Cyborg Speech: Deep Multilingual Speech Synthesis For Generating Segmental Foreign Accent With Natural Prosody

Citation for published version:

Digital Object Identifier (DOI):
10.1109/ICASSP.2018.8462470

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
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 quintaphone 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...
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 but rather like acoustic modelling (predicting 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 representing different phones and languages, enabling speech output with intermediate phonetics or articulation, in mixed languages. 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.

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 Δ² 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.

3.2. 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 representing 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 Δ² 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
4. EXPERIMENTS

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 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.

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 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.

<table>
<thead>
<tr>
<th></th>
<th>Japanese</th>
<th>English</th>
<th>Substitutions made</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>IPA</td>
<td>Open JTalk</td>
<td>IPA</td>
</tr>
<tr>
<td>t</td>
<td>r</td>
<td>z</td>
<td>r</td>
</tr>
<tr>
<td>c</td>
<td>sh</td>
<td>j</td>
<td>z</td>
</tr>
<tr>
<td>dз</td>
<td>z</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>dz</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>te</td>
<td>ch</td>
<td>t</td>
<td>11</td>
</tr>
</tbody>
</table>

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.

Optimisation proceeded using minibatch stochastic gradient descent (SGD) in two stages: first, raw SGD with learning rate $2.5 \times 10^{-6}$ 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 (frame MSE) was obtained from the extracted F0, MGC, and BAP features.

4.1. Stimulus Generation

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 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.
### 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 equivalent to the closest numerical value above them.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Deg.</th>
<th>Syst.</th>
<th>Manip</th>
<th>CHI</th>
<th>M'IN</th>
<th>AUS</th>
<th>IND</th>
<th>USA</th>
<th>Unk.</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>0.0</td>
<td>NAT</td>
<td>none</td>
<td>77</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>None</td>
<td>0.0</td>
<td>VOC</td>
<td>none</td>
<td>72</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>None</td>
<td>0.0</td>
<td>MON</td>
<td>none</td>
<td>50</td>
<td>8</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>None</td>
<td>0.0</td>
<td>BIL</td>
<td>none</td>
<td>51</td>
<td>7</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>1.0</td>
<td></td>
<td>BIL</td>
<td>sh</td>
<td>44</td>
<td>10</td>
<td>5</td>
<td>3</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>1.0</td>
<td></td>
<td>BIL</td>
<td>z</td>
<td>45</td>
<td>7</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>11</td>
</tr>
<tr>
<td>1.0</td>
<td></td>
<td>BIL</td>
<td>ch</td>
<td>47</td>
<td>9</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>11</td>
</tr>
<tr>
<td>1.0</td>
<td></td>
<td>all</td>
<td></td>
<td>45</td>
<td>10</td>
<td>5</td>
<td>3</td>
<td>1</td>
<td>12</td>
</tr>
<tr>
<td>1.0</td>
<td></td>
<td>all</td>
<td></td>
<td>45</td>
<td>10</td>
<td>5</td>
<td>3</td>
<td>1</td>
<td>12</td>
</tr>
</tbody>
</table>

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 regime 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


