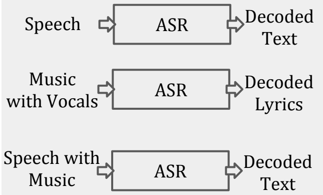
# Music-Speech Separation Evaluation

Justin N., Harshine V., Aarushi W., and Jingfei X.

Carnegie Mellon University

## Motivation



Speech → ASR → Decoded Text

Music with Vocals → ASR → Decoded Lyrics

Speech with Music → ASR → Decoded Text

**Basis:** Standard ASR

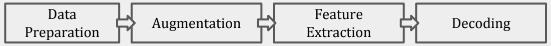
**Relevant work:** genre analysis, & automated lyric transcription

**Proposed work:** SNR, genre, gender, instrument analysis for speech transcription

## Datasets

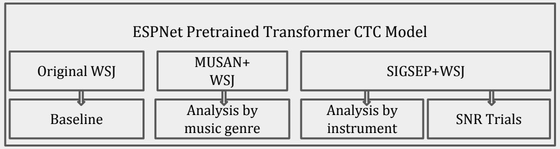
| | |
|---|---|
| WSJ [3] | - primary speech signal<br>- 70 hours of speech, between male and female speakers |
| SiSEC DSD100 [1] (SigSep MUSDB18) | - Single-layer & multi-layer music<br>- Music Layers: Bass, Drums, & Vocals<br>- ~100 songs/instrument layer |
| MUSAN [4] | - Full Music Mixtures<br>- Genres: Rock, Blues, and Soul<br>- ~100 songs/genre |

## Methodology



Data Preparation → Augmentation → Feature Extraction → Decoding

Trial Setup

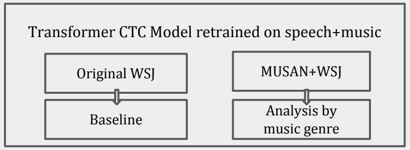


ESPNet Pretrained Transformer CTC Model

Original WSJ → Baseline

MUSAN+ WSJ → Analysis by music genre

SIGSEP+WSJ → Analysis by instrument; SNR Trials



Transformer CTC Model retrained on speech+music

Original WSJ → Baseline

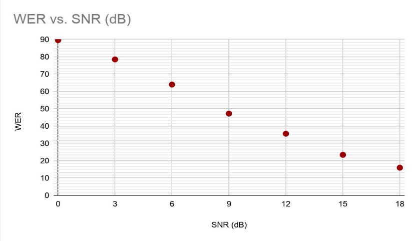
MUSAN+WSJ → Analysis by music genre

Experimental Setup

## Results and Analysis

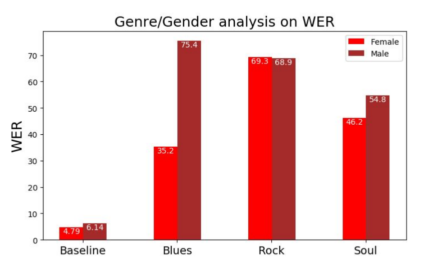
**Pretrained Transformer Model**

### Experiment: SNR



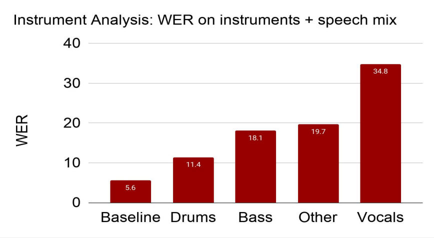
In order to pick a suitable SNR to perform our experiments at, we experimented with various SNRs to find a suitable value which would yield realistic results.

### Experiment: Genres and Gender



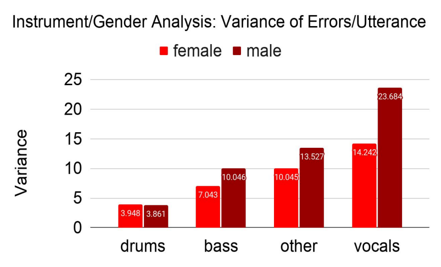
Blues and soul had a greater effect on male speaker scores because they have more energy in low frequency.

