INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Spontaneous conversational speech synthesis from found data

Éva Székely, Gustav Eje Henter, Jonas Beskow, Joakim Gustafson

Division of Speech, Music and Hearing, KTH Royal Institute of Technology, Stockholm, Sweden
{szekely,ghe,beskow,jkgu}@kth.se

Abstract

Synthesising spontaneous speech is a difficult task due to disfluencies, high variability and syntactic conventions different from those of written language. Using found data, as opposed to lab-recorded conversations, for speech synthesis adds to these challenges because of overlapping speech and the lack of control over recording conditions. In this paper we address these challenges by using a speaker-dependent CNN-LSTM breath detector to separate continuous recordings into utterances, which we here apply to extract nine hours of clean single-speaker breath groups from a conversational podcast. The resulting corpus is transcribed automatically (both lexical items and filler tokens) and used to build several voices on a Tacotron 2 architecture. Listening tests show: i) pronunciation accuracy improved with phonetic input and transfer learning; ii) it is possible to create a more fluent conversational voice by training on data without filled pauses; and iii) the presence of filled pauses improved perceived speaker authenticity. Another listening test showed the found podcast voice to be more appropriate for prompts from both public speeches and casual conversations, compared to synthesis from found read speech and from a manually transcribed lab-recorded spontaneous conversation.
Index Terms: Speech synthesis, conversational speech, spontaneous speech, hesitations, disfluencies, found data

1. Introduction

Conversational speaking systems are becoming ever more widespread. However, interaction quality is not reaching its full potential [1], possibly due to issues with the voice. Adapting read-speech voices for synthesising conversations is not straightforward [2] and it stands to reason that interactions might improve if dialogue systems were able to speak truly conversationally, rather than with voices based on written prompts read aloud.

While there has been some past work on building speech synthesisers from spontaneous speech audio [3, 4, 5, 6], this was restricted to small, hand-annotated corpora and statistical parametric speech synthesisers. Bigger corpora can usually be sourced from found speech recordings, but the output quality has been disappointing, e.g., [7]. Instead, the text-to-speech (TTS) field has hitherto concentrated on audiobook data as convenient source of transcribed single-speaker audio, e.g., [8, 9, 10, 11]. Such speech materials often feature acted expressive speaking styles [11], which are likely to to be beneficial for conversational applications [12], but ultimately, they are recordings of a speaker reading written text aloud. Using genuinely conversational speech audio found in the wild has seldom (if ever) been attempted for building a complete TTS system, and for good reason: Up until recently, necessary speech technologies such as ASR, forced alignment and the underlying TTS engines were not sufficiently robust to enable good-quality synthesis from such data. In other words, previous research has considered TTS from found data or spontaneous conversational corpora, but not from speech with both these characteristics.

Found conversational data tends to be messy, with unconventional sentence structures and overlapping speech. We have previously proposed [13] to address these issues through the use of a speaker-dependent breath detector, summarised in Sec. 2.2. We proposed using breath events for segmentation since speakers' respiratory patterns relate to the speech planning process in spontaneous conversations [14] and are highly correlated with major prosodic breaks [15] and turn-taking behaviour [16].

In this paper, we describe and evaluate neural seq2seq TTS from a large corpus of found, spontaneous conversational speech, originating from a podcast. We compare the resulting synthesiser to TTS trained on found (but read) speech recordings as well as spontaneous conversational (but lab-recorded and carefully transcribed) speech. Our experiments investigate the suitability of different speech corpora for synthesising prompts from different genres – in particular, whether TTS from automatically-transcribed, spontaneous conversational speech is appropriate for spontaneous spoken genres – as well as the impact of fluency on speaker perception and appropriateness.

