Cyborgs and other controllable synthesisers: an update on past and planned research

Gustav Eje Henter

ghe@kth.se

Dept. of Speech, Music and Hearing, KTH Royal Institute of Technology, Sweden

2018-11-13

Talk contents

The three parts of today's presentation:

I. Review of some recent publications
II. A more in-depth investigation
III. Planned future work

Unifying threads

How the three parts of today's presentation fit together:

I. Review of some recent publications
   - Technical interest: controllable speech synthesis

II. A more in-depth investigation
   - Technical interest: controllable speech synthesis
   - Application interest: speech perception

III. Planned future work
   - Application interest: speech perception
     (Controllable speech synthesis will be incorporated later)

Unifying threads

How the three parts of today's presentation fit together:

I. Review of some recent publications
   - Technical interest: controllable speech synthesis

II. A more in-depth investigation
   - Technical interest: controllable speech synthesis
   - Application interest: speech perception

III. Planned future work
   - Application interest: speech perception
     (Controllable speech synthesis will be incorporated later)

Part I: Recent publications

Lightning talks on selected articles produced since leaving CSTR:

1. Speaker adaptation and control using input codes
2. Learning controllable TTS from annotated and latent variation
3. Deep encoder-decoder models for unsupervised learning of controllable speech synthesis

Part I: Recent publications

Lightning talks on selected articles produced since leaving CSTR:

1. Speaker adaptation and control using input codes
2. Learning controllable TTS from annotated and latent variation
3. Deep encoder-decoder models for unsupervised learning of controllable speech synthesis

Publication 1

This segment briefly summarises:

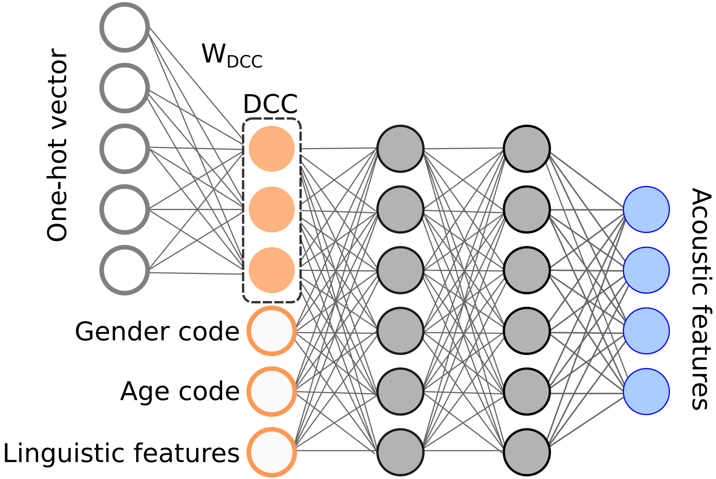
Luong, H.-T., Takaki, S., Henter, G. E., and Yamagishi, J. (2017).

Adapting and controlling DNN-based speech synthesis using input codes.
In *Proc. ICASSP*, pages 4905–4909

Objective

- Investigate the use of different *input codes*. . .
  - Providing different types of speaker information
  - Using different encoding schemes
- . . . for. . .
  a. Multi-speaker synthesis
  b. Speaker adaptation
  c. Speaker morphing and modification
- . . . in statistical parametric speech synthesis (SPSS)

Input codes



One-hot vector
W_{DCC}
DCC
Gender code
Age code
Linguistic features
Acoustic features

Data

- Japanese Voice Bank corpus
- 112 training speakers (56 of each gender)
  - 23 held-out adaptation speakers (9 M, 14 F)
  - $\approx$100 training/adaptation utterances each
  - 10 utterances held-out for every speaker
- Ages 10 through 89
  - 8 per gender and age group (decade) in training data

Input codes considered

- Speaker code encoding schemes:
    - One-hot (112 speakers $\Rightarrow$ 112 dim)
    - Average (in one-hot model)
        - Does not vary across speakers
    - Random (8 or 112 dim)
    - Learned (8 or 112 dim)
        - "Discriminant condition codes" (DCC) (Xue et al., 2014)
        - This learns both a control knob and where to set it
- Gender and age code encoding schemes:
    - One-hot (2 genders; 7 age brackets)
    - Numerical (binary flag; age bracket midpoints)

Multi-speaker synthesis results

- Neural-network acoustic models and (where applicable) input codes were trained to minimise MSE using backpropagation
- Objective findings:
  - Input codes vastly improved MCD and F0 RMSE
  - MCD decreased steadily with increasing DCC size
- Subjective MOS-test findings:
  - Only 9 listeners and 4 random utterances per method, so no statistically significant differences
  - Categorical gender and age codes performed worst in both quality and speaker similarity

Speaker adaptation results

- For adaptation, we keep the network fixed and only learn speaker-specific input codes using backpropagation on a small amount of data from the new speaker
  - Optimally embeds new speakers in the existing speaker space
- Objective findings:
  - Adaptation vastly improved MCD and F0 RMSE
  - Slightly worse numbers than on training speakers
  - MCD and F0 RMSE decreased steadily with increasing DCC size
- Subjective preference-test findings:
  - No adaptation $<$ speaker-code adaptation $<$ speaker-code adaptation with categorical oracle age and gender $<$ speaker-code adaptation with numerical oracle age and gender
  - Speaker-encoding scheme and dimensionality did not matter much

Speaker-trait manipulation results

- No formal evaluation performed
  - No reference to evaluate against
- Listening examples including manipulation and morphing are available at www.hieuthi.com/papers/icassp2017
  - Let's hear a few examples!

Part I: Recent publications

Lightning talks on selected articles produced since leaving CSTR:

1. Speaker adaptation and control using input codes
2. Learning controllable TTS from annotated and latent variation
3. Deep encoder-decoder models for unsupervised learning of controllable speech synthesis

Publication 2

This segment briefly summarises:

Henter, G. E., Lorenzo-Trueba, J., Wang, X., and Yamagishi, J.

