

Utilizing TTS Synthesized Data for Efficient Development of Keyword Spotting Model

Hyun Jin Park¹, Dhruuv Agarwal¹, Neng Chen¹, Rentao Sun¹, Kurt Partridge¹, Justin Chen¹, Harry Zhang¹, Pai Zhu¹, Jacob Bartel¹, Kyle Kastner¹, Gary Wang¹, Andrew Rosenberg¹, Quan Wang¹

¹Google LLC, Mountain View, CA, U.S.A.

{hjpark, dhruuv, nengchen, sunrentao, kep, jstchen, harryz, paizhu, bartel, kkastner, wgary, rosenberg, quanw}@google.com

Abstract

This paper explores the use of TTS synthesized training data for KWS (keyword spotting) task while minimizing development cost and time. Keyword spotting models require a huge amount of training data to be accurate, and obtaining such training data can be costly. In the current state of the art, TTS models can generate large amounts of natural-sounding data, which can help reducing cost and time for KWS model development. Still, TTS generated data can be lacking diversity compared to real data. To pursue maximizing KWS model accuracy under the constraint of limited resources and current TTS capability, we explored various strategies to mix TTS data and real human speech data, with a focus on minimizing real data use and maximizing diversity of TTS output. Our experimental results indicate that relatively small amounts of real audio data with speaker diversity (100 speakers, 2k utterances) and large amounts of TTS synthesized data can achieve reasonably high accuracy (within 3x error rate of baseline), compared to the baseline (trained with 3.8M real positive utterances).

Index Terms: keyword spotting, TTS synthesized training data

1. Introduction

The keyword spotting (KWS) task is to detect spoken keywords while rejecting background speech and noise. KWS has become an important mechanism for activating conversational human-computer interfaces since the advances in ASR. Representative examples include virtual assistants like Alexa, Siri, and Google Assistant, where KWS technology is used to start user-assistant interaction [1–3].

A production level KWS system must cover a huge variety of conditions due to the diversity of populations, pronunciations, and acoustic environments. Also, keyword spotting models should be “always on”—ideally strictly causal for low latency, and with a small computational footprint to limit energy consumption. To meet these requirements, there has been substantial KWS research using neural networks. Prior work has shown significant quality improvement and latency reduction in low-resource settings [1–12].

Production level KWS models are usually trained with large amounts of training data to cover the wide diversity of pronunciation and acoustic environments. Gathering audio data specific to a target keyword incurs significant cost, as it requires human contributors to generate audio recordings.

Recent advancements in TTS (Text To Speech) allow generation of realistic audio at low cost. This has inspired many applications of generated data for the ASR domain [13–19]. For ASR, TTS models enable the use of text-only data, which is much more abundant than labeled audio. This benefit often leads to improved accuracy and reduced data cost.

Following the success in ASR, there have been efforts to utilize TTS-generated data for KWS [20,21]. Lin et al proposed to pre-train an embedding model with real data, and fine-tune attached classifier head models using limited amounts of TTS or real data [20]. Werchniak et al showed initial exploration of TTS data usage for single keyword detection problem [21] where mixing real and TTS data gave the best results.

The current state of the art TTS models can generate large amounts of synthetic data that sound natural to the human ear, but the generated data distribution may not match the distribution of real data. To address this mismatch, we propose TTS-based KWS model training strategies, based on three key components.

Firstly we develop a text generator that generates text phrases tailored for KWS training. The text generator is designed to maximize diversity of TTS synthesized output. Secondly, we utilize state of the art TTS modules that can synthesize speech with large number of voices. The TTS modules provide a large number of pre-trained voices, and support the generation of personalized voices based on input audio.

Finally, we evaluate various strategies to mix synthetic TTS data and real human speech data, with a focus of minimizing data cost while maximizing quality. To minimize cost and time for KWS model development, we evaluated mixing options that uses as small amount of real data as possible, while using large amounts of TTS generated data.

