The Blizzard Challenge 2019

Zhizheng Wu, Zhihang Xie, Simon King[†]

[†]The Centre for Speech Technology Research
University of Edinburgh, UK

wuzhizheng@gmail.com, Simon.King@ed.ac.uk

Abstract

The Blizzard Challenge 2019 is the fifteen annual Blizzard Challenge and is the twelfth consecutive one organised by the University of Edinburgh, with support from the other members of the Blizzard Challenge committee. The task this year was to build a text-to-speech system from the anchor's voice in 8-hour corpus of online talk shows hosted by a single Chinese male celebrity. The recording environment was unconstrained.

Index Terms: Blizzard Challenge, speech synthesis, evaluation, listening test

1. Introduction

Black and Tokuda conceived the Blizzard Challenge in 2005 [1] and there have been annual summary papers like this one every year, plus a one-off retrospective summary-of-summaries covering the first decade [2]. For many previous Challenges, the submitted speech, reference natural samples, raw listening test responses, scripts for running the listening test and scripts for the statistical analysis, can be obtained from the Blizzard Challenge website [3].

2. Participants

37 teams registered for this year's challenge and obtained the data. Of the 37 registered teams, 24 teams submitted entries which is a historic record number. The submitted entries alongside human speech and one benchmark system are summarized in Table 1 .

The DNN parametric benchmark[1] uses the publicly-available Merlin recipe combined with NTU's internal Chinese front-end.

When reporting results, the systems are identified using letters, with A denoting natural speech, B the DNN benchmark system and the remaining letters denoting the systems submitted by participants in the challenge. The system identifiers are assigned randomly each year. Most participating teams reveal their system identifier in their workshop paper. This year, for the first time in the history of the challenge, a few teams failed to submit a paper describing their system: these are noted in Table 1.

3. Voice to build

3.1. Speech database

The speech data is from the Chinese Luogic talk show program "Everyday 60 seconds". In the program, the Chinese male celebrity, Zhenyu Luo, publishes 60 seconds of speech every morning in a WeChat post. The speech is spontaneous and expressive and contains many fillers. The recording conditions and device are unknown and uncontrolled.

For the 2019 challenge, 8 hours of these recordings were available to participants. The recordings were released in their original 60 s duration long-form format without segmenting into utterances. As in all Blizzard Challenges, the organisers held out some material for use as part of the test set. This material was randomly selected from the collected recordings.

3.2. Task

The participants were tasked with building a single synthetic voice from the provided data.

A test set totalling 2546 sentences was synthesised by each participating team, and submitted as 16-bit / 16 kHz or 24 kHz uncompressed audio. These test sentences were drawn from the held-out portion of the corpus, poems, Wikipedia and newspapers.

For testing naturalness, we used only the held-out talk show sentences. Intelligibility was tested used both held-out talk show and Wikipedia sentences. The remaining sentences are intended for future tests and analysis.

3.3. Listener types

Similarly to previous years, various listener types were used in the test. The letters in parenthesis below are the identifiers used for each type in the results distributed to participants:

- Paid Edinburgh University students, all native speakers of Chinese (any accent) and generally aged 18-25. These were recruited in Edinburgh and carried out the test in purpose-built soundproof listening booths using good quality audio interfaces and headphones. (EP)
- Speech experts (self-declared), recruited via participating teams and mailing lists. (EE)
- Volunteers recruited via participating teams, mailing lists, blogs, word of mouth, etc. (ER)

As in all previous challenges, participating teams were asked to help recruit volunteer participants (in categories EE or ER) for the listening test. Table 2 summarises the listeners who participated this year.

3.4. Listening test completion rate

Table 2 gives a breakdown of evaluation completion rates for the naturalness and similarity sections of the listening test, per listener type. Table 4 presents a breakdown of evaluation completion rates for the intelligibility section of the test. The completion rate for MOS and similarity is 100%, as expected. However, the completion rate for PTER and PER are lower. This is because some systems failed to generate intelligible speech.

4. Analysis methodology

For naturalness and similarity, we give results broken down by listener type (paid, speech expert, volunteer) and also for all listeners combined. For intelligibility, we only present results for paid listeners because most volunteers and experts didn't finished this section of the test; including their partial results would risk skewing results, due to the between-subjects design. Analysis by

[1] Thanks to Haihua Xu, NTU, Singapore

Table 1: *The participating systems and their short names. The first row is natural speech (system identifier A) and the second row is the benchmark (system identifier B). The remaining rows are in alphabetical order of the system's short name and not in alphabetical order of system identifier. Systems are categorised as: HMM (Hidden Markov Model statistical parametric), DNN (Deep Neural Network statistical parametric, including architectures such as BLSTM), clustergen (decision tree statistical parametric), waveform concatenation, hybrid (waveform concatenation guided by a statistical parametric model such as a DNN), or neural generative (WaveNet, WaveRNN, Tacotron etc).*

Short name	Details	Method
NATURAL	Natural speech	human
DNN_BM	Merlin + STRAIGHT benchmark	DNN + STRAIGHT vocoder
CMU	Carnegie Mellon University	Seq2Seq + WaveRNN
DeepSound	Guangzhou Deepsound Technology Co. Ltd	Encoder-Decoder + Neural vocoder
DKU	Duke Kunshan University	Encoder-Decoder + Griffin-Lim (GL)
Horizon	Nanjing Horizon Robotics Integrated Circuit Co.,Ltd	Encoder-Decoder + WaveRNN
IIM-USTC	Institute of Intelligent Machines(IIM) Chinese Academy of Sciences + USTC	hybrid
IMU	Computer science department, Inner Mongolia University	Encoder-Decoder + GL
IOA	Institute of Acoustics, Chinese Academy of Sciences	DNN + WORLD vocoder
LINGBAN	Beijing Lingban Technology Co. Ltd	LSTM + WaveNet
Mobvoi	Mobvoi Information & Technology Company	Encoder-Decoder + WaveNet
NLPR	National Laboratory of Pattern Recognition (NLPR), Chinese Academy of Sciences	Tacotron + LPCNet
NTUT	National Taipei University of Technology	HMM + MGC vocoder
Paopao	iQiyi Inc (*No paper submission*)	DNN + GL
PingAnTech	PingAn Technology(*No paper submission*)	E2E (WaveRNN backend)
RoyalFlush	Hithink RoyalFlush Information Network Co. Ltd	Encoder-Decoder + GL
SJTU	ShangHai Jiao Tong University	Tacotron + WaveNet
STC	Speech Technology Center	Encoder-Decoder + LPCNet
SZ-NPU	Sogou Inc + Northwestern Polytechnical University	Tacotron + WaveRNN
T-beta	Tencent Technology Co., Ltd	Encoder-Decoder + WaveNet
TJU	Tianjin University & Didi Chuxing & Huiyan Technology (Tianjin) Co., Ltd.	Tacotron + GL
TL@NTU	Nanyang Technological University, Singapore	Concatenative
USTC	University of Science and Technology of China (USTC)	DNN + WaveNet
UTokyo	The University of Tokyo	DNN + NMF + WORLD
VIVI	Vivo AI Research Center (Shenzhen)	Tacotron + GL
XMU	Xiamen University (*No paper submission*)	E2E + GL

listener type was provided to participants and can be obtained by non-participants by downloading the complete listening test results distribution package via the Blizzard website. Since complete raw listeners scores for every stimulus presented in the listening test are included in this distribution, re-analysis of the data is possible by anyone who wishes to do so. The organizers of the challenge would be interested to hear of any such re-analysis.

Please refer to [4] for a description of the statistical analysis techniques used and justification of the statistical significance techniques employed to produce the results presented here. In all material published by the organizers, system names are anonymised. Individual teams are free to reveal their system identifier if they wish.

5. Results

Standard boxplots are presented for the ordinal data, where the median is represented by a solid bar across a box showing the quartiles; whiskers extend to 1.5 times the inter-quartile range and outliers beyond this are represented as circles. Bar charts are presented for the word error rate type interval data. A single ordering of the systems is employed in all plots. This ordering is in descending order of mean naturalness calculated from the responses of all listeners combined and both sentence-based naturalness sections combined. Note that this ordering is intended only to make the plots more readable by using the same system ordering across all plots for both tasks and *can not be interpreted as a ranking*. In other words, the ordering does not tell us which systems are significantly better than others. Given that the presentation of results as tables, significance matrices, boxplots and bar-charts is now well established, we will not provide a detailed commentary for every result.

Only four systems (B, L, N, U) employed the classic parametric system setup (HMM or DNN with a signal processing-based vocoder such as WORLD or STRAIGHT). From the results, we see that these systems generally achieve a lower naturalness than fully neural systems. Two systems (Q and Z) employed waveform concatenation using a neural network guiding the unit selection, with mixed results: Z achieved above average naturalness and intelligibility while Q was below average.

Unsurprisingly, neural approaches dominated this year's Blizzard Challenge. These systems achieved better naturalness than other approaches. Most achieved sequence-to-sequence regression using an encoder-decoder architecture. A selection of neural waveform generators were used, with Wavenet and WaveRNN being popular choices.

