A Hierarchical Predictor of Synthetic Speech Naturalness Using Neural Networks

Takenori Yoshimura[1], Gustav Eje Henter[2,3], Oliver Watts[2], Mirjam Wester[2], Junichi Yamagishi[2,3] and Keiichi Tokuda[1]
([1]Nagoya Institute of Technology, Nagoya, Japan, [2]The University of Edinburgh, Edinburgh, UK, [3]National Institute of Informatics, Tokyo, Japan)

1. Introduction

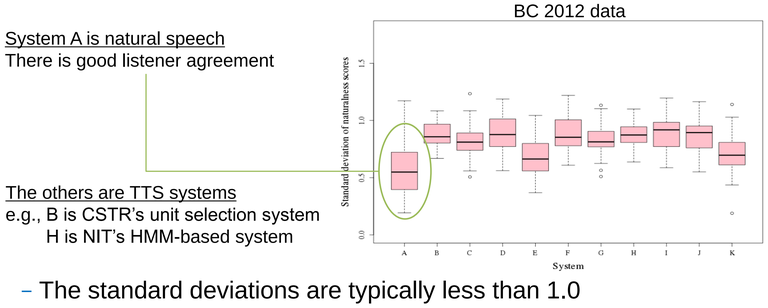
- Objective measures to automatically quantify the naturalness of synthetic speech
  - Distance between parameters extracted from synthetic speech and natural speech (e.g., mel-cepstral distance; MCD)
  - Well-known measures in telecommunications research
  - ⇒ Poor correlation with human perception
  - ⇒ Still require expensive subjective tests for TTS tuning
- Goal: Create a new, more accurate objective measure
  - Trained on the result of large-scale subjective evaluations
  - Hierarchically combine linear regression, feed-forward and convolutional neural networks (CNNs)
  - ⇒ Automatically learn the complex relationship between synthetic speech and a listener's score

2. Blizzard Challenge data

- Blizzard Challenge (BC)
  - Annual challenge to understand and compare research techniques for building speech synthesizers
    - Participants build a synthetic voice from a released speech database and synthesize a given set of test sentences
    - The sentences from each synthesizer are then evaluated through large-scale listening tests
  - The synthetic speech, natural speech, and listener responses are publicly available

	2008	2009	2010	2011	2012	2013
Domain	news novel	news conv	news novel	news novel	news novel	news novel
# Systems	20	17	17	12	10	9
# Stimuli	840	663	612	312	420	477

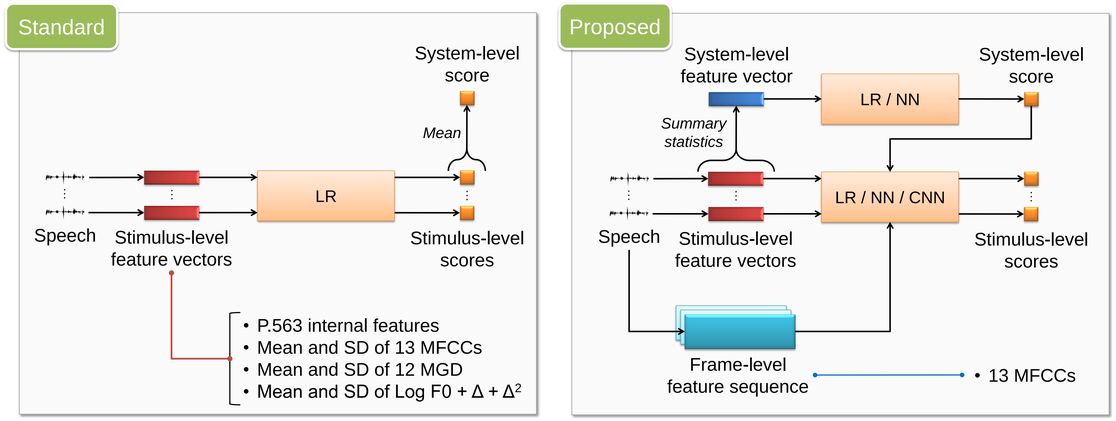
- Investigation of listener agreement
  - Calculate the standard deviation of 5-point opinion scores of naturalness given by listeners for each stimulus