The contributions of this paper are : (1) We explore KWS model training using large amounts of synthetic data and *minimum amounts of real data* to achieve comparable accuracy to the baseline which uses large scale real positive data. (2) We also provide *reports on trade off relationship* between amount of used real data and model accuracy in multiple sweep conditions. (3) We propose a text generator that creates TTS input texts tailored to maximize diversity of TTS output by utilizing experimental *prosody control* feature of Virtuoso TTS (section 4.1).

2. Related works

For improved data diversity, we use two different TTS models.

2.1. Virtuoso TTS

Virtuoso is a multilingual speech-text joint training model that can learn from untranscribed speech, unspoken text, and paired speech-text data sources [22, 23]. This model is capable of generating speech in 139 languages for 726 predefined speaker profiles. We used the simple text-to-speech mode of the Virtuoso model. Given a transcript, the model can generate an utterance for a target language from a designated speaker, with randomized prosody. Experiments with Virtuoso TTS model indicates that punctuation symbols in the text input can be used to control the prosody of the synthesized speech, and we leverage this

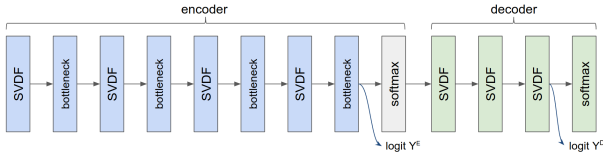


Figure 1: Baseline KWS model architecture

capability of the Virtuoso model to augment our synthesized data.

2.2. AudioLM TTS

AudioLM is an audio generative language model that features long term coherence and high quality [24]. We used a variant of AudioLM-based TTS model that can be conditioned by text and audio, with the key feature of synthesizing audio while retaining the speaker’s characteristics and prosody of the input audio [25]. The diversity of the generated dataset is able to match a rich variety of real human audio prompts.

3. Baseline keyword spotting model

3.1. Input features

In this study, we followed the lead of previous research [3,6] and employed the same input features. Specifically, we extracted a 40-dimensional vector of spectral filterbank energies (calculated over a 25-millisecond window) at every 10-millisecond time frame. To create a 120-dimensional input feature vector X_t for every 20 milliseconds, we stacked and strided three consecutive frames. To enhance the model’s resilience and adaptability, we incorporated data augmentation techniques. We followed the approach in [26], applying established methods like simulated reverberation and noise mixing to the data before feature extraction.

3.2. Architecture

We employed a two-stage model architecture (Fig. 1), as outlined in previous work [3,6]. This architecture comprises seven factored convolution layers (also known as SVDf [3]) and three bottleneck projection layers, organized into encoder and decoder sub-modules connected sequentially. The model is optimized for streaming inference, and has roughly 320,000 parameters.

The encoder module takes as input X_t , a vector of stacked spectral filter-bank energies representing the audio features. It then produces an N -dimensional encoder output Y^E , which is designed to encapsulate N phoneme-like sound units crucial for keyword recognition. This encoded representation is then passed to the decoder module, which generates a 2-dimensional output Y^D trained to predict the presence or absence of the keyword within the audio stream. The final prediction logit, denoted as Y , combines both the encoder and decoder outputs: $Y = [Y^E, Y^D]$. This unified representation enables robust keyword spotting in diverse audio environments.

3.3. Supervised training objective

The baseline training approach leverages two types of supervised losses (Eq. 1). The first loss term directly calculates the cross-entropy between model logits and labels, following the method established in [3]. The second loss term computes the cross-entropy between max-pooled logits and labels, as introduced

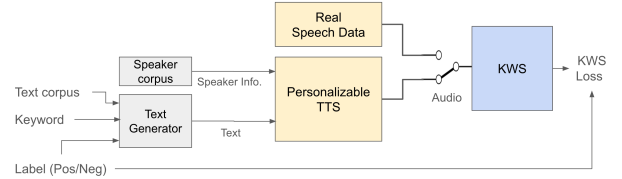


Figure 2: Proposed KWS approach with TTS generated data

in [6]. Both loss terms have distinct components for the encoder and decoder modules, and a weighted combination of these components forms the final loss.

