PROSODY-CONTROLLABLE SPONTANEOUS TTS WITH NEURAL HMMS

Harm Lameris, Shivam Mehta, Gustav Eje Henter, Joakim Gustafson, Éva Székely

Division of Speech, Music and Hearing, KTH Royal Institute of Technology, Stockholm, Sweden

ABSTRACT

Spontaneous speech has many affective and pragmatic functions that are interesting and challenging to model in TTS. However, the presence of reduced articulation, fillers, repetitions, and other disfluencies in spontaneous speech make the text and acoustics less aligned than in read speech, which is problematic for attention-based TTS. We propose a TTS architecture that can rapidly learn to speak from small and irregular datasets, while also reproducing the diversity of expressive phenomena present in spontaneous speech. Specifically, we add utterance-level prosody control to an existing neural HMM-based TTS system which is capable of stable, monotonic alignments for spontaneous speech. We objectively evaluate control accuracy and perform perceptual tests that demonstrate that prosody control does not degrade synthesis quality. To exemplify the power of combining prosody control and ecologically valid data for reproducing intricate spontaneous speech phenomena, we evaluate the system's capability of synthesizing two types of creaky voice.

Index Terms: Speech Synthesis, Prosodic Control, Neural-HMM, Spontaneous speech, Creaky voice

1. INTRODUCTION

In recent years, the quality of end-to-end deep neural-network-based text-to-speech (TTS) architectures have improved to rival human speech [1]. Two issues faced by state-of-the-art TTS are a lack of ecological validity [2], how well the speech data reflects the context of the use-case, and expressivity in the learned prosody. Most architectures are trained on read-speech corpora, e.g., [3, 4] that have limited prosodic coverage, and generate "average" prosody. Moreover, the prosody is generated solely from text, which allows no control over the generated style [5]. Concurrently, spontaneous speech is increasingly used in TTS [6, 7]. Spontaneous speech data is challenging to model, due to disfluencies and large variability [8]; offers high ecological validity for evermore commonplace conversational AI systems, and the varied prosody offered by spontaneous data positively impacts factors such as word recall and attention [9]. For conversational systems it would also be useful to be able to synthesize creaky voice, as this have been found to be a strong turn-yielding cue [10, 11].

Several approaches exist for prosody-controlled TTS. In [12], the authors use a Tacotron 2 architecture for which the decoder is conditioned on a prosodic reference encoder. During training, the prosodic reference encoder generates prosody embeddings from spectrogram slices in an unsupervised manner. At inference, the prosody of a reference audio file is transferred to the target audio. In [13] a quantized fine-grained VAE with an autoregressive prosody prior is used which learns a latent representation of the prosody from the aligned spectrogram.

The approaches closest to this paper appear in [14] and [15], a modified version of [16]. In [14] the authors introduce a hierarchical model based on Tacotron 2 that uses a separate prosody encoder to predict sentence-wise pitch, phone duration, speech energy, and spectral tilt. This allows for the control of these features and the production of a variety of styles, while achieving similar mean opinion scores to a baseline Tacotron 2 model. This hierarchical model requires a large amount of data, for which sufficient-length spontaneous speech corpora do not exist, and requires enormous amount of resources to train. In [15] Tacotron 2 is modified by appending embedding values for the pitch, loudness, and duration to the output of the encoder before being passed to an augmented attention mechanism.

In this paper we study the effect of using spontaneous speech data for TTS with prosodic feature control: We use a method based on neural hidden Markov models (HMM) TTS [17], equivalent to a kind of transducer TTS [18]. Important for spontaneous speech applications, neural HMMs [19] force monotonic alignments between input symbols and output frames, which helps to train rapidly and on smaller, more disorderly datasets than neural TTS based on conventional neural attention [17]. The statewise nature of the neural HMM is also appealing for modelling disfluencies and other speech irregularities that have been transcribed with discrete tokens. Here, there is a possibility to represent speech phenomena such as disfluencies, partial repetitions, under-articulated speech segments, etc., that cause the alignment between speech and transcription to be lower in a sponta-

