9th ISCA Speech Synthesis Workshop
13-15 Sep 2016, Sunnyvale, USA

Multidimensional scaling of systems in the Voice Conversion Challenge 2016

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

This study investigates how listeners judge the similarity of voice converted voices using a talker discrimination task. The data used is from the Voice Conversion Challenge 2016. 17 participants from around the world took part in building voice converted voices from a shared data set of source and target speakers. This paper describes the evaluation of similarity for four of the source-target pairs (two intra-gender and two cross-gender) in more detail. Multidimensional scaling was performed to illustrate where each system was perceived to be in an acoustic space compared to the source and target speakers and to each other.

Index Terms: Voice Conversion Challenge, evaluation

1. Introduction

The Voice Conversion Challenge (VCC) 2016 [1], one of the special sessions at Interspeech 2016, was devised to better understand different voice conversion techniques. This was facilitated by supplying a common dataset for participants to perform speaker identity conversion on and by carrying out an evaluation of the naturalness and similarity (to target and source speakers) of the resulting speech. In total, 17 participants from around the world contributed to the challenge. A description of VCC 2016 including motivation, database, participation rules, and main findings is given in [1]. A more detailed presentation and analysis of the evaluation results can be found in [2]. The current paper expands on the similarity evaluation.

Speaker similarity in the original evaluation [1, 2] was measured using the same/different paradigm. Pairs of stimuli were presented to subjects and their task was to judge whether the stimuli could have been spoken by the same person. For each source-target (ST) pair the 18 voice conversion (VC) systems (17 participants and the baseline system) were compared to both the source and target speakers. This paper extends the similarity evaluation by also comparing all VC systems to each other. By comparing all systems to each other and to the source and target speakers it enables visualisation of the distances between systems using multidimensional scaling (MDS).

The idea is that there is a psychological distance between stimuli in a perceptual space. The greater the distance between stimuli, the greater the dissimilarity in perception of these stimuli is. MDS has been used in speech perception studies to measure the distance between various types of stimuli, e.g., languages [3], tones [4], vowels [5, 6] and consonants [7]. More closely related to the current topic are studies that have compared speakers [8, 9, 10] or synthetic speech systems [11, 12]. In the context of voice conversion, we are not aware of studies that have considered MDS to measure the effectiveness of different VC techniques, or how they relate to each other.

It is intuitively appealing to be able to place VC systems in a perceptual space in relation to source and target speakers as the goal in voice conversion is to convert speech from a source speaker into that of a target speaker. In this study, we aim to map the perceptual space of how listeners perceive VC voices in relation to each other and in relation to the target and source speakers they are based on. We are interested in finding patterns in the perception of VC systems, e.g., groups of techniques or types of systems that lead to similar results. Visualising the distance between VC systems could (possibly) shed further light on the similarity results. How do the VC systems relate to each other? Are there systems that sound very similar? Which systems are very different?

The design of the experiment, including how the four source-target pairs were selected, is given in Section 2. The results section (Section 3) reports similarity to target in the current experiment and how this compares to the original evaluation results. After presenting MDS plots for each of the four ST pairs (Section 3.2) the results are discussed and the paper concludes in Section 5 with a few take-home messages.

2. Experimental set-up

This section describes the procedure used to select ST pairs for evaluation, the design of the listening test and how it was carried out.

2.1. Selection of ST pairs

For VCC 2016, each participant created 25 VC voices using the five source and five target speakers that were distributed in the challenge. Of those 25 voices, 16 ST pairs were evaluated in the VCC evaluation, four for each gender condition (Male-Male, Female-Female, Male-Female and Female-Male). For the current evaluation, four ST pairs were selected, one for each gender condition. The selection was based on spread across systems in terms of similarity to the target speaker in the previous evaluation [1, 9]. We aimed for ST pairs where roughly 50% of systems achieved more than 50% similarity to the target speaker. This resulted in the selection of following ST pairs: SF1-TF1, SM1-TM2, SM1-TF1 and SF1-TM2. Figure 1 shows similarity to the target speaker for the selected ST pairs. These results, per ST pair, are from the original evaluation. In all figures, the letters 'A' ... 'Q' indicate the 17 participants, 'S' is for source speaker, 'T' for target and 'B_' indicates the baseline (B_).

