Task Arithmetic can Mitigate Synthetic-to-Real Gap in Automatic Speech Recognition

Hsuan Su[♡] **Hua Farn**[♡] **Fan-Yun Sun**[♢] **Shang-Tse Chen**[♡] **Hung-yi Lee**[♡]
[♡]National Taiwan University [♢]Stanford University
hsuansu.96@gmail.com

Abstract

Synthetic data is widely used in speech recognition due to the availability of text-to-speech models, which facilitate adapting models to previously unseen text domains. However, existing methods suffer in performance when they fine-tune an automatic speech recognition (ASR) model on synthetic data as they suffer from the distributional shift commonly referred to as the synthetic-to-real gap. In this paper, we find that task arithmetic is effective at mitigating this gap. Our proposed method, *SYN2REAL* task vector, shows an average improvement of 10.03% improvement in word error rate over baselines on the SLURP dataset. Additionally, we show that an average of *SYN2REAL* task vectors, when we have real speeches from multiple different domains, can further adapt the original ASR model to perform better on the target text domain.

1 Introduction

Existing automatic speech recognition (ASR) models have been found to lack generalizability towards domains unseen during training (Bartelds et al., 2023; Radford et al., 2022; Sundar et al., 2023). Existing works, when adapting an ASR model to a previously unseen domain, often rely on synthetic speech data (Su et al., 2024; Bataev et al., 2023; Joshi and Singh, 2022; Zheng et al., 2021; Yuen et al., 2023; Yang et al., 2023) due to its ease of generation and availability. However, this approach often leads to performance degradation due to acoustic mismatches such as intonations, background noise, speaker accents, and environmental sound differences between synthetic and real speech (Su et al., 2024). This distributional shift is often referred to as the synthetic-to-real gap. This paper tackles this problem, particularly when adapting an ASR model from a source domain with text and real speech data to a new target domain with only text data.

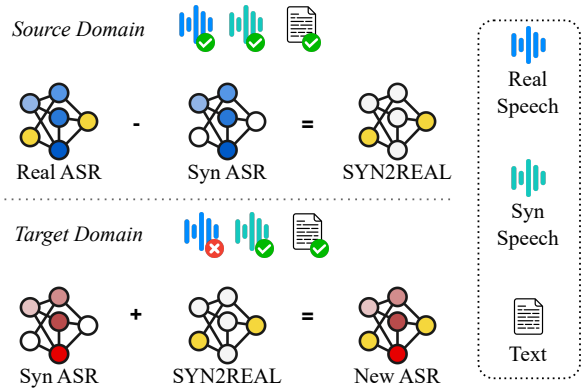

Figure 1: **Overview of the *SYN2REAL* Task Vector Approach.** The pre-trained model is fine-tuned on source domain synthetic and real speech data, separately. The difference between their parameters forms the *SYN2REAL* task vector. The *SYN2REAL* task vector is then added to a model fine-tuned on target synthetic data to overcome the synthetic-to-real gap.

Our idea is that paired synthetic speech and real speech data within a single domain can guide the adaptation of models trained on synthetic data to perform better on real-world data in a new domain. Inspired by the new paradigm of editing pre-trained neural networks by manipulating their weights (Sung et al., 2023; Tam et al., 2024), we propose to bridge the synthetic-to-real gap using task vectors (Huang et al., 2024; Bhardwaj et al., 2024). A task vector is a representation that encodes the difference between two tasks, allowing models to perform arithmetic operations to transition from one task to another. In this paper, we show that we can apply simple arithmetic operations to bridge the synthetic-to-real gap.

Taking inspiration from Ilharco et al. (2023), we propose *SYN2REAL* task vector for synthetic-to-real adaptation in ASR. Figure 1 provides an overview of the *SYN2REAL* task vector approach. The top row illustrates the process of fine-tuning models on synthetic and real speech data separately and then deriving the *SYN2REAL* task vector from

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 8905–8915

the differences in their parameters. The bottom row demonstrates the application of this vector to a model fine-tuned on synthetic target domain data, resulting in an adapted model with improved performance by incorporating the acoustic characteristics of real speech.

We conduct comprehensive experiments and ablation studies to demonstrate the effectiveness of our approach. Applying the *SYN2REAL* task vector results in an average improvement of 10.03% in word error rate (WER) for unseen target domains compared to the model before applying our method. Cosine similarity analysis of *SYN2REAL* task vectors generated by different text-to-speech (TTS) models confirms that *SYN2REAL* task vectors effectively capture domain-specific acoustic information.

2 Related Works

ASR Text-only Domain Adaptation In the context of automatic speech recognition (ASR), "text-only" domain adaptation typically refers to scenarios where the target domain only provides text data for training or fine-tuning the models. Previous work has explored internal language models adaptation that finetune language models in ene-to-end ASR models with CTC loss to improve the generalizability (Chen et al., 2023; Sato et al., 2022; Vuong et al., 2023).

The other direction is to adapt ASR models with synthetic speech. Zheng et al. (2021) develop a method that provides synthetic audio for out-of-vocabulary (OOV) words to boost recognition accuracy. Yang et al. (2023) works on personalizing ASR with synthetic speech. Bataev et al. (2023) focuses on developing a mel-spectrogram generator to improve ASR models.

Task Arithmetic The concept of task vector is introduced in Ilharco et al. (2023). Task vectors are created by subtracting the weights of a fine-tuned model from those of its corresponding pre-trained model. Different task vectors derived from the same pre-trained models can then be adjusted and combined through these simple arithmetic operations such as addition and subtraction to achieve multi-task learning (Zhang et al., 2023) and task forgetting (Daheim et al., 2023).

