Advances in Machinery, Materials Science and Engineering Application IX
M. Chen et al. (Eds.)

doi:10.3233/ATDE230552

CAMP: A Unified Data Solution for Mandarin Speech Recognition Tasks

Zeping MIN [a,1], Qian GE [b] and Zhong LI [c]
[a] *School of Mathematical Sciences, Peking University, Beijing, China*
[b] *Academy for Advanced Interdisciplinary Studies, Peking University, Beijing, China*
[c] *Microsoft Research Asia, Beijing, China*

Abstract. Speech recognition, the transformation of spoken language into written text, is becoming increasingly vital across a broad range of applications. Despite the advancements in end-to- end Neural Network (NN) based speech recognition systems, the requirement for large volumes of annotated audio data tailored to specific scenarios remains a significant challenge. To address this, we introduce a novel approach, the Character Audio Mix- up (CAMP), which synthesizes scenario-specific audio data for Mandarin at a significantly reduced cost and effort. This method concatenates the audio segments of each character's Pinyin in the text, obtained through force alignment on an existing annotated dataset, to synthesize the audio. These synthesized audios are then used to train the Automatic Speech Recognition (ASR) models. Experiments conducted on the AISHELL-3, and AIDATATANG datasets validate the effectiveness of CAMP, with ASR models trained on CAMP synthesized data performing relatively well compared to those trained with actual data from these datasets. Further, our ablation study reveals that while synthesized audio data can significantly reduce the need for real annotated audio specific to each scenario, it cannot entirely replace real audio. Thus, the importance of real annotated audio data in specific application scenarios is emphasized.

Keywords. Speech recognition, audio data synthesis, low resource

1. Introduction

Speech recognition, the conversion of spoken language into written text, is becoming increasingly essential in various applications due to its broad range of uses. It plays a crucial role in diverse fields such as transcription services, voice assistants, telephony services, medical dictation, and accessibility services, all of which heavily depend on efficient and accurate speech recognition systems. The advent of deep learning has shifted speech recognition from traditional HMM- GMM modeling [1] towards end-to-end speech recognition models [2]–[7].

The training of end-to-end Neural Network (NN) based speech recognition systems is a complex task that often re- quires large volumes of annotated audio data. These end-to- end models implicitly learn a language model during training, although the inclusion of an auxiliary language model at the decoding stage often improves results. However, each application scenario presents a unique distribution of

[1] Zeping MIN, Corresponding Author, School of Mathematical Sciences, Peking University, Beijing, China, Email: zpm@pku.edu.cn

transcribed text as shown in figure 1, implying that merely adding a targeted language model at the decoding stage may not yield the desired results.

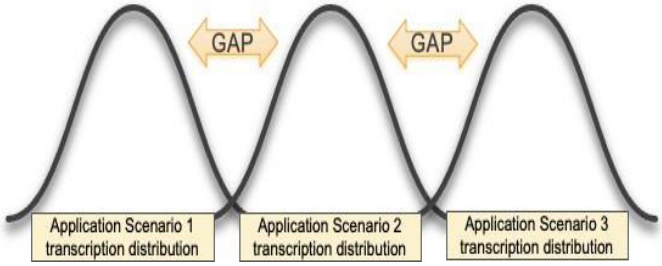

Figure 1. A gap exists in the distribution of transcription across different application scenarios.

The specific requirements of each scenario highlight the need for scenario-specific training data. Even within the same language, different applications may require the collection of well-annotated audio data tailored to specific scenarios. Although this is essential for developing a robust and specialized speech recognition system, it can be labor-intensive and time-consuming.

To address these challenges, we introduce a novel approach, the Character Audio Mix-uP (CAMP). This method generates scenario-specific synthesized audio data for Mandarin at a significantly lower cost and effort, potentially transforming the approach to training Mandarin speech recognition systems. In essence, this method concatenates the audio segments of each character's Pinyin in the text to synthesize the audio. These synthesized audios are then used to train the ASR models. The Pinyin audio segments used for concatenation can be obtained by performing force alignment on an existing annotated dataset, which does not need to be scenario-specific. To evaluate the effectiveness of the CAMP method, we use the AISHELL-1, AISHELL-3, and AIDATATANG datasets to simulate three distinct data scenarios. We sourced the Pinyin audio segments for concatenation from the AISHELL-1 dataset, then synthesized the audio data for the AISHELL- 3 and AIDATATANG datasets using our CAMP method. We found that ASR models trained with CAMP synthesized data performed relatively well compared to those trained with the actual AISHELL-3 and AIDATATANG data. The Character Error Rate (CER) of an ASR model trained on real data from the AISHELL-3 dataset is 8.71, while the CER of a model trained on synthesized data and a small amount of real data reaches 14.74. On the AIDATATANG dataset, the CER of an ASR model trained on real data is 4.72, while the CER of a model trained on synthesized data and a small amount of real data reaches 8.26.

