Automatic Singing Transcription

Victor Shan

November 26, 2023

Abstract

1 Introduction

Songs play an important part of society because they are conduits of cultural expression, foster emotional connection and is a great source of entertainment. Automatic Singing Transcription (AST) is a branch of Automatic Speech Recognition (ASR) can facilitate a deeper way to interact with it. In ASR, the goal is to transcribe normal speech to text, whether that's a conversation between two people, a lecture, or a podcast. Examples of ASR include Youtube's auto-generated captions, assistants like Siri and Alexa, and even live captioning of meetings through applications like Zoom. AST is a similar task, but instead of transcribing normal speech, AST transcribes singing. Applications of AST range from karaoke applications to Music Information Retrieval (MIR) systems. These MIR systems can be used to search for songs based on lyrics, categorize songs based on their lyrics, or even aid in the generation of new songs.

While AST is similar to ASR, there are a few key differences that make it a more challenging task. First, singing is a more complex signal than normal speech. Singing has a more dynamic signal, with more variation in pitch, duration and vibrato [Gupta et al., 2018]. For example, vowels sounds are often held for longer in singing than in normal speech like dragging the last part of "bye" in word "goodbye". Songs also often include music in the background, which can make it more difficult to distinguish between the singer's voice and the background music. Finally, there is a lack of large, publicly available datasets for AST [Meseguer-Brocal et al., 2018]. This is in contrast to ASR, where there are many large datasets available, such as LibriSpeech [Panayotov et al., 2015]. This lack of data makes it especially difficult to train AST models.

There are many possible outputs for AST but one of the most detailed and useful outputs is the time-aligned phoneme sequence. This is a sequence of phonemes, or the smallest unit of sound in a language, that is aligned with the audio. Time-aligned phoneme sequences are useful because it can contains enough information to reconstruct the not only the lyrics, but also how they match with the music. This output can be used to generate a karaoke application that highlights the lyrics as they are sung, or even train a model to sing covers of songs. Word-level alignement is an easier alternative but it misses the crucial information of the duration of phonemes within words. The ideal output is one that includes even more information such as the notes to sing at so that singers would be able to read and sing the song without any additional information. In this paper, we will mainly focus on the phoneme-level alignment but use word-level alignment to compare the performance of our model to other models.

2 Datasets

2.1 Requirements

The requirements of a good dataset are audio files, lyrics and timestamps. The audio files can be in any format such as mp3 or wav. The lyrics can be in any format as well, but the timestamps must be in a format that can be used to align the lyrics with the audio. The timestamps can be in the form of a phoneme sequence or word sequence but should also include the onset and offeset times (beginning and end). Sentence level timestamps are too coarse and would need to be broken down into word or phoneme level timestamps. Most lyrics used by popular music services such as Spotify only use sentence/line level timestamps. Ideally, the dataset would also be large enough to train deep neural networks and publicly available so that other researchers can use it and compare their models to existing models.

2.2 Challenges

With these requirements in mind, it is easy to see why there are so few datasets available. First, it is difficult to create a dataset with lyrics and timestamps. The lyrics must be manually transcribed and the timestamps must be manually aligned with the audio. This is a time consuming process that requires a lot of effort. Second, it is difficult to obtain the rights to use the audio files. Most songs are owned by record labels and it is difficult to obtain the rights to use these files. These two challenges make it difficult to create new datasets and is the reason why there are so few datasets available and why most of the existing datasets are small.

