# ADAPTING SPEECH SEPARATION TO REAL-WORLD MEETINGS USING MIXTURE INVARIANT TRAINING

*Aswin Sivaraman*[1,2,*], *Scott Wisdom*[1], *Hakan Erdogan*[1], *John R. Hershey*[1]

[1]Google Research [2]Indiana University

## ABSTRACT

The recently-proposed mixture invariant training (MixIT) is an unsupervised method for training single-channel sound separation models because it does not require ground-truth isolated reference sources. In this paper, we investigate using MixIT to adapt a separation model on real far-field overlapping reverberant and noisy speech data from the AMI Corpus. The models are tested on real AMI recordings containing overlapping speech, and are evaluated subjectively by human listeners. To objectively evaluate our models, we also devise a synthetic AMI test set. For human evaluations on real recordings, we also propose a modification of the standard MUSHRA protocol to handle imperfect reference signals, which we call MUSHIRA. Holding network architectures constant, we find that a fine-tuned semi-supervised model yields the largest SI-SNR improvement, PESQ scores, and human listening ratings across synthetic and real datasets, outperforming unadapted generalist models trained on orders of magnitude more data. Our results show that unsupervised learning through MixIT enables model adaptation on real-world unlabeled spontaneous speech recordings.

***Index Terms***— source separation, unsupervised learning, mixture invariant training, real-world audio processing

## 1. INTRODUCTION

Extracting estimates of clean speech in the presence of interference is a long-standing research problem in signal processing. This task is referred to as *speech enhancement* when the interference is non-speech, or *speech separation* when the interference can include speech. There has been tremendous progress in recent years with speaker-independent speech separation, made possible using deep learning algorithms [1, 2]. Yet, an important problem remains unsolved regarding the mismatch between training and test domains.

This mismatch problem stems in part from reliance on *supervised training* of speech separation models [3, 4, 5, 6]. Conventionally, supervised training is performed by generating synthetic mixtures out of a set of clean reference signals. The separation model then learns a mapping function between the synthetic mixtures and their composing reference signals. Creating artificial mixtures is necessary since it is generally infeasible to obtain recordings of real mixtures paired with clean references due to cross-talk between the microphones. But, artificial mixing can exacerbate domain mismatch if the generated examples are non-realistic or uncharacteristic of the target domain. Some researchers mitigate the introduced bias by manually crafting the training dataset to mimic the target domain distribution. Many axes of variance must be considered: including speaker characteristics, speaker position and motion, speech activity patterns, noise types, noise patterns, and acoustic reverberation [7, 8]. Other researchers circumvent manual engineering by synthesizing large datasets with enough variety that, it is hoped, some subset will match the target domain [9, 10].

Alternatives to supervised training aim to overcome mismatch by leveraging noisy recordings directly from the target domain, avoiding the need for reference signals. One prior study investigated test-time adaptation of speech enhancement systems using only noisy speaker-specific recordings [11]. However, their approach relied on having a large dedicated non-speech dataset, likely unobtainable within the target domain.

Other recent works propose training exclusively on unlabeled real-world recordings, which may provide a better match to the target domain's characteristics. One work is a generative model using a variational autoencoder [12], although this model was only validated on small domains such as mixtures of MNIST images and spectrograms of a few musical instrument mixtures. For more general unsupervised separation, the recently proposed *mixture invariant training* (MixIT) [13] enables discriminative training on raw mixture audio without labels or ground-truth reference signals.

One caveat to MixIT is that it involves using *mixtures of mixtures* (MoMs) as inputs, which may potentially create a form of mismatch with the target domain, where the input is a single mixture. In contrast, MoMs have more active sources on average than single mixtures, and perhaps some inconsistency in the acoustics between two different recordings. However, it is an empirical question whether the mismatch introduced by MixIT is as detrimental as the forms of mismatch that MixIT alleviates. One way to mitigate both risks is to jointly train on supervised synthetic data, which may better approximate the target domain in terms of the number of sources and consistency of the reverberation.

Previous experiments showed that MixIT performed well at adaptation to reverberation [13]. However, these experiments were conducted using synthetic data as the target domain, and were compared with supervised training data that was strongly mismatched in terms of both source activity as well as reverberation. Thus the benefit of such adaptation for real data remains to be verified.

