Analysis of the Voice Conversion Challenge 2016 Evaluation Results

Mirjam Wester[1], *Zhizheng Wu*[1], *Junichi Yamagishi*[1,2]

[1]The Centre for Speech Technology Research, The University of Edinburgh, UK
[2]National Institute of Informatics, Japan

mwester@inf.ed.ac.uk, zhizheng.wu@ed.ac.uk, jyamagis@inf.ed.ac.uk

Abstract

The Voice Conversion Challenge 2016 is the first Voice Conversion Challenge in which different voice conversion systems and approaches using the same voice data were compared. This paper describes the design of the evaluation, it presents the results and statistical analyses of the results.

Index Terms: Voice Conversion Challenge, evaluation

1. Introduction

The Voice Conversion Challenge (VCC) 2016, one of the special sessions at Interspeech 2016, deals with the well-known task of speaker identity conversion, referred as Voice Conversion (VC). The objective of the VCC is to compare various VC techniques on identical training and evaluation speech data. The full description of VCC 2016, the motivation, the database, the rules, the participants and main findings are presented in [1]. In the current paper, we describe the listening test design in more detail, we present the results of the listening test and the subsequent statistical analyses.

2. Evaluation

Voice converted voices were evaluated in terms of naturalness and similarity. The questions we addressed were:

1. How natural does the voice converted voice sound?
2. How similar does the voice converted voice sound compared to the target speaker and to the source speaker?

2.1. Voice conversion data

Data from five source and five target speakers were provided to the 17 participants (details in [1]) who each created 25 voice converted (VC) voices. In addition to this, there was a baseline system created at CSTR, i.e., 25×18 systems = 450 voices. Instead of evaluating all 450 VC voices, we decided to reduce the number of source-target (ST) pairs from 25 to 16. One source female speaker was removed as her recordings sounded Lombardized and one of the male target voices was removed because he had a significantly slower speaking rate than the other male speakers. Table 1 shows the resulting source-target pairs for each gender condition.

Table 1: *Source (S) - target (T) pairs per gender condition.*

M-M	F-F	M-F	F-M
SM1 - TM2	SF1 - TF1	SM1 -TF1	SF1 - TM2
SM2 - TM1	SF2 - TF2	SM2 -TF2	SF2 - TM1
SM1 - TM1	SF1 - TF2	SM1 -TF2	SF1 - TM1
SM2 - TM2	SF2 - TF1	SM2 -TF1	SF2 - TM2

2.2. Naturalness

The Blizzard evaluation [2, 3] was taken as inspiration for the design of the VCC Challenge. However, note that the current evaluation is considerably more complicated than Blizzard as not only are there 18 different participants/systems, they each provide voices for 25 source-target (ST) pairs. In designing the naturalness evaluation, we had to make a number of decisions: i) What type of test? and ii) How many voices? One of the main constraints we had to consider was the number of samples a listener could judge in an hour.

In line with most previous work in the field, we decided to use MOS tests for naturalness (e.g, [4, 5, 6, 7, 8, 9]). Despite the shortcomings of MOS, we are not aware of any other evaluation technique that can be used to compare 18 different voices. The scale ranged from (1) totally unnatural to (5) completely natural. The subjects were instructed that the score should reflect their opinion of how natural or unnatural the sentence sounded.

If a listener were to judge all voices (18 participants * 16 ST pairs = 288, + 4 source + 4 target = 296 stimuli) it would take roughly 50 minutes. From experience, we know that listeners judge about 6 sentences per minute. However, as we wanted a single listener to judge both the naturalness and similarity of VC voices this would take too long. We decided against an alternative in which listeners would come in for multiple sessions because of the risk of listeners dropping out between tests.

Instead of asking each listener to judge all ST pairs we considered reducing the MOS test to contain one single ST pair. In terms of time this would be an excellent solution. However, each listener would then only encounter one gender condition and listeners needed to encounter the full range of gender conditions as ratings are context-sensitive, i.e., the other voices in a set influence the judgement of each sample [10]. Therefore, we came up with an intermediate solution in which each listener hears eight source-target (ST) pairs, two from each gender condition, to make the two sets as comparable as possible. Two versions of the experiment were made: Set 1 contains the top half of Table 1, above the dashed line, and Set 2 the bottom half. Each set was listened to by 100 subjects, which took roughly 25 minutes. The order of stimuli was random for each listener, with each sentence selected at random with replacement from the pool of 30 test sentences. (Although there were 54 sentences for each target speaker, sentences longer than 5 s or shorter than 2 s were removed for the listening tests.)