2. Corpus

2.1. Found podcast data

The data we use in this study is an untranscribed weekly technology podcast, called the "ThinkComputers" podcast, available in the public domain via the Internet Archive (archive.org). The recordings contain product reviews and discussions of technology news from two male speakers of American English mixed into a single audio channel. At the time of writing, over 150 episodes are available online, each about an hour long. In this paper we selected episodes 129–158 (excluding eps. 137 and 149 because of recording quality issues, and 157 which was missing). Altogether 27 episodes (29 hours of recordings) were used. The audio was downloaded in Ogg Vorbis format (71 kbpbs at 48 kHz) and then converted to raw audio. We selected the speaker with the most air time for corpus building.

2.2. Segmentation

To segment the data into clean, well-defined utterances we used the speaker-dependent breath detection method proposed in [13]. It uses a convolutional-recurrent network on mel-spectrograms and zero-crossing rate, trained on a small amount of coarsely annotated seed data. This classifier proved effective in identifying breath events and speech segments for each individual speaker, as well as segments containing overlapping speech from both speakers. With this method we obtained 8,457 speech segments from the 27 episodes, starting with a breath event from the target speaker. For speech segments longer than 9 seconds without a final breath, we set the ending point at the location of the last silence of a minimum of 100 ms before reaching 9 s. Next, we reviewed the results of the automatic selection by listening to all extracted utterances, to exclude ones

http://dx.doi.org/10.21437/Interspeech.2019-2836

containing noise (e.g., typing), laughter and overlapping speech not spotted by the method, such as short backchannels. 229 segments were found to contain a missed breath in the middle, but because they were a concatenation of two short breath groups, we included them as is. In total, the final database comprised 6,218 breath groups, 549 min (9 h 9 min) of audio. We will henceforth refer to this as TCC, ThinkComputers Corpus.

2.3. Transcription

It is resource-demanding and challenging to transcribe 9 hours of jargon-laden speech by hand. Instead we used the Google Cloud Speech API [17], specifically the enhanced video model with automatic punctuation for US English. Weekly podcast topic keywords, found on archive.org for each episode, were added as 'phrase hints' to help identification of brand and product names and other technology related jargon.

As noted by [18], the Google Speech API generally omits hesitations such as filled pauses (*uh* and *um*). To gain control over filled pauses (FPs) in the TTS, and to distinguish fluent and disfluent segments, it was necessary to identify FPs in the data. This was done using IBM Watson Speech to Text using the US English BroadbandModel, since it has the option to include generic *Hesitation* tokens in the transcription. By combining this transcript with the output of the Gentle forced aligner [19], we were able to differentiate between *uh*, *um* and remaining disfluencies, and also put these tokens back in the Google API transcription, which we perceived to be slightly more accurate than the Watson output. In total, 49.7% of the TCC breath groups were annotated as containing at least one filled pause, with at most four FPs found in a single breath group.

3. TTS

All voices described in this paper were built using the TensorFlow implementation [20] of the Tacotron 2 spectrogram prediction framework [21]. Input files were sampled at 22.1 kHz. We used the hyperparameters described in [21], with the following changes: mel filterbank spanning 55 Hz to 5.5 kHz for male voices and 55 Hz to 7.6 kHz for the female voice. To optimise for training on 4 GPUs (10 GB memory each), a batch size of 48 was used and the number of frames generated at each decoding step was increased to two. For waveform synthesis we used the Griffin-Lim algorithm [22] with pre-emphasis [20].

4. Evaluation

We carried out three separate evaluations: first a more formal, annotation-based evaluation to assess pronunciation issues and conversational features of voices built on the TCC, then a MUSHRA-like listening test investigating the appropriateness of the voice on different spoken genres in comparison with other synthetic voices, and finally a preference test gauging the effect of FPs on the listeners' perception of the speaker. The results of these experiments are discussed and interpreted in Sec. 5.

4.1. Pronunciation and conversational characteristics

4.1.1. Voices and evaluation design

The first evaluation looked at the impact of transfer learning from read speech with transcriptions, and of phonetic versus grapheme-based input encoding. Using a similar methodology, we also assessed how the prevalence of certain spontaneous conversational style elements in the synthesis output was affected by removing disfluent utterances from the corpus.