Recently, task vectors have shown promise in natural language processing (NLP) (Huang et al., 2024; Daheim et al., 2023; Bhardwaj et al., 2024; Zhang et al., 2023). Daheim et al. 2023 used a task vector from a negatively fine-tuned model to mitigate hallucinations. Zhang et al. (2023) proposed combining parameter-efficient fine-tuning (PEFT) modules (Hu et al., 2022; Liu et al., 2022) arithmetically. Huang et al. (2024) obtained the Chat Vector by subtracting the chat version of Llama 2 (Touvron et al., 2023) from its pre-trained version, enhancing dialogue capabilities and safety. Bhardwaj et al. (2024) introduced RESTA, adding a safety vector to re-align safety for models fine-tuned on downstream tasks. The application of task vectors is relatively underexplored in ASR. Ramesh et al. (2024) applied task arithmetic to ASR models and introduced a "task analogy" formulation, improving performance on low-resource tasks using models trained on high-resource tasks. Unlike Ramesh et al. (2024), we focus on using task vector to mitigate the distributional shift between real and synthetic data.

3 Methodology

In the context of automatic speech recognition (ASR) domain adaptation, domain mismatch can be broadly classified into three categories:

1. **Acoustic Variation Mismatch**: This mismatch refers to differences in speech caused by variations in acoustic properties.

2. **Textual Topic Mismatch**: This mismatch involves discrepancies in the subject matter or style of the textual content.

3. **Synthetic vs. Real Speech Mismatch**: This mismatch refers to the acoustic differences between synthesized speech generated from text and actual spoken speech.

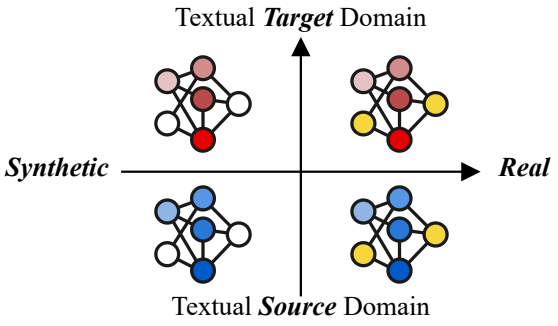

Figure 2: **Domain Shifts in ASR Domain Adaptation.** Illustration of domain adaptation challenges in ASR, showing shifts between synthetic and real speech across source and target textual domains.

In this work, we aim to adapt ASR models from source textual domains with real speech and text data to a new textual domain with only text data. We leverage data synthesized from off-the-shelf text-to-speech (TTS) systems to address this textual topic mismatch. Figure 2 illustrates the domain shifts we focused on in ASR adaptation, depicting the challenges of bridging both the textual gap (source vs. target domain) and the acoustic gap (synthetic vs. real speech).

While previous works (Yang et al., 2023; Su et al., 2024) have shown that adapting ASR models using synthetic data effectively addresses textual topic mismatch, ASR models trained on synthetic data often underperform compared to those trained on real data due to mismatches between synthetic and real speech.

To overcome this limitation, we propose the *SYN2REAL* task vector, a novel approach designed to bridge the acoustic gap between synthetic and real speech data, enhancing the performance of ASR models in domain adaptation.

3.1 Problem Formulation

We assume a problem setting in which we have two domains: a source domain D_s and a target domain D_t. The source domain D_s consists of paired text and speech samples, denoted as T_s and S_s, respectively. The target domain D_t contains only text data, denoted as T_t. This problem setting is common as it is easy to generate synthetic text data, whereas collecting paired real speech data is labor-intensive.

3.2 *SYN2REAL* Task Vector

To adapt ASR models to a previously unseen domain, we employ a common methodology (Joshi and Singh, 2022) that utilizes synthetic data generated from the target text T_t for model adaptation.

Previous work in task arithmetic has demonstrated that vectors can encode distinct capabilities, such as language or domain-specific features. We hypothesize that the differences in acoustic properties between real and synthetic speech are also learnable and can be isolated through parameter arithmetic. Specifically, we assume that we have models fine-tuned on real and synthetic data from the source domain, denoted as θ^S_{real} and θ^S_{syn} respectively. The acoustic disparity between real and synthetic speech is quantified by subtracting the parameter sets of these models:

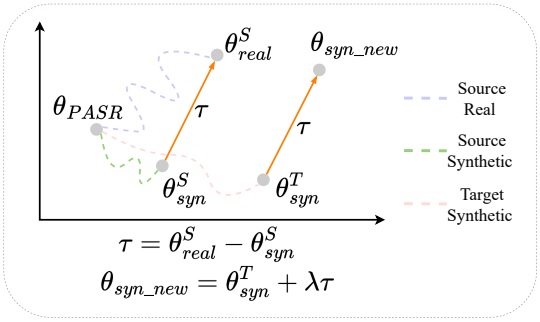

Figure 3: **Framework for *SYN2REAL* task vector in Domain Adaptation for ASR.** The framework illustrates the process of creating the *SYN2REAL* task vector by subtracting the parameter differences between a model fine-tuned on synthetic speech (Source Synthetic) and a model fine-tuned on real speech (Source Real) from pretrained ASR (PASR). This task vector is then applied to the target synthetic domain (Target Synthetic) to improve ASR performance by bridging the gap between synthetic and real speech data.

$$\tau = \theta^S_{real} - \theta^S_{syn} \tag{1}$$

Once the *SYN2REAL* vector τ is computed, we apply it to the model parameters fine-tuned on the synthetic data in the target domain, θ^T_{syn}, thereby enhancing its adaptation to the target domain:

$$\theta_{syn\_new} = \theta^T_{syn} + \lambda\tau \tag{2}$$

Where λ is the scaling factor of *SYN2REAL* task vector.

This adjusted model, θ_{syn_new}, is expected to perform more robustly in the target domain as it incorporates the acoustic characteristics of real speech, making it better suited for practical ASR tasks where real speech is present.

