Evaluation of ASR in Musical Environment

Justin Nguyen
Electrical and Computer Engineering
Carnegie Mellon University
jdnguyen@andrew.cmu.edu

Harshine Visvanathan
Electrical and Computer Engineering
Carnegie Mellon University
hvisvana@andrew.cmu.edu

Aarushi Wadhwa
Electrical and Computer Engineering
Carnegie Mellon University
aarushiw@andrew.cmu.edu

Jingfei Xia
Electrical and Computer Engineering
Carnegie Mellon University
jingfeix@andrew.cmu.edu

Abstract

ASR is utilized in settings when the user may be in a hands-busy or eyes-busy environment. For example, ASR is used to recognize commands when a user is operating a vehicle or in home automation devices where the user may be speaking over music or films. These noisy channels make ASR difficult to perform due to the prevalence of both in-band noise from vocals and out-of-band noise from musical instruments. We propose an extensive analysis of an ASR pipeline utilizing several modern techniques for recognizing speech from music using simulated and real data for training and evaluation.

1 Introduction

Speeches, conferences, presentations, theatrical plays, and numerous other events are environments in which speech, and music commonly exist together. In recording and transcribing speakers at these events, background noise often interferes with with the speech signal alone. This weakens the performance of Automated Speech Recognition (ASR) systems in the task of speech transcription.

In the environments described above, music is often a substantial source of background noise. Thus, this paper aims to evaluate the performance of ASR systems in recognizing speech from musical environment.

In this task, it is important that any background music does not interfere with the speech transcription. Only then will the main speaker at any given time can easily be distinguished for automated speech transcription purposes. We recognize that noise and music at any event can vary drastically and can be interpreted dis-similarly. Thus, we aim to understand how the performance in the task of speech recognition, as performed by an ESPnet model, correlates to each of the following attributes:

- Signal-to-Noise-Ratio (SNR)
- Single-layer instrumental music
- Music genres
- Whether the training data included utterances with or without background music

2 Related Work

Audio source separation is a widely researched topic; as a result, there are numerous methods that have accomplished this task with varying success rates [1, 2, 3, 4, 5]. According to Al-Shoshan, a majority of the methods can be broken into three categorical approaches: time-domain, frequency-domain, and time-frequency domain; all requiring distinct music and speech libraries to serve as their test datasets [6]. Al-Shoshan also iterates that audio signals can be broken into nine categories: speech signal of a single speaker, solely music signal, single talker and music signal mixture, singer and music signal mixture, singer signal, abnormal music tones, multiple speakers, noises that are not music nor speech signals, multi-speaker/multi-singer multi-music signal mixture. Thus, it is clear that there are four types of music signals that could exist in the music library of a music-speech separation problem. A single layer music signal can be further divided based on the source of its audio (such as woodwind, brass, percussion, piano instruments), since they all have differing characteristics. Following this breakdown, our study will analyze the interaction and impact when speech is mixed with both, various instruments and songs.

There are several related works and fields to this analysis on speech recognition with music. These related works look at joint music-speech separation in ASR models, separate models for suppressing in-band noise, and studies looking at the effect of music on ASR performance and augmentations to ASR training to improve the system's WER.

For example, Hughes looks at the music-speech separation problem and looks at the MAX and Algonquin noise suppression models for performing music speech separation [1]. Using 38 hours of manually transcribed utterances form spoken queries to Google Voice Search, Hughes and his team simulated a noisy environment by mixing the clean utterances with a random song from a database of 500 popular songs at their desired SNR (10dB) to form their training dataset. They were able to improve their word error rate on a GMM ASR model by over 8% [1].

Woo et. al. looked at applying end-to-end models for music-mixed speech separation using a jointly trained multi-speaker ASR system to separate the speech from music and recognize the speech [5]; they evaluated their method through ASR experiments using speech data mixed with background music from a wide variety of Japanese animations.

Li proposes a system to separate the singing voice from background accompaniment in monoaural recordings by using vocal detection and pitch detection techniques [7].

The studies by Hughes [1], Woo [5], and Li [7] looked at modeling the differences between speech and the music to separate them before running the speech recognizer on the separated signal. This pre-processing step has been shown to give substantial improvements in WER, but trains a separate model to handle the noise and does not address the impact of noise in the recognizer directly. The field of music-speech separation is also highly related to the field of music lyric transcription which faces the similar problem of in-band and out-of-band noise from music and possibly other vocals and has had more research in studying the direct effects of noise in speech transcribers.