To evaluate pronunciation performance, three voices were built using different settings. For two of the voices we made use of transfer learning by first pre-training a voice on the LJSpeech corpus [10] for 65k iterations, and then changing over to fine-tune on the TCC from this checkpoint. LJSpeech is a corpus of 13,100 utterances (approximately 24 h) from audiobooks read by a female speaker of American English. Two input modalities were considered: grapheme-level input and phoneme-level input obtained using the `g2p_en` front-end [23] with the CMU dictionary [24]. The following voices were compared:

POD-GT-FP	**G**rapheme-level input with **T**ransfer learning from LJSpeech; **FP**s transcribed
POD-PR-FP	**P**honeme-level input and **R**andom initialisation; **FP**s transcribed
POD-PT-FP	**P**honeme-level input with **T**ransfer learning from LJSpeech; **FP**s transcribed

We synthesised 40 sets of 10 phonetically balanced "Harvard sentences" [25] using each of the three voices. Two listeners assessed the number of mispronounced phones in the 400 sentences, counting only instances where both listeners agreed that the pronunciation error resulted from a possible encoder error and not from reduced pronunciations common in spontaneous speech.

To study the impact of training-data fluency on the speech output, we trained a fourth voice – called **POD-PT-FLU** for *fluent* – only on the most fluent segments of the TCC. Specifically, we used the 2,763 TCC breath groups (3 h 31 min) where no filled pauses and a maximum of 1 other disfluency was found by the transcription procedure in Sec. 2.3. For the 400 Harvard sentences as synthesised by POD-PT-FLU and POD-PT-FP, annotators were asked to note the number of deleted function words and the number of repeated syllables. In addition, they also counted two other style markers frequently occurring in conversational speech [26], namely the number of times the determiners 'the' and 'a' were pronounced with a non-reduced vowel (as 'thiy' and 'ei', following the notation of [27]) instead of the reduced schwa vowel. These four aspects were chosen because a high inter-rater agreement can be expected.

4.1.2. Results

The results of the pronunciation error evaluation are summarised in Tab. 1. Almost all mispronounced phones were vowels. Differences in pronunciation error rate were evaluated using a Z-test between samples. Following Bonferroni correction POD-PT-FP was found to have significantly lower pronunciation error rate than POD-GT-FP and POD-PR-FP ($p<0.001$). The difference between POD-GT-FP and POD-PR-FP was not significant.

The results of the annotation-based analysis of the conversational features are provided in Tab. 2. The voice trained on utterances not containing FPs had significantly fewer deletions ($p=0.024$) and repeated syllables ($p<0.001$). There was no significant effect on the number of non-reduced pronunciations of

