Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 6368–6374
Marseille, 11–16 May 2020

Augmented Prompt Selection for Evaluation of Spontaneous Speech Synthesis

Éva Székely, Jens Edlund, Joakim Gustafson
Division of Speech, Music and Hearing, KTH Royal Institute of Technology
Lindstedtsvägen 24, 114 28 Stockholm, Sweden
szekely@kth.se, edlund@speech.kth.se, jkgu@kth.se

Abstract

By definition, spontaneous speech is unscripted and created on the fly by the speaker. It is dramatically different from read speech, where the words are authored as text before they are spoken. Spontaneous speech is emergent and transient, whereas text read out loud is pre-planned. For this reason, it is unsuitable to evaluate the usability and appropriateness of spontaneous speech synthesis by having it read out written texts sampled from for example newspapers or books. Instead, we need to use transcriptions of speech as the target - something that is much less readily available. In this paper, we introduce Starmap, a tool allowing developers to select a varied, representative set of utterances from a spoken genre, to be used for evaluation of TTS for a given domain. The selection can be done from any speech recording, without the need for transcription. The tool uses interactive visualisation of prosodic features with t-SNE, along with a tree-based algorithm to guide the user through thousands of utterances and ensure coverage of a variety of prompts. A listening test has shown that with a selection of genre-specific utterances, it is possible to show significant differences across genres between two synthetic voices built from spontaneous speech.

Keywords: evaluation, spontaneous speech synthesis, human-in-the-loop, intelligence augmentation

1. Introduction

The application areas of speech synthesis are expanding into new domains, from news readers and simple task-oriented dialogues to more elaborate, multi-turn conversations. With this come new needs for more complex affordances such as exhibiting turn-taking cues, performing speaker entrainment, holding the listener's attention or expressing degrees of uncertainty, to name a few. It is expected that agents and dialogue systems will move from using TTS trained on read-speech data to context embedded speech synthesisers, trained on recordings of real-life spontaneous speech. As we move away from citation-style, one-size-fits-all, neutral read speech, we find that more expressive TTS voices might also become more context-dependent (Székely et al., 2019b). For instance, a spontaneous speech synthesiser trained on impromptu, casual conversations might perform really well on tasks within the original domain, but it is unknown if and how well it will transfer to new domains. However, it should not be necessary to record new data and train a new voice for each new application to ensure the appropriateness of the synthesiser's speaking style. Rather, TTS developers should be equipped with rapid evaluation techniques, to measure a system's performance for a particular domain, or compare systems across different genres and domains, and consequently adjust their systems for better suitability.

When preparing materials for listening tests, evidently, the text to be synthesised should be selected prior to synthesis, to avoid "cherry picking" of evaluation stimuli. Test materials for subjective tests of read speech synthesis usually come from randomly selected text from domains like newspapers and novels. No domain is fully neutral however, and a prescreening of the randomly selected sentences is almost always necessary for avoid strange, or potentially offensive content, or just atypical sentences. Domain-dependent subjective tests have long been suggested to be a valid metric for evaluation of spontaneous speech synthesis (Sundaram and Narayanan, 2003), but in practice, the preparation of these listening tests can be very different from evaluating read-speech synthesis on written utterances. Namely, when we evaluate the performance of spontaneous speech synthesis on spontaneously spoken utterances, we need to select the utterances from audio recordings, rather than text. In addition, using transcriptions from in-the-wild, real-world spontaneous speech in our evaluations means that the selection of test materials is most often far from random, but rather a lengthy process of manually choosing utterances using various criteria based on semantics, length, the presence and number of filled pauses and other disfluencies, reduced forms etc. (Andersson et al., 2010; Dall, 2017; Székely et al., 2019b; Székely et al., 2019a). There are several reasons why selecting evaluation prompts for spontaneous TTS tends to be a cumbersome process. Apart from the transcription often not being available, or containing ASR errors, the most important difficulty comes from the fact that in spontaneous speech, there are no sentences, in the conventional sense. Spoken communication often does not conform to the syntactic rules of written language, and consequently, automatic segmentation to standalone semantically coherent units is problematic. There are repetitions, re-starts, filled pauses, a well as reduced and mispronounced words and other phenomena that are difficult to represent in text. Moreover, as system developers, we do not necessarily have a clear idea of the styles present / affordances needed for the synthetic voice to represent a spoken genre or fulfill a particular task in a given domain.