2.2. Listening test design

The goal of the current listening test is to relate each VC system to each other VC system and to the source (S) and target (T) speakers. As there were 17 participants, a baseline system, a target speaker and a source speaker this resulted in a total of 20

http://dx.doi.org/10.21437/SSW.2016-7

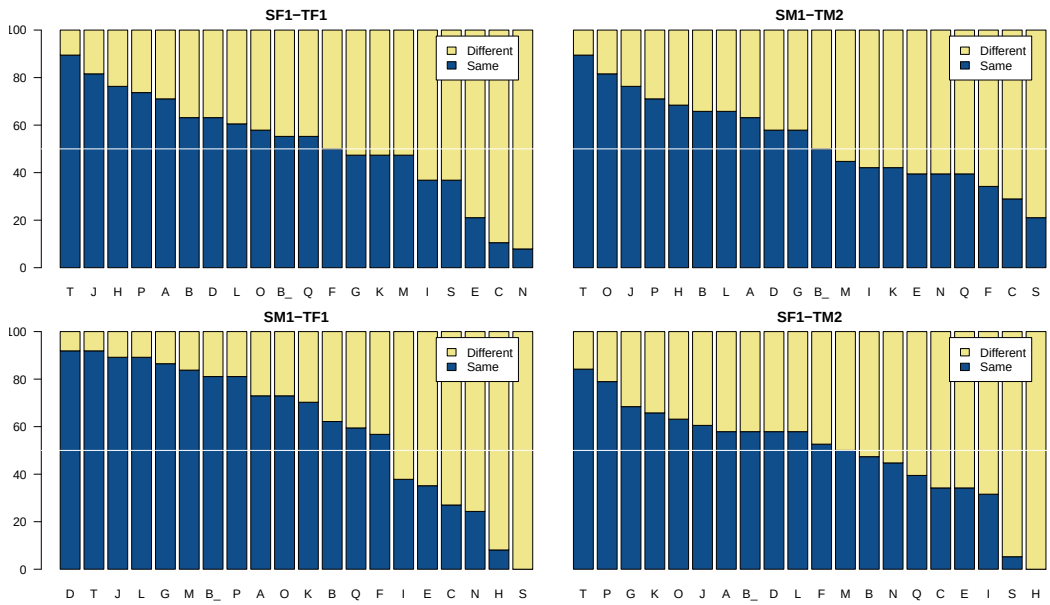

Figure 1: *Similarity (same/different) to target speaker for each selected ST pair.*

"systems" to compare. The number of trials to judge is then 210 per ST pair $((20 * 19)/2 + 20)$. Thus, a subject listens to 420 sentences, and gives 210 judgements, which translates into an estimated listening time of 1 hour.

When designing a discrimination task using the same/different paradigm it is standard practice to balance the amount of "same" and "different" trials in order to limit the bias in listener response strategies. 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 "different". However, strictly speaking there is no real correct answer. Regarding the VC-VC comparisons, some VC systems will sound more like each other, others will be sound more dissimilar. Basically, there is not an absolute same/different correct answer. Thus, in a way, the requirement to have equal same and different trials is met as these VC-VC comparisons can be considered to be perceptually somewhere between same and different. Furthermore, although standard practice, it is not a strict requirement and others have also chosen to work with comparisons that have not been balanced [13, 6].

The order of trials was random for each listener, with each sentence selected at random with replacement from the pool of 30 test sentences (same sentences as [1, 9]). Within a trial the order of the two sentences was random and the linguistic content of the sentences was different. Similar studies supporting the use of different sentences are [9, 11, 14, 6, 15], an exception is Kreiman & Papcun [16] who did use sentences with the same linguistic content, but in their study two acoustic versions of each sentence were available and listeners never compared two identical stimulus tokens.

