INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Synthesis after a couple PINTs: Investigating the Role of Pause-Internal Phonetic Particles in Speech Synthesis and Perception

Mikey Elmers[1], *Johannah O'Mahony*[2], *Éva Székely*[3]

[1]Saarland University, Germany
[2]University of Edinburgh, United Kingdom
[3]KTH Royal Institute of Technology, Sweden

elmers@lst.uni-saarland.de, johannah.o'mahony@ed.ac.uk, szekely@kth.se

Abstract

Pause-internal phonetic particles (PINTs), such as breath noises, tongue clicks and hesitations, play an important role in speech perception but are rarely modeled in speech synthesis. We developed two text-to-speech (TTS) systems: one with and one without PINTs labels in the training data. Both models produced fewer PINTs and had a lower total PINTs duration than natural speech. The labeled model generated more PINTs and longer total PINTs durations than the model without labels. In a listening experiment based on the labeled model we evaluated the influence of various PINTs combinations on the perception of speaker certainty. We tested a condition without PINTs material and three conditions that included PINTs. The condition without PINTs was perceived as significantly more certain than the PINTs conditions, suggesting that we can modify how certain TTS is perceived by including PINTs.

Index Terms: speech synthesis, pause particles, computational paralinguistics

1. Introduction

Pause-internal phonetic particles (PINTs) describe a variety of different phenomena such as tongue clicks, breath noises (i.e., inhalations and exhalations), acoustic-phonetic silence, and filler particles (FPs) like "uh" or "um". Research has shown that PINTs have listener-oriented benefits in synthetic speech. For example, FPs can reduce the cognitive load for the listener [1], silences can help digit recollection [2], and breath noises can aid in sentence recollection [3]. Text-to-speech (TTS) trained on read speech has reached human levels of naturalness. However, PINTs still remain largely unexplored [1, 4] in the modeling and synthesis of spontaneous speech. The implementation of PINTs into TTS can further explore the relationship between listener-oriented benefits and PINTs.

The goal of this study is to model PINTs based on a spontaneous speech corpus, and apply the resulting synthetic speech in a perceptual experiment. First, we present a technological contribution that incorporates PINTs from spontaneous speech into a TTS system. While synthesis of filled pauses and breath events have been the focus of other studies (e.g. [1] [4] [5]), to the best of our knowledge, this is the first synthetic voice that is able to produce discourse clicks. Second, we demonstrate that a variety of PINTs patterns, generated with TTS, can be used as experimental material. This is a contribution to an emerging methodology that uses state-of-the-art neural TTS for stimuli creation, instead of manual manipulations of recorded speech samples [6]. Specifically, we evaluated the effect of PINTs, on perceived certainty of the speaker, via a listening experiment.

2. Method

2.1. TTS Generation

2.1.1. Corpus Information

Our training material is from Open Yale Courses [7], which is a project that provides free and open access to a number of introductory courses from Yale University. We selected lectures that included a high number of spontaneous speech phenomena. Next, we annotated a subset[1] of lectures totaling 3 h 7 min for a single speaker with a diverse PINTs profile. The selected speaker's PINTs material was approximately 40 % of the total lecture time. An example annotation can be found in Figure 1.

2.1.2. Data Preparation

Our training data incorporated transcripts taken from the Open Yale Courses website. We removed all punctuation in the original transcripts, as these are meant to improve the readability and do not correspond to acoustics. Next, we assigned PINTs to the available punctuation labels. For example, silence (,), inhalation (;), exhalation (.), tongue click (tk), filler particle (uh), and filler particle (um). The following is an example transcript with PINTs punctuation inserted: "; the metropolis which uproots people . , tk uh takes them away takes them out of ; traditional cultures , tk ;". Numbers were typed alphabetically (e.g., nineteen twenty two), accented symbols (e.g., Leger vs. Léger) and hyphens (e.g., self consciously) were removed, and acronyms were written out (e.g., r i s).