***SYN2REAL Ensemble* Task Vector** In the previous discussion, we assume no access to domain labels such as 'email' or 'music' from the source domain to emulate real-world situations better. All real speech data falls in the source domain. In scenarios where we have access to data of multiple domains, another approach is to create *SYN2REAL* task vectors for each domain separately and then combine these vectors. This method involves fine-tuning separate ASR models on each individual source domain to obtain domain-specific *SYN2REAL* task vectors, which are then averaged to form a comprehensive *SYN2REAL* task vector.

For each domain i in source domain S. The task vector can be defined as:

$$\tau_i = \theta^{S_i}_{real} - \theta^{S_i}_{syn} \quad (3)$$

Where $\theta^{S_i}_{real}$ and $\theta^{S_i}_{syn}$ represent the model parameters fine-tuned on real and synthetic data for domain i respectively. Once the vector τ_i is computed, we apply it to the model parameters fine-tuned on synthetic target domain data θ^{T}_{syn}, thereby enhancing its adaptation to the target domain:

$$\theta_{syn\_new} = \theta^{T}_{syn} + \frac{\lambda}{|S|} \sum_{i=0}^{|S|} \tau_i \quad (4)$$

Where $|S|$ is the number of source domains, and λ is the scaling factor for the task vector.

4 Experimental Setups

We design our experiments to answer the following questions: **Q1:** What is the efficacy of *SYN2REAL* task vector?, **Q2:** How does *SYN2REAL* task vector perform across different model sizes? **Q3:** Is *SYN2REAL* task vector effective on ASR models other than Whisper?, **Q4:** Can we form *SYN2REAL* task vectors from other TTS models?, **Q5:** What is the impact of the scaling factor λ? , **Q6:** Do *SYN2REAL* task vectors obtained with the same TTS have similar direction?

To address these questions and simulate real-world scenarios, we first create a source domain ASR model by combining synthetic and real speech data from various domains. We then adapt this source domain ASR model to the target domain using data synthesized by TTS models. *SYN2REAL* task vector is constructed by subtracting the weights of an ASR model fine-tuned on synthetic data from the weights of the same ASR model fine-tuned on real data, both using the same pre-trained model as the starting point. Our goal is to improve the performance of an ASR model on the target domain without using any real speech from the target domain.

4.1 Dataset

SLURP (Bastianelli et al., 2020) is a spoken language understanding dataset containing 16521 utterances of human commands towards a virtual agent, based on 200 pre-defined prompts such as "How would you ask for the time." The utterances are categorized into 18 domains (e.g., email, cooking, etc.). In each of our experiments, we select one of these domains as the target domain and combine the remaining 17 domains to form the source domain.

4.2 Text-to-Speech (TTS) Models

In our experiments, we prepare synthetic speech using two off-the-shelf TTS models for text from the target domains.

BARK BARK[1] is a transformer-based (Vaswani et al., 2023) autoregressive model, it is pretrained with similar architecture as AudioLM (Borsos et al., 2023) and Vall-E (Wang et al., 2023). The input of BARK contains prompts, transcription, and users. In our experiments, we do not specify the speaker for BARK.

Speech T5 Speech T5 (Ao et al., 2022) is a unified model framework that employs encoder-decoder pre-training for self-supervised speech/text representation learning. In our experiments, we randomly sample 5 speakers from 7931 speakers.

XTTS XTTS is a SOTA TTS model released by Coqui (Casanova et al., 2024). It is a multi-speaker, end-to-end TTS model capable of synthesizing production-level quality speech. In our experiments, we use its second version, XTTS-v2[2], to synthesize speech for the target domain.

4.3 ASR Models

Whisper Whisper (Radford et al., 2022) is an encoder-decoder Transformer-based (Vaswani et al., 2023) model that supervised finetuned on 680,000 hours of labeled audio data. All experiments are conducted using the Whisper small model, except for the ablation study, where we experiment with models of different sizes, including the base and tiny models, to validate our method.

Wav2Vec2-Conformer Wav2Vec2 (Baevski et al., 2020) is a framework for self-supervised learning of speech representations that masks latent representations of the raw waveform and solves a contrastive task over quantized speech representations. Wav2Vec2-Conformer (Wang et al., 2022) (referred to as Wav2Vec2 in the experiments) follows the same architecture as Wav2Vec2, but replaces the Attention-block with a Conformer-block (Wang et al., 2020) is the conformer (Gulati et al., 2020). We use the large

[1] https://github.com/suno-ai/bark
[2] https://huggingface.co/coqui/XTTS-v2

WER	Target Domains																		Average
Methods	**Alarm**	**Audio**	**Calendar**	**Cooking**	**Datetime**	**Email**	**General**	**IOT**	**Lists**	**Music**	**News**	**Play**	**QA**	**Recommendation**	**Social**	**Takeaway**	**Transport**	**Weather**	
Target Synthetic ASR (Baseline)	16.13	14.69	22.88	14.26	47.16	16.23	27.16	13.67	15.49	23.51	21.31	21.61	24.04	17.54	29.57	21.25	18.91	15.45	21.16
+ *SYN2REAL*	**15.65**	**13.68**	**22.64**	14.36	**40.29**	**16.15**	**16.87**	**12.49**	**15.22**	**17.03**	**21.25**	**20.77**	**23.88**	**15.19**	**21.87**	**18.03**	**16.90**	20.38	**19.04**
Relative WER (%)↑	**2.95%**	**6.87 %**	**1.03 %**	-0.70 %	**14.58 %**	**0.50 %**	**37.89 %**	**8.58 %**	**1.74 %**	**27.57 %**	**0.28 %**	**3.88 %**	**0.64 %**	**13.42 %**	**26.04 %**	**15.14 %**	**10.65 %**	-31.91%	**10.03%**

Table 1: **Word Error Rate (WER) Performance Across Various Target Domains.** Comparison of the baseline Whisper model and the model enhanced with the *SYN2REAL* task vector generated by BARK. The *SYN2REAL* task vector shows an average WER reduction of 10.03% across various target domains. Target Synthetic ASR refers to the baseline that is finetuned on 17 domains (excluding the target domain) real+synthetic data followed by synthetic data from the target domains in the SLURP dataset.

checkpoint[3] with 618M parameters with rotary position embeddings, pretrained and fine-tuned on 960 hours of Librispeech (Panayotov et al., 2015) on 16kHz sampled speech audio to conduct experiments.

