USING SYNCHRONIZED LYRICS AS A NOVEL INPUT INTO VOCAL ISOLATION MODELS

Amir Lankarani

ABSTRACT

While state-of-the-art source separation models tend to employ hybrid architectures, few have supplemented this audio data with any metadata such as lyrics. Thus, this project tests whether synchronized lyrics can be used to create posteriorgrams and lyrical alignments that can improve source separation models. Using the standard MusDB-HQ dataset along with its lyrical annotations, we augmented the dataset by altering the pitch and BPM of random audio stems to remix them while maintaining the lyrics' time-points. Full mixture data was fed into a Kaldi model that pre-isolated the audio, faded out portions that had no lyrics in the annotations, and then re-segmented the utterances at more precise time-point. The resegmented textual data was then used to create alignments of the precise time each word was spoken and posteriorgrams of the predicted chance each phoneme is said at every audio frame. These features were fed into a Pytorch source separation architecture modified from the NUSSL library to see if it could improve traditional source separation models based on extracting the vocals' spectral mask. Posteriorgrams were concatenated onto the spectrogram prior to an RNN chain and lyrical alignments were encoded and embedded using linear models that were re-embedded at the end with the result of the audio RNN chain. The separation showed little to no improvements with either the posteriorgrams and alignments. This is likely because the alignment and posterior data were ambiguous compared to the precise spectral data. However, more data and a more nuanced architecture is needed to truly see if these tools may provide benefit.

***Keywords*—** Separation, vocals, lyrics, alignments

1. INTRODUCTION

It can often be difficult to extract the precise content of a speaker's voice from an audio file as there is an inherent ambiguity to the phonetic qualities of vocals. Factors like visual context, cultural context, and semantic context can make both people and algorithms process the same vocalization in many ways. For instance, the McGurk effect can result in processing the same sound as an F or a B based on how we see the speakers' lip moves [1]. Algorithms often lack this awareness of context whether due to lack of data, the lack of data types, the lack of diversity of data, or the lack of an adequate architecture to precisely recognize this context.

This difficulty is made more difficult by the "cocktail party problem" which describes how algorithms can often have difficulty isolating which sound a stimulus comes from. While human brains can easily isolate a single sound stream from a mixture of sounds based upon what they are focusing on, there is no consistent method to do this for algorithms and instead, we must train algorithms to determine this separation on their own [2].

Music represents perhaps the most difficult case for source separation as songs are often composed of many unique sounds and textures that were created just for that song (eg a pitched sample). Yet the vocals from music often show more consistency than these other sounds as they all stem from the same voices we train speech models on [20]. For this reason, this study explores whether speech processing methods can improve a model's ability to extract vocals using some of the same phonetic extraction tools speech recognition researchers use.

Separating the different sources of sound has many valuable purposes such as helping DJs create remix stems as well as providing a useful data extraction tool for researchers. Sadly, the best source separation models are still too unrealistic to be used extensively for these purposes [3]. The best models are either unable to isolate enough of an audio source's frequencies losing the natural timbres of that source, or they isolate too many frequencies representing other sound sources making these other sounds bleed into the isolated stem [4]. Thus, despite extensive research, significant improvements are still needed.

2. FORMULATION FROM THE STATE OF THE ART

2.1 Hybrid Source Separation Models

In order to compare source separation models, the publicly available dataset MusDB18 has become the standard reference dataset for evaluating 4-source music separation [5]. The Hybrid Demucs model, created as a part of the 2021 Sony MDX challenge, is currently the highest performing model on this dataset as measured by the global signal to distortion ratio [6, 7].

This model is representative of an emerging trend in source separation that uses hybrid methodologies combining both time-domain (T-D) and time-frequency (T-F) modalities into one singular model [6, 8]. However, while hybrid models have become the norm, few models attempt to bolster their separation using any features not already present in the audio signal.

2.2 Text-Enhanced Source Separation Models

Written transcriptions of the linguistic and phonemic content of voices have been used used extensively for almost all tasks in speech research. In contrast, written lyrics have only been used in a few specific fields of music research such as lyric transcription. When using external data, music source separation models tend to instead use melodic metadata like scores and pitch traces [9].

