THE POTENTIAL OF NEURAL SPEECH SYNTHESIS-BASED DATA AUGMENTATION FOR PERSONALIZED SPEECH ENHANCEMENT

Anastasia Kuznetsova, Aswin Sivaraman, Minje Kim

Indiana University, Luddy School of Informatics, Computing, and Engineering, Bloomington, IN, USA

ABSTRACT

With the advances in deep learning, speech enhancement systems benefited from large neural network architectures and achieved state-of-the-art quality. However, speaker-agnostic methods are not always desirable, both in terms of quality and their complexity, when they are to be used in a resource-constrained environment. One promising way is personalized speech enhancement (PSE), which is a smaller and easier speech enhancement problem for small models to solve, because it focuses on a particular test-time user. To achieve the personalization goal, while dealing with the typical lack of personal data, we investigate the effect of data augmentation based on neural speech synthesis (NSS). In the proposed method, we show that the quality of the NSS system's synthetic data matters, and if they are good enough the augmented dataset can be used to improve the PSE system that outperforms the speaker-agnostic baseline. The proposed PSE systems show significant complexity reduction while preserving the enhancement quality.

***Index Terms*—** personalized speech enhancement, data augmentation, speech synthesis

1. INTRODUCTION

Designing algorithms for speech enhancement (SE) is a long-standing research problem in which the current state-of-the-art methods use deep neural networks (DNNs) [1, 2, 3, 4, 5, 6]. DNN-based noise suppression algorithms typically utilize a training set prepared by artificially mixing arbitrary noise sounds with clean speech signals from many different speakers. As a result, fully-trained SE systems attempt to enhance any speech within a given input mixture. These models can generalize to the unseen test speakers if the model's computational capacity is large enough to encompass the variations found in thousands of speakers and noise types. Therefore, generalist models come at the cost of increased test-time complexity.

Recent studies have explored methods for developing target speaker extraction (TSE) models which employ a conditioning mechanism to focus on the target speaker out of a mixture of multiple talkers and noise sources [7, 8]. A straightforward strategy for targeting on a particular speaker is to condition the model with a speaker embedding [9], inferred from around 5 to 10 seconds of enrollment data (i.e., clean speech) of the target speaker [10]. Likewise, embedding-based conditioning frameworks effectively merge the tasks of speaker extraction and denoising from a noisy multi-talker mixture [7, 6], and some of them also call the task personalized speech enhancement.

On the contrary, our approach is based on the single-talker assumption, so that the model can be specialized in the talker, while reducing its complexity and improving the performance. This type of personalized speech enhancement (PSE) seeks the potential benefits of personalizing the model beyond just better performance [11]. Its goal is to reduce the model complexity while preserving the enhancement quality. It is possible because on the one hand, learning to enhance a single speaker is simpler than enhancing many speakers, allowing for less complex models. Therefore, PSE may be seen as a model compression mechanism, as more-efficient PSE models may be deployed in place of larger speaker-agnostic SE models without sacrificing performance [12, 13, 14, 11]. On the other hand, personalization is a challenging optimization task unless it gets help from the enrollment procedure that provides the target speaker's information. However, the reference signals acquired in this way come at the cost of untrustworthy recording quality and privacy concerns. Consequently, reducing the amount of the target speaker's clean speech is an additional constraint for PSE systems toward data efficiency.