Dabike [8], Gupta [9], and Stoller [10] investigate lyric transcription using various end-to-end deep learning models. Dabike was able to curate a dataset using Karaoke performances which yielded a WER or 19.6% [8]. Gupta and Stoller both looked at the Mauch and Jamendo music datasets and were able to build ASR systems were able to achieve WER of 52% [9], and 48.9% [10] respectively. Guptas study performed an analysis of ASR system performance versus genre broadclasses and trained a separate ASR system that was provided the genre a-priori; they found that their ASR system performed almost twice as well on pop genre broadclass compared to the hip hop and metal genre broadclasses in WER and that hip hop and pop had better lyrics alignment compared to the metal genre [9]. The genre-informed model showed 2-4% absolute improvement in WER [9].

Our work will continue on these studies by looking specifically at how instrumentation, genre, speaker, and speaker gender all play a role in speech recognition when mixed with music. We will be combining the approaches of both music-speech separation and automatic lyric-transcription by:

1. Simulating a noisy environment by mixing both instrumental and vocalized music with an utterance, described in Section 3.5

2. Transcribing the utterance using an end-to-end ASR model without the use of a preprocessing step, described in Section 3.6

3. Analyzing the impact of noise on the resulting transcriptions, described in Section 4

3 Methodology

3.1 ESPnet

We use the ESPnet toolkit [11], an open source platform for end-to-end speech processing which focuses on end-to-end automatic speech recognition (ASR), and adopts widely-used dynamic neural network toolkits, Chainer and PyTorch, as a main deep learning engine, to develop our models. This toolkit uses Kaldi ASR toolkit style for data processing, feature extraction/format, and recipes to provide a complete setup for speech recognition and other speech processing experiments.

3.2 Model Training

In our work, we train a Transformer (CTC) Model with two different datasets and perform our analysis on both of these models.

Transformer Model The transfomer model [12] aims to solve sequence-to-sequence tasks while handling long-range dependencies with ease. In our work, we used the transformer model to perform both acoustic and language modelling.

CTC The Connectionist Temporal Classification (CTC) criterion [13] is a criterion which jointly infers the segmentation of the transcription while increasing the overall score of the right transcription. We have used CTC in our work to reduce amount & complexity of code & fudge factors for similar accuracy.

3.3 Analysis Trial Setup

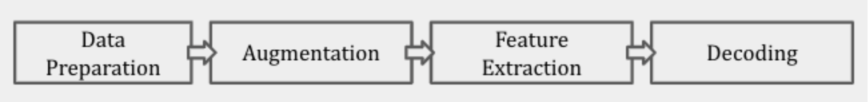


Figure 1: Trial Setup

In our work, the basic steps involved in decoding, like the data preparation, feature extraction and decoding steps are similar to the ones outlined in the ESPnet platform. Additionally, we perform a data augmentation step, to mix our audio and speech signals, which is explained in the Section 3.5.

3.4 Datasets

To perform our analysis, we mixed speech and audio signals. The following are the datasets we used in our work,

Wall Street Journal (WSJ) Dataset The WSJ dataset [14] is used as the primary speech signal on which the music is transposed. It includes approximately 70 hours of speech, between male and female speakers. There are two datasets for training: LDC93S6B as wsj0 and LDC94S13B as wsj1. Each dataset contains the origin video of the soundtrack for several seconds as a wv1 file and a summary of corpora for each utterance.

SigSep MUSDB18 Dataset The SiSEC DSD100 dataset [15] consists of single-layer & multi-layer music. The music layers we have considered in our work are bass, drums, and vocals. This dataset comprises of approximately 100 songs per instrument layer.