Genres of spoken discourse differ from each other in many aspects (Jones, 2016). Regarding speaker-listener relationship we differentiate between symmetric dialogue, asymmetric dialogue and monologue. Regarding the degree of planning, speech can be memorised or scripted, extemporaneous (planned, using an outline) or impromptu.

According to register, discourse can range from formal through semi-formal to informal. Moreover, in different genres various communicative functions might be relevant, such as expressing different degrees of certainty, giving instructions, feedback, reporting or narrating, giving descriptions or introductions etc.

Contextual appropriateness has been proposed as a metric by Wagner et al. (2019) which ideally involves situationally oriented, task-based, interactive and context-embedded evaluations. However, in the early stages of development of a synthesiser, these might not be economical to implement because they tend to be more difficult to design and execute, take more time and resources, as well as be harder to repeat and reproduce when systems get updated, or to compute significance or to get comparable results across experiments (Székely et al., 2012; Wester et al., 2016; Mendelson and Aylett, 2017; Betz et al., 2018; Clark et al., 2019).

The goal of this paper is to develop an augmented prompt selection system to be used in conventional listening tests, for rapid perceptual evaluation of TTS voices for a expected usage domain and obtaining an estimate of the task-based performance of the system. This should also help assess the difference between two similar systems that are both expected to perform quite well in a given domain.

There are a number of constraints on evaluation prompts to be used in MUSHRA-like (Rec, 2003), MOS (ITU, 2016), or pairwise listening tests, although most of these are applied as rules of thumb in the TTS community, and clear guidelines are not available. Because stimuli have to be presented to listeners in randomised order to prevent ordering effects, we cannot rely on context-embedding and simply select consecutive utterances from a recording. Stimuli also have to be semantically coherent, to be understandable to listeners, at least to the degree at which they can estimate from the words, how the utterance should sound or should have sounded. Contents should not be sensitive, controversial or jargon-laden to avoid bias and ensure ease of comprehension (unless these aspects are specifically the aim of the evaluation). The number of stimuli is also limited to how many ratings or comparisons people are able to complete without fatigue, which for example for web-based MUSHRA-like listening tests tends to be capped at 20-30 prompts, depending on the number of systems compared. When evaluating contextual appropriateness, a random selection of such a limited number of evaluation prompts is unlikely to provide a good coverage of the different speech styles present in the target domain or genre, and simultaneously confirm to the semantic constraints mentioned above. The contribution of this work is an interactive visualisation tool, called Starmap[1] to help TTS developers navigate through thousands of audio files of utterances originating from a target domain, and select evaluation materials that are varied, and at the same time representative of the most important communicative acts, pragmatic and expressive functions the speaker needs to perform in that domain. Visualising speech features with dimensionality reduction techniques has been a subject of recent attention (Székely et al., 2018; Fallgren et al., 2019; Tånnander et al., 2019). Here we apply a t-SNE visualisation (Van Der Maaten and Hinton, 2008) on prosodic features of segmented utterances, combined with a tree-based algorithm (Barnes and Hut, 1986) to accelerate the user's search through the possible samples. Our hope is to provide this tool for a stage 1 prescreening type listening test to select the best suited voice, or during system development to allow for the fine-tuning of systems and developing new system variants to be better suited for particular tasks, and then later proceed to more detailed interactive and task-oriented evaluation methodologies.

This paper is organised as follows. Section 2. introduces the components of the proposed augmented prompt selection method, outlining data preparation, feature extraction and visualisation, the acceleration of sample selection and the user interface. Section 3. describes a sample listening test exhibiting a use-case scenario for the proposed evaluation method. Results of the evaluation are discussed in Section 4.

2. Method

2.1. Segmentation

The input data needs to be segmented into utterances. In the case of spontaneous speech, this is not a trivial process. In the experiment, we used a simplified version of the speaker-dependent breath detector presented in Székely et al. (2019c) to segment the speech into breath groups, or sequences of breath groups of a minimum of 2 seconds long. However, depending on the data used, a voice activity detector, or a combination of silence threshold with a set minimum length can be sufficient.