However, due to advances in lyrical extraction and alignment, there has recently been an increased usage of lyrics to inform music research. In 2019, Shultze-Forster et al. proved that including non-sonic information in vocal extraction models can improve extraction quality particularly when extracting from weakly-trained spectral makeups [10]. In reference to this paper, Meseguer-Brocal and Peeters created a successful source separation model using time-aligned phonetic data as a novel input [11]. Jeon, Choi, and Lee further proved that combining phonetically-aligned lyrics with a processed spectrogram can significantly enhance Open-unmix, a popular reference source separation model [12]. Shultze-Forster later found that even unaligned lyrics may improve source separation, as alignment can occur during separation [13].

However, there are still no studies that have used the "synchronized lyrics" metadata to bolster source separation even though this tag is present in most commercial music files and can provide more precise information to a source separation model. Furthermore, no studies have attempted to bolster source separation using posteriorgrams, a matrix representing the probability of each phoneme at each frame of the audio.

3. APPROACH AND FORMULATION

This study evaluates the efficacy of using synchronized lyrics and posteriorgrams, in tandem with other metadata extracted from lyrics, as tools to improve source separation. This is done by using traditional speech recognition tools from the Kaldi toolkit to extract alignments and posteriorgrams from music files as if they were speaker files. We hope to augment these approaches by improving segmentation using the utterance locations of the synchronized lyrics. We then feed this into a standard source separation framework in various ways to test if any of these methods confer benefits. More specifically, we will test if posteriorgrams alone can provide any benefit and then if posteriorgrams alongside alignments can create further improvements.

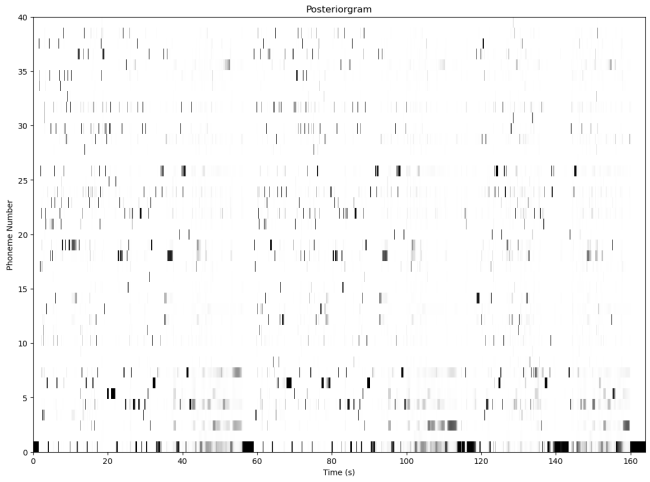

Figure 1: A posteriorgram extracted from a full audio mixture predicting a singers' phonetic content at each time.

4. EXPERIMENTS AND RESULTS

4.1 Data Preparation

This study uses the high-quality wav version of the MusDB-18 dataset, MusDB-HQ [14]. This allows us to compare our model with the state-of-the-art models in their best form. We have also included information from a separate MUSDB lyrics extension composed of hand-annotated text, start, and end time for each line of lyrics for each song in the dataset as LRC files [15]. These lyrics were processed to work within Kaldi by removing its extraneous characters, creating a lexicon of all words present in the dataset, and storing each utterance in a dataset into a text & segments file for that dataset.

Since each song in the MUSDB library has a separated music stem for vocals, drums, bass, and miscellaneous, we can significantly extend the dataset by rearranging various stems from different songs into novel song combinations. Initial testing showed poor alignments when mixing stems at random with no tempo or pitch alterations, so we improved our mashup creation creating an altered version of Demucs' "automix.py" program. This program creates novel songs by randomly shifting the chromas and beat onsets for all the songs each iteration before finding a random combination of any four stems that meets the pitch and tempo cohesion criteria used [8]. This shifting of pitch and frequency allowed our dataset to simulate different vocal ranges and styles of music outside of the dataset. Furthermore, the randomized sample matching simulated artificial noise and volume variations that commonly differ in less professional musical contexts. We further altered this program by: having it output synchronized lyrics that match the placement of vocals

within the mixture track; preventing any splits from being made in the middle of a single line of vocals; and extending the maximum pitch shift and tempo shift to 4 semitones and 25% of the BPM respectively in order to account for the significantly smaller dataset compared to Demucs. Overall, we created 14 sets of 84 additional training songs before removing about 700 of these songs as they were either too short or did not include ample vocal data. This resulted in a total training set of about 600 songs. 14 of the songs in the original dataset make up the validation set. 45 out of 50 songs of the standardized MusDB test set were used as our test set (the other five had no vocals or were not in English).

