Generating Segment-Level Foreign-Accented Synthetic Speech with Natural Speech Prosody

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

Abstract: We present a new application of deep-learning-based TTS, namely multilingual speech synthesis for generating controllable foreign accent. We train an acoustic model on non-accented multilingual speech recordings from the same speaker and interpolate quinphone linguistic features between languages to generate microscopic foreign accent. By copying pitch and durations 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. Experiments on synthetic American-English-accented Japanese confirm the success of the approach.

Keywords: Controllable speech synthesis, foreign accent, multilingual speech synthesis, speech perception

1. Introduction

This paper reports on the development of a deep multilingual speech synthesis system that can be used as a tool for research on speech perception, in particular foreign-accent perception. This is a new application of neural-network-based speech synthesis. The approach we describe is able to generate foreign-accented speech despite only being trained on native-speaker (though multilingual) speech data.

An appealing aspect of using speech synthesis to generate stimuli for speech-perception research is that synthesis allows for a great deal of output control. In our approach the accent can be controlled continuously on the level of individual segments. This capability is central to exploring the microscopic phonetic cues that contribute to foreign-accent perception. While much prior research has considered supra-segmental properties of accent (cf. [1], [2], [3], [4]), listeners instead point to segmental errors as the most important cues for establishing foreign-accentedness [5].

Segmental speech manipulations cannot easily be elicited from human talkers. Our automated approach scales better to than manually splicing recordings together, and also avoids the risk of splicing artefacts. While multilingual statistical speech synthesis has been used to generate foreign-accented speech stimuli before [6], our work extends this to consider deep learning. Research has showed that deep and recurrent neural networks provide significantly greater synthesis quality [7], [8] than earlier paradigms and also allow a great degree of control over the output, successfully manipulating aspects like speaker identity [9], expression [10], and emotion [11].

At the same time – as a second innovation – we propose to replicate the prosody of natural speech in our synthetic stimuli by borrowing durations and pitch contours from natural speech recordings. This avoids the issue of the sometimes inappropriate prosody generated by conventional text-to-speech (TTS) possibly interfering with accent perception.

The methodology we describe can be applied to any language combination and allows for a wide variety of phonetic manipulations. Our experiments confirm that we are able to generate characteristically accented stimuli with similar speech quality to a monolingual baseline system, and with pitch contours that closely match those of natural speech.

2. Method

Our goal is to generate foreign-accented stimuli from unaccented, multilingual recordings. For this reason, we are dependent on multilingual recordings of a speaker native in more than one language. Unlike a regular text-to-speech system, we propose a system that also takes natural speech recordings as input at synthesis time; these recordings provide reference prosody (pitch and durations) for the generated speech to use. For this reason, we require the database to contain natural recordings of the prompts to be used as stimuli, in addition to regular TTS training data. Unusually, we thus have a text-and-speech-to-speech system whose output speech parameters mix human and machine parts, wherefore we call it "cyborg speech".

2.1 Data

For the experiments in this paper, we had access to a database of a male voice talent native in both US English and Japanese. 2000 training utterances (all at 48 kHz, 16 bit) and 20 designated test utterances were used for each language. WORLD [12] with 200 frames per second was used for acoustic analysis and syn-

[1] National Institute of Informatics, Chiyoda, Tokyo 101–8430, Japan
[2] Waseda University, Shinjuku, Tokyo 169–8050, Japan
[3] The University of Edinburgh, Edinburgh EH8 9AB, UK
[a)] gustav@nii.ac.jp
[b)] jyamagis@nii.ac.jp

thesis, except for F0 analysis, which used the GlottDNN pitch extractor [13] to reduce the amount of voicing errors. Spectra and aperiodicities were then converted to 60-dimensional MGCs and 25-dimensional BAPs, along with dynamic features (deltas and delta deltas), following established norms in statistical parametric speech synthesis.

2.2 Input Processing

Text analysis used Flite [14] with the Combilex [15] General American (GAM) dictionary for English text, and Open JTalk [16] for Japanese. HTS [17] (one system trained on each language) was used for forced alignment of training data and for determining durations in the prosodic reference recordings.

Since durations are given as input, no duration model was needed, and only the frame-wise acoustic parameters MGCs and BAPs had to be predicted, on the basis of linguistic and other input features. Among our extracted linguistic features, only the quinphone context features were used in by our neural networks. The remaining linguistic features were discarded, since they differ in nature between languages and are mainly intended for (here unnecessary) prosody prediction. Counting the phones of each language as distinct, we obtained a total of 98 phones in the combined phone set. The phones were represented with a one-hot encoding in neural network quinphone inputs. By interpolating between phone identities, pronunciation can be adjusted for each segment, or even each frame. In doing so across languages, we can replicate segmental foreign accent.

In addition to the binary quinphone features, three duration-based features (like in [8]), a voiced-unvoiced flag and static and dynamic log-pitch values (interpolated in unvoiced regions), plus a binary language flag, were provided as inputs for each frame. All except the language flag were mean and variance normalised.

