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]

gustav@nii.ac.jp, jyamagis@nii.ac.jp

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

2018-04-18

Synopsis

- We generate foreign-accented synthetic speech audio
    - …with native prosody
    - …having finely controllable accent
    - …as a new application of deep-learning-based speech synthesis
    - …using multilingual techniques
    - …from non-accented speech data alone

Overview

Overview

Studying foreign accent

What makes speech sound foreign-accented?

- A question of speech perception research
  - Empirical method: Measure how listeners respond to speech stimuli with carefully controlled differences
- Useful knowledge for improving foreign-language instruction

Cues to foreign accent

What makes speech sound foreign-accented?

- Supra-segmental properties
  - Intonation and pauses (Kang et al., 2010)
  - Nuclear stress (Hahn, 2004)
  - Duration (Tajima et al., 1997)
  - Speech rate (Munro and Derwing, 2001)
  - And more...
- Segmental properties
  - Pronunciation errors
    - Listeners often consider this the most important aspect! (Derwing and Munro, 1997)
    - Worthwhile to correct even if not

Studying segmental foreign accent

- Need speech stimuli isolating and interpolating segmental effects
    - Only specific segments should be affected
    - Without supra-segmental effects

Studying segmental foreign accent

- Need speech stimuli isolating and interpolating segmental effects
  - Only specific segments should be affected
  - Without supra-segmental effects
- Method 1: Record deliberate mispronunciations
  - Difficult/impossible to elicit

Studying segmental foreign accent

- Need speech stimuli isolating and interpolating segmental effects
  - Only specific segments should be affected
  - Without supra-segmental effects
- Method 1: Record deliberate mispronunciations
  - Difficult/impossible to elicit
- Method 2: Cross-language splicing
  - Labour-intensive manual work
  - Artefacts at joins

Studying segmental foreign accent

- Need speech stimuli isolating and interpolating segmental effects
  - Only specific segments should be affected
  - Without supra-segmental effects
- Method 1: Record deliberate mispronunciations
  - Difficult/impossible to elicit
- Method 2: Cross-language splicing
  - Labour-intensive manual work
  - Artefacts at joins
- Method 3: Synthesise stimuli
  - Data-driven, automated approach
  - No joins
  - New tool; unusual application of speech synthesis

Our approach

- Methods for synthesising foreign-accented stimuli
  - Multilingual HMM-based TTS (García Lecumberri et al., 2014)
  - Multilingual deep learning (this presentation!)
    - We improve on (García Lecumberri et al., 2014) in two ways:

Our approach

- Methods for synthesising foreign-accented stimuli
  - Multilingual HMM-based TTS (García Lecumberri et al., 2014)
  - Multilingual deep learning (this presentation!)
    - We improve on (García Lecumberri et al., 2014) in two ways:
- Improvement 1: Deep learning
  - Improved signal quality (Watts et al., 2016), meaning it better replicates the perceptual cues in natural speech
  - Enables easy control of the output synthesis (Watts et al., 2015; Luong et al., 2017)

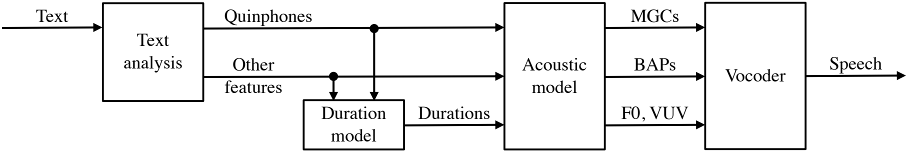
Our approach

- Methods for synthesising foreign-accented stimuli
  - Multilingual HMM-based TTS (García Lecumberri et al., 2014)
  - Multilingual deep learning (this presentation!)
    - We improve on (García Lecumberri et al., 2014) in two ways:
- Improvement 1: Deep learning
  - Improved signal quality (Watts et al., 2016), meaning it better replicates the perceptual cues in natural speech
  - Enables easy control of the output synthesis (Watts et al., 2015; Luong et al., 2017)
- Improvement 2: Use reference prosody (pitch and duration)
  - Can be taken from natural speech, or predicted by a separate system
  - Allows us to impose native-like suprasegmental properties