If the audio of the test material contains speech from more than one participant, speaker diarisation needs to be performed. To ensure that all components of the augmented prompt selection presented in this paper are language-independent, in the experiment, we used the speaker diarisation method described in Patino et al. (2018) which is unsupervised and does not require transcription.

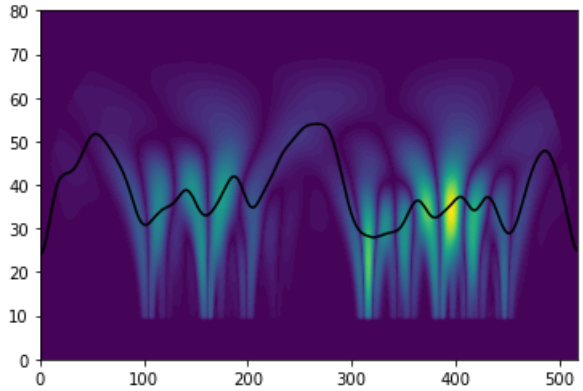

Figure 1: Speech rate approximation on an example utterance (*'And um if you say that'*) by the peaks of the wavelet matrix (Suni et al., 2016).

[1] http://sprakbanken.speech.kth.se/tools/starmap

2.2. Visualisation of the distribution of prosodic features with t-SNE

The visualisation of the prosodic feature distribuition in the target genre was implemented using t-SNE, an embedding technique that is commonly used for the visualisation of high-dimensional data in scatter plots (Van Der Maaten and Hinton, 2008).

For feature extraction, we used the Wavelet Prosody Analyzer toolkit[1], which implements the Continuous Wavelet Transform (CWT) based hierarchical prosody representation and estimation method described in Suni et al. (2017).

The following prosodic features were extracted for each utterance, normalised and used as input to the t-SNE:

- f0 mean
- f0 standard deviation
- energy mean
- duration
- estimated speech rate (see Figure 1)
- location of main prominence peak (see Figure 2)

Figure 3 illustrates the distribution of each feature on one of our evaluation datasets (see Section 3.3.). Speech rate was estimated using an unsupervised wavelet-based rate estimation method (Suni et al., 2016), which can be applied without the need for transcription. The speech is preprocessed as described in the original paper, including the smoothing out of sub-syllable phonetic details using a Gaussian filter. A CWT energy scalogram (see Figure 2) is calculated using a Mexican Hat mother wavelet. (A scalogram is a visual representation of the wavelet transform, having axes for time, scale and coefficient value.) The maximum energy scale is used to estimate speech rate; peaks in this signal correlate with the location of syllables, see Figure 1. Estimated speech rate is then calculated by the number of peaks (as approximation of the number of syllables) in the speech part of the signal divided by the duration of the speech signal between breaths.

[1] https://github.com/asuni/wavelet_prosody_toolkit

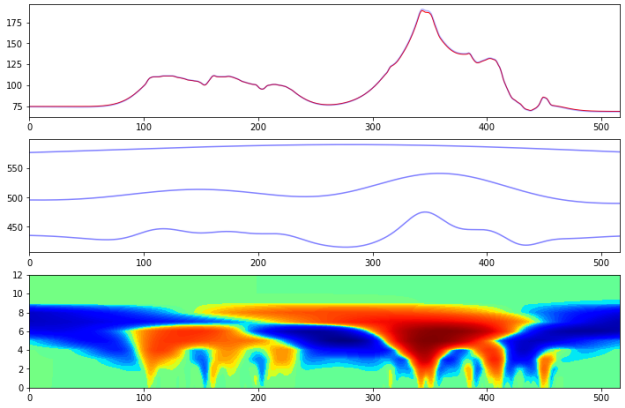

Figure 2: Prominence features of the same utterance as in Figure 1 (Suni et al., 2017). From top to bottom: f0, prominence scales, f0 scalogram.