2.3. Similarity

Measuring speaker similarity in a meaningful way is obviously a key aspect of the evaluation of any type of voice conversion. Using mean opinion scores to evaluate similarity, although a widely-used technique, is not without problems: judging how similar voices are on a scale from 1 to 5 may not be all that

meaningful. Judging the similarity of one speaker compared to another speaker is a rather unusual task, it is not an element of a person's regular, everyday speech perception. Recognising speakers, however, is something we do all the time. Therefore, we felt the same/different paradigm, which arguably is measuring something more akin to speaker recognition, would be more appropriate for similarity judging. (See [11, 12, 13, 14, 15] for other studies that use the same/different paradigm).

Listeners were given pairs of stimuli and given the following overall instructions: *"Do you think these two samples could have been produced by the same speaker? Some of the samples may sound somewhat degraded/distorted. Please try to listen beyond the distortion and concentrate on identifying the voice. Are the two voices the same or different? You have the option to indicate how sure you are of your decision."* The scale for judging was: "Same: absolutely sure", "Same: not sure", "Different: not sure" and "Different: absolutely sure".

Not only were VC stimuli compared to the target speaker but also to the source speaker. Table 2 shows the trials and number of occurrences in one ST pair test set. The final column gives the "correct" answer. As the objective in voice conversion is to sound like the target the T-VC trials should be classed as "same" and the S-VC trials as "diff". Strictly speaking there is no real correct answer. Same/different trials were roughly balanced. Each listener was given three ST pairs to judge, one within-gender, one cross-gender and one at random ensuing all ST pairs were covered across listeners. As before the sentences were selected at random (never the same sentence within a trial) from the pool of 30 sentences with replacement and the order of the trials was random.

Table 2: *Stimuli for each ST pair for similarity judgement.*

trials	answer	#
S - S	same	1
T - T	same	1
S - T	diff	1
S - VC1, S - VC2, ..., S - VC18	"diff"	18
T - VC1, T - VC2, ..., T - VC18	"same"	18
		39

2.4. Listeners

Experiments were carried out using a web interface. The listeners were seated in sound isolated booths and listened to the samples using Beyerdynamic DT 770 PRO headphones. Listeners were remunerated for their time and effort.

After completing the experiment, listeners filled out a short questionnaire with questions regarding gender, native language, accent and whether they were speech experts or not. 200 (52 male and 148 female) subjects took part in the experiment. Table 3 gives a breakdown of the age categories and (self-reported) accents of subjects.

Table 3: *Age and accent of subjects.*

Age	#	Accent	#
18-20	39	British	124
20-29	146	North American	45
30-39	10	other	21
40-59	5	not given	10

3. Results

3.1. Naturalness

Figure 1 shows a boxplot of MOS values over all ST pairs for each system, ordered by the mean (red dots). The letters A ... Q indicate the 17 participants [1] and S = source, T = target and B_ = baseline. Figure 2 shows the separate results for Set 1 and Set 2. A one-way ANOVA revealed that the overall means for the two sets (different materials, and different listeners) are significantly different [$F(1, 31998) = 29.59, p < 10^{-8}$].

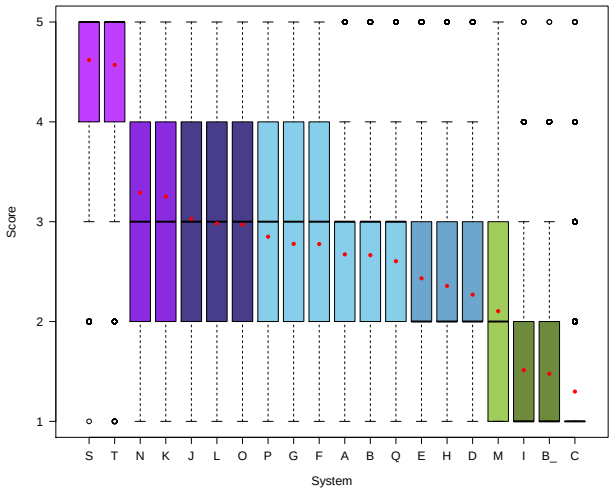

Figure 1: *Naturalness MOS for all ST pairs and all systems, ordered by mean (red dots). Black lines indicate medians.*

