Strategies for developing a conversational speech dataset for Text-To-Speech Synthesis

Adigwe, Adaeze

International Speech Communications Association
2022

Adigwe, A & Klabbers, E 2022, 'Strategies for developing a conversational speech dataset for Text-To-Speech Synthesis', Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, vol. 2022-September, pp. 2318-2322. https://doi.org/10.21437

http://hdl.handle.net/10138/355892
10.21437/Interspeech.2022-10802

publishedVersion

Interspeech 2022
18-22 September 2022, Incheon, Korea

Strategies for developing a conversational speech dataset for Text-To-Speech Synthesis

Adaeze Adigwe[1,2], *Esther Klabbers*[2]

[1]University of Helsinki, Finland
[2]ReadSpeaker, the Netherlands

adaeze.adigwe@helsinki.fi, esther.judd@readspeaker.com

Abstract

There have been many efforts to improve the quality of speech synthesis systems in conversational AI. Although state-of-the-art systems are capable of producing natural-sounding speech, the generated speech often lacks prosodic variation and is not always suited to the task. In this paper, we examine dialogue data collection methods to use as training data for our acoustic models. We collect speech using three different setups: (1) Random read-aloud sentences; (2) Performed dialogues; (3) Semi-Spontaneous dialogues. We analyze prosodic and textual properties of the data collected in these setups and make some recommendations to collect data for speech synthesis in conversational AI settings.

Index Terms: conversational text-to-speech, speaking styles, prosody, speech corpus

1. Introduction

Despite conversational AI systems gaining traction across various applications, they still largely fall short of the user's expectations because of their inability to realize conversational fluency that is intrinsic to communication between humans. Text-to-Speech (TTS) systems play a major role in computer voice interaction with humans, however off-the-shelf general purpose TTS systems are very limited in their ability to naturally synthesize and yield the vast characteristics of spoken conversational interactions [1]. A text-to-speech model that is contextually-aware and able to generate appropriate prosody will be a preferable system for conversational text-to-speech when compared to current models. One way to improve the quality of conversational text-to-speech is to record a speech dataset that has a good coverage of the prosodic devices that are utilized naturally in conversations.

The conventional way of developing a speech corpus for TTS is to record read-aloud speech from a single speaker in a professional studio. In most cases, the voice talents are instructed on the speaking style that they should assume when reading-out the transcript material. This method ensures clean and high quality speech as well as a phonetically balanced corpus as the transcript material is selectively curated beforehand to ensure sufficient phonetic coverage of the language. However, when recording data sets, recording setups are likely to pose trade-offs between control and naturalness over the recorded material [2]. When developing speech data sets, depending on the extent of control in the recording setup, many speculate that the naturalness of the recorded speech declines even in instances where the talents are asked to portray a given speaking style, as such prompted methods may lead to exaggeration [3]. Additionally, previous corpus-based studies have highlighted significant prosodic differences between read and spontaneous/naturally occurring speech [4, 5]. Unlike general purpose TTS, conversational TTS voices should ideally emulate speaking styles that are closer to natural occurring and spontaneous speech.

In this paper, we are motivated by the question: what are the considerations for developing conversational speech corpora for text-to-speech synthesis and how to strike a balance between control and perceived naturalness of the recording material? We hypothesize that a traditional read-aloud single speaker recording method is not suitable for conversational speech data collection as it does not reflect the dynamics of dialogues interactions. To this end, we are interested in investigating the differences between speech produced in three proposed recording setups by analyzing the acoustic-prosodic differences and the synthesized speech per setup. Specifically, we are interested in answering the following questions: (Q1) Is there a difference between speech collected in a multi-speaker interactive setting compared to single-speaker settings?; (Q2) Are performing dialogues by two voice talents able to simulate dialogues that are prosodically similar to actual natural-occurring conversations or are their features still edging towards read-aloud speech?

2. Related Work

Growing research interest in expressive speech synthesis for conversational speech has led towards efforts solving two sub-problems. On the one hand, recent research has looked into building context-aware neural TTS models by incorporating and conditioning contextual information such as audio and linguistic features during training [6, 7]. This has also led to newly proposed evaluation paradigms, which aim to move from rating naturalness of speech in isolation to rating appropriateness of speech in context [8]. On the other hand, research in conversational speech synthesis has also led to various approaches towards either recording conversational speech corpora or selectively utilizing existing found conversational data.