In this paper, our goal is to train a model on a target domain for which we lack matching training data with supervised reference signals. In particular, we experiment with training a neural network-based speech separation system targeting the AMI Corpus dataset [14], where no matching supervised training data exists. To this end, we train our system using either: (1) supervised training with synthetic reverberant data, or (2) unsupervised MixIT training using AMI data (i.e. matched domain), or (3) a *semi-supervised* combination of the two. Lastly, we also investigate the benefits of pretraining the model using MixIT on AudioSet [15], a very large open-domain dataset, prior to the above configurations or in isolation.

Evaluating our models on real-world data presents a challenge: objective metrics cannot be used due to the lack of reference signals. To address this, we perform human listening tests using the real AMI

[*] Work done during an internship at Google.

978-1-6654-0540-9/22/$31.00 ©2022 IEEE

Corpus test set data. To handle the lack of perfect reference signals for the real data, we propose an extension of the MUSHRA (multiple stimuli with hidden reference and anchors) [16] protocol which we call MUSHIRA (multiple stimuli with hidden *imperfect* reference and anchors), where headset recordings containing some cross-talk are used as an imperfect reference. In order to measure objective metrics, we also construct a synthetic AMI mixture dataset, which leverages the synchronized headset and distant microphone recordings in addition to word boundary annotations. Synthetic AMI is a proxy dataset of pseudo-references, created using a linear time-invariant filter that projects audio from headset microphones to distant microphones. Our evaluations confirm that MixIT is helpful for adaptation, with the best results produced by a combination of supervised and unsupervised training.

## 2. TRAINING METHODS

One approach to supervised source separation for sources of the same or ambiguous class is permutation-invariant training (PIT) [4, 5]. Given a mixture $x \in \mathbb{R}^T$ of reference sources $\mathbf{s} \in \mathbb{R}^{M\times T}$ and separated sources $f_\theta(x) = \hat{\mathbf{s}} \in \mathbb{R}^{M\times T}$, the PIT objective is

$$\mathcal{L}_{\text{PIT}}\left(\mathbf{s}, \hat{\mathbf{s}}\right) = \min_{\mathbf{P}} \sum_{m=1}^{M} \mathcal{L}\left(s_m, [\mathbf{P}\hat{\mathbf{s}}]_m\right), \tag{1}$$

where $\mathbf{P}$ is an $M \times M$ permutation matrix and $\mathcal{L}$ is a signal-level loss function.

To train a source separation model on real-world data which lacks reference sources requires an unsupervised learning algorithm. The recently proposed mixture-invariant training (MixIT) [13] accomplishes this using a form of self-supervision. In this approach, the model inputs are mixtures of mixtures (MoMs), which are the sum of 2 reference mixtures, (i.e., $\bar{x} = x_1 + x_2$ where $x_n \in \mathbb{R}^T$). Given reference mixtures and separated sources $\hat{\mathbf{s}} = f_\theta(\bar{x})$, the MixIT loss estimates a mixing matrix $\mathbf{A} \in \mathbb{B}^{2\times M}$:

$$\mathcal{L}_{\text{MixIT}}\left(\{x_n\}, \hat{\mathbf{s}}\right) = \min_{\mathbf{A}\in\mathbb{B}^{2\times M}} \sum_{n=1}^{2} \mathcal{L}\left(x_n, [\mathbf{A}\hat{\mathbf{s}}]_n\right) \tag{2}$$

where $\mathbb{B}^{2\times M}$ is the set of $2 \times M$ binary matrices where each column sums to 1 (i.e., the set of matrices which assign each separated source $\hat{s}_m$ to one of the reference mixtures $x_n$), and $\mathcal{L}$ is a signal-level loss function between reference mixtures and their estimates. Training is thus discriminative with respect to the individual mixtures, and the individual source estimates emerge as latent variables.

In this paper, for both PIT and MixIT, we use negative thresholded SNR as the signal-level loss function:

$$\mathcal{L}(y, \hat{y}) = -10\log_{10} \frac{\|y\|^2}{\|y-\hat{y}\|^2 + \tau\|y\|^2} \tag{3}$$

where $\tau = 10^{-\text{SNR}_{\text{max}}/10}$ acts as a soft threshold that clamps the loss at $\text{SNR}_{\text{max}}$. We empirically select $\text{SNR}_{\text{max}} = 30$ dB.

