sharpness, Harmonicity	15 + deltas
V	Voicing	F0, Voicing, Jitter (local, delta), Shimmer, Logarithmic HNR	6 + deltas

3. Model adaptation for domain mismatch

Our goal is to build a framework to automatically align lyrics to the polyphonic music audio. With an acoustic model trained on solo-singing data, we can adapt the model towards the test data in two ways: (a) by making the test data closer to the trained solo-singing acoustic models by applying vocal separation on polyphonic test data, (b) by adapting the acoustic models to polyphonic data. In [11], the former approach was explored. But source separation algorithms are known to introduce artifacts in the extracted vocal, thus the pipeline gets dependent on the reliability of the source separation algorithm. In this work, we investigate the latter approach, i.e. adapting the acoustic model using a small amount of in-domain polyphonic data to reduce the domain mismatch. Model adaptation is achieved by initializing the hidden layers using the neural network acoustic model trained on the solo-singing data and retraining this model by performing extra forward-backward passes only using the available polyphonic training data for a small number of epochs and possibly with a smaller learning rate.

As discussed earlier, acoustic modeling of singing vocals in the presence of background music is constrained by a lack of lyrics annotated data. Recently, a multimodal DALI dataset [32] was introduced, that consists of 5,000+ polyphonic songs with note annotations and weak word-level, line-level, and paragraph-level lyrics annotations. It was created with a set of initial manual annotations of time-aligned lyrics made by non-expert users of Karaoke games, where the audio was not available. The corresponding audio candidates were then retrieved from the web, and an iterative method of obtaining a large-scale lyrics annotated polyphonic music data was proposed. However, the reliability of these lyrics annotations have not been verified. The authors have released 105 songs as the ground-truth data, where the annotations are manually checked and corrected. In this work, we make use of this ground-truth data for domain adaptation.

4. Experimental setup

We conduct two sets of experiments to study the impact of our proposed acoustic modeling strategies for lyrics alignment: (1) we first assess the effect of the speech and music informed features on lyrics alignment in solo-singing, and (2) then we investigate the effects of these features in polyphonic music lyrics alignment, along with model adaptation techniques. In this section, we detail the datasets used for the experiments, acoustic model architecture, the system configurations, and evaluation metrics for assessing the quality of the boundaries.

4.1. Datasets

All datasets used in the experiments are summarized in Table 2. The training data for solo-singing acoustic modeling is approximately 50 hours of the DAMP dataset [5, 9] that has weak line-level lyrics transcription. We use the DALI ground-truth data for domain adaptation of the acoustic models to the poly-

Table 2: *Dataset description. (solo: solo-singing; poly: singing mixed with music)*

Name	Audio type	Content	Lyrics Ground-Truth	Avg Word Length(s)/# words
Training/Adaptation data				
DAMP train [5]	solo	35,662 lines	line-level weak transcription	-
DALI train [32]	poly	70 songs	word and line-level boundaries	
Test data				
DAMP test [5]	solo	1697 lines	line-level transcription	-
Hansen-solo [33]	solo	7 songs	word-level boundaries	0.485 / 2212
Hansen-poly [33]	poly	7 songs	word-level boundaries	0.485 / 2212
Mauch-poly [2]	poly	20 songs	word-level boundaries	0.871 / 5052
DALI dev [32]	poly	9 songs	word and line-level boundaries	0.471 / 2305
DALI test [32]	poly	20 songs	word and line-level boundaries	0.442 / 5260

Table 3: *System configurations. Baseline acoustic models are trained on DAMP subset-train (Table 2). AECSV are the feature group IDs from Table 1.*

Configs	Adaptation data	Features
C1	-	MFCC, i-vectors
C2	-	MFCC, i-vectors, AECSV
C3	vocal-extracted DALI	MFCC, i-vectors
C4	vocal-extracted DALI	MFCC, i-vectors, AECSV
C5	polyphonic DALI	MFCC, i-vectors
C6	polyphonic DALI	MFCC, i-vectors, AECSV

phonic music. It consists of 99 songs[2], that we divided into train, development (dev), and test, in the ratio of 70:9:20.

We evaluated our alignment systems on two datasets - 7 songs[3] from the Hansen's a capela and polyphonic datasets [33], and 20 songs of the Mauch's polyphonic dataset [2]. These datasets were used in the MIREX lyrics alignment challenges of 2017 and 2018. These datasets consist of Western pop songs with manually annotated word-level boundaries. We tune our model adaptation scheme on the DALI-dev set, and also report alignment results on the DALI-test set.