Prominence is calculated using CWT performed on a composite signal combining f0, energy and speech rate as described in Suni et al. (2017), applying the original parameter settings, but with the speech rate approximated as above. Multiple hierarchical scales are separated (see second panel in Figure 2), which at the highest levels summarise an utterance-level phrase structure, whereas lower scales tend towards representing microprosody. To form an input feature to the t-SNE, continuous prominence values are summarised by the relative location of the peak in the top scale of the scalogram, indicating the location of the main prominence peak. For each utterance, a continuous value between 0 (start of utterance) and 1 (end of utterance) indicating the relative location of the peak is calculated.

2.3. Optimisation of sample selection with the Barnes-Hut algorithm

To help ensure the selection of a variety of utterances, and to accelerate the user's navigation through the sample space and we apply a guided, weighted random sampling method on the t-SNE. For this approach, we take inspiration from the acceleration of t-SNE using tree-based algorithms presented in Van Der Maaten (2014). Barnes-Hut (Barnes and Hut, 1986) simplifies the computation of the forces acting on a body in large data sets by grouping points that are far away and using their center of mass in these calculations. To ensure that the weighted random sampling takes into account elements that have previously been selected or discarded, we use this principle to adjust the weight in the sampling of each point as between the selection of new samples as follows. The weight of each data point is initialised at 1. A force exerted on each unselected data point is calculated using Barnes-Hut, setting the weight of a selected datapoint to a large multiple of the unselected data points (sum of all other weights). The force is normalised to a 0 to 1 scale over all observations, and the reduction in the weight of each datapoint is calculated by the normalised force amount, multiplied by a standard factor (2.1 was used in our experiments). Updated weights are floored at 0, and the weights of samples already played are set to 0. Figure 4 shows the updated sampling weights after 6 samples have been selected by the user.

2.4. Interface

In the current version of Starmap, the user has the following options to propagate through the data and select some of the utterances recommended by the weighted random sampling:

1. play the next sample suggested by the algorithm
2. repeat last sample
3. discard (based on e.g. semantics) and play a sample nearby
4. play together with following utterance
5. play together with previous utterance
6. add current sample(s) to final selection
7. undo last selection
8. select sample by coordinate and play
9. show all scalograms of selection
10. exit (and save final selection of samples)

