Blind audio source separation is an active area of research. In this problem, the input to the model is a sound wave with sounds produced from multiple sources. The paper suggests a novel ontology based weighted ensemble architecture to improve the accuracy of current models. We created a sample such architecture consisting of a Transformer to identify ontological categories of sound sources followed by an ensemble of Wave-U-Net Models. The final result is taken after weighing the results based on the predicted categories. In addition, the paper introduces a novel dataset consisting of 53 categories. The new model is tested on the new datasets and better training and validation curves are obtained as compared to a single Wave-U-Net model. In addition, while this is not discussed in the paper, the architecture allows for training a model by intermixing difference datasets: only models specialized for the respective sounds can be trained on the new datasets. Hence, even if the resulting dataset is imbalanced, this does not adversely effect the training procedure.

CCS Concepts: • Computing methodologies → Language resources.

Additional Key Words and Phrases: blind audio source separation, neural networks, audio, transformer, sound classification

1 INTRODUCTION

Blind signal separation (BSS) is the separation of a set of source signals from a set of mixed signals, without the aid of information (or with very little information) about the source signals or the mixing process. [1] Due to the inherent difficulty of the problem, the accuracy of previous models remain a problem. This paper suggests a novel end-to-end ontology based weighted ensemble model. The first part of the architecture identifies the different categories present in the sound, and the second consists of an ensemble of models used to separate the mixture into its source components. The result obtained from the classification model is used to determine the weights assigned to the output of the separation models in the ensemble, hence training some models more than others for the same sound source.

We provide a sample implementation of the suggested architecture using a Transformer model [6] for the classification phase followed by a ensemble of Wave-U-Net models [3] for source separation. In particular, our transformer model identifies the presence of 6 main categories(Human Sounds, Animal Sounds, Music, Source-Ambiguous Sounds, Sounds of things and Natural Sounds ). The result of this is then is forwarded to an ensemble of 6 Wave-U-Net models, each of which tries to separate the mixture sound into the component sources. This allows different models in the ensemble to specialize on different kinds of sounds, instead of expecting each model to work on each category equally well.

We tested this model on a novel dataset that consists of atomic sounds from 53 categories, and observed a substantial improvement over a singular Wave-U-Net model tested over the same hyperparameters.

To summarize, the main contributions of this paper are the following:

(1) Introduction of a novel dataset based on manual annotations on the Google AudioSet
(2) Introduction of a novel ontology based weighted ensemble that is

The paper first discusses previous work done in blind source separation(Section 2), introduces the new dataset(Section 3), and then proceeds to explain the model architecture(Section 4). The last three sections(Sections 5, 6, 7) discuss the initial results obtained after experimentation on the new dataset. The results show that the model is more stable while training and testing in the earlier epochs as compared to a single model. However, over the long run no substantial improvement is observed in the overall the loss values. Nevertheless, the suggested model can be very beneficial to blind audio source separation since the architecture provides better
flexibility and allows effortless combination of different datasets.

2 RELATED WORK

Current methods for audio source separation can be divided into two main categories:

(1) Operating on short-term Fourier Transforms of the inputs
(2) Working directly on the sound wave

Each method has advantages and disadvantages. Spectral representation has been used in various previous researches[2], [5]. Working on STFT allows the model to directly access the magnitude and the phase information of the sound. However, a huge disadvantage of this approach is the dependence of the model's accuracy on the size and overlap of audio frames, making the results largely dependent on properly tuning these parameters in addition to the hyperparameters of the machine learning model. Usually, these parameters are left fixed and the focus is only on the hyperparameters of the model. Further, the source phase is often not estimated, and it is assumed to be equal to the mixture phase, which is normally not true. Wave-U-Net[3] aimed to alleviate this problem by directly working on the sound source without the transformation. The original paper, however, only works on 1 second clips, and on one type of sounds—musical instruments. This problem is substantially easier, and the performance of Wave-U-Net alone doesn’t extend to more complex datasets such as the introduced Wild-Mix that have multiple sources of sounds. The model also fails to work well on longer clips.