In this work, we investigate a novel data augmentation strategy for PSE systems using neural speech synthesis (NSS) data. The NSS based data augmentation was previously used in ASR and other related tasks [15]. To the best of our knowledge this work is the first attempt to use synthetic speech for PSE. We consider the scenario in which a PSE model has access to a very limited amount of speech data from the target speaker, e.g., 5 seconds of clean reference speech. We investigate whether this amount is good enough for some off-the-shelf NSS systems to generate clean speech signals, while preserving the speaker's identity. Their synthesis quality must vary depending on the model's performance, and there are various ways to evaluate the perceptual quality of synthesized speech, such as naturalness, intelligibility, personality, emotions, etc. Hence, our goal is to analyze the correlation between the varying quality of synthesized speech and the PSE performance using this augmented dataset. We compare two different NSS-augmented datasets and their usability on PSE models. Furthermore, we also compare the proposed approach against non-personalized SE models as well as self-supervised PSE models [11]. Our results show that NSS dataset augmentation is useful for PSE, especially in cases where the model is too small to generalize well to the test speaker. We also observe that the quality of synthesized speech impacts PSE performance. Our findings are consistent across different model sizes, where NSS-augmented PSE models outperform speaker-agnostic models. This suggests that sub-optimal NSS synthesis is still advantageous in the context of personalizing speech enhancement systems with a simpler training framework compared to existing methods.

2. PROPOSED METHOD

The generalists: A fully-supervised framework for training SE models defines a large set $\mathbb{G}$ of clean utterances from many anonymous speakers. They are mixed with various noise signals sampled from noise corpus $\mathbb{N}$ at random signal-to-noise ratios (SNRs) to

This material is based upon work supported by the National Science Foundation under Grant No. 2046963.

A. Kuznetsova, A. Sivaraman and M. Kim, "The Potential of Neural Speech Synthesis-Based Data Augmentation for Personalized Speech Enhancement," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 2023, pp. 1-5, doi: 10.1109/ICASSP49357.2023.10096601.

Set	Subset	Duration	Quantity	Description	Corpus
$\mathbb{G}$	$\mathbb{G}_{\text{tr}}$ $\mathbb{G}_{\text{vl}}$	443 h 8h	1132 spkrs 20 spkrs	Clean speech from many anonymous speakers	LibriSpeech [16]
$\mathbb{S}^{(1:20)}$	$\mathbb{S}^{(1:20)}_{\text{tr+vl}}$ $\mathbb{S}^{(1:20)}_{\text{te}}$	5×30 sec/spkr 30 sec/spk	20 spkrs 20 spkrs	Target speakers set used for synthesis Clean speech target speaker set only used for evaluation	LibriSpeech [16]
$\mathbb{S}^{(21:30)}$	$\mathbb{S}^{(21:30)}_{\text{tr+vl}}$ $\mathbb{S}^{(21:30)}_{\text{te}}$	3 sec/spkr $\sim$8 sec/spk	10 spkrs 10 spkrs	Target speaker set used for synthesis Clean speech target speaker set only used for evaluation	AudioLM examples
$\tilde{\mathbb{S}}^{(1:20)}$	$\tilde{\mathbb{S}}^{(1:20)}_{\text{tr}}$ $\tilde{\mathbb{S}}^{(1:20)}_{\text{vl}}$	60 sec/spkr 30 sec/spkr	20 spkrs 20 spkrs	Synthesized target speaker utterances (YourTTS)	
$\tilde{\mathbb{S}}^{(21:30)}$	$\tilde{\mathbb{S}}^{(21:30)}_{\text{tr}}$ $\tilde{\mathbb{S}}^{(21:30)}_{\text{vl}}$	21 (AudioLM) or 30 (YourTTS) sec/spkr 7 (AudioLM) or 10 (YourTTS) sec/spkr	10 spkrs 10 spkrs	Synthesized target speaker utterances (using either AudioLM or YourTTS)	
$\mathbb{N}$	$\mathbb{N}_{\text{tr}}$ $\mathbb{N}_{\text{vl}}$ $\mathbb{N}_{\text{te}}$	5h 0.5h 0.5h	616 noises 60 noises 60 noises	Injection noises used during SE and PSE training Injection noises not seen during training, used to prepare speaker-specific test sets	MUSAN [17]
$\mathbb{T}$			88156 sentences	Sentences used for synthesis in both {tr, vl} partitions of $\tilde{\mathbb{S}}^{(1:20)}$ and $\tilde{\mathbb{S}}^{(21:30)}$.	VCTK [18]

Table 1. Description of the datasets used in experiments.