(2017). Principles for learning controllable TTS from annotated and latent variation.
In *Proc. Interspeech*, pages 3956–3960

Objectives

- Point out that many approaches for unsupervised learning of TTS control use the same training heuristic
  - "DCC" (Luong et al., 2017) and "sentence-level control vectors" (Watts et al., 2015) are mathematically identical
  - Both try to learn a control knob and per-example control-knob settings that explains the data as well as possible
- Provide a theoretical interpretation of this heuristic
  - Based on the theory of latent (unobserved) variables
- Demonstrate the use of the approach for learning to control unannotated nuances in emotional expression

Heuristic training criterion

- Assume a statistical model (joint density function) $f_{\boldsymbol{X},\,\boldsymbol{Z}|\boldsymbol{L}}(\boldsymbol{x},\,\boldsymbol{z}\mid\boldsymbol{l};\,\boldsymbol{\theta}) = f_{\boldsymbol{X}|\boldsymbol{L},\,\boldsymbol{Z}}(\boldsymbol{x}\mid\boldsymbol{l},\,\boldsymbol{z};\,\boldsymbol{\theta})\,f_{\boldsymbol{Z}|\boldsymbol{L}}(\boldsymbol{z}\mid\boldsymbol{l};\,\boldsymbol{\theta})$, where:
  - $\boldsymbol{X}$ is the speech
  - $\boldsymbol{Z}$ are the unknown (latent) control parameters
  - $\boldsymbol{L}$ are the given linguistic features we condition on
  - $\boldsymbol{\theta}$ contains the model parameters (network weights)
- Let the training data be $\mathcal{D} = \{(\boldsymbol{l}_n,\,\boldsymbol{x}_n)\}$
- Simultaneously estimate network weights and unknown control parameters through the criterion $\widetilde{\mathcal{L}}$:

$$\left\{\widehat{\boldsymbol{\theta}},\,\widehat{\boldsymbol{z}}_n\right\} = \operatorname*{argmax}_{\{\boldsymbol{\theta},\,\boldsymbol{z}_n\}} \widetilde{\mathcal{L}}\left(\{\boldsymbol{\theta},\,\boldsymbol{z}_n\}\mid\mathcal{D}\right)$$

$$= \operatorname*{argmax}_{\{\boldsymbol{\theta},\,\boldsymbol{z}_n\}} \sum_n \ln f_{\boldsymbol{X}|\boldsymbol{L},\,\boldsymbol{Z}}\left(\boldsymbol{x}_n\mid\boldsymbol{l}_n,\,\boldsymbol{z}_n;\,\boldsymbol{\theta}\right)$$

The principled method

- A more principled approach would be to use maximum-likelihood (MLE) and maximum a-posteriori (MAP) estimation:

$$\begin{aligned}
\widehat{\boldsymbol{\theta}}_{\mathrm{ML}} &= \operatorname*{argmax}_{\boldsymbol{\theta}} \mathcal{L}\left(\boldsymbol{\theta} \mid \mathcal{D}\right) \\
&= \operatorname*{argmax}_{\boldsymbol{\theta}} \sum_{n} \ln f_{\boldsymbol{X} \mid \boldsymbol{L}}\left(\boldsymbol{x}_n \mid \boldsymbol{l}_n;\, \boldsymbol{\theta}\right) \\
&= \operatorname*{argmax}_{\boldsymbol{\theta}} \sum_{n} \ln \int f_{\boldsymbol{X},\, \boldsymbol{Z} \mid \boldsymbol{L}}\left(\boldsymbol{x}_n,\, \boldsymbol{z} \mid \boldsymbol{l}_n;\, \boldsymbol{\theta}\right) \mathrm{d}\boldsymbol{z} \\
\widehat{\boldsymbol{z}}_{\mathrm{MAP}n} &= \operatorname*{argmax}_{\boldsymbol{z}} f_{\boldsymbol{Z} \mid \boldsymbol{L},\, \boldsymbol{X}}\left(\boldsymbol{z} \mid \boldsymbol{l}_n,\, \boldsymbol{x}_n\right)
\end{aligned}$$

  - The integral (marginalisation) is usually infeasible to compute

Main result

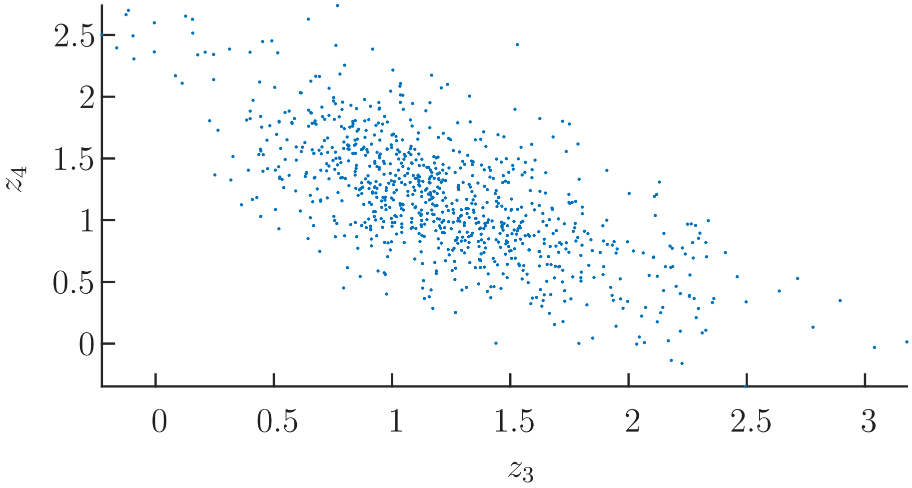
- Assume:
  1. Flat prior: $f_{\boldsymbol{Z}|\boldsymbol{L}}(\boldsymbol{z} \mid \boldsymbol{l};\, \boldsymbol{\theta}) = \text{const.}$ for relevant $\boldsymbol{z}$ and $\boldsymbol{\theta}$
  2. Peaked posterior: $f_{\boldsymbol{Z}|\boldsymbol{L},\boldsymbol{X}}(\boldsymbol{z} \mid \boldsymbol{l},\, \boldsymbol{x};\, \boldsymbol{\theta})$ is a Dirac spike at $\widehat{\boldsymbol{z}}_{\mathrm{MAP}}$
- Then any change in $\boldsymbol{\theta}$ or $\{\boldsymbol{z}_n\}$ that increases $\widetilde{\mathcal{L}}$ also increases $\mathcal{L}$
  - Derived using EM-techniques/Jensen's inequality
  - Assuming iterated optimisation
- Implications:
  - $\widetilde{\mathcal{L}}$ performs approximate likelihood maximisation
  - $\widetilde{\mathcal{L}}$ produces approximate MAP estimates of $\boldsymbol{Z}_n$
  - Unlike MLE, the heuristic $\widetilde{\mathcal{L}}$ does not account for uncertainty in the latent space

Experiment

- Emotional speech database:
  - Japanese-language acted emotional speech
  - 7 emotions (neutral, happy, sad, calm, insecure, excited, angry)
  - 8400 utterances (17 hours) split 80% train, 10% dev., 10% test
- Systems compared:
  - Baseline acoustic model with only emotional category control
  - Proposed model learning heuristic 2D control within each emotional category
- Findings from crowdsourced listening test:
  - The heuristic method learned control parameters that provide perceptually salient control within emotions
  - Learning to control emotional nuance did not degrade emotion recognition compared to the baseline