3 NOVEL DATASET

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Year</th>
<th>#Class</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>TIMIT</td>
<td>1986</td>
<td>1</td>
<td>CP</td>
</tr>
<tr>
<td>WSJ0</td>
<td>1993</td>
<td>1</td>
<td>CP</td>
</tr>
<tr>
<td>RWC</td>
<td>2003</td>
<td>2-4</td>
<td>Music</td>
</tr>
<tr>
<td>MASS</td>
<td>2004</td>
<td>2-4</td>
<td>Music</td>
</tr>
<tr>
<td>VCTK</td>
<td>2009</td>
<td>1</td>
<td>CP</td>
</tr>
<tr>
<td>DREAMSS</td>
<td>2013</td>
<td>2</td>
<td>Music</td>
</tr>
<tr>
<td>MIR-1k</td>
<td>2013</td>
<td>2-4</td>
<td>Music</td>
</tr>
<tr>
<td>TSP</td>
<td>2014</td>
<td>1</td>
<td>CP</td>
</tr>
<tr>
<td>DT</td>
<td>2015</td>
<td>1</td>
<td>CP</td>
</tr>
<tr>
<td>iKala</td>
<td>2015</td>
<td>2-4</td>
<td>Music</td>
</tr>
<tr>
<td>DSD100</td>
<td>2016</td>
<td>5</td>
<td>Music</td>
</tr>
<tr>
<td>VoiceBank</td>
<td>2016</td>
<td>1</td>
<td>CP</td>
</tr>
<tr>
<td>JADE</td>
<td>2017</td>
<td>1</td>
<td>CP</td>
</tr>
<tr>
<td>MUSDB18</td>
<td>2018</td>
<td>5</td>
<td>Music</td>
</tr>
<tr>
<td>AVSpeech</td>
<td>2018</td>
<td>1</td>
<td>CP</td>
</tr>
<tr>
<td>MUSDB-HQ</td>
<td>2019</td>
<td>5</td>
<td>Music</td>
</tr>
</tbody>
</table>

Table 1: Common datasets used for Blind Audio Source Separation

The Google audio-visual dataset consists of 2.1 million annotated videos with 5800 hours of audio over a total of 527 categories. Each sound class is organized in a hierarchy. All these sounds are weakly annotated and each class is given an approximate ‘quality score’. A quality score of x% implies a x% of the videos consist of mentioned sound in video. However, most sounds are overlapping, making the raw data unusable for blind source separation.

A new dataset called WildMix[7] was introduced by the MultiComp Lab that originally consisted of 25 categories. This dataset has now been further expanded to include more categories. The new dataset consists of 100 sounds each from 53 categories that were manually annotated with the start and the end of an atomic sound (absence of any other sound). This is one of the largest such dataset created yet and contains a wide variety of sounds. Hence, the models trained on this are expected to be more general in contrast to models trained only on a few categories previously, thus making it more useful in real world applications.

Since only atomic sounds are used, the comparison of the ground truths with the output of the model are accurate. An average of 10-15% of sounds in AudioSet yielded such values (i.e had atomic sounds for more that 1s). All the work conducted in this paper uses this dataset.

4 SAMPLE MODEL ARCHITECTURE(TRANSFORMER + WAVE-U-NET)

The proposed novel hybrid model consists of two phases.

4.1 Transformer Model

4.1.1 Background. Introduced in the paper, Attention is All You Need[6], the transformer model is largely regarded as the state of the art for working on inputs represented as sequences. The model tries to identify ‘important’ places that help the model perform the task at hand, without the presence of a recurrent neural network. The Fig 1 describes the multiheaded attention layer that helps the model encode the input.

4.2 Details of Implementation

In this paper, the transformer model is used to solve the multilabel ontological classification problem. The model is supposed to identify
the various top level hierarchies present in the sound, namely:

- Human Sound
- Source-ambiguous sounds
- Animal
- Sounds of things
- Music
- Natural sounds

These are based on the 6 top hierarchies present in the AudioSet. Top level is favored over lower levels to allow for scalability across multiple datasets created from WildMix, and providing flexibility when additional categories are annotated. Further, this is not expected to adversely affect the accuracy of the overall model since the models will still be specializing over similar sounds. Moreover, different models for lower levels would not be possible due to memory constraints.

1. A short term fourier transform is used on the mixture sound (None of the disadvantages of STFT mentioned in the related work are expected to be harmful for classification as the mixture phase information should suffice for the classification problem)
2. An Multi-Headed attention layer is used to create an encoding of the STFT input
3. A fully connected layer followed by Sigmoid layer is used to get predicted classes. The use of Sigmoid Layer allows the model to predict multiple labels since sounds from multiple classes are mixed together.
4. The outputs from all encoded frames are averaged to obtain the final output

4.3 Wave-U-Net Ensemble

4.3.1 Background. The Wave-U-Net model [3] works directly on the time domain audio signal as input. The basic outline of a single wave-u-net model is outlined in Figure 2.