In this year's challenge, when combining the opinions of all listeners, no system was as natural as natural speech (Figures 2 and 3), or as similar to the target speaker. M was significantly more natural than all other systems. Interestingly, self-declared ”speech experts” were collectively of the opinion that system M was as natural as human speech (Figures 4 and 5).

In the intelligibility test, there was no comparison with natural speech this year, and there are a group of systems with equally low error rates (Figures 18 and 19).

6. Acknowledgements

In addition to those people already acknowledged in the text, we wish to thank a number of additional contributors without whom running the challenge would not be possible. Vasilis Karaiskos provided essential advice and wisdom, accumulated over many previous years of the challenge, and sanity-checked the listening test implementation. Rob Clark designed and implemented the scripts used to perform statistical analysis; Dong Wang wrote the WER program. Apple and DataBaker provided financial support. The listening test scripts are based on earlier versions provided by previous organisers of the Blizzard Challenge. Thanks to all participants and listeners.

7. References

[1] Alan W. Black and Keiichi Tokuda, "The Blizzard Challenge - 2005: Evaluating corpus-based speech synthesis on common datasets," in *Proc Interspeech 2005*, Lisbon, Portugal, September 2005.

[2] Simon King, "Measuring a decade of progress in Text-to-Speech," *Loquens*, vol. 1, no. 1, 2014.

[3] "The Blizzard Challenge website," http://www.synsig.org/index.php/Blizzard_Challenge.

[4] R. A. J. Clark, M. Podsiadło, M. Fraser, C. Mayo, and S. King, "Statistical analysis of the Blizzard Challenge 2007 listening test results," in *Proc. Blizzard Workshop (in Proc. SSW6)*, Bonn, Germany, August 2007.

In the tables on the following pages, the footnotes in the captions specify whether the numbers in that table are based on listener feedback [2] or on the listening test results themselves. [3]

[2] These numbers are calculated from the feedback forms that listeners complete at the end of the test. As this is optional, many listeners decide not to fill it in. If they do, they do not always reply to all the questions in the form.

[3] These numbers are calculated from the database where the results of the listening tests are stored.

	Registered	No response at all	Partial evaluation	Completed Evaluation
EE	209	84	0	125
EP	142	1	0	141
ER	216	121	0	95
ALL	**567**	**206**	**0**	**361**

Table 2: *Listener registration and evaluation completion rates for naturalness and similarity evaluation.* [3]

Age	under 20	20-29	30-39	40-49	50-59	60-69	70-79	over 80
EE	2	88	33	0	2	0	0	0
EP	0	135	6	0	0	0	0	0
ER	1	85	8	0	0	0	1	0
Total	**3**	**308**	**47**	**0**	**2**	**0**	**1**	**0**

Table 3: *Age of listeners whose results were used in naturalness and similarity (completed the evaluation fully).* [3]

	Registered	No response at all	Partial evaluation	Completed Evaluation
PER	142	5	20	117
PTER	142	5	26	111

Table 4: *Listener registration and evaluation completion rates for intelligibility in terms of Pinyin (with tones) error rate and Pinyin (without tones) error rate respectively.* [3]

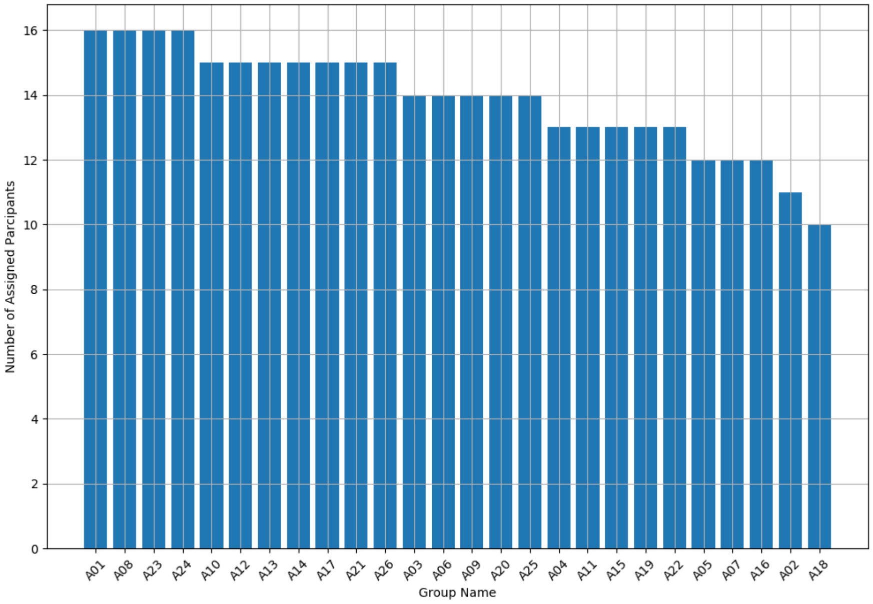

Figure 1: *Number of assigned participants to each listening group in naturalness and similarity.*