Overview

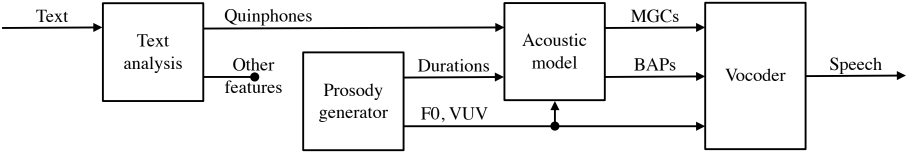
Building the synthesiser

Traditional text-to-speech:



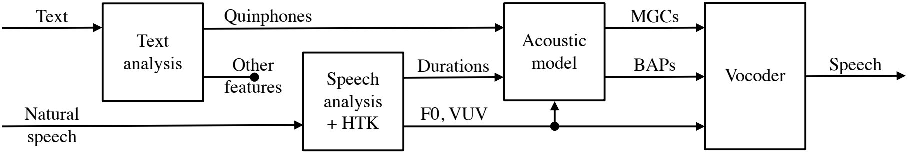
Building the synthesiser

Speech synthesis with arbitrary prosody:



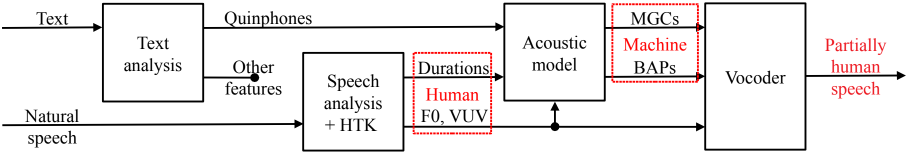
Building the synthesiser

Speech synthesis with natural prosody:

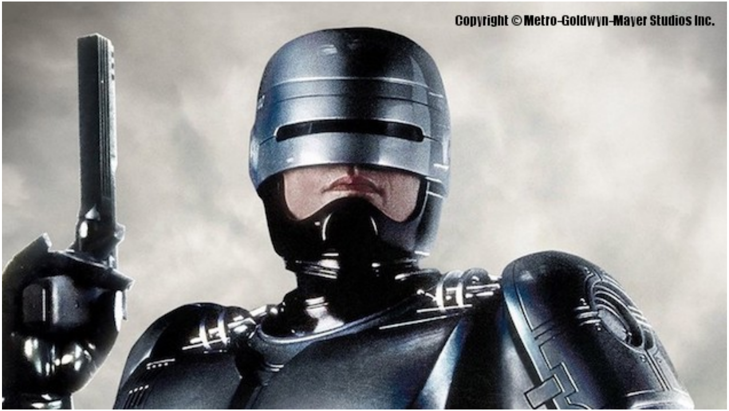


Building the synthesiser

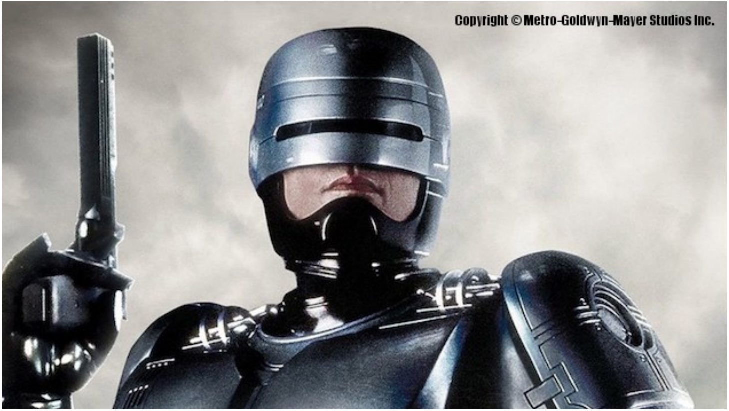
Speech synthesis with natural prosody:



"Cyborg speech"



"Cyborg speech"



- Cyborg: A being with both organic and biomechatronic body parts
  - Our acoustic parameters are a combination of man and machine

Making it foreign

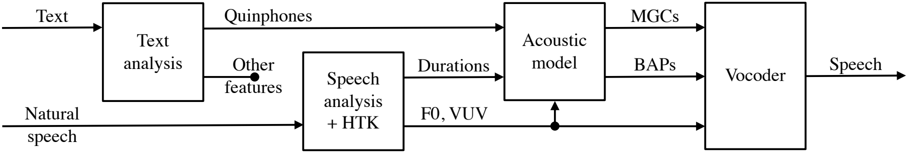
- Segmental foreign accent through multilingual speech synthesis:
  - Teach a single model to synthesise several languages natively
  - During synthesis, interpolate specific phones in the spoken language towards phones in the accent language
  - Maintain the same voice across languages
    - In this case by using data from a multilingually native speaker

Making it foreign

- Segmental foreign accent through multilingual speech synthesis:
  - Teach a single model to synthesise several languages natively
  - During synthesis, interpolate specific phones in the spoken language towards phones in the accent language
  - Maintain the same voice across languages
    - In this case by using data from a multilingually native speaker
- Running example: American English and Japanese
  - Combilex GAM (Richmond et al., 2009): 54 English phones
  - Open JTalk (Oura et al., 2010): 44 Japanese phones
  - Combined, bilingual phoneset: $54 + 44 = 98$ phones

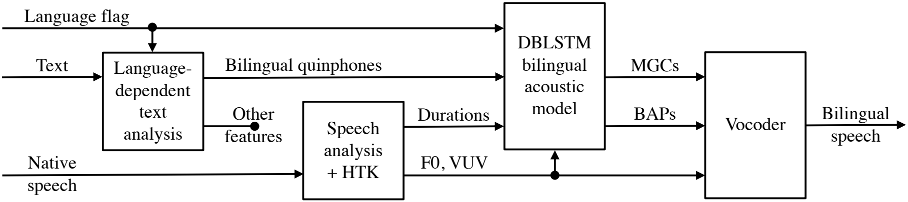
Synthesising foreign accent

Cyborg speech:



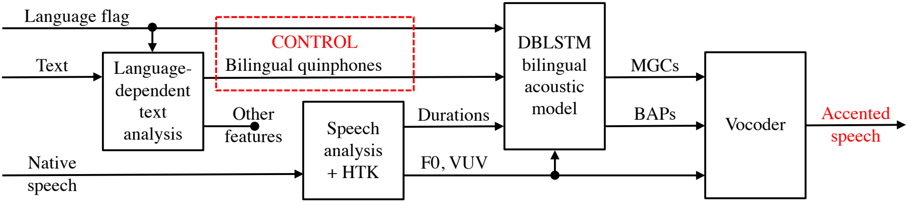
Synthesising foreign accent

Bilingual cyborg speech synthesis:



Synthesising foreign accent

Foreign-accented speech synthesis:



Overview

Data and processing

- Male voice talent native in both US English and Japanese
  - 2000 utterances per language
    - US English example
    - Japanese example
  - 20 pre-recorded test utterances in each language
    - Source of reference pitch and durations
  - 48 kHz at 16 bits

Data and processing

- Male voice talent native in both US English and Japanese
  - 2000 utterances per language
    - US English example
    - Japanese example
  - 20 pre-recorded test utterances in each language
    - Source of reference pitch and durations
  - 48 kHz at 16 bits
- WORLD vocoder (Morise et al., 2016)
- Forced alignment using HTS (Zen et al., 2007)
  - Separate systems for each language