5 Results & Discussion

Here, we discuss our results in relation to the questions we set out to answer.

5.1 What is the efficacy of *SYN2REAL* task vector?

To answer **Q1**, we apply our method by comparing the word error rate (WER) across various target domains. We select one of these domains as the target domain and combine the remaining 17 domains to form the source domain Table 1 presents the WER results for both the baseline ASR model fine-tuned on synthetic speech data and the model enhanced with the *SYN2REAL* task vector.

The baseline model, fine-tuned solely on synthetic data, exhibits varying WERs across different target domains, with an average WER of 21.16. This performance highlights the challenge of adapting ASR models to real-world data when trained on synthetic speech, primarily due to acoustic mismatches.

The application of the *SYN2REAL* task vector significantly reduces WER across most target domains. The *SYN2REAL*-enhanced model achieves an average WER of 19.04, representing an average relative WER reduction of 10.03%. This improvement demonstrates the *SYN2REAL* task vector's effectiveness in bridging the gap between synthetic and real speech data, enhancing the model's adaptability to diverse real-world scenarios.

To further validate our method's effectiveness in solving the synthetic vs. real speech mismatch, we also report two additional baselines. We conducted experiments evaluating the pretrained Whisper small model and the Whisper small model fine-tuned on the source domain (real + synthetic) on the target domain, with average WERs of 33.30 and 27.13, respectively. We address the acoustic variation mismatch by fine-tuning the pretrained Whisper on source domain speech. Further fine-tuning on target domain synthetic data addresses the textual topic mismatch, achieving a WER of 21.16. Finally, applying the *SYN2REAL* task vector mitigates the synthetic vs. real speech mismatch, resulting in a WER of 19.13.

The *SYN2REAL* task vector shows particularly notable improvements in domains such as 'Music' (27.57% reduction), 'Takeaway' (15.14% reduction), and 'Social' (26.04% reduction). These results suggest that the task vector effectively captures domain-specific acoustic variations, enabling the ASR model to generalize better to unseen target domains. In the following experiments we select the four domains includes two highest improved domains ('Music' & 'Social'), and the two lowest improved domains ('Weather' & 'Cooking') to conduct the experiments.

However, it is important to note that some domains, such as 'Cooking' and 'Weather,' exhibit marginal improvements or slight degradation in WER. These variations indicate that while the *SYN2REAL* task vector generally enhances performance, further fine-tuning and domain-specific adjustments may be necessary to optimize results across all target domains.

Overall, the results demonstrate that the *SYN2REAL* task vector is a promising approach for improving ASR domain adaptation, showing significant improvements over both the baseline and a recent work, AdaBERT-CTC (Vuong et al., 2023), which only achieved a WER of 26.1 on the SLURP test set. By addressing the acoustic mismatches between synthetic and real speech data, our method significantly enhances the performance of ASR models in real-world applications.

[3] facebook/wav2vec2-conformer-rope-large-960h-ft

Relative WER ↑	Cooking	Music	Social	Weather	Average
Tiny	**41.11%**	-13.47%	2.60%	**30.42%**	**19.48%**
Base	1.49%	**37.80%**	5.00%	6.82%	14.70%
Small	-0.70%	27.56%	**26.04%**	-31.91%	12.43%

Table 2: **Relative WER Improvement Across Different Model Sizes after applying *SYN2REAL* task vector.** This table shows the relative WER improvement compared to the Target Synthetic ASR for Whisper models of various sizes (Tiny, Base, and Small).

5.2 How does *SYN2REAL* task vector perform across different model sizes?

To answer **Q2**, we analyze the effect of model size on the performance of ASR adaptation using the *SYN2REAL* task vector. Table 2 presents the relative word error rate improvements across different model sizes (Tiny, Base, Small) and various target domains.

The results indicate that the Base model achieves the highest average relative WER improvement of 14.70% across all target domains. This model size shows substantial gains, particularly in the 'Music' (37.80%) and 'Social' (5.00%) domains, demonstrating its robustness in adapting to diverse acoustic characteristics using the *SYN2REAL* task vector.

The Tiny model, while achieving a higher average improvement of 19.48%, shows considerable performance gains in the 'Cooking' (41.11%) and 'Weather' (30.42%) domains. However, it experiences a performance degradation in the 'Music' domain (-13.47%). This suggests that while the Tiny model can benefit significantly from the *SYN2REAL* task vector in certain domains, its overall adaptability might be limited compared to larger models due to its reduced model size.

Interestingly, the Small model exhibits an average relative WER improvement of 12.43%, with substantial gains in the 'Social' (26.04%) and 'Music' (27.56%) domains. However, it shows a notable degradation in the 'Weather' domain (-31.91%), indicating potential overfitting or sensitivity to specific acoustic variations.

These results highlight the importance of model size in ASR adaptation using the *SYN2REAL* task vector. The Base model consistently provides balanced performance across most domains, suggesting it strikes a good balance between model size and performance. In contrast, the Tiny and Small models show varying degrees of effectiveness.