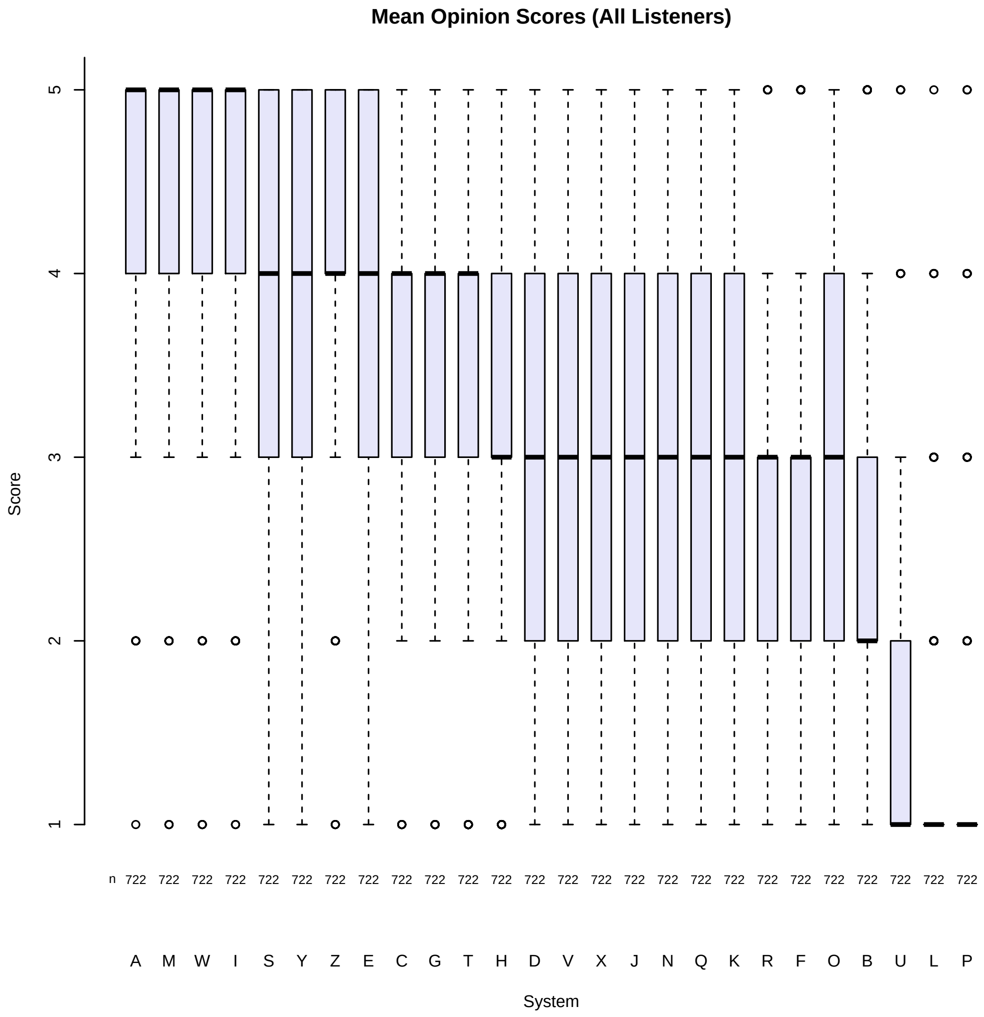

Figure 2: *Naturalness as judged by all listeners.*

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
A		■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
B	■		■	■	■		■	■	■	■	■	■	■	■		■	■	■	■	■	■	■	■	■	■	■
C	■	■		■		■		■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
D	■	■	■		■	■	■		■		■	■	■	■	■	■	■	■	■	■	■		■		■	■
E	■	■		■		■	■	■	■	■	■	■	■	■	■	■	■	■		■	■	■	■	■		
F	■		■	■	■		■	■	■	■		■	■	■		■			■	■	■	■	■	■	■	■
G	■	■		■	■	■		■	■	■	■	■	■	■	■	■	■	■	■		■	■	■	■	■	■
H	■	■	■		■	■	■		■	■	■	■	■	■	■	■	■	■	■		■	■	■	■	■	■
I	■	■	■	■	■	■	■	■		■	■	■	■	■	■	■	■	■	■	■	■	■		■	■	■
J	■	■	■		■	■	■	■	■		■	■	■		■	■	■	■	■	■	■		■		■	■
K	■	■	■	■	■		■	■	■	■		■	■			■			■	■	■	■	■	■	■	■
L	■	■	■	■	■	■	■	■	■	■	■		■	■	■		■	■	■	■		■	■	■	■	■
M	■	■	■	■	■	■	■	■	■	■	■	■		■	■	■	■	■	■	■	■	■	■	■	■	■
N	■	■	■	■	■	■	■	■	■			■	■		■	■		■	■	■	■	■	■		■	■
O	■		■	■	■		■	■	■	■		■	■	■		■			■	■	■	■	■	■	■	■
P	■	■	■	■	■	■	■	■	■	■	■		■	■	■		■	■	■	■	■	■	■	■	■	■
Q	■	■	■	■	■		■	■	■	■		■	■			■			■	■	■	■	■	■	■	■
R	■	■	■	■	■		■	■	■	■		■	■	■		■			■	■	■	■	■	■	■	■
S	■	■	■	■		■	■	■	■	■	■	■	■	■	■	■	■	■		■	■	■	■	■		
T	■	■	■	■	■	■			■	■	■	■	■	■	■	■	■	■	■		■	■	■	■	■	■
U	■	■	■	■	■	■	■	■	■	■	■		■	■	■	■	■	■	■	■		■	■	■	■	■
V	■	■	■		■	■	■	■	■		■	■	■	■	■	■	■	■	■	■	■		■		■	■
W	■	■	■	■	■	■	■	■		■	■	■	■	■	■	■	■	■	■	■	■	■		■	■	■
X	■	■	■		■	■	■	■	■		■	■	■		■	■	■	■	■	■	■		■		■	■
Y	■	■	■	■		■	■	■	■	■	■	■	■	■	■	■	■	■		■	■	■	■	■		
Z	■	■	■	■		■	■	■	■	■	■	■	■	■	■	■	■	■		■	■	■	■	■		

Figure 3
Significant differences in naturalness by all listeners between systems are indicated by solid black boxes.

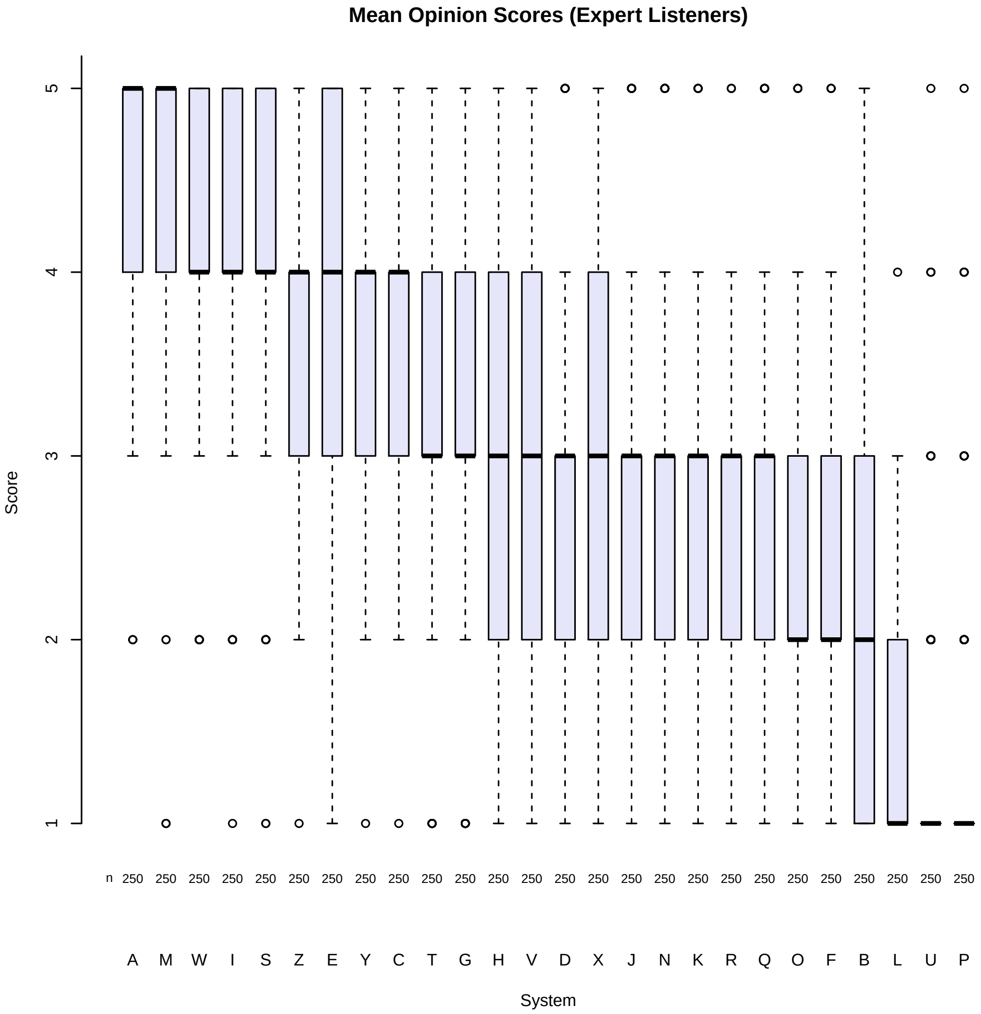

Figure 4: *Naturalness as judged by "speech expert" listeners.*

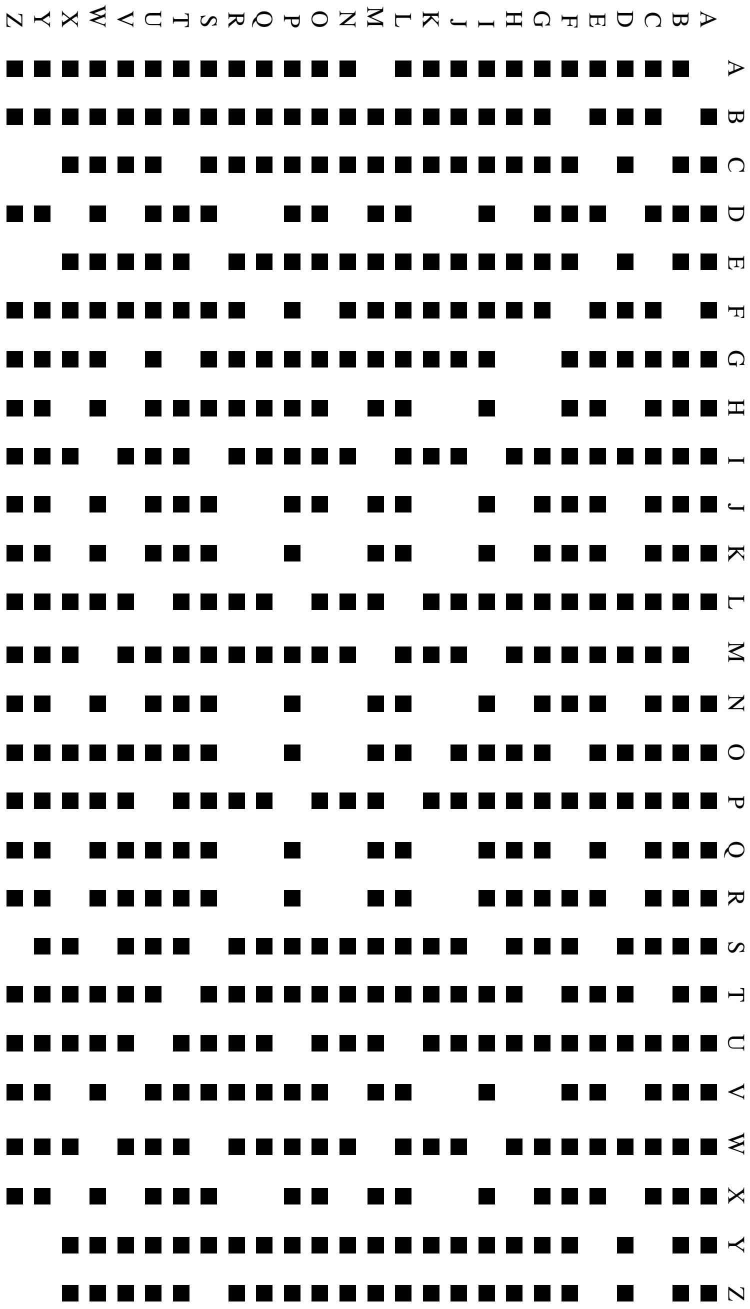

Figure 5
Significant differences in naturalness by expert listeners between systems are indicated by solid black boxes.