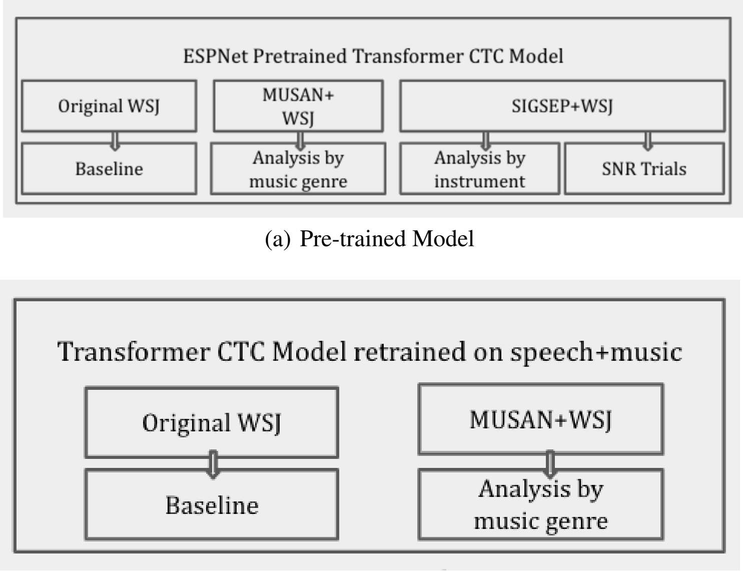
MUSAN Dataset The MUSAN dataset [16] consists of full music mixtures/ songs of various genres. The genres we have considered for our work are rock, blues, and soul. This dataset also comprises of approximately 100 songs per genre.

3.5 Music-Speech Simulation

The music and speech mixes were simulated by mixing the two audio sources prior to feature extraction. The WSJ dataset utterances were raw 16-bit 16kHz mono audio recorded without any mastering or preamplifiers, so the output levels were relatively low at about -30dB[14]. The music on the other hand has already been mastered, but tracks from both dataset were recorded in stereo at 24-bit 44.1kHz[15, 16]. In addition, the utterances are around 4 - 15 seconds long, much shorter than the songs which were around 3 minutes long. To properly get a reliable mix, the selected songs from the SIGSEP dataset were mixed at 15 seconds in for instrumental and mixed tracks when all instruments were playing. For the vocal isolated tracks in the SIGSEP dataset, a clip taken from the 90 second mark was taken since in the majority of the selected songs, the vocalist would be singing at this point. For the MUSAN dataset, the clip of audio to be mixed with speech was chosen by examining at the energy spectrum of the audio signal. A 20 second audio of high energy is chosen from the full-length audio signal and mixed with the speech signal to generate the samples for full music mixture trials.

After the mixing start points were described for music samples in both datasets, the following mixing procedure was used for each song:

1. Select a random song to mix with this utterance
2. Downsample the music track to 16-bit 16kHz mono
3. Slice the downsampled music track starting from the previously described start point to the length of the mixed utterance
4. Normalize the downsampled and sliced audio track to the negative of the desired SNR (dB) peak level
5. Normalize the utterance to 0dB peak level
6. Mix the normalized audio track and utterance
7. Normalize the mixed track to 0dB peak level
8. Feed the output to the recognizer



(a) Pre-trained Model

(b) Retrained Model

Figure 2: Experimental Setup.

3.6 Experimental Setup

We developed our models based on the ASR implementation in ESPnet toolkit by Watanabe et al [11]. In evaluating speech recognition capabilities from environments with background music, we curated a dataset that mimicked this environment. There are two components in these signal inputs: the speech itself, and the music. The music dataset includes various permutations of characteristics: genres, layers of music, noise. In evaluating single-layer and multi-layer music, we turned to the SIGSEP MUSDB18 [15] as a data source and for full music mixtures, we used the MUSAN dataset [16]. The energies of these initial audio files from these datasets were manually evaluated and we thoroughly vetted 20 seconds of each audio component that provided sufficient data. We trained two models for our analysis, both with the baseline architecture. One was trained using only the utterances from the WSJ dataset [14] and the other network was trained on a full music mixture dataset, consisting of 100 different songs mixed with the WSJ dataset. Our analysis includes results on various SNR trials, single-layer music trials and full music mixture trials, along with gender and genre analysis for every trial.