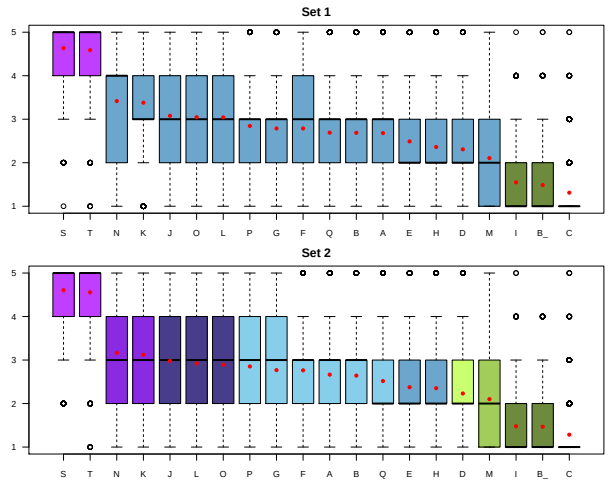

Figure 2: *Naturalness MOS for Sets 1 & 2 ST pairs, all systems.*

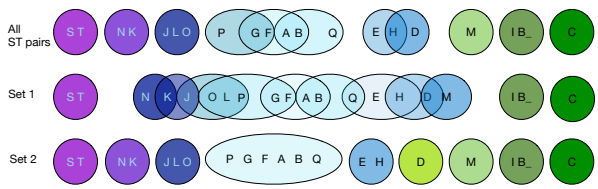

Figure 3: *Groupings of systems that do not differ significantly from each other in naturalness, for all ST pairs, Sets 1 & 2.*

Wilcoxon signed-rank tests with Bonferroni correction ($\alpha = 0.01$) confirm that all systems are rated significantly less

natural than the source (S) and target (T) speakers. Furthermore, most pairwise comparisons are significantly different from each other. However, in this case we are also interested in which systems are *not* significantly different from each other. This is roughly illustrated in Figures 1 & 2 by the colour of the boxes and graphically in Figure 3 by grouping the systems with the same naturalness scores together. This illustrates that although the ANOVA indicates a significant difference between sets 1 and 2, the rankings of the systems do not change much across the sets.

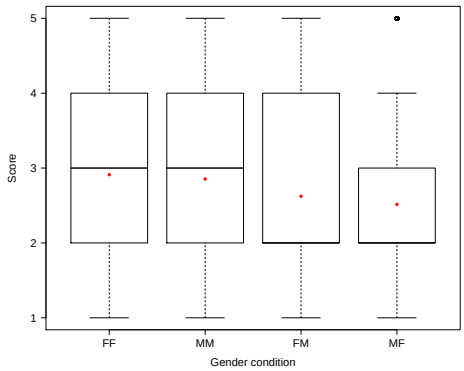

Figure 4: *Overall naturalness MOS per gender condition.*

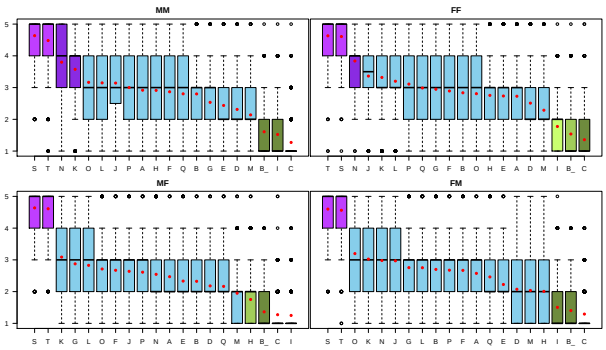

Figure 5: *Naturalness MOS per gender condition all systems.*

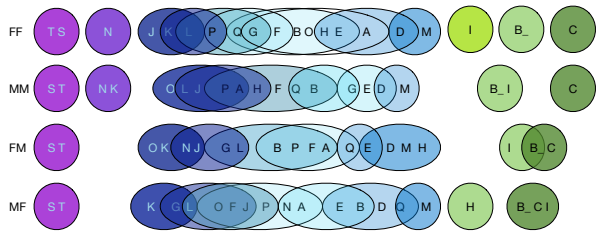

Figure 6: *Groupings of systems that do not differ significantly from each other in naturalness, per gender condition.*

