Overview of Methods for Data Augmentation for Speech-to-Text and Text-to-Speech

Blagica Penkova[1,2],Maja Mitreska[1,2], Kostadin Mishev[1,2] and Monika Simjanoska[1,2]
[1]iReason, LLC, Skopje, N. Macedonia
[2]Ss. Cyril and Methodius University,
Faculty of Computer Science and Engineering,
Skopje, N. Macedonia
{blagica.penkova, maja.mitreska, kostadin.mishev, monika.simjanoska}@ireason.mk,
{blagica.penkova, maja.mitreska.1}@students.finki.ukim.mk
{kostadin.mishev, monika.simjanoska}@finki.ukim.mk

***Abstract*—In the field of machine learning and deep learning, data augmentation is a widely used technique to expand the amount of training data available. This involves altering existing data instances or generating new synthetic data, with the aim of enhancing the quantity and variability of the training set.It has shown to be especially useful when working with low-resource languages and domains, where datasets are limited. This paper provides an overview of the data augmentation methods used for speech-related tasks, specifically for speech-to-text and text-to-speech applications.The goal of this paper is to provide researchers and practitioners with a comprehensive understanding of the data augmentation methods available for speech-related tasks, their strengths and potential applications.**

***Index Terms*—Data augmentation, Speech-to-text, Text-to-speech**

I. Introduction

Speech-To-Text (STT) and Text-to-Speech (TTS) are two essential components of spoken language processing.The performance of these systems largely depends on the quantity and quality of the training data available.However, data insufficiency remains a major challenge in building these systems, making data augmentation an increasingly crucial strategy to overcome this issue. Data augmentation in STT and TTS serves as a method to improve the robustness of the models and to avoid over-fitting on a given dataset. However, the selection of an appropriate data augmentation strategy can be challenging since it depends on the architecture being used. Different models require different data representations, and not all augmentation techniques may be suitable for a particular model. Therefore, it is crucial to carefully choose and implement the most suitable data augmentation methods for a given architecture. This will ensure that the model is trained on a diverse set of data and can better generalize to new data.

In section II and III, we will provide a comprehensive overview of the latest advancements and commonly used techniques in data augmentation for speech-to-text and text-to-speech technologies, respectively.

II. Methods for data augmentation for Speech-to-Text

One of the earliest approaches of data augmentation for Speech-to-Text was to vary the speed of the audio signal by speeding it up or slowing it down, a technique known as speed perturbation [1].In the years that followed, researchers developed various other data augmentation techniques, such as adding background noise, pitch shifting, and time stretching. However, with the advent of deep learning, researchers began to focus on developing more advanced and efficient techniques.

SpecAugment is one of the most popular methods for data augmentation in speech recognition and it operates on the log-mel spectrogram of the audio input and involves three modifications: time-wrapping, time masking, and frequency masking. Time-wrapping involves wrapping a chosen point on the log-mel spectrogram to the left or right along the time axis, effectively stretching or compressing the audio signal in time. Time and frequency masking are applied to a selected number of consecutive time steps or mel frequency channels, randomly masking out certain portions of the spectrogram to force the STT system to be more robust to missing information. In the paper [2] published in 2019, the authors applied SpecAugment to the LibriSpeech dataset and evaluated its effectiveness using a Listen, Attend and Spell (LAS) network.The results showed that SpecAugment can significantly improve the performance of STT systems, achieving state-of-the-art results on the LibriSpeech dataset and on other datasets as well.

Interestingly in 2021, there is an upgraded version of SpecAugment called SpecAgument++ [3] that applied data augmentation techniques not only on the input spectograms, but also on the data in the hidden space to enhance the intermediate representation of the feature vectors. For the intermediate state, two approaches are applied, frequency masking of the channels and masking blocks of time frames. Moreover, experiments are made with three different masking schemes, zero-masking (masking consecutive units with zeros), mini-batch based mixture and mini-batch based cutting masking which included additional noise to the dataset.

Three notable approaches from 2019 that also gained attention are synthetic data, WaveNet and data augmentation using Generative Adversarial Networks (GAN).
In this paper [4], the authors present a simple approach that has proven to be useful, which involves expanding a natural speech dataset with synthetic speech. Training the Tacotron-2 with GST model on the MAILABS English-US dataset, they obtained model that successfully learnt all 3 different speaking styles and accents. The authors then used this model to synthesize speech data from the LibriSpeech dataset, resulting in a new dataset that contains examples spoken by multiple speakers with multiple accents.

Another approach to data augmentation in STT is proposed in [5], which focuses on augmenting data with different voices of the same utterances. The authors of this paper use a generative model, WaveNet [6], to convert the voice characteristics of one speaker to another. By reducing the usage of acoustic features, the approach synthesizes speech with different pitch patterns.