Furthermore, our ablation study reveals that while synthesized audio data can significantly reduce the need for real annotated audio specific to each scenario, it cannot entirely replace real audio. In other words, the contribution of a small amount of real data cannot be fully replaced. We still need to collect real annotated audio data for each specific application scenario, but the CAMP method can significantly reduce the quantity of real audio data required. We conjecture that this is because training models solely with synthesized audio makes it difficult for the model to learn robust features.

The subsequent sections delve into our proposed method- ology in detail, provide an overview of the experiments con- ducted, discuss the results, and finally, conclude based on our findings.

In summary, our contributions are as follows:

- We introduce a novel approach, the Character Audio Mix- up (CAMP), for synthesizing scenario-specific audio data for Mandarin, significantly reducing the need for scenario- specific annotated data.
- Through experiments on the AISHELL-3, and AI- DATATANG datasets, we validate the effectiveness of CAMP. The results indicate that ASR models trained on CAMP synthesized data only suffer a minor CER loss compared to those trained with actual data from these datasets.

Our ablation study shows that while synthesized audio data can substantially lessen the requirement for real an- notated audio in each scenario, it cannot entirely replace it. Hence, we emphasize the continuing importance of real annotated audio data in specific application scenarios.

2. Method

Considering the varying distribution of transcribed text across different application scenarios and the time-consuming and costly nature of data collection, we propose the Character Audio Mix-uP (CAMP) method. The fundamental operation of CAMP is to concatenate the audio segments of each character's Pinyin in the text, customized to the text in the target application scenario, to synthesize audio. These synthesized audios are then used to train the Automatic Speech Recognition (ASR) models intended for the target scenario. The audio segments of each character's Pinyin used for concatenation can be obtained through force alignment on arbitrarily available collected Mandarin audio datasets. Therefore, CAMP can function as an audio data solution for numerous practical application scenarios of ASR models in Mandarin Chinese. A graphical representation of our CAMP method is provided in figure 2.

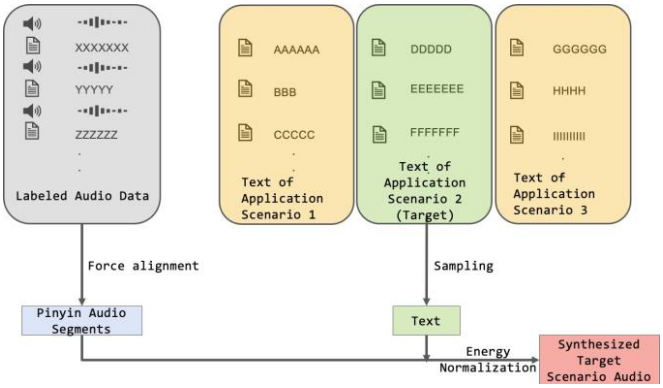

Figure 2. Illustration of the CAMP method. The process begins with the extraction of Pinyin audio segments from an existing dataset via force alignment. These segments are then concatenated according to the text of the target scenario to create synthesized audio data. Before concatenating, the audio segments are further refined through energy normalization. The synthesized audio data is ultimately used for training the ASR model.

Additionally, to improve the quality of the synthesized audios from CAMP, we suggest a straightforward post-processing method known as energy normalization.