In addition to the overall results, the results for each of the gender conditions is of interest. Figure 4 shows the overall MOS per gender condition, all systems combined. Figure 5 shows the results per system. Finally, Figure 6 illustrates the significance groupings of systems. A one-way ANOVA with gender condition as the within-group factor shows there is a significant effect of gender condition [$F(3, 31996) = 181.3, p < 10^{-16}$]. Post-hoc Tukey tests show, as expected, that the naturalness ratings follow the order of the means in Figure 4. Intra-gender VC scores significantly higher than cross-gender VC, in terms of naturalness.

3.2. Similarity

The similarity scores are represented in Figure 7 by stacked barplots with the listeners' confidence included. The top barplot shows VC compared to the target and the bottom shows VC compared to the source speaker. Figure 8 gives the same results with the degree of confidence omitted. Barnard's exact test, with Bonferroni correction was used to calculate significance between systems on the binary same/different data, illustrated in Figure 9. Finally, results per gender condition are presented in Figure 10.

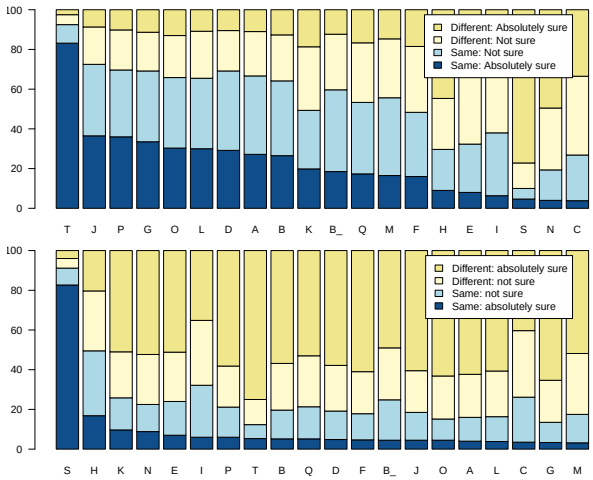

Figure 7: *Similarity (with listener confidence) to target speaker (top) and to source speaker (bottom) over all ST pairs.*

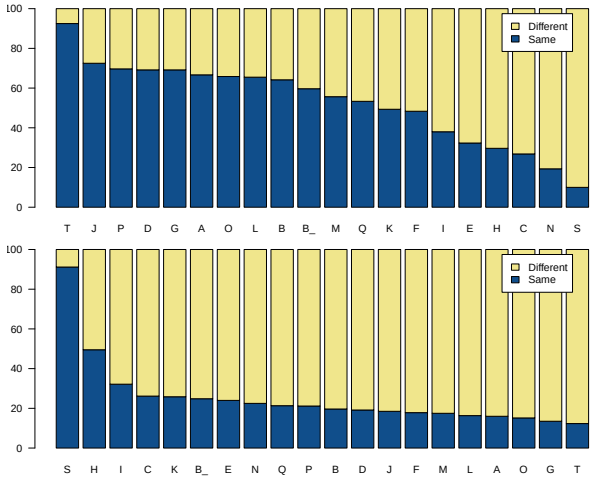

Figure 8: *Similarity same/different to target speaker (top) and to source speaker (bottom) over all ST pairs.*

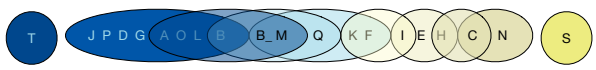

Figure 9: *Groupings of systems that do not differ significantly from each other in terms of similarity to target.*