The top level loss is a weighted combination of the cross-entropy and max-pooled losses (Eq. 1). This combination helps prevent overfitting and improves the model’s ability to perform well on unseen data, ultimately enhancing its robustness and effectiveness in keyword spotting tasks.

$$\mathcal{L}_{\text{sup}} = \sum_{t=1..n} [(1 - \alpha)L_{\text{CE}}(Y(X_t, \theta), c_t) + \alpha L_{\text{MP}}(Y(X_t, \theta), \omega_{\text{end}})] \quad (1)$$

$Y(X_t, \theta)$ denotes the combined encoder and decoder model output given input X_t and parameter set θ . L_{CE} represents the end-to-end loss proposed in Alvarez [3]. We use the implementation as defined by Eq. 2 in Park et al. [6] where c_t is the per-frame target label for CE-loss. L_{MP} represents the max-pool loss proposed in Park et al. [6], which was defined by Eq. 12. ω_{end} represents the end-of-keyword position label for the max-pool loss. α is a loss-weighting hyper-parameter determined empirically. Refer to [3,6] for details of L_{MP} and L_{CE} .

4. Proposed approach

Fig. 2 shows the high level view of the proposed TTS based KWS training approach. In this approach, the baseline KWS model can take input audio examples from either real speech data or TTS generated data sources. We explore training recipes that mix both real and synthesized data. The mixing ratio is a hyper-parameter which we explore in sweep experiments. We use TTS models that can generate speech samples in various speaker types and locales. The TTS models are conditioned by both speaker information and text. Speaker information can be either an index of the prefixed speaker (Virtuoso), or audio samples from any speaker (AudioLM). Text input to TTS is generated by a text generator, which combines target keyword and randomized negative text from a text corpus depending on target label (positive or negative). Details of the major components of the proposed approach are discussed below.

4.1. Text augmentation by text generator

We introduces a text generator module, which generates positive or negative phrase examples given target label (pos/neg), target keywords (for example, “hey google” or “ok google”) and a random text corpus (which constitutes negative phrases in the form of user queries following the keyword).

We define a keyword as a combination of *prefix* and *key_name* (keyword := (*prefix*, *key_name*)). For example *prefix* can be “Hey” and *key_name* can be “Google”. Based on given keyword (prefix and key_name) and random query text, we build a positive phrase by concatenating. Negative phrases

can be constituted by using any textual corpus and filter out the keyword. We also randomly add some prosody control symbols (Table 1) to vary the TTS output. Those prosody control symbols are experimental features of Virtuoso TTS model. We first define a set of text templates with variables (prefix, key_name, and query) and prosody control symbols (Table 2). Variables can be replaced with actual provided texts. Then those text templates are randomly sampled to generate TTS input texts.

Table 1: *Prosody control symbols*

Controlled texts	Effects
<i>text</i>	default pronunciation of <i>text</i>
<i>(text)</i>	speak <i>text</i> slowly
<i>text:</i>	insert pause after <i>text</i>
<i>text?</i>	increase pitch at the end of <i>text</i>
<i>text!</i>	speak <i>text</i> loudly

Table 2: *Text generation templates*

Text templates	Notes
{ <i>prefix</i> } { <i>key_name</i> } { <i>query</i> }	positive phrase
{ <i>prefix</i> } ({ <i>key_name</i> }) { <i>query</i> }	positive phrase
({ <i>prefix</i> }): ({ <i>key_name</i> }) { <i>query</i> }	positive phrase
{ <i>prefix</i> }: ({ <i>key_name</i>)? { <i>query</i> }	positive phrase
{ <i>prefix</i> }: { <i>key_name</i> }! { <i>query</i> }	positive phrase
{ <i>query</i> }	negative phrase

4.2. TTS data generation by Personalizable TTS

The TTS models used in our approach provide capabilities to generate diverse and personalized speech. The Virtuoso system supports 726 pretrained high quality speakers. AudioLM-based TTS model can generate speech with a voice matching with the input audio example. We maximally utilize such personalization capability of the TTS to generate training examples with diverse voices.

Also Virtuoso is highly multilingual model supporting 139 languages. We also utilize this feature to generate speech with different language targets from fixed English phrases. The resulting TTS output sounds like an accented version of English, adding diversity to the synthesized output.