4.3.2 Details of Implementation of a single model. Each downsampling block and upsampling block consists of a Conv1D layer followed by a leaky relu activation. The mixture audio is passed through a series of 'Downsampling blocks' whose output is sent to both the next downsampling block and a 'Upsampling' block. Thus, the input to each 'Upsampling' block consists of the input from the previous upsampling block and the output of a corresponding 'Downsampling block'. In essence, the downsampling blocks try to encode the source by iteratively taking coarser time frames, and the upsampling blocks try to decode the input iteratively working on both a encoding and a decoding of the same size. Finally, the encoded result and the original model input are concatenated and passed into the final output layer that produces the source sounds.

4.3.3 Ensemble. The ensemble model consists of 6 distinct wave-u-net models built as described above. Each model returns the source separated results. In addition to the sound input, the model also receives the predicted ontological categories as an input. A function $\sigma$ is applied to the ontological categories to get the respective weights of each model for that particular input. A weighted average of the concatenated outputs of the 6 models is then taken to receive the final output. The function $\sigma$ is regarded as a hyperparameter and experiments using different alternatives are conducted. The details
about the various functions tested are outlined in Section 5. The complete architecture is shown in Fig 3.

Fig. 3. Sample Suggested Model Architecture.
Symbol \( W_i \) is used to symbolize the \( i^{th} \) wave-u-net model in the ensemble.

5 EXPERIMENTS
The atomic sounds in WildMix are artificially mixed in the following manner:

1. The sample rate of the sound is converted from 44100 Hz to 16000 Hz.
2. \( X \) random categories are selected from the dataset.
3. \( Y \) sounds from \( X \) categories are mixed together by randomly selecting the start point and the duration of each sound, producing each clip for a total length of 4 seconds.

The initial experiments were conducted on \( X = 10, Y = 2 \). A total of 20000 such samples are created for the train data and 2000 are used for the validation dataset.

The same dataset is used for both the classification and segregation part of the model. In the next two sections we discuss the experiments conducted on the the two parts separately.

5.1 Classification Model using Transformer

5.1.1 Training Procedure. A frame size of 256 and a hop length of 192 is used while creation of the short term fourier transforms of the sound. Cross Entropy Loss is used as the loss function. We use the ADAM Optimizer with a learning rate of \( 3e^{-4} \) and decay rates \( \beta_1 = 0.9, \beta_2 = 0.999 \). The batch size, number of heads, number of layers and the output size of the encoder model are regarded as hyperparameters. Early stopping after 20 epochs of no improvement is deployed. The best model is then selected and then experiments are conducted to select a threshold \((t)\). The one hot encoding \((L)\) is then created from the model output \('O'\) as follows:

\[
L_i = \begin{cases} 
1 & \text{if } O_i > t \\
0 & \text{otherwise}
\end{cases}
\]

5.1.2 Wave-U-Net Model. Experiments are first conducted on a single Wave-U-Net model to obtain good hyper parameters. CSA Loss function is after the model output is converted to a Spectrogram representation, comparing it to the ground truth spectrogram representation. [4] is used as the loss function. The loss function utilizes phase information of the desired signal to improve the separation performance.

We use the ADAM Optimizer with a learning rate of \( 3e^{-4} \) and decay rates \( \beta_1 = 0.9, \beta_2 = 0.999 \). Early stopping after 20 epochs of no improvement is deployed. The model with the best validation loss is chosen.

After the selection of hyperparameters, we proceed to training the ensemble. Four distinct experiments are conducted to select the best way to use the function that converts the predicted one-hot-embeddings to weights:

- Using a linear layer, using ontological classes as input.
- Using the outputs of the transformer model directly
- Using threshold to predict the classes and averaging the results of the respective models
- Using threshold to predict the classes and using a Sigmoid layer on this. The output is the average output of only models corresponding to the classes present

6 RESULTS

6.1 Classification Model
After deploying grid search, the best transformer model hyperparameters were found to be:

- Batch size: 128
- Size of Output of encoder: 128
- Number of Heads: 4
- Number of Layers: 8

The following notation is used below:

- \( TP_x \): True Positives
- \( FP_x \): False Positives
- \( FN_x \): False Negatives
- \( TN_x \): True Negatives

Precision = \( \frac{TP}{TP + FP} \)
Recall = \frac{TP}{TP+FN}
F1 = 2\frac{Recall \times Precision}{Recall + Precision}
Acc = \frac{TP+TN}{TP+TN+FP+FN}

where the values used here are the respective sums of the values over all categories.

