SSW11 Speech Synthesis Workshop
Aug 26-28 2021
hte

How do Voices from Past Speech Synthesis Challenges Compare Today?

Erica Cooper, Junichi Yamagishi

August 28, 2021

National Institute of Informatics, Japan

Table of contents

1. Introduction
2. Listening Test
3. Results
4. Analysis of Natural Speech
5. Discussion and Future Work

Introduction

Project

- **A very large-scale listening test combining samples from many past Blizzard Challenges and Voice Conversion Challenges**
  - MOS results from separate past listening tests cannot be meaningfully combined and compared because the set of systems and therefore the context of the test are completely different.
  - Conducting a new test with these samples combined will enable more direct comparisons.
  - The data gathered can be used to train MOS prediction models.

Motivation

- How **reliable** and **reproducible** are MOS scores?
- How do past listening test results **compare** to ratings gathered in the present day?
- Will the results **correlate** even when the listening test context has changed?
- What **observations** can we make about speech synthesis and voice conversion systems over the years?
- What are the effects of the **speaker** of the dataset on synthesized speech quality?
- **Collect a very large-scale database of a variety of synthesizer outputs and their MOS ratings for the purpose of training MOSnet-like systems for automatic MOS prediction**

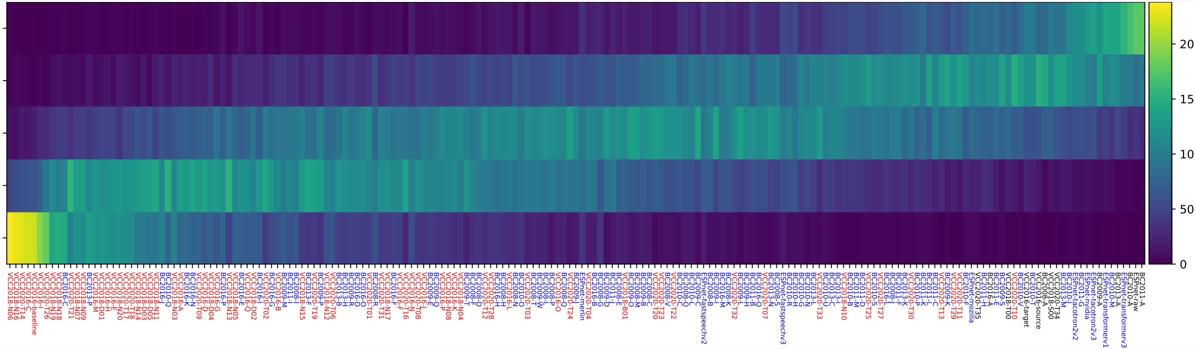
Listening Test

Large-scale listening test

- **187 different systems** from past BC, VCC, and ESPnet TTS
  - BC 2008, 2009, 2010, 2011, 2013, 2016 (English **EH1** tasks only)
  - VCC 2016, 2018, 2020 (**same-language** task only)
  - ESPnet TTS: samples from **ICASSP 2020** trained on LJSpeech
- **38 utterances per system**
  - Samples are balanced over genre where relevant
  - Samples are balanced over source and target speakers for VCC
  - Only genres included in the original naturalness tests are included
- **Test design**
  - One sample from each system per set
  - One listener rates one set containing 187 samples
  - Coverage of **8 Japanese listeners per set; 304 total listeners**
  - **MOS rating for naturalness** on a scale of 1-5
- Significant differences: Mann-Whitney U test ($p<0.05$) with Bonferroni correction

Results

Listening test results



Listening test results

Best systems

- ESPnet-transformerv3
- BC2010-M
- ESPnet-transformerv1
- ESPnet-tacotron2v3
- ESPnet-nvidia

Worst systems

- VCC2018-N06
- VCC2018-N16
- VCC2020-T14
- VCC2016-C
- VCC2016-baseline

Listening test results

Best systems

- ESPnet-transformerv3
- BC2010-M
- ESPnet-transformerv1
- ESPnet-tacotron2v3
- ESPnet-nvidia