The original annotations included an "other" category, which comprised a variety of phenomena such as laughter. The "other" labels from the annotations were not included in the training transcript because they comprised rare cases that were too infrequent to reliably model. We exclusively used punctuation and textual labels for PINTs, as opposed to introducing new symbols or phonemes. This ensures that our TTS system is capable of interpreting automatically generated input that is trained on text alone. In particular, this enables the fine-tuning of large language models on TTS corpora, as demonstrated in [8], to generate synthesis prompts that produce the distribution of PINTs in the training data. For example, inserting semicolons in places where the speaker is likely to take a breath, or 'tk' tokens when a speaker is likely to use a tongue click.

The training data was segmented into breath groups following [5, 9], which meant that audio snippets began and ended with an inhalation label. If the duration of the utterance was greater than 11 seconds, a constraint of Tacotron 2 [10], the audio was cut at a silence label instead. PINTs are often mod-

[1] Lectures 1, 7, 13, and 24 from https://oyc.yale.edu/english/engl-310

10.21437/Interspeech.2023-2178

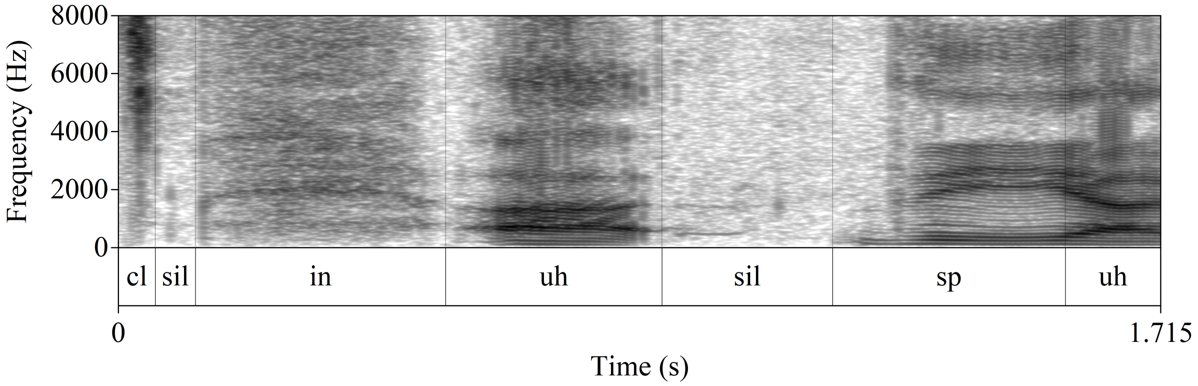

Figure 1: *Example section from speaker. Annotations of PINTs: silence (sil), inhalation noise (in), exhalation noise (ex), filler particles (uh) and (um), tongue click (cl), and other (o). Speech is annotated as "sp".*

eled beyond single sentences. However, due to the limitations of Tacotron2, we've only investigated single-sentence environments. All utterances were at least 4 seconds long. In total, we included 1224 breath group utterances, with 1128 in the training data and 96 held out for validation.

2.1.3. TTS Training

The TTS system was trained using a PyTorch implementation[2] of the sequence-to-sequence neural TTS engine Tacotron 2 [10]. The models (with 28.2M parameters) were trained using transfer learning on a pre-trained model based on a large read speech corpus, LJSpeech [11]. This approach has been beneficial to TTS quality when training on a limited size spontaneous corpus. Specifically, in reducing the number of mispronunciations and increasing speed of convergence [4]. We trained two models on the data: ControlledPINT, where all transcribed PINTs are included with their own lexical token, and AutoPINT, where we removed the transcriptions of the PINTs. Phoneme-level input is used for training and synthesis and is obtained from the transcripts using the `g2p_en` package [12]. Both voices were trained for 70k iterations on top of the published read speech model, on 3 GPUs each, for 67 hours, with a batch size of 28. The speech signal is decoded from the model output using the neural vocoder HiFi-GAN [13].

To evaluate the PINT insertion of the ControlledPINT and the AutoPINT models, we compared their outputs to natural speech using five sentences that were excluded from the training data. For the ControlledPINT model, we designed the input to match the type and location of the PINTs in the natural sentence. The AutoPINT model used only the textual material. We synthesized multiple versions using each model and selected the versions with minimal distortions or errors, without regard to PINTs production. We avoided versions[3] that included metallic reverberations that would sometimes occur due to the recording conditions.