We outline some of the various approaches that have been used to collect conversational speech data. Following the conventional read-aloud recording method, researchers in [9] developed a conversational TTS corpus for a US English male speaker, RyanSpeech, using transcript material collected from various conversational domains. Considering the perceived difficulty of a voice talent to enact conversational material with natural conversational prosody within a single-speaker setting, researchers in [6] proposed an approach that extended the recording setup to two speakers enacting dialogues in context. This approach was used to collect a Mandarin Chinese conversational speech data set of task-oriented dialogues where two female voice talents played the roles of a conversational voice agent and a user. The actors were not constrained to the transcript material and were free to modify content and insert other spontaneous behaviors as they saw fit. The third strategy

10.21437/Interspeech.2022-10802

is to utilize found naturally occurring conversational speech, this lifts off the control exerted by recording setups and so is perceived to have high naturalness and adequately employed conversational prosody. Researchers in [10] were successfully able to extract clean single speaker conversational speech from found podcast data to train a TTS model. The drawback with found data is that it is often ill-suited for text-to-speech synthesis, as there are frequent occurrences of disfluencies, restarts, repetitions, fillers, fragmented sentence structures and overlaps between interlocutors. This makes the post-processing of the data a very tedious process. Our experimental setup aims to collect conversational data related to all three aforementioned setups and analyse the differences between all three.

Section 3 details the three setups for data collection. Section 4 discusses the analysis of the differences between the three setups in terms of prosody and acoustics. In Section 4.7 we discuss the implications for collecting conversational speech data for TTS.

3. Corpus Design

Our experiment involves collecting conversational speech from voice talents across three recording procedures as illustrated in Figure 1.

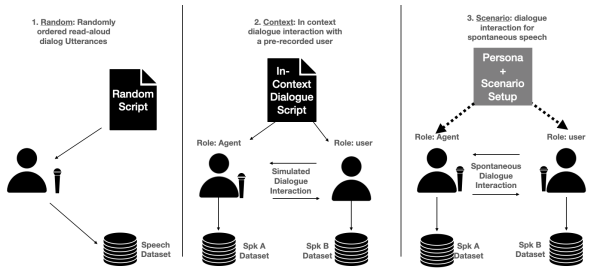

Figure 1: *Overview of the three recording setups 1) Read-Aloud Single Speaker Setting 2) Performing Dialogues 3) Semi-Spontaneous Dialogues*

3.1. Transcript

For this experiment we mainly focused on task-oriented dialogues and we utilized existing dialogues from the Google Taskmaster dataset [11], online language learning websites and manually bootstrapped dialogues based on a complaints database. Examples of task-oriented dialogues include movie ticketing, ordering pizza, car appointment scheduling and filing fraud complaints. Additionally, we also included dialogues that assumed a measure of familiarity between the speakers, such as party invitations and dialogue between work colleagues. In total we gathered 23 dialogues to be recorded.

3.2. Recording Setups

The speech recordings consisted of three different setups: (1) Random read-aloud sentences; (2) Performed dialogues; (3) Semi-Spontaneous dialogues.

- ***Setup 1*** *Random Read-Aloud Sentences*: Each speaker separately recorded their part of the dialogue presented sentence-by-sentence in random order.
- ***Setup 2*** *Performed Dialogues*: Both speakers, assuming their role play, enacted their dialogue parts from a shared screen in context. This setup was designed to be similar to the proposed approach in [6]. Similarly, the voice talents were free to insert spontaneous behaviours or modify the scripts in a way they thought would enable them sound more conversational in the given context.
- ***Setup 3*** *Semi-Spontaneous Dialogues*: To elicit spontaneous speech, the voice talents were presented first with a scenario which described the scene for the dialogue context, as well as persona narratives for the agent/user roles beforehand. This approach was derived from [12] for eliciting emotional speech. Secondly, the data collector's screen also displayed cue cards with words or phrase prompts related to the intent of the turn in the dialogue to help the speakers co-create the dialogues easily. The dialogues in this setup were based on dialogues from Setup 2 which was recorded prior to Setup 3.