4.2. ASR architecture

The ASR system used in these experiments is trained using the Kaldi ASR toolkit [34]. A context dependent GMM-HMM system is trained with 40k Gaussians using 39 dimensional MFCC features including the deltas and delta-deltas to obtain the alignments for neural network training. The frame rate and length are 10 and 25 ms, respectively. A factorized time-delay neural network (TDNN-F) model [35] with additional convolutional layers (2 convolutional, 10 time-delay layers followed by a rank reduction layer) was trained according the standard Kaldi recipe (version 5.4). An augmented version of the solo-singing training data described in Section 4.1 is created by reducing (x0.9) and increasing (x1.1) the speed of each utterance [36]. This augmented training data is used for training the neural network-based acoustic model. The default hyperparameters provided in the standard recipe were used and no hyperparameter tuning was performed during the acoustic model training. The baseline acoustic model is trained using 40-dimensional MFCCs as acoustic features that are combined with i-vectors [37]. During the training of the neural network [38], the frame subsampling rate is set to 3 providing an effective frame shift of 30 ms. A duration-based modified pronunciation lexicon is employed which is detailed in [10].

4.3. System configurations

The baseline acoustic model (C1) is trained on solo-singing DAMP subset-train with the 40-dimensional MFCCs and 100-dimensional i-vectors. To test the performance of the additional features, extracted using OpenSMILE toolbox [22], we append the five feature groups with a total dimension of 154 to the 140-dimensional baseline feature vector (C2). We also analyse the contribution of each feature group by appending only one feature subset, eg. C2-V, C2-A, C2-E etc. We adapt the baseline model with the vocal-extracted DALI-train data (C3, C4), and polyphonic DALI-train data (C5, C6). We use the state-of-the-art implementation of the Wave-U-Net based audio source separation [39] for vocal extraction from the polyphonic audio.

4.4. Evaluation metrics

Mean AE is the absolute error or deviation in seconds from the predicted to the true word start times, averaged over all words in a dataset. Previous studies have reported this metric, but mean AE is affected drastically by outliers. Therefore, to gauge the distribution of alignment errors, we also present median (Med.), standard deviation (Std.) of the absolute boundary errors. Moreover, we measure the percentage of hypothesized word boundaries that are within an acceptable tolerance interval around the ground-truth boundary (i.e. %Correct or %C). Observing the range of average word durations in Table 2, we set this acceptable tolerance interval as approximately half the average duration of words, i.e. the percentage of word-start boundaries within 250 ms of the ground-truth.

Table 4: *Mean AE performance on Hansen's solo-singing data with models trained on DAMP solo-singing data. The median of absolute boundary errors in all cases in this table is 0.03s.*

Config	Mean(s)	Std.(s)	%C
C1	0.20	0.75	91.5
C2	0.13	0.63	94.1
C2-A	0.17	0.95	92.3
C2-E	0.30	1.73	91.7
C2-C	0.32	1.84	90.7
C2-S	0.24	1.36	92.7
C2-V	0.48	1.75	87.6

5. Results and discussion

5.1. Performance on solo-singing

In the first set of experiments, we explore the effect of each of the speech and music informed feature groups combined with MFCCs and i-vectors. The alignment results provided by different feature configurations on the Hansen's solo-singing dataset is shown in Table 4. Training the solo-singing acoustic models with the additional features reduces the average boundary error from 200 ms to 130 ms, while the standard deviation and the %C also improve. We also observe that the auditory and the spectral feature groups individually contribute to the improved performance. Many songs in this solo-singing dataset contain chorus sections, where other singers and the main singer may sing different lyrics at the same time. The robust RASTA features in auditory group is observed to be helpful in such cases. Moreover, the individual groups perform worse than their combination, which implies that the groups provide exclusive information that complement each other.

5.2. Performance on polyphonic audio

To reduce the domain mismatch between solo-singing acoustic models and the polyphonic test data, we adopt three approaches:

Table 5: *Mean AE for various adaptation configurations (LR: same initial learning rate; 0.5LR: half of initial learning rate).*

Config ->	C1	LR, epoch1	LR, epoch2	LR, epoch3	0.5LR, epoch1	0.5LR, epoch2	0.5LR, epoch3
DALI-dev	0.288	**0.170**	0.182	0.173	0.171	0.198	0.201
DALI-test	0.343	0.159	0.162	0.163	0.156	0.176	0.174

[2]There are a total of 105 songs in the ground-truth data, out of which the audio file links to 6 songs are not accessible from Singapore.