2.2. Perceptual Study

Using synthesized samples generated by the ControlledPINT model, we developed a perceptual experiment that uses generated audio to evaluate how PINTs influence certainty scores. In this study the participants listened to audio samples and evaluated how "certain" the speaker sounded of their opinion.

2.2.1. Stimuli Creation

The textual material consisted of 10 sentences of similar syntactic structure, where the speaker describes their observations and opinions about artwork. For example, "The brush strokes in this painting contribute to a feeling of liveliness and energy". The semantic content of the utterances allowed for perceived hedging, indicating uncertainty. A Likert scale was used for evaluation with 1 representing "completely uncertain" and 7 representing "completely certain". Listeners heard a total of 40 audio stimuli, consisting of 10 different sentences synthesized in 4 different conditions (see Table 1). The "PINTsless" condition did not insert PINTs during synthesis. The "long silence" condition inserted a longer silence by including 3 silence symbols in a row. The "filler particle" condition inserted a silence and "um". And the "combinatory" condition inserted a silence, um, tongue click, and inhalation. Sentence final inhalations were removed, since the stimuli were evaluated in isolation. The tongue click in the "combinatory" version was surrounded by other PINTs because previous research has found that they co-occur alongside other PINTs in word-searching [14, 15].

Table 1: *Description of conditions used in perceptual study. The inserted material (punctuation labels) during generation is included.*

condition	**punctuation**
PINTsless	*N/A*
long silence	, , ,
filler particle	, *um*
combinatory	, *um tk in*

Our first hypothesis was that the PINTsless condition would be rated as more certain than the conditions that included

[2] `https://github.com/NVIDIA/tacotron2`

[3] Sample audio used for TTS comparison and perceptual experiment can be found at `https://mikeyelmers.github.io/paper_interspeech23ttsdemo/`

PINTs. Our second hypothesis was that the combinatory condition would be rated as more certain than the filler particle condition. A FP may indicate that the speaker has encountered word search (e.g., lexical retrieval) problems and the following tongue click may signal that the word was found. The long silence condition was included as a distractor to prevent the participants from developing overly simple heuristics in their certainty ratings.

2.2.2. Experimental Task

In an initial questionnaire, participants were asked about hearing impairment and age. All participants listened to the same set of stimuli, the order of which was randomized. The experiment required the use of headphones. Participants were asked to rate *'How certain does the speaker sound?'* on a 7-point Likert scale. The audio began automatically and the participants could click a "Play" button to hear the audio up to two more times before making their decision.

2.2.3. Participants

The perception study was created using a web-based experiment platform, Labvanced [16], which presented the audio material and collected responses. We recruited participants using the crowd-sourcing platform Prolific [17]. Fifty native English participants from the UK took part (mean age 40.7 years; age range 20–70 years, reflecting a diverse range of ages that represents a broad population). None of the participants self-reported a hearing impairment. Participants were paid for their participation.

3. Results

3.1. Evaluation of TTS Model Performance

3.1.1. Quantitative Analysis

Using the five sentences that were excluded from the training data, we annotated three versions and measured the duration of their PINTs material (see Table 2). Our results[4] showed that the natural condition had the longest duration of PINTs material, which closely matched the overall PINTs profile proportion of the speaker (40%). The ControlledPINT model produced the second longest durations and second largest proportion, while the AutoPINT version produced the shortest durations and smallest proportion. These findings were expected, but it was noteworthy that the ControlledPINT version closely resembled the natural version, and that the AutoPINT version could generate PINTs durations and proportion that were half of natural speech without any explicit labels. This is in line with the findings of [4], where filled pauses were automatically synthesized with a similar method.