4.3. TTS and Real data mixing strategy

Although current state of the art TTS can generate realistic human speech, the distribution of TTS generated data and real speech data might mismatch. TTS generated data might still have artifacts that does not exist in real speech recordings, and it might not generate all the variations present in real speech. To compensate for such mismatch we explore the strategy of mixing real data with TTS based data, by training models with various data mixing options, and evaluate them on real data.

5. Experimental Setup

5.1. Input Data

We trained KWS models for "Hey/OK Google" detection task using real and synthesized data. For real speech data, we used

anonymized utterances collected in accordance to Google’s Privacy and AI Principles [27, 28]. TTS data is generated using Virtuoso and AudioLM variant TTS models. Multi-style data augmentation [29] is applied during training. Table 3 summarizes the number of utterances used. TTS data numbers include both Virtuoso and AudioLM sources.

Table 3: *Data types and sizes*

Data Types	Utterance counts
Real Positive Utts	3.8 M
Real Negative Utts	14.1 M
Synthesized Positive Utts	7.5 M
Synthesized Negative Utts	5.1 M

5.2. Evaluation Data and Metric

We evaluate the model performance on real Hey/OK Google data sets. Our primary metric to compare model performance is false reject rate (FRR) and the model threshold is selected to optimize FRR while keeping the maximum allowed false accept per hour fixed at 0.133 which is a typical operational condition.

5.3. Different experiment sweeps

- Baseline model performance on different data-sets
In this experiment, we aim to establish the baseline metric scores when the model is trained on different data-sets including synthesized and real data.
We also explore addition of real negative data. Obtaining positive data, is constrained by the selected the key-word(s) while negative data can be obtained from virtually any data source as long as it doesn’t contain the target keyword. Hence, we explored adding such negative sets, here-by called as *base* real negative data, to improve baseline performance.
- TTS with incremental amounts of real positive data
In this experiment we wish to understand the need of real data for model training when quality TTS data is available. To do so, we compare performance of pure TTS trained model, and gradually increase the real positive data to 100k. We also test the effect of addition of real *base* negatives here, mentioned above, and see how it can offset the need for real data drastically.
- TTS data and varying amount of speakers in real data
We train models with the baseline of fixed TTS configuration, while adding real data uniformly per speaker and gradually increasing speaker count. Uniformly sampling real utterances from speakers should provide data-diversity to help model train better and faster ideally.
- TTS data and varying number of utterances per speaker in real data
Similar to the above experiment, we keep the TTS data as fixed, but now vary the number of utterances per speaker while keeping the speaker count as fixed. Instead of increasing diversity of speakers, we aim to see how many utterances per speaker helps the model.

6. Experimental Results

6.1. Baseline KWS models with simple mixing options

Table 4 shows the evaluation results of models trained with different combinations of training data. We show FRR’s at a

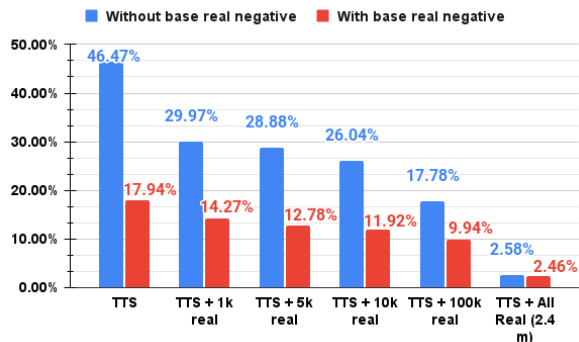


Figure 3: FRR over different amounts of real data. Blue bars indicate the baseline configs, and red bars have additional base negative data. Medium sized model used here.)

fixed false accept rate (section 5.2). TTS data trained FRR’s are high (46.47%) compared to real data trained baseline FRR (3.17%).

However, we observe that adding real negative data (~11 M) to all sweep experiments improves FRR numbers of all TTS based models dramatically (the second half of Table 4). Based on the observation and, we decide to keep using real negative data as *base* negative data for all other experiments. We also see that mixing all TTS and real data gives the best results.