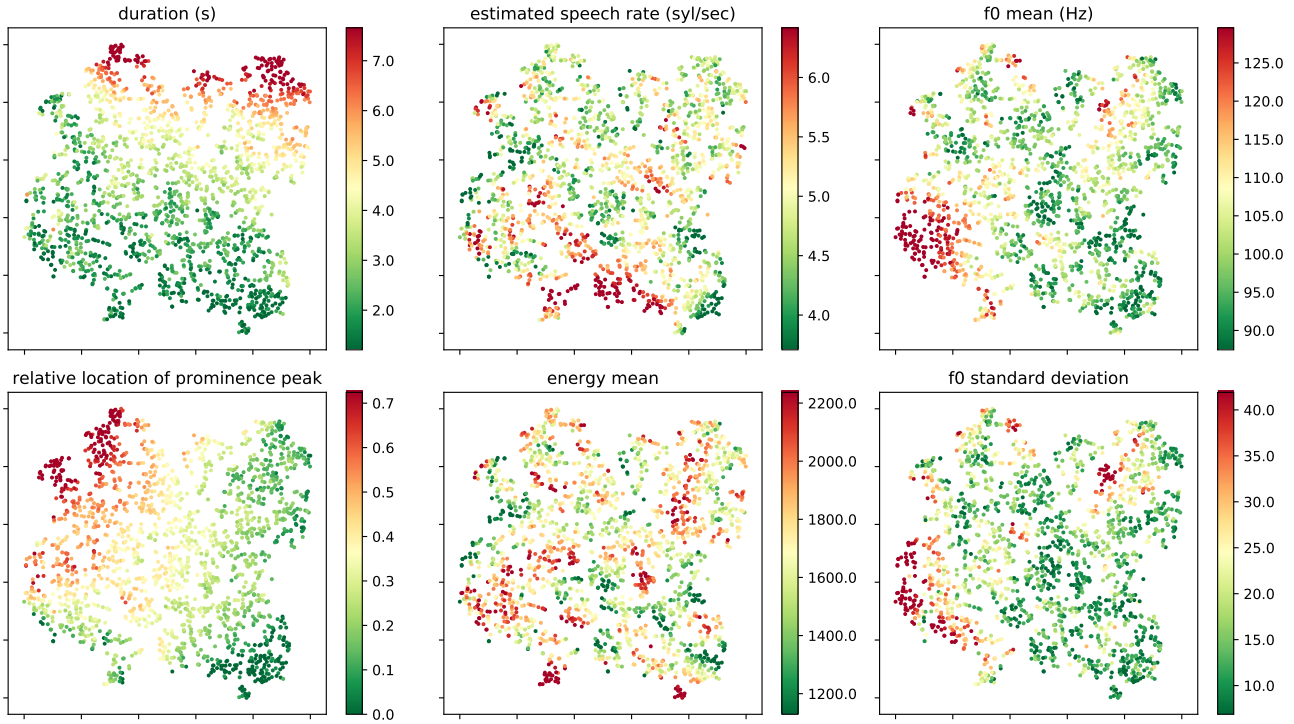

Figure 3: t-SNE visualisation of prosodic features of 1924 utterances.

The t-SNE visualisation is updated after each time an utterance is added to the final selection by the user: displaying the location of the already selected samples, and the updated weights accordingly, with lower weights appearing in a lighter shade on a colour scale. The selected sample is displayed in green, while the previous and following samples are colored blue and yellow, respectively. This allows the user to listen to speech segments in their context, and if necessary for comprehension, append two segments to be used together as one evaluation prompt. It is also possible for the user to discard an utterance based on for example semantic constraints, but query the system (option 3) to instead play a speech segment with similar prosodic characteristics (closest neighbour in the tSNE). For an illustration of Starmap in use, see Figure 4.

The user also has the option to display the f0 scalogram of the current and already selected samples, to provide an additional visual overview of the prosodic variety across the utterances already in the selection.

3. Evaluation

3.1. Aim of evaluation

Since the augmented prompt selection tool is novel exploratory in nature, instead of a formal evaluation, we here conduct an experiment, where we showcase the usage of the tool on an example evaluation scenario; using two spontaneous TTS voices trained on different data and two different target domains of spoken discourse.

3.2. Databases and synthesis

In this evaluation, we compare two synthetic voices, trained on different genres of spontaneous speech. For the first voice, which we will refer to as TGD, we used the audio from the Trinity Speech-Gesture Dataset (Ferstl and McDonnell, 2018), consisting of 25 impromptu monologues, on average 10.6 minutes long, performed over multiple recording sessions by a male speaker of Irish English. The actor is speaking in a colloquial style, spontaneously and without interruption, on topics such as hobbies, daily activities, and interests. During the monologues, he addresses a person seated behind the cameras who is giving visual, but no verbal feedback.

The second voice, here called TCC, was built on the ThinkComputers Corpus, originating from publicly available recordings of a conversational podcast by two male speakers of American English, described in (Székely et al., 2019b). The podcast features product reviews and discussions of technology news, presented in a conversational, extemporaneous style. Utterances from the speaker with the most air time were selected for the TTS corpus.

The speaker-dependent breath detector proposed by Székely et al. (2019c) was employed to detect breath events in the recordings. With this method, which we were able to segment the audio of the TGD fully automatically, to produce a TTS corpus of 3,487 breath-group utterances, totalling 4.5 hours. The TTC corpus required a manual review of the automatically selected breath groups, to remove utterances that contained overlapping speech with feedback tokens from the conversation partner. In total, the final database comprised 6,218 breath groups, 9 hours of audio.

Both corpora were transcribed using the enhanced video model of Google Cloud Speech-to-Text API (Google LLC, 2019). The Gentle forced aligner (Ochshorn and Hawkin, 2017) and the US English BroadbandModel of the IBM Watson Speech-to-Text were used to detect and disambiguate between filled pauses *uh* and *um* in order