## 3. EXPERIMENT SETUP

Our speech separation model is the "improved time-domain convolutional neural network" (TDCN++) [17], which estimates a fixed number of masks for a particular basis representation of the input mixture. This input representation is multiplied by the masks, and the result is inverted back to a set of time-domain separated waveforms. For the basis representation, we either use a learned basis [1] with 2.5 ms window and 1.25 ms hop, or a short-time Fourier transform (STFT) with 32 ms window and 8 ms hop. As our experiment pertains to domain adaptation through both supervised and unsupervised training, we hold this model architecture constant and vary training schemes. All models were trained with the Adam optimizer [18] with batch size 128 and learning rate $10^{-3}$ on 32 Google Cloud TPU v3 cores, with early stopping on the validation loss.

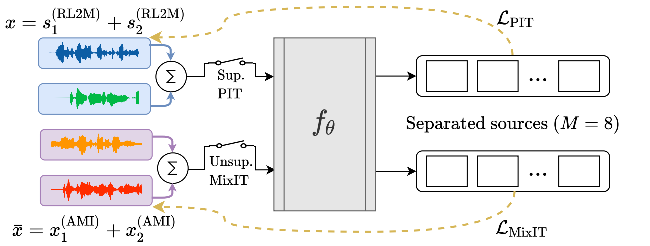
$x = s_1^{(\text{RL2M})} + s_2^{(\text{RL2M})}$
$\mathcal{L}_{\text{PIT}}$
Σ
Sup. PIT
Σ
Unsup. MixIT
$f_\theta$
...
Separated sources ($M = 8$)
...
$\bar{x} = x_1^{(\text{AMI})} + x_2^{(\text{AMI})}$
$\mathcal{L}_{\text{MixIT}}$

**Fig. 1**. Experiment configuration. We train a separation model $f_\theta$ using either one or both datasets paired with appropriate loss functions: supervised PIT with reverberant Libri2Mix (RL2M), or unsupervised MixIT with AMI. Model parameters $\theta$ can either be randomly initialized or warm-started from another model pretrained with unsupervised MixIT on AudioSet.

Our training configurations are illustrated in Figure 1. For supervised data, we use anechoic and reverberant versions of Libri2Mix [19, 20]. The anechoic version is the official clean two-speaker mixtures, and the reverberant version RLibri2Mix [13] uses synthetic impulse responses using a simulator described in previous work [20].

For unsupervised data, raw 20 s clips of AMI Corpus distant microphone audio are used. During training, random 10 s clips are sampled from these 20 s clips to increase diversity of training examples. When creating MoMs from this raw AMI Corpus data, we only combine mixtures from the same location (Edinburgh, TNO, or Idiap). If this consistency between locations is not used, we found that the separation model works poorly in practice, since it exploits the provided cue during training. As in recent work [21], we found that pretraining the separation model on AudioSet [15] was very helpful. For all experiments, we used the same checkpoints trained for about 2.7M steps using MixIT on 10 s clips of AudioSet with batch size 256 and learning rate $10^{-3}$ on 64 Google Cloud TPU v3 cores.

In terms of training dataset durations, supervised (reverberant) Libri2Mix is 212 h, unsupervised matched AMI is 34 h (Edinburgh), 23 h (TNO), and 14 h (Idiap), and unsupervised open-domain AudioSet is 5800 h. Note that the amount of AMI data is one order of magnitude less than the amount of supervised data, and two orders of magnitude less than the amount of AudioSet data.

## 4. EVALUATION

To estimate the objective performance of our methods, we use synthetic mixtures of AMI data, as well as real AMI recordings. We also propose an extension of the MUSHRA listening test to handle imperfect reference signals.

### 4.1. Synthetic Overlapping AMI

To enable performance measurement for our methods using objective measures on data that is exactly matched to AMI as much as possible, we constructed a synthetic evaluation set with isolated sources from the AMI recordings. As a first step, we used the provided transcript annotations to segment utterances based on word boundaries, which yielded 17 h of isolated non-overlapping speech.