4.2 Extracting Alignments

In the next step, we found alignments and phoneme probabilities at each time-point by creating a Kaldi script combining aspects of the ASA and ALTA Kaldi recipes with some modifications [16, 17]. Pre-trained models used were derived from the standard pipeline of the ALTA recipe [18]. Both recipes predict phoneme alignments from music using a publicly available set of singing phones, the CMU dictionary, and triphone GMM-HMMs adapted from the Librispeech recipe to better accommodate singing voices [16]. Like the ASA recipe, we start by isolating the vocals from each song using a pre-trained Demucs model improving our alignment results as it allows Kaldi to carry out Vocal Activity Detection (VAD) with minimal noise. We then use a grapheme-to-phoneme phonetisaurus model to extend the ASA lexicon to recognize the phonetic qualities of words in our dataset that it was not trained to recognize.

Next audio was re-segmented into new utterance positions and split into subcomponents using a TDNN model from the ALTA recipe. We found that the segmentation of this model was more precise and accurate than the hand-annotated boundaries as it could find exact segment points at the frame level. However, these segment locations had a much higher variance than the hand-annotated segment locations. Thus, to reduce this variance, we faded out portions of the separated audio where annotators said there were no lyrics. This caused Kaldi to segment along these artificial boundary points and resulted in splits that were both accurate and consistent.

Lastly, MFCCs and i-vectors were extracted using standard Kaldi methodology. The i-vectors were used to get precise alignments of the time where each lyric is spoken output as a text file. The MFCCs are used alongside a decision tree to get exact predictions of the probability of each phoneme at each time-frame as represented by a 2D matrix posteriorgram. Qualitatively, these alignments and posteriorgrams appeared to match rather well with the audio itself but a more thorough testing would be useful.

4.3 Enhanced Vocal Separation

Historically, the most common technique for source separation has involved visualizing the mixture spectrogram as an image that is comprised of the layering of multiple different spectral source images known as masks [19]. At ISMIR 2020, a tutorial was held outlining a baseline approach to source separation by mask estimation that we have mimicked as a point of comparison [19].

More specifically, our baseline model only learned to extract the vocal file, as that is the only stem relevant to this study, although we also measured the accuracy of the combination of all non-vocals during the testing phase. Upon initialization of a sample in the model, it carries out an FFT using a window of 512 and a hop length of 128 to approximate an input spectrogram. We used a batch size of 8 and selected a little over 20 seconds worth of audio for each sample to make training manageable. During each forward pass, the amplitude is converted to decibels before batch normalizing its values to make all the bins more meaningful. The output of the normalization is fed into a three-layer stack of recurrent neural networks each with a hidden size of 512 and a dropout rate of .3. Lastly, this output is fed into an embedding layer that converts our model into a finalized mask. Each activation layer uses Sigmoid since it has been shown to have the best results for source separation.

We then altered this baseline in multiple ways to test if our alignments could improve it. Our first method extended the posteriorgram values so that it had the same number of time frames as the spectrogram. Then, we concatenated all 41 rows of the posteriorgrams on top of the normalized spectral magnitudes along the time dimension to see if the model could relate phonetic predictions with certain vocal spectral mask qualities. We tried this with and without the following: batch normalization of the posteriorgram, a larger hidden size to accommodate the newly concatenated rows, and a dx & a ddx of the posteriogram also concatenated. Our final model settled on no normalization, a smaller hidden size, and no extra posteriorgram layers. We also tested if any separation could occur using only the posteriorgram and no audio data - this performed poorly as it simply learned to consistently map all mid-range frequencies at all time points where vocals should occur.

Lastly, we used aligned lyrics as a final input. Each word in the dataset's dictionary was encoded as a unique integer. Then, a vector was made for the encoding of each frame of the song and the last word spoken prior to that frame. We then fed this into an embedding layer which was concatenated with the embedding layer of the concatenated audio / posteriorgram before undergoing one final layer to get an output mask.