Table 4: Model baselines trained on different datasets

Baseline models per train data	FRR
Virtuoso data only	53.10%
AudioLM data only	46.50%
TTS (Virtuoso + AudioLM)	46.47%
Real data only	3.17%
Improved baselines models per train data	FRR
Virtuoso + <i>base real negative data</i>	17.75%
AudioLM + <i>base real negative data</i>	16.59%
TTS + <i>base real negative data</i>	17.94%
TTS + Real data	2.46%

6.2. TTS + Incremental amounts of real positive data

In this setup, we train models with all the TTS data and gradually increasing amounts of real positive data (randomly sampled) as shown in Fig. 3. The blue bars show the FRR numbers of the models trained without real negative data, while the red bars show FRR numbers of models trained with *base* real negative data. Blue and Red bars show similar improvements when real negative data is added as *base* training data.

In Fig. 3 we see that the model FRR’s improves (decreases) monotonically, as we add more real *positive* data to the training data. From the Fig., we can observe that all the real negative data improved FRR number from 46.7% down to 17.94% with TTS only baseline. On top of that adding all the real positive data to TTS only baseline improved FRR from 17.94% to 2.46%. We can conclude that both negative and positive real data have a significant impact to the TTS only baseline model. And we note that about 100k of real positive data samples with *base* real negative data and TTS data gives 9.94% FRR, which is about ~3 times the FRR of real data only baseline (3.17%).

6.3. TTS + Real data with varying number of speakers

In this setup, we use all the TTS data with the real negative data as the base training data, and mixes small amounts of real positive data that has increasing number of speakers. We sample fixed number (10) of utterances per each speaker. The first half of Table 5 shows the result, where the FRR numbers improves as the number of speakers in the real positive data grows. We observe that models trained with number of speakers 100 or higher, gives performance with FRR similar to or less than 3 times the FRR of the baseline (3.17%). Note that the baseline model is trained with 3.8M real positive data, while the TTS+100 speaker model was trained with only 1k real positive utterances.

Table 5: Speaker variation experiments. First we increase speaker count while fixed utterances per speaker to 10, and second, we fix the speaker count to 100, and increase utterances per speaker. All exp uses full TTS data and base negative data.

Speakers with 10 utterances each	FRR
TTS + 1 Speaker (10 utts)	15.28%
TTS + 10 Speakers (100 utts)	14.94%
TTS + 100 Speakers (1k utts)	9.78%
TTS + 200 Speakers (2k utts)	9.90%
TTS + 500 Speakers (5k utts)	7.63%
Speaker count fixed at 100	FRR
TTS + 2 utts/speaker (200 utts)	10.99%
TTS + 6 utts/speaker (600 utts)	10.95%
TTS + 12 utts/speaker (1.2k utts)	10.71%
TTS + 20 utts/speaker (2k utts)	9.47%
TTS + 200 utts/speaker (20k utts)	7.99%

6.4. TTS + Real data with varying number of utterances per speaker

As a complimentary experiment to section 6.3, we keep the speaker count constant (100) and increase the number of utterances per speaker with the real positive data. Second half of Table 5, shows that increasing number of utterances improves FRR relatively *slowly* given fixed number of speakers. From this observation, we can conclude that increasing diversity of speakers as the previous sweep has more impact than increasing number of utterances with fixed speaker count.

7. Conclusion

We explore and evaluate the benefits of TTS synthesized data for training keyword spotting models. To maximize diversity of generated data, we applied prosody controlled text generation, and used TTS generation conditioned by audio samples. Finally we evaluated various mixing conditions of real data that ranges from zero to a scale, where we evaluated the effects of number of speakers, and number of utterances per speakers. Experiments showed that we are able to obtain reasonably high accuracy (~ 3 times FRR of baseline) with a fractional number of real positive samples (2k) compared to the baseline (3.8M).

8. Acknowledgements

The authors would like to acknowledge the support from Charles Yoon, Pedro Meningbar, Bhuvana Ramabhadran, and Françoise Beaufays.

