CYBORG SPEECH: DEEP MULTILINGUAL SPEECH SYNTHESIS FOR GENERATING SEGMENTAL FOREIGN ACCENT WITH NATURAL PROSODY

Gustav Eje Henter[1], *Jaime Lorenzo-Trueba*[1], *Xin Wang*[1], *Mariko Kondo*[2], *Junichi Yamagishi*[1,3]

[1]National Institute of Informatics, Tokyo, Japan [2]Waseda University, Tokyo, Japan
[3]The University of Edinburgh, Edinburgh, UK

ABSTRACT

We describe a new application of deep-learning-based speech synthesis, namely multilingual speech synthesis for generating controllable foreign accent. Specifically, we train a DBLSTM-based acoustic model on non-accented multilingual speech recordings from a speaker native in several languages. By copying durations and pitch contours from a pre-recorded utterance of the desired prompt, natural prosody is achieved. We call this paradigm "cyborg speech" as it combines human and machine speech parameters. Segmentally accented speech is produced by interpolating specific quinphone linguistic features towards phones from the other language that represent non-native mispronunciations. Experiments on synthetic American-English-accented Japanese speech show that subjective synthesis quality matches monolingual synthesis, that natural pitch is maintained, and that naturalistic phone substitutions generate output that is perceived as having an American foreign accent, even though only non-accented training data was used.

Index Terms— Multilingual speech synthesis, phonetic manipulation, foreign accent, DNN

1. INTRODUCTION

Not all speech synthesis systems are created with the primary goal of facilitating the interaction between man and machine – the ability for a computer to mimic human speech also opens a window into human perception of speech. This paper describes how modern deep-learning-based speech synthesis can provide new tools for perception research. Specifically, we demonstrate a novel application of neural-network-based speech synthesis, which is to synthesise foreign-accented speech from unaccented speech recordings alone.

The synthesis approach we describe allows phonetic output control at the segmental level while recreating the prosody of natural speech recordings. This is particularly appealing for "microscope studies" of speech perception which explore the effect of carefully-defined phonetic alterations that are well resolved in time and degree. The speech stimuli we generate embody pronunciation manipulations that are difficult to elicit from human speakers, and complement more labour-intensive stimulus-creation tools from speech perception such as splicing or manipulating recorded speech signals in a parametric/vocoder representation.

Compared to existing speech synthesisers, especially in perception and foreign-accent research as described in Sec. 2, our main achievements are:

1. *Using deep learning to generate high-quality speech with controllable, segment-level foreign accent*, despite using only native multilingual speech recordings as input.
2. *Replicating prosodically-relevant speech properties* (pitch and phone durations) *as extracted from natural speech recordings*, thus circumventing any concerns over the sometimes inadequate prosody of conventional text-to-speech.

Our approach – detailed in Sec. 3 – is applicable to any number of languages in combination (assuming appropriate multilingual data is available) and a multitude of different phonetic manipulations. We show in objective and subjective experiments (Sec. 4) that our proposal follows the pitch contour of the natural speech and generates noticeably and characteristically accented speech while matching the segmental quality of a monolingual baseline synthesiser.

2. BACKGROUND

2.1. Foreign-Accent Research

By comparing listeners' reactions to speech designed to differ in specific aspects, researchers can shed light on how humans perceive speech. This methodology can be used to study the phonemic and perceptual basis of foreign accent (FA), and to disentangle the individual cues that underpin it. Historically, most FA research has focused on supra-segmental accent properties, including intonation and pauses [1], nuclear stress [2], duration [3], and speech rate [4]. Interestingly, however, [5] instead highlighted segmental errors as the phonetic cues most strongly responsible for conveying FA.

Unfortunately, it is not straightforward to create speech stimuli that isolate the effect of different segmental mispronunciations on accentedness. Even professional voice talents find it difficult to produce speech with isolated nonnative mispronunciations, not to mention speaking in intermediate degrees of accent. Cross-language splicing is labour-intensive and prone to artefacts at the joins. Consequently, [6] introduced the idea of stimulus generation through multilingual speech synthesis. Our work can be seen as an evolution of that idea, but using deep learning to generate foreign accent.