4 Results

4.1 SNR Trial

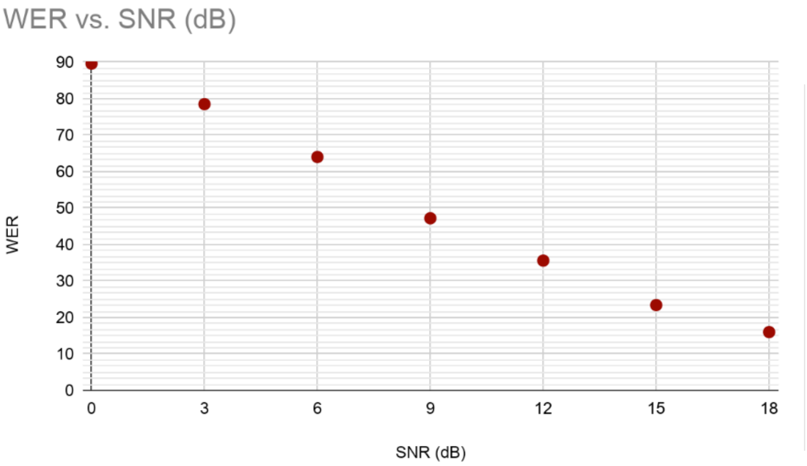


Figure 3: WER across SNRs from 0dB to 18dB in steps of 3dB.

The WSJ utterance and the music can be mixed at many different levels. To keep a consistent SNR-dependent error rate for all of our analysis we performed a SNR trial to find the ideal SNR value to mix our audio sources at. For all the following experiements, a 12dB SNR was used as the mixing level. The mixing process was described previously in Section 3.5. The result from the SNR trial is shown in Figure 3.

4.1.1 Analysis

For our first trials using the pretrained ESPNet CTC Transformer Model, we found that a SNR from 12dB to 15dB yielded the most reasonable transcription results where the music was not too loud that the the ASR system was unable to discern the speech, but not so soft that there were too few errors to analyze the effect of music on the recognizer. At 12dB and 15dB SNR, the recognizer would begin to transcribe both the speech and occasionally transcribe vocals from the music as well which proved to be a good balance for analyzing how music affected the transcription. This result matches that of Hughes who found that a 10dB SNR closely matched the sound of moderately loud background music [1].

While at 18dB SNR, the music was extremely quiet, the WER with the pretrained model was still twice as high compared to the original results without any music mixed. We think the larger WER can be attributed to the normalization applied to the utterance which, while normalized the peak level, can still result in clipping for extremly quick changes in the PCM codes.

4.2 Instrumental Trial

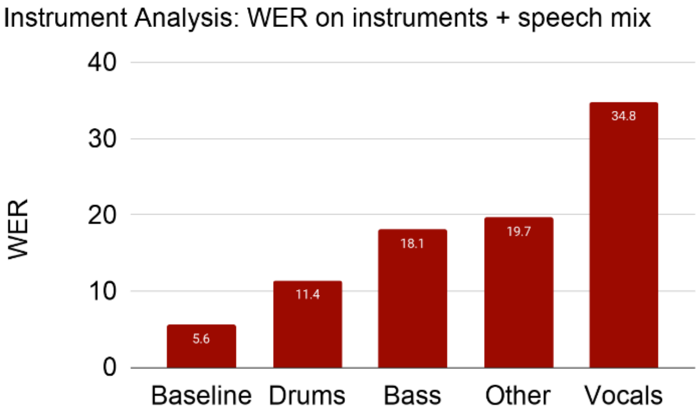


Figure 4: WER of recognized utterance when mixed with different instruments from a song. Baseline result is from the model decoding the utterance with no music mixed in.

After performing the SNR trial, we then performed an analysis of the effect on WER when utterances are mixed with instrument in hopes that the low-level analysis would help our analysis of the result when the utterance is mixed with full songs described later. The instrumental tracks are from the SIGSEP dataset [15] which breaks down the songs in to bass, drum, other, and vocal tracks for every song in the dataset. The other track is guitar, violin, or some other mid-ranged instrument in the song. The WER result for each instrument's mix is shown in Figure 4. As expected, when mixing instrumental noise with the utterance, the WER went up, with the vocal noise yielding the greatest impact on WER.

4.2.1 Analysis

While our results did validate our existing preconception that vocals would affect *speech* recognition the most, it is important to note the progression of instruments as the WER increases. Compared to the baseline, drums has the least effect on WER, followed by bass and other instruments which had similar impact on WER. From our analysis and listening of the drum and speech mixtures, at the ideal background music SNR of 12dB, our perception of the the lower pitched drums (toms and kick) were much fainter than our perceived intensity of higher pitched percussion (hi-hat, snare); in addition the higher pitched percussion sounded almost like white-noise at these SNRs leading to the least impact on WER.