2.3 Dataset Augmentation

2.3.1 SpecAugment

Due to the lack of datasets, it is more important to make the most of the existing datasets. One way to do this is to augment the existing datasets. SpecAugment is a series of techniques that augment the audio spectrogram to improve the performance of ASR models. A frequent intermediate representation are Mel-Frequency Cepstral Coefficients (MFCC). These are image representations of energy at certain frequencies on a scale that more closely matches human hearing [Raissi, 2021]. These techniques include time warping, frequency masking and time masking [Park et al., 2019]. Time warping is a technique that stretches or compresses the audio spectrogram in the time dimension. Frequency masking is a technique that masks a random number of frequency channels in the audio spectrogram. Time masking is a technique that masks a random number of time steps in the audio spectrogram. These techniques are used to augment the audio spectrogram before it is fed into the ASR model. [Park et al., 2019]

2.3.2 Transforming Existing Datasets Into Pseudo-Singing Datasets

Another way to augment the existing datasets is to transform the existing datasets. This can be done by shifting the pitch, duration or vibrato [Kruspe, 2015]. The advantage of this technique is that is can also be applied to speech datasets and transform them into pseudo-singing datasets. The disadvantage is that the results will contain artifacts from the transformation based on the techiniques applied. Neural network models showed an almost 15% improvement on the transformed TIMIT dataset than ones trained on the original TIMIT dataset [Kruspe, 2015].

2.3.3 Transforming Utterance level datasets into Phoneme level datasets

A technique from ASR that can also be applied to AST is transforming utterance/sentence level datasets into phoneme level datasets. There are many ASR datasets that contain single utterances such as LibriSpeech and MUSDB18

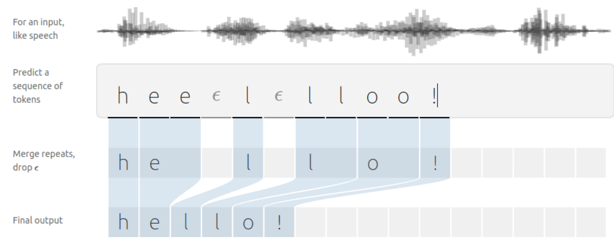

Figure 1: CTC Function collapsing repetitions [Hannun, 2017]

[Panayotov et al., 2015, Rafii et al., 2017]. These datasets can be effectively transformed into phoneme level datasets by using a phoneme dictionary such as the CMU Pronouncing Dictionary [CMU, 2023]. This dictionary contains a mapping from words to phonemes. Using this dictionary, the utterances can be transformed into phoneme sequences. These sequences can then be used with a Connectionist Temporal Classification (CTC) loss function to train AST models [Hannun, 2017]. CTC allows for the model to output a sequence of phonemes per time step and the repetitions are collapsed into a single phoneme as shown in Figure 1. This technique was used to allow a model to incorperate the utterance level LibriSpeech dataset into the training for a phoneme ASR model [Baevski et al., 2020]. Timestamps can be retrieved from the pre-CTC output that had time-aligned phoneme classifications. The same process can be applied to AST models to allow them to use the utterance level song datasets.

2.3.4 Teacher-Student Approach

The teacher-student approach is inspired by the technique of the same name that was intended to reduce the size of large deep neural networks. The idea was first train a large deep neural network, the teacher, and then train a smaller deep neural network, the student, to mimic the teacher [Abbasi et al., 2019]. However, this technique can also be applied in cases of low labeled data availablility but high unlabeled data availablility. To start off, a model would be trained on a small dataset of labeled data. Then, the model would be used to label a large dataset of unlabeled data. Finally, a new model would be trained on the newly labeled dataset [Meseguer-Brocal et al., 2018]. This has proven to be effective for transcribing drums in music with the student model outperforming the teacher model [Wu and Lerch, 2017].

2.4 Existing Datasets

2.4.1 TIMIT (ASR)

TIMIT is a dataset of speech recordings of 630 speakers of eight major dialects of American English with time-aligned phoneme sequences [Garofolo et al., 1993]. It is a popular dataset for ASR and has been used to train many ASR models. However, since this is not a singing dataset, it does not contain any of the characteristics of singing such as pitch, duration or vibrato. This is an excellent dataset to apply transformation into a pseudo-singing dataset metioned in 2.3.2. The popularity of this dataset makes it a good candidate for applying transformations to create a pseudo-singing dataset and also for a general benchmark to compare against other models.