### Experiment: Instruments



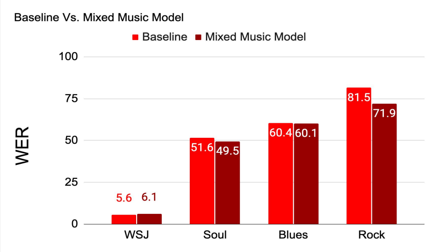
The instrumental mixes perform worse than the baseline. Drums imitate white noise at low SNRs resulting in the least impact compared to tonal noise.

### Experiment: Instruments and Gender



Male speakers had significant variance in errors except when mixed with drums with statistically equal change.

**Retrained model with mixed music and speech**



### Experiment: WSJ training dataset

This model trained on the dataset containing music mixtures, shows an improvement in the WER for specific genres but no improvement in the case of the original WSJ dataset.

## Conclusion

We performed an in-depth analysis of the effects of varying genres, instruments, and SNRs of music in an ASR system to inform our development of an end-to-end model which performs ASR on far-field speech, and music-speech separation. We see an improvement using the model trained with the mixture music dataset.

## Future Work

- Expand upon the instrumental analysis with flute, trombone, piano, guitar, etc.
- Expand upon the genre analysis with jazz, theatre, opera, etc.
- An in-depth analysis of how various categories of background *noise* impacts music-speech separation and recognition
- Re-evaluate the ESPnet model with permutation invariant training using the WSJ2mix dataset with a given the ground truth lyrics from the music
- Vary attention models in time and/or frequency

## References

1. Antoine Liutkus, Fabian-Robert Stöter, Zafar Rafii, Daichi Kitamura,Bertrand Rivet, Nobutaka Ito, Nobutaka Ono, and Julie Fontecave 13th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA 2017)
2. C. Gupta, E. Yılmaz and H. Li, "Automatic Lyrics Alignment and Transcription in Polyphonic Music: Does Background Music Help?," ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020, pp. 496-500, doi: 10.1109/ICASSP40776.2020.9054567.
3. D. B. Paul and J. M. Baker, “The design for the Wall Street Journal-based CSR corpus,” in Proceedings of the workshop on Speech and Natural Language. Association for Computational Linguistics, 1992, pp. 357–362.
4. Snyder D. et. al. "Musan: A music, speech, and noise corpus." arXiv preprint arXiv:1510.08484 (2015)

# Music-Speech Separation Evaluation

**Justin N., Harshine V., Aarushi W., and Jingfei X.**

Carnegie Mellon University

## Motivation

### Problem Statement:

- Speeches, conferences, presentations, theatrical plays, etc. are environments in which speech, and music  commonly exist together
- In recording and transcribing these events, it is important that any background music is disregarded

### Proposed Approach:

- The main speaker at any given time is can then easily be distinguished for automated speech transcription purposes
- We recognize that noise and music at any event can vary drastically and can be interpreted dissimilarly. Thus, we aim to understand how the performance in the task of music-speech separation, as performed by an ESPnet model, correlates to each of the following attributes:

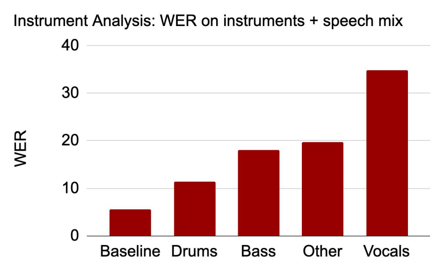
SNR, Musical instruments, Genres, Whether the training data included utterances with/without background music

## Methodology

- Our models are based on the ESPnet implementation by Watanabe et al. [1]
- 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 music by instrument, we turned to the SIGSEP DSD100[4] as a data source. In evaluating genres of full music mixtures, we used the MUSAN dataset[2].
- 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[3] 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 on unseen songs, along with gender and genre analysis for every trial.