[3]The word boundary ground-truth of the songs *clocks* and *i kissed a girl* were not accurate, hence excluded from this study

Table 6: *AE performance on **vocal-extracted** Hansen-poly and Mauch-poly data.*

	Hansen-poly				Mauch-poly			
	Med.(s)	**Mean(s)**	**Std.(s)**	**%C**	**Med.(s)**	**Mean(s)**	**Std.(s)**	**%C**
C1	0.23	2.33	5.10	51.4	1.49	14.31	22.37	32.8
C2	0.15	0.94	2.76	69.9	0.26	4.05	8.30	49.0
C3	0.82	6.84	11.92	41.1	1.61	12.47	20.39	34.9
C4	0.21	2.35	5.22	59.6	0.36	5.19	9.52	41.8

Table 7: *AE performance on Hansen-poly and Mauch-poly data.*

	Hansen-poly				Mauch-poly			
	Med.(s)	**Mean(s)**	**Std.(s)**	**%C**	**Med.(s)**	**Mean(s)**	**Std.(s)**	**%C**
C1	30.10	36.20	31.85	14.5	20.33	39.70	48.55	10.5
C2	2.88	9.57	13.38	27.7	2.93	14.69	22.59	25.8
C5	0.08	1.82	5.72	71.8	0.15	3.78	9.98	60.9
C6	0.11	2.37	6.85	64.7	0.18	**1.93**	5.90	57.5

(a) vocal extraction of the polyphonic test data, as done in previous studies [4, 7, 11], (b) adapt the models with vocal extracted polyphonic data, and (c) adapt the models with polyphonic data. We used DALI-train for adaptation, and DALI-dev to optimize the alignment performance (mean AE) by adjusting the initial learning rate (LR) and the number of epochs, as shown in Table 5. We choose the setting that performs the adaptation using the same initial LR within a single epoch as it gives the best performance on the development set. The best result reported on the DALI-test set is also obtained using this setting. Please note that the DALI-test data contains short lines or utterances of 3-10s, which is different from the other test sets in which the entire song of 2-3 mins. is given to the system. The short duration of the DALI-test set results in relatively smaller mean AE values.

5.2.1. On vocal-extracted polyphonic test data

Table 6 summarizes the performance of solo-singing models (C1, C2) and adapted models with extracted vocals (C3, C4) with and without the additional features on the vocal extracted Hansen-poly and Mauch-poly test datasets. We observe that model adaptation does only a slight difference in the performance (cf. C1, C3), but the additional features improve the performance by a large margin (cf. C1, C2). MFCC features are known to be sensitive to background noise [40]. So, domain adaptation with the extracted vocals containing distortion and artifacts is a possible reason for the poor performance of the adapted models. On the other hand, the additional features are designed to be robust to noise, thus improving the performance.

5.2.2. On polyphonic test data (without vocal extraction)

Table 7 shows the lyrics alignment performance of the unadapted (C1, C2) and polyphonic data adapted (C5, C6) acoustic models on the Hansen-poly and Mauch-poly data. The poor performance of the solo-singing models (C1, C2) on polyphonic data is expected due to domain mismatch. But here, the domain adaptation (C5, C6) gives a considerable improvement in performance. A comparison of Table 6 and 7 shows that domain adaptation *without* vocal extraction performs better. This suggests that domain adaptation with a small amount of polyphonic data helps the acoustic model capture the spectro-temporal variations of singing vocals, which offers a simple, but effective solution in scenarios with limited polyphonic singing data.

One main difference between the Hansen's and Mauch's datasets is that the songs in the Mauch's dataset are rich in long-duration musical interludes that have no singing vocals, while Hansen's has only a few of such interludes. We observe that the content-informed features and domain adaptation help to improve the boundaries next to these long interludes. Thus, the improvement in alignment performance is more evident in the Mauch's dataset, than in the Hansen's dataset.