Overall, the analysis demonstrates that while the *SYN2REAL* task vector significantly improves ASR performance across different model sizes, the extent of improvement is influenced by the model's capacity.

5.3 Is *SYN2REAL* task vector effective on ASR models other than Whisper?

To validate the effectiveness of the *SYN2REAL* task vector on other ASR models, we conduct additional experiments using the Wav2vec2-Conformer large model.

Wav2Vec2-Conformer	Cooking	Music	Social	Weather	Average
Target Synthetic ASR (Baseline)	21.26	17.41	25.84	16.74	20.31
+ *SYN2REAL*	**18.88**	**14.33**	**21.48**	**13.36**	**17.01**
Relative WER ↑	**11.21%**	**17.66%**	**16.87%**	**20.22%**	**16.25%**

Table 3: **WER of *SYN2REAL* task vector on Wav2Vec2-Conformer.** This table shows the WER and relative WER improvement across different target domains on Wav2Vec2-Conformer model before and after applying *SYN2REAL* task vector.

Table 3 presents the WER results across various target domains, including 'Cooking', 'Music', 'Social', and 'Weather', comparing the baseline model finetuned on synthetic speech with the model enhanced by the *SYN2REAL* task vector. The Table 3 shows a significant reduction in WER when the *SYN2REAL* task vector is applied. The average WER drops from 20.31 to 17.01, representing an overall relative improvement of 16.25%.

The most notable improvement is observed in the 'social' domain, with a relative WER reduction of 16.87%. The 'Music' domain also shows a substantial improvement of 17.66%, indicating that the task vector successfully captures and mitigates the acoustic variability associated with music-related speech.

In the 'Cooking' and 'Weather' domains, the WER reductions are 11.21% and 20.22%, respectively. While the improvement in the 'Cooking' domain is more modest, it still indicates that the *SYN2REAL* task vector enhances the model's adaptability to domain-specific acoustic characteristics.

Overall, the application of the *SYN2REAL* task vector significantly enhances the performance of the Wav2vec2-Conformer large model across all tested domains. These results validate the effectiveness of the *SYN2REAL* approach in bridging the gap between synthetic and real speech data, ultimately improving the robustness and versatility of ASR systems in diverse real-world scenarios.

5.4 Can we form *SYN2REAL* task vector from other TTS models?

To answer **Q4**, we conducted experiments using the Whisper Small model with synthetic data generated by the Speech T5 model and XTTS model. Table 4 presents the WER results across various target domains, including 'Cooking', 'Music', 'Social', and 'Weather', comparing the baseline model finetuned on synthetic speech with the model enhanced by the *SYN2REAL* task vector.

TTS Model		Cooking	Music	Social	Weather	Average
Speech T5	Baseline (Synthetic ASR)	16.94	16.04	53.34	16.27	25.65
	+ *SYN2REAL*	**16.00**	**15.75**	**52.95**	**15.97**	**25.17**
	Relative WER ↑	**5.57%**	**1.77%**	**0.73%**	**1.82%**	**1.86%**
XTTS	Baseline (Synthetic ASR)	14.70	**15.89**	23.49	15.52	17.40
	+ *SYN2REAL*	**13.51**	16.37	**22.50**	**15.06**	**16.86**
	Relative WER ↑	**8.11%**	-2.98%	**4.19%**	**2.93%**	**3.10%**

Table 4: **WER on Whisper Small with *SYN2REAL* task vector using Speech T5 and XTTS models.** The table shows WER and relative WER improvement across different target domains.

The results indicate that applying the *SYN2REAL* task vector leads to a reduction in WER across most of tested domains. The average WER drops from 25.65 to 25.17 for Speech T5 and from 17.40 to 16.86 for XTTS, representing an overall relative improvement of 1.86% and 3.10% respectively.

The 'Cooking' domain shows the highest relative WER reduction of 5.57% and 8.11%, suggesting that the *SYN2REAL* task vector effectively adapts the model to this specific domain.

However, the improvement in the 'social' domain is relatively modest for Speech T5, with a relative WER reduction of only 0.73 %. This could be attributed to the high baseline WER in this domain, suggesting that the synthetic data from Speech T5 might have limitations.

Overall, the application of the *SYN2REAL* task vector to the Whisper Small model with synthetic data from different TTS models demonstrates consistent improvements in average WER, albeit with varying degrees of improvement across different domains. These results validate the flexibility and effectiveness of our approach in improving ASR models trained with synthetic data from different TTS models.

5.5 What is the impact of the scaling factor λ?

This section investigates the effect of scaling the *SYN2REAL* task vector on the WER of different ASR models. Figure 4 illustrates the WER as a function of the scaling factor λ for various ASR models and synthetic data, including Whisper Tiny with BARK, Whisper Base with BARK, Whisper Small with BARK, Whisper Small with Speech T5, and W2V2-Conformer with BARK.

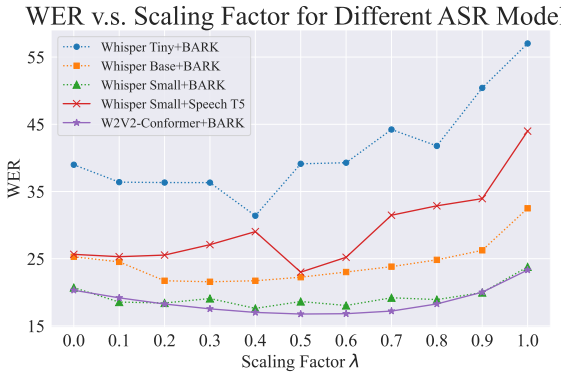

Figure 4: **WER vs. Scaling Factor across Different ASR Models & Different TTS Models** The plot shows the average WER on 'Cooking', 'Music', 'Social', and 'Weather' target domains as a function of the scaling factor λ for various ASR models (Whisper and W2V2-conformer) and the TTS models (BARK and Speech T5) to make *SYN2REAL* task vectors. We denote it as '{ASR+TTS}', such as 'Whisper Tiny+BARK' in the figure. The scaling factor adjusts the magnitude of the SYN2REAL task vector applied to each model.