3.3. Corpus Recording

The corpus was recorded with two professional voice talents, a male and female speaker, both native speakers of British English with a previous working relationship with each other. In task-oriented dialogues, the male actor assumed the role of the agent while the female actor assumed the role of the user. The decision was made because the male speaker had already recorded a speech database with us for use in TTS, so this will aid future synthesis experiments. Both actors were compensated for their services. After the recording session, both actors were asked to fill out a survey describing their individual experience and preferences with all three different recording setups.

The recording sessions were administered via a video conferencing platform, during which each voice talent was situated in their own professional recording studio and recorded their part of the conversations on their own computer. To simulate an exclusively audio recording sessions, the voice talents were asked to turn off their videos. Both speakers could see the data collector's shared screen which displayed the dialogue transcript material designed for the different recording setups. The sequence of the recording session was as follows: Setup 2: Performed dialogues followed by Setup 3: Semi-Spontaneous dialogues then lastly Setup 1: Random Read-Aloud sentences.

4. Analysis

Previous research shows that a single feature is insufficient to distinguish between speaking styles but many diverse aspects can be analyzed to differentiate between styles [13]. In our analysis we utilize features covering acoustic, phonetic and prosodic aspects of speech produced solely by the male speaker in our corpus, resulting in a total of 869 sentences consisting of 264, 290 and 313 sentences for setup 1, 2 and 3 respectively. Each interlocutor's turn is treated as a sentence in our analysis. A one-way ANOVA test showed that the features analyzed per setup are statistically significant different with $p<0.002$.

4.1. Speech Timing

We extracted durational features from all the sentences in the corpus. In the calculations, we excluded segments of spontaneous behaviours such as fillers (*uhm, eh*) from the sentences as we feared they might influence the results. The features extracted per sentence include speaking rate (phonemes/sec), Sentence duration, and pause-to-speech ratio across the three different setups which are plotted in Fig 2

In previous literature, reported speaking rates for different speaking styles were not always consistent in corpus anal-

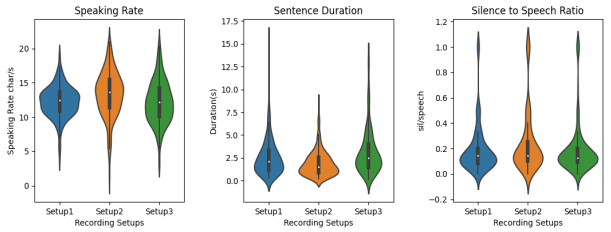

Figure 2: *Speaking Rate, Sentence Duration, Pause-to-speech ratio by recording setup*

ysis [14]. For our corpus, in Setup 1, we take into account that the voice actor's focus in that exercise is to ensure that they are articulating words clearly and are deliberately audible when recording the data in this format therefore leading to a lower speaking rate. This is likely to differ in Setup 2, where both actors are performing dialogues and both of them can read the full dialogue on the screen at the same time so communicating clearly might not be the speaker's focus. Likewise, in Setup 3, speakers generate dialogues semi-spontaneously with cues, this setup is likely to account for a slower speech rate as the speakers are constructing sentences in the dialogues each turn. Violin plots of sentence duration shows setup 1 and 3, with longer tails and wider distribution in comparison to setup 2 which has more sentences less than 2.5 seconds long. Lastly we observe longer tails in the violin plots of silence-to-speech ratio in setup 2 and 3 compared to setup 1, potentially indicating a more paced-speaking style in a multi-speaker setups than in the single-speaker setup.

4.2. Intonation

The F_0 values were extracted using Praat's python wrapper implementation [15]. We set the floor and ceiling frequency values to 75Hz and 400Hz respectively. Utilizing an approach implemented in [16] we filtered out frequency values that were outside 1.5 times the interquantile range for the speaker. The extracted intonational features include F_0mean, F_0std, and F_0range and are plotted in Figure 3. Comparable to the analyses found in [14, 16], the F_0 range declined across all three setups, with the read-aloud sentences in Setup 1 having the highest F_0 range and the semi-spontaneous having the lowest F_0 mean and range values as shown in Figure 1. Previous analysis in [16] cites lack of turn-taking organization in setup 1 as reason for larger F_0 range. However, we do observe longer tails F_0 range and F_0 mean in setup 3, likely accounting for more expressive sentences in the semi-spontaneous dialogues. In general, we expect more intonational variation in setup 3 compared setups 1 and 2.

