Interspeech 2022
18-22 September 2022, Incheon, Korea

Transplantation of Conversational Speaking Style with Interjections in Sequence-to-Sequence Speech Synthesis

Raul Fernandez[1], *David Haws*[1], *Guy Lorberbom*[1], *Slava Shechtman*[1], *Alexander Sorin*[1]

IBM Research AI

{fernanra,dhaws}@us.ibm.com, guy.lorberbom@ibm.com, {slava,sorin}@il.ibm.com

Abstract

Sequence-to-Sequence Text-to-Speech architectures that directly generate low level acoustic features from phonetic sequences are known to produce natural and expressive speech when provided with adequate amounts of training data. Such systems can learn and transfer desired speaking styles from one seen speaker to another (in multi-style multi-speaker settings), which is highly desirable for creating scalable and customizable Human-Computer Interaction systems. In this work we explore one-to-many style transfer from a dedicated single-speaker conversational corpus with style nuances and interjections. We elaborate on the corpus design and explore the feasibility of such style transfer when assisted with Voice-Conversion-based data augmentation. In a set of subjective listening experiments, this approach resulted in high-fidelity style transfer with no quality degradation. However, a certain voice persona shift was observed, requiring further improvements in voice conversion.
Index Terms: sequence to sequence speech synthesis, conversational speech synthesis, synthesis of interjections, voice conversion, data augmentation, speaking style transplantation

1. Introduction

Text-to-speech (TTS) systems based on sequence-to-sequence (S2S) and neural vocoding techniques [1, 2, 3, 4] produce state-of-the-art performance that approach or match human levels, and have led to a ubiquitous deployment of the technology in speech interfaces. In tandem with the evolving quality of such systems, we see a move toward solving more complex cases that arise in Human-Computer Interaction (HCI), including the need to synthesize different styles for various usage contexts, and to transfer this ability across various speakers and languages under often limited-data conditions. This work pursues several of these goals by creating fluid, conversational voices that can be naturally deployed within dialogue systems, are capable of a range of expressive devices observed in human-to-human communication (including the generation of non-lexical interjections), and can be learned from a conversational corpus and extended to novel voices lacking such training material. Central to the approach is the use of a style-preserving Voice Conversion (VC) algorithm that replicates a carefully-designed conversational TTS corpus in new target voices and enables transplantation of conversational characteristics (acoustic, lexical and non-lexical) across speakers. In order to allow this approach to be successful, this work also describes a strategy for collecting high-quality expressive data under controlled conditions that still respects the naturalness and spontaneity of live data.

Previous work in expressive S2S TTS has focused on varying notions of expressions, ranging from the implicit expression contained in a prosodic reference [5] to the use of explicit, but fairly polar, styles that may be insufficient for dialogues [6]. Learning conversational styles from spontaneous dialogues without dedicated TTS corpora was tackled in [7], illustrating the difficulty and limitations of the approach, and justifying the work presented here using controlled data. The closest research to ours to explicitly target conversational TTS is the work of [8] for voice agents in Mandarin. That work also importantly deals with the issue of designing appropriate training corpora though our approach differs substantially from theirs in the explicit use of Dialog Act tags and collection methodology (we cover more of those differences in § 2). The synthesis of interjections finds precedent in earlier work by [9] for unit-selection architectures, and more recently by [10] for S2S frameworks. Our work differs in that it situates the generation within the conversational domain and targets a larger inventory. Another feature needed for effective synthesis of dialogue turns is the ability to realize narrow emphatic focus on demand, which we have previously presented in [11] and continue to use here.

The idea of using VC as data-augmentation has been previously explored within tasks like the classification of symbolic prosodic events [12], recognition of low-resourced accented speech [13], and more recently in [14] to address a similar goal as in this work to mitigate the lack of expressive material for a variety of target speakers. Our work differs in the fact we tackle a more fine-grained set of conversational events with more expressive range, and in the way we allocate training resources (we use a larger corpus with a richer set of tags from a single training speaker vs. smaller corpora from several speakers).

We cover the details of designing and collecting expressive dialogues in §2. In §3 we present a S2S architecture sensitive to style transplant, and cover various techniques to enable this in §4. The experimental setup in §5 describes the training mechanism for building and selecting the models evaluated in §6 before concluding with observations and future directions in §7.