2.2. DNNs and Speech Synthesis for Perception Research

When creating synthetic speech stimuli for perception research, high segmental quality is a priority: research suggests that human perception processes differ between natural and (formant) synthesised speech due to a dearth of natural acoustic cues in the synthetic audio [7, 8]. While formant synthesis controlled by rules [9, 10] is the classic synthesis paradigm in perception research, a handful of recent studies (e.g., [11, 6]) have considered hidden Markov model (HMM) text-to-speech (TTS) coupled with decision-tree acoustic models for controllable speech synthesis with speech perception applications. Synthesis based on deep recurrent neural networks (RNNs) [12, 13] improve on the segmental quality of decision trees [14, 15] through

This work was partially supported by MEXT KAKENHI Grant Numbers (15H02729, 17K12720).

statistical learning methods that better replicate the acoustic properties (and cues) in natural speech recordings.

Another important advantage of deep learning is that it also provides great flexibility for controlling the output speech signal, e.g., to vary speaker identity [16], expression [17], or emotion [18]. A single deep neural network acoustic model can even be trained to perform multilingual TTS [19], though that requires redesigning the architecture of the synthesiser. This paper similarly describes a new RNN synthesiser architecture capable of multilingual speech synthesis with segmental control of foreign accent, while maintaining the prosodic characteristics of natural speech. In particular, we propose to extract durations and pitch contours from native, natural speech recordings, and use these as inputs rather than outputs in a deep LSTM-based acoustic model. Absent such recordings, one can always predict durations and pitch from text using dedicated, high-accuracy methods like [20].

3. METHOD

We will now describe how to practically realise our goal of creating a multilingual speech synthesiser with phonetic control and given prosodic characteristics. Our plan is to re-use the durations (phone timings) and the pitch contour (fundamental frequency F0 and voiced/unvoiced decision) of natural, native speech productions of the text prompt to be generated. This requires that such recordings are available prior to synthesis, either by only synthesising held-out prompts from the training data, or by custom-recording natural productions of the prompts to be synthesised. Alternatively, durations and F0 can be predicted externally, e.g., through conventional TTS methods, but the prosody will then be defined by a machine.

Since our speech durations are given, no duration model is needed. Our task is instead like acoustic modelling (predicting frame-wise inputs to a vocoder) in conventional TTS, except that pitch values also are provided. The vocoder parameters that remain to be predicted are source aperiodicities and filter spectra, which capture speech phonation quality and articulation; our approach is to learn to predict these with an RNN. Unusually, the system that results from this setup is neither text-to-speech nor speech-to-speech (voice conversion), but requires both text *and* speech input to generate audio. As the resulting speech output is a chimeric combination (cf. [21]) of man and machine – natural and predicted speech parameters – we dub this paradigm *cyborg speech*.

3.1. Data

For this study, we had access to high-quality studio recordings of a male professional voice talent speaking both US English and Japanese natively. The recordings comprised 2000 utterances in each language (3 h 29 min of Japanese and 4 h 15 min minutes of English) plus a separate set of 20 test utterances in each language. All audio recordings were downsampled to 48 kHz at 16 bits per sample, normalised to -6 dB below clipping, and limited to a single channel.

Acoustic analysis and synthesis were performed using WORLD [22] with 5 ms frame step. However, due to frequent voicing errors in the WORLD F0 extraction, particularly in Japanese devoiced vowels, we used pitch contours from the GlottDNN pitch extractor in [23] instead. This substantially improved analysis-synthesis quality. The obtained WORLD spectra and aperiodicities were subsequently reduced to 60 mel-generalised cepstrum coefficients (MGCs) and 25 band aperiodicities (BAPs) plus their Δ and Δ^2 coefficients.

To associate each audio frame with a phone, we performed forced alignment using two monolingual synthesisers based on HTS [24], one for each language. English text processing and linguistic feature extraction was performed using Flite [25] with the Combilex [26] General American (GAM) phonetic dictionary (54 metaphonemes), while the Japanese text used Open JTalk [27] with a standard phone set (44 phones).

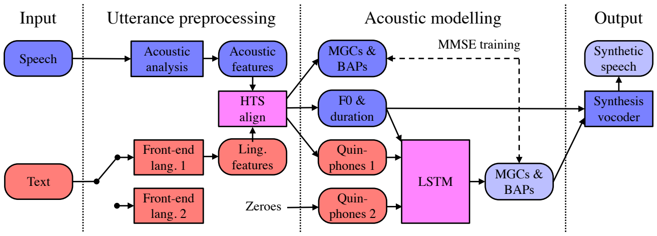

Fig. 1. Overview of bilingual cyborg speech synthesis system.