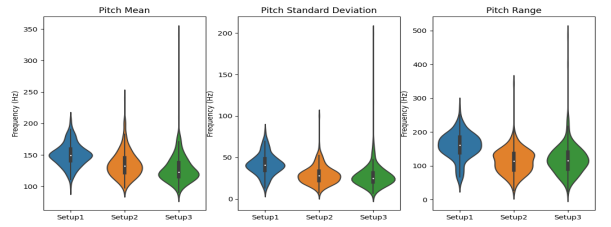

Figure 3: *F_0 mean, F_0 std, F_0 range by recording setups*

4.3. Pitch Decomposition

Pitch contours can give us a lot of information about intonational phenomena observed in the speech. The previous paragraph has shown some interesting differences in F_0 mean and range between the different setups. But there is more information to be extracted from F_0 contours. Using ideas from the superpositional intonation model developed by [17] we have devised a simple strategy to subtract a phrase curve from the F_0 contour, thus leaving the residual F_0 contour. The phrase curve is modeled by a straight line for the entire phrase, such that no F_0 values fall below the phrase curve. The residual contour can be used to detect whether there is a rise at the end of a phrase or sentence. In [18] the F_0 contour is decomposed into phrase curves, accent curves, and continuation rise curves, but this method requires prior knowledge of the prosodic structure of the sentence, to indicate the location of accents and phrase boundaries. In our recordings, the locations of phrase boundaries are indicated in the label files, but accentuation is more challenging to detect automatically. One major difference between text material selected for general-purpose synthesis versus conversational synthesis is the presence of an adequate amount of questions. They tend to be underrepresented in standard TTS text material. In English, there are also intonational differences between questions that start with a wh-word such as where, why, what, who, and how and other questions. In Table 1 we have noted how many sentences in each setup were declarative, wh-questions or other questions and how many of those ended in a rise in F_0 (where the residual value is $>$ 15 Hz). As can be seen, wh-questions often end without a rise which corroborates earlier findings of English intonation [19]. But there is a difference in the amount of questions and wh-questions which end in a rise. Setups 2 and 3 both show more frequent rises in F_0 at the end of questions and wh-questions than Setup 1.

Setup	Decl.	Question	Wh-question
Setup 1	7/161 (4.34%)	20/52 (38.5%)	5/52 (9.6%)
Setup 2	19/190 (10%)	31/53 (59.6%)	12/51 (23.5%)
Setup 3	33/200 (16.5%)	24/50 (48%)	20/64 (31.3%)

Table 1: *Distribution of rises and nr of questions, wh-questions and declarative sentences in the three setups*

4.4. Textual Differences

We compared text, manually transcribed from speech collected in Setup 2 and Setup 3. We chose to analyse text differences because the transcript material used in Setup 2 were gathered mainly from text-based dialogues chats which were then used as transcript for a spoken communication setting. However, we believe there are fundamental differences in communicating through both mediums. When engaged in spontaneous talk, speakers have the full expressive power of spoken language to construct sentences with the help of prosodic devices to communicate a message, meaning that the same sentence can be produced in different ways, conveying various meaning.

We selected the subset of dialogues that were recorded in both Setup 2 and Setup 3. Next, we extracted features such as average turns per dialogues, word count per sentence and frequency of fillers in the dialogue. On average the turn lengths of dialogues for both setups is 20. The average length of an sentence in a dialogue are 11 words and 14 words for Setup 2 and Setup 3 respectively therefore Setup 3 tended to have a higher word count compared to Setup 2. Lastly, in Setup 3 the speaker used fillers in his speech, with the average fillers count

per dialogue being 2.5 compared to the absence of fillers found in Setup 2 dialogues.