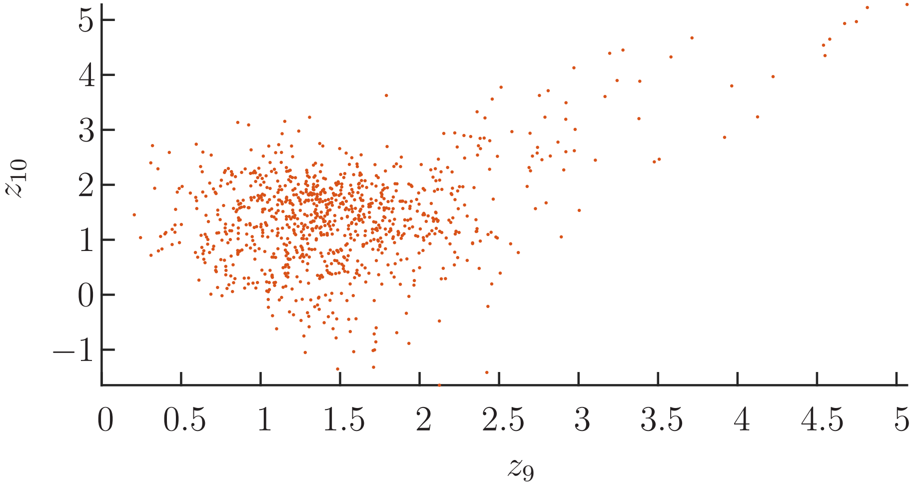
Latent space examples

Learned $\widehat{\boldsymbol{z}}_n$-vectors for happy speech:



Latent space examples

Learned $\widehat{\boldsymbol{z}}_n$-vectors for sad speech:



Part I: Recent publications

Lightning talks on selected articles produced since leaving CSTR:

1. Speaker adaptation and control using input codes
2. Learning controllable TTS from annotated and latent variation
3. Deep encoder-decoder models for unsupervised learning of controllable speech synthesis

Publication 3

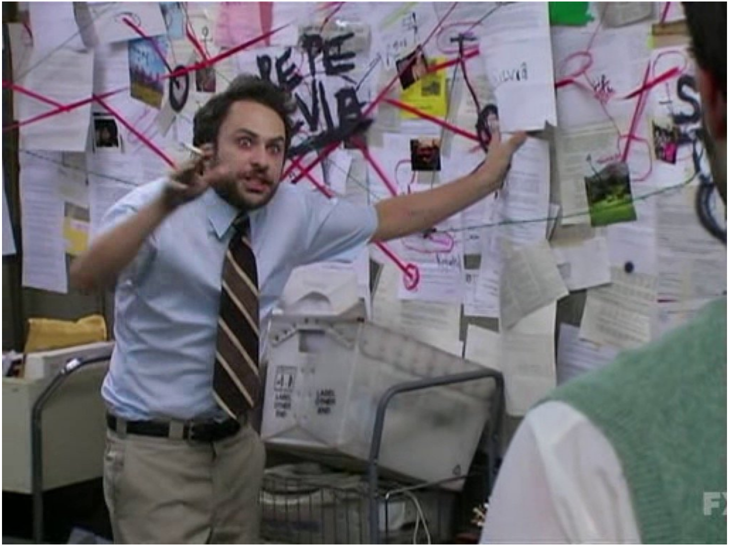
This segment briefly summarises:

Henter, G. E., Lorenzo-Trueba, J., Wang, X., and Yamagishi, J. (2018b). Deep encoder-decoder models for unsupervised learning of controllable speech synthesis.
arXiv preprint arXiv:1807.11470

Objectives

- Survey recent publications on
    - TTS output control, and how to learn it
    - Autoencoders in TTS
- Give a nicer derivation of the same result as in publication 2
- Show that the heuristic method(s) in publication 2 can be interpreted as autoencoders
- Give a probabilistic interpretation of so-called VQ-VAEs
- Relate the heuristics to VQ-VAEs
- Compare the approaches in an application to the same emotional-speech data used in publication 2

"Related work"



Update of previous result

- Under the same assumptions (flat prior, sharp posterior) as in publication 2, it is shown that any change in $\{\boldsymbol{\theta}, \boldsymbol{z}_n\}$ that increases $\widetilde{\mathcal{L}}$ also increases a lower bound on $\mathcal{L}$
  - Derived using variational techniques (evidence lower bound, ELBO)
  - This result allows joint optimisation

Heuristics as autoencoders

- Simple observation:

$$\max_{\boldsymbol{x},\,\boldsymbol{z}} g\left(\boldsymbol{x},\ \boldsymbol{z};\ \boldsymbol{\theta}\right) = \max_{\boldsymbol{x}} g\left(\boldsymbol{x},\ \boldsymbol{z}_{\max}\left(\boldsymbol{x};\ \boldsymbol{\theta}\right);\ \boldsymbol{\theta}\right)$$
$$\text{where } \boldsymbol{z}_{\max}\left(\boldsymbol{x};\ \boldsymbol{\theta}\right) = \operatorname*{argmax}_{\boldsymbol{z}} g\left(\boldsymbol{x},\ \boldsymbol{z};\ \boldsymbol{\theta}\right)$$

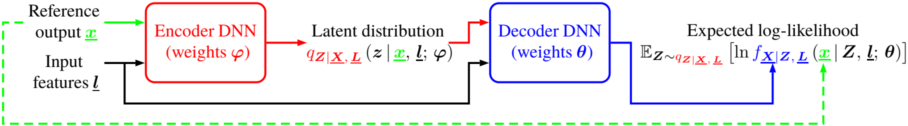
- Assuming most-likely parameter generation (MLPG), we therefore have the (conditional) autoencoder structure:

$$\widehat{\boldsymbol{z}}_{\mathrm{ENC}n}\left(\boldsymbol{x}_n \mid \boldsymbol{l}_n;\ \boldsymbol{\theta}\right) = \operatorname*{argmax}_{\boldsymbol{z}} \ln f_{\boldsymbol{X}\mid\boldsymbol{L},\,\boldsymbol{Z}}\left(\boldsymbol{x}_n \mid \boldsymbol{l}_n,\ \boldsymbol{z};\ \boldsymbol{\theta}\right)$$
$$\widehat{\boldsymbol{x}}_{\mathrm{DEC}n}\left(\widehat{\boldsymbol{z}}_n \mid \boldsymbol{l}_n;\ \boldsymbol{\theta}\right) = \operatorname*{argmax}_{\boldsymbol{x}} \ln f_{\boldsymbol{X}\mid\boldsymbol{L},\,\boldsymbol{Z}}\left(\boldsymbol{x} \mid \boldsymbol{l}_n,\ \widehat{\boldsymbol{z}}_n;\ \boldsymbol{\theta}\right)$$

  - Note: If $f_{\boldsymbol{X}\mid\boldsymbol{L},\,\boldsymbol{Z}}$ is fixed-variance isotropic Gaussian, training minimises the squared error $\sum_n \left(\widehat{\boldsymbol{x}}_{\mathrm{DEC}n} - \boldsymbol{x}_n\right)^2$

Autoencoder schematic

Building blocks of a (variational) autoencoder:



Observations

Some observations regarding the autoencoder interpretation:

- The heuristic method $\widetilde{\mathcal{L}}$ can be seen as an autoencoder where:
  - The encoder and decoder both use the same network, $f_{\boldsymbol{X} \mid \boldsymbol{L}, \boldsymbol{Z}}$
  - The encoder and decoder both include an explicit optimisation operation
    - This can be slow to compute in practice
- The $\widetilde{\mathcal{L}}$ autoencoder is optimised using variational principles
  - We are driven to explore connections to *variational autoencoders*

Variational autoencoders

Variational autoencoders (VAEs) (Kingma and Welling, 2014; Rezende et al., 2014):

- Latent-variable models with a variational posterior for $\boldsymbol{Z}$ that maximise a lower bound (ELBO) on the likelihood
- The decoder describes how $\boldsymbol{Z}$ influences $\boldsymbol{X}$
- The encoder *learns* to perform (approximate) *inference*
  - This is called "amortised inference"
  - Fast at test time and more straightforward to optimise
  - Sub-optimal compared to brute optimisation
    - "Amortisation gap" (Cremer et al., 2018)
- Training often fails because the $\boldsymbol{Z}$-prior term in the objective function dominates the likelihood term
  - The VAE then does not learn any useful control

VQ-VAEs

Vector-quantised VAEs – VQ-VAEs (van den Oord et al., 2017):

- Quantise the encoder-net output $\boldsymbol{z}_e \in \mathbb{R}^D$ into $\boldsymbol{z}_q \subset \mathcal{Z} \in \mathbb{R}^D$
  - $\mathcal{Z}$ is a learned vector-quantisation codebook
- These are less prone to failed learning than regular VAEs
  - VQ-VAE training does not incentivise adherence to the $\boldsymbol{Z}$-prior
  - Trained fine on our emotional speech database, unlike regular VAEs
- Their objective function mixes geometric and probabilistic terms
  - No obvious probabilistic interpretation

New probabilistic interpretation

- Let the latent variable be $\boldsymbol{Z} = (\boldsymbol{Z}_q, \boldsymbol{Z}_e)$ and assume:
  - $\boldsymbol{Z}_q$ is discrete and uniform over $\mathcal{Z}$
  - $\boldsymbol{Z}_e$ is a fixed-variance isotropic Gaussian given $\boldsymbol{z}_q$ with mean $\boldsymbol{z}_q$
    - The latent variable prior $f_{\boldsymbol{Z}}$ is then a GMM
  - $\boldsymbol{X}$ is conditionally independent of $\boldsymbol{Z}_e$ given $\boldsymbol{Z}_q$
  - Variational posteriors are point masses
- We show that variational inference in this model is mathematically equivalent to a VQ-VAE with $\beta = 1$
- The VQ-VAE dependence structure differs from previous VAEs with GMM latent variables
  - Graphical model of VQ-VAE dependencies: $\boldsymbol{Z}_e \leftarrow \boldsymbol{Z}_q \rightarrow \boldsymbol{X}$
  - GMM VAE (Nalisnick et al., 2016): $\boldsymbol{Z}_q \rightarrow \boldsymbol{Z}_e \rightarrow \boldsymbol{X}$

Connections

The heuristic $\widetilde{\mathcal{L}}$ for learning unsupervised control and VQ-VAEs are closely related

- Similarities:
  - Both can be seen as autoencoders
  - Both relate to variational approaches with flat priors and peaked posteriors
  - Neither allows latent-variable uncertainty
- Differences:
  - The $\widetilde{\mathcal{L}}$ heuristic does not perform quantisation
  - VQ-VAEs amortise inference
    - Fast at test time but yields sub-optimal likelihood

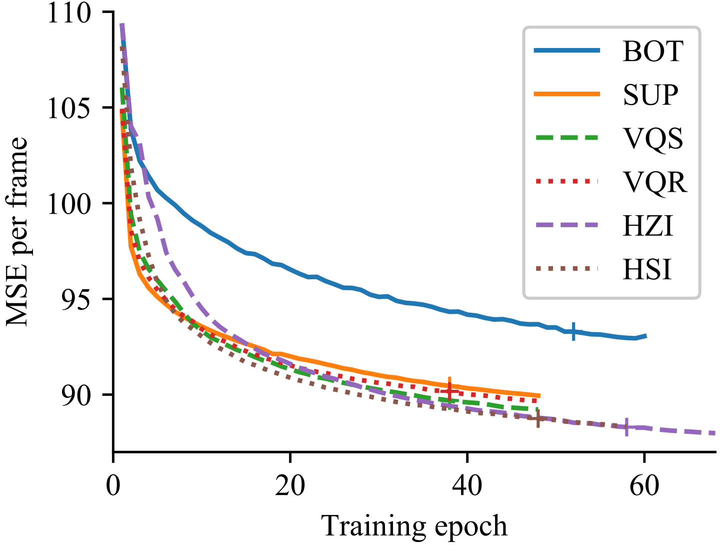
Experimental comparison

To compare the studied techniques, several SPSS systems were trained on the data from publication 2:

- **BOT** Bottom line with no emotional control
- **SUP** The best supervised system trained on this data in (Lorenzo-Trueba et al., 2018a)
  - Unlike all other systems, this system knows the emotional categories and strength of each utterance
- **HZI** $\widetilde{\mathcal{L}}$ heuristic with zero initialisation
- **HSI** $\widetilde{\mathcal{L}}$ heuristic initialisation with the supervised control values from SUP
- **VQR** VQ-VAE with transposed ("reverse") encoder structure
- **VQS** VQ-VAE with non-transposed ("same") encoder structure

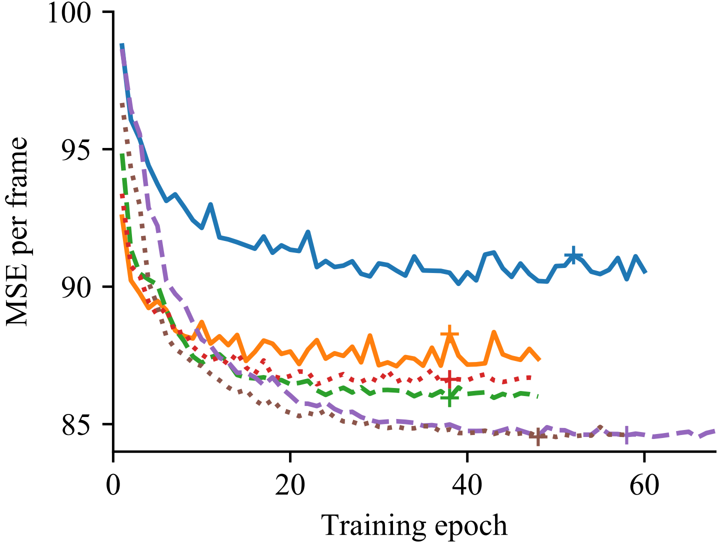
Learning curves

Learning curves on training set:



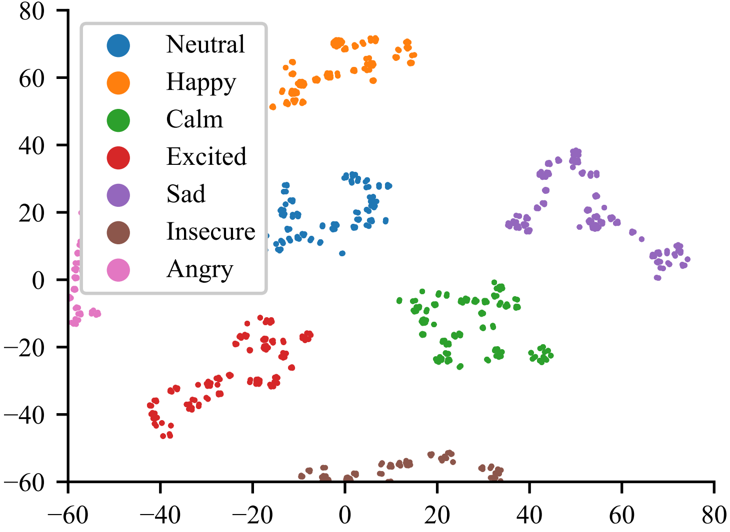
Learning curves

Learning curves on test set:



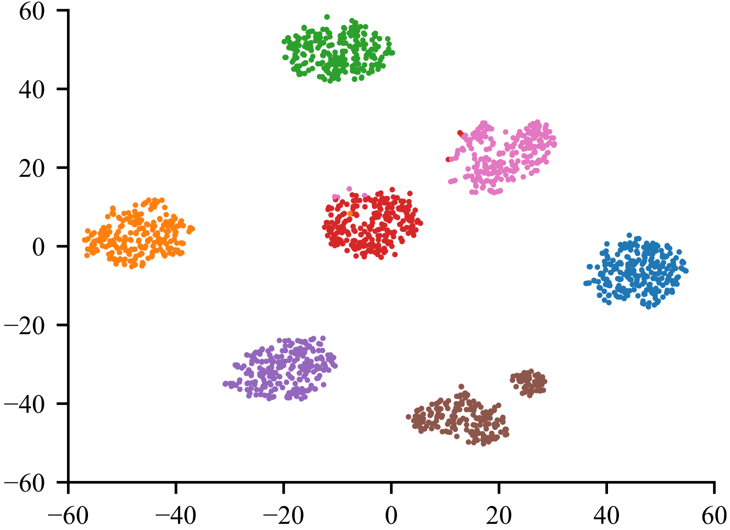
Visualising the control space

2D t-SNE visualisation of 8D SUP vectors on the test-set:



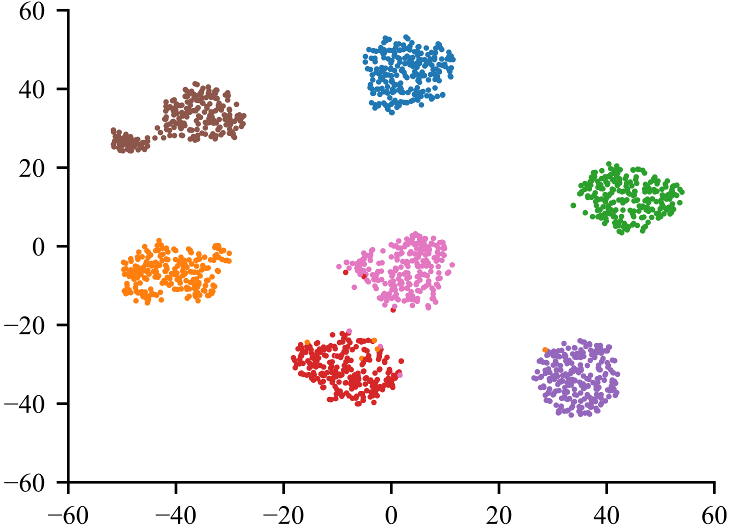
Visualising the control space

2D t-SNE visualisation of 8D HSI vectors on the test-set:



Visualising the control space

2D t-SNE visualisation of 8D HZI vectors on the test-set:



Experimental results

- Objective results:
  - Both heuristic and VQ-VAE synthesisers identified and separated the emotions
  - Unsupervised methods outperformed SUP in terms of MSE
    - Since they can learn better control inputs than sup
  - There is a small but noticeable amortisation gap
  - Heuristic methods took more epochs to terminate
  - Heuristic approaches were not sensitive to initialisation
- Subjective results from a crowdsourced listening test:
  - SUP, HSI, HZI, VQS, and VQR were all comparable in terms of:
    - Perceived speech quality
    - Emotion recognition
    - Perceived emotional strength

Conclusions

What have we learned from the three publications in part I?

- Unsupervised learning of TTS output control is possible, as well as supervised control
  - Both for speaker characteristics and emotional expression
- Many reasonable setups perform similarly in practice
  - VQ-VAEs and the heuristic method(s) can both be interpreted probabilistically as autoencoder setups optimised using variational inference
- VQ-VAEs might be preferred over DCC/"sentence-level control vectors" for future controllable synthesisers
  - Encoding uses forward propagation rather than optimisation and backpropagation, making them easier to train and faster to use

Talk contents

The three parts of today's presentation:

I. Review of some recent publications
II. A more in-depth investigation
III. Planned future work

Part II: Co-author credit

This part is based on:

Henter, G. E., Lorenzo-Trueba, J., Wang, X., Kondo, M., and Yamagishi, J. (2018a). Cyborg speech: Deep multilingual speech synthesis for generating segmental foreign accent with natural prosody.
In *Proc. ICASSP*, pages 4799–4803

Thanks also to Prof. María Luisa García Lecumberri, Prof. Martin Cooke, and Rubén Pérez Ramón on the Diacex project.