This research was supported by the Swedish Research Council projects Connected (VR-2019-05003), Perception of speaker stance (VR-2020-02396), the Riksbankens Jubileumsfond project CAPTivating (P20-0298), and by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

neous speech corpus than in a read speech corpus [20]. We extend neural HMM TTS to learn a control space for the mean and variation of the fundamental frequency (f0) and speech rate (syllables per second). By using spontaneous data and with the control of the f0, we can implicitly manipulate voice quality through the synthesis of creaky voice, a type of phonation where the vocal cords are relaxed to create a low, irregular pitch with a low harmonics to noise ratio [21]. Creaky voice has not been synthesized using neural TTS, as its non-periodicity makes it complex to measure. Opting for low mean and variation in f0 synthesis enables us to produce stylistic and end-of-turn creaky voice. We perform a data analysis, an objective evaluation of the control of the feature modification space, an expert analysis of the creaky voice quality produced by the system, and a subjective evaluation of the synthesized speech of our system. Audio samples are provided at www.speech.kth.se/tts-demos/prosodic-hmm.

2. METHOD

2.1. Data

We used three datasets to train systems. For our base-model we used a scripted conversational corpus, RyanSpeech corpus [22]. This corpus contains 10 hours (11,279 utterances) of a male speaker of US English reading textual materials from conversational settings. The spontaneous model was trained on a corpus created from the audio of the Trinity Speech-Gesture dataset [23], which consists of 25 impromptu monologues by a male voice actor speaking Hiberno-English, using an impromptu, colloquial style. For one of the evaluations, we also used the industry standard LJSpeech[1].

We pre-processed the spontaneous corpus by segmenting the monologues into breath groups, i.e., single stretches of speech between two breath events, as was performed in [6, 24]. We opted for breath groups as a unit since we hypothesize that minimal style changes occur within a single breath group. Using breath groups also enables the possibility to change style within a given utterance by inserting a breath. The breath groups were combined into bigrams to create audio files of up to 11 seconds [25]. We extracted the mean and standard deviation of the fundamental frequency and mean speech rate per breath group using the Wavelet Prosody Toolkit (WPT) [26] to create three prosodic features: f0 variability (per-utterance standard deviation of f0), pitch (mean f0), and speech rate (syllables per second). Other prosodic features like energy and spectral tilt could also be included. The feature values were z-standardized before training.

2.2. Model architecture

A modified version of neural HMM TTS [17] was used, which is an auto-regressive TTS architecture that synthesises

[1]https://keithito.com/LJ-Speech-Dataset/

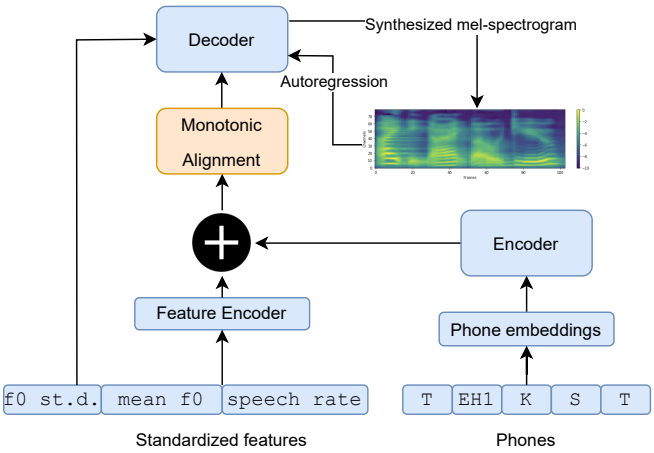


Fig. 1: Model architecture

mel-spectrograms conditioned on input text. It follows an encoder-decoder architecture similar to [3] but instead of a cumulative attention, it uses a left-to-right no-skip HMM (defined by a neural network) to force monotonic alignments between text inputs and mel spectrogram frames. In addition to the already present CNN + Bi-LSTM based encoder in neural HMM we added a *feature encoder* which contains a single feed-forward layer (Fig. 1) to project features into a 512-dimensional space. This modification is used to project the audio features into a more expressive control space. After standardizing the features, we use a two-step conditioning method to incorporate the prosodic features. The output of the feature encoder is first projected into the same dimensionality as the phone embeddings, and the two are concatenated to define the final states of the HMM. Additionally, we append a skip connection that adds the standardized prosodic features to the outputs of the encoder. The skip connection provides more robust control over the synthesis, as both encoder and decoder are conditioned on the prosodic features.