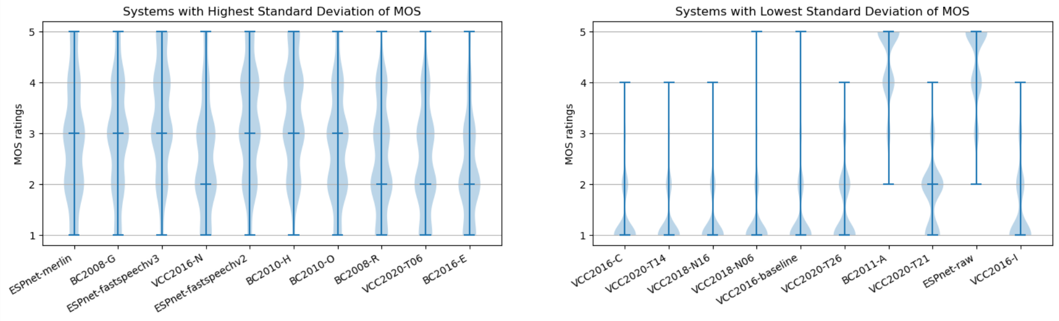
Worst systems

- VCC2018-N06
- VCC2018-N16
- VCC2020-T14
- VCC2016-C
- VCC2016-baseline

VCC typically has a smaller amount of data

Listening test: Agreement

Agreement = 0.5 (Krippendorff's Alpha and Intra-Class Correlation)



Merlin has the largest variation in scores

Do new MOS results correlate with original ones?

System-level and utterance-level Pearson's correlation coefficient, Spearman Rank correlation coefficient, and root mean squared error between original and new listening test results by challenge or set of systems

	System-level			Utterance-level		
Challenge	**PCC**	**SRCC**	**RMSE**	**PCC**	**SRCC**	**RMSE**
BC2008	0.93	0.89	0.33	0.70	0.67	0.62
BC2009	0.97	0.95	0.48	0.76	0.72	0.64
BC2010	0.93	0.98	0.66	0.74	0.73	0.85
BC2011	0.91	0.90	0.76	0.76	0.67	0.87
BC2013	0.97	0.98	0.49	-	-	-
BC2016	0.97	0.93	0.40	-	-	-
VCC2016	0.97	0.92	0.42	0.56	0.53	1.12
VCC2018	0.96	0.91	0.77	0.55	0.53	1.10
VCC2020	0.98	0.96	0.23	0.87	0.87	0.48
ESPnet	0.99	0.98	0.09	0.73	0.61	0.59

Do new MOS results correlate with original ones?

System-level and utterance-level Pearson's correlation coefficient, Spearman Rank correlation coefficient, and root mean squared error between original and new listening test results by challenge or set of systems

	System-level			Utterance-level		
Challenge	**PCC**	**SRCC**	**RMSE**	**PCC**	**SRCC**	**RMSE**
BC2008	0.93	0.89	0.33	0.70	0.67	0.62
BC2009	0.97	0.95	0.48	0.76	0.72	0.64
BC2010	0.93	0.98	0.66	0.74	0.73	0.85
BC2011	0.91	0.90	0.76	0.76	0.67	0.87
BC2013	0.97	0.98	0.49	-	-	-
BC2016	0.97	0.93	0.40	-	-	-
VCC2016	0.97	0.92	0.42	0.56	0.53	1.12
VCC2018	0.96	0.91	0.77	0.55	0.53	1.10
VCC2020	0.98	0.96	0.23	0.87	0.87	0.48
ESPnet	0.99	0.98	0.09	0.73	0.61	0.59

Do new MOS results correlate with original ones?