simulate arbitrary contaminated speech, i.e., $x = s + n$ where $s \in \mathbb{G}$ and $n \in \mathbb{N}$. The SE model is a mapping function $f(\cdot)$ with trainable parameters $\mathcal{W}_{\text{SE}}$ that aims to recover s from x, i.e., $f(x; \mathcal{W}_{\text{SE}}) = y \approx s$. Our experiments use negative signal-to-distortion ratio (SDR) [19] as the loss function for the SE system:

$$\mathcal{L}_{\text{Neg-SDR}}(\hat{v}\|v) = -10 \log_{10} \left[\frac{\sum_t (v_t)^2}{\sum_t (v_t - \hat{v}_t)^2} \right], \qquad (1)$$

where v is the clean signal and $\hat{v}$ is the estimated signal. We refer to such models as generalists and use them as a baseline.

PSE using NSS: The *transfer learning* approach can introduce a bias towards a particular speaker. As opposed to random initialization, transfer learning deems to be highly beneficial for specialists if the speaker-specific clean speech is available for finetuning, which we denote by $\mathbb{S}^{(i)}$ with a speaker index i. We assume $|\mathbb{S}^{(i)}| \ll |\mathbb{G}|$ due to any technical challenges in acquiring clean recordings or privacy concerns of the user, e.g., a few seconds. Therefore, we propose the data augmentation technique for PSE using neural speech synthesis (NSS) systems. Here, we assume that the NSS system is a text-to-speech (TTS) system that can generate any utterances that sound similar to the target speaker.

Data augmentation via NSS: Our goal in using the NSS systems is to generate as many new utterances per speaker as needed using an NSS model $g(\cdot)$ with pretrained parameters $\mathcal{W}_{\text{NSS}}$. A sentence is sampled from a large set $t \in \mathbb{T}$ fed to the NSS model as $\tilde{s} = g(t; \mathcal{W}_{\text{NSS}}|s)$, where s works as a condition to preserve target speaker's personality as $s \in \mathbb{S}^{(i)}$. Thus, we get a synthesized set of target speaker utterances $\tilde{\mathbb{S}}^{(i)}$ of any predefined size. In training phase $\tilde{s}$ is used the same way as s to construct a noisy mixture $\tilde{x} = \tilde{s} + n$. We define the task of data augmented PSE as $f(\tilde{x}; \mathcal{W}_{\text{PSE}}) = \tilde{y} \approx \tilde{s}$. We hypothesize that the quality of the speech synthesis system affects the performance of the PSE, where two main factors are involved in defining the NSS systems' performance: the general quality of the speech signal and personality. While these two concepts are not straightforward to quantify, we empirically show that the two NSS systems, in comparison, are with different performances, and they are correlated to their usefulness in the PSE task. Our experiments address the following research questions: (a) does the quality of an NSS system impact its usefulness towards PSE? (b) for each of the tested NSS systems, how much generated data is needed in order for a transfer learning-based PSE system to perform comparably to large generalist models? (c) how much lossless compression can we achieve with NSS-augmented PSE models in comparison with generalist SE models?

3. EXPERIMENTAL SETUP

In our experiments, we take the baseline speaker-agnostic generalist model proposed in [11] and finetune it with utterances synthesized by two off-the-shelf NSS models that exhibit varying performances. We repeat the finetuning process for multiple target speakers to comprehensively assess the proposed method. The speech enhancement model complexity between four preset sizes in order to investigate the merit of PSE in terms of model compression.

3.1. NSS Models

Our first NSS system is YourTTS [20], a multi-lingual multi-speaker TTS model composed of a transformer-based encoder, a normalizing flow decoder, and a HiFi-GAN vocoder. Second, we choose AudioLM [21], which is a novel auto-regressive speech synthesis system based on a language modeling (LM) approach. As opposed to YourTTS, AudioLM does not require textual data to synthesize speech, because the LM generates a word sequence on the fly. Instead of attempting to reproduce this model, whose pretrained version is unavailable, we use the small number of examples published in the authors' website[1]. Although these examples are a small dataset, their relative higher quality provides an interesting comparison point to the large quantity of YourTTS results.