The instructions given to the listeners were *"The purpose of this study is to judge whether two given 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. For each box, please listen to the two samples and decide whether the samples are spoken by the same speaker."* The options for judgement available to the listeners were: "Same" and "Different".

2.3. Listeners

Eighty subjects took part in the experiment. Each listener was given one of the ST pairs to rate. They were seated in sound isolated booths and listened to the samples using Beyerdynamic DT 770 PRO headphones. The experiment was carried out using a web interface. It took listeners on average 45 minutes to complete the experiment, slightly faster than was estimated. Listeners were remunerated for their time and effort. The experimental set-up was designed to result in 20 listeners per ST pair. A mix-up in assigning the correct test to two of the listeners, however, resulted in 19 subjects for SF1-TF1 and SM1-TF1.

3. Results

Non-metric multi-dimensional scaling (MDS) was used to visualise the same/different judgements. Sammon's non-linear mapping [17], a form of non-metric multidimensional scaling was used. All the solutions are two-dimensional computed using Sammon in R [18]. This implementation chooses a two-dimensional configuration to minimize the stress. The sum of squared differences between the input distances and those of the configuration are weighted by the distances. The whole sum is then divided by the sum of input distances to make the stress scale-free.

Table 1: *Statistics on the differences between the original similarity scores and current scores.*

ST pair	Pearson's r	% drop
SF1-TF1	0.77	-300
SM1-TM2	0.76	-364
SM1-TF1	0.75	-564
SF1-TM2	0.84	-282