  - The standard deviations are typically less than 1.0
  - ⇒ Mean opinion scores (MOSs) are meaningful prediction targets

3. Speech naturalness prediction using neural networks

- Prediction framework
  - System-level score prediction: the score is typically predicted as the average of predicted stimulus-level scores
  - ⇒ Directly predicting a system-level score using rich features may be effective
  - Stimulus-level score prediction: the score cannot be predicted well compared to system-level prediction
  - ⇒ Combine the two predictions to leverage the robust system-level prediction
- Overcoming the limitations of conventional measures
  - Frame-wise nature: Global patterns are ignored
  - Local and global degradation: Difficult to detect local artifacts such as discontinuities
  - ⇒ CNN is suited for automatically capturing various degradations of synthetic speech by stacking multiple convolutional-pooling layers



LR — Elastic net linear regression
- Regularization weights: automatically determined by 5-fold cross-validation

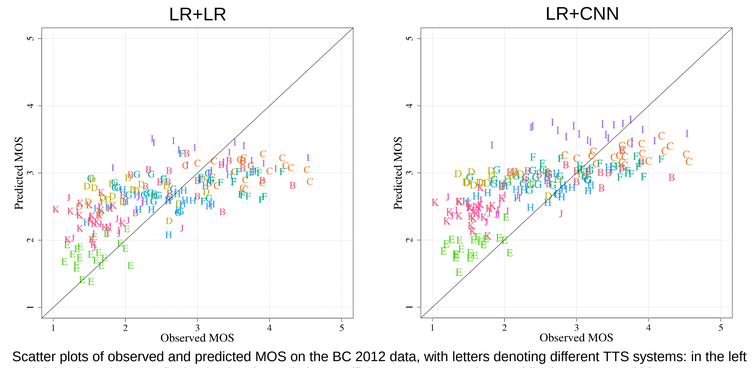
NN — Feed-forward neural network
- System-level prediction: 290-64-1
- Stimulus-level prediction: 96-16-17-1

CNN — Convolutional neural network
- 1st conv. layer: 4 filters (13 x 15)
- 1st pool. layer: max-pooling over 2 units
- 2nd conv. layer: 26 filters (1 x 15)
- 2nd pool. layer: max-pooling over all units

- Leave-one-year-out cross-validation test

		LR	LR+LR	LR+NN	LR+CNN	NN+NN
System-level	RMSE	0.52		0.43		**0.33**
	ρ_s	0.55		**0.74**		0.72
Stimulus-level	RMSE	0.78	0.68	**0.68**	0.69	**0.68**
	ρ_s	0.40	0.56	0.57	**0.58**	0.57
Stimulus-level (within-system)	ρ_s	-	0.11	-	**0.17**	-

Root mean square error (RMSE) and Spearman's rank correlation coefficient ρ_s between listener MOS and predicted MOS, where the values are the average over six years



Scatter plots of observed and predicted MOS on the BC 2012 data, with letters denoting different TTS systems: in the left and right plots, the overall stimulus-level correlation coefficients are 0.73 and 0.79 while the average within-system stimulus-level correlation coefficients are 0.04 and 0.18, respectively

- Hierarchical prediction outperforms standard prediction
  - Hierarchy structure can consider the relation between the two levels
- NN+NN obtains the lowest system-level RMSE
  - Simultaneous optimization might compensate for the lack of training data
- CNN does not work well on the stimulus-level
  - Lack of training data (only less than 3000 stimuli) leads to overfitting
  - Large acoustic variation caused by speakers, domains, TTS systems, etc.
  - Difficult to capture linguistic context
  - ⇒ May require additional information
    e.g., linguistic label, degradation annotation, and acoustic features
- Within system, some success is observed for CNN compared to LR
  - CNN may be able to extract important signals affecting human perception

4. Conclusions

- We investigated hierarchical and CNN approaches for synthetic-speech naturalness prediction
  - Improved several aspects, but the prediction is still challenging
  - Limited data and various acoustic/linguistic factors make the prediction difficult
- Future work
  - Append linguistic information to input features