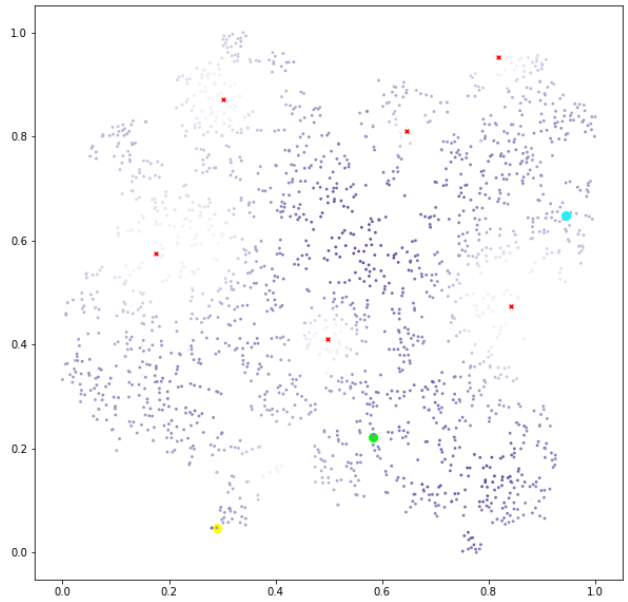

Figure 4: Illustration of Starmap in use, with 6 samples selected; the current sample under consideration is marked green with its preceding and following sample marked in blue and yellow respectively (if available).

to obtain an annotation as close as possible to the original speech. For more details on the annotation pipeline please refer to (Székely et al., 2019b). All punctuation was removed from the automatic transcription, as ASR inserts punctuation with the main goal of making text more readable, not necessarily corresponding to the prosody of the realised speech. Both synthetic voices were built using the Tacotron 2 spectrogram-prediction framework (Shen et al., 2018), implementation by Mama (2018). Audio was sampled at 22.05 kHz. The Griffin-Lim algorithm (Griffin and Lim, 1984) was used for waveform synthesis. Transfer learning on a pretrained model of read speech (Ito, 2017) and front-end-based phonetisation were used as these have been shown to reduce the pronunciation errors produced by voices trained on corpora transcribed by ASR (Székely et al., 2019b).

3.3. Use-case scenarios

Two spoken genres were selected to demonstrate an example use of Starmap with the TTS voices described in Section 3.2.. The first target domain selected is a solo-host podcast, specifically the "Lexicon Valley podcast" by John McWhorter, which is a show about language-related topics such as etimology, neurolinguistics and syntax. Apart from a few embedded sound tracks, it is the hosts task is to keep the audience engaged, speaking in an entertaining, extemporaneous style using a prepared outline but producing speech on the fly. In the evaluation, we used 5 episodes totalling 193 minutes, segmented into 2609 utterances of a minimum length of 2 seconds, using breath event detection (Székely et al., 2019c). If two breath events were less than 2 seconds apart, the segment was concatenated with the following breath group.

The second target domain in this evaluation is the role of the interviewer in interview podcasts, where the host is equipped with a set of prepared introductions and questions, but maintains a relatively free-flowing conversation with the guest, where new topics and questions may arise spontaneously. 12 episodes were selected from "The TED Interview podcast", hosted by Chris Anderson for evaluation. A total of 1924 utterances (145 minutes) were included from the interviewer, using speaker diarisation and breath detection.

3.4. Utterance selection using Starmap

From each target domain, 20 samples were chosen using the Starmap described in Section 2. The process took approximately 70 minutes. The selected samples were transcribed through ASR and then corrected manually before fed into the synthesier. No punctuation was used in the input prompts.

3.5. Listening experiment

Subjects were given an ABX-style listening test implemented using the WebMUSHRA codebase (Schoeffler et al., 2018), and asked which voice they preferred for each sample, with the option to indicate if they had no preference for either of the samples. Five pairs of samples were repeated throughout the test, to allow checking users' answers for consistency. The evaluation stimuli can be listened to under: http://www.speech.kth.se/tts-demos/LREC20/

4. Results

26 participants, recruited through Prolific Academic completed the listening tests successfully, taking an average time of 10 minutes. All listeners were native speakers of English and reported to have been wearing headphones throughout the test. If a participant provided an inconsistent answer on more than 2 of the 5 repeated sample pairs, their results were excluded from the analysis.

The results showed that for the interview podcast domain, neither voice was preferred significantly more often than the other, see Figure 5. However, a significant majority of listeners preferred the TCC voice for the solo-host podcast domain over the TGD voice ($p \ll 0.001$).