3.2. Acoustic Model Inputs

Typical TTS front-ends extract a diverse set of features from the text. This includes phonetic context (quinphones) and a variety of other additional, typically language-dependent, features. While the use of phonetic context is widespread across the most popular front-ends, other linguistic features vary substantially across languages and systems, making them hard to reconcile in a multilingual application. However, since these additional features primarily are used for improving the prediction of prosody, they are not necessary for our application; hence we ignored any such extra front-end features, keeping only five-hot quinphone context as our only text-derived RNN input, as phones are necessary for enabling phonetic control.

To obtain multilingual quinphones, we took the union of the phone sets in all languages considered, treating duplicate symbols across languages as distinct phones. This gave a total of 98 phones for our bilingual system. In practise, our scheme is implemented by zero-padding conventional five-hot monolingual quinphone vectors features to reach the dimensionality of the target multilingual quinphone feature vector. A more advanced setup could encode multilingual phones based on IPA and articulatory features. This would open the door to even more discriminate pronunciation control.

Aside from the quinphone features, we added a binary language flag indicating the language of the current frame. This flag can be generalised to a one-hot vector in applications with more than two languages. While the language flag in our case is redundant given that our multilingual identities already encode language information, that would not necessarily be the case if an articulatory or IPA-based phonetic input representation was used. Regardless of encoding, it is straightforward to interpolate between the input feature values representing different phones and languages, enabling speech output with intermediate phonetics or articulation, in mixed languages.

Since we have access to reference durations and F0 values, it makes sense to provide these as inputs to the acoustic model, especially since this benefits subjective output quality [14]. Following [15], we used three duration-dependent features as input for each frame. For pitch input, we used the voicing flag and (interpolated) log F0 and its Δ and Δ^2 features as frame inputs, which enables us to learn and replicate dependencies between voicing frequency and the acoustic features of the speech. A schematic overview of the complete system and its training is provided in Fig. 1.

3.3. RNN Training and Synthesis

With the above setup, one obtains an acoustic modelling problem where each frame has output dimensionality 255 (static and dynamic

Japanese		English		Substitutions made	
IPA	Open JTalk	IPA	Combilex GAM	Max	No. prompts
ɾ	r	ɹ	r	9	19
ɕ	sh	ʃ	S	8	13
dz	z	z	z	5	7
dʑ	j	dʒ	dZ	3	8
tɕ	ch	tʃ	tS	2	11

Table 1. Cross-language phone substitutions to emulate American-accented Japanese. The last two columns count the number of substitutions made in the most affected prompt, and the number of test prompts with at least one substitution.

MGCs and BAPs, but no F0) and input dimensionality 228 (Japanese), 278 (English), or 498 (bilingual). We used a DBLSTM with the same design and dimensions as in [28] to learn a mapping from input sequences to output sequences. Network weights were initialised with small random values and then trained using CURRENNT [29] to minimise mean-squared error (MSE). Approximately five percent of the training-data utterances were randomly held out for validation, while the 20 designated test sentences were kept aside for final evaluation. All input and output features except the binary language flag were normalised to zero mean and unit variance.

Optimisation proceeded using minibatch stochastic gradient descent (SGD) in two stages: first, raw SGD with learning rate 2.5×10^{-6} and no momentum was applied for 160 epochs; then AdaGrad [30] with learning rate 0.001 was applied to the previously obtained network. Both stages used early stopping, though this really only affected the AdaGrad stage, where it always terminated training within 30 epochs. The model with the best validation-set performance across all of training was used for the experiments.

Three different systems were trained: one on the Japanese data, one on the English data, and one on the bilingual data over the joint phone set. The final validation-set performance (frame MSE) was 83.51 on the Japanese data, 85.74 on the English data, and 84.87 on the combined data, falling between the two monolingual systems.

Predicted static and dynamic parameter sequences were reconciled using MLPG [31] and then fed into WORLD to generate synthetic speech stimuli for the test set utterances. Standard postfiltering was applied to enhance the clarity of the speech.

4. EXPERIMENTS

To verify the abilities of our approach to generate native and foreign-accented speech with maintained prosody, we evaluated several aspects of our trained multilingual synthesiser. While our approach is able to speak in, and mix accents from, any language provided as training data, we concentrated on Japanese stimuli, since we only had easy access to native Japanese listeners for the subjective evaluation. We contrasted performance against the trained monolingual Japanese system, not the English monolingual system, due to the absence of appropriate listeners.