3.2. Datasets

Table 1 describes all the datasets used in our experiments. The subscripts 'tr', 'vl', and 'te' denote training, validation, and test subsets, respectively. We chose to work with two different subsets of speakers $\mathbb{S}^{(1:20)}$ and $\mathbb{S}^{(21:30)}$ for better comparison. $\mathbb{S}^{(1:20)}$ contains 20 speakers from LibriSpeech's [16] *train-clean-100* subset to match the training setup in [11]. The second set $\mathbb{S}^{(21:30)}$ is based on the high-quality audio samples generated by AudioLM available online.

[1] https://google-research.github.io/seanet/audiolm/examples/

We generate the augmented version $\tilde{\mathbb{S}}^{(1:20)}_{\text{YourTTS}}$ using YourTTS. For a given speaker ID i, the system takes a 5-second-long clean reference audio of the target speaker from $s \sim \mathbb{S}^{(i)}$ and a random sentence $t \sim \mathbb{T}$, which are the input pair (s, t) to YourTTS that generates $\tilde{s}$. We repeat the process until we collect a pre-defined duration of speech for the target speaker.

Data augmentation for the second subset $\mathbb{S}^{(21:30)}$ is conducted using both NSS systems. First, as for AudioLM, we conveniently repurpose their publicly available synthesis results. The 7-second-long AudioLM utterances in $\tilde{\mathbb{S}}^{(21:30)}_{\text{AudioLM}}$ are generated from a 3 seconds of audio prompt; four such synthesized examples are available per speaker, making 28 seconds of synthesized audio. We use the 3 second propmt in training thus the total amount is 31 seconds of speech per speaker. We generate another augmented set $\tilde{\mathbb{S}}^{(21:30)}_{\text{YourTTS}}$ using the same 3-second long prompt, but this time we can synthesize as long utterances as we want because we have access to the pre-trained model. All the utterances are resampled to 16 kHz.

For noise subsets we use *sound-bible* partition of MUSAN [17] only for test-time mixtures $\mathbb{N}_{\text{te}}$, 60 noise files from *free-sound* folder are set aside for validation mixtures $\mathbb{N}_{\text{vl}}$ and the rest of the signals from *free-sound* are used for training time noisy mixtures $\mathbb{N}_{\text{tr}}$. Mixturre SNR is chosen from $[-5, 5]$ dB at random.

3.3. Implementation

All models in our experiment are based on the well-known monoaural time-domain source separation DNN, ConvTasNet (CTN) [22]. Following [11], we define four different sizes of the model: large, medium, small and tiny. With each size variant the number of channels in the bottleneck module and convolutional blocks is reduced by factor of 2. Consequently, the number of trainable parameters in each model is 138.8K, 224.1K, 437.8K, and 1M, from the tiny to the large models, respectively. We assume smaller models are more suitable for on-device speech enhancement.

The generalist models are pretrained using the Asteroid implementation of ConvTasNet [23] as in [11] and then finetuned using the proposed augmented datasets. We use Adam optimizer [24] with a learning rate of $1\text{e}{-6}$ and batch size of 8. After seeing 500 mixtures the model is validated on a fixed set of mixtures depending on the size of the training set. For $\tilde{\mathbb{S}}^{(1:20)}$ the validation set consists of 30 $\times$ 4-second long mixtures. For $\tilde{\mathbb{S}}^{(21:30)}_{\text{AudioLM}}$ and $\tilde{\mathbb{S}}^{(21:30)}_{\text{YourTTS}}$ a synthesized utterance is held out (7 and 10 seconds for AudioLM and YourTTS, respectively) to construct the validation set of 10 $\times$ 4-second long mixtures. Since AudioLM uses only the first three seconds of the original speaker's prompt, we use the rest of it for testing both AudioLM and YourTTS experiments by generating 10 mixtures each. To this end, we choose the AudioLM examples with the longest enrollment signals. This setup is used for our Experiment #2 (Sec. 3.5).