We also looked at count information for the individual PINTs grouped by condition (see Table 3). The natural condition has the highest count values, with many more silences, especially edge silences that are adjacent to other PINTs, than material generated by either of the TTS systems. The ControlledPINT model produces more of the filler particle "uh" than the AutoPINT condition, but both systems produced the same number of "um" filler particles. The ControlledPINT system sometimes produce multiple PINTs from a single label, rendering more "uh" PINTs than was present in the natural speech. Only the natural material had tongue clicks or other labels.

Table 2: *Duration information for the different TTS models and natural speech for five sentences excluded from training. Both the total PINTs duration (PINTs dur) and the total audio duration (total dur) are measured in seconds. The proportion (prop) is measured out of 100%.*

condition	**PINTs dur**	**total dur**	**prop**
$natural$	15.96	41.57	38.39
$ControlledPINT$	13.82	40.82	33.86
$AutoPINT$	7.95	36.11	22.00

Table 3: *Count information for the different TTS models and natural speech: silence (sil), inhalation (in), exhalation (ex), filler particles (uh) and (um), tongue click (cl), and other (o).*

condition	**sil**	**in**	**ex**	**uh**	**um**	**cl**	**o**
$natural$	43	23	2	10	–	2	8
$ControlledPINT$	14	14	1	17	1	–	–
$AutoPINT$	10	13	1	4	1	–	–

3.1.2. Qualitative Analysis

Both the ControlledPINT and AutoPINT systems sometimes generate exhalations without a label. These exhalations were often near other PINTs in the data and this close association might be the cause of their unlabeled inclusion. Further evidence to support this theory comes from the ControlledPINT system sometimes producing a sequence of PINTs from just one or two labels in the input. Overall, the system is able to generate PINTs well, mirroring the PINTs pattern of the speaker. Occasionally, in cases with 5 or more PINTs in a row, the system struggles to perfectly recreate the PINTs sequence.

Our observations also revealed that tongue clicks were only realized by Tacotron2 when adjacent to silences or breath events. This is likely due to the fact that tongue clicks were one of the rarer PINTs in the training data and were almost always adjacent to other PINTs. Without an inhalation or silence in the prompt, the synthesizer would attempt to pronounce the tongue click symbol (tk) phonetically. The quality and loudness of the synthesized audio was also variable, likely due to differences in recording conditions across lectures. Originally, we expected the models to be quite probabilistic in their PINTs generation, however, they were more consistent than expected. Sometimes versions differed in their PINTs content but more often the differences were due to prosody and pronunciation variations.

3.2. Evaluation of the Perceptual Study

We created material, generated by the ControlledPINT system, for a perceptual experiment to evaluate the certainty of sentences in four conditions. The results for the perceptual study are in Table 4. Participants used the full scale in all conditions. We incorporated three measures of central tendency: mean, median, and mode. Each of these measurements highlights a different aspect of the data. For example, the mode for the PINTless condition was 7, indicating that the most common value was the highest possible rating.

The results confirmed our initial hypothesis that the PINTsless version would be rated more certain than the conditions with PINTs. The mean, median, and mode all clearly indicate that the PINTsless version sounded most certain. We also

[4] All data and code for the results can be accessed at `https://github.com/MikeyElmers/paper_interspeech23`

Table 4: *Descriptive statistics for the different conditions.*

conditon	**mean**	**median**	**mode**	**sd**
$PINTsless$	5.90	6	7	1.10
$long\,silence$	4.31	4	4	1.30
$filler\,particle$	3.72	4	4	1.24
$combinatory$	3.51	3	4	1.27

hypothesized that the combinatory version would be rated as slightly more certain than the filler particle condition. However, the data did not support this. All three PINTs conditions had similar certainty ratings but the long silence condition had the highest mean of the three PINTs conditions. The long silence condition also had more certainty scores in the 5-7 range than the other two PINTs conditions. The certainty scores for both the filler particle and combinatory conditions are similar but the filler particle condition has marginally higher certainty scores.

Statistical modeling was conducted with cumulative link mixed models (clmm) from the ordinal [18] (Version 2022.11-16) package in R [19] (Version 4.1.2). A post-hoc analysis was conducted using emmeans [20] (Version 1.8.4-1).