4.5. Preliminary TTS Experiment

The goal of our corpus analysis is to recommend suitable methods for developing conversational speech corpora for TTS. To test the validity of the speech data collected in our experiment, we conducted a preliminary synthesis experiment using a multi-style speech synthesis model. We utilized a variant of the FastSpeech2 end-to-end speech synthesis architecture [20] and encoded the male speaker's speech data from each of the three setups as a style embedding. This was then added to the output vectors of the phoneme encoder for training. The training data also contained a larger set of neutral style training data from the same male UK English speaker. The weights for the styles were adjusted to account for the skewed distribution of the training material. We had more than 8000 neutral sentences versus approximately 275 sentences per setup. A vocoder based on MelGAN architecture [21] is used to reconstruct the waveform from the Mel-spectrogram features. Post-training, we projected the t-distributed stochastic neighbor embedding (t-SNE) derived from GST embeddings from each of the setups in our training data. Although the data set collected through this experiment is not very large, the plot in Figure 4 shows that even with small amount of conversational speech ($<$ 20 minutes per setup), the acoustic model is still capable of partially separating the perceived speaking style from all three conversational speech setups.

Figure 4: *T-SNE Clustering of GST embeddings of sentences from the all 3 recording setups*

There are noticeable perceived differences in the prosody of the synthesized speech from each of the three setups. We measured the F_0 mean and speaking rate and observed that they correlated with the representative features in the training data. For future research we plan to conduct a subjective evaluation with the synthesized sentences from all three setups.

4.6. Voice Talent Preferences

In the post-recording survey both talents rated the semi-spontaneous recording, Setup 3, as the best setup in helping them produce dialogue speech naturally. They also both commented that the cue-words were useful in helping them produce more natural responses whilst maintaining a dialogue structure. Interestingly both talents rated Setup 2, Performing Dialogues, as their best setup in terms of overall experience, followed by Setup 3, Spontaneous dialogues then lastly read-aloud setting. From this feedback we believe that producing dialogue speech in a multi-speaker setting, is most preferred by voice talents.

4.7. Discussion

In the previous paragraphs we examined linguistic and acoustic differences between three data collection methods for conversational speech synthesis. We found that the voice talents had the best overall experience using Setup 2 Performed Dialogues, while they felt that Setup 3 Semi-Spontaneous Dialogues gave them the best platform to produce naturally flowing dialogue speech. We found differences across setups in F_0 mean and range with the read-aloud setup having larger values. Setup 3 was the most challenging to annotate as it contained fillers and re-starts. A preliminary synthesis experiment showed that we can reliably synthesize the speaking style of each setup. The intonation analysis showed the importance of selecting diverse dialogue data containing a balanced mix of declarative sentences, wh-questions, and other questions.

5. Conclusion and Future Research

In this paper we have explored three setups for developing conversational speech corpora for speech synthesis. The three approaches were random read-aloud sentences, performed dialogues and semi-spontaneous dialogues. We analyzed prosodic and textual features extracted from sentences across all three setups and synthesized some samples using a multi-style speech synthesis model. Our analysis showed that performed dialogues have prosodic features more similar to spontaneous speech than that of read speech. By our analysis we recommend that multi-speaker settings are best suited for collecting clean conversational speech. However for speech data that truly mimics human speaking style, for instance in human-robot social interactions, spontaneous dialogues are likely more ideal for such use-cases.

In the future we intend to perform a subjective evaluation of the synthesized styles. While we have collected some meaningful information about the three setups, the findings also lead to more questions. While Setup 3 gave the voice talents the best platform for producing natural dialogue speech, it brings some challenges to the data preparation and training of a synthesizer. Setup 2 gave the best overall experience but given the fact that two speakers are needed to read a dialogue, it does mean double the work compared to Setup 1. However, having the data from both speakers from Setup 2 and 3 available gives us the chance to perform more analysis on the interaction between the two dialogue partners and evaluating the synthesis in context. We also intend to extend the synthesis framework with a VAE or similar module to allow for random but meaningful prosodic variation in the synthesis output.

We have tried to select a variety of dialogues in our recordings from customer complaints to pizza ordering to buying movie tickets. While we have separated out the three Setups as speaking styles, we have not looked at the intents behind the agents sentences. In customer complaint calls, there are certain sentences that call for an apologetic tone, whereas most other sentences could be considered as friendly. We didn't have enough material to analyze it to the fullest. We do believe however, that voice talents can more realistically evoke these intents if the text material is appropriate. Further research should shine more light on this issue.

6. Acknowledgements

The first author has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska Curie grant agreement No 859588. The authors are thankful to Maaike Groenewege, Johannah O'Mahony and ReadSpeaker's R&D team whose suggestions and discussions have been instrumental in shaping the direction of this paper.