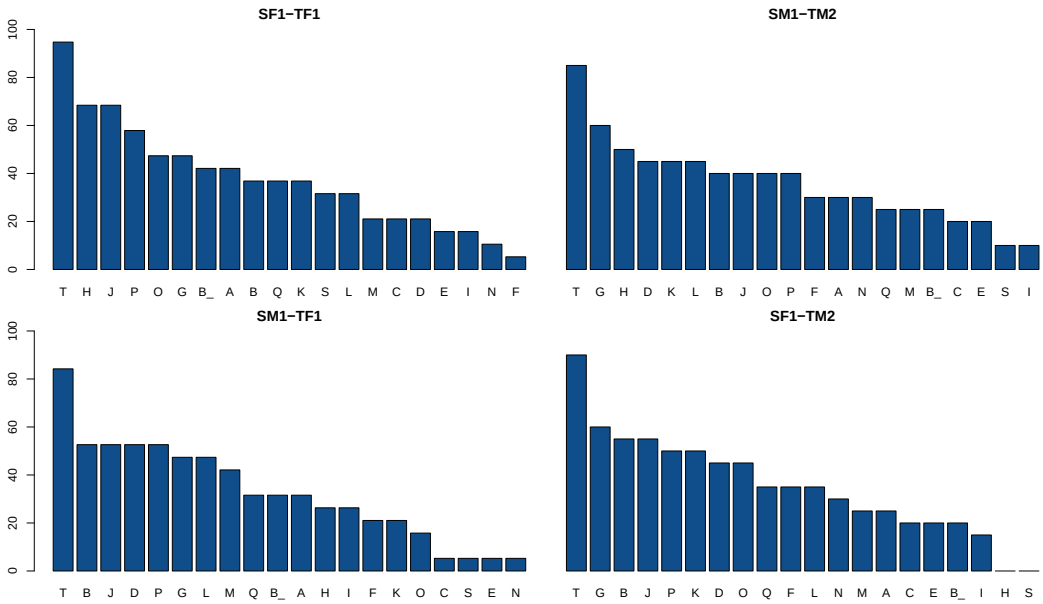

Figure 2: *% correct (same as target) for the four ST pairs*

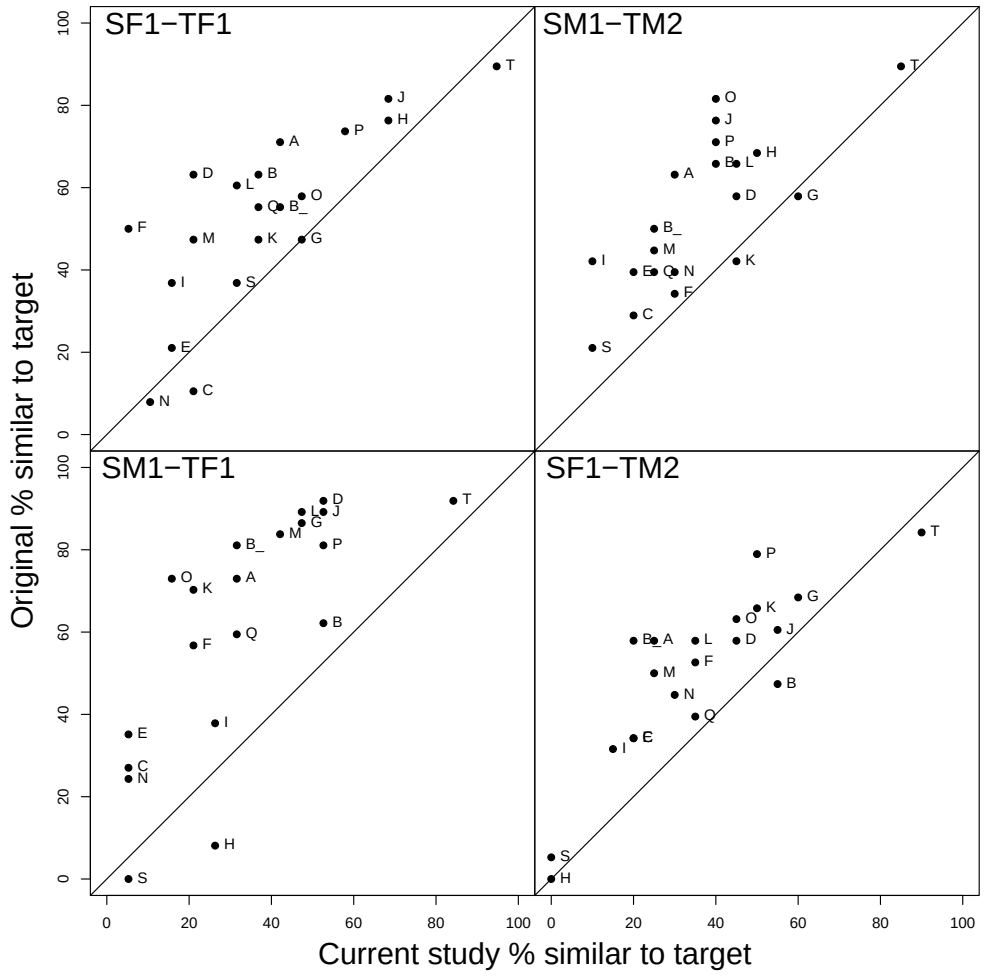

Figure 3: *Original vs current study % correct (same as target) for the four ST pairs*