3. EXPERIMENTS

3.1. Experimental setup

To investigate the level of prosodic variation, we first performed a data analysis comparing the per-utterance mean natural logarithm of the f0 (log f0), as well as the per-utterance speech rate for each dataset. We then trained two spontaneous TTS voice models, one baseline model trained on standard Neural HMM TTS, and our proposed model with prosody control, by pre-training on RyanSpeech [22] for 24,000 iterations with batch size 32, and then finetuning the models on spontaneous speech audio from the Trinity Speech and Gesture Dataset [23] for 9,500 iterations with batch size 20.

To examine how the control space for the prosodic features was learned, we performed an objective analysis of the synthesized utterances from the proposed model and the base-

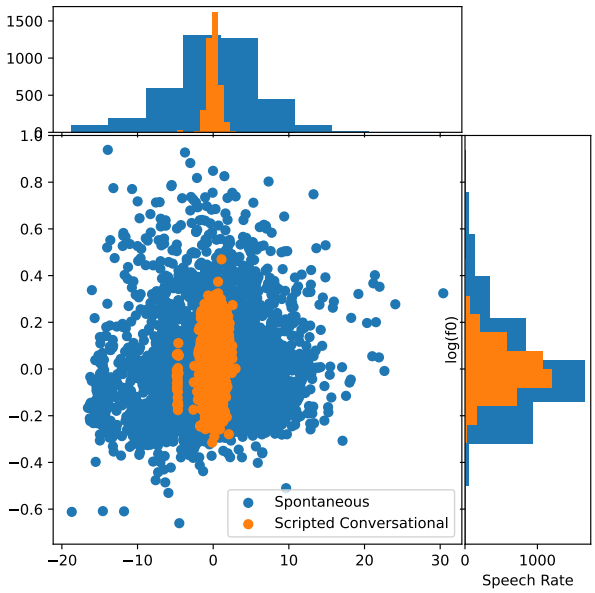


Fig. 2: The mean per-utterance log f0 and mean per-utterance speech rate for the spontaneous and scripted conversational corpora

line, in which we investigate the distribution of the features that were compared in the data analysis using gradual increments of the prosodic feature control. We synthesized 280 utterances per prosodic feature, corresponding to 40 utterances for each feature setting, modifying each standardized input feature between [-3, 3] standard deviations from the mean, while keeping the other features constant. We also conducted two subjective listening tests in which we compared the proposed system to the baseline, and tested for quality degradation of the modified utterances by comparing them to a reference of the mean features in a MUSHRA-like CMOS.

To exemplify the system's capability to synthesize characteristic spontaneous speech with a range of voice qualities, we conducted an additional perceptual evaluation focusing on two different types of creaky voice which are implicitly achievable by adjusting the prosodic features.

3.2. Data analysis

To compare the distribution of the fundamental frequency and speech rate, we randomly selected 4009 audiofiles from the scripted conversational corpus Ryanspeech [22] to compare to the spontaneous speech corpus TSGD [23]. Figure 2 shows the distribution of the mean of the per-utterance log f0, as well as the speech rate per utterance for the spontaneous and scripted conversational corpora. The values are centred around the corpus mean f0 and speech rate. In the figure, one can see that especially the speech rate is much more variable in the spontaneous speech corpus than in the scripted conversational corpus, with the speech rate for the scripted conversational corpus being closely centred around the mean, while the spontaneous corpus is a widely clustered around its mean. This is reflected in the peakier shape of the distribution for the scripted corpus. The mean log f0 also displays more diversity and a larger range of values in the spontaneous corpus, especially for the higher-pitched datapoints. The log f0 is similarly distributed for both corpora, although it covers a larger range of values for the spontaneous corpus.