Network and training

- Acoustic model network topology followed (Wang et al., 2017):
  - 2 logistic sigmoid feed-forward layers
  - 2 bidirectional LSTM layers

Network and training

- Acoustic model network topology followed (Wang et al., 2017):
  - 2 logistic sigmoid feed-forward layers
  - 2 bidirectional LSTM layers
- Minibatch training to minimise frame mean-square error
  - Plain SGD followed by AdaGrad (Duchi et al., 2011) with early stopping
  - Using the C++ framework CURRENNT (Weninger et al., 2015)

Systems

- Natural speech (NAT)
- Analysis-synthesis (VOC)
- Monolingual Japanese cyborg system (MON)
- Bilingual cyborg system (BIL)
    - Only this system can interpolate phones across languages

Cross-language substitutions

Consonant substitutions inspired by common mispronunciations among native American English speakers (L1) learning Japanese (L2):

Japanese		English		Substitutions	
IPA	Open JTalk	IPA	Combilex GAM	Max	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

(Manipulations in the other direction allow BIL to generate Japanese-accented English instead)

Example stimuli

System	NAT	VOC	MON	BIL
ID 12	▶	▶	▶	▶
ID 13	▶	▶	▶	▶

System	BIL	BIL	BIL	BIL	BIL	BIL
Substitution	r	sh	z	j	ch	all
ID 12	▶	▶	▶	▶	▶	▶
ID 13	▶	▶	▶	▶	▶	▶

(How perceptible the differences are depends on your native language; they might be more obvious to non-Japanese listeners)

Overview

Listening test

- Crowdsourced, web-based listening test
  - 131 native Japanese listeners
  - Rating balanced sets of utterances
  - 599 ratings per condition (system and manipulation)

Listening test

- Crowdsourced, web-based listening test
  - 131 native Japanese listeners
  - Rating balanced sets of utterances
  - 599 ratings per condition (system and manipulation)
- Responses collected per stimulus presentation:
  - Speech quality: 1 (poor) to 5 (excellent)
  - Strength of foreign accent: 1 (native-like) to 7 (very strong)
  - Foreign accent classification: 5 nationalities (CHI, KOR, AUS, IDN, and USA), "none", and "unknown"

Strength of perceived foreign accent

System	Substitution	Accent strength	Change
NAT	none	1.60 ± 0.046	-
VOC	none	1.73 ± 0.050	0.13 vs. NAT
MON	none	2.42 ± 0.064	0.69 vs. VOC
BIL	none	2.39 ± 0.063	-0.03 vs. MON
BIL	r	3.38 ± 0.071	0.99 vs. none
BIL	sh	2.53 ± 0.064	0.14 vs. none
BIL	z	2.42 ± 0.064	0.03 vs. none
BIL	j	2.48 ± 0.064	0.09 vs. none
BIL	ch	2.45 ± 0.062	0.06 vs. none
BIL	all	3.55 ± 0.071	1.16 vs. none

(Ranges are 95% mean accent strength confidence intervals)

Strength of perceived foreign accent

System	Substitution	Accent strength	Change
NAT	none	1.60 ± 0.046	-
VOC	none	1.73 ± 0.050	0.13 vs. NAT
MON	none	2.42 ± 0.064	0.69 vs. VOC
BIL	none	2.39 ± 0.063	-0.03 vs. MON
BIL	r	3.38 ± 0.071	0.99 vs. none
BIL	sh	2.53 ± 0.064	0.14 vs. none
BIL	z	2.42 ± 0.064	0.03 vs. none
BIL	j	2.48 ± 0.064	0.09 vs. none
BIL	ch	2.45 ± 0.062	0.06 vs. none
BIL	all	3.55 ± 0.071	1.16 vs. none

(Ranges are 95% mean accent strength confidence intervals)