The scaling factor λ adjusts the magnitude of the *SYN2REAL* task vector applied to the ASR models. We evaluated a range of scaling factors from 0.1 to 1.0 to determine the optimal balance that minimizes WER.

The results show that different models respond variably to changes in the scaling factor. For Whisper Tiny+BARK, the curve is steeper, indicating that smaller models may be more sensitive to larger adjustments from the *SYN2REAL* task vector. In contrast, Whisper Base+BARK maintains relatively stable WER values across different scaling factors, suggesting a more robust performance.

Notably, Whisper Small+BARK and Whisper Small + Speech T5 exhibit a U-shaped trend, where moderate scaling factors (around $\lambda = 0.3$ to 0.5) yield the lowest WER. This indicates that an optimal scaling factor exists for these models, which balances the incorporation of real speech characteristics without overwhelming the model with excessive parameter adjustments. The Wav2vec2-Conformer model consistently shows lower WER values across all scaling factors, with the best performance at $\lambda = 0.5$.

Overall, the analysis suggests that the optimal scaling factor λ varies depending on the ASR

model's architecture and size. While smaller models like Whisper Tiny+BARK may benefit from lower scaling factors, larger and more robust models like W2V2-Conformer+BARK can effectively leverage higher scaling factors. These findings highlight the importance of tuning the scaling factor to achieve the best domain adaptation performance for different ASR models.

5.6 Do *SYN2REAL* task vectors obtained with the same TTS have similar directions?

To further validate the *SYN2REAL* approach, we conducted a cosine similarity analysis between *SYN2REAL* task vectors derived by different text-to-speech (TTS) models: BARK (denoted as B_) and Speech T5 (denoted as S_). Figure 5 presents the cosine similarity heatmap between these *SYN2REAL* task vectors.

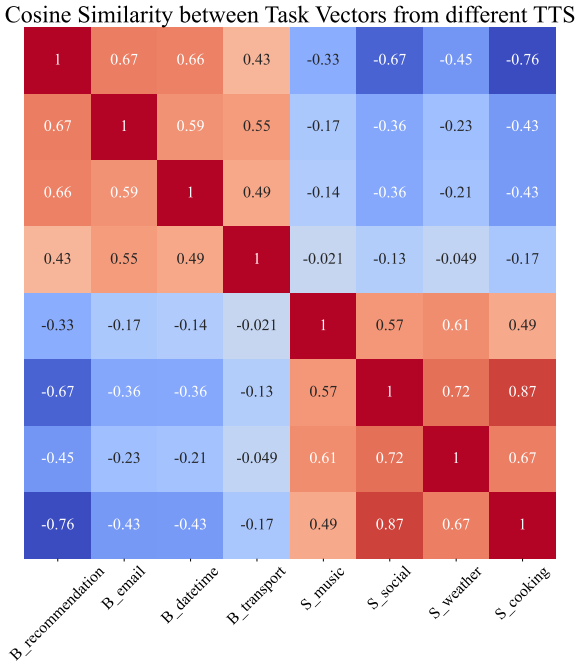

Figure 5: **Cosine Similarity between task vectors derived from Different TTS Models.** This heatmap shows the cosine similarity between task vectors generated by BARK (B_) and Speech T5 (S_) models. Higher similarity values between vectors from similar domains indicate effective acoustic-specific information transfer by the *SYN2REAL* method.

The heat map reveals that *SYN2REAL* task vectors from similar TTS exhibit higher cosine similarity, indicating that the *SYN2REAL* task vector effectively captures the distributional shifts between different acoustic domains.

Moreover, the negative similarities between certain *SYN2REAL* task vectors, such as 'B_recommendation' and 'S_music' (-0.67), highlight the distinct acoustic features between these TTS synthetic data, further emphasizing the effectiveness of the *SYN2REAL* approach in distinguishing and adapting to different acoustic environments.

The overall trend observed in the heatmap supports the hypothesis that the *SYN2REAL* task vectors not only bridge the gap between synthetic and real data but also maintain consistency within similar acoustic environments. This consistency is crucial for enhancing ASR performance across diverse target domains, as it ensures that the task vectors can generalize well to new, unseen domains.

6 *SYN2REAL* Task Vector given Domain Labels

In this section, we explore an alternative approach to generating *SYN2REAL* task vectors, assuming we have access to domain labels for the data in the source domains. This approach, which we refer to as *SYN2REAL Ensemble* task vector, involves generating separate *SYN2REAL* task vectors for each source domain and then combining them to enhance the adaptation of the ASR model to the target domain.

SYN2REAL Ensemble	**Cooking**	**Music**	**Social**	**Weather**	**Average**
Target Synthetic ASR (Baseline)	14.26	23.51	29.57	15.45	20.70
+ *SYN2REAL Ensemble*	14.46	**16.98**	**21.13**	**15.11**	**16.92**
Relative WER ↑	-1.40%	**27.78%**	**28.54%**	**2.20%**	**18.25%**

Table 5: **WER Performance on Whisper Small Model with *SYN2REAL Ensemble* task vectors.** This table compares the word error rate (WER) of the baseline ASR model fine-tuned on synthetic speech data with the WER of the model enhanced with the *SYN2REAL Ensemble* task vectors across four target domains: 'Cooking', 'Music', 'Social', and 'Weather'.