System	MOS	Confidence Interval
NHMM	3.60	[3.51,3.69]
Proposed	3.48	[3.40,3.57]

Table 1: The results of the subjective MOS evaluation

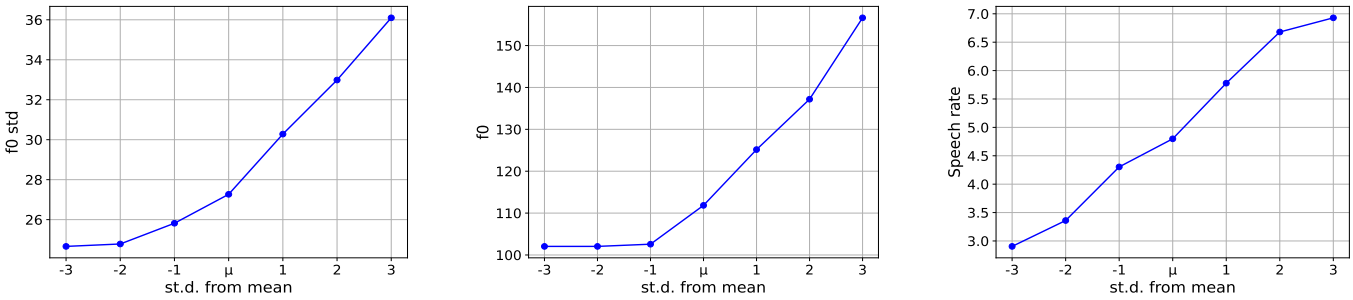
3.3. Objective analysis

For the objective evaluation we generated 50 utterances from a held out set for various points in the feature space ranging from [-3, 3] per feature. From these utterances we computed f0 and speech rate again with WPT. The results of this evaluation are shown in Figure 3. As can be seen in 3a, control over the f0 variability is especially predictable in the [-2, 3] standard deviations from the mean range. The smaller differences for the [-3, -2] range can be explained by the low concentration of data present (2.14%).

For the mean f0, figure 3b shows there is predictable control across the [-1, 3] st.d. from the mean range. After listening to the utterances between [-3, -1] st.d. and examining the output from WPT, the cause of this is the creaky voice quality of both the actor's and synthesized speech in this range. Creaky voice does not produce accurate f0 readings due to its lack of periodicity. We perceived -3 st.d. from the mean as having a more intense creak than -2 st.d. from the mean, suggesting that instead of lower pitch, this modification changes the level of creakiness. Figure 3c indicates the control over the speech rate, and shows predictable control throughout the [-3, 3] st.d. from the mean range. Informal evaluation also highlighted an absence of interaction between these features, indicating the ability to vary the feature space individually for each feature.

3.4. Perceptual evaluation

In the subjective evaluation, 44 native English speakers recruited through Prolific were presented with 40 samples of spontaneous synthesized speech, to be rated on a 5-point MOS scale. The stimuli consisted of 20 samples created with the baseline system without prosodic control, and 20 samples where the proposed system's feature values match the mean and variation of f0 and the speech rate of the non-modified stimulus, as extracted by WPT. Table 1 shows the participants' ratings. The confidence intervals on the results



(a) f0 st.d. (b) Mean f0 (c) Speech rate

Fig. 3: The effect of input feature modification on the corresponding speech property in the output speech

show that the addition of prosodic control did not result in a degradation of quality.

We conducted an additional experiment to further test the hypothesis that prosody modification did not decrease audio quality. For this evaluation, we recruited 30 participants for a MUSHRA-like 7-point CMOS of the quality of the individual feature modifications on utterances. In the evaluation, participants were asked to rate the quality of 4×10 sentences synthesized at [-1.5, -0.5, 0.5, 1.5] st.d. from the mean to a reference which was synthesized at the mean for each feature. Wilcoxon signed-rank tests found only significant differences in favour of the modified utterances, indicating that prosodic changes did not harm synthesis quality.