Headset audio for these isolated speech segments is very clean, with almost no background noise. The distant microphone audio for these segments contains background noise in addition to speech. To remove this background noise and create clean reverberant spatial images of speech at the distant microphone, we estimated a linear time-invariant filter $\hat{h}$. This filter is estimated using least-squares to map the segmented headset speech to the noisy distant microphone observation. The filter is found using the following equation:

$$\hat{h} = \arg\min_{h \in \mathbb{R}^P} \|x * h - y\|^2 , \tag{4}$$

where $h$ is an $P$-point causal finite impulse response filter, $x$ is the headset microphone signal, $y$ is the distant microphone signal, and $*$ denotes convolution. The filtered headset signal is given by $x * \hat{h}$. The filter was chosen to be $200\,\mathrm{ms}$ long ($P = 3200$ at 16 kHz sampling rate), and the estimation was done in the time domain. This filtering is not perfect: if the filtered headset is subtracted from the microphone array, the residual still contains some speech content, likely due to movement and imperfect estimation. However, the filtered audio generally sounds perceptually reasonable.

Using noisy distant microphone audio and filtered headset audio, we constructed synthetic mixtures as follows. First, we generate short clips from the segments: segments shorter than $5\,\mathrm{s}$ are considered short "complete" clips, and segments longer than $5\,\mathrm{s}$ are chopped into non-overlapping $5\,\mathrm{s}$ clips, where any remainders of a segment less than $5\,\mathrm{s}$ are considered "remainder" clips. To construct an example, a $5\,\mathrm{s}$ clip is randomly chosen. The filtered headset for this clip is used as one reference source, and the residual (subtracting the filtered headset from the noisy distance microphone) is used as an imperfect noise reference that still contains traces of speech.

Next, we sample a second clip, with overlap that matches the distribution of AMI. To achieve this, we measured overlaps of two or more speakers, identified using word-level annotations, and fit a log-normal distribution on these overlaps. The overlap $\gamma$ for each clip is sampled from this distribution. A different random speaker is selected (with care to ensure an equal balance of speaker gender), and a clip is selected for this speaker that avoids unnatural breaks of words and matches sampled overlap $\gamma$, either by placing a complete clip with duration close to $\gamma \cdot 5$ in the middle, or using part of a clip with duration greater than $\gamma \cdot 5$ at the beginning or end. The filtered headset for this clip is used as a second speech reference source.

In summary, synthetic AMI examples consist of three reference sources: imperfect noise reference active for the whole clip, distant speech reference 1 (S1) active for the whole clip, and distant speech reference 2 (S2) active for the proportion $\gamma$ of the clip.

### 4.2. Real Overlapping AMI

To test our models' performance on real overlapping AMI data, we selected segments (with a minimum duration of $2.5\,\mathrm{s}$) from the AMI Corpus test set meetings wherever the corpus annotations indicated a two-speaker overlap. We took care to balance gender of the two speakers for Edinburgh (TNO test meetings only contain male speakers, and Idiap test meetings only contain female speakers). We then annotated these segments by how much cross-talk was present in each speaker's headset audio. Segments with no or minor cross-talk were selected for evaluation. This was done to obtain an imperfect reference for the target speaker (i.e., the most prominent voice or "foreground" speaker). Although some segments contained minor amounts of cross-talk in the headset recordings, overall, the speaker's relative volumes sounded roughly equal in the distant microphone audio. We manually identified 92 such examples.

### 4.3. MUSHRA and MUSHIRA

The MUSHRA listening test depends on a pristine reference signal to indicate what the target signal is and to calibrate ratings (raters are required to annotate the hidden reference as 100). Because of this, MUSHRA is unsuitable to use for real recordings where we do not have pristine ground-truth. For example, MUSHRA is unsuitable to be used for real AMI recordings, because cross-talk from other speakers is present in AMI headset audio.

To solve this problem, we propose a modified version of MUSHRA which we call MUSHIRA. MUSHIRA is a slight modification of MUSHRA where raters are not required to rate the hidden reference as 100. They are instructed to rate clips according to the most prominent speech from a single speaker (the "foreground speech"), and that the presented imperfect reference signal is one of the treatments presented. MUSHIRA is useful when a particular model can outperform the reference, e.g. if the reference is a headset with cross-talk, a perfect recovery of the foreground speech by the separation model may score higher than the reference.

For MUSHRA on synthetic AMI, we chose a subset of the full synthetic AMI set, which consists of 70 examples from each of the 3 AMI rooms. For each example, we create two items to be rated: one with speaker 1 as the reference, and one with speaker 2 as the reference. The close-talking headset is the reference, filtered headset and noisy distant mixture are anchors, and we evaluate a set of 10 models (the last 10 rows of Table 1). For MUSHIRA on real AMI, we use the headset with cross-talk as the imperfect reference, filtered headset and noisy distant mixture as anchors, and the same 10 models used for MUSHRA on synthetic AMI. For both MUSHRA and MUSHIRA, we collect 5 ratings per item.