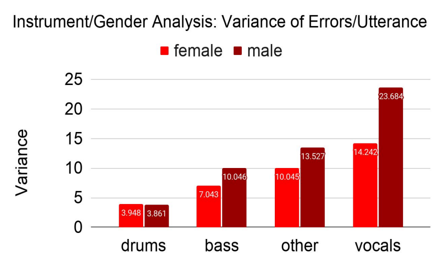
## Results and Analysis

### Experiment: Instruments



After training the ESPnet model on the utterance data excluding music, we see that that all instrumental mixes perform worse than the baseline, but the  vocals has the greatest WER and the drums have the least.

### Experiment: Instruments and Gender

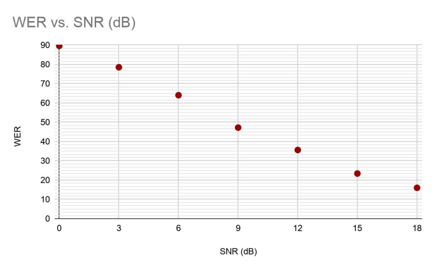


After training the ESPnet model on the utterance data excluding music, we see that the female speech and vocals mix has a large variance compared to the male's.

### Experiment: WSJ training dataset (*right*)

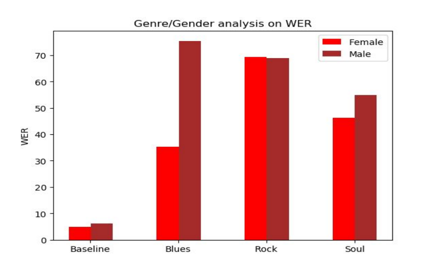
This model trained on the dataset containing music mixtures, shows an improvement in the WER for specific genres but no improvement in the case of the original WSJ dataset.

### Experiment: SNR

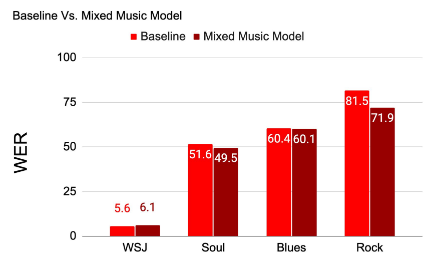


In order to pick a suitable SNR to perform our experiments at, we tried various SNRs to find a suitable value which would give us realistic results.

### Experiment: Genres and Gender



After training the ESPnet on the utterance data excluding music, we see that in Blue and Soul music, male has larger WER but in Rock music WER is almost the same.



## Discussion

- We performed an in-depth analysis of the effects of different instruments, SNRs, and genres of music in an ASR system to aid the development of an end-to-end model which can perform music and speech separation and ASR on far-field speech.

## Future Work

- Future studies can expand upon our instrumental analysis and learn the task's performance with other instruments such as flute, trombone, piano, guitar, etc.
- Future studies can expand upon our genre analysis  and learn the task's 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 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

## Related Work and References

1. Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, and Tsubasa Ochiai, "ESPnet: End-to-End Speech Processing Toolkit," Proc. Interspeech'18, pp. 2207-2211 (2018)
2. Snyder, David, Guoguo Chen, and Daniel Povey. "Musan: A music, speech, and noise corpus." arXiv preprint arXiv:1510.08484 (2015)
3. D. B. Paul and J. M. Baker, "The design for the Wall Street Journal-based CSR corpus," in Proceedings of the workshop on Speech and Natural Language. Association for Computational Linguistics, 1992, pp. 357–362.
4. Antoine Liutkus, Fabian-Robert Stöter, Zafar Rafii, Daichi Kitamura,Bertrand Rivet, Nobutaka Ito, Nobutaka Ono, and Julie Fontecave 13th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA 2017)

# Planning - Motivation

- Speeches, conferences, presentations, theatrical plays, etc. are environments in which speech, music, and noise commonly exist together
- In recording and transcribing these events, it is important that any background noises are disregarded
  - Background noises often include music, or can be other noises modeled as music
- The main speaker at any given time is then easily distinguished for automated transcription purposes
- We recognize that noise and music at any event can vary drastically. Thus, we aim to understand how the performance in the task of music-speech separation, as performed by an ESPnet model, correlates to each of the following attributes:
  - SNR
  - Musical instruments
  - Genres
  - Whether the training data included utterances with/without background music/noise