6.1 Performance of *SYN2REAL Ensemble* task vector

To evaluate the effectiveness of the *SYN2REAL Ensemble* task vector, we conducted experiments using the Whisper Small model with synthetic speech generated by the BARK TTS model. The experiments were carried out on four target domains: 'Cooking', 'Music', 'Social', and 'Weather,' using 17 source domains to create the *SYN2REAL Ensemble* task vectors. The results are presented in Table 5. The results indicate that the *SYN2REAL Ensemble* task vector provides significant improve-

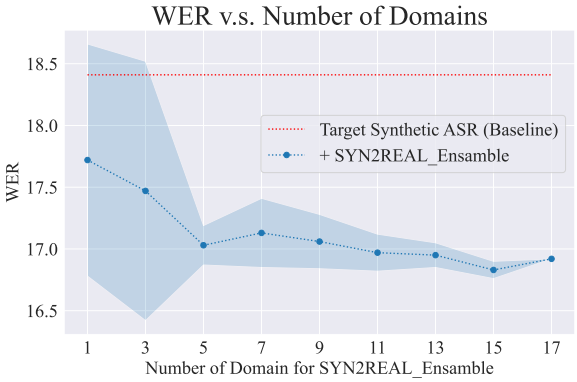

Figure 6: **WER vs. Number of Source Domains for *SYN2REAL Ensemble* task vector.** This plot shows the word error rate (WER) of the Whisper small model and the number of source domains used to generate the *SYN2REAL Ensemble* task vector with BARK model. The x-axis represents the number of source domains, and the y-axis represents the WER on average of the four domains ('Cooking', 'Music', 'Social', and 'Weather').

ments in WER for several target domains compared to the baseline method. The average WER is reduced from 20.70 to 16.92, representing an overall relative improvement of 18.25%. This highlights the effectiveness of using domain-specific task vectors to capture detailed acoustic characteristics, leading to enhanced model adaptation.

Comparing the *SYN2REAL Ensemble* task vector to the original *SYN2REAL* task vector, we find that *SYN2REAL Ensemble* task vector generally outperforms the original approach. The detailed domain-specific information captured by the distinct task vectors enhances model adaptation in most domains. However, in real-world scenarios, we often do not have access to labels finer-grained domain labels.

6.2 Impact of the number of domains on the performance of *SYN2REAL Ensemble* task vector

Figure 6 shows the average WER across four target domains ('Cooking', 'Music', 'Social', and 'Weather') when we use different numbers of source domain data to generate *SYN2REAL Ensemble* task vectors.

The results indicate that increasing the number of source domains generally improves ASR performance. The WER decreases from 17.8 to 16.8 as the number of source domains increases from 1 to 17. Notably, significant improvements are observed within the incorporation of the first 5 source domains. This trend suggests that incorporating more source domains helps capture diverse acoustic characteristics, leading to better domain adaptation.

7 Conclusion

This paper introduces *SYN2REAL* task vector to address mismatches between synthetic and real speech data. Future work will refine this approach and extend its application to other types of tasks and data such as visual data, contributing to more reliable speech and vision recognition systems. Experiments showed significant WER reductions, averaging 10.03% across 18 domains. We test various models, including Whisper Small, Whisper Base, Whisper Tiny, and Wav2Vec2-Conformer, with *SYN2REAL* task vector showing robust performance. We also explore *SYN2REAL Ensemble* approach further enhanced domain-specific adaptation but required domain labels. Overall, *SYN2REAL* task vector is a promising solution for improving ASR models with synthetic data.

8 Limitations

Domain-Specific Performance Variations While the *SYN2REAL* task vector shows significant improvements in many target domains, certain domains, such as 'Cooking' and 'Weather,' exhibit marginal improvements or slight degradation in word error rate (WER). This suggests that the task vector's effectiveness may vary based on the specific characteristics of different domains, indicating a need for further domain-specific fine-tuning and adjustments.

Scaling Factor Sensitivity The performance of the *SYN2REAL*-enhanced models is sensitive to the scaling factor λ. Finding the optimal scaling factor requires careful tuning, and the best value can vary between different ASR models and target domains. This adds a layer of complexity to the implementation and may limit the approach's generalizability without additional adaptive scaling strategies.

Synthetic Data Quality The approach relies heavily on the quality of synthetic speech data generated by TTS systems. Variations in the quality and acoustic properties of synthetic data across different TTS systems can impact the effectiveness of the *SYN2REAL* task vector. Ensuring consistent quality in synthetic data is crucial for achieving robust domain adaptation.

Model-Specific Dependencies The observed improvements are model-dependent, with larger mod-

els like Wav2Vec2-Conformer showing more substantial gains compared to smaller models like Whisper Tiny. This indicates that the *SYN2REAL* task vector's effectiveness might be influenced by the underlying model architecture and size, potentially limiting its applicability to a wider range of ASR models without further optimization.

9 Acknowledgements

We specifically thank Ting-Yao Hu and Dianna Yee for all the insightful discussions and constructive suggestions for this work.

This work was supported in part by the National Science and Technology Council under Grants NSTC 112-2634-F-002-006, MOST 110-2222-E-002-014-MY3, NSTC 113-2222-E-002-004-MY3, NSTC 113-2634-F-002-001-MBK.

References

Junyi Ao, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing Li, Yu Zhang, Zhihua Wei, Yao Qian, Jinyu Li, and Furu Wei. 2022. SpeechT5: Unified-modal encoder-decoder pre-training for spoken language processing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5723–5738, Dublin, Ireland. Association for Computational Linguistics.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Preprint*, arXiv:2006.11477.

Martijn Bartelds, Nay San, Bradley McDonnell, Dan Jurafsky, and Martijn Wieling. 2023. Making more of little data: Improving low-resource automatic speech recognition using data augmentation. *Preprint*, arXiv:2305.10951.