For the audio segments $a_1, a_2, \ldots, a_n$ to be concatenated, this method can be mathematically represented as:

$$E = \frac{1}{n}\sum_{i=1}^{n}\|a_i\|_2, \quad \tilde{a}_i = \frac{a_i}{\|a_i\|_2}E \tag{1}$$

2.1. Audio Segment of Pinyin

In labeled audio datasets, we can employ force alignment technology to associate characters with audio segments. By using a Pinyin tool, we can map these characters to Pinyin, resulting in a correspondence between Pinyin and audio segments, thereby obtaining the audio segment of Pinyin. In this section, we discuss the benefits of using Pinyin to retrieve audio segments. Given the large number of characters in Mandarin and the disparity in transcribed text distribution across different scenarios, using characters to retrieve audio segments presents certain challenges. If the target scenario contains characters that are not included in the transcriptions of the labeled audio dataset, we would be unable to retrieve the corresponding audio segments accurately. However, Pinyin represents the basic phonetic units of Mandarin, significantly mitigating the issues encountered when using characters to retrieve audio segments.

3. Experiments

To validate the effectiveness of our CAMP method, we con- duct numerical experiments on automatic speech recognition (ASR) tasks. We construct audio segments of Pinyin using the aishell-1 dataset [8]. Subsequently, we use the aishell-3 [8] and aidatatang datasets to represent two other application scenarios. We employ our CAMP method to synthesize the audio of the training split transcriptions in aishell-3 and aidatatang, and then use these synthesized audios to individually train ASR models.

It's important to note that in the experiments using synthesized audio data by CAMP, the training data includes a small amount of real speech data. Details about the data we used are presented in table 1. We found that when using synthesized audio, the inclusion of a small amount of real audio is crucial for improving ASR performance. We validated this through ablation experiments. This might be due to the model's difficulty in learning robust acoustic features from only synthesized audio.

To demonstrate the contribution of CAMP-synthesized audio data in training ASR models, we establish two comparison groups: one uses the real training split data to train the ASR model, and the other only uses the small amount of real audio included to train the ASR model. In all setups, we use the same model architecture (two-pass CTC and AED joint architecture) to ensure fairness.

Table 1. Details about the small amount of real speech data included.

Dataset	Small amount of real speech data included
AISHELL-3	7500 utterances (randomly selected from train split)
AIDATATANG	dev split

3.1. Setup

The experiments are conducted using WeNet [3], which is a two-pass CTC and AED joint architecture, as shown in figure 3. A shared encoder, which includes multiple conformers [9], extracts information from speech data and encodes it into high-dimensional embeddings. The CTC decoder and attention decoder are trained jointly. The attention decoder consists of multiple transformer decoder layers. The experiments on AISHELL-3 and AIDATATANG are performed with a respective batch size of 16. The Adam optimizer with a learning rate of 0.002 is used during the training process. We adopt 12 conformer layers as the shared encoder and 6 transformer layers as the attention decoder. The embedding dimensions of transformer layers are set as 256 with 4 attention heads.

During inference, we use the advanced attention rescoring decode mode.

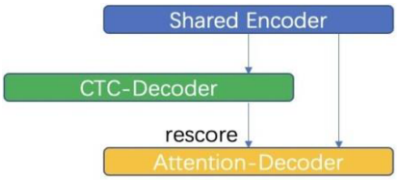

Figure 3. WeNet architecture.

3.2. Results

The primary results of our experiments are presented in table 2.

Here, the shorthand *'small amount real data only'* indicates that the training is only performed on the small amount of real data, while the shorthand *'small amount real data + synthesized data'* denotes that the training set consists of the small amount of real data and CAMP synthetic data. For comparison, we also train with full real audio data in the train split of aishell-3 and aidatatang. Note that we conduct all experiments without using any language models.

From table 2, several observations can be made. If we only use a small amount of real data, the final character error rate is quite high on both aishell-3 and aidatatang datasets. A CER larger than 60 on test sets implies that the sentences are barely understandable. This poor performance is expected due to the lack of training data. With the aid of CAMP synthetic data, the CER on test sets decreases significantly on both datasets, which validates the effectiveness of our CAMP methods. Compared to the results obtained on the whole training sets, our results are still competitive despite the limited usage of real data.

Moreover, when we examine the incorrect examples of inference texts on test audio, most of the wrong words have similar pronunciations to the correct ones, and the whole sentences can be easily understood and corrected.

Table 2. The CER results on test sets under different training sets. it can be observed that by adding synthesized data generated by camp, the performance of the ASR models can significantly improve.

