Response to reviews of paper 3681 – "Blow: a single-scale hyperconditioned flow for non-parallel raw-audio voice conversion".

We are very grateful to all reviewers for their positive assessment of our work and their constructive feedback. We have considered all minor remarks and typos and already updated our version of the manuscript accordingly.

Two reviewers comment on the forward-backward conversion procedure and our design hypothesis that the $\mathbf{z}$ space should be condition-free (or identity-neutral in the case of voice conversion). We believe the suggestions made in this regard are very interesting, and thank the reviewers for that. We have acted in two ways to improve our work. On the one hand, we have rephrased the "Forward-backward conversion" section (Sec. 4.3), removing some terms that complicated the understanding of the procedure and improving the wording of the remaining ones. On the other hand, following the reviewers' suggestion, we have performed an additional experiment to bring evidence that the $\mathbf{z}$ space is identity-neutral (the details of this experiment are already included in our current version of the manuscript). Our reasoning for the experiment that, if $\mathbf{z}$ vectors carried information about the speaker, then a powerful-enough classifier on top of $\mathbf{z}$ vectors should be able to perform speaker identification much better than random chance, which is 1.1% for the considered data set and split. In the extreme case, if $\mathbf{z}$ vectors carried as much information about speaker identity as the original $\mathbf{x}$ vectors, the classifier should approach the accuracy of the classical-feature linear classifier we use for the Spoofing metric, which is 99.3% for the considered data set and split. Hence, on top of $\mathbf{z}$ vectors, we trained both a random forest classifier with 50 estimators and a multilayer perceptron with 3 layers, and obtained accuracies of 1.8 and 1.4%, respectively; both marginally above random chance. This gives us an indication that that the $\mathbf{z}$ space has limited speaker identity information or that, in the best of cases, such information is not apparent or complex to extract.

Finally, one reviewer points out that we should compare to "AUTOVC: Zero-shot Voice Style Transfer with Only Autoencoder Loss (ICML 2019)". We thank the reviewer for pointing us out to this very recent and interesting work, and we will include it in our "Related work" section (Sec. 2). At the moment of writing this author feedback response, we are working with the provided code to make it work under our setting in order to, in the future, be able to provide additional results.