More in-depth analysis bass, other, and vocals on the speech recognition require looking more closely at their interaction with male and female speakers which will be discussed in Section 4.4.1.

4.3 Genre Trial

We have also conducted experiments to analyze effect on WER when mixing speech with different genres of music. We use Blues, Rock and Soul music in MUSAN dataset [16]. Each piece of music is extracted from the middle of a complete song and lasts 20 seconds. Then we compare WER when speech is mixed with different types of music as before. The result is shown in Figure 5.

4.3.1 Analysis

From the plot we can find that after adding music as background noise WER becomes much larger than that in clean speech as we have expected. It is also noted that Rock music has greatest effect on WER, followed by Blues music and Soul music. If the speech is mixed with Rock music, most of the texts are covered by the strong beats of the drum and bass, which makes it very hard to identify. However in Blues and Soul music there are not as many beats as in Rock music, so the error rate is much lower. Another reason is that Rock music has high energy across a wide range of frequency domain while Blues and Soul music are mainly concentrated on low frequency domain. So Rock

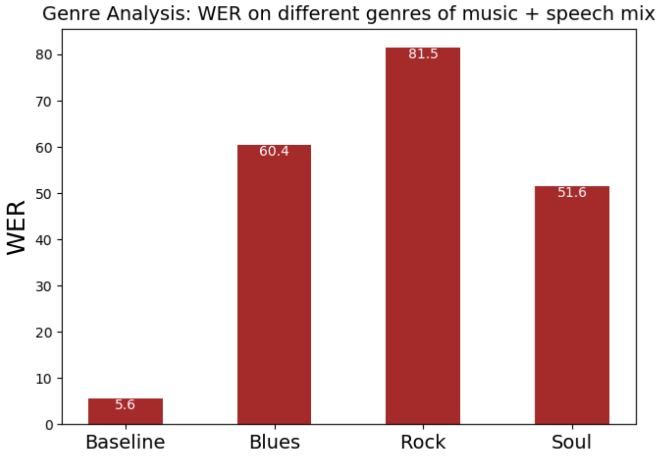


Figure 5: WER of different genres of music

music will have a large impact on all kinds of voice in speech while Blues and Soul music covers only the voice in low frequency domain such as male speakers' voice. We will discuss this in detail later.

We can also see that in Soul music the error rate is smaller than in Blues music. The difference is probably caused by the vocal part of different piece of music. As we have mentioned before, vocals would affect speech recognition the most. In Blues it contains more words in the lyrics than Soul music which would make speech recognition much harder and thus produce larger error rate.

4.4 Gender

To understand the nuanced affects of how speech is perceived by an ASR system, we further analyzed how speaker gender impacts WER. We perform two gender-based transcription experiments. The first analyzes the impact that the gender has on the ASR system with background single-layered instrumental music - as evaluated in Section the Instrumental Trial. The second analyzes the impact that the gender has on the ASR system with mixed music of varying genres - as evaluated in the Genre Trial.

4.4.1 Gender Instrument Analysis

Figure 6 shows the average WER as seen across each evaluated instrumental layer, and gender. From the figure, we note that the performance of the ASR system on these instrumental layers increases in the following order: bass, drums, other, vocals. It is further telling that the difference between WER for each genre between males and females is marginal. However, the impact of gender on ASR performance with single-layered instrumentals in the background is not negligible in many edge cases.

This result can be seen this in Figure 7 which views variance across all decodings while Figure 8 shows the individual error breakdowns for a 1.5 IQR scale (2.7σ). The exclusion of outliers brings the male and female decodings back in line with one another. Across all instrumental layers, we can see that male speakers have equal or greater global variance in the ASR system's performance than female speakers, most notably in mixtures with bass, other, and vocals.