2. Design of a TTS Conversational Corpus

An objective in this work is to assemble a suitable corpus of recorded utterances to develop conversational TTS systems, particularly for conversational voice agents in customer-care (CC) settings. There are opposing constraints at work here. On one hand, we strive for a style that feels spontaneous, lively, and with more relaxed articulation. Corpora of spontaneous, natural dialogues with these characteristics already exist [15]. However, unconstrained conversations also exhibit behaviors that pose challenges when building TTS systems and severely degrade its quality, like disfluencies, repetitions, hypo-articulation, self-corrections, intonationally "ill-posed" utterances, etc. Our solution is to adopt the classical approach of recording professional voice talent under controlled conditions, but to carefully draft a script that facilitates and elicits a range of expressiveness and discourse-level behaviors, and to work with

[1]Authors are listed alphabetically and contributed equally.

Copyright © 2022 ISCA

10.21437/Interspeech.2022-388

a voice actor (VA) that can convincingly bring the script to life in a style that matches the CC setting. To that end, we have designed a script with the following:

(1) Appropriate structure and contextual history: The main part of the script consists of two-part dialogues with the VA playing the role of an *Agent* interacting with a *User*. The VA has access to both sets of turns, but records only their lines while silently reading over the *User* responses. Giving the VA access to the history allows them to follow a narrative arch and react expressively to the ever-evolving context. The split between *Agent* and *User* is approximately 75% -25%. Though this figure can differ substantially from real-life interactions, it is done to maximize the amount of audio collected within a given budget. The *User* turns merely provide context and keep the interaction moving along, creating a space for the VA to perform.

(2) Appropriate semantic context: The script has been sourced from real dyads (text-based and transcribed conversations) from proprietary materials reflecting a variety of CC topics, as well as from open-source dialogues from the Taskmaster-2 corpus [16]. These have been edited as needed for semantic clarity, to eliminate any personal information, and in order to reduce the artifacts mentioned (repetitions, corrections, etc.).

(3) Dialog-Act Tagging: The script has been annotated with ten tags reflecting various communicative and dialog-managing functions. This is done both to exploit higher-level features during modeling (see § 3), but also to aid the VA (who has access to the marked-up text while recording) understand the structure of the dialogues and to guide them toward particular stylistic performances when enacting individual tags. Prior to recording the script, the VA and producer held training sessions to establish a common understanding of the meaning of these tags, and to workshop some distinct stylistic renderings for them.

(4) Interjections: A substantial portion of the script contains a set of seven interjections (§ 2.1) to mirror several exclamations, fillers, etc. observed in naturally occurring dialogues, and to inject convincing naturalness and spontaneity into the dialogues.

A strategy was proposed in [8] for collecting dialogues for TTS with similar goals of maximizing spontaneity while reducing performance artifacts. Both that approach and ours eschew using isolated turns in favor of contextual dialogues as an elicitation paradigm. However, [8] allows freer improvisation during the recording between two speakers whereas we allow the VA to deviate from the script only in minor ways (corrected afterwards), and provide guidance with DA tags to hit desired expressiveness targets. With this design 100% of the recorded audio is usable to train an *Agent*-like model whereas [8] uses only the *Agent* turns amounting to about 50% of the collected data. For our work, we collected about 7 hours of audio; this includes 6 hours of conversations, and 1 hour of non-conversational, but expressive material for synthesizing focal emphasis.

2.1. Dialog Acts and Interjections

In this work we propose a set of ten *dialog tags* to help describe, control, and guide appropriate expressive realizations of the events happening in a dialogue. We approached the problem of creating such an inventory from both theoretical and empirical angles. The DIT++ taxonomy of Dialog Acts (DAs) [17], which provides a multi-dimensional framework for analyzing the communicative functions of acts (and has a focus on both human-human (H2H) and human-machine (HM) interactions) was consulted to have an initial set of descriptors. This inventory is quite large and likely over-specialized for synthesis goals, so the next step was to cross-reference these labels against actual HM dialogues to empirically select or modify the tags to arrive at a more concise set. A fundamental selection principle was to focus on commonly occurring DAs that would require, or benefit from, a dedicated stylistic shaping in terms of both acoustic realization (i.e., prosody and voice quality) and linguistic content (i.e., formulaic and idiomatic constructions that can serve as affective carriers, and allow for the inclusion of interjections. This led to the following inventory: {*Agreement*, *Farewell*, *Greeting*, *Empathy*, *Instruction*, *Positive Feedback*, *Surprise*, *Thanks*, *Uncertainty*, *Waiting* }.