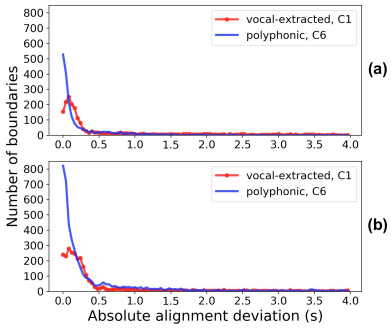

Figure 1: *Comparison of word boundary alignment error distribution between C1 on extracted vocals test data and C6 on polyphonic test data on (a) Hansen's and (b) Mauch's datasets.*

Table 8: *Comparison of mean AE (s) with existing literature.*

	MIREX 2017		MIREX 2018	ICASSP 2019		
	AK [7]	**GD [8, 41]**	**CW [12]**	**DS [13]**	**CG [11]**	**Ours**
Training data	6,000 songs (DAMP) (solo)	6,000 songs (DAMP) (solo)	7,300 songs (KKBOX) (poly)	44,232 songs (Spotify) (poly)	35,662 lines (DAMP) (solo)	35,662 lines (DAMP) (solo) + 70 songs (DALI) (poly)
Architecture	DNN-HMM	DNN-HMM	GMM-HMM	UNet based end-to-end	SAT DNN-HMM	CNN-TDNN-F
Hansen-poly	7.34	10.57	2.07	-	1.39	0.93 (median: 0.15)
Mauch-poly	9.03	11.64	4.13	0.35	6.34	1.93 (median: 0.18)

Although the mean AE of the boundaries is more than a second, the median of errors is less than 180 ms for the best performing systems. A comparison of the boundary error distribution of C1 on extracted vocals, and C6 on polyphonic data (Figure 1) shows a large increase in the number of boundaries towards zero error, for both the datasets. This also means that the there are hypothesized boundaries that are far away from the the true boundaries, that needs to be investigated in future.

5.3. Comparison with existing literature

In Table 8, we compare our best results with the past studies, and find that our strategy provides better results than all previous work, except for the end-to-end system [13]. An end-to-end system requires a large amount of data for reliable output which we do not have access to. Our proposed strategies show a way to fuse knowledge-driven and data-driven methods to address the problem of lyrics-to-audio alignment in a low-resourced setting.

6. Conclusions

In this study, we discuss two strategies to obtain improved acoustic modeling for the task of lyrics-to-audio alignment. Particularly, we propose to (1) employ additional features with speech- and music-related information together with conventional MFCCs, and (2) adapt solo-singing acoustic model using small amount of in-domain polyphonic data. We validated the robustness of these features to background music and ability to capture the spectro-temporal variations in polyphonic singing vocals. The alignment experiments demonstrate that applying the described strategies reduces the mean AE to 1.9s on the Mauch's dataset which is better than all results reported in the MIREX lyrics alignment challenge.

7. Acknowledgments

This research is supported by Ministry of Education, Singapore AcRF Tier 1 NUS Start-up Grant FY2016, Non-parametric approach to voice morphing.

8. References

[1] H. Fujihara, M. Goto, J. Ogata, and H. G. Okuno, "Lyricsynchronizer: Automatic synchronization system between musical audio signals and lyrics," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 6, pp. 1252–1261, 2011.

[2] M. Mauch, H. Fujihara, and M. Goto, "Integrating additional chord information into HMM-based lyrics-to-audio alignment," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 1, pp. 200–210, 2012.

[3] M. McVicar, D. P. Ellis, and M. Goto, "Leveraging repetition for improved automatic lyric transcription in popular music," in *Proc. ICASSP*, 2014, pp. 3117–3121.

[4] A. Mesaros and T. Virtanen, "Automatic recognition of lyrics in singing," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2010, p. 4, 2010.

[5] C. Gupta, R. Tong, H. Li, and Y. Wang, "Semi-supervised lyrics and solo-singing alignment," in *Proc. ISMIR*, 2018.

[6] M. Ramona, G. Richard, and B. David, "Vocal detection in music with support vector machines," in *2008 Proc. ICASSP*. IEEE, 2008, pp. 1885–1888.

[7] A. M. Kruspe, "Bootstrapping a system for phoneme recognition and keyword spotting in unaccompanied singing," in *Proc. ISMIR*, 2016, pp. 358–364.

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

[9] S. Sing!, "Smule.digital archive mobile performances(damp)," https://ccrma.stanford.edu/damp/, 2010 (accessed March 15, 2018).

[10] C. Gupta, H. Li, and Y. Wang, "Automatic pronunciation evaluation of singing," *Proc. INTERSPEECH*, pp. 1507–1511, 2018.

[11] B. Sharma, C. Gupta, H. Li, and Y. Wang, "Automatic lyrics-to-audio alignment on polyphonic music using singing-adapted acoustic models," in *Proc. ICASSP*. IEEE, 2019, pp. 396–400.