Data setting	Character Error Rate (CER)	
	AISHELL-3	AIDATATANG
Small amount real data only	>60	20.38
Small amount real data + synthesized data	14.74	8.26
Full real data	8.71	4.72

4. Ablation Studies

4.1. Importance of Real Data

To further investigate the impact of synthesized data, we also design corresponding ablation experiments. On one hand, we remove the small amount of real data and only use the CAMP synthetic data. The CER results on test sets are shown in table 3. The observation is that the small amount of real data is indeed important, and its absence will lead to a significant decrease in the model performance.

According to the experiment results, we conjecture that adding a small amount of real data helps the model learn a robust feature, which indicates the importance of real data.

Table 3. The CER results on test sets under different data settings. It can be observed that a small amount of real data is important. After removing the real data, the performance of the ASR model drops significantly.

Dataset	Character Error Rate (CER)	
	Small Amount Real Data + Synthesized Data	Synthesized Data Only
AISHELL-3	14.74	30.73
AIDATATANG	8.26	54.78

4.2. Size of Synthesized Data

To better understand the contribution of the synthesized audio to training, we vary the amount of synthesized audio by CAMP and observe the changes in model performance. The training performances under different amounts of synthesized audio by CAMP are shown in table 4.

Table 4. ASR results under different amounts of synthesized utterances. The results further illustrate the fact that synthesized utterances indeed help ASR models.

Data setting	Character Error Rate (CER)	
	AISHELL-3	AIDATATANG
Small amount of real data only	>60	20.38
Small amount of real data + 10,000 synthesized utterances	35.51	13.44
Small amount of real data + 20,000 synthesized utterances	30.22	12.27
Small amount of real data + all synthesized utterances	14.74	8.26

5. Related Work

End-to-end Speech recognition has been a subject of intensive research over the past few years [2]–[7]. Neural network- based speech recognition systems have been shown to implicitly learn a language model during the training process [10], [11]. However, the distribution of transcribed text varies considerably across different application scenarios, suggesting that simply incorporating a targeted language model at the decoding stage [11], [12] may not yield the desired outcome. Therefore, it is often

necessary to collect specific speech data for training in each scenario. Considering the substantial cost of data collection, several solutions have been proposed, including data augmentation [13], [14], and synthesized-label semi-supervised learning [15], [16]. Nonetheless, current approaches have limitations that hinder their wide adoption. For example, synthesized-label semi-supervised learning often requires a decent initial synthesized labeler, and data augmentation methods struggle to expand the diversity of transcribed text.

The use of synthesizing audio data for specific scenarios has also been explored. A common approach is to train a Text-to-Speech (TTS) system to generate synthetic audio [17]–[25]. However, these TTS systems are typically NN-based, implying extensive computational resources required for training and generating synthetic audios. Worse, they might necessitate corresponding labeled data for training, limiting their feasibility under extremely low-resource conditions.

In this paper, we offer a unified data solution for numerous application scenarios in Mandarin: synthesizing new audio by concatenating Pinyin audio segments. This approach is straightforward, efficient, and easy to implement. It can significantly reduce the need for real audio data for training ASR systems in various practical application scenarios, thereby having the potential for broad impact.

6. Conclusion and Future Work

In this work, we introduced the Character Audio Mix-uP (CAMP) method, a novel approach for synthesizing scenario- specific audio data for Mandarin. This method significantly reduces the need for scenario-specific annotated data, thereby addressing the challenges associated with the collection of well-annotated audio data tailored to specific scenarios.

Our experiments on the AISHELL-3 and AIDATATANG datasets validated the effectiveness of CAMP. The results indicated that ASR models trained on CAMP synthesized data only suffered a minor Character Error Rate (CER) loss compared to those trained with actual data from these datasets. This demonstrates the potential of CAMP to revolutionize the approach towards training speech recognition systems.

However, our ablation study revealed that while synthesized audio data can substantially lessen the requirement for real annotated audio in each scenario, it cannot entirely replace it. This underscores the continuing importance of real annotated audio data in specific application scenarios.

As for future work, we recognize that the current implementation of CAMP is dependent on Pinyin and is thus limited to Mandarin. We aim to extend this method to other languages, broadening its applicability and further enhancing its potential to transform the field of speech recognition. This expansion will require addressing the pronunciation rules of each language, which presents an exciting avenue for further research.

References

[1] Povey D, Ghoshal A, Boulianne G, et al. The kaldi speech recognition toolkit. IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, IEEE Signal Processing Society, 2011.