More advanced methodologies propose usage of Generative Adversarial Networks (GAN) for data augmentation. In these approaches the transformations are not performed directly on the raw audio input as in the more traditional approaches. In [7], the created speech samples are based on spectrum feature level and are generated frame by frame with no dependencies between them. Moreover, the augmented data contains no real labels and all the generated samples are independent from each other. The GAN model consists of generator, which produces samples from a data distribution and usually that is a low dimensional random noise, and discriminator, which decide whether the sample is real data or generated noise.

After the introduction of the SCADA method [8] in 2020, this approach has gained attention for improving STT systems. SCADA represents stochastic, consistent and adversarial data augmentation and all of the augmentation are applied on the feature domain i.e. on the mel filterbank outputs. The experiments are done using two stacks of data augmentation methods. The first one comprises RandAugment [9] policy that randomly samples between three augmentations: identity, low pass and scaled Gaussian noise (RA-pre). And the second one, combines the RandAugment with the SpecAugment method (RA-spec). Given an input mel spectrogram RA-pre is applied, and than RA-spec. The authors incorporate various consistency measures like encoder consistency, consistency loss, Jensen-Shannon Divergence and virtual adversarial noise to the ASR encoder and decoder to obtain substantial WER reduction.

Two noteworthy techniques that gained attention in 2021 are SpliceOut and SapAugment.

SpliceOut [10] is a data augmentation method that performs a simple modification to time masking, and can be used in combination with other data augmentation methods. For a given log-mel spectogram with T time steps, SpliceOut selects N intervals that are removed from the input. The remaining parts are joined and they represent it the augmented sample.

A Sample-Adaptive Policy for Augmentation (SapAugement) is proposed in [11] and explores a novel approach of

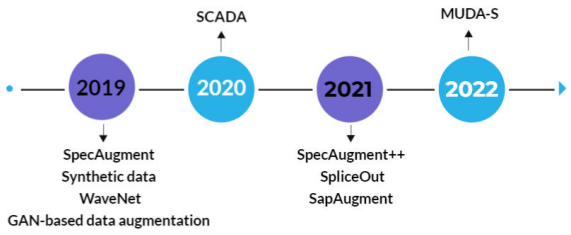

Fig. 1. Timeline of Speech-To-Text Data Augmentation Methods

augmenting data. Despite the recent models that perform the same method on all of the training samples, SapAugemnt proposes applying the augmentation parameters based on the training loss of the data samples. The sample-adaptive strategy takes as input the training loss and produces a scalar that sets the intensity of the augmentation. A sample with minimal loss receives a greater augmentation, whereas a sample with significant loss receives a softer augmentation. To the existing three augmentation policies from SpecAugment, time and frequency masking, and time stretching on the log-mel spectograms (feature domain augmentation), SapAugment adds two more methods for raw speech domain augmentation. Those are SampleParsing and CutMix initially developed for image augmentation. The SapAugment framework training is defined as determining a policy for selecting and adapting augmentation parameters that enhances validation accuracy.

A recent paper [12] published in 2022 presents a novel data augmentation technique for improving the performance of speech recognition systems. The proposed approach generates synthetic speech utterances with varying speaking rates and applies speed perturbations to the original data. This technique, called MUDA-S, is designed to expose speech recognition systems to a more diverse set of training data and improve their robustness to variations in speaking rate. Experimental results on several speech recognition tasks show that MUDA-S can significantly improve the accuracy of end-to-end speech recognition systems, particularly for tasks that involve speech from non-native speakers or speech that is distorted by noise or reverberation.

Lately, many methodologies are trying to incorporate data augmentation on-the-fly to optimize the computational performances of the models. In [13], a dynamic time stretching is performed on the feature vectors and sub-sequence sampling.

The paper focuses on improvement of the sequence-to-sequence speech recognition systems. The other paper, [14], introduces real time speech enhancement in the waveform domain where the Remix augmentation shuffles the noises within the batch and Band Mask to remove 20% of the frequencies uniformly.

Also, there is a growing trend of using text-to-speech (TTS) systems to augment automatic speech recognition (ASR) systems. TTS systems can generate speech data from text-only data, which can then be used to augment ASR systems [15].

III. Methods for Data Augmentation for Text-to-Speech

As for data augmentation methods in SST, there are also data augmentation methods for the text-to-speech techniques. Text-to-speech systems generally consists of two part: a sequence-to-sequence acoustic model and a neural vocoder.

Several papers in 2020 introduced innovative techniques for Text-To-Speech (TTS) data augmentation, including Copycat prosody transfer, quantized VAE for TTS, and speed factor-based data augmentation.