2.3 Network Design and Training

Our proposed method uses deep, recurrent neural networks to predict static and dynamic MGCs and BAPs (total 255 dimensions) from framewise inputs (228 dimensions in Japanese, 278 in English, and 498 for a bilingual system). Our network design followed the DBLSTMs in [18]. The systems, one bilingual and one Japanese monolingual, were initialised randomly and then trained to minimise mean-squared error using CURRENNT [19]. The optimisation used 160 epochs of raw SGD followed by AdaGrad [20]. Approximately 5% of the training data were held out as a validation set for early stopping, which terminated the AdaGrad stage in 30 epochs or less in all runs.

3. Experiments

For the experiments, we explored the generation of Japanese speech, specifically the 20 test sentences, with a (segmental) US English foreign accent generated by the bilingual system (labelled BIL). The properties of such speech were compared against those of the Japanese monolingual system (MON) and of natural (NAT) and analysis-synthesised (VOC) speech.

A number of cross-language phonetic substitutions were performed with BIL on the test-prompt input quinphones, changing certain phones to US English approximations in order to replicate common pronunciation errors among US English natives speaking Japanese as a second language. Details are provided in **Table 1**. The language flag was also altered based on the phone set of the centre phone. We additionally generated BIL stimuli with no substitutions, as well as stimuli making all substitutions in the table simultaneously. MLPG and postfiltering were used by all systems when synthesising speech stimuli from predicted acoustic parameters.

Table 1 Foreign-accent phone substitutions considered.

Subst.	Japanese		US English		No. affected
ID	IPA	Open JTalk	IPA	Combilex	prompts
r	ɾ	r	ɹ	r	19 of 20
sh	ɕ	sh	ʃ	S	13 of 20
z	dz	z	z	z	7 of 20
j	dʑ	j	dʒ	dZ	8 of 20
ch	tɕ	ch	tʃ	tS	11 of 20

Table 2 Results from numerical evaluations in this study.

System	Subst.	logF0 corr.	Quality MOS	Accent strength
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	0.961	3.07±0.036	3.38±0.071
BIL	sh	0.965	3.27±0.035	2.53±0.064
BIL	z	0.965	3.31±0.035	2.42±0.064
BIL	j	0.965	3.31±0.036	2.48±0.064
BIL	ch	0.965	3.28±0.035	2.45±0.062
BIL	all	0.965	3.01±0.037	3.55±0.071

3.1 Objective Evaluation

To objectively verify that the prosody of synthesised speech matched that of the reference recordings, we computed the correlation coefficients between the log-pitch contours of NAT and the different systems with and without phone substitutions on the 20 test prompts. The results are tabulated in **Table 2**. All correlations are close to 1.0, indicating that pitch contours in natural and synthetic stimuli are closely matched.

3.2 Subjective Evaluations

To explore the perceptual characteristics of our synthesised speech, we performed a web-based listening test with paid Japanese native listeners sourced through CrowdWorks$^{\text{LTD}}$. The listeners were presented with a stimulus and requested to score its quality (1, Bad, through 5, Excellent) and the strength of foreign accent (1, native-like, through 7, very strong) on traditional Likert scales. They also classified the language of the foreign accent, choosing among "No accent", "Chinese", "Korean", "Australian", "Indonesian", "American", and "Don't know", based on the most populous groups of foreign citizens living in Japan.

Each listener scored between one and six sets of stimuli, each set containing one example of each system and substitution combination. In total each combination was rated 599 times by 131 different listeners. From the responses we computed the mean speech quality and mean strength of foreign accent (with 95% confidence intervals from a Gaussian approximation). These are reported in the final columns of Table 2. The distribution of the responses to the foreign-accent classification task are meanwhile tabulated in **Table 3**. (Three response categories that never ex-

Table 3 Perceived foreign accent distribution, in percent.

System	Subst.	None	CHI	USA	Other	Unknown
NAT	none	77	3	5	4	12
VOC	none	72	3	8	4	13
MON	none	50	8	9	7	27
BIL	none	51	7	10	8	24
BIL	r	23	9	29	11	28
BIL	sh	44	10	10	9	27
BIL	z	48	7	11	7	28
BIL	j	47	9	11	8	26
BIL	ch	45	10	12	7	26
BIL	all	19	10	33	11	28

ceeded 5% individually have for simplicity been combined to form the "Other" category.)

4. Discussion

The subjective tests support a number of interesting observations. For the quality ratings, the mean opinion scores in Table 2 clearly degraded with vocoding and additionally (but to a lesser extent) when predicted speech parameters were used; both these differences are statistically significant at the level 0.05, according to Student's t-tests corrected for multiple comparisons. The quality difference between MON and VOC, on the other hand, is negligible. We conclude that building a bilingual instead of a monolingual synthesiser did not affect quality substantially. Among substituted stimuli, the biggest effects in quality are seen with the substitutions r and all, dropping MOS by 0.25 and 0.31 points, respectively. This could be due to somewhat reduced signal quality, or it could be an effect of listeners rating foreign-accented speech intrinsically less favourably than native-like speech.

