MODEL SELECTION FOR DEEP AUDIO SOURCE SEPARATION VIA CLUSTERING ANALYSIS

Alisa Liu, Prem Seetharaman, Bryan Pardo

Northwestern University
Computer Science Department
Evanston, IL

ABSTRACT

Audio source separation is the process of separating a mixture into isolated sounds from individual sources. Deep learning models are the state-of-the-art in source separation, given that the mixture to be separated is similar to the mixtures the deep model was trained on. This requires the end user to know enough about each model’s training to select the correct model for a given audio mixture. In this work, we propose a confidence measure that can be broadly applied to any clustering-based separation model. The proposed confidence measure does not require ground truth to estimate the quality of a separated source. We use our confidence measure to automate selection of the appropriate deep clustering model for an audio mixture. Results show that our confidence measure can reliably select the highest-performing model for an audio mixture without knowledge of the domain the audio mixture came from, enabling automatic selection of deep models.

Index Terms—source separation, deep learning, performance prediction, ensemble methods, deep clustering

1. INTRODUCTION

Audio source separation is the process of separating a mixture into isolated sounds from individual sources. It enables tasks where it would be valuable to attend to or manipulate individual sounds in a mixture. Examples include automatic speech recognition with multiple speakers, music manipulation, and content-based audio search.

Deep learning models are state-of-the-art for source separation, given that the mixture to be separated is similar to the mixtures the model was trained on [1]. Unfortunately, a deep separation model trained on one domain does not generalize well to others. For example, using a model trained on music examples to isolate a single voice from a recording of multiple concurrent voices will not produce usable results. Thus, the end user must know enough about each model’s training to select the appropriate model for a given audio mixture.

This limits how models can be deployed and imposes a bottleneck on adding source separation to an automation chain. A method to automatically select the best model for the current mixture would transform the range of applications where source separation can be applied. For example, imagine a hearing aid that automatically switches models when the user moves from an outdoor construction site (a model trained to separate speech from environmental noise) to an indoor restaurant (a model trained to separate speech from speech). Moreover, simply knowing when no source separation model available for the current audio scene is useful. When source separation would produce inappropriate output that might be worse than not applying separation at all, a device could simply turn off source separation.

In this work, we develop a confidence measure that can be applied to systems that perform clustering on embedding spaces to do source separation, as embodied by the family of deep clustering [2] methods. These have been among the most successful source separation approaches in recent years [3, 4]. We use this confidence measure to automatically select the model output with the best predicted separation quality. A system overview is shown in Figure 1.
Figure 2: 2D visualization of the embedding of a single music mixture produced by three different models. Visually, the embedding of the music model is far more clusterable than the speech and environmental model. Therefore the mixture may be easier to separate with the music model.

2. PRIOR WORK

Kim et al. [5] produces a selection method for speech enhancement by training an ensemble of autoencoders and selecting the model with the lowest reconstruction error. While speech enhancement is related to source separation, the underlying assumption is that the desired signal constitutes the majority of the audio. Since, source separation requires removal of large portions of the input audio, low reconstruction error does not correspond to good source separation.

In computer vision, failure detection has employed a measure of confidence that is learned during training [6] or by comparing the the input image to the training data [7, 8]. Our work is for audio, does not require a specific model for error estimation, and does not need access to the training distribution. The clusterability of audio representations was considered in [9], where clusterability was computed when the information about true label assignments is already known. Here we use unsupervised clusterability measures where the label assignments are estimated.

There has been work in combining multiple algorithms for source separation [10, 11, 12]. In particular, [13] trained a separate deep net to estimate source separation quality in order to guide switching between separation algorithms. In contrast, our confidence measure does not require any training. Moreover, all of these ensemble methods were applied to multiple methods that work for one application domain (e.g. two separation approaches, both applied to speech separation) whereas our method chooses which model is the most appropriate for a given mixture across multiple domains. Our confidence measure builds on the one in [14], which was designed to predict the performance of a direction-of-arrival algorithm for source separation. In that work, the confidence measure was designed particularly for speech mixtures and required the clustering to be based on Gaussian Mixture Models. The confidence measure here can be applied anywhere that embedding spaces are clustered. In this paper we apply it to an ensemble of deep clustering networks.