Emanuele Bastianelli, Andrea Vanzo, Pawel Swietojanski, and Verena Rieser. 2020. SLURP: A spoken language understanding resource package. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7252–7262, Online. Association for Computational Linguistics.

Vladimir Bataev, Roman Korostik, Evgeny Shabalin, Vitaly Lavrukhin, and Boris Ginsburg. 2023. Text-only domain adaptation for end-to-end asr using integrated text-to-mel-spectrogram generator. *Preprint*, arXiv:2302.14036.

Rishabh Bhardwaj, Do Duc Anh, and Soujanya Poria. 2024. Language models are homer simpson! safety re-alignment of fine-tuned language models through task arithmetic. *Preprint*, arXiv:2402.11746.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour. 2023. Audiolm: a language modeling approach to audio generation. *Preprint*, arXiv:2209.03143.

Edresson Casanova, Kelly Davis, Eren Gölge, Görkem Göknar, Iulian Gulea, Logan Hart, Aya Aljafari, Joshua Meyer, Reuben Morais, Samuel Olayemi, and Julian Weber. 2024. Xtts: a massively multilingual zero-shot text-to-speech model. *Preprint*, arXiv:2406.04904.

Chang Chen, Xun Gong, and Yanmin Qian. 2023. Efficient text-only domain adaptation for ctc-based asr. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–7.

Nico Daheim, Nouha Dziri, Mrinmaya Sachan, Iryna Gurevych, and Edoardo M. Ponti. 2023. Elastic weight removal for faithful and abstractive dialogue generation. *Preprint*, arXiv:2303.17574.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented Transformer for Speech Recognition. In *Proc. Interspeech 2020*, pages 5036–5040.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Shih-Cheng Huang, Pin-Zu Li, Yu-Chi Hsu, Kuang-Ming Chen, Yu Tung Lin, Shih-Kai Hsiao, Richard Tzong-Han Tsai, and Hung yi Lee. 2024. Chat vector: A simple approach to equip llms with instruction following and model alignment in new languages. *Preprint*, arXiv:2310.04799.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2023. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations*.

Raviraj Joshi and Anupam Singh. 2022. A simple baseline for domain adaptation in end to end ASR systems using synthetic data. In *Proceedings of the Fifth Workshop on e-Commerce and NLP (ECNLP 5)*, pages 244–249, Dublin, Ireland. Association for Computational Linguistics.

Haokun Liu, Derek Tam, Muqeeth Mohammed, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. 2022. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. In *Advances in Neural Information Processing Systems*.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE*

International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 5206–5210.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *Preprint*, arXiv:2212.04356.

Gowtham Ramesh, Kartik Audhkhasi, and Bhuvana Ramabhadran. 2024. Task vector algebra for asr models. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12256–12260.

Hiroaki Sato, Tomoyasu Komori, Takeshi Mishima, Yoshihiko Kawai, Takahiro Mochizuki, Shoei Sato, and Tetsuji Ogawa. 2022. Text-Only Domain Adaptation Based on Intermediate CTC. In *Proc. Interspeech 2022*, pages 2208–2212.

Hsuan Su, Ting-Yao Hu, Hema Swetha Koppula, Raviteja Vemulapalli, Jen-Hao Rick Chang, Karren Yang, Gautam Varma Mantena, and Oncel Tuzel. 2024. Corpus synthesis for zero-shot asr domain adaptation using large language models. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12326–12330.

Anirudh S. Sundar, Chao-Han Huck Yang, David M. Chan, Shalini Ghosh, Venkatesh Ravichandran, and Phani Sankar Nidadavolu. 2023. Multimodal attention merging for improved speech recognition and audio event classification. *ArXiv*, abs/2312.14378.

Yi-Lin Sung, Linjie Li, Kevin Lin, Zhe Gan, Mohit Bansal, and Lijuan Wang. 2023. An empirical study of multimodal model merging. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1563–1575, Singapore. Association for Computational Linguistics.

Derek Tam, Mohit Bansal, and Colin Raffel. 2024. Merging by matching models in task parameter subspaces. *Transactions on Machine Learning Research*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need. *Preprint*, arXiv:1706.03762.

Tyler Vuong, Karel Mundnich, Dhanush Bekal, Veera Elluru, Srikanth Ronanki, and Sravan Bodapati. 2023. AdaBERT-CTC: Leveraging BERT-CTC for text-only domain adaptation in ASR. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 364–371, Singapore. Association for Computational Linguistics.

Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Dmytro Okhonko, and Juan Pino. 2020. Fairseq S2T: Fast speech-to-text modeling with fairseq. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 33–39, Suzhou, China. Association for Computational Linguistics.

Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Sravya Popuri, Dmytro Okhonko, and Juan Pino. 2022. fairseq s2t: Fast speech-to-text modeling with fairseq. *Preprint*, arXiv:2010.05171.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. 2023. Neural codec language models are zero-shot text to speech synthesizers. *Preprint*, arXiv:2301.02111.

Karren Yang, Ting-Yao Hu, Jen-Hao Rick Chang, Hema Swetha Koppula, and Oncel Tuzel. 2023. Text is all you need: Personalizing asr models using controllable speech synthesis. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Kwok Chin Yuen, Li Haoyang, and Chng Eng Siong. 2023. Asr model adaptation for rare words using synthetic data generated by multiple text-to-speech systems. In *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1771–1778.

Jinghan Zhang, Shiqi Chen, Junteng Liu, and Junxian He. 2023. Composing parameter-efficient modules with arithmetic operation. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Xianrui Zheng, Yulan Liu, Deniz Gunceler, and Daniel Willett. 2021. Using synthetic audio to improve the recognition of out-of-vocabulary words in end-to-end asr systems. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5674–5678.