The training continues until there is no validation SDR improvement for 5000 mixtures. We report SDR, PESQ [25] and eSTOI [26] for both test and validation sets.

3.4. Experiment #1: Comparison with other PSE methods

The goal of the experiments with $\tilde{\mathbb{S}}^{(1:20)}$ is to show the benefit of synthetic data augmentation compared to speaker-agnostic models. We generate two sets using YourTTS, with length 60 or 120 seconds, in order to determine the minimal amount of synthesized data needed to achieve enhancement performance comparable to the generalist's. These experiments use the same test set as generalist and specialist models in [11], allowing for a direct comparison.

Subset	MOS (est.)	Cosine Similarity
$\tilde{\mathbb{S}}^{(1:20)}_{\text{YourTTS}}$; 60 sec.	3.78	0.80
$\tilde{\mathbb{S}}^{(1:20)}_{\text{YourTTS}}$; 120 sec.	3.81	0.80
$\tilde{\mathbb{S}}^{(21:30)}_{\text{AudioLM}}$; 21 sec.	**4.35**	**0.96**
$\tilde{\mathbb{S}}^{(21:30)}_{\text{YourTTS}}$; 30 sec.	4.18	0.87
$\tilde{\mathbb{S}}^{(21:30)}_{\text{YourTTS}}$; 60 sec.	4.12	0.88

Table 2. Quality evaluation of synthesized speech for YourTTS and AudioLM generated sets.

3.5. Experiment #2: Comparison between NSS schemes

Experiments with $\tilde{\mathbb{S}}^{(21:30)}$ aim to determine the impact of the quality of the synthesized speech as well as the amount of high-quality data needed for fine-tuning as opposed to lower-quality counterparts. To this end, we finetune the generalist with the two synthesized sets $\tilde{\mathbb{S}}^{(21:30)}_{\text{YourTTS}}$ and $\tilde{\mathbb{S}}^{(21:30)}_{\text{AudioLM}}$, resulting in 10×2 PSE models.

4. RESULTS AND DISCUSSION

4.1. NSS Systems Performance

A subjective evaluation of the synthesized utterances $\tilde{\mathbb{S}}$ is infeasible due to the difficulty in conducting a listening test on a large-scale synthesized dataset. Additionally, objective metrics, such as PESQ and STOI, cannot be applied due to the lack of clean references.

We measure the quality difference between the two NSS models using non-intrusive objective quality evaluation metrics. Instead of the mean opinion scores (MOS), we use an open-source neural network MOS estimator [27], with which we can indirectly compare the speech quality of the synthesized results. In addition, to compare the personality-preservation performance, we extract x-vectors with the neural speaker encoder developed by the SpeechBrain project [28]. Then, we compute the average cosine similarity between the synthesized utterances' x-vector and that of the ground-truth target speaker. As can be seen in Table 2, the estimated MOS scores of YourTTS samples are lower than the quality of AudioLM samples. Furthermore, the high cosine similarity indicates that speaker personality is better preserved using AudioLM.

Note that these results do not necessarily imply the overall performance of the two NSS systems in comparison, as their direct comparison is unfair due to various reasons. For example, the AudioLM's demo examples we collected from the authors' website might not correctly represent the model's overall quality. Meanwhile, the test speakers in $\mathbb{S}^{(1:20)}$ were already seen by YourTTS during its pretraining, so the results might be better than its actual performance on unseen speakers. However, if we limit the comparison to examples used in this paper, it is convincing that AudioLM's examples are better than YourTTS's. Next, we will see their influence on PSE.