Part II: 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

Part II: Cyborg speech

Part II: Cyborg speech

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)

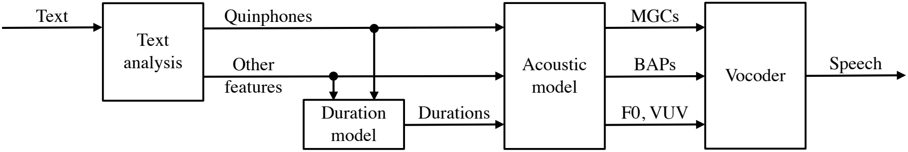
Our approach

- 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

Part II: Cyborg speech

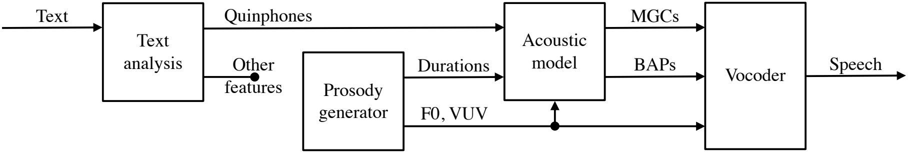
Building the synthesiser

Traditional text-to-speech:



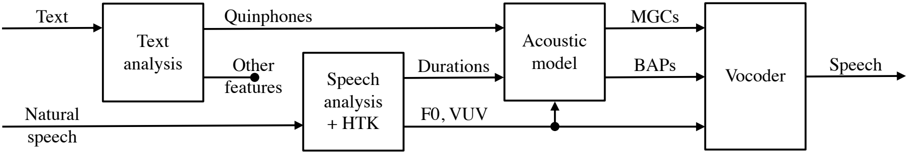
Building the synthesiser

Speech synthesis with arbitrary prosody:



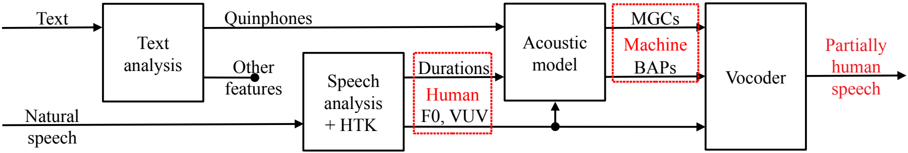
Building the synthesiser

Speech synthesis with natural prosody:

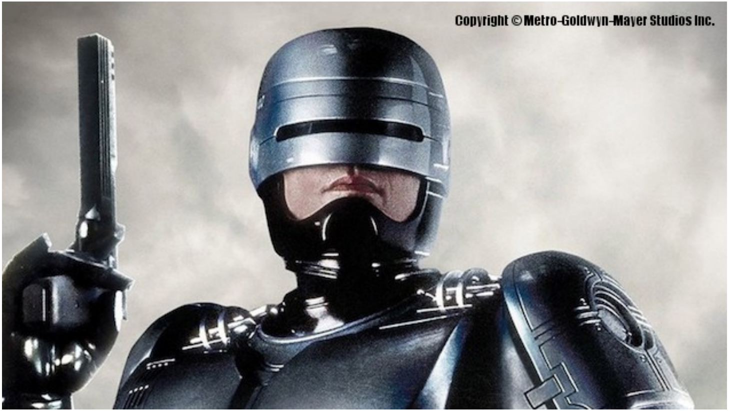


Building the synthesiser

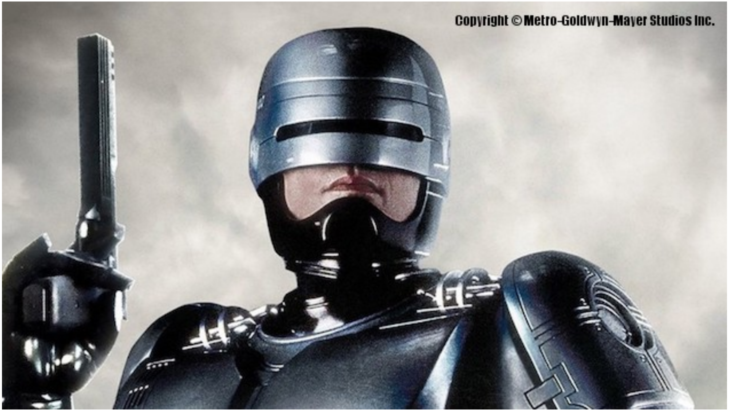
Speech synthesis with natural prosody:



"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

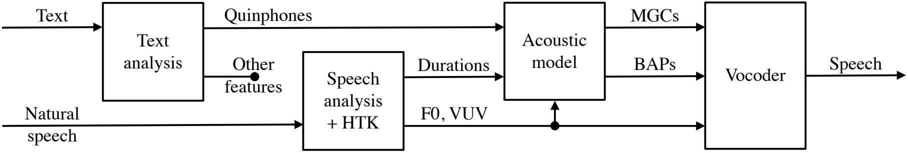
- Segmental foreign accent through multilingual speech synthesis:
  - Teach a single model to synthesise several languages natively
  - During synthesis, interpolate specific phones in the spoken language towards phones in the accent language
  - Maintain the same voice across languages
    - In this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

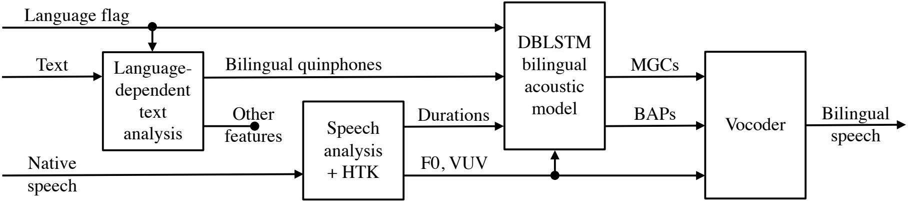
Synthesising foreign accent

Cyborg speech:



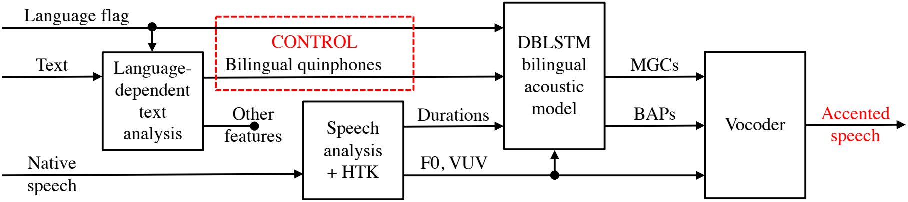
Synthesising foreign accent

Bilingual cyborg speech synthesis:

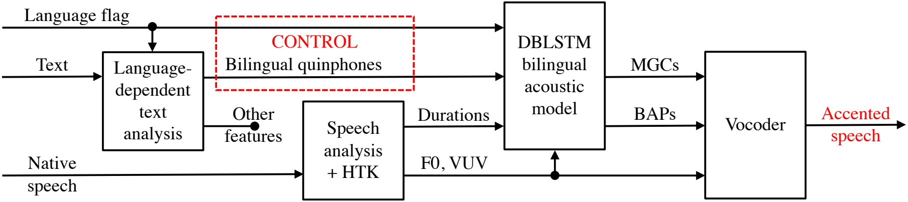


Synthesising foreign accent

Foreign-accented speech synthesis:



Foreign-accented speech synthesis:



Synthetic mispronunciations through cross-language interpolation between 98-dimensional one-hot phone encodings in the quinphones

Part II: Cyborg speech

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)
  - GlottDNN (Airaksinen et al., 2016) pitch extractor
    - Fewer VUV errors
  - Static and dynamic features (MLPG)
- 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	▶	▶	▶	▶	▶	▶

(Note: How perceptible the differences are depends on your native language)

Part II: Cyborg speech

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"