The Copycat model [16] is a fine-grained prosody transfer model that maintains prosody when converting from one speaker style to another and is composed of three parts: phoneme encoder, prosody bottleneck encoder and parallel decoder. A modification is incorporated in the model, as a concatenation is made to the upsampled phonemes before the phoneme encoder. The used TTS model is Tacotron-based with additional variational auto-encoder (VAE) and first, it is trained on the whole dataset, and fine-tuned only on the synthetic one.

The paper [17] introduces TSS system that uses quantized fine-grained variational autoencoder (VAE) structure for discretizing the latent features and using the result it trains an auto-regressive (AR) prior model. The architecture is Tacatron 2-based with quantized fine-grained VAE. As stated, the quantized representations improve the naturalness of the audio samples produced in the latent vector space.

In [18] a full TTS system are proposed that leverages the use of data augmentation technique based on speed factor. Additional synthetic speech is generated with chancing the pitch and the speed of the utterances.

In 2021, [19] a data augmentation method is proposed by utilizing previously recorded samples in a desired speaking style from other speakers. That synthetic dataset is furthermore used as expansion of the original real data. The goal of Voice Conversion is to transform an utterance spoken by a source speaker into an utterance read by a separate target speaker while retaining all other linguistic properties.

Interesting approach is proposed in [20], where the augmented audio were created on-the-fly with the latest TTS model during the training phase. In particular, augmented data were generated by 'forcing' a speaker to reproduce the utterances of the other three speakers by requiring their attention alignment matrices to be as comparable as possible. The TTS model is based on GST-Tacotron 2, and the augmentation is based on the attention alignment matrix which is usually dropped by-product that may consist useful rhythmic information of the input. The augmented data imitates the alignment matrix of another speaker with different speaking style.

There are systems that incorporates TTS systems to improve the quality of other non-autoregressive TTS systems [21]. More precisely, a source AR TTS model is trained to generate high-quality data and used on a large unseen text corpus to create synthetic dataset. The newly created dataset is used to train non-AR TTS system which outperforms the current state-of-the-art models.

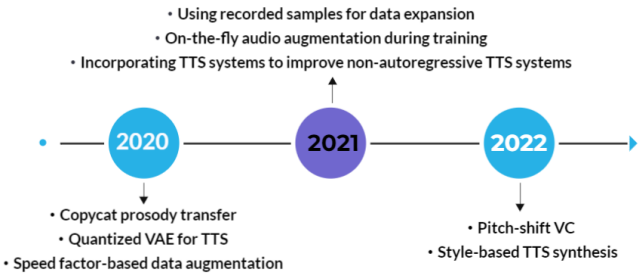

Fig. 2. Timeline of Text-To-Speech Data Augmentation Methods

The most recent paper published in 2022, [22] explores the VC method with pitch-shift for data augmentation. Using the pitch-shift (PS) technique, various pitch dynamics can be acquired from single speaker and the VC system play role in transforming neutral speech to speech with emotional attributes. The PS augmentation is first applied to the source and target speaker's neutral recordings and results with generating 15 times larger data. All the augmented data is used for training the VC component that is based on a non-parallel Scyclone model [23]. For the optimization of the VC training process a short-time Fourier transform (STFT) regularization loss is applied. After the VC model is trained on the augmented data, it is used to augment the original data once more, before passing it to the TTS architecture for further training.

The same year, [24] published StyleTTS [24] which represents a style-based generative model for TTS synthesis. To obtain natural speech synthesis, they leverage the novel Transferable Monotonic Aligner (TMA) and duration-invariant data augmentation techniques. A style encoder collects style vectors from reference audio in the architecture, and the style vectors are given to the decoder and prosody predictors through adaptive normalization. To learn natural prosody independently of phoneme duration estimate, a unique duration-invariant data augmentation technique is used. The technique generates lifelike prosodic patterns and emotive tones similar to the reference audio using stylization. The same text is connected with different speaking styles using several reference audios and enables one-to-many mapping. The framework consists of eight components: text encoder, text aligner, style encoder, a pitch extractor and a decoder.

IV. Conclusion

In this paper, we have provided a comprehensive overview of the latest data augmentation methods in the domains of Text-to-Speech and Speech-to-Text. Our review highlights the importance of data augmentation techniques in improving the performance and accuracy of these systems. It is evident from the research covered that there is significant potential for exploration and further development in these areas, especially with the rise in demand for voice and conversational AI assistants by multilingual customers.

references

References

[1] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Sixteenth annual conference of the international speech communication association*, 2015.

[2] D. S. Park, W. Chan, Y. Zhang, *et al.*, "SpecAugment: A simple data augmentation method for automatic speech recognition," Sep. 2019. DOI: 10.21437/interspeech.2019-2680.