After analyzing the speaker audio, we noticed that the male speakers had varying ways of speaking. The male speakers could be categorized into the following buckets: reporter, narrator, southern-accent, monotonous. Whereas, the female speakers could not audibly be characterized nor clustered because they all spoke in similar styles (this held true for all except one female speaker who could be distinguished as an elderly lady). This distinction in the gender-based speaker groups explains

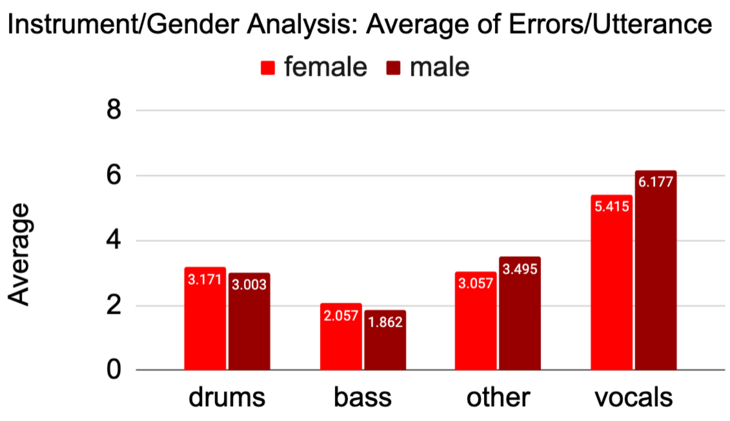


Figure 6: *WER average* of different genders on different instrumental music layers

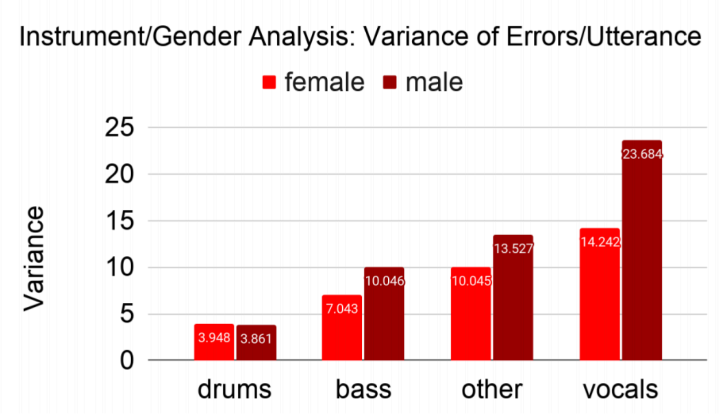
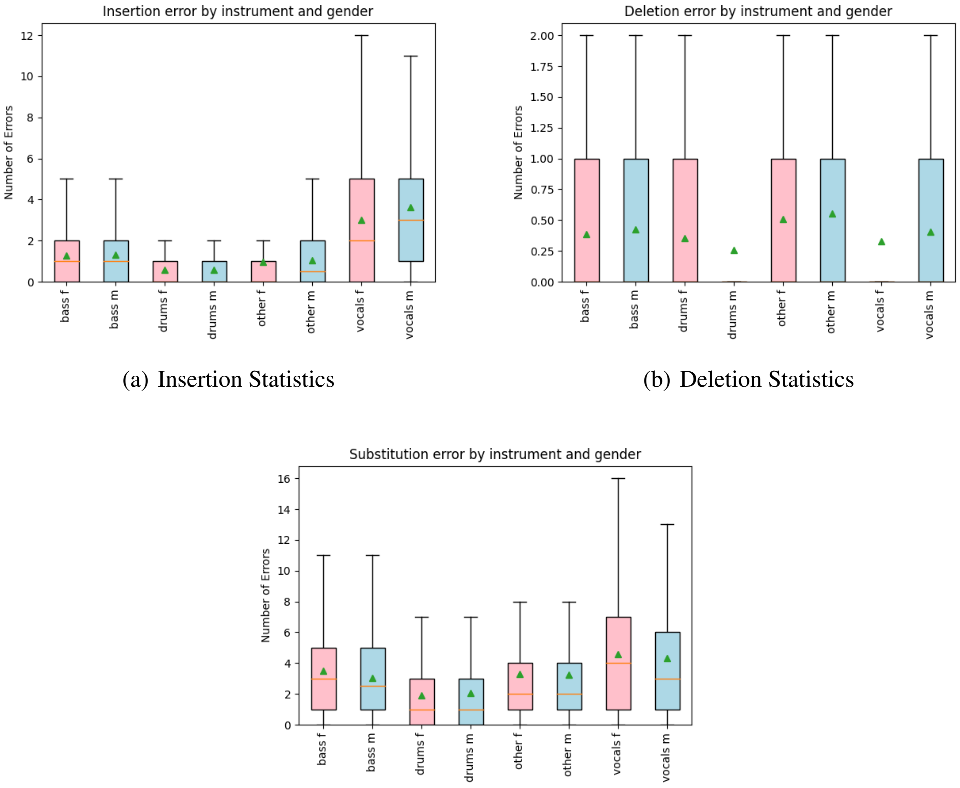


Figure 7: *Variance of error counts* of different genders on different instrumental music layers

the greater variance in the male spoken group, than the female spoken group. The analysis overall shows that the ASR system *is* sensitive to intonations in the targeted speech transcription.