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

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

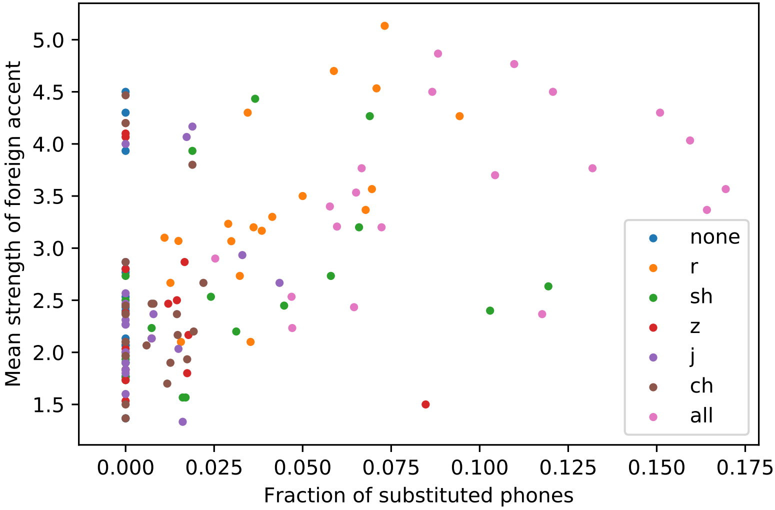
Distribution of perceived accent

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)

Part II: Cyborg speech

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
- Bilingual training did not degrade perception vs. monolingual
- Natural prosody was maintained (high correlation)
- Modelling artefacts were perceived as an “unknown” accent

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
    - Currently in progress at Waseda University together with NII

Talk contents

The three parts of today's presentation:

I. Review of some recent publications
II. A more in-depth investigation
III. Planned future work

Idea for follow-up work

Idea: Continue exploring and expanding the utility of speech synthesis for speech sciences research

- Speech sciences helped TTS get started – now it's time for TTS to return the favour
- Simon King argued for this in his ICPhS 2015 keynote
    - Speech synthesis has evolved rapidly since then
    - Yet there is scant adoption of anything newer than formant synthesis (Klatt, 1980), PSOLA (Charpentier and Stella, 1986), or STRAIGHT (Kawahara, 2006)
- Our plan is to show rather than tell
    - Cyborgs are only the beginning!

Why hasn't this happened already?

Why isn't synthetic speech more commonly used in speech sciences such as speech perception research?

- Is it because speech scientists are unfamiliar with speech technology?
  - CSTR and TMH have what it takes to compensate for this
  - Fine output control is now both possible and learnable
    - Whether accuracy and precision suffice has not been studied
- Is it because research shows that synthetic and natural speech are perceived very differently (Winters and Pisoni, 2004)
  - Casts a shadow of doubt over the generalisability of perception results from TTS studies
  - These results pertain to rule-based formant synthesisers
  - Is this still true today?
- Other hypotheses welcome!

Why hasn't this happened already?

Why isn't synthetic speech more commonly used in speech sciences such as speech perception research?

- Is it because speech scientists are unfamiliar with speech technology?
  - CSTR and TMH have what it takes to compensate for this
  - Fine output control is now both possible and learnable
    - Whether accuracy and precision suffice has not been studied
- Is it because research shows that synthetic and natural speech are perceived very differently (Winters and Pisoni, 2004)
  - Casts a shadow of doubt over the generalisability of perception results from TTS studies
  - These results pertain to rule-based formant synthesisers
  - Is this still true today?
- Other hypotheses welcome!

Known perceptual differences

General findings on rule-based formant speech synthesis lifted from the review in (Winters and Pisoni, 2004):

1. Synthetic speech is less intelligible than natural speech
2. Perception of synthetic speech requires more cognitive resources
3. Perception of synthetic speech interacts with higher-level linguistic knowledge
4. Synthetic speech is more difficult to comprehend than natural speech
5. Perception of synthetic speech improves with experience
6. Alternative populations process synthetic speech differently
7. Prosodic cues, naturalness, and acceptability differences

Recent synthesis improvements

- Intelligibility
  - TTS intelligibility is at ceiling in quiet (King, 2014)
  - Not necessarily true in noise (Cooke et al., 2013)
- Quality/naturalness
  - TTS naturalness has been improving steadily (King, 2014)
  - Neural networks improved SPSS further (Watts et al., 2016)
  - End-to-end approaches improved on SPSS (van den Oord et al., 2016) and can rate close to recorded speech (Shen et al., 2018)
- Speaker similarity
  - Improved hugely (along with naturalness) in voice conversion in the last year (Lorenzo-Trueba et al., 2018b)
- Output control
  - Already discussed in parts I and II of this talk
  - Very impressive style (Wang et al., 2018) and prosody control (Skerry-Ryan et al., 2018) possible with leading end-to-end TTS

Initial research question

Is the output from modern speech-synthesis methods perceived similarly to natural speech recordings?

- For "vanilla" output as well as modified/controlled stimuli

Initial research question

Is the output from modern speech-synthesis methods perceived similarly to natural speech recordings?

- For "vanilla" output as well as modified/controlled stimuli
- We hypothesise:
    1. The gap in perception has closed substantially
    2. Any remaining gap is sufficiently small that robust conclusions now may be drawn from research on synthesised speech

Retesting agenda

Any and all topics in (Winters and Pisoni, 2004) bear revisiting:

1. Synthetic speech is less intelligible than natural speech
2. Perception of synthetic speech requires more cognitive resources
3. Perception of synthetic speech interacts with higher-level linguistic knowledge
4. Synthetic speech is more difficult to comprehend than natural speech
5. Perception of synthetic speech improves with experience
6. Alternative populations process synthetic speech differently
7. Prosodic cues, naturalness, and acceptability differences

Retesting agenda

Any and all topics in (Winters and Pisoni, 2004) bear revisiting:

1. Synthetic speech is less intelligible than natural speech
2. Perception of synthetic speech requires more cognitive resources
3. Perception of synthetic speech interacts with higher-level linguistic knowledge
4. Synthetic speech is more difficult to comprehend than natural speech
5. Perception of synthetic speech improves with experience
6. Alternative populations process synthetic speech differently
7. Prosodic cues, naturalness, and acceptability differences

Measuring cognitive processing demands

- Classic method: Measure differences in response time for different stimuli
  - Reaction times in response to words vs. nonwords (Pisoni, 1981)
  - The measure generalises to other tasks
    - Proposal: Modified rhyming test (MRT) in noise or quiet
- Newer method: Measure pupil dilation during stimulus presentation
  - Already being explored at CSTR (Govender and King, 2018; Simantiraki et al., 2018)
- Should also be coupled with, e.g., a MUSHRA test for speech quality