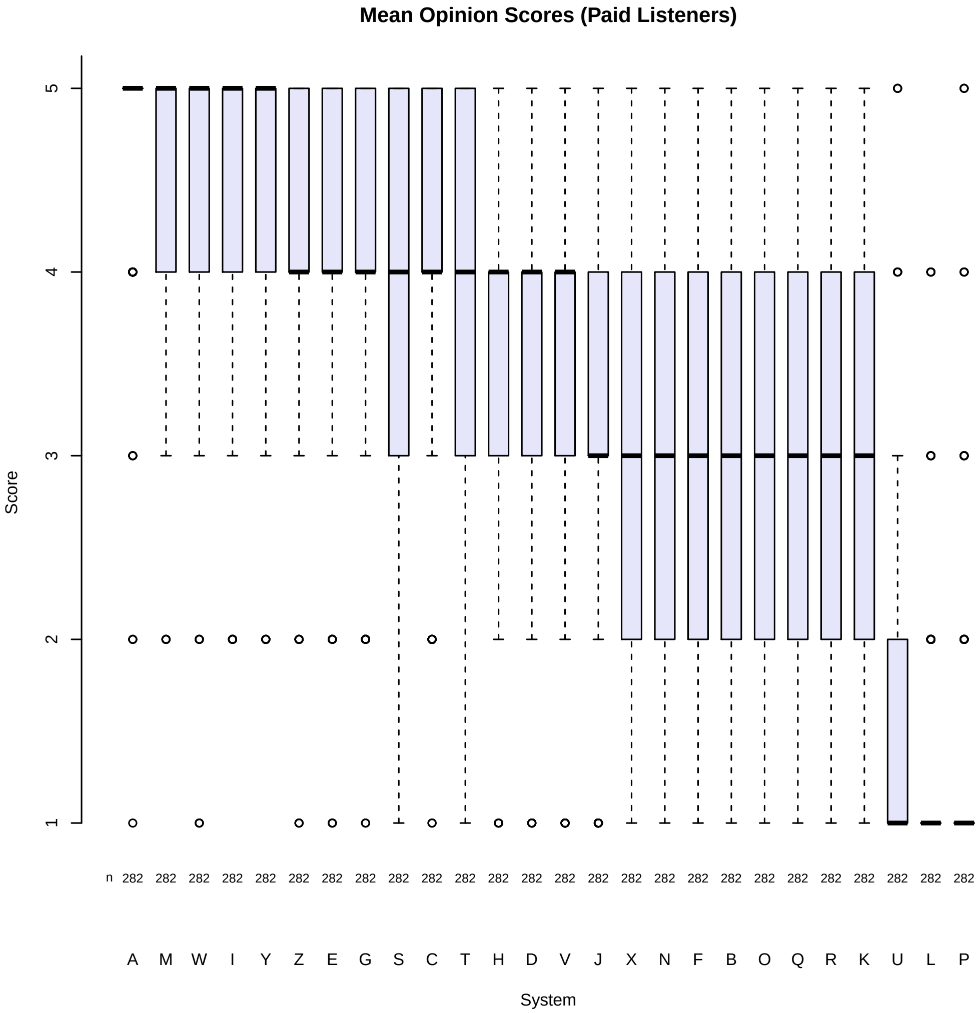

Figure 6: *Naturalness by paid listeners.*

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
A		■	■	■	■	■	■	■	■	■	■	■		■	■	■	■	■	■	■	■	■	■	■	■	■
B	■		■	■	■		■	■	■			■	■			■			■	■	■	■	■		■	■
C	■	■		■		■		■	■	■	■	■	■	■	■	■	■	■			■	■	■	■	■	
D	■	■	■		■	■	■		■		■	■	■		■	■	■	■	■		■		■		■	■
E	■	■		■		■		■		■	■	■	■	■	■	■	■	■		■	■	■	■	■		
F	■		■	■	■		■	■	■			■	■			■			■	■	■	■	■		■	■
G	■	■		■		■		■	■	■	■	■	■	■	■	■	■	■			■	■	■	■		
H	■	■	■		■	■	■		■	■	■	■	■	■	■	■	■	■			■		■	■	■	■
I	■	■	■	■		■	■	■		■	■	■	■	■	■	■	■	■	■	■	■	■		■		
J	■		■		■		■	■	■		■	■	■			■		■	■	■	■		■		■	■
K	■		■	■	■		■	■	■	■		■	■			■			■	■	■	■	■	■	■	■
L	■	■	■	■	■	■	■	■	■	■	■		■	■	■		■	■	■	■	■	■	■	■	■	■
M		■	■	■	■	■	■	■	■	■	■	■		■	■	■	■	■	■	■	■	■		■	■	■
N	■		■		■		■	■	■			■	■			■			■	■	■		■		■	■
O	■		■	■	■		■	■	■			■	■			■			■	■	■	■	■		■	■
P	■	■	■	■	■	■	■	■	■	■	■		■	■	■		■	■	■	■	■	■	■	■	■	■
Q	■		■	■	■		■	■	■			■	■			■			■	■	■	■	■		■	■
R	■		■	■	■		■	■	■	■		■	■			■			■	■	■	■	■	■	■	■
S	■	■		■		■			■	■	■	■	■	■	■	■	■	■			■	■	■	■	■	
T	■	■			■	■			■	■	■	■	■	■	■	■	■	■			■		■	■	■	■
U	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■		■	■	■	■	■
V	■	■	■		■	■	■		■		■	■	■		■	■	■	■	■		■		■		■	■
W	■	■	■	■	■	■	■	■		■	■	■		■	■	■	■	■	■	■	■	■		■		
X	■		■		■		■	■	■		■	■	■			■		■	■	■	■		■		■	■
Y	■	■	■	■		■		■		■	■	■	■	■	■	■	■	■	■	■	■	■		■		
Z	■	■		■		■		■		■	■	■	■	■	■	■	■	■		■	■	■		■		

Figure 7
Significant differences in naturalness by paid listeners between systems are indicated by solid black boxes.