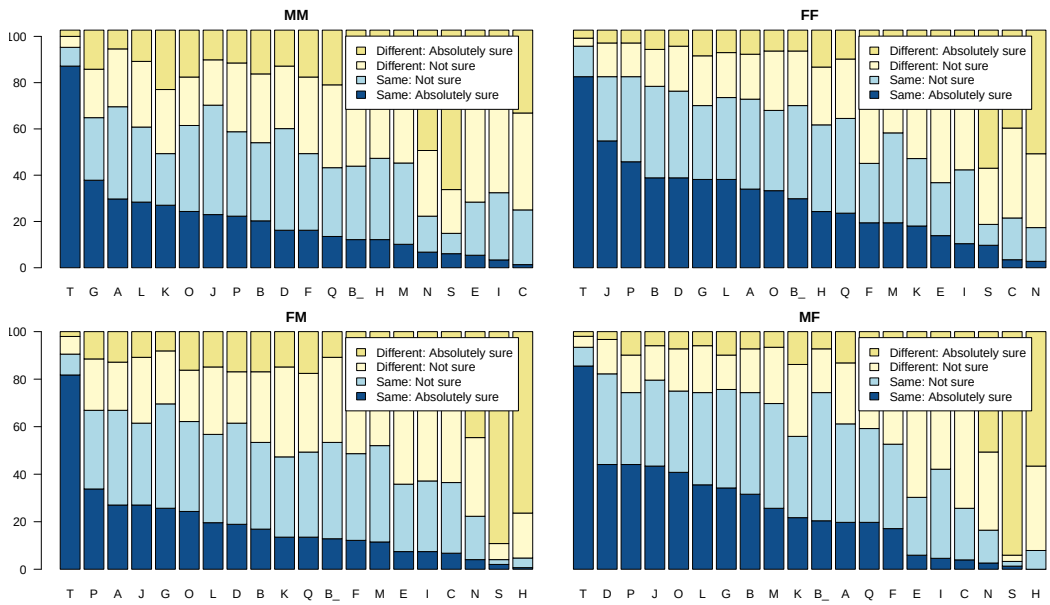

Figure 10: *Similarity (with listener confidence) to target speaker, per gender condition.*

Figure 8 shows that – viewing the binary decision – speaker similarity of the top systems is around 70%, which is quite high. However, the 'not sure' portions of even the top systems is high (Figure 7). For example, for system "J", the proportion same is 72.5% of this 36.5% is 'sure' and 36% is 'not sure'. For all other systems, the proportion of 'not sure' is larger than 'sure'.

A one-way ANOVA on the similarity data with gender condition as the within-group factor again shows a significant effect of gender condition [$F(3, 11996) = 41.21, p < 10^{-16}$]. The post-hoc Tukey HSD test shows that FM and MM do not differ significantly from each other nor do FF and MF. It further reveals that, FF and MF score significantly higher in terms of similarity to the target than MM and FM. This suggests that conversion to the female target voices is more successful.

4. Discussion

The VCC results indicate that there is still a lot of work to be done in voice conversion, it is not a solved problem. Achieving both high levels of naturalness and a high degree of similarity to a target speaker –within one VC system– remains a formidable task, succinctly summarised in Figure 11. The fact that so much of the similarity score is made up of listeners not being sure whether or not the samples were from the same speaker or not is an indication of the difficulty of the task, both for listeners and for the VC community. That said, the results for female to female conversion show that it is possible to achieve very high levels of similarity 80%, with certainty above 50%.

Carrying out an evaluation of this size is a complex task and compromises were inevitable. For example, distributing the ST pairs across two sets for naturalness rating. Although we attempted to make the sets as comparable as possible –by balancing across gender conditions– the ratings of the systems are still context sensitive and whether or not they should be compared is a disputable point [10]. For similarity evaluation, multi-dimensional scaling (MDS) of the systems compared to both target and source would have been compelling, however,