Strength of perceived foreign accent

System	Substitution	Accent strength	Change
NAT	none	1.60±0.046	-
VOC	none	1.73±0.050	0.13 vs. NAT
MON	none	2.42±0.064	0.69 vs. VOC
BIL	none	2.39±0.063	−0.03 vs. MON
BIL	r	3.38±0.071	0.99 vs. none
BIL	sh	2.53±0.064	0.14 vs. none
BIL	z	2.42±0.064	0.03 vs. none
BIL	j	2.48±0.064	0.09 vs. none
BIL	ch	2.45±0.062	0.06 vs. none
BIL	all	3.55±0.071	1.16 vs. none

(Ranges are 95% mean accent strength confidence intervals)

Strength of perceived foreign accent

System	Substitution	Accent strength	Change
NAT	none	1.60±0.046	-
VOC	none	1.73±0.050	0.13 vs. NAT
MON	none	2.42±0.064	0.69 vs. VOC
BIL	none	2.39±0.063	−0.03 vs. MON
BIL	r	3.38±0.071	0.99 vs. none
BIL	sh	2.53±0.064	0.14 vs. none
BIL	z	2.42±0.064	0.03 vs. none
BIL	j	2.48±0.064	0.09 vs. none
BIL	ch	2.45±0.062	0.06 vs. none
BIL	all	3.55±0.071	1.16 vs. none

(Ranges are 95% mean accent strength confidence intervals)

Distribution of perceived accent

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

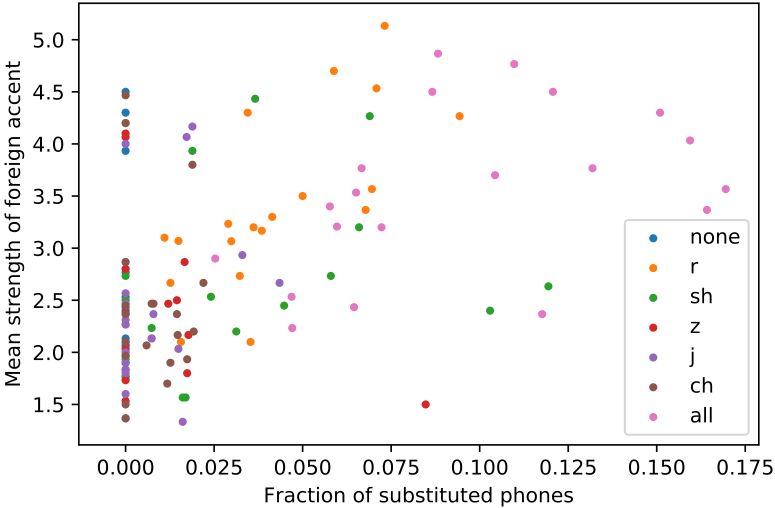
Distribution of perceived accent

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

Distribution of perceived accent

Condition		Accent language (%)				
System	**Substitution**	**None**	**USA**	**CHI**	**Other**	**Unk.**
NAT	none	77	5	3	4	12
VOC	none	72	8	3	4	13
MON	none	50	9	8	7	27
BIL	none	51	10	7	8	24
BIL	r	23	29	9	11	28
BIL	sh	44	10	10	9	27
BIL	z	48	11	7	7	28
BIL	j	47	11	9	8	26
BIL	ch	45	12	10	7	26
BIL	all	19	33	10	11	28

Scatterplot of BIL stimuli



(The overall Pearson correlation coefficient is 0.43)

Overview

Empirical conclusions

- Substituting the phone “r” (in r and all) produced distinctly American-accented Japanese speech
- Other substitutions were less noticeable
  - But also less numerous in the test sentences
- Modelling artefacts were perceived as an “unknown” accent
- Bilingual training did not degrade perception vs. monolingual