System-level and utterance-level Pearson's correlation coefficient, Spearman Rank correlation coefficient, and root mean squared error between original and new listening test results by challenge or set of systems

	System-level			Utterance-level		
Challenge	**PCC**	**SRCC**	**RMSE**	**PCC**	**SRCC**	**RMSE**
BC2008	0.93	0.89	0.33	0.70	0.67	0.62
BC2009	0.97	0.95	0.48	0.76	0.72	0.64
BC2010	0.93	0.98	0.66	0.74	0.73	0.85
BC2011	0.91	0.90	0.76	0.76	0.67	0.87
BC2013	0.97	0.98	0.49	-	-	-
BC2016	0.97	0.93	0.40	-	-	-
VCC2016	0.97	0.92	0.42	0.56	0.53	1.12
VCC2018	0.96	0.91	0.77	0.55	0.53	1.10
VCC2020	0.98	0.96	0.23	0.87	0.87	0.48
ESPnet	0.99	0.98	0.09	0.73	0.61	0.59

Do new MOS results correlate with original ones?

System-level and utterance-level Pearson's correlation coefficient, Spearman Rank correlation coefficient, and root mean squared error between original and new listening test results by challenge or set of systems

	System-level			Utterance-level		
Challenge	**PCC**	**SRCC**	**RMSE**	**PCC**	**SRCC**	**RMSE**
BC2008	0.93	0.89	0.33	0.70	0.67	0.62
BC2009	0.97	0.95	0.48	0.76	0.72	0.64
BC2010	0.93	0.98	0.66	0.74	0.73	0.85
BC2011	0.91	0.90	0.76	0.76	0.67	0.87
BC2013	0.97	0.98	0.49	-	-	-
BC2016	0.97	0.93	0.40	-	-	-
VCC2016	0.97	0.92	0.42	0.56	0.53	1.12
VCC2018	0.96	0.91	0.77	0.55	0.53	1.10
VCC2020	0.98	0.96	0.23	0.87	0.87	0.48
ESPnet	0.99	0.98	0.09	0.73	0.61	0.59

Do MOS scores improve year by year?

Best system in each challenge compared to the previous challenge's best system

Year : Best system	MOS	Improved?	Significant?
BC2008 : J	3.63		
BC2009 : S	3.87	✓	x
BC2010 : M	4.27	✓	✓
BC2011 : G	4.12	x	x
BC2013 : M	4.01	x	x
BC2016 : L	3.63	x	✓
VCC2016 : O	2.86		
VCC2018 : N10	3.55	✓	✓
VCC2020 : T10	3.88	✓	x
ESPnet : transformerv3	4.33		

Do MOS scores improve year by year?

Best system in each challenge compared to the previous challenge's best system

Year : Best system	MOS	Improved?	Significant?
BC2008 : J	3.63		
BC2009 : S	3.87	✓	x
BC2010 : M	4.27	✓	✓
BC2011 : G	4.12	x	x
BC2013 : M	4.01	x	x
BC2016 : L	3.63	x	✓
VCC2016 : O	2.86		
VCC2018 : N10	3.55	✓	✓
VCC2020 : T10	3.88	✓	x
ESPnet : transformerv3	4.33		

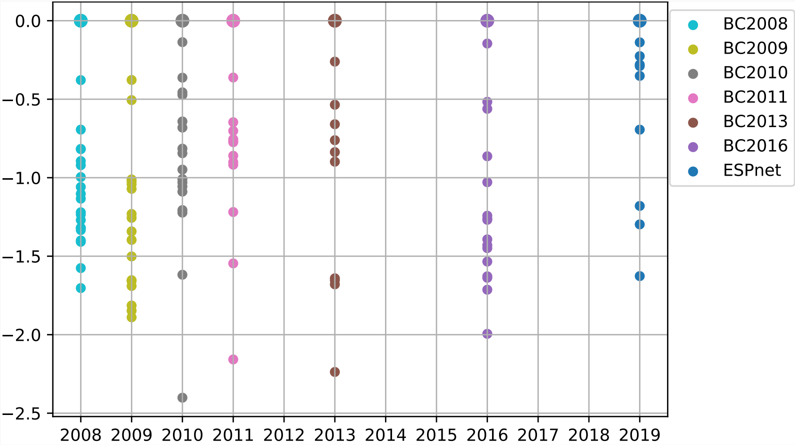
At what point did TTS quality reach that of natural speech?

Is the year's best system significantly different from that year's natural speech?