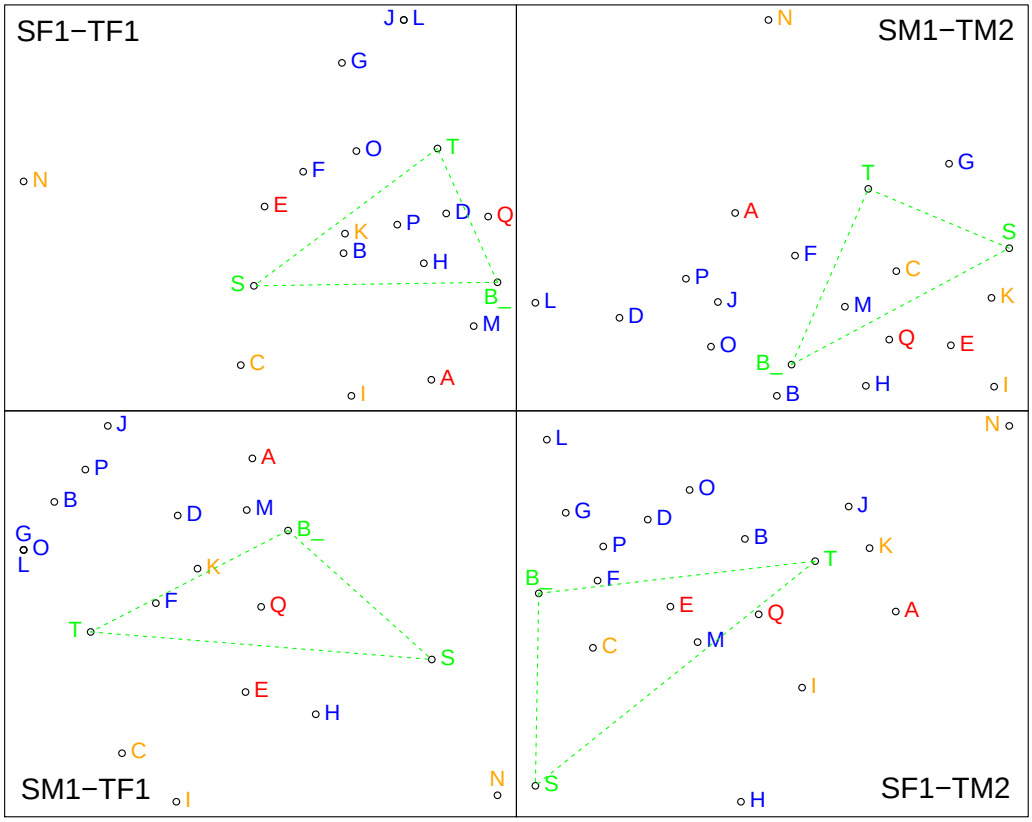

Figure 4: *MDS plots per ST pair. Color indicates vocoder type: red: Ahocoder, blue: STRAIGHT, orange: other.*

3.1. Similarity to target

The percentage correct compared to target per ST pair is shown in Figure 2. Comparison to Figure 1 shows that the orders of systems have changed somewhat and the percentages similar to target are higher in the original test than in the current evaluation for all four ST pairs. To more clearly illustrate the differences in results, Figure 3 shows the relationship between the original similarity to target results and those found in the current evaluation. In addition to that, Table 1 shows Pearson's correlations between the original and current evaluations for the different ST pairs as well as the overall percentage drop in performance.

3.2. Multidimensional scaling

Figure 4 shows MDS plots for the four ST pairs. The source (S), target (T) and baseline (B_) are connected by green dotted lines to facilitate interpretation of the figures. The stress values are: SF1-TF1: 0.06, SM1-TM2: 0.07, SM1-TF1: 0.05, and SF1-TM2: 0.05. There are a number of general observations that can be made:

- There is no obvious correspondence between the results for the different ST pairs.
- The distance between S and T is larger in the across-gender conditions than in the within-gender conditions, as you would expect.
- System N is a clear outlier in all four cases.

Table 2 gives a general overview of the vocoder, parametrisation and models used in the different systems. Note, this table is a simplification of the techniques used in the various systems and is based on the responses provided by the participants to a questionnaire from the VCC organisers.

If we consider categorisation according to type of vocoder used, we can distinguish three groups, indicated using color in Figure 4: Ahocoder: (A, E, Q), STRAIGHT: (B, D, F, G, H, J, L, M, O, P) and other: (C, I, K, N). Using this categorisation highlights the following relationships. In the cross-gender conditions, the systems using Ahocoder show some systematic behaviour. The red systems group along the x-axis dimension for SM1-TF1, and along the y-axis dimension for SF1-TM2. The orange group do not only distinguish themselves by not using Ahocoder or STRAIGHT, but also by not using MGC. From this group, C and I cluster together for target TF1 (left side of Figure 4). They are relatively far away from most other systems and near to S. As so many of the systems use STRAIGHT, a clear group is less obvious. However, there are sub-sets of systems (J, P, B, G, O, L) that cluster together and also score the highest in terms of similarity to target, cf. Figure 2.