## 5. RESULTS

In our initial experiments, we compared using the original Libri2Mix dataset [19] versus our reverberant version, RLibri2Mix, for supervised training. Using the original Libri2Mix resulted in poor performance compared to RLibri2Mix in terms of SI-SNRi on our full synthetic AMI dataset: training with Libri2Mix yielded $-2.3\,\mathrm{dB}$ without AudioSet warm-start and $-1.8\,\mathrm{dB}$ with AudioSet warm-start, and RLibri2Mix yielded $1.3\,\mathrm{dB}$ without AudioSet warm-start and $1.7\,\mathrm{dB}$ with AudioSet warm-start. Thus, we only use RLibri2Mix as supervised data in our main experiments.

Although both datasets contain reverberant speech, RLibri2Mix is mismatched from AMI in other ways. RLibri2Mix has synthetic reverberation rather than real, it consists of read rather than spontaneous speech, it contains no background noise, and the pattern of overlap between speakers is different. Our goal with using RLibri2Mix was to use a relatively standard supervised dataset and demonstrate that models trained with mismatched data can be adapted using unsupervised learning.

With more engineering effort and knowledge of the target domain, better-matching synthetic datasets could be constructed. For example, we could use our procedure from section 4.1 to construct a matched synthetic training dataset for AMI, since AMI includes detailed annotations to detect speaker overlap. A model trained on such matched supervised data would likely exceed the performance of our proposed training methods. However, such an approach is not generally feasible to apply to other domains, as gathering annotations and constructing a synthetic supervised dataset for a specific domain is unrealistic and cost-prohibitive in practice. Thus, we consider this approach to be out of scope for the present investigation, since our goal is to train a model on a target domain for which we lack matching training data with supervised reference signals.

**Table 1**. Averaged results over synthetic and real AMI datasets. “S1” refers to the full-duration speaker and “S2” refers to the overlapping speaker. For full synthetic AMI, the absolute input SI-SNRs are $0.5\,\mathrm{dB}$ for S1 and $-9.2\,\mathrm{dB}$ for S2, which are used in the SI-SNRi computation. “Warm Start” indicates pretraining the model with MixIT on $5800$ hours of AudioSet (AS) data. We denote the reference signal for each metric as (H) for headset or (FH) for headset filtered to the distant microphone. The 95% confidence interval for MUSHRA is $\pm 1.1$, and for MUSHIRA is $\pm 2.0$. MUSHRA scores with a $^*$ indicate scores from a different preliminary MUSHRA study with 95% confidence interval 1.0, with S1(H) and S2(H) scores of 64.8 and 61.6 for FH anchor, and 34.4 and 21.8 for distant mic anchor.