2.4.2 LibriSpeech (ASR)

LibriSpeech is a 1000 hour dataset of audiobook recordings where each recording has a matching sentence [Panayotov et al., 2015]. This dataset is also a popular dataset for ASR and has been used to train many ASR models including recent breakthroughs like wav2vec 2.0 [Baevski et al., 2020]. This dataset is also a good candidate for training AST models because it is large and publicly available. This dataset is special because of the clarity of the audio and previous success by other models such as the wav2vec 2.0 model in detecting phonemes in this dataset when fine tuned with the TIMIT dataset [Baevski et al., 2020]. This is done through the CTC technique mentioned in 2.3.3. Since speech and singing both use the same phonemes, because they are in the same language, this large dataset can be used to train base models before fine-tuning them AST models.

2.4.3 Jamendo Dataset

This dataset one of the most popular datasets for AST and has been used by many state-of-the-art AST models. It contains 20 English songs and 60 songs in other languages with word-aligned timestamped sequences [Stoller et al., 2019]. This dataset is a good candidate for training AST models because it is publicly available and it's popularity makes it an excellent benchmark to compare against other models. However, it is still a relatively small dataset and does not contain any phoneme sequences. The authors of this dataset were able to achieve a 77.8% Word Error Rate (WER) which still leaves a lot of room for improvement. Remember to record the performance of the state of the art models on this dataset

2.4.4 MUSDB18

MUSEDB18 is a dataset of 150 songs with isolated vocals and accompaniment tracks [Rafii et al., 2017]. This dataset is a good candidate for training AST models because it is publicly available and it has clean isolated vocals. The downside of this dataset is that it doesn't contain any word level timestamps. However, with the CTC technique mentioned in 2.3.3, this dataset can still be used to train AST models.

2.4.5 NUS Dataset

This dataset is one of the few datasets that contains phoneme level timestamps [Duan et al., 2013]. There are 169 minutes of 20 unique English songs by 12 different people. The CMUDict was used for the phoneme vocabulary and timestamps were manually annotated. This dataset is the ideal dataset type for training AST models due to this level of detail. It also includes a mix of slow to fast melodies annd a balanced gender distribution [Duan et al., 2013].

2.4.6 Free Music Archive

The Free Music Archive (FMA) dataset is a dataset of 106,574 tracks with 161 genres [Defferrard et al., 2017]. This dataset is not a good candidate for direclty training AST models because it does not contain any lyrics and some music may not even contain any singing at all. However, it is a good source of unlabeled songs that could be labeled through the teacher-student technique in 2.3.4. It can also provide a good source of general singing audio that can be used in the training of wav2vec2.0 models that will be described in section 3.3.

2.4.7 VocalSet

VocalSet is a 10 hour dataset of a capella singing from 20 professional singers demonstrating a variety of singing techniques [Wilkins et al., 2018]. This dataset is a good candidate for training Voice Activity Detection (VAD) models because it contains onset and offset timestamps for each vocal segment. This is also a good a good dataset to help train AST models to know what singing sounds like.

2.4.8 Other Datasets

Many datasets were considered but left out due to the lack of availablility. Some of the most popular datasets such as Mauch's Dataset [MIREX, 2021] and Hansen's Dataset [Hansen, 2012] are not publicly available anymore. Newer datasets such as DALI [Meseguer-Brocal et al., 2018] and DAMP! [Smule, 2018] are are hidden behind institutional logins and require manually requesting access.

While both the Mauch's Dataset and Hansen's Datasets are quite small (Mauch has 20 songs, Hansen has 9 [MIREX, 2021]), the DALI and DAMP! datasets are much larger. The DALI dataset in particular used a version of the Teacher Student technique mentioned in 2.3.4 to label 105 songs with timestamps for the word and phoneme level [Meseguer-Brocal et al., 2018]. The DAMP! dataset is even larger with 300x30x2 song dataset. Both of these datasets would be excellent candidates for training AST models from their size alone.