4.1. Stimulus Generation

Our test material comprised 20 natural Japanese-language utterances between 7.4 and 10.8 seconds in duration. Their durations, pitch contours, and corresponding text prompts were used as inputs to the Japanese monolingual (denoted "MON") and bilingual synthesiser (denoted "BIL") to generate 20 synthetic speech stimuli for each system. We also compared against the 20 natural recordings from the test set (denoted "NAT") and analysis-synthesis stimuli (denoted "VOC") obtained from the extracted F0, MGC, and BAP features.

Among the four stimulus generators considered, only BIL is capable of interpolating between languages as necessary for generating foreign accent. We decided to investigate this novel feature in a simple application to creating American English-accented Japanese speech stimuli. Specifically, we considered five distinct consonant substitutions inspired by common mispronunciations among native speakers of American English (L1) speaking Japanese as a foreign language (L2). All of these mispronunciations can be emulated by simply replacing specific Japanese phones in the test prompt phonetisations by a corresponding phone in the Combilex General American phone set, without changing pitch or duration. The chosen substitutions and their prevalence are described in Tab. 1.

The manipulations in Tab. 1 represent five different directions in which Japanese speech can be altered to simulate an American accent. We also created stimuli combining all substitutions simultaneously, a manipulation labelled "all". Two degrees of interpolation were considered for each manipulation: either completely replacing all quinphone references to the substituted phone by a phone from the other language (degree 1.0), or half-way interpolation by averaging the original and substituted input vectors (degree 0.5). The language flag was similarly altered on frames where the centre phone came from the English phone set. This produced a total of twelve synthetically accented stimuli for each of the 20 test-set prompts.

We use the term *condition* to refer to the combination of system (NAT, VOC, MON, BIL), manipulation (none, r, sh, z, j, ch, all), and degree of interpolation (0.0, 0.5, 1.0). Each of the 16 conditions considered is associated with one stimulus (waveform) per text prompt, for a total of 320 stimuli included in the evaluation.

4.2. Pitch Reproduction Experiment

Even though we used durations and pitch contours directly estimated from natural speech recordings, this is no guarantee that these prosodic features are faithfully reproduced by the vocoder. To check this, we re-ran the GlottDNN pitch extractor on the WORLD-synthesised speech files, giving a new pitch contour for each stimulus. If pitch is accurately maintained by the synthesiser, these should be similar to the pitch contours extracted from NAT. Heading (a) of Tab. 2 quantifies the mean per-prompt Pearson correlation between mutually voiced frames of $\ln$ F0 contours from NAT and from other configurations. The correlations are much higher than F0 correlations seen in pure TTS, e.g., [20].

4.3. Subjective Quality Evaluation

Next, we estimated the subjective segmental quality of speech generated by the different systems through a web-based listening test. We used native Japanese listeners crowdsourced through CrowdWorks$^{\mathrm{LTD}}$ to gather assessments of all stimuli on the classic opinion score Likert scale from 1 (Bad) through 5 (Excellent). Stimuli were presented to listeners for assessment in a randomised but balanced design, such that each set of 16 stimuli that a listener heard contained one example from each condition, and that each distinct stimulus would be rated 30 times throughout the course of the evaluation. Listeners were remunerated for each complete set of 16 stimuli they rated, and were limited to rating at most six utterance sets.

After excluding a single listener who reported not being able to properly perform the task, 599 subjective ratings remained for each condition, provided by a total of 131 participating individuals. The results of the evaluation of subjective quality are reported under heading (b) of Tab. 2, in the form of mean opinion score (MOS) for each condition and two-tailed 95% confidence intervals (rounded outwards) for these means, based on a Gaussian approximation.