# Planning - Motivation

- Speeches, conferences, presentations, theatrical plays, etc. are environments in which speech, music, and noise commonly exist together
- In recording and transcribing these events, it is important that any background music is disregarded
- The main speaker at any given time is can then easily be distinguished for automated speech transcription purposes
- We recognize that noise and music at any event can vary drastically and can be interpreted dissimilarly. Thus, we aim to understand how the performance in the task of music-speech separation, as performed by an ESPnet model, correlates to each of the following attributes:
  - SNR
  - Musical instruments
  - Genres
  - Whether the training data included utterances with/without background music/noise

# Planning - Methodology

- We developed our models based on the ESPnet implementation by Watanabe et al.
- 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 as a data source and for full music mixtures, we used the MUSAN dataset.
- 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.

# Planning - Methodology - Cont.

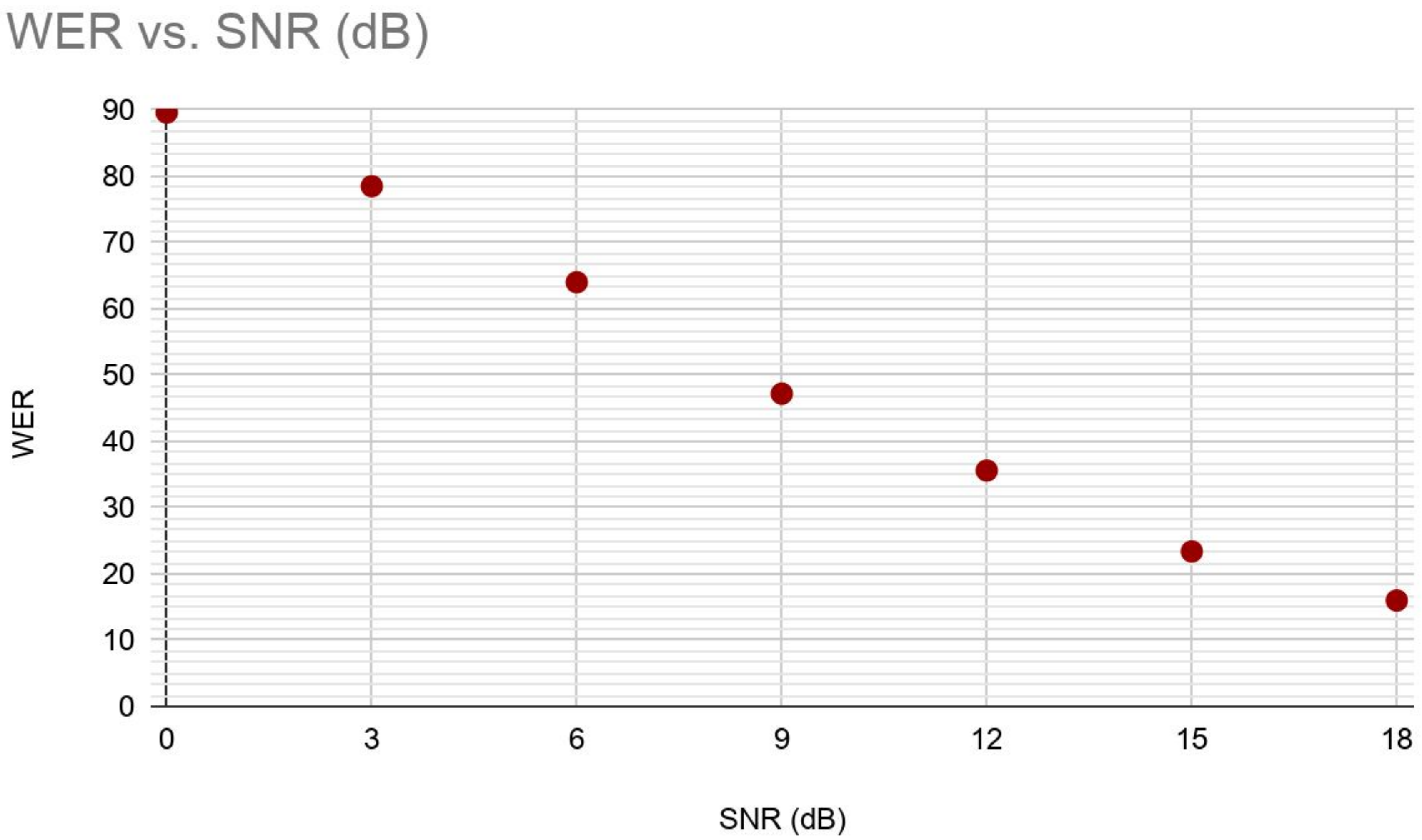
- We trained two models for our analysis, both with the baseline architecture. One was trained using only the utterances from the WSJ dataset and the other network was trained on a full music mixture dataset, consisting of 100 different songs mixed with the WSJ dataset.
- Our analysis includes results on various SNR trials, single-layer music trials and full music mixture trials, along with gender and genre analysis for every trial.