We compared a base clmm model, $clmm(certain \sim (1 \mid id) + (1 \mid stimuli)$, and a condition model, $clmm(certain \sim condition + (1 \mid id) + (1 \mid stimuli)$. The condition model predicts the certainty score with a single predictor, condition, as a fixed effect. For random effects, both subject id and stimuli with intercepts was included. An `anova()` was used to compare the two models. Table 5 shows that the model with condition as a predictor provides a significantly better fit than the base model as determined by both the Akaike information criterion (AIC) [21] and log-likelihood.

Table 5: *ANOVA comparison of base model and model with condition (cond) as a predictor. Includes number of parameters (par), AIC, log-likelihood (logLik), likelihood ratio test statistic (LR), degrees of freedom (df), and p-value (p).*

	par	**AIC**	**logLik**	**LR**	**df**	**p**
$base$	8	5718.5	−2851.2	−	−	−
$cond$	11	5637.2	−2807.6	87.24	3	< 0.001

A post-hoc analysis for pairwise comparisons was conducted. The PINTsless condition was significantly different (p < 0.001) from all PINTs conditions. The long silence condition was significantly different from both the filler particle condition ($p < 0.01$) and the combinatory condition ($p < 0.001$). However, the filler particle and combinatory conditions were *not* significantly different (p = 0.451).

4. Discussion

In our listening experiment, we expected the combinatory condition to indicate higher certainty scores than the other PINTs conditions, but this was not the case. Surprisingly, the long silence condition received slightly higher certainty scores than the other two PINTs conditions. One possible explanation is that the long silence condition might be less obtrusive than the filler particle or combinatory conditions. However, the long silence condition still disrupts the flow of speech more than the PINTsless condition, thereby reducing the listener's certainty. All PINTs were evaluated equally even though each PINT has different realizations that can influence certainty. Additionally, evaluating the effects of dialect, age, and gender for the interpretation of PINTs was outside the scope of this experiment.

Tongue clicks exhibit a number of functions such as: introducing a new sequence or topic, word search, maintaining a turn, backchanneling, stance marking, and repair [14, 22]. The acoustic realizations of these tongue clicks are highly variable [23], which means that the tongue clicks the TTS engine rendered might behave differently from our intended function. The fact that tongue clicks did not improve certainty by signaling a successful word search affirms that the production and perception of different PINTs patterns might require more elaborate experimental design, such as in-context perceptual evaluations. Future research could provide insights for audio enhancement tools, to reveal which tongue clicks can be removed from the recording and which are necessary for retaining the speaker's original intent.

We created two different TTS systems that were able to produce PINTs. The annotations for the TTS corpus were made manually, and therefore a time-consuming process. One limitation of the manual annotations was that we were only able to evaluate a single speaker. Automatic detection of PINTs is a challenging task, especially since some particles are less common than others [24]. Improvements could be made by including more data and more consistent audio quality. This study modeled PINTs in single-sentence environments. Future work should explore multi-sentence environments, which are more representative of the way PINTs occur in natural speech. The experiment in our paper is one possible example of how generative modeling can be used to create materials and test hypotheses, in this case improving our understanding of the functional properties of PINTs. Using generative modeling to distill knowledge is not going to replace the need for corpus-based research, but it is becoming a useful and necessary addition.

5. Conclusion

We developed an annotation scheme that uses plain text punctuation symbols to describe a speaker's PINTs pattern, which focused on consistency for successful generative modeling. Using these annotations, we trained two synthetic voices: ControlledPINT and AutoPINT. ControlledPINT used overt PINTs labels in the training material. AutoPINT did not include any PINTs labels and relies on the probabilistic rendering of Tacotron2 to insert them automatically. The novelty of our models is that they are the first to produce tongue clicks. Using the output of the ControlledPINT model, we conducted a perceptual experiment to evaluate how certain a synthetic speaker sounds in 4 different conditions. Importantly, we have shown that by incorporating natural phenomena (e.g., clicks), we are able to create manipulated experimental material. We hope that this line of research will contribute towards a deeper understanding of these complex and latent speech phenomena.