4.2. Experiment #1

The results of Experiment #1 are summarized in Table 3. First, on the test set, we can see that the proposed PSE models, in general, underperform the baseline generalist model except for a few large model cases. Meanwhile, their performance on the validation set is indeed consistently better than the baseline. Since the validation set was also built based on YourTTS's synthesized utterances, the

Experiment #1	Size	SDRI (te)	SDR (te)	eSTOI (te)	PESQ (te)	SDRI (vl)	SDR (vl)	eSTOI (vl)	PESQ (vl)
Baseline Generalist, trained from $\mathbb{G}$, tested on $\mathbb{S}^{(i)}$ ($i \in \{1,\dots,20\}$)	L	9.84	10.38	0.70	**1.68**				
	M	9.74	10.25	0.69	**1.59**				
	S	8.75	9.29	0.66	1.50				
	T	7.89	8.43	0.64	1.42				
PSE Model, finetuned from $\tilde{\mathbb{S}}^{(i)}_{\text{YourTTS}}$, $\lvert\tilde{\mathbb{S}}^{(i)}_{\text{YourTTS}}\rvert$ =**60** sec. ($i \in \{1,\dots,20\}$)	L	10.28	10.79	0.70	1.65	**12.68**	12.06	**0.78**	1.96
	M	9.59	10.10	0.68	1.57	**12.16**	11.55	**0.76**	1.85
	S	8.69	9.20	0.65	1.48	**11.34**	10.73	**0.73**	1.7
	T	8.08	8.59	0.63	1.40	**10.59**	9.98	**0.70**	1.57
PSE Model, finetuned from $\tilde{\mathbb{S}}^{(i)}_{\text{YourTTS}}$ $\lvert\tilde{\mathbb{S}}^{(i)}_{\text{YourTTS}}\rvert$ =**120** sec. ($i \in \{1,\dots,20\}$)	L	10.27	10.78	0.70	1.66	12.54	**12.57**	**0.78**	**2.02**
	M	9.60	10.11	0.69	1.58	11.96	**11.99**	**0.76**	**1.88**
	S	8.80	9.32	0.66	1.49	11.07	**11.09**	**0.73**	**1.74**
	T	8.08	8.60	0.64	1.41	10.25	**10.28**	**0.70**	**1.59**
PseudoSE + DP [11], trained via SSL, tested on $\mathbb{S}^{(i)}$ ($i \in \{1,\dots,20\}$)	L	**10.40**	**10.91**	**0.72**	1.62				
	M	**10.19**	**10.70**	**0.71**	1.58				
	S	**9.88**	**10.39**	**0.70**	**1.55**				
	T	**9.40**	**9.91**	**0.68**	**1.49**				

Table 3. Experiment #1 results. Best results are indicated in bold. Durations represent the amount of synthesized data used for training.

Experiment #2	Size	SDRI (te)	SDR (te)	eSTOI (te)	PESQ (te)	SDRI (vl)	SDR (vl)	eSTOI (vl)	PESQ (vl)
Baseline Generalist, trained from $\mathbb{G}$, tested on $\mathbb{S}^{(i)}$ ($i \in \{21,\dots,30\}$)	L	10.58	10.00	**0.63**	1.55	11.13	10.59	0.67	1.48
	M	10.16	9.57	**0.62**	1.50	10.58	10.04	0.65	1.41
	S	9.52	8.93	**0.59**	1.40	10.21	9.68	0.62	1.33
	T	8.67	8.08	0.58	1.35	9.19	8.65	0.60	1.28
PSE Model, finetuned from $\tilde{\mathbb{S}}^{(i)}_{\text{AudioLM}}$ $\lvert\tilde{\mathbb{S}}^{(i)}_{\text{AudioLM}}\rvert$ =**21** sec. ($i \in \{21,\dots,30\}$)	L	**11.16**	**10.58**	**0.63**	**1.56**	12.29	**11.75**	0.67	1.52
	M	**10.75**	**10.16**	**0.62**	**1.52**	11.81	**11.27**	0.65	1.45
	S	**10.04**	**9.45**	0.58	**1.41**	11.07	**10.53**	0.62	1.36
	T	**9.32**	**8.74**	0.56	**1.36**	10.31	**9.77**	0.59	1.29
PSE Model, finetuned from $\tilde{\mathbb{S}}^{(i)}_{\text{YourTTS}}$ $\lvert\tilde{\mathbb{S}}^{(i)}_{\text{YourTTS}}\rvert$ =**30** sec. ($i \in \{21,\dots,30\}$)	L	10.58	9.96	**0.63**	1.55	**12.47**	11.69	**0.78**	**1.94**
	M	10.08	9.49	**0.62**	1.50	**11.96**	11.19	**0.77**	**1.80**
	S	9.37	8.78	0.58	1.39	**11.11**	10.33	**0.73**	**1.68**
	T	8.68	8.09	0.58	1.35	**10.34**	9.56	**0.71**	**1.58**
PSE Model, finetuned from $\tilde{\mathbb{S}}^{(i)}_{\text{YourTTS}}$ $\lvert\tilde{\mathbb{S}}^{(i)}_{\text{YourTTS}}\rvert$ =**60** sec. ($i \in \{21,\dots,30\}$)	L	10.56	9.97	**0.63**	1.55	12.06	11.67	0.75	1.67
	M	10.07	9.49	**0.62**	1.50	11.48	11.09	0.75	**1.80**
	S	9.42	8.84	0.58	1.40	10.51	10.13	0.68	1.44
	T	8.63	8.04	0.57	1.34	9.78	9.40	0.68	1.44