# Planning - Methodology

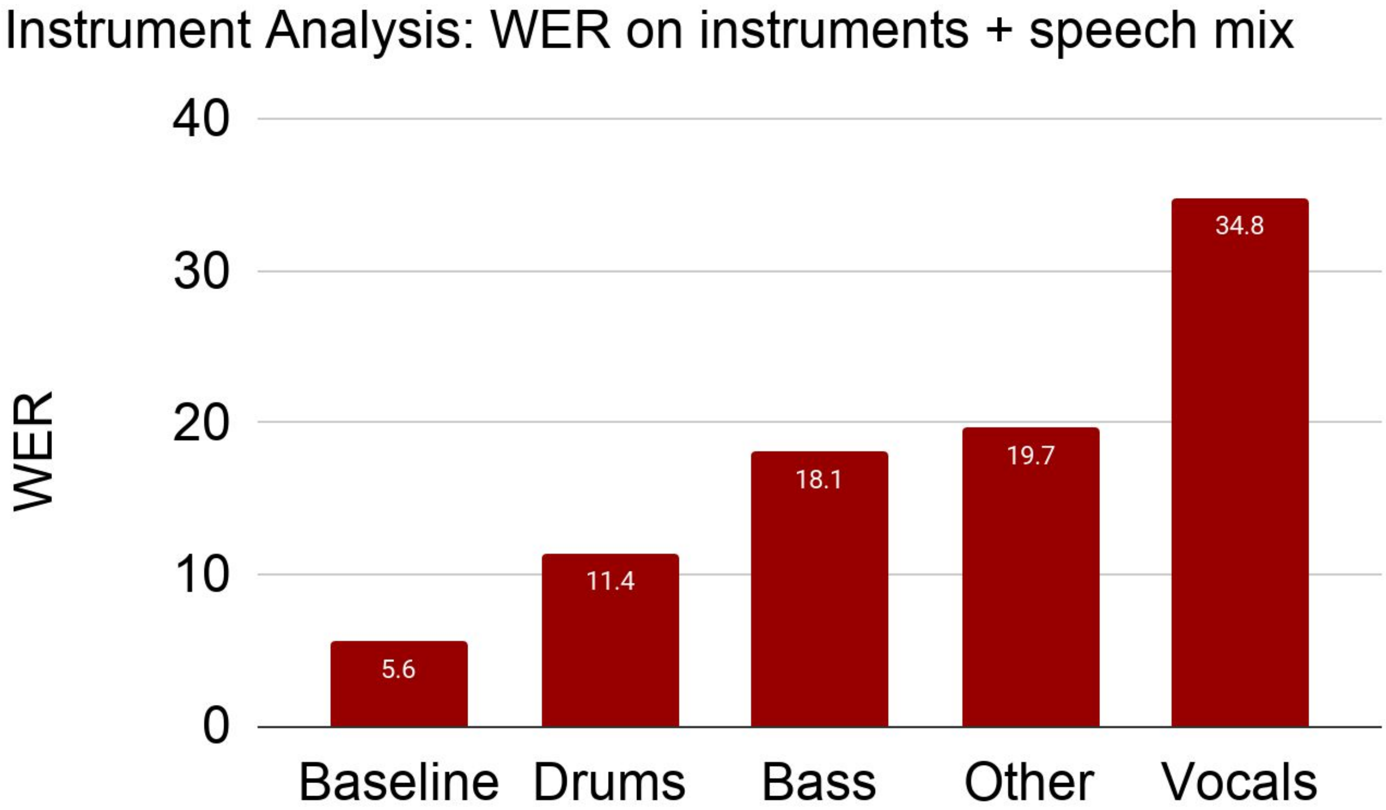
- We developed our models based on the ESPnet implementation by Watanabe et al.
- In evaluating speech recognition capabilities from environments with background music, we curated a dataset that mimicked this environment.
- There are two components in these signal inputs: the speech itself, and the music. 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 as a data source and for full music mixtures, we used the MUSAN dataset.
- 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.