Table 1: *Pronunciation assessment of 400 Harvard sentences.*

Voice	Pronunciation errors
POD-GT-FP	49
POD-PR-FP	43
POD-PT-FP	13

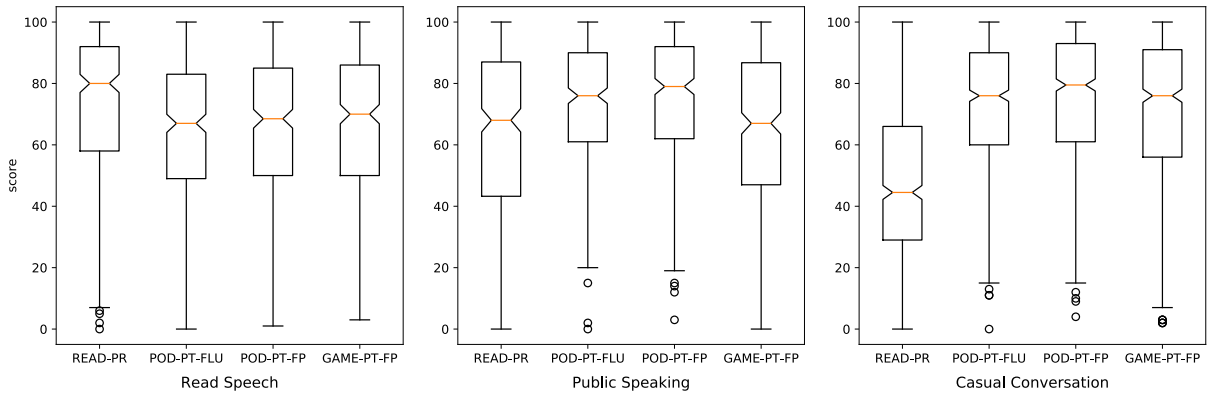

Figure 1: *Boxplot for evaluation of appropriateness of speaking style for read speech, public speaking and casual conversation.*

determiners *the* and *a*. We also measured the speech rate of these two voices on the 400 samples, finding that the fluent voice, POD-PT-FLU, had a significantly higher speech rate over POD-PT-FP (4.14 vs. 3.84 syllables per second; $p<0.001$) even though there were no FPs present in these prompts.

Table 2: *Disfluencies and conversational characteristics in 400 Harvard sentences.*

Voice	Deletions	Repeated syllables	'the' as 'thiy'	'a' as 'ei'
POD-PT-FP	15	32	40 / 422	18 / 98
POD-PT-FLU	5	6	57 / 422	13 / 98

4.2. MUSHRA-like listening test on genre appropriateness

4.2.1. Voices and evaluation design

The goal of this evaluation was to see how the podcast voice was perceived, compared with voices trained on read speech and on lab-recorded, manually annotated spontaneous conversational speech. Specifically, we assessed the appropriateness of these voices for synthesising speech in different spoken genres, namely audiobooks, public speaking and casual conversation. These three genres were selected to cover a range of registers (formal–informal) and requiring different degrees of speech planning, from scripted through planned to completely spontaneous. Four voices were included in this evaluation:

POD-PT-FP	TCC; **FP**s transcribed (as before)
POD-PT-FLU	**Flu**ent breath groups from TCC; FPs neither present nor synthesised (as before)
READ-PR	**P**honeme-input voice with **R**andom init trained on LJSpeech [10]; no FPs
GAME-PT-FP	Lab-recorded **P**honeme-input voice with **T**ransfer learning from LJSpeech; **FP**s transcribed

The GAME-PT-FP voice used 1,767 single-speaker breath groups (1 h 33 min) from a male speaker of Irish English playing a cognitively-challenging game requiring dialogue. This was recorded in a lab over separate channels, manually transcribed and segmented as described in [6]. The voice had noticeably better naturalness over the Merlin [28] voice in [6].

The stimulus prompts selected for this test were as follows:

Read speech	10 utterances originating from the Arctic Corpus [29]
Public speaking	10 utterances transcribed from political speeches and keynote talks; 7 of the prompts contain FPs
Casual conversation	20 utterances from a corpus of casual spontaneous conversations collected for the purposes of speech synthesis by [4]; 11 of the prompts contain FPs

The prompt list and the experiment stimuli are available under http://www.speech.kth.se/tts-demos/. The test utterances were chosen such that the genre would be apparent to the listener from the words alone, e.g., "Nevertheless we found ourselves once more in the high seat of abundance" (read), "And once again uh thank you it's an honour to receive the medal" (public speaking), "But um yeah we were just kind of driving around and" (casual conversation). We used 20 instead of 10 prompts from casual spontaneous conversation to better cover various reduced forms, discourse markers and syntactic forms not following the conventions of written language.

We carried out a MUSHRA-like listening test with these prompts using a modified version of WebMUSHRA [30]. Participants were recruited through Prolific Academic[1] and could listen to and rate the same prompt as spoken by the four different systems in parallel. The order of the prompts and systems were randomised for each listener. Participants were asked to rate the stimuli based on *how well the speaking style matched the content of the utterance*, on a scale from 0 (Bad) to 100 (Excellent), while ignoring differences of gender and accent.

4.2.2. Results

34 subjects completed the experiment, taking on average 22 minutes to complete it. All participants reported having used headphones for the test.