[3] H. Wang, Y. Zou, and W. Wang, "Specaugment++: A hidden space data augmentation method for acoustic scene classification," *arXiv preprint arXiv:2103.16858*, 2021.

[4] J. Li, R. Gadde, B. Ginsburg, and V. Lavrukhin, *Training neural speech recognition systems with synthetic speech augmentation*, 2018. DOI: 10.48550/ARXIV.1811.00707.

[5] J. Wang, S. Kim, and Y. Lee, "Speech augmentation using wavenet in speech recognition," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6770–6774. DOI: 10.1109/ICASSP.2019.8683388.

[6] A. v. d. Oord, S. Dieleman, H. Zen, *et al.*, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[7] H. Hu, T. Tan, and Y. Qian, "Generative adversarial networks based data augmentation for noise robust speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5044–5048. DOI: 10.1109/ICASSP.2018.8462624.

[8] G. Wang, A. Rosenberg, Z. Chen, Y. Zhang, B. Ramabhadran, and P. J. Moreno, "Scada: Stochastic, consistent and adversarial data augmentation to improve asr.," in *INTERSPEECH*, 2020, pp. 2832–2836.

[9] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "Randaugment: Practical automated data augmentation with a reduced search space," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp. 702–703.

[10] A. Jain, P. R. Samala, D. Mittal, P. Jyothi, and M. Singh, "Spliceout: A simple and efficient audio augmentation method," *CoRR*, vol. abs/2110.00046, 2021. arXiv: 2110.00046.

[11] T.-Y. Hu, A. Shrivastava, J.-H. R. Chang, *et al.*, "Sapaugment: Learning a sample adaptive policy for data augmentation," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 4040–4044. DOI: 10.1109/ICASSP39728.2021.9413928.

[12] T. Zhong, Z. Liu, H. Wang, Y. Wang, and X. Zhang, "Audio Style Transfer: A Unified Approach to Voice Style Transfer and Music Style Transfer," in *Proc. Interspeech 2022*, 2022, pp. 2385–2389. DOI: 10.21437/Interspeech.2022-1151.

[13] T.-S. Nguyen, S. Stueker, J. Niehues, and A. Waibel, "Improving sequence-to-sequence speech recognition training with on-the-fly data augmentation," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 7689–7693.

[14] A. Defossez, G. Synnaeve, and Y. Adi, "Real time speech enhancement in the waveform domain," *arXiv preprint arXiv:2006.12847*, 2020.

[15] M. Mimura, S. Ueno, H. Inaguma, S. Sakai, and T. Kawahara, "Leveraging sequence-to-sequence speech synthesis for enhancing acoustic-to-word speech recognition," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, 2018, pp. 477–484.

[16] S. Karlapati, A. Moinet, A. Joly, V. Klimkov, D. Sáez-Trigueros, and T. Drugman, "Copycat: Many-to-many fine-grained prosody transfer for neural text-to-speech," *arXiv preprint arXiv:2004.14617*, 2020.

[17] G. Sun, Y. Zhang, R. J. Weiss, *et al.*, "Generating diverse and natural text-to-speech samples using a quantized fine-grained vae and autoregressive prosody prior," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 6699–6703.

[18] R. Liu, X. Wen, C. Lu, and X. Chen, "Tone learning in low-resource bilingual tts.," in *Interspeech*, 2020, pp. 2952–2956.

[19] G. Huybrechts, T. Merritt, G. Comini, B. Perz, R. Shah, and J. Lorenzo-Trueba, "Low-resource expressive text-to-speech using data augmentation," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 6593–6597.

[20] R. Chung and B. Mak, "On-the-fly data augmentation for text-to-speech style transfer," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, IEEE, 2021, pp. 634–641.

[21] M.-J. Hwang, R. Yamamoto, E. Song, and J.-M. Kim, "Tts-by-tts: Tts-driven data augmentation for fast and high-quality speech synthesis," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 6598–6602.

[22] R. Terashima, R. Yamamoto, E. Song, *et al.*, "Cross-speaker emotion transfer for low-resource text-to-speech using non-parallel voice conversion with pitch-shift data augmentation," *arXiv preprint arXiv:2204.10020*, 2022.

[23] A. Kanagaki, M. Tanaka, T. Nose, R. Shimizu, A. Ito, and A. Ito, "Cyclegan-based high-quality non-parallel voice conversion with spectrogram and wavernn," in *2020 IEEE 9th Global Conference on Consumer Electronics (GCCE)*, IEEE, 2020, pp. 356–357.

[24] Y. A. Li, C. Han, and N. Mesgarani, "Styletts: A style-based generative model for natural and diverse text-

to-speech synthesis," *arXiv preprint arXiv:2205.15439*, 2022.