# Planning - Methodology - Cont.

- We trained two models for our analysis, both with the baseline architecture. One was trained using only the utterances from the WSJ dataset 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.


WER vs. SNR (dB)
WER
90
80
70
60
50
40
30
20
10
0
0
3
6
9
12
15
18
SNR (dB)

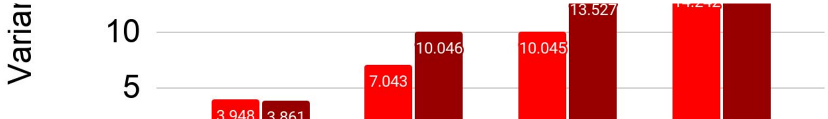
# Planning - Results



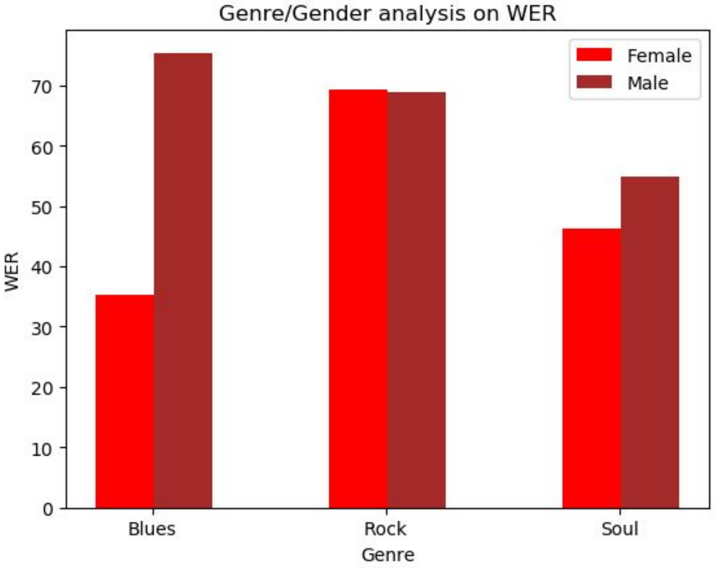
Primary plot fc