Table 4. Experiment #2 results. Best results are indicated in bold. Durations represent the amount of synthesized data used for training.

performance improvement on the validation set appears to signify an overfitting case. In other words, the finetuning-based personalization was done on the wrong speaker due to the mismatch between the target speaker and the synthesized speech in terms of personality. This trend does not change even if we double the size of synthesized speech to 120 seconds—the larger augmented set actually worsens the situation. For comparison, we reproduce the results of one of the self-supervised learning (SSL) methods from [11], namely pseudo speech enhancement (PseudoSE) and data purification (DP), which shows the best results even though they do not use any clean speech of the target speaker. Experiment #1 results suggest that a well-designed SSL can outperform the finetuning-based PSE if the data augmentation does not maintain the personality of the target speaker.

4.3. Experiment #2

Experiment #2's results are shown in Table 4. We see that the NSS-based data augmentation, if better NSS results from AudioLM are used, improves the PSE performance in almost every way. We note a clear gap between the PSE models depending on which augmented set they are finetuned from, i.e., $\tilde{\mathbb{S}}^{(i)}_{\text{AudioLM}}$ vs. $\tilde{\mathbb{S}}^{(i)}_{\text{YourTTS}}$. This result resonates with Table 2, where the AudioLM examples showed better speech quality and personalization performance in most metrics. In this smaller subset with ten speakers, it appears that YourTTS adapts to the target speakers better than it does to the first speaker set in Experiment #1. Yet, there is a clear gap between the two NSS systems' impact on PSE. Note that the direct comparison to the PseudoSE+DP results shown in Table 3 is impossible due to the mismatch of the two subsets.

5. CONCLUSION

Our work investigated the potential of neural speech synthesis methods for adapting speech enhancement models towards a particular speaker. We employed two off-the-shelf neural synthesis systems (YourTTS and AudioLM) to synthesize new speech utterances as if they were spoken by the target speaker. Because YourTTS and AudioLM vary in performance depending on the conditioning mechanism and their own model capacity, we assessed the output speech quality and personality-preservation of both systems using non-intrusive metrics. Our experiments demonstrate that speech synthesis quality does correlate with usefulness towards PSE. Using the best-quality synthesis dataset, we show that it is possible to implement efficient PSE systems via a simple finetuning approach. Examples are available at: `https://saige.sice.indiana.edu/research-projects/PSE_NSS`.

6. REFERENCES

[1] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An Experimental Study on Speech Enhancement Based on Deep Neural Networks," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 65–68, 2014.

[2] F. Auger and P. Flandrin, "Improving the readability of time-frequency and time-scale representations by the reassignment method," *IEEE Transactions on Signal Processing*, vol. 43, no. 5, pp. 1068–1089, May 1995.