System	Manipulation	(a) Mean log-F0 correlation			(b) Quality MOS			(c) Mean strength of foreign accent		
		Deg. 0.0	Deg. 0.5	Deg. 1.0	Deg. 0.0	Deg. 0.5	Deg. 1.0	Deg. 0.0	Deg. 0.5	Deg. 1.0
NAT	none	1	-	-	4.43±0.031	-	-	1.60±0.046	-	-
VOC	none	0.990	-	-	3.71±0.040	-	-	1.73±0.050	-	-
MON	none	0.986	-	-	3.34±0.035	-	-	2.42±0.064	-	-
BIL	none	0.965	-	-	3.33±0.035	-	-	2.39±0.063	-	-
BIL	r	”	0.967	0.961	”	3.27±0.035	3.07±0.036	”	2.55±0.062	3.38±0.071
BIL	sh	”	0.963	0.965	”	3.30±0.035	3.27±0.035	”	2.42±0.063	2.53±0.064
BIL	z	”	0.962	0.965	”	3.30±0.035	3.31±0.035	”	2.38±0.062	2.42±0.064
BIL	j	”	0.976	0.965	”	3.33±0.035	3.31±0.036	”	2.41±0.064	2.48±0.064
BIL	ch	”	0.965	0.965	”	3.29±0.035	3.28±0.035	”	2.45±0.064	2.45±0.062
BIL	all	”	0.961	0.965	”	3.23±0.035	3.01±0.037	”	2.66±0.065	3.55±0.071

Table 2. Results of numerical evaluations of pitch contour reproduction and of subjective quality and strength of foreign accent. “Deg.” denotes the degree of interpolation within a column. Hyphens indicate cells whose values are undefined, while quotation marks indicate configurations not evaluated explicitly, but whose performance is theoretically equivalent to the closest numerical value above them.

Looking at Tab. 2, we note that the natural recordings are as high-quality, but that analysis-synthesis (VOC) shows notable degradation (0.72 points on the five-point scale). Compared to VOC, the further drop in rating introduced by the two synthesisers MON and BIL without manipulation is smaller (at most 0.38 points). Most importantly, however, the difference in rated quality (MOS) between MON and BIL is only 0.01 points. We can conclude that our approach is not far from the segmental-quality of stimuli created through vocoder-domain signal modification, and that there effectively is no quality penalty for building a system capable of multilingual speech synthesis and foreign accent generation, compared to a monolingual one. Holm-Bonferroni corrected Student’s t-tests show that the differences in quality of (NAT, VOC) and (VOC, MON) are statistically significant at the 0.05 level, whereas (MON, BIL) is not.

For manipulated conditions (BIL), the subjective quality ratings tend to lie somewhat below BIL with no manipulation. Most of the time, the difference is insignificant, but for the manipulations *r* and *all* at full interpolation the drop is 0.25 and 0.31 MOS points, respectively. This might either reflect an actual decrease in segmental signal quality, or it could be that listeners (partially or fully) are unable to perceptually separate their judgment of accent from that of segmental quality, meaning that accented speech may be more likely to be rated as having intrinsically lower quality due to the accent alone. The latter would be consistent with findings that foreign accent can affect attitudes and value judgements towards a speaker [32]. An evaluation with English speech and listeners comparing BIL monolingual system could assess the quality of the English aspect of BIL, but would still not be able to disentangle the effects of foreign accent and segmental quality during interpolation.

4.4. Foreign-Accent Evaluation

Each time a test stimulus was presented, the listener was also instructed to indicate the strength of foreign accent, scored on a seven-point Likert scale from 1 (native-like) to 7 (very strong), as in [6]. As that study only investigated short, isolated words with consonant substitutions (whereas we tested entire sentences with few, sometimes no substitutions), and they did not keep the prosody across languages, we expect our manipulations to produce less extreme effects on perceived accent than they observed. In return for using longer stimuli, we were able to collect reliable quality ratings in Sec. 4.3.

Per-condition average ratings of the strength of foreign accent, together with confidence intervals like before, are reported under heading (c) of Tab. 2. It is obvious that speech with the most pervasive phonetic substitutions r and sh (Deg. 1.0) was perceived as substantially more accented than BIL without interpolation. In particular, manipulations involving the r-phoneme (r and all) provoked highly significant differences of 0.99 or more on the seven-point scale, despite the limited fraction of phone tokens that were modified. The intermediate degree of interpolation produced smaller rating differences. The fact that some manipulations were perceived as noticeably more accented than others echoes the findings in [6].

Condition			Perceived foreign accent (% of responses)						
Syst.	Manip.	Deg.	None	CHI	KOR	AUS	IDN	USA	Unk.
NAT	none	0.0	77	3	2	1	1	5	12
VOC	none	0.0	72	3	2	1	1	8	13
MON	none	0.0	50	8	4	2	1	9	27
BIL	none	0.0	51	7	4	3	1	10	24
BIL	r	1.0	23	9	3	5	3	29	28
BIL	sh	1.0	44	10	5	3	1	10	27
BIL	z	1.0	48	7	3	2	2	11	28
BIL	j	1.0	47	9	5	2	1	11	26
BIL	ch	1.0	45	10	4	2	1	12	26
BIL	all	1.0	19	10	4	5	2	33	28

Table 3. Response distributions of perceived foreign accent. The percentages in each row may not sum to exactly 100 due to rounding.