4.4.2 Gender Genre Analysis

Figure 9 shows WER difference between male and female when we use different genres of background music. As we can see from the Figure, if we use only clean speech dataset, there is not much difference between male speakers and female speakers. If we add Rock music as the background noise, WER for male and female speakers are still almost the same. However, if we add Blues and Soul music, the model makes much more mistakes in retrieving speech text for male, especially when we use Blues as our background noise.



(a) Insertion Statistics

(b) Deletion Statistics

(c) Substitution Statistics

Figure 8: Breakdown for each type of error in the instrument trial by instrument and gender. Medians are shown in yellow and means in green.

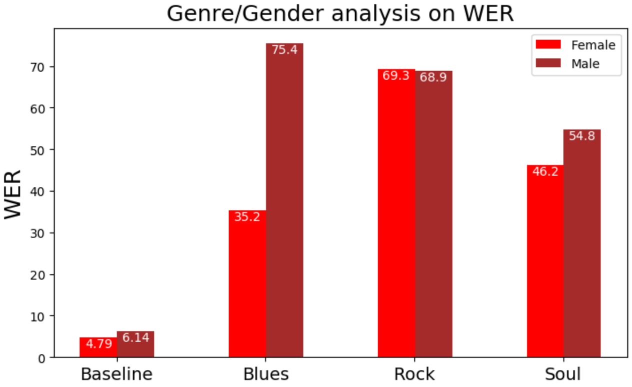
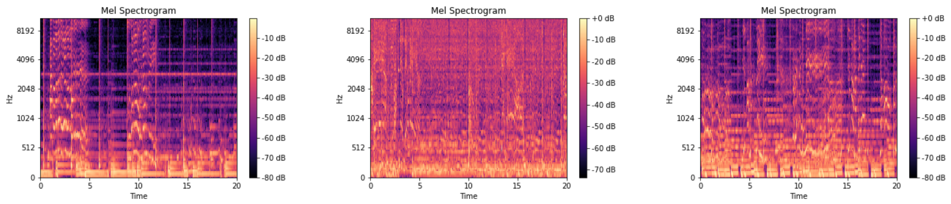


Figure 9: WER of different genders on different genres of music

From our analysis, we were able to see a clear distinction in the results because the different genres have different frequency spectra. If we examine at the mel-spectrogram for different genres of music in Figure 10, we can see that in Rock music the energy of spectrum is almost the same everywhere. It is similar to white noise and it has the same effects on both a male and a female speaker. For Blues and Soul music, more energy is concentrated on lower frequency. As we know male voice lies on lower frequency than female voice. As a result, Blues and Soul music affect male speech more and cause the WER to be higher. We can also see that with more energy concentrated in lower frequency, WER difference between male and female will become larger. This is why in Blues music it has larger male-female difference than in Soul music.



(a) Blues music

(b) Rock music

(c) Soul music

Figure 10: Mel-spectrogram for different genres of music

4.5 Retrained Model

To examine the results of our model trained on pure speech data, we trained another CTC Transformer Model on a mixture of speech and audio. The audio samples were chosen from the MUSAN dataset [16] and comprised of 100 different audio samples, 50 Rock, 25 Blue and 25 Soul. These songs were chosen randomly and mixed with the WSJ Dataset [14] and the retrained model was trained. The results we obtained are shown in the Figure 11.