[12] C.-C. Wang, "Mirex2018: Lyrics-to-audio alignment for instrument accompanied singings," in *MIREX 2018*, 2018.

[13] 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*. IEEE, 2019, pp. 181–185.

[14] R. J. Zatorre and S. R. Baum, "Musical melody and speech intonation: Singing a different tune," *PLoS biology*, vol. 10, no. 7, p. e1001372, 2012.

[15] S. Zhang, R. C. Repetto, and X. Serra, "Study of the similarity between linguistic tones and melodic pitch contours in Beijing opera singing." in *Proc. ISMIR*, 2014, pp. 343–348.

[16] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, May 2011.

[17] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293–302, 2002.

[18] J. Park, D. Kim, J. Lee, S. Kum, and J. Nam, "A hybrid of deep audio feature and i-vector for artist recognition," *arXiv preprint arXiv:1807.09208*, 2018.

[19] M. Mandel and D. Ellis, "Song-level features and support vector machines for music classification," in *Proc. ISMIR*, 2005.

[20] H. Fujihara and M. Goto, "Lyrics-to-audio alignment and its application," in *Dagstuhl Follow-Ups*, vol. 3. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2012.

[21] A. Loscos, P. Cano, and J. Bonada, "Low-delay singing voice alignment to text." in *Proc. ICMC*, 1999.

[22] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the Munich versatile and fast open-source audio feature extractor," in *Proc. ACM Multimedia*. ACM, 2010, pp. 1459–1462.

[23] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi *et al.*, "The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism," in *Proc. INTERSPEECH*, 2013.

[24] F. A. Raposo, D. M. de Matos, and R. Ribeiro, "An information-theoretic approach to machine-oriented music summarization," *Pattern Recognition Letters*, 2019.

[25] A. Alajanki, Y.-H. Yang, and M. Soleymani, "Benchmarking music emotion recognition systems," *PLOS ONE*, pp. 835–838, 2016.

[26] J. Böhm, F. Eyben, M. Schmitt, H. Kosch, and B. Schuller, "Seeking the superstar: Automatic assessment of perceived singing quality," in *2017 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2017, pp. 1560–1569.

[27] M. J. Carey, E. S. Parris, and H. Lloyd-Thomas, "A comparison of features for speech, music discrimination," in *Proc. ICASSP*, vol. 1. IEEE, 1999, pp. 149–152.

[28] C. Panagiotakis and G. Tziritas, "A speech/music discriminator based on RMS and zero-crossings," *IEEE Transactions on Multimedia*, vol. 7, no. 1, pp. 155–166, 2005.

[29] H. Hermansky and N. Morgan, "Rasta processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, 1994.

[30] T. George, E. Georg, and C. Perry, "Automatic musical genre classification of audio signals," in *Proc. ISMIR*, 2001.

[31] D. Ellis, "Classifying music audio with timbral and chroma features," in *Proc. ISMIR*, 2007.

[32] 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.

[33] J. K. Hansen, "Recognition of phonemes in a-cappella recordings using temporal patterns and mel frequency cepstral coefficients," in *9th Sound and Music Computing Conference (SMC)*, 2012, pp. 494–499.

[34] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The Kaldi speech recognition toolkit," in *in Proc. ASRU*, 2011.

[35] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohammadi, and S. Khudanpur, "Semi-orthogonal low-rank matrix factorization for deep neural networks," in *Proc. INTERSPEECH*, 2018, pp. 3743–3747.

[36] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Proc. INTERSPEECH*, 2015, pp. 3586–3589.

[37] G. Saon, H. Soltau, D. Nahamoo, and M. Picheny, "Speaker adaptation of neural network acoustic models using i-vectors," in *Proc. ASRU*, Dec 2013, pp. 55–59.

[38] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely sequence-trained neural networks for ASR based on lattice-free MMI," in *Proc. INTERSPEECH*, 2016, pp. 2751–2755.

[39] D. Stoller, S. Ewert, and S. Dixon, "Wave-u-net: A multi-scale neural network for end-to-end audio source separation," in *Proc. ISMIR*, 2018.

[40] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 745–777, April 2014.

[41] G. Dzhambazov, "Knowledge-based probabilistic modeling for tracking lyrics in music audio signals," Ph.D. dissertation, Universitat Pompeu Fabra, 2017.