| **Model Configuration** | | | | **Full Synthetic AMI** | | **Subset Synthetic AMI** | | | | | | **Real AMI** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Sup. PIT | Unsup. MixIT | Warm Start | Basis | SI-SNRi S1(FH) | SI-SNRi S2(FH) | PESQ S1 (H) | PESQ S1 (FH) | PESQ S2 (H) | PESQ S2 (FH) | MUSHRA S1(H) | MUSHRA S2(H) | MUSHIRA |
| Headset (H) | | | | – | – | 4.50 | 2.18 | 4.50 | 2.56 | 100.0 | 100.0 | 89.7 |
| Headset filtered to distant mic (FH) | | | | ∞ | ∞ | 2.38 | 4.50 | 2.79 | 4.50 | 62.2 | 60.1 | 46.8 |
| Distant mic | | | | 0.0 | 0.0 | 1.69 | 1.97 | 1.62 | 1.69 | 36.0 | 25.2 | 38.0 |
| RL2M | – | – | learn | -0.1 | 2.8 | 1.68 | 1.93 | 1.69 | 1.81 | 31.0$^*$ | 19.9$^*$ | – |
| RL2M | – | – | STFT | 1.3 | 4.3 | 1.65 | 1.87 | 1.65 | 1.79 | – | – | – |
| RL2M | – | AS | learn | 1.7 | 5.7 | 1.84 | 2.21 | **1.88** | 2.10 | 41.4$^*$ | 26.8$^*$ | – |
| RL2M | – | AS | STFT | -0.6 | 2.8 | 1.73 | 1.98 | 1.76 | 1.90 | – | – | – |
| – | – | AS | learn | 4.0 | 10.6 | 1.90 | 2.31 | **1.88** | 2.06 | 47.4 | 27.4 | 42.5 |
| – | – | AS | STFT | 3.7 | 9.5 | 1.86 | 2.23 | 1.86 | 2.04 | 46.0 | 28.0 | 43.8 |
| – | AMI | – | learn | 3.9 | 10.6 | 1.80 | 2.20 | 1.81 | 2.01 | 44.5 | 27.5 | 40.5 |
| – | AMI | – | STFT | 1.0 | 5.7 | 1.60 | 1.82 | 1.61 | 1.73 | 34.3 | 22.4 | 35.5 |
| – | AMI | AS | learn | 3.8 | 9.7 | 1.85 | 2.26 | 1.79 | 1.98 | 45.5 | 27.0 | 42.1 |
| – | AMI | AS | STFT | 3.7 | 10.9 | 1.83 | 2.24 | 1.80 | 2.02 | 45.7 | 28.8 | 42.5 |
| RL2M | AMI | – | learn | 4.2 | 11.4 | 1.86 | 2.27 | 1.84 | 2.04 | 46.0 | 29.3 | 41.9 |
| RL2M | AMI | – | STFT | 2.6 | 8.2 | 1.78 | 2.11 | 1.76 | 1.93 | 42.5 | 26.5 | 41.5 |
| RL2M | AMI | AS | learn | **4.9** | **12.4** | **1.93** | **2.39** | **1.88** | **2.13** | 48.3 | 29.7 | 43.9 |
| RL2M | AMI | AS | STFT | 3.9 | 10.8 | 1.89 | 2.33 | **1.88** | 2.12 | **49.7** | **29.8** | **44.4** |

We also performed an initial evaluation with a MUSHRA listening test on a different, preliminary version of the synthetic AMI dataset. This evaluation only included models using learnable basis, with purely-supervised training on RLibri2Mix, with or without an AudioSet warm start, as well as models using unsupervised and semi-supervised training. The purely-supervised models were the lowest scoring models in terms of MUSHRA for this evaluation: $25.5$ without AudioSet warm-start, $34.1$ with AudioSet warm-start, compared to $28.1$ for distant microphone anchor, $35.6$ for the next best model (unsupervised MixIT on AMI without AudioSet warm-start), and $63.2$ for the filtered headset anchor. Thus we decided to exclude them in our final MUSHRA and MUSHIRA evaluations so that we could directly compare STFT and learnable basis.

Table 1 shows our overall results in terms of SI-SNRi on the full synthetic AMI dataset (using filtered headset as reference, which is analogous to a `bss_eval`-like filtering of the reference signal [22]), PESQ and MUSHRA scores evaluated on the subset of synthetic AMI examples used for the MUSHRA listening test, and MUSHIRA scores for real overlapping AMI examples. We measure PESQ using either the headset (H) or filtered headset (FH) as the reference signal, to match MUSHRA and SI-SNRi computations, respectively.

Note that SI-SNRi is lower for S1 compared to S2. This is because S1 has a higher average input SI-SNR of 0.5 dB, versus -9.2 dB for S2. Also, the S1 filtered-headset reference is imperfect, since the time-invariant filtering projection from headset to distant microphone is not perfect. The perceptual PESQ and listening test scores indicate that models are able to separate these filtered-headset references relatively well, although no method exceeds the performance of the filtered headset anchor.

Overall, using a warm-start from pretraining on AudioSet provides a consistent improvement across training configurations in terms of all metrics. These pretrained models exhibit surprisingly strong performance even without tuning on additional datasets. Training on AMI alone with MixIT only slightly underperforms training on open-domain AudioSet data, which is remarkable given that the AMI data is about two orders of magnitude smaller (tens of hours, versus 5800 hours of AudioSet data), highlighting the benefit of in-domain training. In future work, we intend to study the effect of relative dataset sizes on separation performance.

Generally, using a learnable basis yields higher SI-SNRi compared to STFT, but learnable basis models produced slightly lower MUSHRA and MUSHIRA scores. Informal listening identified more artifacts in learnable basis models as a possible cause.