[2] Watanabe S, Hori T, Karita S, et al. Espnet: End-to-end speech processing toolkit. arXiv Preprint arXiv:1804.00015, 2018.
[3] Yao Z, Wu D, Wang X, et al. Wenet: Production oriented streaming and non-streaming end-to-end speech recognition toolkit. arXiv Preprint arXiv:2102.01547, 2021.
[4] Dong L, Xu S and Xu B. Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2018; pp. 5884–5888.
[5] Chan W, Jaitly N, Le QV and Vinyals O. Listen, attend and spell. arXiv preprint arXiv:1508.01211, 2015.
[6] Graves A. Sequence transduction with recurrent neural networks. arXiv preprint arXiv:1211.3711, 2012.
[7] Graves A, Ferna´ndez S, Gomez F and Schmidhuber J. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. Proceedings of the 23rd International Conference on Machine learning. 2006; pp. 369–376.
[8] Shi Y, Bu H, Xu X, Zhang S and Li M. Aishell-3: A multi-speaker mandarin TTS corpus and the baselines. arXiv preprint arXiv:2010.11567, 2020.
[9] Gulati A, Qin J, Chiu CC, et al. Conformer: Convolution-augmented transformer for speech recognition. 2020.
[10] Battenberg E, Chen JR, Child, et al. Exploring neural transducers for end-to-end speech recognition. 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), IEEE, 2017; pp. 206–213.
[11] Shan C, Weng C, Wang G, et al. Component fusion: Learning replaceable language model component for end-to-end speech recognition system. ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019; pp. 5361–5635.
[12] Sriram A, Jun H, Satheesh S and Coates A. Cold fusion: Training seq2seq models together with language models. arXiv preprint arXiv:1708.06426, 2017.
[13] Park DS, Chan W, Zhang Y, et al. Specaugment: A simple data augmentation method for automatic speech recognition. arXiv preprint arXiv:1904.08779, 2019.
[14] Ko T, Peddinti V, Povey D and Khudanpur S. Audio augmentation for speech recognition. Sixteenth Annual Conference of the International Speech Communication Association, 2015.
[15] Higuchi Y, Moritz N, Le Roux J and Hori T. Momentum pseudo-labeling: Semi-supervised ASR with continuously improving pseudo-labels. IEEE Journal of Selected Topics in Signal Processing. 2022; 16, (6) pp. 1424–1438.
[16] Likhomanenko T, Collobert R, Jaitly N and Bengio S. Continuous soft pseudo-labeling in ASR. arXiv preprint arXiv:2211.06007, 2022.
[17] Laptev A, Korostik R, Svischev A, Andrusenko A, Medennikov I and Rybin S. You do not need more data: Improving end-to-end speech recognition by text-to- speech data augmentation. 2020 13th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), IEEE, 2020; pp. 439–444.
[18] Rossenbach N, Zeyer A, Schlu¨ter R and Ney H. Generating synthetic audio data for attention-based speech recognition systems. ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2020; pp. 7069‑7073.
[19] Sun G, Zhang Y, Weiss RJ, et al. Generating diverse and natural text-to-speech samples using a quantized fine-grained VAE and autoregressive prosody prior. ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2020; pp. 6699–6703.
[20] Li J, Gadde R, Ginsburg B and Lavrukhin V. Training neural speech recognition systems with synthetic speech augmentation. arXiv preprint arXiv:1811.00707, 2018.
[21] Ueno S, Mimura M, Sakai S and Kawahara T. Data augmentation for ASR using TTS via a discrete representation. 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), IEEE, 2021; pp. 68– 75.
[22] Rosenberg A, Zhang Y, Ramabhadran B, et al. Speech recognition with augmented synthesized speech. 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), IEEE, 2019; pp. 996–1002.
[23] Tjandra A, Sakti S and Nakamura S. Listening while speaking: Speech chain by deep learning. 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), IEEE, 2017; pp. 301–308.
[24] Zevallos R. Text-to-speech data augmentation for low resource speech recognition. arXiv preprint arXiv:2204.00291, 2022.
[25] Xue S, Tang J and Liu Y. Improving speech recognition with augmented synthesized data and conditional model training. 2022 13th International Symposium on Chinese Spoken Language Processing (ISCSLP), IEEE, 2022; pp. 443–447.