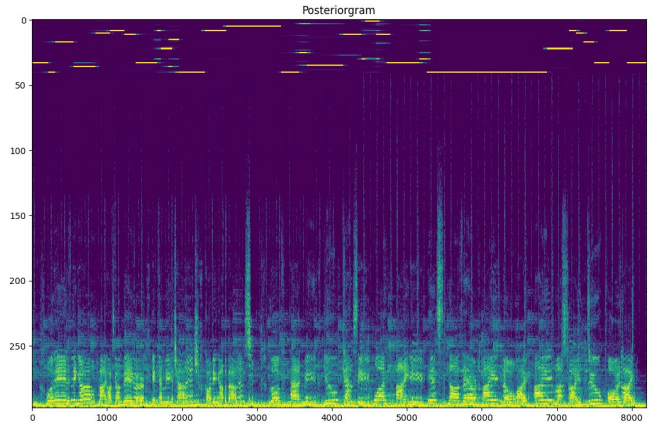

Figure 2: Posteriorgram concatenated onto the spectral content of a full song after normalizing the values of both.

4.4 Evaluation

Group	Source	SDR	SIR	SAR	SNR	SRR
Oracle	Vocals	78	78	151.9	NaN	78
Audio	Vocals	0.4	9.5	1.2	3.3	4.5
Audio + Posterior	Vocals	0.6	9	1.4	3.5	3
Audio + Posterior + Alignments	Vocals	1.2	8.7	2.2	3.3	9.5
Oracle	Non-vocals	76.4	76.4	151.8	NaN	76.4
Audio	Non-vocals	9.5	15.5	12.1	10	20.6
Audio + Posterior	Non-vocals	9.7	13.9	12.3	10.2	19.9
Audio + Posterior + Alignments	Non-vocals	9.5	17.8	10.5	10	17

Figure 3: Evaluation scores of the various models using source-invariant: source-to-distortion ratio, source-to-interference-ratio, source-to-artifacting ratio, source-to-noise ratio, and source-to-residual ratio

To facilitate better comparison, we compared each model after 50 epochs. While none of the models had converged at this point, this was a good stopping point that allowed us to compare multiple well-trained models at a comparable time point in our given budget.

Sadly, we found only marginal improvements from our augmented data within the given training conditions. The vocals' source-to-distortion-ratio, the most common evaluator in source separation, improved by .2 when adding the posteriors and .8 when adding alignments reflecting a slight reduction in distortion. Similarly, using the posteriors improved vocal source-to-inference ratio by a small value of .2 while incorporating the lyrics resulted in a larger improvement of 1.0 indicating a decrease in artifacting. The most notable improvement was in the source-to-residual ratio which showed a large 5-point increase when adding lyrics but scored lower by 1.5 if only posteriors were added. However, the normal audio model had the best source to interference ratio and the alignments model had the worst reflecting how the additional layers may have caused the model to learn to include more audio not actually a part of the model. Overall, these differences are all too small to make any claims so more testing will be needed on more fully trained models before making a conclusion. Regardless, these values are far below the state-of-the-art which boasts a SDR of 10.01 on the dataset [7].

Observing the models qualitatively, the masks derived from the augmented models appeared to better recognize where there were no vocals tending to exclude these sections entirely along the no vocals sections of the posteriorgram. However, the posteriorgrams introduced a trained ambiguity as each phoneme bin takes up multiple frames and can also be represented spectrally in many ways. This can be seen as a larger blur for the masks from this model. The model with time-aligned lyrics seemed even better than the posteriorgram model at finding boundaries as it tended towards bounding mask points along the point where a new lyric should be; yet it similarly suffered from an increased blurring likely due to having poorly trained associations for many of the words in the corpus. Furthermore, the inclusion of the time-aligned lyrics may have improved the model solely because it added more training layers to the model as it performed the worst during training despite its scores in evaluation.

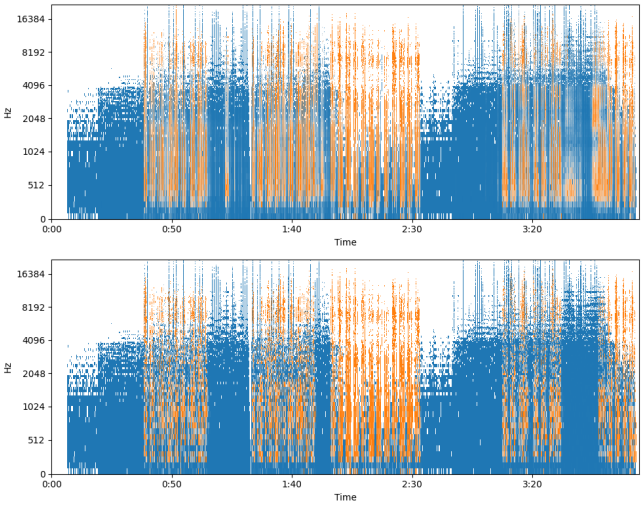

Figure 4: Top is the mask of the predicted vocals relative to predicted non-vocals for the posteriorgram-audio model. Bottom is the mask of the actual vocals relative to the actual non-vocals. Orange is the vocals while blue is non-vocals.