Categories based on the models/techniques used in systems are difficult to establish as most systems are not easily defined by just one technique. Notwithstanding, for some of the ST pairs it looks like systems that use LSTM (G, L, M, O) cluster together to some degree.

4. Discussion

Although it is an attractive idea to visualize the distances between the target speaker, source speaker and VC systems using multidimensional scaling, interpretation of the results in not

Table 2: *System details. For acronyms/abbreviations, see voice conversion literature.*

System	Vocoder	Parametrisation	Model
A	Ahocoder	MGC	GMM + MGE
B	STRAIGHT	MGC	exemplar
C	Other (LPC)	LSF	DNN/GMM
D	STRAIGHT	MGC	MDN + GMM
E	Ahocoder	MGC	BLFW
F	STRAIGHT	MGC	phonetic posteriorgram
G	STRAIGHT	MGC	LSTM
H	STRAIGHT	MGC	WSOLA/Deep corr. network
I	Other (HSM)	LSF	i-vector + GMM
J	STRAIGHT	MGC	direct modification + GMM
K	Other	LSF	GMM
L	STRAIGHT	MGC	fusion (incl. LSTM)
M	STRAIGHT	MGC	LSTM
N	Other	LP coef.	Speech Filing System (SFS)
O	STRAIGHT	MGC	GTDNN/LSTM
P	STRAIGHT	MGC	MLPG /GMM
Q	Ahocoder	MGC	frame selection/MLPG

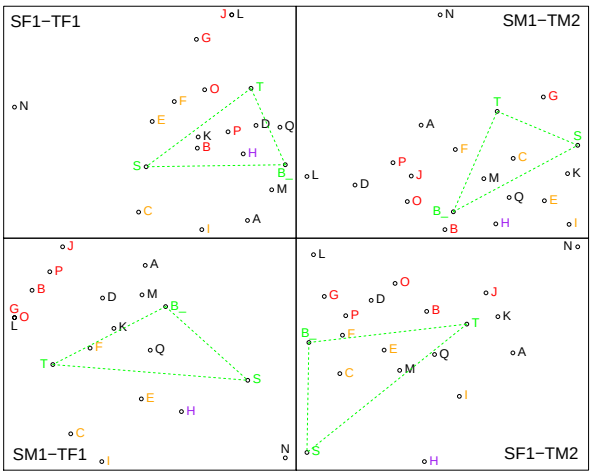

Figure 5: *MDS plots per ST pair. Color indicates score type: red: high score, orange: low score, purple: high & low*

straightforward.

VC systems do not necessarily behave the same on cross-gender ST pairs and intra-gender pairs. It can even be down to the specifics of the source speaker and the target speaker that influence the successfulness of a certain system. Nevertheless, there are a few systems (J, B, P, G, O) that consistently perform at the top end of the similarity scale, and others that languish at the lower end of the scale (I, C, E, F). One that bucks the trend is system H which is at the top for intra-gender conversion, but near the very bottom for cross-gender conversion. Figure 5 illustrates how the categorisation into high and low scoring systems forms quite distinct groups. It is interesting to note how the red group clusters. For instance, for SM1-TM2, systems J, B, O, & P cluster together near the baseline whereas G is quite distant from this cluster but nearer the target. Thus, the MDS and similarity percentage correct results suggest, in this case, that J, B, O, & P sound more similar to each other than to the target.

The difference in similarity to target results between the original experiment [1, 2] and the current experiment raises the question whether or not the task that the listeners were set was do-able. The overall drop in performance was on average 375 percentage points per ST pair, roughly 19% per system. What might explain this drop in performance?

In the original experiment, listeners only ever compared VC system to target or source, i.e., synthetic speech was always compared to natural speech. In the MDS test, the listeners compared each VC system to each other VC system (i.e., synthetic - synthetic) and to source and target (synthetic - natural speech). The combination of both types of trials in one experiment (although inevitable) may have made listeners less likely to judge a system same as target (or source) when the comparison was between natural and synthetic speech. Comparing synthetic speech to natural speech has been shown to lead to a drop in performance in speaker similarity [19]. How it might be affecting judgements in the current experiment is an unknown.