Methods to compare

- Test how research-grade modern speech synthesisers compare against:
  - Natural speech recordings
  - Classic rule-based formant synthesis
    - With speaker-adapted pitch range and formants
- Two synthesis paradigms:
  - LSTM-based SPSS
  - End-to-end system
- Two types of speech control:
  - Speech in, speech out (SISO), e.g., copy synthesis
    - Common starting point for creating modified speech stimuli for perception research
  - Text in, speech out (TISO), e.g., TTS
- (Modified/controlled speech stimuli not considered at this time)

Proposed system list

- Natural speech recordings
- SISO:
  - MagPhase (Espic et al., 2017) copy synthesis
  - GlotNet (Juvela et al., 2018) copy synthesis
- TISO:
  - Merlin (Wu et al., 2016) with MagPhase
  - Phone-input DCTTS (Tachibana et al., 2018) or Tacotron 2 with GlotNet
  - A speaker-adapted rule-based formant synthesiser
    - If possible

Concrete plans

- Collaborators: Zofia Malisz at KTH, Oliver and Cassia at CSTR
  - Possibly more
- Target: ICPhS 2019
  - Deadlines Dec 4 (abstract) and 11 (full paper)
- Required: A synthesis database with recordings of MRT or utterances with words/nonwords
  - Nick Hurricane data has MRT, but is quite small
  - Other suggestions?

Long-term programme

- Develop...
- Validate...
- Use...

...controllable SISO/TISO tools for speech sciences

The end

The end

Thank you for listening!

The end

Question time!

References I

Airaksinen, M., Bollepalli, B., Juvela, L., Wu, Z., King, S., and Alku, P. (2016).
GlottDNN – A full-band glottal vocoder for statistical parametric speech synthesis.
In *Proc. Interspeech*, pages 2473–2477.

Charpentier, F. J. and Stella, M. G. (1986).
Diphone synthesis using an overlap-add technique for speech waveforms concatenation.
In *Proc. ICASSP*, pages 2015–2018.

Cooke, M., Mayo, C., Valentini-Botinhao, C., Stylianou, Y., Sauert, B., and Tang, Y. (2013).
Evaluating the intelligibility benefit of speech modifications in known noise conditions.
Speech Commun., 55(4):572–585.

Cremer, C., Li, X., and Duvenaud, D. (2018).
Inference suboptimality in variational autoencoders.
In *Proc. ICLR Workshop Track*.

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

References II

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

Espic, F., Valentini-Botinhao, C., and King, S. (2017).
Direct modelling of magnitude and phase spectra for statistical parametric speech synthesis.
In *Proc. Interspeech*, pages 1383–1387.

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.

Govender, A. and King, S. (2018).
Using pupillometry to measure the cognitive load of synthetic speech.
In *Proc. Interspeech*, pages 2838–2842.

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

References III

Henter, G. E., Lorenzo-Trueba, J., Wang, X., Kondo, M., and Yamagishi, J. (2018a).
Cyborg speech: Deep multilingual speech synthesis for generating segmental foreign accent with natural prosody.
In *Proc. ICASSP*, pages 4799–4803.

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

Henter, G. E., Lorenzo-Trueba, J., Wang, X., and Yamagishi, J. (2018b).
Deep encoder-decoder models for unsupervised learning of controllable speech synthesis.
arXiv preprint arXiv:1807.11470.

Juvela, L., Tsiaras, V., Bollepalli, B., Airaksinen, M., Yamagishi, J., and Alku, P. (2018).
Speaker-independent raw waveform model for glottal excitation.
In *Proc. Interspeech*, pages 2012–2016.

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 IV

Kawahara, H. (2006).
STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds.
Acoust. Sci. Technol., 27(6):349–353.

King, S. (2014).
Measuring a decade of progress in text-to-speech.
Loquens, 1(1).

Kingma, D. P. and Welling, M. (2014).
Auto-encoding variational Bayes.
In *Proc. ICLR*.

Klatt, D. H. (1980).
Software for a cascade/parallel formant synthesizer.
J. Acoust. Soc. Am., 67(3):971–995.

Lorenzo-Trueba, J., Henter, G. E., Takaki, S., Yamagishi, J., Morino, Y., and Ochiai, Y. (2018a).
Investigating different representations for modeling and controlling multiple emotions in DNN-based speech synthesis.
Speech Commun., 99:135–143.

References V

Lorenzo-Trueba, J., Yamagishi, J., Toda, T., Saito, D., Villavicencio, F., Kinnunen, T., and Ling, Z. (2018b).
The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods.
arXiv preprint arXiv:1804.04262.

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.

Nalisnick, E. T., Hertel, L., and Smyth, P. (2016).
Approximate inference for deep latent Gaussian mixtures.
In *Proc. NIPS 2016 Workshop Bayesian Deep Learn.*

References VI

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

Pisoni, D. B. (1981).
Speeded classification of natural and synthetic speech in a lexical decision task.
J. Acoust. Soc. Am., 70:S98.

Rezende, D. J., Mohamed, S., and Wierstra, D. (2014).
Stochastic backpropagation and approximate inference in deep generative models.
In *Proc. ICML*, number 2, pages 1278–1286.

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.

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.

References VII

Simantiraki, O., Cooke, M., and King, S. (2018).
Impact of different speech types on listening effort.
In *Proc. Interspeech*, pages 2267–2271.

Skerry-Ryan, R., Battenberg, E., Xiao, Y., Wang, Y., Stanton, D., Shor, J., Weiss, R. J., Clark, R., and Saurous, R. A. (2018).
Towards end-to-end prosody transfer for expressive speech synthesis with Tacotron.
In *Proc. ICML*, pages 4693–4702.

Tachibana, H., Uenoyama, K., and Aihara, S. (2018).
Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention.
In *Proc. ICASSP*, pages 4784–4788.

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

van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K. (2016).
WaveNet: A generative model for raw audio.
arXiv preprint 1609.03499.

References VIII

van den Oord, A., Vinyals, O., and Kavukcuoglu, K. (2017).
Neural discrete representation learning.
In *Proc. NIPS*, pages 6309–6318.

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

Wang, Y., Stanton, D., Zhang, Y., Skerry-Ryan, R., Battenberg, E., Shor, J., Xiao, Y., Ren, F., Jia, Y., and Saurous, R. A. (2018).
Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis.
In *Proc. ICML*, pages 5180–5189.

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 IX

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.

Winters, S. J. and Pisoni, D. B. (2004).
Perception and comprehension of synthetic speech.
Research on Spoken Language Processing Progress Report No. 26 (2003–2004).
Speech Research Laboratory, Department of Psychology, Indiana University, Bloomington, Indiana.

Wu, Z., Watts, O., and King, S. (2016).
Merlin: An open source neural network speech synthesis system.
In *Proc. SSW*, pages 218–222.

Xue, S., Abdel-Hamid, O., Jiang, H., Dai, L.-R., and Liu, Q. (2014).
Fast adaptation of deep neural network based on discriminant codes for speech recognition.
IEEE/ACM T. Audio Speech, 22(12):1713–1725.

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.

Part II: 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)