3.5. Evaluation of synthesis of creaky voice

To evaluate the presence and naturalness of creaky voice, we synthesized 10 utterances for our spontaneous voice, as well as for an LJSpeech voice finetuned for 9k iterations over the spontaneous model in 3 intended styles: modal voice quality, creaky voice as a stylistic expression (throughout an utterance), and end-of-turn creak by adjusting the features until the intended style was produced. To verify the presence of creaky voice in the samples, we performed an analysis of the creakiness with COVAREP [27]. In addition, we conducted an expert listening evaluation asking 15 people with training in linguistics or speech technology to rate what percentage of the utterances is creaky using a sliding scale with the range of 0-100 (Table 2).

Participants were also asked to rate the naturalness of the creak and supply general comments. Nearly all participants rated the creakiness as natural. The experts commented on the influence of creakiness on the perception of the speaker's mood, and mentioned the ability to distinguish the *strength* as well as the *extent* of the creak.

A one-way ANOVA with post-hoc Tukey showed a significant difference between no creak and the two styles of creak ($p<0.001$) and no significant difference between the two styles of creak for the spontaneous voice, whereas all categories were significantly different for LJSpeech (all $p<0.01$). Some utterances designed as non-creaky were still perceived as containing some creak, possibly due to the vocal characteristics of the speaker or vocoder artefacts, although these were always rated as less creaky than the creaky styles.

Creak Type	Expert mean score		Measured creak	
Dataset	LJSpeech	TSGD	LJSpeech	TSGD
none	22.6±6.6	33.6±4.9	5.7	0.4
stylistic	72.3±4.4	59.7±5.0	23.2	8.1
end-of-turn	45.0±3.8	53.8±5.0	18.2	7.5

Table 2: The mean reported creakiness with confidence intervals and avg. creaky segments % measured in Covarep.

4. CONCLUSIONS

We present an architecture for the prosodic modification of spontaneous speech, which is difficult to model with data-hungry attention-based architectures due to the highly complex and varied nature of spontaneous speech. We demonstrate that spontaneous speech is more varied than scripted conversational speech for per-utterance mean f0 and speech rate. Our objective analysis shows that the modelling of prosodic features provides control over the variation and mean of f0 and the speech rate of synthesized speech. To show that the prosody-modifiable feature of the synthesizer does not degrade quality, we carried out two perceptual evaluations, in which the modified synthesized speech was rated similarly or better than non-modified speech. Finally, we conducted an additional experiment to showcase the system's ability to exhaust the possibilities in varied speech data by synthesizing natural sounding creaky voice: both as a stylistic feature, as well as in utterance-final position.

This work provides a framework for future research to utilize more spontaneous speech corpora, which most closely correspond to real conversational speech. The demonstrated capability of synthesizing naturalistic and varied creaky voice lends itself to the investigation of more explicit control for voice quality dimensions.

5. REFERENCES

[1] X. An, F. K. Soong, S Yang, and L. Xie, "Effective and direct control of neural TTS prosody by removing interactions between different attributes," *Neural Networks*, vol. 143, pp. 250–260, 2021.

[2] M. Wester, O. Watts, and G. E. Henter, "Evaluating comprehension of natural and synthetic conversational speech," in *Proc. Speech Prosody*, 2016, pp. 736–740.

[3] J. Shen, R. Pang, R.J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, et al., "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," in *Proc. ICASSP*, 2018, pp. 4779–4783.

[4] J. Kim, S. Kim, J. Kong, and S. Yoon, "Glow-TTS: A generative flow for text-to-speech via monotonic alignment search," *Proc. NeurIPS*, pp. 8067–8077, 2020.

[5] D.S.R. Mohan, V.J. Hu, T.H. Teh, A. Torresquintero, C.G.R. Wallis, M. Staib, et al., "Ctrl-P: Temporal control of prosodic variation for speech synthesis," in *Proc. Interspeech*, 2021, pp. 3875–3879.