<table>
<thead>
<tr>
<th>Threshold</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
<th>Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.3</td>
<td>0.668</td>
<td>0.839</td>
<td>0.743</td>
<td>0.742</td>
</tr>
<tr>
<td>0.4</td>
<td>0.716</td>
<td>0.769</td>
<td>0.741</td>
<td>0.761</td>
</tr>
<tr>
<td>0.5</td>
<td>0.762</td>
<td>0.697</td>
<td>0.727</td>
<td>0.767</td>
</tr>
</tbody>
</table>

Table 1: Results of classification model on validation dataset

Threshold values [0.3, 0.4, 0.5] are tested for F1 scores, precision and recall. Table 1 compares the precision, recall, f1 and accuracy values of the best classification model using the different threshold values. The accuracy under all thresholds is good and close to each other (76%). Hence, we based our decision on the other values which vary significantly between the thresholds. The best precision is obtained at a threshold of 0.5, but at the value the recall is lower. After consideration, the threshold value of 0.4 is chosen despite having a lower precision as compared to 0.5 for the since the recall is much higher here. In this case, false positives are better than false negatives: This is because over-training some models is better than under-training a model that should specialize in the required category.

At a threshold of 0.4, it was observed that most classes had a high precision and recall (above 70%). The animal class seemed to have the lowest precision and recall value of 60%. This is probably because of the varied nature of animal sounds, for eg. the sound of bird is very different from dog barking; arguably more similar to a bell as compared to a dog. NOTE: the validation dataset did not have instances of sound of things category. However, to maintain consistency across different datasets, we are still using the 6 categories as advertised.

6.2 Wave-U-Net Ensemble

As mentioned in the section above, different ontology based weighing mechanisms were used. The following results were obtained:

- Using a linear layer, using ontological classes as input:
  This leads to a problem similar to mode collapse, such that the model after some iterations gives a very high weight to a single model. This is because once a model starts performing well, the model continues to train it, amplyfying the problem.

- Using the outputs of the transformer model directly:
  It was expected that using a sigmoid layr on top of outputs directly might help with the accuracy since the model would then also have information about the confidence of the desired outputs. However, after experimentation it was observed that thee results were very similar to a single wave-u-net model.

A probable explanation for this behavior is that due to the large number of models(6), such a input would divide the weights almost equally, hence, the information about the classes present is not properly exploited.

- Using threshold to predict the classes and averaging the results of the respective models:
  This led to a large variability in the validation loss. Overall, the model performed worse than a single wave-u-net model. A few possible reasons for this include: over dependence on the outputs of the previous model, and lack of training examples of a similar kind.

- Using threshold to predict the classes and using a Sigmoid layer on this. The output is the average output of only the classes present
  This led to substantially better results as compared to a single Wave-U-Net model in the earlier epochs. In this situation, all models are trained to some degree for each input. However, the models corresponding to the categories present a sample are trained more as compared to other models

After the initial experimentation, the last method was deemed to be the most fit for the current situation. Further analysis of this model was done to get a better understanding of the impacts of using a ontology based weighing mechanism(A sample such training and validation curve is displayed in Fig 4)

- A single Wave-U-Net model displays a lot more variance in both the training and the validation curves. Further it converges much later for the same hyperparameters
  In a single Wave-U-Net Model the same model is supposed to separate say a piano and a trumpet, and an engine and a speaker. However, there are substantial inherent differences in these sounds. As a result, it is expected that the model would display lack of stability especially in the beginning few epochs. In contrast, the suggested architecture forces different models to specialize in different kinds of sounds.

- The converged points are very close to each other.
  Unfortunately, no significant difference in the loss was obtained in the long term.

7 DISCUSSION AND CONCLUSION

The main purpose of the project was to observe the addition of information about categories improves the performance of models. While the primary goal of the project is to improve the current state of the art, this paper only discusses the effects of the specified augmentation on the model accuracy. Hence, less focus was placed on the absolute values of the loss function and more focus is placed on the overall trends.

While the results were not as good as hoped, the performance in the earlier epochs using the last weighing approach seems to be promising. In addition, another benefit of this approach is the fact that different datasets can be combined together and only the respective models can be trained. Future work can be done to see
(a) Training Curve

(b) Validation Curve

Fig. 4. Sample Comparison between a single Wave-U-Net and the suggested weighted ensemble model. The Batch size used is 8, number of levels in each wave-u-net is 6, and a learning rate of 3e-4 is used.

how this improves the model. A major reason for no substantial improvements in the accuracy in the long term could be the lack of data of similar kind. Since many datasets are available that have specific kinds of sounds, the model provides an ideal framework to combine the datasets together. Sound from different datasets can be mixed with each other, and passed through our model without much consideration about equitable split. This allows greater flexibility in training since we do not need to worry about the relative splits of data as only the respective models are trained, not effecting the accuracy of the other models in the ensemble. In contrast, it is expected that if a specific kind of sound dominates a dataset, any other model would likely only specialize on that sound category. Further effects of adding this augmentation on other types of models will be conducted to see if the same trends are observed in other models.

REFERENCES