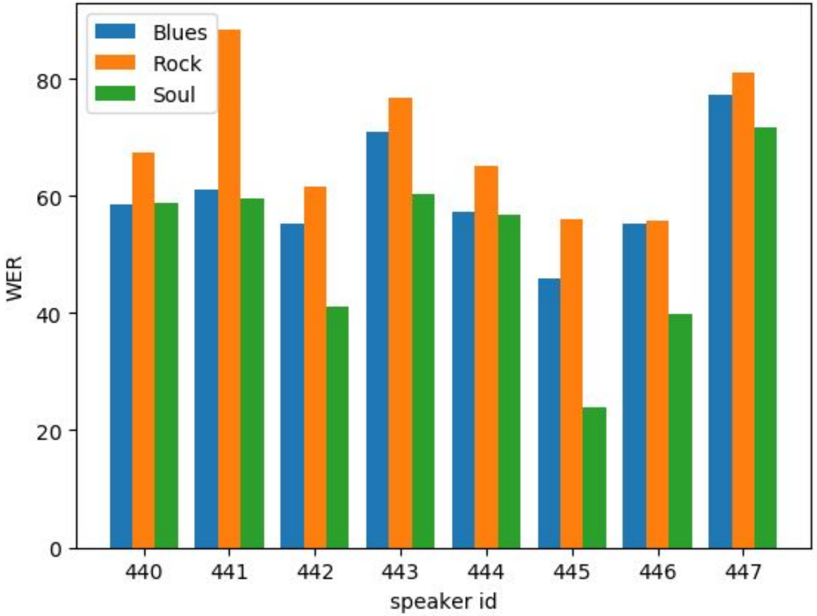
If we could fit t

Would be nice, but not necessary

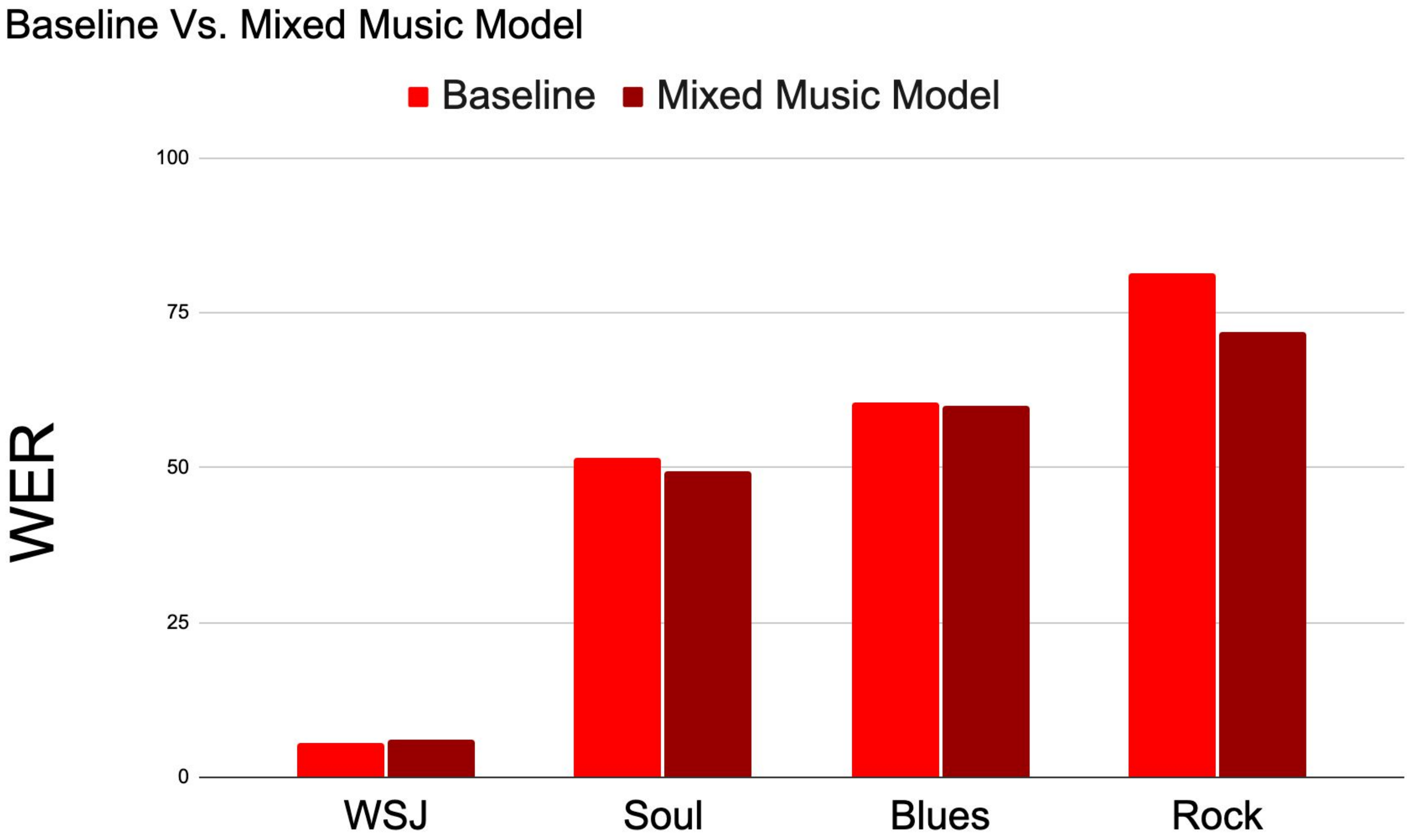


# Planning - Results





# Planning - Results



# Planning - Results

Experiment 2:

This model trained on a dataset containing full music mixtures, 100 different songs mixed with the WSJ dataset, shows an improvement in the WER scores of the specific genres but no improvement in the case of the original WSJ dataset.