By interjections we mean a set of often non-lexical items that are used as exclamations, fillers, and backchannels, and which are prevalent, and indeed necessary, in conversational settings. Interjections can differ from standard lexical items in their segmental and suprasegmental characteristics; for instance: unusual durations or pitch excursions governed by pragmatic factors (and not fully predictable from syntax), little to no voicing during production, etc. Even when we limit the inventory to items that can be transcribed phonetically fairly accurately (excluding more complex sounds like laughter), their high-quality synthesis from a corpus without enough exemplars remains challenging because of the differences just mentioned. To enable their successful synthesis during a dialogue, we have included the following set of interjections in the design of the corpus, where they appear in combination with a variety of DA tags previously listed: {*aha*, *oh*, *hmm*, *huh*, *uh*, *uh-huh*, *um*}. This inventory is by no means exhaustive and has been designed to attempt coverage of various desirable functional roles within a dialogue while trying to retain a manageable inventory[1].

3. Architecture Description

Our model (Fig. 2) is a variant of the non-attentive Tacotron2 [2] architecture (NAT2) from [4] with an extended set of symbolic input embeddings, augmented with hierarchical prosodic controls (HPCs) [11, 6] to improve its controllability. We previously showed that these HPCs are able to realize lexical focus [11], and facilitate speaking-style transfer (for *good news* and *apology*) in a *attentive* Tacotron2-like architecture [6] coupled with an LPCNet [18] neural vocoder. In this work we combine NAT2 with the controllability provided by HPC [11, 6]. Large scale crowd MOS tests (up to 100 subjects, 40 samples per system, 25 votes per sample) revealed that the proposed HPC-controlled *NAT2* architecture produces speech with the same quality and naturalness as the HPC-controlled *attentive* Tacotron2 architecture [11, 6] while almost entirely eliminating stuttering, repetitions, early stopping and other stability problems affecting the attentive Tacotron2 models in some edge cases. We also ensured that the proposed HPC-controlled NAT2 system is responsive to HPC offsets [6] and suitable for local word focus realization, as in [11], and for "simpler" speaking style transfer, as in [6], and is thus a fair baseline for the challenging conversational style-transplantation task.

The proposed S2S model is fed with phonetic symbols including pauses and word boundaries, enriched with lexicographic stress, phrase type (as extracted from the input text by our proprietary TTS Front-End) and additional prosodic directives, that were exposed to some of the voice talents during some of the speech recording sessions (i.e. word emphasis [11], speaking styles [6] and unique for this work *dialog tags* and *interjection types*) and generates an acoustic feature

[1] More details on the definitions and usage of the individual DA tags and interjections can be found in http://ibm.biz/IS22-S2SConv.

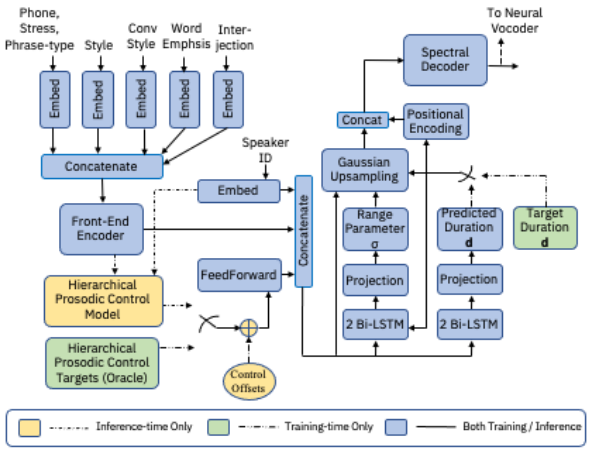

Figure 1: *The proposed HPC-controllable non-attentive Tacotron2 acoustic model, predicting LPCNet acoustic features from the extended phone sequence.*

vector sequence that is then fed to an independently trained LPCNet [18] neural vocoder to generate high quality samples in real time [19]. The sequence to sequence acoustic model (see Fig. 2) is first trained with ground truth HPCs, minimizing the acoustic regression loss, as in [6], but based on $L_1 + L_2$ loss operator [2], plus the duration prediction loss [4], then the HPC prediction sub-model is trained [6].