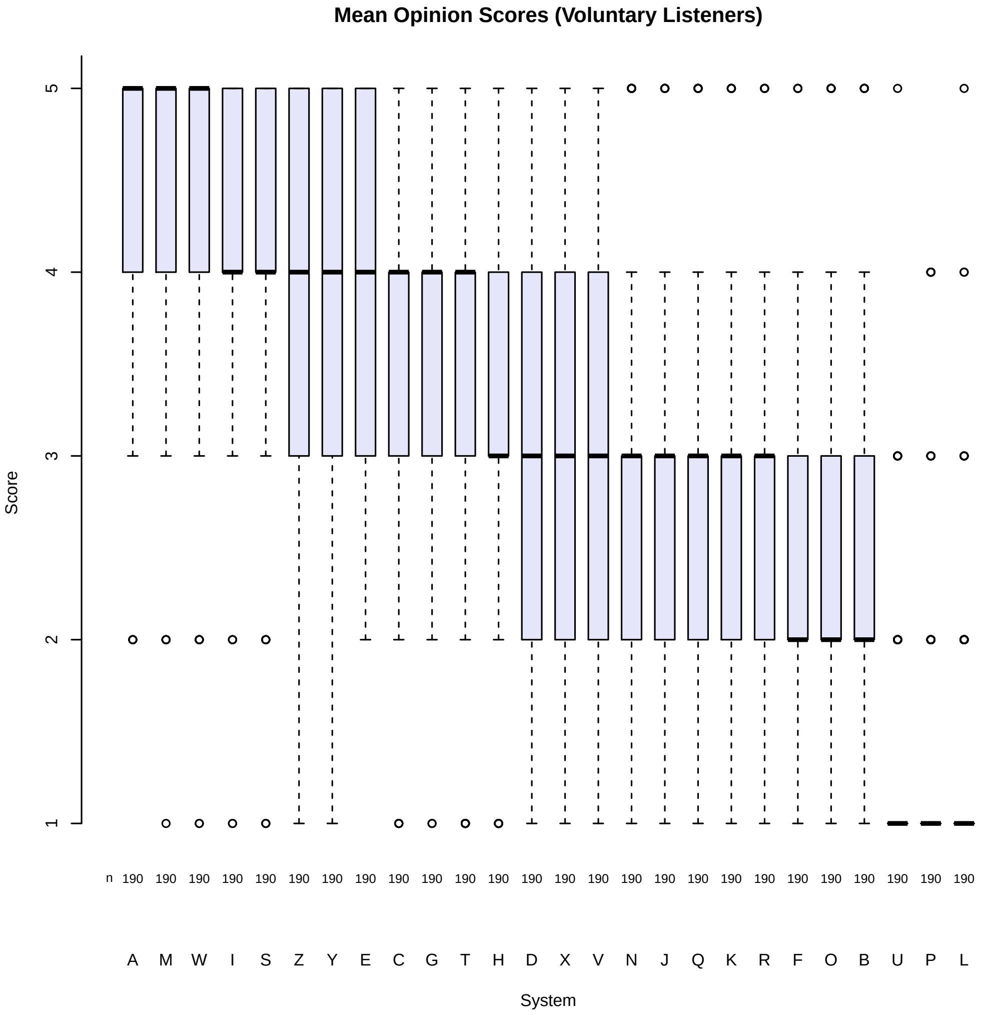

Figure 8: *Naturalness by voluntary listeners.*

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
A		■	■	■	■	■	■	■		■	■	■		■	■	■	■	■	■	■	■	■		■	■	■
B	■		■	■	■		■	■	■	■	■	■	■	■		■	■		■	■	■	■	■	■	■	■
C	■	■		■		■			■	■	■	■	■	■	■	■	■	■	■		■	■	■	■		
D	■	■	■		■	■	■		■		■	■	■		■	■		■	■		■		■		■	■
E	■	■		■		■		■	■	■	■	■	■	■	■	■	■	■		■	■	■	■	■		
F	■		■	■	■		■	■	■			■	■			■			■	■	■	■	■	■	■	■
G	■	■		■		■			■	■	■	■	■	■	■	■	■	■	■		■	■	■	■	■	■
H	■	■			■	■			■	■	■	■	■	■	■	■	■	■	■		■	■	■		■	■
I		■	■	■	■	■	■	■		■	■	■		■	■	■	■	■		■	■	■		■	■	■
J	■	■	■		■		■	■	■			■	■		■	■			■	■	■		■		■	■
K	■	■	■	■	■		■	■	■			■	■			■			■	■	■		■		■	■
L	■	■	■	■	■	■	■	■	■	■	■		■	■	■		■	■	■	■		■	■	■	■	■
M		■	■	■	■	■	■	■		■	■	■		■	■	■	■	■	■	■	■	■		■	■	■
N	■	■	■		■		■	■	■			■	■		■	■			■	■	■		■		■	■
O	■		■	■	■		■	■	■	■		■	■	■		■			■	■	■	■	■	■	■	■
P	■	■	■	■	■	■	■	■	■	■	■		■	■	■		■	■	■	■		■	■	■	■	■
Q	■	■	■		■		■	■	■			■	■			■			■	■	■		■		■	■
R	■		■	■	■		■	■	■			■	■			■			■	■	■		■		■	■
S	■	■	■	■		■	■	■		■	■	■	■	■	■	■	■	■		■	■	■		■		
T	■	■			■	■			■	■	■	■	■	■	■	■	■	■	■		■	■	■	■	■	■
U	■	■	■	■	■	■	■	■	■	■	■		■	■	■		■	■	■	■		■	■	■	■	■
V	■	■	■		■	■	■	■	■			■	■		■	■			■	■	■		■		■	■
W		■	■	■	■	■	■	■		■	■	■		■	■	■	■	■		■	■	■		■	■	■
X	■	■	■		■	■	■		■			■	■		■	■			■	■	■		■		■	■
Y	■	■		■		■	■	■	■	■	■	■	■	■	■	■	■	■		■	■	■	■	■		
Z	■	■		■		■	■	■	■	■	■	■	■	■	■	■	■	■		■	■	■	■	■		

Figure 9
Significant differences in naturalness by voluntary listeners between systems are indicated by solid black boxes.

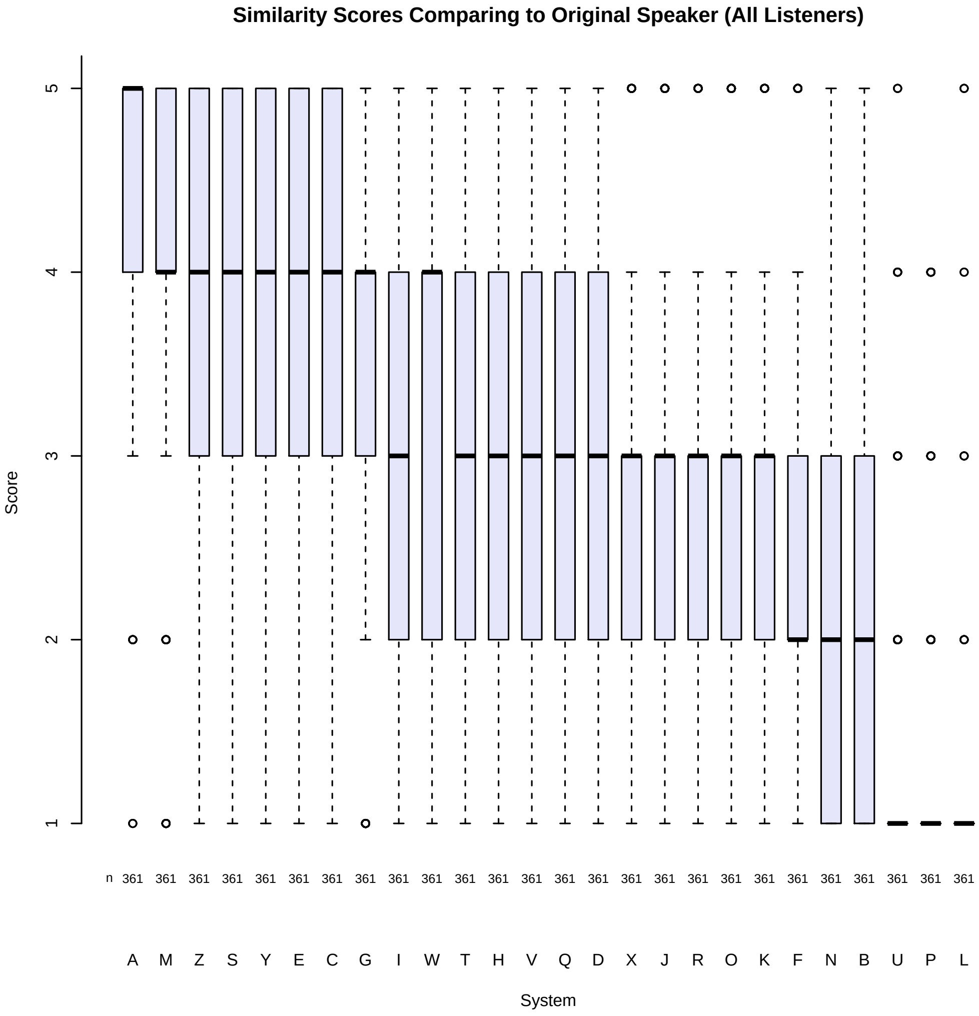

Figure 10: *Similarity to original speaker by all listeners.*

Z	Y	X	W	V	U	T	S	R	Q	P	O	N	M	L	K	J	I	H	G	F	E	D	C	B	A	
■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■		A
■	■	■	■	■	■	■	■	■	■	■	■		■	■	■	■	■	■	■	■	■	■	■		■	B
		■		■	■	■		■	■	■	■	■	■	■	■	■		■		■		■		■	■	C
■	■		■		■		■			■		■	■	■			■		■	■	■		■	■	■	D
		■	■	■	■	■		■	■	■	■	■	■	■	■	■	■	■	■	■		■		■	■	E
■	■		■	■	■	■	■		■	■			■	■			■	■	■		■	■	■	■	■	F
■	■	■		■	■		■	■	■	■	■	■	■	■	■	■		■		■	■	■		■	■	G
■	■	■			■		■	■		■	■	■	■	■	■	■			■	■	■		■	■	■	H
■	■	■		■	■		■	■	■	■	■	■	■	■	■	■				■	■	■		■	■	I
■	■		■		■	■	■			■		■	■	■			■	■	■		■		■	■	■	J
■	■		■	■	■	■	■		■	■		■	■	■			■	■	■		■		■	■	■	K
■	■	■	■	■	■	■	■	■	■	■	■	■	■		■	■	■	■	■	■	■	■	■	■	■	L
	■	■	■	■	■	■		■	■	■	■	■		■	■	■	■	■	■	■	■	■	■	■	■	M
■	■	■	■	■	■	■	■	■	■	■	■		■	■	■	■	■	■	■		■	■	■		■	N
■	■		■	■	■	■	■			■		■	■	■			■	■	■		■		■	■	■	O
■	■	■	■	■		■	■	■	■		■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	P
■	■		■		■		■			■		■	■	■	■		■		■	■	■		■	■	■	Q
■	■		■		■	■	■			■		■	■	■			■	■	■		■		■	■	■	R
		■	■	■	■	■		■	■	■	■	■		■	■	■	■	■	■	■		■		■	■	S
■	■	■			■		■	■		■	■	■	■	■	■	■				■	■		■	■	■	T
■	■	■	■	■		■	■	■	■		■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	U
■	■				■		■			■	■	■	■	■	■		■		■	■	■		■	■	■	V
■	■	■			■		■	■	■	■	■	■	■	■	■	■				■	■	■		■	■	W
■	■		■		■	■	■			■		■	■	■			■	■	■		■		■	■	■	X
		■	■	■	■	■		■	■	■	■	■	■	■	■	■	■	■	■	■		■		■	■	Y
		■	■	■	■	■		■	■	■	■	■		■	■	■	■	■	■	■		■		■	■	Z

Figure 11
Significant differences in similarity comparing to original speaker by all listeners between systems are indicated by solid black boxes.