Year : Best system	Significant difference from natural speech?
BC2008 : J	✓
BC2009 : S	✓
BC2010 : M	x
BC2011 : G	✓
BC2013 : M	x
BC2016 : L	x
VCC2016 : O	✓
VCC2018 : N10	x
VCC2020 : T10	x
ESPnet : transformerv3	x

At what point did TTS quality reach that of natural speech?

Difference of each system from natural speech, computed from averaged z-score-normalized ratings by listener for each challenge



Correlation with objective measures

- SNR: r=0.17
- P.563: r=0.05
- MOSnet trained on ASVspoof: r=0.03

Correlation with objective measures

- SNR: r=0.17
- P.563: r=0.05
- MOSnet trained on ASVspoof: r=0.03

Room for improvement of objective measures

Analysis of Natural Speech

Natural speech preferences and effects of corpus on TTS

- The effect of speech corpus on perceived TTS quality is well-documented:
  - J. Williams, J. Rownicka, P. Oplustil, and S. King, “Comparison of speech representations for automatic quality estimation in multi-speaker text-to- speech synthesis,” 2020
  - F. Hinterleitner, C. Manolaina, and S. Moller, “Influence of a voice on the quality of synthesized speech,” 2014
- Since every challenge uses a different corpus, this is a confounding factor to making meaningful direct comparisons across challenges, but it is still important to capture preferences regarding these factors for training a MOS prediction model.

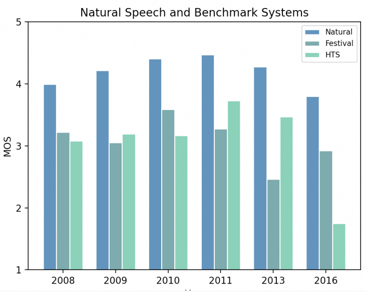
Natural speech metadata and genre

- **Professional speakers** were rated as significantly more natural than non-professional speakers
- **Female speakers** had a marginally-significantly (p=0.05) higher MOS than male speakers
- **No significant differences** between British and American speakers
- Genres: news, book, conversational
  - News rated as most natural (MOS=4.36)
  - Conversational: MOS=4.14
  - Book: MOS=4.09 (significantly lower)

Speaker characteristics

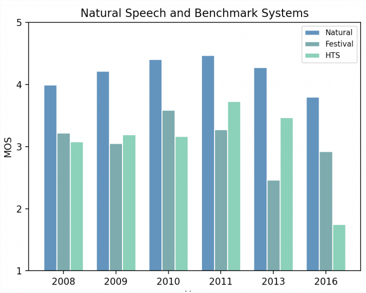
- Standard Praat features:
  - Minimum, maximum, mean, and standard deviation of f0 and energy
  - NHR, jitter, shimmer
- Moderate **negative** correlations with MOS for **shimmer** (r=-0.46), **NHR** (r=-0.41), and **mean energy** (r=-0.37)
- Moderate **positive** correlations with MOS for **standard deviation of energy** (r=0.41)

Effect of corpus on benchmark systems



- Moderate correlations for Festival
    - Pearson r=0.33
    - Spearman r=0.54
- Strong correlations for HTS
    - Pearson r=0.87
    - Spearman r=0.90

Effect of corpus on benchmark systems



- Moderate correlations for Festival
    - Pearson r=0.33
    - Spearman r=0.54
- Strong correlations for HTS
    - Pearson r=0.87
    - Spearman r=0.90

HTS output quality more closely matches the quality of the training data.

Discussion and Future Work

Discussion and future work

- We have a large dataset for training MOSnet-type systems
- Strong correlations with past listening tests
- Choice of speaker for training data is very important
- Will repeating the test with English listeners reveal language-dependent or cultural factors?
- Some systems have clear agreements whereas others have a wider distribution of scores.
    - What makes certain systems so "controversial"?
    - Are certain types of artifacts or unnaturalness more salient to some listeners than to others?
    - Analysis of listener differences
    - Incorporate variance of scores into MOSnet

Questions?