3 Related Works

3.1 Early HMM-Based Acoustic Models

3.2 Music Informed Models

3.3 wav2vec 2.0 and Transfer Learning

3.4 Whisper Word-Level Alignment

4 Method

1. Preprocess datasets
2. Augment datasets
3. Create frankenstein dataset
4. Train model
5. Evaluate model
6. Label unlabeled datasets
7. train student model on newly labeled datasets
8. evaluate student model on original, manually labeled datasets
9. repeat

5 Results

5.1 WER

5.2 PER

6 Discussion

7 Future Work

7.1 Adversarial Training

8 Conclusion

References

[Abbasi et al., 2019] Abbasi, S., Hajabdollahi, M., Karimi, N., and Samavi, S. (2019). Modeling teacher-student techniques in deep neural networks for knowledge distillation.

[Baevski et al., 2020] Baevski, A., Zhou, H., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations.

[CMU, 2023] CMU (2023). Cmu pronouncing dictionary.

[Defferrard et al., 2017] Defferrard, M., Benzi, K., Vandergheynst, P., and Bresson, X. (2017). FMA: A dataset for music analysis. In *18th International Society for Music Information Retrieval Conference (ISMIR)*.

[Duan et al., 2013] Duan, Z., Fang, H., Li, B., Sim, K. C., and Wang, Y. (2013). The nus sung and spoken lyrics corpus: A quantitative comparison of singing and speech. In *2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, pages 1–9.

[Garofolo et al., 1993] Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., Dahlgren, N. L., and Zue, V. Z. (1993). Semi-supervised lyrics and solo-singing alignment.

[Gupta et al., 2018] Gupta, C., Tong, R., Li, H., and Wang, Y. (2018). Semi-supervised lyrics and solo-singing alignment.

[Hannun, 2017] Hannun, A. (2017). Sequence modeling with ctc. *Distill*. https://distill.pub/2017/ctc.

[Hansen, 2012] Hansen, J. K. (2012). Recognition of phonemes in a-cappella recordings using temporal patterns and mel frequency cepstral coefficients.

[Kruspe, 2015] Kruspe, A. (2015). Training phoneme models for singing with "songified" speech data.

[Meseguer-Brocal et al., 2018] Meseguer-Brocal, G., Cohen-Hadria, A., and Peeters, G. (2018). Dali: a large dataset of synchronized audio, lyrics and notes, automatically created using teacher-student machine learning paradigm. In ISMIR, editor, *19th International Society for Music Information Retrieval Conference*.

[MIREX, 2021] MIREX (2021). 2021:automatic lyrics transcription.

[Panayotov et al., 2015] Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (2015). Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

[Park et al., 2019] Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D., and Le, Q. V. (2019). Specaugment: A simple data augmentation method for automatic speech recognition. In *Interspeech 2019*. ISCA.

[Rafii et al., 2017] Rafii, Z., Liutkus, A., Stöter, F.-R., Mimilakis, S. I., and Bittner, R. (2017). The MUSDB18 corpus for music separation.

[Raissi, 2021] Raissi, M. (2021). Mel-spectrogram and mfccs.

[Smule, 2018] Smule, I. (2018). Damp-mvp: Digital archive of mobile performances - smule multilingual vocal performance 300x30x2.

[Stoller et al., 2019] Stoller, D., Durand, S., and Ewert, S. (2019). End-to-end lyrics alignment for polyphonic music using an audio-to-character recognition model.

[Wilkins et al., 2018] Wilkins, J., Seetharaman, P., Wahl, A., and Pardo, B. (2018). Vocalset: A singing voice dataset. In *International Society for Music Information Retrieval Conference*.

[Wu and Lerch, 2017] Wu, C.-W. and Lerch, A. (2017). Automatic drum transcription using the student-teacher learning paradigm with unlabeled music data.