3. PROPOSED METHOD

3.1. Deep clustering

In deep clustering, a neural network is trained to map each time-frequency bin in a magnitude spectrogram of an audio mixture to a higher-dimensional embedding, such that bins that primarily contain energy from the same source are near each other and bins whose energy primarily come from different sound sources are far from each other. Given a good mapping, the assignment of bin to source can then be determined by a simple clustering method, such as K-Means. All members of the same cluster are assigned to the same source. Because deep clustering performs separation via clustering, our method relies on an analysis of the embedding space produced by deep clustering to establish a confidence measure.

3.2. Confidence measure

The core insight behind the confidence measure is that the distribution of the embedded time-frequency points of a mixture is predictive of the performance of the algorithm. Figure 2 shows a visualization of the confidence measure as applied to the distribution of points in a mixture produced by three trained deep clustering networks, each trained on a different domain. The input is a music mixture. The speech (left) and environmental (right) models return distributions with no clear clusters. The music model (middle) shows a more clusterable distribution, which is reflected by a higher confidence
score. The confidence measure $C(X)$ combines the silhouette score $S(X)$ and posterior strength $P(X)$ through multiplication so that it is high only when both are high. That is, $C(X) = S(X) \cdot P(X)$.

### 3.2.1. Silhouette score

The silhouette score [15] captures how far apart clusters are (intercluster distance) and how dense they are (intracluster distance). Define $X$ as the set of embeddings for every time-frequency point in an audio mixture, where $x_i$ is the embedding of one point. $X$ is partitioned into $K$ clusters $C_k$: $X = \bigcup_{k=1}^{K} C_k$. Given a data point $x_i$ assigned to cluster $C_k$:

$$a(x_i) = \frac{1}{|C_k| - 1} \sum_{j \in C_k, j \neq i} d(x_i, x_j)$$

$$b(x_i) = \min_{\ell \neq k} \frac{1}{|C_\ell|} \sum_{j \in C_\ell} d(x_i, x_j)$$

Here, the intracluster distance $a(x_i)$ is the mean distance (using a distance function $d$) between $x_i$ and all other points in $C_k$, and the intercluster distance $b(x_i)$ is the mean distance between $x_i$ and all the points in the nearest cluster $C_\ell$. Compute the silhouette score of $x_i$

$$s(x_i) = \frac{b(x_i) - a(x_i)}{\max\{a(x_i), b(x_i)\}}$$

Note $s(x_i)$ ranges from $-1$ to $1$. In the silhouette score, every point in every cluster must be compared to every point in every other cluster. Since a time-frequency representation can easily contain upwards of a million points, computing the silhouette score for every point in embedding space is intractable. Therefore, we select a constant number of points ($N = 1000$) from the loudest 1% of all time-frequency bins. This is because the assignment of louder time-frequency bins is more important perceptually than softer time-frequency bins. We take the mean silhouette score across the sample to estimate the silhouette score $S(X)$ for the mixture.

### 3.2.2. Posterior strength

For every point $x_i$ in a dataset $X$, the clustering algorithm, soft K-Means, produces $\gamma_{ik} \in [0, 1]$, which indicates the membership of the point $x_i$ in some cluster $C_k$, also called the posterior of the point $x_i$ in regards to the cluster $C_k$. The closer that $\gamma_{ik}$ is to 0 (not in the cluster) or 1 (in the cluster), the more sure the assignment of that point. We compute the posterior strength of $x_i$ as follows:

$$P(x_i) = \frac{K \left( \max_{k \in \{0,...,K\}} \gamma_{ik} \right) - 1}{K - 1}$$

The equation maps points that have a max posterior of $\frac{1}{K}$ (equal assignment to all clusters) to 0, and points that have a max posterior of 1 to 1. The overall posterior strength $P(X)$ is the mean $P(x_i)$ for the loudest 1% of time-frequency bins.

We apply this in conjunction with the silhouette score to predict separation performance across multiple domains. An implementation is available\(^1\) and integrated into nussl [16].

### 4. EXPERIMENTAL DESIGN

Our experiments are designed to investigate two facets of the proposed confidence measure. The first is whether or not the confidence measure correlates with ground truth separation performance, measured via signal-to-distortion ratio (SDR) [17]. The second is whether we can use the confidence measure to choose the appropriate deep clustering model given an audio mixture that comes from any of the three domains we consider: environmental, speech, and musical sounds.