The systems were compared in pairs using a Wilcoxon signed-rank test with Bonferroni correction, for a total of 6 pairwise comparisons per genre. For read speech, READ-PR was rated significantly above each of the spontaneous voices ($p<0.001$ for each).

For public speaking, there was no significant difference between READ-PR and GAME-PT-FP, but both these voices

[1] https://prolific.ac/

were rated significantly below the POD voices ($p<0.001$ for each). Again, no significant difference was found between the POD-PT-FLU and POD-PT-FP.

For casual conversation prompts, the READ-PR voice was rated significantly lower than each of the voices built from spontaneous speech corpora ($p<0.001$ for each). The POD-PT-FP voice was also rated significantly above the GAME-PT-FP voice ($p<0.001$), but there was no significant difference between the POD voices with and without FPs. Looking at only the prompts containing actual FPs (11 in the casual conversation and 7 in the public speaking genres) did not seem to influence the results.

4.3. Preference tests on the impact of filled pauses

4.3.1. Voices and evaluation design

To further investigate the effect of filled pauses on the perception of synthetic voices, we carried out two pairwise preference tests between the POD-PT-FP and POD-PT-FLU voices. The prompts were transcriptions of the first 20 and 22 breath groups of two Interspeech keynote speeches, concatenated into chunks of 5 and 3 breath groups, respectively, between 10 and 27 seconds long. The reason for this concatenation was to enable listeners to form an overall impression of the speaker without being distracted by specific prosodic realisations of individual utterances, without overloading their auditory memory with long speech samples. The POD-PT-FP prompts included 33 filled pauses, of which 16 *um*s and 17 *uh*s.

We conducted two separate experiments on the Figure-Eight crowd-sourcing platform[2], one asking "Which speaker sounds more engaging?" and another asking "Which speaker sounds more authentic?" Listeners were given short descriptions of the concepts (e.g., engaging means a speaker who is easy to listen to; authentic refers to a speaker who is presenting their own genuine content, not acting or reading). It was possible to answer "They both sound the same".

4.3.2. Results

Two groups of 25 native English speakers each participated in the listening test, taking an average of 10 minutes to complete it. Tests showed no clear listener preference for one voice over the other in terms of engagement. However, a significant majority of the listeners found the POD-PT-FP voice to sound more authentic, see Fig. 2.

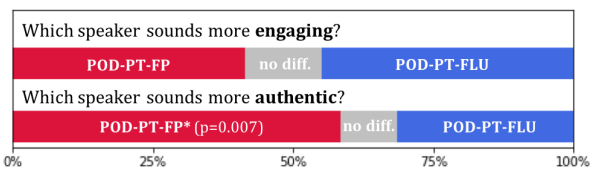

Figure 2: *Preference test results. p-values were calculated using the exact binomial test on the null hypothesis that the fluent voice is more engaging (thus adding no-preference votes to its count).*

5. Discussion

The results of the pronunciation evaluation show that the combination of phoneme-level input and initialising the synthesiser on a model of a voice trained on read speech (with transcriptions) via transfer learning reduced the number of pronunciation errors significantly. It is not straightforward to tell to what degree remaining pronunciation errors are attributable to the use of automated transcriptions. We also saw that training a voice on only the fluent parts of the corpus results in a significant decrease of deleted function words and repeated syllables, but does not seem to affect the number of non-reduced pronunciation of determiners. This is in line with the findings of [27], who have shown that human speakers not only choose to say 'thiy' to signal imminent problems of the speech planning, but also as a strategy to maintain as much fluency as possible. Together with the higher speech rate identified for POD-PT-FLU, we conclude that this version of the voice is representative of a faster, more fluent speaking style but still retains conversational characteristics in the synthesised speech. This is confirmed by the perceptual evaluation in 4.2, where no significant differences were seen between the performance of POD-PT-FP and POD-PT-FLU, despite a high number of filled pauses in the evaluation data.