In addition to the strength of foreign accent, we also prompted listeners to indicate the language of the foreign accent of each stimulus presented, choosing between “No accent”, “Chinese”, “Korean”, “Australian”, “Indonesian”, “American”, and “Don’t know”; this list was based on the most populous groups of non-citizens residing in Japan. A breakdown of the resulting responses for a subset of conditions is presented in Tab. 3. The table data supports the following observations and conclusions: i) The recorded Japanese speech is not foreign-accented (77% “no accent” responses). ii) The effects of vocoding (VOC) and especially RNN-based prediction (MON and BIL) interfere with human accent classification (51% or less “no accent”; 24% or greater “unknown” responses). iii) On every row, “American” was the most commonly perceived specific accent. iv) Speech with manipulated ‘r’s was obviously identifiable as an American foreign accent. (“American” was the most common response to the manipulations r and all, with “no accent” at 23% or less.)

5. CONCLUSION

We have described a new application of deep-learning-based speech synthesis to create speech stimuli with controllable foreign accent only using native (non-accented) multilingual speech recordings. Using a novel system architecture, we are able to mimic the prosody of natural speech and achieve a signal quality not far below that of vocoded audio. Empirical tests confirm the efficacy of the approach.

Our next step is to apply our method to perform research on segmental foreign accent. This includes a more varied regimen of objective and subjective tests, together with extensions to other languages and manipulations. Compelling refinements of the method include more meaningful encodings of the phonetic inputs, e.g., place of articulation, and considering multispeaker data.

6. REFERENCES

[1] Okim Kang, Don Rubin, and Lucy Pickering, "Suprasegmental measures of accentedness and judgments of language learner proficiency in oral English," *Mod. Lang. J.*, vol. 94, no. 4, pp. 554–566, 2010.

[2] Laura D. Hahn, "Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals," *TESOL Quart.*, vol. 38, no. 2, pp. 201–223, 2004.

[3] Keiichi Tajima, Robert Port, and Jonathan Dalby, "Effects of temporal correction on intelligibility of foreign-accented English," *J. Phonetics*, vol. 25, no. 1, pp. 1–24, 1997.

[4] Murray J. Munro and Tracey M. Derwing, "Modeling perceptions of the accentedness and comprehensibility of L2 speech," *Stud. Second Lang. Acq.*, vol. 23, no. 4, pp. 451–468, 2001.

[5] Tracey M. Derwing and Murray J. Munro, "Accent, intelligibility, and comprehensibility," *Stud. Second Lang. Acq.*, vol. 19, no. 1, pp. 1–16, 1997.

[6] María Luisa García Lecumberri, Roberto Barra Chicote, Rubén Pérez Ramón, Junichi Yamagishi, and Martin Cooke, "Generating segmental foreign accent," in *Proc. Interspeech*, 2014, pp. 1303–1306.

[7] Susan A. Duffy and David B. Pisoni, "Comprehension of synthetic speech produced by rule: A review and theoretical interpretation," *Lang. Speech*, vol. 35, no. 4, pp. 351–389, 1992.

[8] Stephen J. Winters and David B. Pisoni, "Perception and comprehension of synthetic speech," *Research on Spoken Language Processing Progress Report No. 26 (2003–2004)*, 2004, Speech Research Laboratory, Department of Psychology, Indiana University, Bloomington, Indiana 47405.

[9] Dennis H. Klatt, "Software for a cascade/parallel formant synthesizer," *The Journal of the Acoustical Society of America*, vol. 67, no. 3, pp. 971–995, 1980.

[10] Dennis H. Klatt, "Review of text-to-speech conversion for English," *The Journal of the Acoustical Society of America*, vol. 82, no. 3, pp. 737–793, 1987.

[11] Ming Lei, Junichi Yamagishi, Korin Richmond, Zhen-Hua Ling, Simon King, and Li-Rong Dai, "Formant-controlled HMM-based speech synthesis," in *Proc. Interspeech*, 2011, pp. 2777–2780.

[12] Yuchen Fan, Yao Qian, Feng-Long Xie, and Frank K. Soong, "TTS synthesis with bidirectional LSTM based recurrent neural networks," in *Proc. Interspeech*, 2014, pp. 1964–1968.