In terms of accent perception, as reported in the final column of Table 2 and the entirety of Table 3, we can note that natural and vocoded speech are both perceived as unaccented by a majority of listeners. However, speech parameters predicted by MON and by BIL without substitutions were rated as noticeably more accented than VOC, although those who perceived an accent mostly rated it as "Unknown".

Once phonetic substitutions were introduced, the average rated strength of foreign accent increased, and the fraction of listeners who responded "No accent" decreased. The effect was particularly stark for the stimuli including the r-substitution (r and all), where foreign-accent strength in Table 2 increased by at least 0.99 points and the "USA" column dominates all others in Table 3, while "No accent" responses decreased to 23% or less. This supports a conclusion that the system successfully generated Japanese-language speech stimuli with a characteristic US English foreign accent.

5. Conclusion

In summary, we have shown how LSTM-based speech synthesis can be used to create speech stimuli with controllable, segment-level foreign accent while maintaining natural prosody. We were able to do this only using non-accented multilingual speech recordings. Experiments confirm that a clear and distinctive accent was achieved. The most compelling future work is to apply the method for new research into foreign-accent perception.

References

[1] Kang, O., Rubin, D. and Pickering, L.: 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] Hahn, L. D.: Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals, *TESOL Quart.*, Vol. 38, No. 2, pp. 201–223 (2004).

[3] Tajima, K., Port, R. and Dalby, J.: Effects of temporal correction on intelligibility of foreign-accented English, *J. Phonetics*, Vol. 25, No. 1, pp. 1–24 (1997).

[4] Munro, M. J. and Derwing, T. M.: Modeling perceptions of the accentedness and comprehensibility of L2 speech, *Stud. Second Lang. Acq.*, Vol. 23, No. 4, pp. 451–468 (2001).

[5] Derwing, T. M. and Munro, M. J.: Accent, intelligibility, and comprehensibility, *Stud. Second Lang. Acq.*, Vol. 19, No. 1, pp. 1–16 (1997).

[6] García Lecumberri, M. L., Barra Chicote, R., Pérez Ramón, R., Yamagishi, J. and Cooke, M.: Generating segmental foreign accent, *Proc. Interspeech*, pp. 1303–1306 (2014).

[7] Watts, O., Henter, G. E., Merritt, T., Wu, Z. and King, S.: From HMMs to DNNs: where do the improvements come from?, *Proc. ICASSP*, pp. 5505–5509 (2016).

[8] Zen, H., Senior, A. and Schuster, M.: Statistical parametric speech synthesis using deep neural networks, *Proc. ICASSP*, pp. 7962–7966 (2013).

[9] Luong, H.-T., Takaki, S., Henter, G. E. and Yamagishi, J.: Adapting and Controlling DNN-based Speech Synthesis Using Input Codes, *Proc. ICASSP*, pp. 4905–4909 (2017).

[10] Watts, O., Wu, Z. and King, S.: Sentence-level control vectors for deep neural network speech synthesis, *Proc. Interspeech*, pp. 2217–2221 (2015).

[11] Henter, G. E., Lorenzo-Trueba, J., Wang, X. and Yamagishi, J.: Principles for learning controllable TTS from annotated and latent variation, *Proc. Interspeech*, pp. 3956–3960 (2017).

[12] Morise, M., Yokomori, F. and Ozawa, K.: 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).

[13] Juvela, L., Wang, X., Takaki, S., Kim, S., Airaksinen, M. and Yamagishi, J.: The NII speech synthesis entry for Blizzard Challenge 2016, *Proc. Blizzard Challenge Workshop* (2016).

[14] HTS Working Group: The English TTS System 'Flite+hts_engine' (2014).

[15] Richmond, K., Clark, R. A. J. and Fitt, S.: Robust LTS rules with the Combilex speech technology lexicon, *Proc. Interspeech*, pp. 1295–1298 (2009).

[16] Oura, K., Sako, S. and Tokuda, K.: Japanese Text-to-Speech Synthesis System: Open JTalk, *Proc. ASJ Spring*, pp. 343–344 (2010).

[17] Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A. W. and Tokuda, K.: The HMM-based speech synthesis system (HTS) version 2.0, *Proc. SSW*, pp. 294–299 (2007).

[18] Wang, X., Takaki, S. and Yamagishi, J.: An autoregressive recurrent mixture density network for parametric speech synthesis, *Proc. ICASSP*, pp. 4895–4899 (2017).

[19] Weninger, F., Bergmann, J. and Schuller, B. W.: Introducing CURRENNT: The Munich open-source CUDA recurrent neural network toolkit, *J. Mach. Learn. Res.*, Vol. 16, No. 3, pp. 547–551 (2015).

[20] Duchi, J., Hazan, E. and Singer, Y.: Adaptive subgradient methods for online learning and stochastic optimization, *J. Mach. Learn. Res.*, Vol. 12, pp. 2121–2159 (2011).