[3] S. Pascual, A. Bonafonte, and J. Serrà, "SEGAN: Speech Enhancement Generative Adversarial Network," in *Proc. Interspeech*, 2017, pp. 3642–3646.

[4] 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, pp. 1702–1726, 2018.

[5] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie, "DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement," in *Proc. Interspeech*, 2020.

[6] S. E. Eskimez, T. Yoshioka, H. Wang, X. Wang, Z. Chen, and X. Huang, "Personalized Speech Enhancement: New Models and Comprehensive Evaluation," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2022, pp. 356–360.

[7] Q. Wang, H. Muckenhirn, K. Wilson, P. Sridhar, Z. Wu, J. R. Hershey, R. A. Saurous, R. J. Weiss, Y. Jia, and I. L. Moreno, "VoiceFilter: Targeted Voice Separation by Speaker-Conditioned Spectrogram Masking," in *Proc. Interspeech*, 2019, pp. 2728–2732.

[8] K. Žmolíková, M. Delcroix, K. Kinoshita, T. Ochiai, L. Nakatani, T.and Burget, and J. Černocký, "Speakerbeam: Speaker aware neural network for target speaker extraction in speech mixtures," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 800–814, 2019.

[9] R. Giri, S. Venkataramani, J.-M. Valin, U. Isik, and A. Krishnaswamy, "Personalized PercepNet: Real-Time, Low-Complexity Target Voice Separation and Enhancement," in *Proc. Interspeech*, 2021, pp. 1124–1128.

[10] T. Zhou, Y. Zhao, and J. Wu, "ResNeXt and Res2Net Structures for Speaker Verification," in *IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 301–307.

[11] A. Sivaraman and M. Kim, "Efficient Personalized Speech Enhancement Through Self-Supervised Learning," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1342–1356, 2022.

[12] S. Kim and M. Kim, "Test-time adaptation toward personalized speech enhancement: Zero-shot learning with knowledge distillation," in *Proc. of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2021.

[13] A. Sivaraman and M. Kim, "Sparse Mixture of Local Experts for Efficient Speech Enhancement," in *Proc. Interspeech*, 2020, pp. 4526–4530.

[14] A. Sivaraman and M. Kim, "Zero-shot personalized speech enhancement through speaker-informed model selection," in *Proc. of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2021.

[15] A. Rosenberg, Y. Zhang, B. Ramabhadran, Y. Jia, P. Moreno, Y. Wu, and Z. Wu, "Speech recognition with augmented synthesized speech," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 996–1002.

[16] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[17] D. Snyder, G. Chen, and D. Povey, "MUSAN: A Music, Speech, and Noise Corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[18] Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald, "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92)," 2019, University of Edinburgh. The Centre for Speech Technology Research (CSTR).

[19] E. Vincent, C. Fevotte, and R. Gribonval, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[20] E. Casanova, J. Weber, C. D. Shulby, A. C. Junior, E. Gölge, and M. A. Ponti, "YourTTS: Towards Zero-Shot Multi-Speaker TTS and Zero-Shot Voice Conversion for Everyone," in *International Conference on Machine Learning*. PMLR, 2022, pp. 2709–2720.

[21] Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour, "Audiolm: a language modeling approach to audio generation," *arXiv preprint arXiv:2209.03143*, 2022.

[22] 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, pp. 1256–1266, 2019.

[23] M. Pariente et al., "Asteroid: The PyTorch-Based Audio Source Separation Toolkit for Researchers," in *Proc. Interspeech*, 2020.

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

[25] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2001, vol. 2, pp. 749–752.

[26] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2010.

[27] G. Mittag, B. Naderi, A. Chehadi, and S. Möller, "NISQA: A Deep CNN-Self-Attention Model for Multidimensional Speech Quality Prediction with Crowdsourced Datasets," in *Proc. Interspeech*, 2021.

[28] Mirco Ravanelli et al., "SpeechBrain: A general-purpose speech toolkit," 2021, arXiv:2106.04624.