The phone identity, lexical stress and phrase type are represented by a single embedding, while the speaking style (e.g. *neutral* or *conversational*), the word emphasis [11], the dialog tags and the interjection types (corresponding, respectively, to the DA and interjection inventories from § 2.1) each has its own embedding (see Fig. 2). For brevity we will refer to the concatenation of the front-end (FE) encoder outputs, the speaker-id embedding, and the HPC embedding as the *encoder output*.

The *encoder output* is fed to the phone-level Duration Predictor and the Range predictor that facilitates the Gaussian upsampling of the *encoder-output* sequence to match the length of the spectral output (as determined by ground-truth phone durations) [4]. The upsampled encoder output is concatenated with the positional encoder output and fed to the Tacotron2-like autoregressive LSTM decoder [4], containing two LSTM stacks, followed by two parallel Linear layers predicting the 80-dim. mel-cepstral features (serving for the autoregressive decoder feedback) and the 22-dim. LPCNet features (20 for spectral envelope and 2 for pitch). The predicted LPCNet features are further processed with two post-nets (one for the cepstrum, and one for the pitch parameters) and are fed to the separately trained LPCNet neural vocoder, as detailed in [19].

4. Conversational Style Transplantation

In this work we aim to transplant the conversational style with interjections from a single speaker to others. This is challenging due to asymmetric data conditions, as the conversational-style and speaker-ID labels are highly correlated (over 80% of the conversational-speaker data is of *conversational* style) and also associated with specific textual patterns, such as common phrases and interjections that only arise in the conversational corpus. Thus the plain training of the proposed HPC-controlled NAT2 system alone (§3) does not attain acceptable style transfer to new speakers (§6) and leads to occasional persona switching for some inputs. To decouple the speaker ID from the rest of FE inputs and labels we explored the following Speaker Adversarial training and VC-based data-augmentation approaches.

4.1. Speaker Adversarial Classifier

To prevent speaker ID leakage into the speaker-independent FE Encoder, we deploy a phone-level adversarial speaker classifier [20] with inverse gradient layer [21] applied to the encoder outputs during training. The cross-entropy of speaker predictions is added to the total training loss as an additional optimization task. However, while improving the speaker similarity, this method significantly deteriorates the style similarity (see §6).

4.2. Data Augmentation by Voice Conversion

The undesired coupling between the conversational style and the conversational speaker ID can be implicitly reduced by training with more multi-speaker conversational data. In [14] VC was proposed as a data-augmentation technique to increase the existing amount of multi-speaker data with a desired style. The trained multi-speaker multi-style TTS system is further fine-tuned into a single-style single-speaker system with a small amount (0.5 hours) of the target style and target speaker recordings available [14]. In the current work, however, no target-speaker recordings with target-style rendering exist, so we explore style transplantation solely relying on VC-generated data: we apply VC on the conversational style corpus to augment the training set with synthetic conversational style utterances for all the existing non-conversational speakers, and remove the original conversational voice recordings from the training completely. To that end, we developed a prosody-preserving VC system, similar to [22], but based on the pretrained HuBERT [23] speech representation, thus not requiring the phonetic sequence as an input. The proposed VC system is tailored to the S2S TTS architecture, has a simpler training procedure than in [24], and exploits an alternative HuBERT feature discretization technique to facilitate Speaker ID disentanglement.

Let $X = [x_1, \cdots, x_T]$ denote a speech utterance of T frames and let S denote its speaker identity. Our VC system is a conditional Discrete Variational Autoencoder. During training, s corresponds to the actual speaker of X whereas at inference time, the conversion is achieved by forcing s to be a different speaker. The model is composed of the following parts:

- Encoder: The encoder is a pretrained HuBERT-Base model [23]. The first 6 layers are fixed during training while the rest 6 layers are fine-tuned. In addition, Gaussian upsampling [4] is used to resample from HuBERT frame rate to the TTS frame rate, thus preserving the original phone alignment.
- Discretization: In order to produce more stable representation, the encoder output is mapped from 768 into $N \times 2$ space, and discretized into $N = 32$ binary variables. To keep the discrete variable differentiable, we used the Gumbel-Softmax relaxation [25, 26] with a fixed temperature of 1. Let $h(x_t)$ denote the discrete outcome of the frame x_t
- Speaker integration: The discrete representation is forwarded to the decoder conditioned on the speaker identity s. We represent s as a one-hot vector and compute the latent $z_t = s_t \otimes h(x_t)$ where $\otimes$ is the outer product operation. We found this integration technique more effective than concatenation.
- Spectral decoder: The architecture for the decoder and the reconstruction loss function are the same as used in the proposed TTS model (§3), generating the acoustic features for the augmented training data.