# Planning - Conclusion

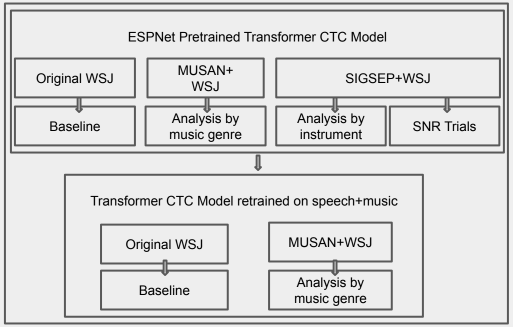
- From our analysis we can see a significant improvement on the mixture music dataset using the model trained with the full mixture music dataset.
- We have performed an in-depth analysis of the effects of different instruments, different SNRs, and different genres of music in an ASR system to aid the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

# Planning - Future Work

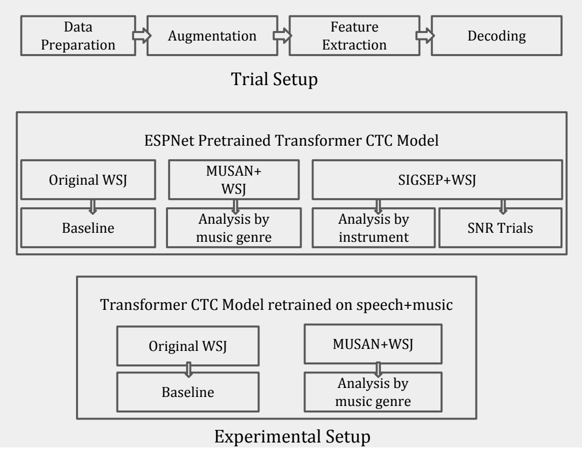
- In our analyses, the considered musical instruments were bass, drums, and vocals. Future studies with additional resources can expand upon this work and learn the task's performance with other instruments such as flute, trombone, piano, guitar, etc.
- In our analyses, the considered genres were rock, soul, and blues. Future studies with additional resources can expand upon this work and learn the task's performance with other genres such as jazz, pop, opera, theatre, classical, etc.
- An in-depth analysis of how various categories of background *noise* impacts music-speech separation and speech recognition
- Permutation invariant training using the WSJ2mix dataset (given the ground truth lyrics from the music)
- Attention models in time
- Attention models in frequency

# Planning - Future Work

- In our analyses, the considered musical instruments were bass, drums, and vocals. Future studies with additional resources can expand upon this work and learn the task's performance with other instruments such as flute, trombone, piano, guitar, etc.
- In our analyses, the considered genres were rock, soul, and blues. Future studies with additional resources can expand upon this work and learn the task's performance with other genres such as jazz, pop, opera, theatre, classical, etc.
- An in-depth analysis of how various categories of background *noise* impacts music-speech separation and speech recognition
- Permutation invariant training using the WSJ2mix dataset (given the ground truth lyrics from the music)
- Attention models in time
- Attention models in frequency



Experimental Setup


Data Preparation
Augmentation
Feature Extraction
Decoding
Trial Setup
ESPNet Pretrained Transformer CTC Model
Original WSJ
MUSAN+ WSJ
SIGSEP+WSJ
Baseline
Analysis by music genre
Analysis by instrument
SNR Trials
Transformer CTC Model retrained on speech+music
Original WSJ
MUSAN+WSJ
Baseline
Analysis by music genre
Experimental Setup

# Planning - Datasets

| Dataset | |
|---|---|
| WSJ | ● primary speech signal<br>● 70 hours of speech, between male and female speakers |
| SIGSEP MUSDB18 | ● Single-layer & multi-layer Music<br>● Music Layers: Bass, Drum, and Vocal signal |
| MUSAN | ● Full Music Mixtures<br>● Genres: Rock, Blues, and Soul |