7. References

[1] M. Marge, C. Espy-Wilson, N. G. Ward, A. Alwan, Y. Artzi, M. Bansal, G. Blankenship, J. Chai, H. Daumé III, D. Dey *et al.*, "Spoken language interaction with robots: Recommendations for future research," *Computer Speech & Language*, vol. 71, p. 101255, 2022.

[2] M. E. Beckman, "A typology of spontaneous speech," in *Computing prosody*. Springer, 1997, pp. 7–26.

[3] E. Douglas-Cowie, N. Campbell, R. Cowie, and P. Roach, "Emotional speech: Towards a new generation of databases," *Speech communication*, vol. 40, no. 1-2, pp. 33–60, 2003.

[4] G. P. Laan, "The contribution of intonation, segmental durations, and spectral features to the perception of a spontaneous and a read speaking style," *Speech Communication*, vol. 22, no. 1, pp. 43–65, 1997.

[5] J. Hirschberg and C. H. Nakatani, "A prosodic analysis of discourse segments in direction-giving monologues," in *34th Annual Meeting of the Association for Computational Linguistics*, 1996, pp. 286–293.

[6] H. Guo, S. Zhang, F. K. Soong, L. He, and L. Xie, "Conversational end-to-end tts for voice agents," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 403–409.

[7] P. Oplustil-Gallegos, J. O'Mahony, and S. King, "Comparing acoustic and textual representations of previous linguistic context for improving text-to-speech," in *Proc. 11th ISCA Speech Synthesis Workshop (SSW 11)*, 2021, pp. 205–210.

[8] J. O'Mahony, P. O. Gallegos, C. Lai, and S. King, "Factors affecting the evaluation of synthetic speech in context," in *The 11th ISCA Speech Synthesis Workshop (SSW11)*. International Speech Communication Association, 2021, pp. 148–153.

[9] R. Zandie, M. H. Mahoor, J. Madsen, and E. S. Emamian, "Ryanspeech: A corpus for conversational text-to-speech synthesis," *arXiv preprint arXiv:2106.08468*, 2021.

[10] É. Székely, G. E. Henter, J. Beskow, and J. Gustafson, "Spontaneous conversational speech synthesis from found data." in *INTERSPEECH*, 2019, pp. 4435–4439.

[11] B. Byrne, K. Krishnamoorthi, C. Sankar, A. Neelakantan, D. Duckworth, S. Yavuz, B. Goodrich, A. Dubey, K.-Y. Kim, and A. Cedilnik, "Taskmaster-1: Toward a realistic and diverse dialog dataset," 2019.

[12] F. Enos and J. Hirschberg, "A framework for eliciting emotional speech: Capitalizing on the actors process," in *First international workshop on emotion: Corpora for research on emotion and affect (international conference on language resources and evaluation (LREC 2006))*, 2006, pp. 6–10.

[13] M. Eskenazi, "Trends in speaking styles research," in *Third European Conference on Speech Communication and Technology*, 1993.

[14] M. Eskénazi, "Changing speech styles: Strategies in read speech and casual and careful spontaneous speech." 1992.

[15] Y. Jadoul, B. Thompson, and B. De Boer, "Introducing parselmouth: A python interface to praat," *Journal of Phonetics*, vol. 71, pp. 1–15, 2018.

[16] P. Wagner and A. Windmann, "Re-enacted and spontaneous conversational prosody—how different?" *Proceedings of Speech Prosody 2016*, pp. 518–522, 2016.

[17] J. P. Van Santen and B. Möbius, "A quantitative model of f_0 generation and alignment," in *Intonation*. Springer, 2000, pp. 269–288.

[18] M. E. Langarani, "A generalized intonation model for english," Ph.D. dissertation, Oregon Health & Science University, Department of Science & Engineering, 2020.

[19] J. Hirschberg, "A corpus-based approach to the study of speaking style," in *Prosody: Theory and experiment*. Springer, 2000, pp. 335–350.

[20] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," *arXiv preprint arXiv:2006.04558*, 2020.

[21] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brébisson, Y. Bengio, and A. C. Courville, "Melgan: Generative adversarial networks for conditional waveform synthesis," *Advances in neural information processing systems*, vol. 32, 2019.