As for the effect of fluency on the impression of the speaker, the pairwise listening test did not indicate that either of the podcast voices sounded more engaging than the other. However, the presence of FPs seemed to increase the perception of authenticity. This suggests that – as long as they are correctly placed and with appropriate prosody – one can insert filled pauses into synthetic speech in order to make it come across as more authentic, without sounding less engaging or less appropriate.

In the subjective evaluation of different genres, the system trained only on read speech was rated as significantly more appropriate for the read speech genre, while the podcast voices scored the highest in both of the other two categories, public speaking and casual conversation. This supports our central hypothesis that training on spontaneous and conversational speech data is beneficial for appropriately synthesising these genres.

6. Conclusions

To summarise, we have shown that, using our new approach, spontaneous conversational speech synthesis from automatically transcribed found data is able to outperform both read-speech synthesis from found data, and TTS from spontaneous conversational speech recorded in a lab and manually transcribed; in terms of appropriateness for spoken genres in two out of three categories. As far as we are aware, this has not been achieved before.

Conversational podcasts appear to be a promising data source for spontaneous speech synthesis with versatile characteristics that are appropriate for synthesising both monologues and conversational speech. Our approach unlocks *genuine* and authentic speech recordings for speech synthesis applications, enabling TTS to move beyond data that is read, prompted, acted or however else elicited in a lab. Our hope is that using recordings of genuine speech will be a new focus area of speech synthesis and we are excited to learn what differences these voices might make for, e.g., dialogue systems and performative applications of synthetic speech [31].

7. Acknowledgements

This research was supported by the Swedish Research Council Project Incremental Text-To-Speech Conversion VR (2013-4935) and by the Swedish Foundation for Strategic Research project EACare (RIT15-0107).

[2]https://www.figure-eight.com/

8. References

[1] L. Clark, N. Pantidi, O. Cooney, P. Doyle, D. Garaialde, J. Edwards, B. Spillane, C. Murad, C. Munteanu, V. Wade, and B. R. Cowan, "What makes a good conversation? challenges in designing truly conversational agents," in *Proc. CHI*, 2019.

[2] M. Wester, O. Watts, and G. E. Henter, "Evaluating comprehension of natural and synthetic conversational speech," in *Proc. Speech Prosody*, 2016, pp. 766–770.

[3] S. Andersson, J. Yamagishi, and R. A. J. Clark, "Synthesis and evaluation of conversational characteristics in HMM-based speech synthesis," *Speech Commun.*, vol. 54, no. 2, pp. 175–188, 2012.

[4] R. Dall, "Statistical parametric speech synthesis using conversational data and phenomena," Ph.D. dissertation, School of Informatics, The University of Edinburgh, Edinburgh, UK, 2017.

[5] T. Nagata, H. Mori, and T. Nose, "Dimensional paralinguistic information control based on multiple-regression HSMM for spontaneous dialogue speech synthesis with robust parameter estimation," *Speech Commun.*, vol. 88, pp. 137–148, 2017.

[6] É. Székely, J. Mendelson, and J. Gustafson, "Synthesising uncertainty: the interplay of vocal effort and hesitation disfluencies," *Proc. Interspeech 2017*, pp. 804–808, 2017.

[7] J. Lorenzo-Trueba, F. Fang, X. Wang, I. Echizen, J. Yamagishi, and T. Kinnunen, "Can we steal your vocal identity from the Internet?: Initial investigation of cloning Obama's voice using GAN, WaveNet and low-quality found data," in *Proc. Odyssey*, 2018, pp. 240–247.

[8] S. King and V. Karaiskos, "The Blizzard Challenge 2012," in *Proc. Blizzard Challenge Workshop*, 2012.

[9] ——, "The Blizzard Challenge 2016," in *Proc. Blizzard Challenge Workshop*, 2016.

[10] K. Ito, "The LJ Speech dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.

[11] Y. Wang, D. Stanton, Y. Zhang, R. Skerry-Ryan, E. Battenberg, J. Shor, Y. Xiao, F. Ren, Y. Jia, and R. A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *Proc. ICML*, 2018, pp. 5180–5189.