The best training configuration is semi-supervised (PIT on reverberant Libri2Mix, MixIT on AMI), warm-starting from weights pretrained with MixIT on AudioSet. We think that these training methods are complementary: even though it is mismatched, supervised data provides examples of isolated speech that are lacking in the unsupervised AMI MoMs, while the matched AMI MoMs help adapt the model towards the target domain.

Audio demos are available online at *https://ami-mixit.github.io*, in addition to recipes for synthetic and real AMI data.

## 6. CONCLUSION

In this paper, we used MixIT to train separation models targeted towards real-world meeting data. Our best results used pretraining with MixIT on a large amount of open-domain data from AudioSet [15], followed by fine-tuning with PIT on supervised data (a reverberant version of Libri2Mix [19]) and MixIT on unsupervised data (real distant microphone recordings from AMI [14]). To estimate objective performance, we constructed a synthetic version of AMI that takes advantage of its annotations and parallel headset and distant microphone recordings. We also proposed a generalization of MUSHRA called MUSHIRA to facilitate human evaluation of source separation systems with imperfect reference signals.

We hope to extend this work by further investigating dereverberation, as well as taking advantage of multiple microphones (the AMI data has a 8-microphone circular array that is consistent across locations). Fine-tuning our models trained with MixIT on AudioSet on other downstream tasks is another interesting avenue of future work.

**Acknowledgements:** Thanks to Kevin Wilson for helpful comments. on the manuscript.

## 7. REFERENCES

[1] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, 2019.

[2] J. Chen, Q. Mao, and D. Liu, "Dual-Path Transformer Network: Direct Context-Aware Modeling for End-to-End Monaural Speech Separation," in *Proc. Interspeech*, 2020.

[3] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 12, Dec 2015.

[4] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2016.

[5] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multitalker speech separation," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2017.

[6] D. L. Wang and J. Chen, "Supervised Speech Separation Based on Deep Learning: An Overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, 2018.

[7] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, E. Habets, R. Haeb-Umbach, V. Leutnant, A. Sehr, W. Kellermann, R. Maas *et al.*, "The REVERB challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2013.

[8] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop*, 2015.

[9] M. Maciejewski, G. Sell, L. P. Garcia-Perera, S. Watanabe, and S. Khudanpur, "Building Corpora for Single-Channel Speech Separation Across Multiple Domains," *arXiv preprint arXiv:1811.02641*, 2018.

[10] E. Manilow, G. Wichern, P. Seetharaman, and J. Le Roux, "Cutting music source separation some Slakh: A dataset to study the impact of training data quality and quantity," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2019.

[11] A. Sivaraman, S. Kim, and M. Kim, "Personalized Speech Enhancement through Self-Supervised Data Augmentation and Purification," in *Proc. Interspeech*, 2021.

[12] J. Neri, R. Badeau, and P. Depalle, "Unsupervised blind source separation with variational auto-encoders," in *Proc. European Signal Processing Conference (EUSIPCO)*, 2021.

[13] S. Wisdom, E. Tzinis, H. Erdogan, R. J. Weiss, K. Wilson, and J. R. Hershey, "Unsupervised sound separation using mixture invariant training," in *Advances in Neural Information Processing Systems*, 2020.

[14] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner, "The AMI Meeting Corpus: A Pre-announcement," in *Machine Learning for Multimodal Interaction*, 2006.

[15] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio Set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2017.

[16] ITU, "Method for the subjective assessment of intermediate quality level of audio systems," *BS.1534*, 2014.

[17] I. Kavalerov, S. Wisdom, H. Erdogan, B. Patton, K. Wilson, J. Le Roux, and J. R. Hershey, "Universal Sound Separation," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2019.

[18] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. International Conference on Learning Representations (ICLR)*, 2015.

[19] J. Cosentino, M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, "LibriMix: An open-source dataset for generalizable speech separation," *arXiv preprint arXiv:2005.11262*, 2020.

[20] Z.-Q. Wang, H. Erdogan, S. Wisdom, K. Wilson, and J. R. Hershey, "Sequential Multi-Frame Neural Beamforming for Speech Separation and Enhancement," in *Proc. Spoken Language Technology Workshop (SLT)*, 2021.

[21] S. Wisdom, A. Jansen, R. J. Weiss, H. Erdogan, and J. R. Hershey, "Sparse, Efficient, and Semantic Mixture Invariant Training: Taming In-the-Wild Unsupervised Sound Separation," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2021.

[22] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, 2006.