The results reflect that the extemporaneous podcast style of TTC matched the target genre of solo-host podcast better than the impromptu style of TGD, and in this case, this seems to be a more important factor in successful genre-transfer than the communicative setting (monologue or conversational). As expected, when looking at the

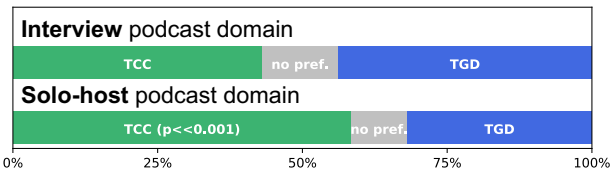

Figure 5: Preference test results. p-values were calculated using the exact binomial test on the null hypothesis that the other voice is preferred (thus adding no-preference votes to that voice count).

prosodic features of the selected utterances by the original speakers, which TTS voice is preferred for an individual utterance is not easily explained by how close an individual feature is to that feature in the original utterance. Speech rate appears to show a weak correlation (r=-0.27) but still no significant relationship with preference ratings.

5. Discussion

Both synthetic voices were built from recordings of male speakers, but apart from the domain of the recordings (conversational podcast vs. impromptu monologue) the voices also differed in the speakers' age, accent, the recording conditions and the amount of the training data. The fact that we have seen a significant difference between the target genres in the evaluation shows that the style was probably the most important factor differentiating between the voices, but it is difficult to infer to what extent other aspects such as accent and voice quality played a role in listener's judgments. When comparing two TTS voices trained on different data, it is seldom possible to bring the differences down to one variable. Moreover, we wanted to include a realistic example of domain-oriented TTS evaluation, where both compared voices are presented in their best possible version, rather than for example reducing available training to make the comparison between the voices more equal.

6. Conclusions

In this paper, we introduced Starmap, a tool to help TTS developers design evaluations using utterances originating from spontaneous speech. The aim was to develop a rapid evaluation method that sheds light on how well spontaneous TTS performs different aspects of spoken discourse and domain-specific affordances a present in speech from a particular genre. The example evaluation included in this study shows that it is possible to find significant differences in spontaneous TTS voices with as little as 20 utterances selected from two genres. We have made the tool available for the research community and we will be adding new features in the future. Our hope is that this method will prove useful for TTS researchers to rapidly assess the usability of their systems for a given genre during the development phase, before moving on to more elaborate domain-oriented interactive evaluations.

7. Acknowledgements

This research is supported by the Swedish Research Council project Connected: Context-aware speech synthesis for conversational AI. VR (2019-05003) and by the Riksbankens Jubileumsfond funded project TillTal (SAF16-0917:1). The results of this work are made accessible through the national infrastructure Språkbanken Tal under funding from the Swedish research Council (2017-00626).

8. Bibliographical References

Andersson, S., Georgila, K., Traum, D., Aylett, M., and Clark, R. A. J. (2010). Prediction and realisation of conversational characteristics by utilising spontaneous speech for unit selection. In *Proc. Speech Prosody*.

Barnes, J. and Hut, P. (1986). A hierarchical O (N log N) force-calculation algorithm. *Nature*, 324(6096):446.

Betz, S., Carlmeyer, B., Wagner, P., and Wrede, B. (2018). Interactive hesitation synthesis: Modelling and evaluation. *Multimodal Technologies and Interaction*, 2(1).

Clark, R., Silen, H., Kenter, T., and Leith, R. (2019). Evaluating Long-form Text-to-Speech: Comparing the Ratings of Sentences and Paragraphs. In *Proc. SSW*, pages 99–104.

Dall, R. (2017). *Statistical Parametric Speech Synthesis Using Conversational Data and Phenomena*. Ph.D. thesis, School of Informatics, The University of Edinburgh, Edinburgh, UK.

Fallgren, P., Malisz, Z., and Edlund, J. (2019). How to Annotate 100 Hours in 45 Minutes. In *Proc. Interspeech 2019*, pages 341–345.

Ferstl, Y. and McDonnell, R. (2018). Investigating the use of recurrent motion modelling for speech gesture generation. In *Proc. IVA*, pages 93–98.