9. References

- [1] S. Panchapagesan, M. Sun, A. Khare, S. Matsoukas, A. Mandal, B. Hoffmeister, and S. Vitaladevuni, "Multi-task learning and weighted cross-entropy for DNN-based keyword spotting," in *Interspeech*, 2016.
- [2] S. Team, "Hey Siri: An On-device DNN-powered Voice Trigger for Apple's Personal Assistant," <https://machinelearning.apple.com/2017/10/01/hey-siri.html>, Apple Inc., 2017, accessed: 2018-10-06. [Online]. Available: <https://machinelearning.apple.com/2017/10/01/hey-siri.html>
- [3] R. Alvarez and H. J. Park, "End-to-end Streaming Keyword Spotting," *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6336–6340, 2019.
- [4] G. Chen, C. Parada, and G. Heigold, "Small-footprint keyword spotting using deep neural networks," in *ICASSP*, 2014.
- [5] M. Sun, D. Snyder, Y. Gao, V. K. Nagaraja, M. Rodehorst, S. Panchapagesan, N. Strom, S. Matsoukas, and S. Vitaladevuni, "Compressed time delay neural network for small-footprint keyword spotting," in *INTERSPEECH*, 2017.
- [6] H. J. Park, P. Violette, and N. Subrahmanya, "Learning to detect keyword parts and whole by smoothed max pooling," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7899 – 7903, 2020.
- [7] A. Gruenstein, R. Álvarez, C. Thornton, and M. Ghodrati, "A cascade architecture for keyword spotting on mobile devices," *ArXiv*, vol. abs/1712.03603, 2017.
- [8] Y. He, T. N. Sainath, R. Prabhavalkar, I. McGraw, R. Alvarez, D. Zhao, D. Rybach, A. Kannan, Y. Wu, R. Pang, Q. Liang, D. Bhatia, Y. Shanguan, B. Li, G. Pundak, K. C. Sim, T. Bagby, S.-Y. Chang, K. Rao, and A. Gruenstein, "Streaming end-to-end speech recognition for mobile devices," *ICASSP 2019*, pp. 6381–6385, 2018.
- [9] A. Gruenstein, R. Alvarez, C. Thornton, and M. Ghodrati, "A cascade architecture for keyword spotting on mobile devices," in *31st Conference on Neural Information Processing Systems (NIPS 2017)*, 2017. [Online]. Available: <https://arxiv.org/abs/1712.03603>
- [10] Y. He, R. Prabhavalkar, R. K., W. Li, A. Bakhtin, and I. McGraw, "Streaming small-footprint keyword spotting using sequence-to-sequence models," *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 474–481, 2017.
- [11] T. Sainath and C. Parada, "Convolutional neural networks for small-footprint keyword spotting," in *Proceedings of Annual Conference of the International Speech Communication Association (Interspeech)*, 2015, pp. 1478–1482.
- [12] S. Panchapagesan, M. Sun, A. Khare, S. Matsoukas, A. Mandal, B. Hoffmeister, and S. Vitaladevuni, "Multi-task learning and weighted cross-entropy for DNN-based keyword spotting," in *INTERSPEECH*, 2016.
- [13] S. Xue, J. Tang, and Y. Liu, "Improving speech recognition with augmented synthesized data and conditional model training," *2022 13th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pp. 443–447, 2022. [Online]. Available: <https://api.semanticscholar.org/CorpusID:256669482>
- [14] Z. Chen, A. Rosenberg, Y. Zhang, G. Wang, B. Ramabhadran, and P. J. Moreno, "Improving speech recognition using gan-based speech synthesis and contrastive unspoken text selection," in *Interspeech*, 2020. [Online]. Available: <https://api.semanticscholar.org/CorpusID:226206599>
- [15] Y. Deng, R. Zhao, Z. Meng, X. Chen, B. Liu, J. Li, Y. Gong, and L. He, "Improving rnn-t for domain scaling using semi-supervised training with neural tts," in *Interspeech*, 2021. [Online]. Available: <https://api.semanticscholar.org/CorpusID:239659443>