[12] É. Székely, J. P. Cabral, M. Abou-Zleikha, P. Cahill, and J. Carson-Berndsen, "Evaluating expressive speech synthesis from audiobooks in conversational phrases," in *Proc. LREC*, 2012, pp. 3335–3339.

[13] É. Székely, G. E. Henter, and J. Gustafson, "Casting to corpus: Segmenting and selecting spontaneous dialogue for TTS with a CNN-LSTM speaker-dependent breath detector," in *Proc. ICASSP*, 2019, pp. 6925–6929.

[14] M. Włodarczak and M. Heldner, "Respiratory constraints in verbal and non-verbal communication," *Front. Psychol.*, vol. 8, no. 708, pp. 1–11, 2017.

[15] P. J. Price, M. Ostendorf, and C. W. Wightman, "Prosody and parsing," in *Speech and Natural Language: Proceedings of a Workshop Held at Cape Cod*, 1989, pp. 5–11.

[16] A. Rochet-Capellan and S. Fuchs, "Take a breath and take the turn: how breathing meets turns in spontaneous dialogue," *Phil. Trans. R. Soc. B*, vol. 369, no. 20130399, 2014.

[17] Google LLC, "Google Cloud Speech API video model," https://cloud.google.com/speech, accessed: 2019-03-18.

[18] T. Baumann, C. Kennington, J. Hough, and D. Schlangen, "Recognising conversational speech: What an incremental ASR should do for a dialogue system and how to get there," in *Dialogues with Social Robots*. Springer, 2017, pp. 421–432.

[19] R. M. Ochshorn and M. Hawkin, "Gentle forced aligner," https://github.com/lowerquality/gentle, 2017, accessed: 2019-02-14.

[20] R. Mama, "Tacotron-2 Tensorflow implementation," https://github.com/Rayhane-mamah/Tacotron-2, 2018, accessed: 2019-02-14.

[21] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," in *Proc. ICASSP*, 2018, pp. 4779–4783.

[22] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE T. Acoust. Speech*, vol. 32, no. 2, pp. 236–243, 1984.

[23] K. Park and J. Kim, "g2p_en: A simple Python module for English grapheme to phoneme conversion," https://github.com/Kyubyong/g2p, 2018, accessed: 2019-02-14.

[24] "CMUDICT. Carnegie Mellon Pronunciation Dictionary," http://www.speech.cs.cmu.edu/cgi-bin/cmudict, 1998, accessed: 2019-02-14.

[25] E. H. Rothauser *et al.*, "IEEE recommended practice for speech quality measurements," *IEEE T. Acoust. Speech*, vol. 17, no. 3, pp. 225–246, 1969.

[26] J. E. Arnold, M. Fagnano, and M. K. Tanenhaus, "Disfluencies signal theee, um, new information," *J. Psycholinguist. Res.*, vol. 32, no. 1, pp. 25–36, 2003.

[27] J. E. Fox Tree and H. H. Clark, "Pronouncing "the" as "thee" to signal problems in speaking," *Cognition*, vol. 62, no. 2, pp. 151–167, 1997.

[28] Z. Wu, O. Watts, and S. King, "Merlin: An open source neural network speech synthesis system," in *Proc. SSW*, vol. 9, 2016, pp. 218–223.

[29] J. Kominek and A. W. Black, "The CMU Arctic speech databases," in *Proc. SSW*, 2004, pp. 223–224.

[30] M. Schoeffler, S. Bartoschek, F.-R. Stöter, M. Roess, S. Westphal, B. Edler, and J. Herre, "webMUSHRA—A comprehensive framework for web-based listening tests," *J. Open Res. Softw.*, vol. 6, no. 1, 2018.

[31] M. P. Aylett, B. R. Cowan, and L. Clark, "Siri, Echo and performance: You have to suffer darling," in *CHI Extended Abstracts*, 2019, p. alt08.