Google LLC. (2019). Google Cloud Speech API video model.

Griffin, D. and Lim, J. (1984). Signal estimation from modified short-time Fourier transform. *IEEE T. Acoust. Speech*, 32(2):236–243.

Ito, K. (2017). The LJ Speech Dataset. `https://keithito.com/LJ-Speech-Dataset`.

(2016). Mean opinion score terminology. ITU-R Recommendation P.800.1, International Telecommunication Union.

Jones, R. (2016). *Spoken discourse*. Bloomsbury Publishing.

Mama, R. (2018). Tacotron-2 Tensorflow implementation. `https://github.com/Rayhane-mamah/Tacotron-2`.

Mendelson, J. and Aylett, M. P. (2017). Beyond the Listening Test: An Interactive Approach to TTS Evaluation. In *Proc. Interspeech*, pages 249–253.

Ochshorn, R. M. and Hawkin, M. (2017). Gentle forced aligner. `https://github.com/lowerquality/gentle`.

Patino, J., Delgado, H., and Evans, N. W. (2018). The EURECOM Submission to the First DIHARD Challenge. In *Proc. Interspeech*, pages 2813–2817.

Rec, I. (2003). bs. 1534-1 method for the subjective assessment of intermediate quality level of coding systems. Technical report, Technical report.

Schoeffler, M., Bartoschek, S., Stöter, F.-R., Roess, M., Westphal, S., Edler, B., and Herre, J. (2018). WebMUSHRA – A comprehensive framework for web-based listening tests. *J. Open Res. Softw.*, 6(1).

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

Sundaram, S. and Narayanan, S. (2003). An empirical text transformation method for spontaneous speech synthesizers. In *Proc. Eurospeech*, pages 1221–1224.

Suni, A., Simko, J., and Vainio, M. (2016). Bound-

ary detection using continuous wavelet analysis. *Speech Prosody 2016*, pages 267–271.

Suni, A., Šimko, J., Aalto, D., and Vainio, M. (2017). Hierarchical representation and estimation of prosody using continuous wavelet transform. *Computer Speech & Language*, 45:123–136.

Székely, É., Cabral, J. P., Abou-Zleikha, M., Cahill, P., and Carson-Berndsen, J. (2012). Evaluating expressive speech synthesis from audiobooks in conversational phrases. In *Proc. LREC*, pages 3335–3339.

Székely, Éva., Wagner, P., and Gustafson, J. (2018). The Wrylie-board: mapping acoustic space of expressive feedback to attitude markers. In *Proc. IEEE Spoken Language Technology Workshop*.

Székely, É., Henter, G. E., Beskow, J., and Gustafson, J. (2019a). How to train your fillers: uh and um in spontaneous speech synthesis. In *Proc. SSW*, pages 245–250.

Székely, É., Henter, G. E., Beskow, J., and Gustafson, J. (2019b). Spontaneous Conversational Speech Synthesis from Found Data. In *Proc. Interspeech 2019*, pages 4435–4439.

Székely, É., Henter, G. E., and Gustafson, J. (2019c). Casting to corpus: Segmenting and selecting spontaneous dialogue for TTS with a CNN-LSTM speaker-dependent breath detector. In *Proc. ICASSP*, pages 6925–6929.

Tånnander, C., Fallgren, P., Edlund, J., and Gusafsson, J. (2019). Spot the Pleasant People! Navigating the Cocktail Party Buzz. In *Proc. Interspeech*, pages 4220–4224.

Van Der Maaten, L. and Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605.

Van Der Maaten, L. (2014). Accelerating t-SNE using tree-based algorithms. *Journal of Machine Learning Research*, 15:3221–3245.

Wagner, P., Beskow, J., Betz, S., Edlund, J., Gustafson, J., Henter, G. E., Le Maguer, S., Malisz, Z., Székely, É., Tånnander, C., and Voße, J. (2019). Speech Synthesis Evaluation – State-of-the-Art Assessment and Suggestion for a Novel Research Program. In *Proc. SSW*, pages 105–110.

Wester, M., Watts, O., and Henter, G. E. (2016). Evaluating comprehension of natural and synthetic conversational speech. In *Proc. Speech Prosody*, pages 766–770.