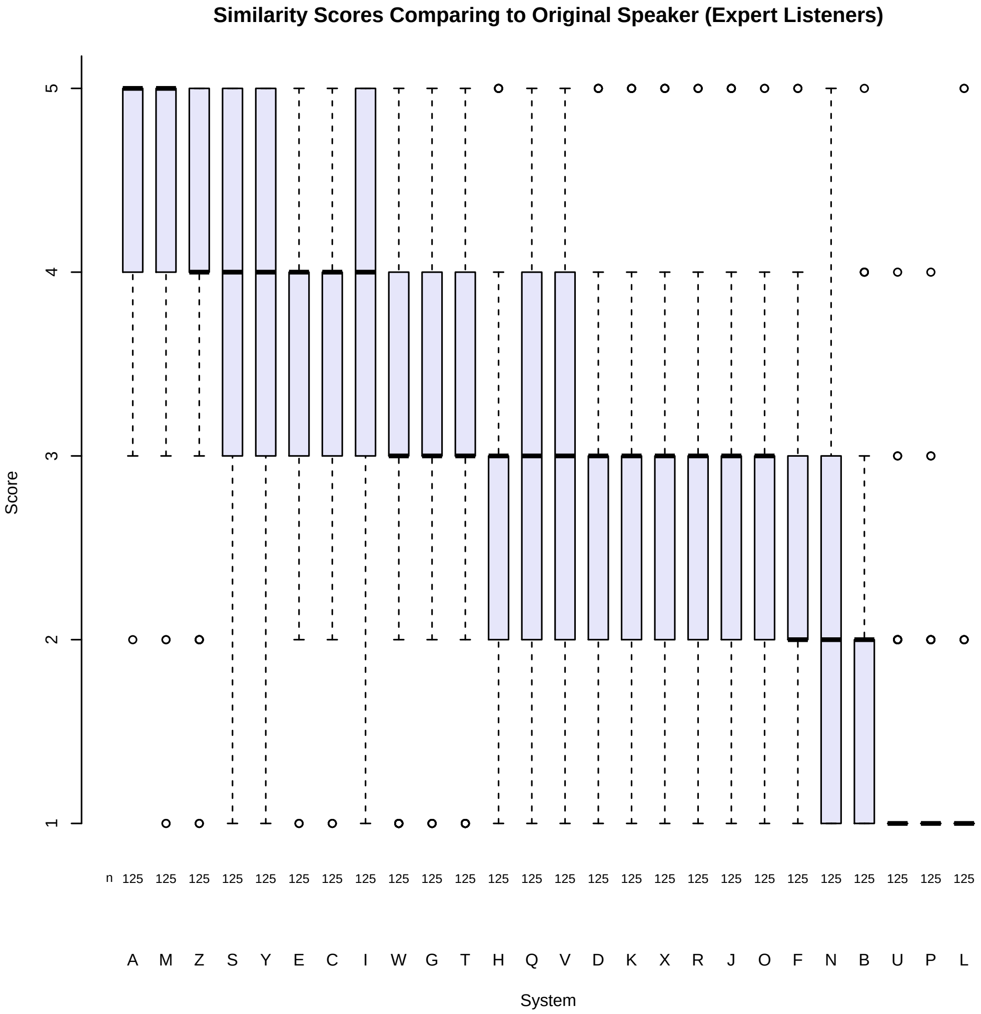

Figure 12: *Similarity to original speaker by expert listeners.*

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
A		■	■	■	■	■	■	■	■	■	■	■		■	■	■	■	■	■	■	■	■	■	■	■	■
B	■		■	■	■	■	■	■	■	■	■	■	■		■	■	■	■	■	■	■	■	■	■	■	■
C	■	■		■		■		■		■	■	■	■	■	■	■	■	■			■	■		■		
D	■	■	■		■				■			■	■	■		■			■		■		■		■	■
E	■	■		■		■		■		■	■	■	■	■	■	■	■	■			■	■		■		
F	■	■	■		■		■	■	■			■	■			■	■		■	■	■	■	■		■	■
G	■	■				■				■	■	■	■	■	■	■		■	■		■			■		■
H	■	■	■		■	■			■			■	■	■		■			■		■				■	■
I	■	■		■		■		■		■	■	■	■	■	■	■	■	■			■	■		■		
J	■	■	■		■		■		■			■	■			■			■	■	■		■		■	■
K	■	■	■		■		■		■			■	■	■		■			■		■				■	■
L	■	■	■	■	■	■	■	■	■	■	■		■	■	■		■	■	■	■		■	■	■	■	■
M		■	■	■	■	■	■	■	■	■	■	■		■	■	■	■	■	■	■	■	■	■	■	■	
N	■		■	■	■		■	■	■		■	■	■			■	■	■	■	■	■	■	■	■	■	■
O	■	■	■		■		■		■			■	■			■			■		■		■		■	■
P	■	■	■	■	■	■	■	■	■	■	■		■	■	■		■	■	■	■		■	■	■	■	■
Q	■	■	■		■	■			■			■	■	■		■			■		■				■	■
R	■	■	■		■		■		■			■	■	■		■			■	■	■		■		■	■
S	■	■		■		■	■	■		■	■	■	■	■	■	■	■	■		■	■	■		■		
T	■	■				■				■		■	■	■	■	■		■	■		■				■	■
U	■	■	■	■	■	■	■	■	■	■	■		■	■	■		■	■	■	■		■	■	■	■	■
V	■	■	■		■	■			■			■	■	■		■			■		■				■	■
W	■	■		■		■				■		■	■	■	■	■		■			■			■		■
X	■	■	■		■		■		■			■	■	■		■			■		■		■		■	■
Y	■	■		■		■		■		■	■	■	■	■	■	■	■	■		■	■	■		■		
Z	■	■		■		■	■	■		■	■	■		■	■	■	■	■		■	■	■	■	■		

Figure 13
Significant differences in similarity comparing to original speaker by expert listeners between systems are indicated by solid black boxes.