- [16] Z. Huang, G. Keren, Z. Jiang, S. Jain, D. Goss-Grubbs, N. Cheng, F. Abtahi, D. Le, D. Zhang, A. D'Avirro, E. Campbell-Taylor, J. Salas, I.-E. Veliche, and X. Chen, "Text generation with speech synthesis for asr data augmentation," *ArXiv*, vol. abs/2305.16333, 2023. [Online]. Available: <https://api.semanticscholar.org/CorpusID:258947654>
- [17] K. D. Yang, T. Yao Hu, J.-H. R. Chang, H. S. Koppula, and O. Tuzel, "Text is all you need: Personalizing asr models using controllable speech synthesis," *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023. [Online]. Available: <https://api.semanticscholar.org/CorpusID:257766638>
- [18] A. Rosenberg, Y. Zhang, B. Ramabhadran, Y. Jia, P. Moreno, Y. Wu, and Z. Wu, "Speech recognition with augmented synthesized speech," in *2019 ASRU*. IEEE, 2019, pp. 996–1002.
- [19] Z. Chen, Y. Zhang, A. Rosenberg, B. Ramabhadran, P. Moreno, and G. Wang, "Tts4pretrain 2.0: Advancing the use of text and speech in ASR pretraining with consistency and contrastive losses," in *Proc. ICASSP*, 2022.
- [20] J. Lin, K. Kilgour, D. Roblek, and M. Sharifi, "Training keyword spotters with limited and synthesized speech data," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7474–7478, 2020. [Online]. Available: <https://api.semanticscholar.org/CorpusID:211021054>
- [21] A. Werchniak, R. Barra-Chicote, Y. Mishchenko, J. Droppo, J. Condal, P. Liu, and A. Shah, "Exploring the application of synthetic audio in training keyword spotters," *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7993–7996, 2021. [Online]. Available: <https://api.semanticscholar.org/CorpusID:233272947>
- [22] T. Saeki, H. Zen, Z. Chen, N. Morioka, G. Wang, Y. Zhang, A. Bapna, A. Rosenberg, and B. Ramabhadran, "Virtuoso: Massive multilingual speech-text joint semi-supervised learning for text-to-speech," *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2022. [Online]. Available: <https://api.semanticscholar.org/CorpusID:253157880>
- [23] T. Saeki, G. Wang, N. Morioka, I. Elias, K. Kastner, A. Rosenberg, B. Ramabhadran, H. Zen, F. Beaufays, and H. Shemtov, "Extending multilingual speech synthesis to 100+ languages without transcribed data," *ArXiv*, vol. abs/2402.18932, 2024. [Online]. Available: <https://api.semanticscholar.org/CorpusID:268063592>
- [24] Z. Borsos, R. Marinier, D. Vincent, E. Kharitonov, O. Pietquin, M. Sharifi, D. Roblek, O. Teboul, D. Grangier, M. Tagliasacchi, and N. Zeghidour, "Audiolm: A language modeling approach to audio generation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2523–2533, 2022. [Online]. Available: <https://api.semanticscholar.org/CorpusID:252111134>
- [25] E. Kharitonov, D. Vincent, Z. Borsos, R. Marinier, S. Girgin, O. Pietquin, M. Sharifi, M. Tagliasacchi, and N. Zeghidour, "Speak, read and prompt: High-fidelity text-to-speech with minimal supervision," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 1703–1718, 2023.
- [26] C. Kim, A. Misra, K. Chin, T. Hughes, A. Narayanan, T. Sainath, and M. Bacchiani, "Generation of large-scale simulated utterances in virtual rooms to train deep-neural networks for far-field speech recognition in google home," *INTERSPEECH 2017*, pp. 379–383, 2017.
- [27] "Google's privacy principles," <https://googleblog.blogspot.com/2010/01/googles-privacy-principles.html>, accessed: 2022-10-17.
- [28] "Artificial intelligence at Google: Our principles," <https://ai.google/principles>, accessed: 2022-10-17.
- [29] C. Kim, E. Varnani, A. Narayanan, and M. Bacchiani, "Efficient implementation of the room simulator for training deep neural network acoustic models," *CoRR*, vol. abs/1712.03439, 2017. [Online]. Available: <http://arxiv.org/abs/1712.03439>