[6] J Gustafson, J. Beskow, and É. Székely, "Personality in the mix-investigating the contribution of fillers and speaking style to the perception of spontaneous speech synthesis," in *Proc. SSW*, 2021, pp. 48–53.

[7] A.O. Adigwe and E. Klabbers, "Strategies for developing a conversational speech dataset for text-to-speech synthesis," in *Proc. Interspeech*, 2022, pp. 2318–2322.

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

[9] E. Rodero, R. F. Potter, and P. Prieto, "Pitch range variations improve cognitive processing of audio messages," *Hum. Commun. Res.*, vol. 43, no. 3, pp. 397–413, 2017.

[10] R. Ogden, "Turn transition, creak and glottal stop in Finnish talk-in-interaction," *Journal of the International Phonetic Association*, vol. 31, no. 1, pp. 139–152, 2001.

[11] A. Gravano and J. Hirschberg, "Turn-taking cues in task-oriented dialogue," *Comput. Speech Lang.*, vol. 25, no. 3, pp. 601–634, 2011.

[12] R. J. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. Weiss, et al., "Towards end-to-end prosody transfer for expressive speech synthesis with Tacotron," in *Proc. ICML*, 2018, pp. 4693–4702.

[13] G. Sun, Y. Zhang, R. J Weiss, Y. Cao, H. Zen, A. Rosenberg, et al., "Generating diverse and natural text-to-speech samples using a quantized fine-grained VAE and autoregressive prosody prior," in *Proc. ICASSP*, 2020, pp. 6699–6703.

[14] T. Raitio, J. Li, and S. Seshadri, "Hierarchical prosody modeling and control in non-autoregressive parallel neural TTS," in *Proc. ICASSP*, 2022, pp. 7587–7591.

[15] Slava Shechtman and Alex Sorin, "Sequence to sequence neural speech synthesis with prosody modification capabilities," in *Proc. SSW*, pp. 275–280.

[16] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," in *Proc. ICLR*, 2020.

[17] S. Mehta, É. Székely, J Beskow, and G. E. Henter, "Neural HMMs are all you need (for high-quality attention-free TTS)," in *Proc. ICASSP*, 2022, pp. 7457–7461.

[18] Y. Yasuda, X. Wang, and J. Yamagishi, "Initial investigation of encoder-decoder end-to-end TTS using marginalization of monotonic hard alignments," in *Proc. SSW*, 2019, pp. 1–6.

[19] K. M. Tran, Y. Bisk, A. Vaswani, D. Marcu, and K. Knight, "Unsupervised neural hidden Markov models," in *Proc. SPNLP*, 2016, pp. 63–71.

[20] É. Székely, J. Edlund, and J. Gustafson, "Augmented prompt selection for evaluation of spontaneous speech synthesis," in *Proc. LREC*, 2020, pp. 6368–6374.

[21] Patricia A Keating, Marc Garellek, and Jody Kreiman, "Acoustic properties of different kinds of creaky voice.," in *ICPhS*, 2015, vol. 2015, pp. 2–7.

[22] R. Zandie, M. H. Mahoor, J. Madsen, and E. S. Emamian, "Ryanspeech: A corpus for conversational text-to-speech synthesis," in *Proc. Interspeech*, 2021.

[23] Y. Ferstl and R. McDonnell, "Investigating the use of recurrent motion modelling for speech gesture generation," in *Proc. IVA*, 2018, pp. 93–98.

[24] 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, pp. 245–250.

[25] É. Székely, G. E. Henter, J. Beskow, and J. Gustafson, "Breathing and speech planning in spontaneous speech synthesis," in *Proc. ICASSP*, 2020, pp. 7649–7653.

[26] A. Suni, J. Šimko, D. Aalto, and M. Vainio, "Hierarchical representation and estimation of prosody using continuous wavelet transform," *Comput. Speech Lang.*, vol. 45, pp. 123–136, 2017.

[27] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "COVAREP—a collaborative voice analysis repository for speech technologies," in *Proc. ICASSP*, 2014, pp. 960–964.