5. CONCLUSIONS AND FUTURE WORK

We were able to accurately extract audio alignments and posteriorgrams from music using speaker recognition tools. This extraction resulted in segmentation

more accurate yet less precise than the hand-annotated segments. Artificially altering audio files using the hand-annotated segments in synchronized lyrics resulted in segmentation were both more precise and more accurate, as well as better posteriorgrams ad alignments. However, a deeper evaluation of these values using labelled alignment and phoneme annotations is necessary.

These posteriorgrams and alignments were unable to confer notable improvements in our source separation models. This appears to largely be because the exact frequency contents associated with the posteriorgrams varies from person to person and context to context while the exact frequency contents of the mixture spectrogram are directly related to the frequency contents of its singers' voices. Similarly, for lyrical alignments, the spectral makeups associated with any word may vary from person to person making it a poor predictor of precise frequency values. The lack of improvement may also be because the posteriorgrams and alignments were created from a significantly larger window size than our FFT so we had to repeat the frames of the posterior to match the spectrogram. This resulted in imprecise phonemes being predicted at certain frames based on the prior frame likely causing the blurred output masks. Future recreation would benefit from matching these window sizes even if it may result in increased processing time.

Due to time and resource constraints, this study was lacking in several areas that may have resulted in more tangible improvements from the alignments and posteriorgrams. Most notably, due to the lack of isolated song stems that also have lyrical positions annotated, this study was only able to train on about 100 original songs. This also created difficulties in expanding our dataset since remixing programs do not account for lyrical positions so we needed to create a rather slow algorithm that only resulted in only about 600 songs in our expanded training set. Furthermore, each model was only trained for 50 epochs, at which point none of the models had yet converged, so it is possible that differences between the models may have emerged upon convergence. Ultimately, it is likely that the variations in the precise associations between phonemes and frequency expression could be better modelled with a dataset that had a larger range of speakers' voices as this would allow it to find more nuanced associations between the mixture spectrogram and the vocal spectrogram without as much overfitting. Similarly, for lyrical encodings, many words in our dataset were only said once or a few times so the model was bound to either ignore this text data or overfit on this text data as there would be little evidence to prove or disprove an association.

Furthermore, our model was very simplified and it is likely that the posteriorgrams and lyrical alignments could have more of an effect in a more nuanced architecture. State-of-the-art hybrid source separation models often tend to use three separate pathways all combined at the end: one for the TF domain, one for the TD domain, and one for the TF & TD domain in tandem [6, 8]. With more time and resources, it would be useful to use the same methodology for posteriorgrams as this may maintain the benefits of a regular audio model by itself while also allowing it to find associations with the posteriorgram as well as allowing it to get precise understanding of where vocals do not exist based on the posteriorgram alone. For lyrical encodings, it would be useful to use a grapheme-to-phoneme phonetisaurus like in the Kaldi model we used so that the model can associate the phonemes in a lyric with the frequency content instead of the word itself. Alternatively, it may be useful to use some large language model embeddings for the word such as using Bert to allow for a more nuanced understanding of word-audio associations than our simple label encoder.

Lastly, this study was segmented into two parts which is unrealistic for actual use cases. Future implementations would benefit from directly interconnecting the Kaldi audio alignments into the source separation training pathways so that both operations are interconnected and seamless. This would create an unprecedented source separation model that could truly understand how to use textual content to improve its separation capabilities as it uses state-of-the-art vocal language modeling tools that are underexplored in the music context. However, with the current implementation of Kaldi and most source separation libraries, this is a difficult step that would demand a large-scale restructuring of both Kaldi and the separation libraries.

6. REFERENCES

[1] K. Tiippana, 'What is the McGurk effect?', Frontiers in Psychology, vol. 5, 2014.

[2] S. Haykin and Z. Chen, 'The cocktail party problem', Neural computation, vol. 17, no. 9, pp. 1875–1902, 2005.