Within each audio domain, we consider a specific task. For speech, the task was to separate two speakers talking simultaneously. For music, the task was to separate singing voice from accompaniment (drums, bass, and other instruments). For environmental sounds, the task was to separate two environmental sounds from one another (e.g. isolating a dog bark from a car horn). For each domain, we trained a source separation network with a deep clustering objective. Each network is identical, consisting of 2 BLSTM layers with 300 hidden units in both directions. The embedding size of each network is 20 with sigmoid activation and is trained for 80 epochs with the Adam optimizer [18] (learning rate of 2e-4).

We create a dataset for each domain. The speech dataset is constructed using the Wall Street Journal dataset [19], the

\(^1\)https://git.io/Jf3y5
First, we investigate whether the confidence measure correlates well to ground-truth performance in terms of SDR for a single model. Figure 3 shows a clear relationship between confidence and SDR for speech mixtures separated by the speech model. We see that both confidence and performance are a function of the mixture type: same-sex mixtures are harder to separate. This was also observed by [23], where they used domain-specific knowledge about the psychoacoustic differences between male and female speech. Here, we have uncovered that same relationship only by analyzing the clusterability of embeddings. We observe similarly strong correlations on the music (r-value of .46) and environmental (r-value of .70) domains.

Next, we compare the confidence-based ensemble approach with other approaches to separating each dataset. The top three rows of Table 2 show the performance of three ensemble approaches. Oracle, which switches with knowledge of the best system, is the upper bound. Random, which selects randomly with equal probability, is the lower bound. Confidence uses our confidence measure to select the model: in this method, all three models are run and confidence measures are computed for each separation, and the output with the highest confidence is chosen.

We observe that the confidence-based ensemble significantly outperforms the random ensemble. In the case of music mixtures, the confidence-based model achieves almost oracle performance. In Table 4, we show a confusion matrix to compare the true domain with the predicted domain. The precision of picking the best model when the best model is speech, music, and environmental sound are .97, .98, and .93, respectively. The recall rates are .72, .98, and .57.

6. CONCLUSION

We have presented a method for constructing ensembles of deep clustering models by using off-the-shelf clustering analysis techniques such as the silhouette score and posterior analysis. This confidence measure can be applied to ensembles of any clustering-based separation algorithms. Our work can be integrated into interfaces that use source separation by easing the burden on the user for selecting an appropriate model, as well as alerting them to system failure.

### Table 1: Confusion matrix comparing the true domain with the domain-specific model selected by the confidence measure. We see that the ensemble fairly reliably picks the correct model for each domain (the diagonal of the confusion matrix).

<table>
<thead>
<tr>
<th>Domain</th>
<th>True domain</th>
<th>Speech</th>
<th>Music</th>
<th>Environ.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speech</td>
<td>2186</td>
<td>25</td>
<td>31</td>
<td></td>
</tr>
<tr>
<td>Music</td>
<td>717</td>
<td>2950</td>
<td>1235</td>
<td></td>
</tr>
<tr>
<td>Environ.</td>
<td>97</td>
<td>25</td>
<td>1734</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>3000</td>
<td>3000</td>
<td>3000</td>
<td></td>
</tr>
</tbody>
</table>

### Table 2: Performance (mean SDR) of different approaches on test datasets from three domains. Higher values are better. The top three rows correspond to the different ensembles. The bottom three rows show the result of using a single model across all the domains.

<table>
<thead>
<tr>
<th>Approach</th>
<th>Speech</th>
<th>Music</th>
<th>Environ.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oracle ensemble</td>
<td>8.3</td>
<td>6.5</td>
<td>12.2</td>
</tr>
<tr>
<td>Confidence ensemble</td>
<td>7.6</td>
<td>6.4</td>
<td>10.5</td>
</tr>
<tr>
<td>Random ensemble</td>
<td>4.8</td>
<td>4.2</td>
<td>2.8</td>
</tr>
<tr>
<td>Speech model</td>
<td>8.2</td>
<td>2.0</td>
<td>3.0</td>
</tr>
<tr>
<td>Music model</td>
<td>1.4</td>
<td>6.5</td>
<td>2.5</td>
</tr>
<tr>
<td>Environ. model</td>
<td>2.1</td>
<td>1.7</td>
<td>11.9</td>
</tr>
</tbody>
</table>
7. REFERENCES