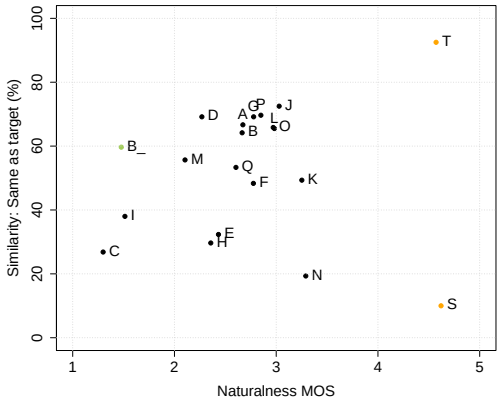

Figure 11: *Overall results for naturalness versus similarity to target speaker for all ST pairs.*

the number of trials needed to place an ST pair in such a space was prohibitive. In this case, measuring similarity for all 16 ST pairs took priority over a more exhaustive evaluation of only a few ST pairs. Future work will include a full MDS analysis for a select set of ST pairs. Finally, our listener population is disproportionally female and British-English (see Table 3). It cannot be ruled out that listener gender has had an effect on the results, which could be an explanation for the higher similarity results for female target voices. Furthermore, British-English listeners may be insensitive to American-English prosody thus missing out on subtle speaker identity cues. Further investigation is needed to ascertain whether different types of listeners do indeed rate similarity differently.

Acknowledgements We are grateful to COLIPS for sponsoring the evaluation of the VCC. This work was supported by the EPSRC under Programme Grant EP/I031022/1 (Natural Speech Technology) and EP/J002526/1 (CAF). The VCC database and listening test results are available at `http://dx.doi.org/10.7488/ds/1430`.

5. References

[1] T. Toda, L. Chen, D. Saito, F. Villavicencio, M. Wester, Z. Wu, and J. Yamagishi, "The Voice Conversion Challenge 2016," in *(submitted to) Interspeech*, 2016.

[2] M. Fraser and S. King, "The Blizzard Challenge 2007," in *Proc. Blizzard Challenge (in Proc. SSW6)*, 2007.

[3] R. A. Clark, M. Podsiadlo, M. Fraser, C. Mayo, and S. King, "Statistical analysis of the Blizzard Challenge 2007 listening test results," in *Proc. Blizzard Challenge (in Proc. SSW6)*, 2007.

[4] T. Toda, H. Saruwatari, and K. Shikano, "Voice conversion algorithm based on Gaussian mixture model with dynamic frequency warping of STRAIGHT spectrum," in *Proc. ICASSP*, vol. 2, 2001, pp. 841–844.

[5] D. Sündermann, H. Höge, A. Bonafonte, H. Ney, A. Black, and S. Narayanan, "Text-independent voice conversion based on unit selection," in *Proc. ICASSP*, 2006.

[6] S. Desai, E. V. Raghavendra, B. Yegnanarayana, A. W. Black, and K. Prahallad, "Voice conversion using artificial neural networks," in *Proc. ICASSP*, 2009, pp. 3893–3896.

[7] D. Erro, A. Moreno, and A. Bonafonte, "Voice conversion based on weighted frequency warping," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 922–931, 2010.

[8] E. Helander, H. Silén, T. Virtanen, and M. Gabbouj, "Voice conversion using dynamic kernel partial least squares regression," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 806–817, 2012.

[9] S. Aryal and R. Gutierrez-Osuna, "Accent conversion through cross-speaker articulatory synthesis," in *Proc. ICASSP*, 2014, pp. 7694–7698.

[10] J. P. van Santen, "Perceptual experiments for diagnostic testing of text-to-speech systems," *Computer Speech & Language*, vol. 7, no. 1, pp. 49–100, 1993.

[11] O. Baumann and P. Belin, "Perceptual scaling of voice identity: common dimensions for different vowels and speakers," *Psychological Research*, vol. 74, no. 1, pp. 110–120, 2010.

[12] M. Wester, "Talker discrimination across languages," *Speech Communication*, vol. 54, no. 6, pp. 781–790, 2012.

[13] A. Kain and M. W. Macon, "Design and evaluation of a voice conversion algorithm based on spectral envelope mapping and residual prediction," in *Proc. ICASSP*, vol. 2, 2001, pp. 813–816.

[14] A. Schmidt-Nielsen and D. P. Brock, "Speaker recognizability testing for voice coders," in *Proc. ICASSP*, vol. 2, 1996, pp. 1149–1156.

[15] J. Kreiman and G. Papcun, "Comparing discrimination and recognition of unfamiliar voices," *Speech Communication*, vol. 10, no. 3, pp. 265–275, 1991.