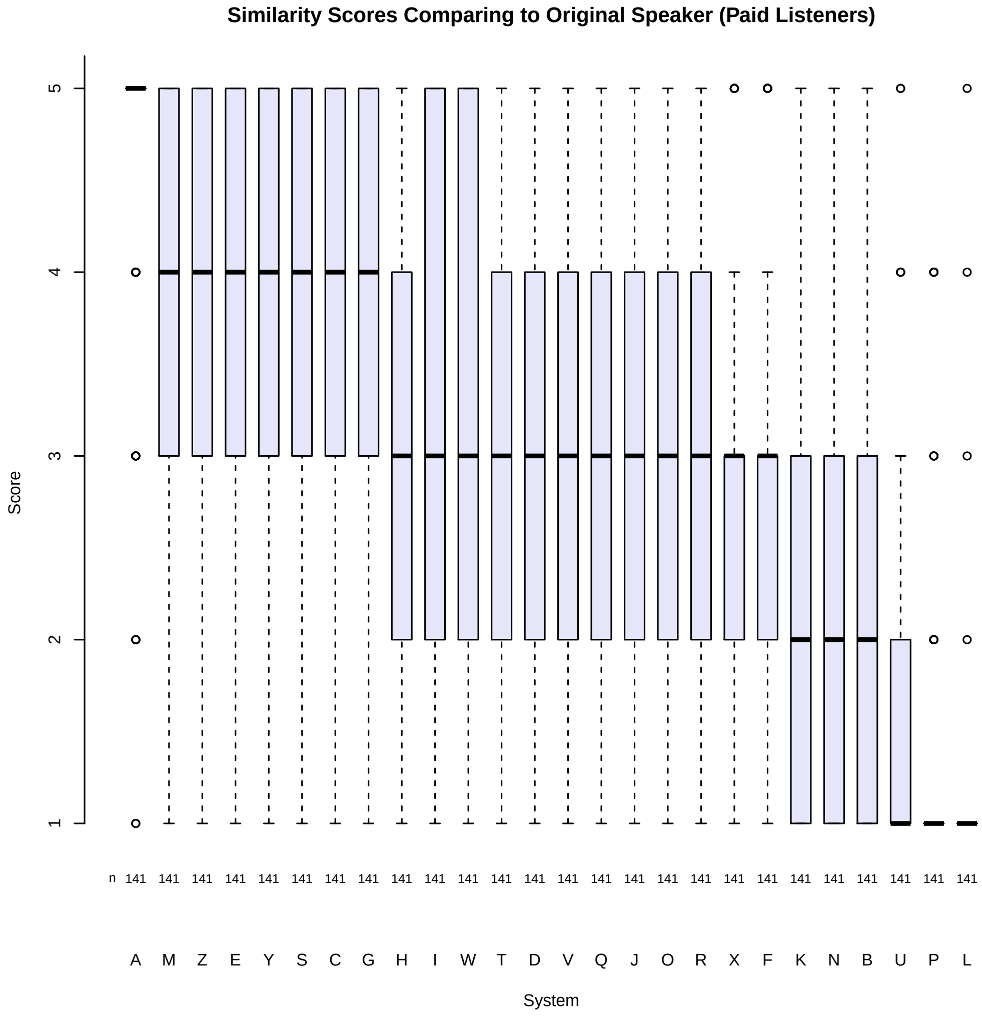

Figure 14: *Similarity to original speaker by paid listeners.*

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
A		■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
B	■		■	■	■		■	■	■	■		■	■		■	■	■		■	■	■	■	■		■	■
C	■	■		■		■				■	■	■		■	■	■	■	■			■	■		■		
D	■	■	■		■						■	■	■	■		■			■		■				■	■
E	■	■		■		■		■	■	■	■	■		■	■	■	■	■		■	■	■	■	■		
F	■		■		■		■	■				■	■			■			■		■				■	■
G	■	■				■				■	■	■		■	■	■	■	■			■			■		
H	■	■			■	■					■	■	■	■		■			■		■				■	■
I	■	■			■						■	■	■	■		■					■				■	■
J	■	■	■		■		■					■	■			■			■		■				■	■
K	■		■	■	■		■	■	■			■	■			■			■	■	■	■	■		■	■
L	■	■	■	■	■	■	■	■	■	■	■		■	■	■		■	■	■	■	■	■	■	■	■	■
M	■	■		■		■		■	■	■	■	■		■	■	■	■	■		■	■	■	■	■		
N	■		■	■	■		■	■	■			■	■			■			■	■	■	■	■		■	■
O	■	■	■		■		■					■	■			■			■		■				■	■
P	■	■	■	■	■	■	■	■	■	■	■		■	■	■		■	■	■	■		■	■	■	■	■
Q	■	■	■		■		■					■	■			■			■		■				■	■
R	■		■		■		■					■	■			■			■		■				■	■
S	■	■		■		■		■		■	■	■		■	■	■	■	■		■	■	■	■	■		
T	■	■			■						■	■	■	■		■			■		■				■	■
U	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■		■	■	■	■		■	■	■	■	■
V	■	■	■		■						■	■	■	■		■			■		■				■	■
W	■	■			■						■	■	■	■		■			■		■				■	■
X	■		■		■		■					■	■			■			■		■				■	■
Y	■	■		■		■		■	■	■	■	■		■	■	■	■	■		■	■	■	■	■		
Z	■	■		■		■		■	■	■	■	■		■	■	■	■	■		■	■	■	■	■		

Figure 15
Significant differences in similarity comparing to original speaker by paid listeners between systems are indicated by solid black boxes.

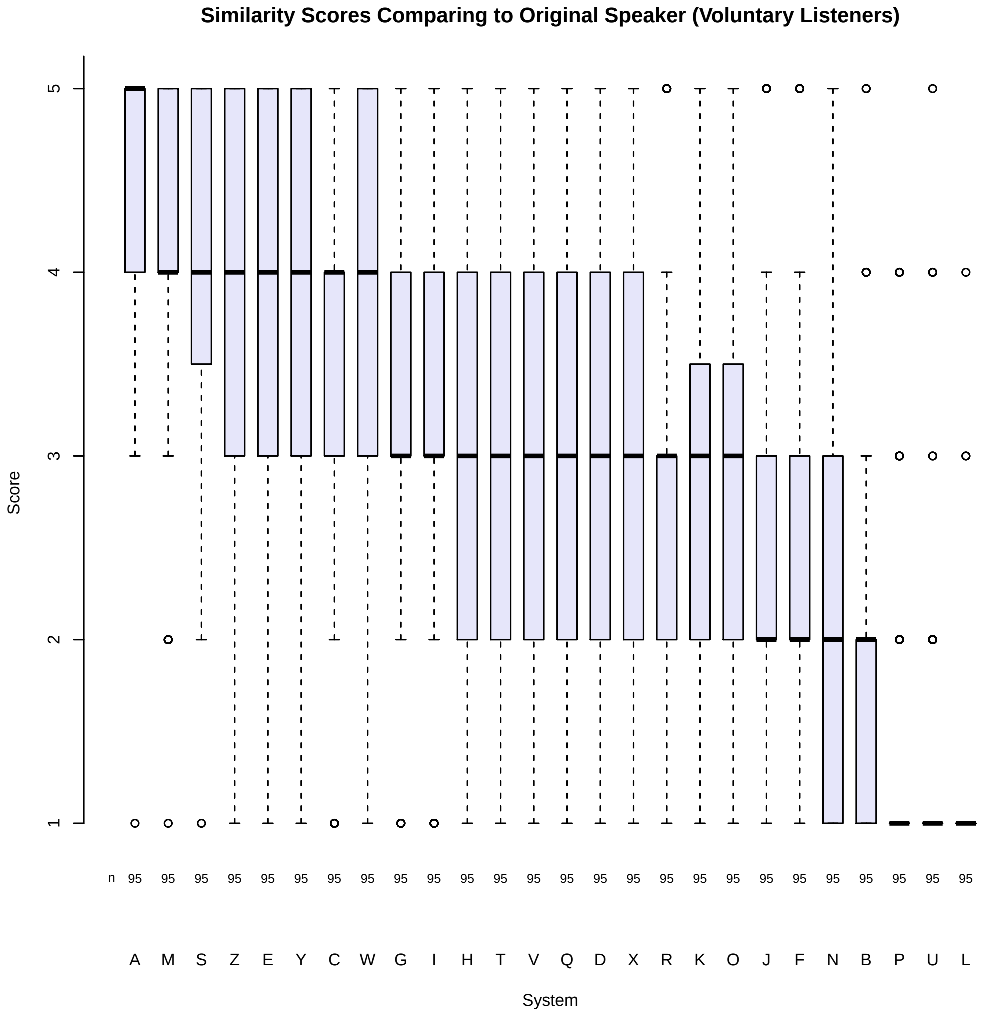

Figure 16: *Similarity to original speaker by voluntary listeners.*

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
A		■	■	■	■	■	■	■	■	■	■	■		■	■	■	■	■		■	■	■	■	■	■	■
B	■		■	■	■		■	■	■	■	■	■	■		■	■	■	■	■	■	■	■	■	■	■	■
C	■	■		■		■				■	■	■		■	■	■	■	■			■	■		■		
D	■	■	■		■							■	■			■			■		■				■	■
E	■	■		■		■		■		■	■	■		■	■	■	■	■			■	■		■		
F	■		■		■		■		■			■	■			■			■		■		■		■	■
G	■	■				■				■		■	■	■	■	■		■			■					
H	■	■			■							■	■	■		■			■		■					■
I	■	■				■				■		■	■	■		■					■					
J	■	■	■		■		■		■			■	■			■			■		■		■		■	■
K	■	■	■		■							■	■			■			■		■				■	■
L	■	■	■	■	■	■	■	■	■	■	■		■	■	■		■	■	■	■		■	■	■	■	■
M		■		■		■	■	■	■	■	■	■		■	■	■	■	■		■	■	■		■		
N	■		■		■		■	■	■			■	■			■			■	■	■	■	■		■	■
O	■	■	■		■		■					■	■			■			■		■		■		■	■
P	■	■	■	■	■	■	■	■	■	■	■		■	■	■		■	■	■	■		■	■	■	■	■
Q	■	■	■		■							■	■			■			■		■				■	■
R	■	■	■		■		■					■	■			■			■		■				■	■
S		■		■		■		■		■	■	■		■	■	■	■	■		■	■	■		■		
T	■	■										■	■	■		■			■		■					■
U	■	■	■	■	■	■	■	■	■	■	■		■	■	■		■	■	■	■		■	■	■	■	■
V	■	■	■		■							■	■	■		■			■		■				■	■
W	■	■				■				■		■		■	■	■					■					
X	■	■	■		■							■	■			■			■		■				■	■
Y	■	■		■		■				■	■	■		■	■	■	■	■			■	■		■		
Z	■	■		■		■		■		■	■	■		■	■	■	■	■		■	■	■		■		

Figure 17
Significant differences in similarity comparing to original speaker by voluntary listeners between systems are indicated by solid black boxes.