6. Acknowledgments

This research was funded in part by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-IDs MO 597/10-1 & TR 468/3-1. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 859588 and from the Riksbankens Jubileumsfond project CAPTivating – Comparative Analysis of Public speaking with Text-to-speech (P20-0298).

7. References

[1] R. Dall, M. Tomalin, and M. Wester, "Synthesising filled pauses: Representation and datamixing." in *Proc. SSW*, 2016, pp. 7–13.

[2] M. Elmers, R. Werner, B. Muhlack, B. Möbius, and J. Trouvain, "Evaluating the effect of pauses on number recollection in synthesized speech," in *Elektronische Sprachsignalverarbeitung 2021, Tagungsband der 32. Konferenz*, ser. Studientexte zur Sprachkommunikation. Berlin: TUD Press, 2021, pp. 289–295.

[3] ——, "Take a breath: Respiratory sounds improve recollection in synthetic speech," in *Proc. Interspeech*, 2021, pp. 3196–3200.

[4] É. Székely, G. E. Henter, J. Beskow, and J. Gustafson, "How to train your fillers: uh and um in spontaneous speech synthesis," in *Proc. SSW*, 2019.

[5] ——, "Breathing and speech planning in spontaneous speech synthesis," in *Proc. ICASSP*, 2020, pp. 7649–7653.

[6] A. Kirkland, H. Lameris, É. Székely, and J. Gustafson, "Where's the uh, hesitation? The interplay between filled pause location, speech rate and fundamental frequency in perception of confidence," in *Proceedings of Interspeech*, 2022, pp. 18–22.

[7] "Open Yale courses," https://oyc.yale.edu/, accessed: June 6th, 2022.

[8] S. Wang, J. Gustafson, and É. Székely, "Evaluating sampling-based filler insertion with spontaneous tts," in *Proc. LREC*, 2022, pp. 1960–1969.

[9] 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.

[10] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions," in *Proc. ICASSP*, 2018, pp. 4779–4783.

[11] K. Ito and L. Johnson, "The LJ speech dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.

[12] 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.

[13] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 022–17 033, 2020.

[14] R. Ogden, "Clicks and percussives in English conversation," *Journal of the International Phonetic Association*, vol. 43, no. 3, p. 299–320, 2013.

[15] J. Moreno, "[!] What's the name of it? [!] : Phonetic Clicks in Word Search Strategies in Glasgow," in *Proc. of the 19th International Congress of Phonetic Sciences*, 2019, pp. 1823–1827.

[16] H. Finger, C. Goeke, D. Diekamp, K. Standvoß, and P. König, "Labvanced: a unified JavaScript framework for online studies," in *International Conference on Computational Social Science (Cologne)*, 2017.

[17] "Prolific," Oxford, UK, 2014, accessed: 24.02.2023. [Online]. Available: https://www.prolific.co

[18] R. H. B. Christensen, "Ordinal—regression models for ordinal data," 2022, r package version 2022.11-16. https://CRAN.R-project.org/package=ordinal.

[19] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2021. [Online]. Available: https://www.R-project.org/

[20] R. V. Lenth, *Emmeans: Estimated Marginal Means, aka Least-Squares Means*, 2023, r package version 1.8.4-1. [Online]. Available: https://CRAN.R-project.org/package=emmeans

[21] H. Akaike, "Information theory and an extension of the maximum likelihood principle," in *International Symposium on Information Theory*, 1973, pp. 267–281.

[22] M. Zellers, "An overview of discourse clicks in Central Swedish," in *Proc. Interspeech 2022*, 2022, pp. 3423–3427.

[23] J. Trouvain and Z. Malisz, "Inter-speech clicks in an Interspeech keynote," in *INTERSPEECH 2016*. International Speech Communication Association, 2016, pp. 1397–1401.

[24] M. Elmers, "Comparing detection methods for pause-internal particles," in *Elektronische Sprachsignalverarbeitung 2022, Tagungsband der 33. Konferenz*, ser. Studientexte zur Sprachkommunikation. Sonderborg: TUD Press, 2022, pp. 204–211.