[3] Y. Zhang, Y. Xiao, W.Q. Zhang, X. Tan, L. Lei, and S. Wang, "*Mixing or Extracting? Further Exploring Necessity of Music Separation for Singer Identification*," in *13th Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, Tokyo, Japan, 2021.

[4] "Four of The Best Stem Separation Tools," Attack Magazine. Available: https://www.attackmagazine.com/reviews/the-best/four-of-the-best-stem-separation-tools/ (accessed Oct. 21, 2022).

[5] Z. Rafii, A. Liutkus, F.R. Stöter, S. I. Mimilakis, and R. Bittner, "MUSDB18 - a corpus for music separation," Dec. 17, 2017. Distributed by *Zenodo*, DOI: 10.5281/zenodo.1117372

[6] AIcrowd, virtual. *Music Demixing | Post-Challenge Town Hall | Idea Sharing & Winner Announcements*. Aug. 25, 2021. Available: https://www.youtube.com/watch?v=TntPVZ4ajIk&t=3717s (accessed Oct. 21, 2022).

[7] "Music Source Separation on MUSDB18," Papers with Code. Available: paperswithcode.com/sota/music-source-separation-on-musdb18 (accessed Oct. 21, 2022).

[8] A. D'efossez, "Hybrid Spectrogram and Waveform Source Separation," in *Proceedings of the MDX Workshop*, virtual, 2021, arXiv: 2111.03600v3 [eess.AS].

[9] S. Ewert, B. Pardo, M. Müller, and M. Plumbley, "Score-Informed Source Separation for Musical Audio Recordings: An Overview," in *Institute of Electrical and Electronics Engineers Signal Processing Magazine*, Vol. 31, Issue: 3, May 2014, pp. 116-124.

[10] K. Schulze-Forster, C. Doire, G. Richard, and R. Badeau. "Weakly Informed Audio Separation," presented in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, U.S.A, 2019, hal-02332689

[11] G. Meseguer-Brocal and G. Peeters. "Content Based Singing Voice Separation via Strong Conditioning Using Aligned Phonemes," in *Proc. of the 21st Int. Society for Music Information Retrieval Conf.*, Montréal, Canada, 2020, arXiv:2008.02070v1 [eess.AS]

[12] C.B. Jeon, H.S. Choi, and K. Lee, "Exploring Aligned Lyrics-Informed Singing Voice Separation," in *Proc. of the 21st Int. Society for Music Information Retrieval Conf.*, Montréal, Canada, 2020, arXiv:2008.04482v1 [eess.AS]

[13] K. Schulze-Forster, C. Doire, G. Richard, and R. Badeau, "Joint phoneme alignment and text-informed speech separation on highly corrupted speech," in *Proceedings of The International Conference on Acoustics, Speech, & Signal Processing*, Barcelona, Spain, 2020, hal-02457075.

[14] Z. Rafii, A. Liutkus, F.R. Stöter, S. I. Mimilakis, and R. Bittner, "MUSDB18-HQ – an uncompressed version of MUSDB," Aug. 1, 2019. Distributed by *Zenodo*, DOI: 10.5281/zenodo.3338373

[15] K. Schulze-Forster, C. Doire, G. Richard, and R. Badeau, "MUSDB18 lyrics extension," Mar. 15, 2021. Distributed by *Zenodo*, DOI: 10.5281/zenodo.3989267.

[16] E. Demirel, S. Ahlbäck, and S. Dixon, "MSTRE-NET: Multistreaming Acoustic Modeling for Automatic Lyrics Transcription" in *Proc. of the 22nd Int. Society for Music Information Retrieval Conf.*, Online, 2021, arXiv:2108.02625v1.

[17] E. Demirel, S. Ahlbäck, and S. Dixon, 'Low Resource Audio-To-Lyrics Alignment from Polyphonic Music Recordings', in ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021, pp. 586–590.

[18]
E. Demirel, S. Ahlbäck, and S. Dixon, 'Automatic Lyrics Transcription using Dilated Convolutional Neural Networks with Self-Attention', in 2020 International Joint Conference on Neural Networks (IJCNN), 2020, pp. 1–8.

[19] E. Manilow, P. Seetharman, and J. Salamon, Open Source Tools & Data for Music Source Separation. https://source-separation.github.io/tutorial, 2020.

[20] J. Merrill and P. Larrouy-Maestri, 'Vocal Features of Song and Speech: Insights from Schoenberg's Pierrot Lunaire'. Front Psychol. vol. 8, 2017