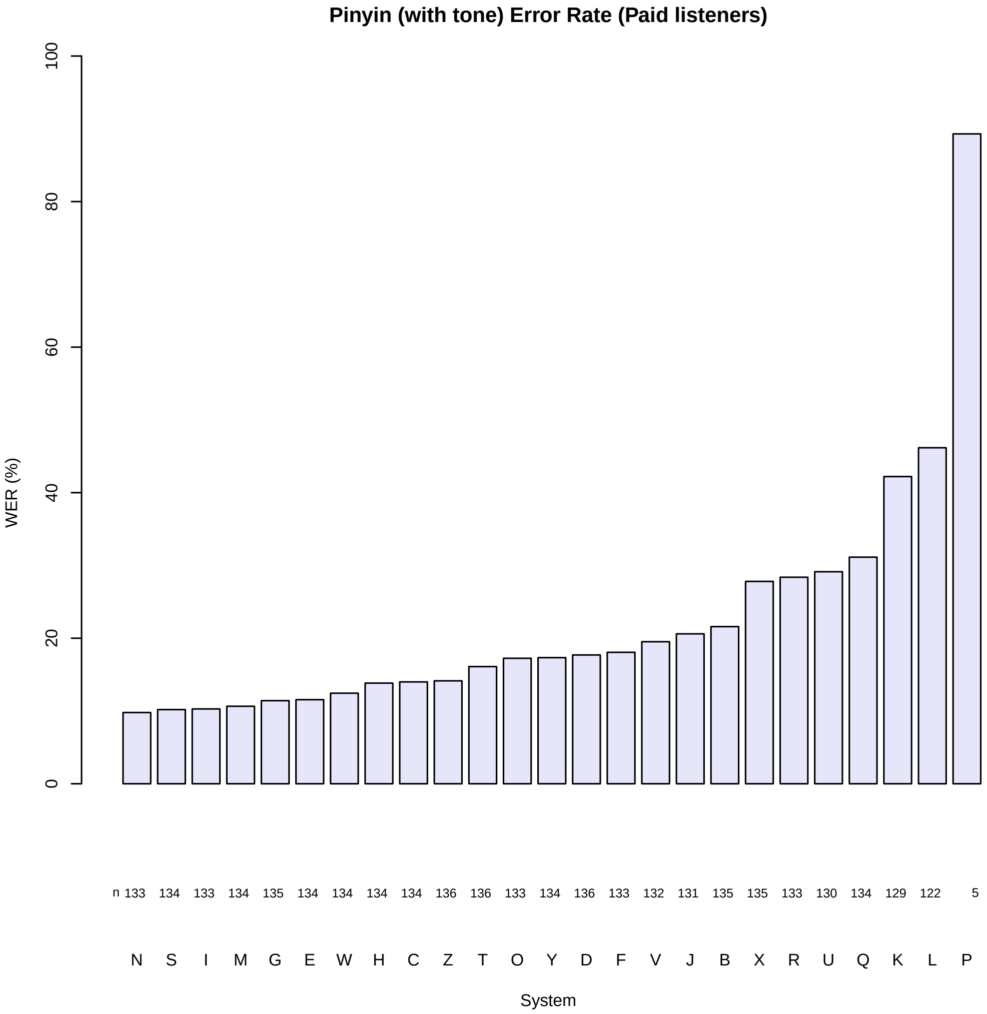

Figure 18: *Intelligibility in Pinyin (with tones) error rate by paid listeners.*

	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
B				■		■		■		■	■	■	■			■	■	■				■			■
C										■	■					■	■			■			■		
D						■		■		■	■	■	■			■	■	■		■			■		
E	■				■				■	■	■					■	■			■	■		■	■	
F				■		■		■		■	■	■	■			■	■	■		■		■	■		
G	■		■		■				■	■	■					■	■			■	■		■	■	
H									■	■	■					■	■			■			■		
I	■		■		■				■	■	■					■	■			■	■		■	■	
J				■		■	■	■		■	■	■	■			■		■				■			
K	■	■	■	■	■	■	■	■	■			■	■	■		■	■	■	■	■	■	■	■	■	■
L	■	■	■	■	■	■	■	■	■			■	■	■		■	■	■	■	■	■	■	■	■	■
M	■		■		■				■	■	■			■		■	■			■	■		■	■	
N	■		■		■				■	■	■					■	■			■	■		■	■	
O										■	■	■				■	■	■		■			■		
P																									
Q	■	■	■	■	■	■	■	■	■	■	■	■	■	■				■	■		■	■		■	■
R	■	■	■	■	■	■	■	■		■	■	■	■	■				■	■			■		■	■
S	■		■		■				■	■	■			■		■	■			■	■		■	■	
T										■	■					■	■			■			■		
U		■	■	■	■	■	■	■		■	■	■	■	■				■	■		■	■		■	■
V				■		■		■		■	■	■	■			■		■		■			■		
W	■				■				■	■	■					■	■			■			■		
X		■	■	■	■	■	■	■		■	■	■	■	■				■	■		■	■		■	■
Y				■		■		■		■	■	■	■			■	■	■		■			■		
Z	■									■	■					■	■			■			■		

Figure 19
Significant differences in intelligibility by paid listeners between systems in terms of Chinese Pinyin (with tones) are indicated by solid black boxes.

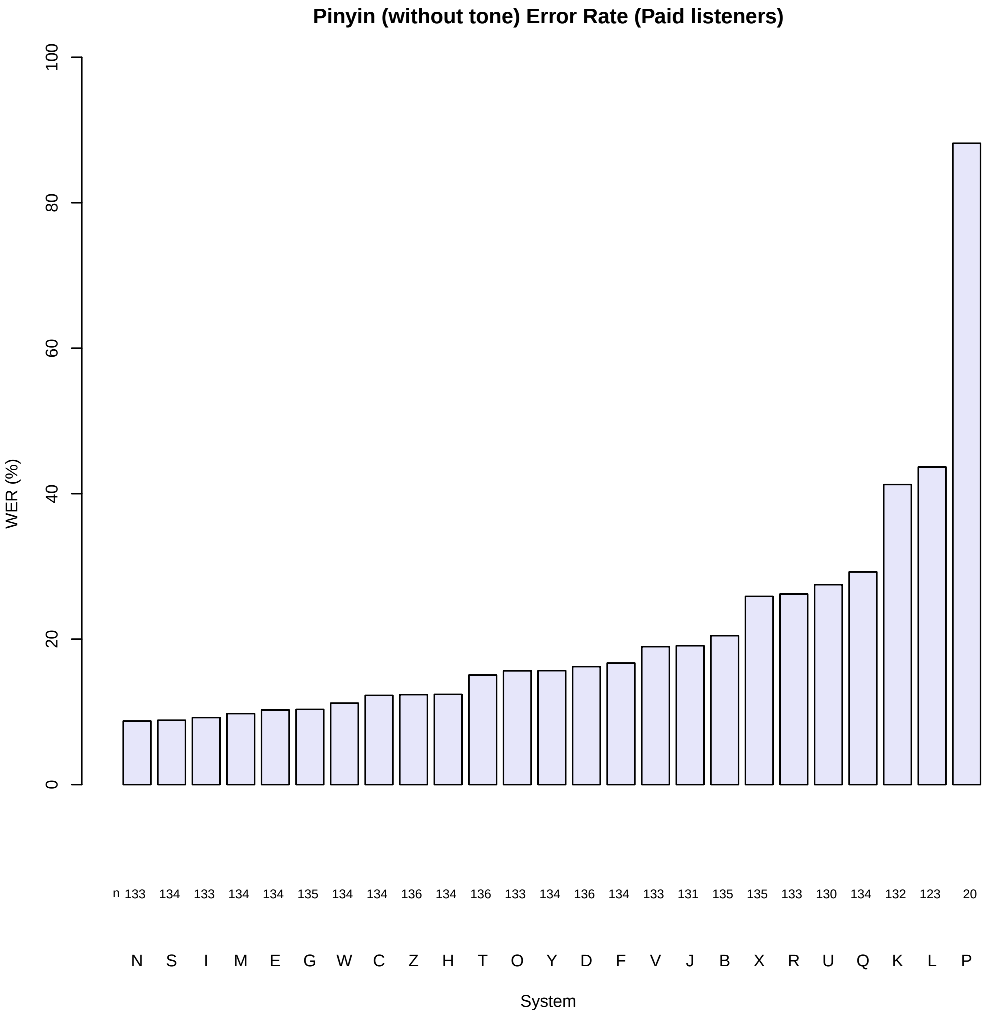

Figure 20: *Intelligibility in Pinyin (without tones) error rate by paid listeners.*

	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
B				■		■	■	■		■	■	■	■			■	■	■				■			■
C									■	■	■					■	■			■			■		
D						■		■		■	■	■	■			■	■	■		■			■		
E	■				■				■	■	■					■	■			■	■		■	■	
F				■		■		■		■	■	■	■			■	■	■		■			■		
G	■		■		■				■	■	■					■	■			■	■		■	■	
H	■								■	■	■					■	■			■			■		
I	■		■		■				■	■	■					■	■			■	■		■	■	
J		■		■		■	■	■		■	■	■	■			■		■				■			■
K	■	■	■	■	■	■	■	■	■			■	■	■		■	■	■	■	■	■	■	■	■	■
L	■	■	■	■	■	■	■	■	■			■	■	■	■	■	■	■	■	■	■	■	■	■	■
M	■		■		■				■	■	■			■		■	■			■	■		■	■	
N	■		■		■				■	■	■					■	■			■	■		■	■	
O										■	■	■				■	■	■		■			■		
P											■									■			■		
Q	■	■	■	■	■	■	■	■	■	■	■	■	■	■				■	■		■	■		■	■
R	■	■	■	■	■	■	■	■		■	■	■	■	■				■	■			■		■	■
S	■		■		■				■	■	■			■		■	■			■	■		■	■	
T										■	■					■	■			■			■		
U		■	■	■	■	■	■	■		■	■	■	■	■	■			■	■			■		■	■
V				■		■		■		■	■	■	■			■		■							
W	■								■	■	■					■	■			■			■		
X		■	■	■	■	■	■	■		■	■	■	■	■	■			■	■			■		■	■
Y				■		■		■		■	■	■	■			■	■	■		■			■		
Z	■								■	■	■					■	■			■			■		

Figure 21
Significant differences in intelligibility by paid listeners between systems in terms of Chinese Pinyin (without tones) are indicated by solid black boxes.