Another factor that may have influenced the listeners' judgements is the boredom-factor. Although subjects did not complain about the arduous nature of the task, it may have been playing a role. Each listener was asked to judge one ST pair, and although they were not (explicitly) aware of this, the lack of variety in source and target speakers may have influenced their judgements. Furthermore, the MDS experiment took around 45 mins to complete whereas in the original experiment, each listener judged three ST-pairs, each taking less than 10 min.

5. Conclusions

Listeners rate VC systems less similar to target speakers when confronted with many more comparisons to judge. Whether this is due to fatigue, boredom or possibly the confounding effect of synthetic-synthetic and natural-synthetic trials within one experiment remains to be investigated.

The large number of participants in the VCC is a double-edged sword. It is clear there is a great interest in shared data sets and evaluations so participants can measure the effectiveness of their system compared to other systems. However, the large number of systems in the VCC makes it difficult to conduct effective evaluations. Future challenges will need to keep this in mind and devise alternative approaches to VC evaluation. For instance, some kind of cascaded set of experiments

could be designed, where in a first round, systems using similar techniques are measured and ranked, followed by a listening test in the second round, in which the top systems from the first round are compared.

Using multidimensional scaling adds to the interpretation of similarity scores by enabling slightly different types of comparisons between systems than mere ranking based on percentage correct. For instance, the type of vocoder used in a system has an audible effect which is visible in the cross-gender MDS plots. Figure 5 further illustrates that systems that score high on similarity to target are grouped together. For some ST pairs, the MDS plot indicates that there are VC systems that sound more like each other than like the target.

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 permanently available at `http://dx.doi.org/10.7488/ds/1430`.

6. 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. Wester, Z. Wu, and J. Yamagishi, "Analysis of the Voice Conversion Challenge 2016 evaluation results," in *(submitted to) Interspeech*, 2016.

[3] V. Stockmal, D. R. Moates, and Z. S. Bond, "Same talker, different language," *Applied Psycholinguistics*, vol. 21, no. 03, pp. 383–393, 2000.

[4] J. T. Gandour and R. A. Harshman, "Crosslanguage differences in tone perception: A multidimensional scaling investigation," *Language and Speech*, vol. 21, no. 1, pp. 1–33, 1978.

[5] R. A. Fox, J. E. Flege, and M. J. Munro, "The perception of English and Spanish vowels by native English and Spanish listeners: A multidimensional scaling analysis," *The Journal of the Acoustical Society of America*, vol. 97, no. 4, pp. 2540–2551, 1995.

[6] 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.

[7] J. D. Harnsberger, "The perception of Malayalam nasal consonants by Marathi, Punjabi, Tamil, Oriya, Bengali, and American English listeners: A multidimensional scaling analysis," *Journal of Phonetics*, vol. 29, no. 3, pp. 303–327, 2001.

[8] R. E. Remez, J. M. Fellowes, and D. S. Nagel, "On the perception of similarity among talkers," *The Journal of the Acoustical Society of America*, vol. 122, no. 6, pp. 3688–3696, 2007.

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

[10] J. Lindh and A. Eriksson, "Voice similarity - a comparison between judgements by human listeners and automatic voice comparison," *Proceedings from FONETIK 2010, Working Papers*, vol. 54, pp. 63–69, 2010.

[11] 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.

[12] C. Mayo, R. A. Clark, and S. King, "Listeners weighting of acoustic cues to synthetic speech naturalness: A multidimensional scaling analysis," *Speech Communication*, vol. 53, no. 3, pp. 311–326, 2011.

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

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

[15] 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.

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

[17] J. Sammon, "A nonlinear mapping for data structure analysis," *IEEE Transactions on Computers*, vol. C-18, no. 5, pp. 401–409, 1969.

[18] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2010, ISBN 3-900051-07-0. [Online]. Available: http://www.R-project.org

[19] M. Wester and H. Liang, "Cross-lingual speaker discrimination using natural and synthetic speech," in *Proc. Interspeech*, Florence, Italy, 2011.