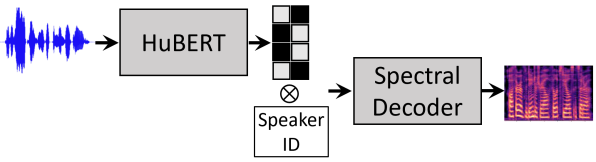

Figure 2: *Voice Conversion System for **Aug** TTS*

5. Experimental Setup

The training material comprises wide-band (22.05 kHz) speech corpora from three professional native speakers of US English uttered mostly in neutral speaking style in various domains, two females and one male, of 10-17k sentences each, and the conversational US English female voice corpus of 4k sentences (§2). Both of the non-conversational female voices were selected to evaluate conversational speaking-style transplantation. We trained the following multi-speaker TTS systems sharing the described architecture to assess different transplantation techniques. All systems were fine-tuned by adjusting the constant Prosody Control Offsets (as described in [6]).

- **Base**: A TTS system trained just on the recorded data.
- **Adv**: The baseline TTS system with Adversarial Speaker Classifier [20], applied on the Front-End Encoder output (trained just on the recorded data).
- **Aug**: The baseline TTS system, trained on the recorded voice corpora, excluding the conversational voice corpus, and augmented with the VC-generated conversational data (for all the three neutral speakers in the training corpora).

5.1. Automated Checkpoint Selection

Due to training-synthesis conditions mismatch in autoregressive architectures, selecting a final configuration and checkpoint can be time-consuming and require subjective listening [27]. To facilitate this, we introduce an automated-selection technique applied to each system. The central element of this technique is an objective assessment of a model checkpoint based on synthesizing speech from the texts corresponding to a held-out set of audio recordings taken as reference. For each synthesized sentence a correct falling/rising intonation pattern is assessed. To this end, for each synthesized-and-reference pair, the pitch contour slope is estimated within the time interval occupied by the last two words. The slope is estimated using a weighted linear regression of the f_0 contour with the weights set to the voicing values. Opposite signs of TTS and reference slope values suggest unnatural intonation patterns. All checkpoints, possibly corresponding to different model configurations, are ranked based on the counts of unnatural patterns found in all synthesized utterances. One or few top-ranked checkpoints are then selected for validation by human listeners.

6. Evaluation

We designed a set of subjective evaluations to assess how well the conversational style is transferred within various multi-speaker TTS systems using a test script of 70 conversational utterances simulating real CC use cases (for a speaker-similarity test only half were used), and manually annotated with interjections, dialog tags and word emphasis labels[2].

[2]Audio samples along with the corresponding annotated text inputs are available at http://ibm.biz/IS22-S2SConv.

Table 1: *Speaking style similarity ABX (Preference scores [%])*

Speaker	Sys1/Sys2	Sys1	No Pref.	Sys2	*p*-value
F1	**Aug/Base**	**61.11**	10.51	28.39	<0.01
	Aug/Adv	**63.33**	8.93	27.74	<0.01
	Adv/Base	43.10	11.83	45.08	0.232
F2	**Aug/Base**	**54.68**	12.99	32.34	<0.01
	Aug/Adv	**60.09**	12.05	27.86	<0.01
	Adv/Base	36.65	15.86	**47.49**	<0.01