[13] Heiga Zen and Haşim Sak, "Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis," in *Proc. ICASSP*, 2015, pp. 4470–4474.

[14] Oliver Watts, Gustav Eje Henter, Thomas Merritt, Zhizheng Wu, and Simon King, "From HMMs to DNNs: where do the improvements come from?," in *Proc. ICASSP*, 2016, pp. 5505–5509.

[15] Heiga Zen, Andrew Senior, and Mike Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Proc. ICASSP*, 2013, pp. 7962–7966.

[16] Hieu-Thi Luong, Shinji Takaki, Gustav Eje Henter, and Junichi Yamagishi, "Adapting and controlling DNN-based speech synthesis using input codes," in *Proc. ICASSP*, 2017, pp. 4905–4909.

[17] Oliver Watts, Zhizheng Wu, and Simon King, "Sentence-level control vectors for deep neural network speech synthesis," in *Proc. Interspeech*, 2015, pp. 2217–2221.

[18] Gustav Eje Henter, Jaime Lorenzo-Trueba, Xin Wang, and Junichi Yamagishi, "Principles for learning controllable TTS from annotated and latent variation," *Proc. Interspeech*, pp. 3956–3960, 2017.

[19] Bo Li and Heiga Zen, "Multi-language multi-speaker acoustic modeling for LSTM-RNN based statistical parametric speech synthesis," in *Proc. Interspeech*, 2016, pp. 2468–2472.

[20] Xin Wang, Shinji Takaki, and Junichi Yamagishi, "An RNN-based quantized F0 model with multi-tier feedback links for text-to-speech synthesis," in *Proc. Interspeech*, 2017, pp. 1059–1063.

[21] Gustav Eje Henter, Thomas Merritt, Matt Shannon, Catherine Mayo, and Simon King, "Measuring the perceptual effects of modelling assumptions in speech synthesis using stimuli constructed from repeated natural speech," in *Proc. Interspeech*, 2014, pp. 1504–1508.

[22] Masanori Morise, Fumiya Yokomori, and Kenji Ozawa, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE T. Inf. Syst.*, vol. 99, no. 7, pp. 1877–1884, 2016.

[23] Lauri Juvela, Xin Wang, Shinji Takaki, SangJin Kim, Manu Airaksinen, and Junichi Yamagishi, "The NII speech synthesis entry for Blizzard Challenge 2016," in *Proc. Blizzard Challenge Workshop*, 2016.

[24] Heiga Zen, Takashi Nose, Junichi Yamagishi, Shinji Sako, Takashi Masuko, Alan W. Black, and Keiichi Tokuda, "The HMM-based speech synthesis system (HTS) version 2.0," in *Proc. SSW*, 2007, pp. 294–299.

[25] HTS Working Group, "The English TTS system 'Flite+hts_engine'," 2014, http://hts-engine.sourceforge.net/.

[26] Korin Richmond, Robert A. J. Clark, and Susan Fitt, "Robust LTS rules with the Combilex speech technology lexicon," in *Proc. Interspeech*, 2009, pp. 1295–1298.

[27] Keiichiro Oura, Shinji Sako, and Keiichi Tokuda, "Japanese text-to-speech synthesis system: Open JTalk," in *Proc. ASJ Spring*, 2010, pp. 343–344.

[28] Xin Wang, Shinji Takaki, and Junichi Yamagishi, "An autoregressive recurrent mixture density network for parametric speech synthesis," in *Proc. ICASSP*, 2017, pp. 4895–4899.

[29] Felix Weninger, Johannes Bergmann, and Björn W. Schuller, "Introducing CURRENNT: The Munich open-source CUDA recurrent neural network toolkit," *J. Mach. Learn. Res.*, vol. 16, no. 3, pp. 547–551, 2015.

[30] John Duchi, Elad Hazan, and Yoram Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *J. Mach. Learn. Res.*, vol. 12, pp. 2121–2159, 2011.

[31] Keiichi Tokuda, Takayoshi Yoshimura, Takashi Masuko, Takao Kobayashi, and Tadashi Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. ICASSP*, 2000, pp. 1315–1318.

[32] Andrew J. Pantos and Andrew W. Perkins, "Measuring implicit and explicit attitudes toward foreign accented speech," *J. Lang. Soc. Psychol.*, vol. 32, no. 1, pp. 3–20, 2013.