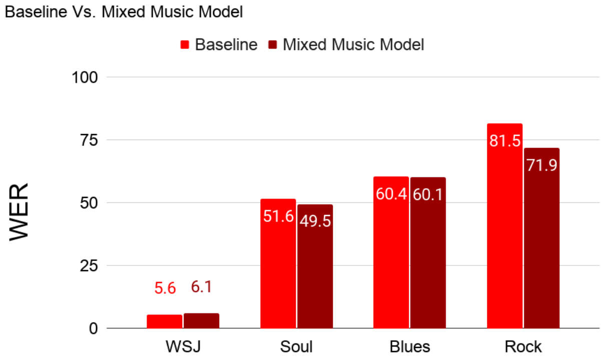


Figure 11: Baseline Vs. Mixed Music Model

4.5.1 Analysis

Since the mel spectograms and the audio samples for blues and soul genres were similar, we chose to train our model with more songs from the rock genre than the other two. The analysis reported was done on 20 new audio samples from the MUSAN dataset, 10 Rock, 5 Blue, and 5 Soul, mixed with the WSJ dataset. From this analysis we can see that there is a distinct improvement in the performance of the model for the Rock genre and also a slight increase in the performance of the Soul and Blues genres, which can be accounted to the fact that the model was trained on more songs from the rock genre and is able to perform better in that setting than the others. We also see a slight performance decrease for the pure WSJ dataset, which can be explained by the fact that we have trained the model with only music mixed with speech and does not include pure speech data.

5 Conclusion and Future Work

From our analysis we can see a significant improvement on the mixture music dataset using the model trained with the full mixture music dataset. We have performed an in-depth analysis of the effects of different instruments, different SNRs, and different genres of music in an ASR system to aid the development of an end-to-end model which can perform ASR on far-field speech.

In summary, we have produced several analysis on our speech recognition model. While our results have provided extensive new insights and proof for expected intuition, there are numerous additional steps that can be taken to expand upon, and dig deeper into, our analyses. These suggestions for future work are listed below.

- Future studies can expand upon our instrumental analysis and learn the tasks performance with other instruments such as flute, trombone, piano, guitar, etc.
- Future studies can expand upon our genre analysis and learn the tasks performance with other genres like jazz, theatre, opera, etc.
- Future studies can consider a new analysis: an in-depth analysis of how various categories of background noise impacts music-speech separation and recognition.
- Future studies can consider a new analysis: an in-depth analysis of how specific types of speech intonations (such as narration, reporting, monotonic, etc.) impact music-speech separation and recognition.
- Future studies can re-evaluate the ESPnet model with permutation invariant training using the WSJ2mix dataset with a given the ground truth lyrics from the music.
- Future studies can vary attention models in time and/or frequency.

References

[1] T. Hughes and T. Kristjansson, "Music models for music-speech separation," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4917–4920, 2012.

[2] Guoning Hu and DeLiang Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Transactions on Neural Networks*, vol. 15, no. 5, pp. 1135–1150, 2004.

[3] A. J. R. Simpson, G. Roma, and M. D. Plumbley, "Deep karaoke: Extracting vocals from musical mixtures using a convolutional deep neural network," 2015.

[4] N. Takahashi and Y. Mitsufuji, "Multi-scale multi-band densenets for audio source separation," pp. 21–25, 10 2017.

[5] J. Woo, M. Mimura, K. Yoshii, and T. Kawahara, "End-to-end music-mixed speech recognition," 2020.

[6] A. Al-Shoshan, "Speech and music classification and separation: A review," *Eng. Sci*, vol. 19, pp. 95–133, 01 2006.

[7] Y. Li and D. Wang, "Separation of singing voice from music accompaniment for monaural recordings," *The Journal of the Acoustical Society of America*, vol. 122, p. 2989, 01 2007.

[8] G. R. Dabike and J. Barker, "Automatic Lyric Transcription from Karaoke Vocal Tracks: Resources and a Baseline System," in *Proc. Interspeech 2019*, pp. 579–583, 2019.

[9] C. Gupta, E. Yılmaz, and H. Li, "Automatic lyrics alignment and transcription in polyphonic music: Does background music help?," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 496–500, IEEE, 2020.

[10] D. Stoller, S. Durand, and S. Ewert, "End-to-end lyrics alignment for polyphonic music using an audio-to-character recognition model," pp. 181–185, 05 2019.

[11] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. E. Y. Soplin, J. Heymann, M. Wiesner, N. Chen, *et al.*, "Espnet: End-to-end speech processing toolkit," *arXiv preprint arXiv:1804.00015*, 2018.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017.

[13] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, pp. 369–376, 2006.

[14] D. B. Paul and J. Baker, "The design for the wall street journal-based csr corpus," in *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*, 1992.

[15] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner, "The MUSDB18 corpus for music separation," Dec. 2017.

[16] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.