All the tests in this work are designed to be parallel, so corresponding high quality synthetic TTS outcomes serve as test references. Specifically, we assess overall quality and naturalness by 5-scale MOS tests with neutral synthetic references (generated by **Adv**); speaker similarity by 5-scale ([-2:2]) MOS similarity test with two speaker references: *A*-reference (conversational speaker's synthetic samples, generated by **Base**) and *B*-reference (target speaker's neutral synthetic samples, generated by **Adv**), where subjects were presented with four scale categories from [28] plus *neither-A-nor-B* category in the middle (zero-value); and speaking style similarity by ABX tests with style reference (synthesized by **Base** with the conversational speaker ID). In the latter case, subjects were shown groups of three parallel audio samples, containing a pair of test samples generated with a target speaker ID and a reference-style sample generated with the conversational-speaker ID. (The reference-style samples were found to convincingly express the conversational speaking style, scoring 4.4 in a separate MOS test). All tests were conducted in crowd-sourced settings where subjects were screened with a transcription test to filter out insufficient English proficiency and/or bad listening conditions. Eventually, about 1000 votes per system were used in the style-similarity and MOS-quality tests, and about 500 votes per system were used in the speaker-similarity tests.

Table 2: *Speaker similarity [-2:2] with 95% conf. interval*

Spk/Sys	**Base**	**Adv**	**Aug**
F1	0.40 ± 0.11	**1.35** ± 0.09	0.42 ± 0.11
F2	0.92 ± 0.11	**1.41** ± 0.09	0.82 ± 0.11

Table 3: *Quality & naturalness MOS with 95% conf. interval*

Spk/Sys	Neutral	**Base**	**Adv**	**Aug**
F1	3.77±0.05	3.58±0.06	3.79±0.05	**3.96**±0.05
F2	3.91±0.05	3.77±0.05	3.89±0.05	3.94±0.05

7. Conclusions

The style similarity (Table 1) and quality (Table 3) evaluations reveal that **Aug** system significantly outperforms all the other systems in terms of transplanted style similarity while not degrading the overall naturalness. However, a speaker identity shift in **Aug** is still perceivable (Table 2). It is clear, though, that further improvement in the VC system quality would directly contribute to the resulting TTS speaker similarity improvement. This defines the main direction of our ongoing and future research, which also includes addressing the more difficult case of female-to-male style transplantation. We have also assumed in our evaluation a test case where user-input mark-up drives the different conversational sub-style renderings. A natural future extension to this work is to do this automatically to either suggest sensible defaults or to obviate this step from the user.

8. References

[1] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. Weiss, N. Jaitly, Z. Yang, Y. Ying Xiao, Z. Chen, B. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: A fully end-to-end text-to-speech synthesis model," *CoRR*, vol. abs/1703.10135, 2017. [Online]. Available: http://arxiv.org/abs/1703.10135

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

[3] S. Shechtman and A. Sorin, "Sequence to Sequence Neural Speech Synthesis with Prosody Modification Capabilities," in *Proc. 10th ISCA Speech Synthesis Workshop (SSW10)*, Vienna, Austria, 2019, pp. 275–280.

[4] J. Shen, Y. Jia, M. Chrzanowski, Y. Zhang, I. Elias, H. Zen, and Y. Wu, "Non-attentive Tacotron: Robust and controllable neural TTS synthesis including unsupervised duration modeling," *CoRR*, vol. abs/2010.04301, 2020.

[5] R. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. Weiss, R. Clark, and R. Saurous, "Towards end-to-end prosody transfer for expressive speech synthesis with Tacotron," *CoRR*, vol. abs/1803.09047, 2018. [Online]. Available: http://arxiv.org/abs/1803.09047

[6] S. Shechtman, R. Fernandez, A. Sorin, and D. Haws, "Synthesis of expressive speaking styles with limited training data in a multi-speaker, prosody-controllable sequence-to-sequence architecture," in *Proc. Interspeech*, 2021, pp. 4693–4697.

[7] A. Ben-David and S. Shechtman, "Acquiring conversational speaking style from multi-speaker spontaneous dialog corpus for prosody-controllable sequence-to-sequence speech synthesis," in *Proc. 11th ISCA Speech Synthesis Workshop (SSW 11)*, 2021, pp. 66–71.

[8] H. Guo, S. Zhang, F. Soong, L. He, and L. Xie, "Conversational End-to-End TTS for voice agents," in *IEEE Spoken Language Technology Workshop (SLT)*, Shenzhen, China, January 2021, pp. 403–409.