Summary of achievements

- We have generated synthetic speech audio with a foreign accent
  - ... that is distinct and recognisable
  - ... having fine accent control
  - ... while maintaining native prosody
  - ... as a new application of deep-learning-based speech synthesis
  - ... using multilingual techniques
  - ... from non-accented speech data alone

Possible extensions

- Use a neural vocoder to improve signal quality
  - This can mitigate both vocoding and modelling artefacts, as demonstrated in Tacotron 2 (Shen et al., 2018)
- Consider other phone encodings beyond one-hot
  - IPA place/manner of articulation? Formant frequencies?
  - Offer more intuitive and general pronunciation control
- Apply the work in foreign-accent research

The end

The end

Thank you for listening!

The end

Any questions?

Acknowledgement

This research has been supported by the Diacex project, in collaboration with Prof. María Luisa García Lecumberri, Prof. Martin Cooke, and Mr. Rubén Pérez Ramón.

References I

Derwing, T. M. and Munro, M. J. (1997).
Accent, intelligibility, and comprehensibility.
Stud. Second Lang. Acq., 19(1):1–16.

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

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

Hahn, L. D. (2004).
Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals.
TESOL Quart., 38(2):201–223.

Kang, O., Rubin, D., and Pickering, L. (2010).
Suprasegmental measures of accentedness and judgments of language learner proficiency in oral English.
Mod. Lang. J., 94(4):554–566.

References II

Luong, H.-T., Takaki, S., Henter, G. E., and Yamagishi, J. (2017).
Adapting and controlling DNN-based speech synthesis using input codes.
In *Proc. ICASSP*, pages 4905–4909.

Morise, M., Yokomori, F., and Ozawa, K. (2016).
WORLD: A vocoder-based high-quality speech synthesis system for real-time applications.
IEICE T. Inf. Syst., 99(7):1877–1884.

Munro, M. J. and Derwing, T. M. (2001).
Modeling perceptions of the accentedness and comprehensibility of L2 speech.
Stud. Second Lang. Acq., 23(4):451–468.

Oura, K., Sako, S., and Tokuda, K. (2010).
Japanese text-to-speech synthesis system: Open JTalk.
In *Proc. ASJ Spring*, pages 343–344.

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

References III

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerry-Ryan, R., Saurous, R. A., Agiomyrgiannakis, Y., and Wu, Y. (2018).
Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions.
In *Proc. ICASSP*, pages 4799–4783.

Tajima, K., Port, R., and Dalby, J. (1997).
Effects of temporal correction on intelligibility of foreign-accented English.
J. Phonetics, 25(1):1–24.

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

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

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

References IV

Weninger, F., Bergmann, J., and Schuller, B. W. (2015).
Introducing CURRENNT: The Munich open-source CUDA recurrent neural network toolkit.
J. Mach. Learn. Res., 16(3):547–551.

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

Subjective quality

System	Substitution	Quality MOS	Change
NAT	none	4.43±0.031	-
VOC	none	3.71±0.040	−0.72 vs. NAT
MON	none	3.34±0.035	−0.37 vs. VOC
BIL	none	3.33±0.035	−0.01 vs. MON
BIL	r	3.07±0.036	−0.26 vs. none
BIL	sh	3.27±0.035	−0.06 vs. none
BIL	z	3.31±0.035	−0.02 vs. none
BIL	j	3.31±0.036	−0.02 vs. none
BIL	ch	3.28±0.035	−0.05 vs. none
BIL	all	3.01±0.037	−0.32 vs. none

(Ranges are 95% MOS confidence intervals)

Prosodic faithfulness

Correlation between NAT and test stimuli pitch (log F0):

System	Substitution?	Pearson correlation
NAT	no	1
VOC	no	0.990
MON	no	0.986
BIL	no	0.965
BIL	yes	0.961–0.965

- These numbers are much higher than for standard TTS
    - Despite pitch extractor/vocoder mismatch (GlottDNN/WORLD)
    - The residual is dominated by pitch doublings in individual frames