[9] J. Adell, D. Escudero, and A. Bonafonte, "Production of filled pauses in concatenative speech synthesis based on the underlying fluent sentence," *Speech Communication*, vol. 54, no. 3, pp. 459–476, March 2012.

[10] É. Székely, G. E. Henter, J. Beskow, and J. Gustafson, "How to train your fillers: uh and um in spontaneous speech synthesis," in *Proc. 10th ISCA Speech Synthesis Workshop (SSW10)*, Vienna, Austria, Sept. 2019, pp. 245–250.

[11] S. Shechtman, R. Fernandez, and D. Haws, "Supervised and unsupervised approaches for controlling narrow lexical focus in sequence-to-sequence speech synthesis," in *IEEE Spoken Language Technology Workshop (SLT)*, Shenzhen, China, January 2021, pp. 431–437.

[12] R. Fernandez, A. Rosenberg, A. Sorin, B. Ramabhadran, and R. Hoory, "Voice-transformation-based data augmentation for prosodic classification," in *Proc. ICASSP*, New Orleans, March 2017, pp. 5530–5534.

[13] T. Fukuda, R. Fernandez, A. Rosenberg, S. Thomas, B. Ramabhadran, A. Sorin, and G. Kurata, "Data augmentation improves recognition of foreign accented speech," in *Proc. Interspeech*, Hyderabad, India, 2018, pp. 2409–2418.

[14] G. Huybrechts, T. Merritt, G. Comini, B. Perz, R. Shah, and J. Lorenzo-Trueba, "Low-resource expressive text-to-speech using data augmentation," in *Proc. ICASSP*, Toronto, 2021, pp. 6593–6597.

[15] Switchboard, Details available at http://www.isip.piconepress.com/projects/switchboard/, 1997.

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

[17] "DIT++ Taxonomy of Dialogue Acts, Annotation Scheme, and DiAML Markup Language," https://dit.uvt.nl/#dit_taxonomy, 2019, release 5.2.

[18] J. M. Valin and J. Skoglund, "LPCNET: Improving neural speech synthesis through linear prediction," in *ICASSP*, Brighton, England, 2019, pp. 5891–5895.

[19] S. Shechtman, R. Rabinovitz, A. Sorin, Z. Kons, and R. Hoory, "Controllable sequence-to-sequence neural TTS with LPCNET backend for real-time speech synthesis on CPU," *CoRR*, 2020. [Online]. Available: http://arxiv.org/abs/2002.10708

[20] T. Nekvinda and O. Dušek, "One model, many languages: Meta-learning for multilingual text-to-speech," *Proc. Interspeech 2020*, pp. 2972–2976, 2020.

[21] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The journal of machine learning research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[22] S. Karlapati, A. Moinet, A. Joly, V. Klimkov, D. Sáez-Trigueros, and T. Drugman, "CopyCat: Many-to-many fine-grained prosody transfer for neural text-to-speech," in *Proc. Interspeech*, Shanghai, China, 2020, pp. 4387–4391.

[23] W.-N. Hsu, Y.-H. H. Tsai, B. Bolte, R. Salakhutdinov, and A. Mohamed, "HuBERT: How much can a bad teacher benefit ASR pre-training?" in *Proc. ICASSP*. Toronto, Canada: IEEE, 2021, pp. 6533–6537.

[24] A. Polyak, Y. Adi, J. Copet, E. Kharitonov, K. Lakhotia, W.-N. Hsu, A. Mohamed, and E. Dupoux, "Speech resynthesis from discrete disentangled self-supervised representations," in *Proc. Interspeech*, Brno, Czech Republic, 2021, pp. 3615–3619.

[25] C. J. Maddison, A. Mnih, and Y. W. Teh, "The concrete distribution: A continuous relaxation of discrete random variables," *arXiv preprint arXiv:1611.00712*, 2016.

[26] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with gumbel-softmax," *arXiv preprint arXiv:1611.01144*, 2016.

[27] S. Shechtman, D. Haws, and R. Fernandez, "Stable checkpoint selection and evaluation in sequence to sequence speech synthesis," in *Proc. ICASSP*, Toronto, Canada, 2021, pp. 7093–7097.

[28] M. Wester, Z. Wu, and J. Yamagishi, "Analysis of the voice conversion challenge 2016 evaluation results." in *Proc. Interspeech*, San Francisco, USA, 2016, pp. 1637–1641.