Breaking the wall of unfamiliarity of one's own voice

Sun Ruikang

sunruikang2000@gmail.com

College of Pharmacy, Shandong University, P.R.China

(Speech video is available here: https://www.bilibili.com/video/av69615843)

Abstract

The voice someone heard inside their head is generally different from it heard on record. This problem makes it difficult to sing correctly for many people.

Due to the development of Transfer Learning and Neural Networks, we can transfer the voice we really made to the voice we intend to with models like Generative Adversarial Networks (GAN).

Speech Draft (English ersion, translated by Google Translate)

Dear judges and friends here:

Hello everyone!

I am Sun Ruikang from the School of Pharmacy of Shandong University. I am sharing my broken project today with you: crack the difference between the sound you hear and the actual sound.

We know that due to the structure of the human skull, the actual sound we make is different from the sound we hear. Because of the bone conduction of the skull, people hear their own voices, which is lower than the actual sound. Take me for example. I listen to my own voice and think that I am a singer, but my friends sound like they think I am like a Donald Duck. Of course, not only do I have such troubles, but many people face the same problem. So why can't we correct this difference based on neural networks?

We can use either a one-dimensional array to represent the sound at a certain moment, or a two-dimensional tensor to represent the sound over a period of time. Admittedly, we will get better results with the latter, but which one to use depends on the project's funding and the computational power of the landing application scenario. We recruited a group of volunteers to adjust their voices recorded in the microphone to make them more like the sound they heard when they made a sound. We convert this sound data into a one-dimensional array, input a fully connected neural network, and after training, we can get the sound output we want.

If simple neural networks still do not meet our needs, we can apply digital audio to existing more advanced migration learning frameworks. For example, the now popular image migration learning frameworks such as DcGAN, Pix2Pix, etc., we can convert human voices into just the same spectral map, that is, a two-dimensional tensor to adapt to the existing framework, and then further processing. . If the funds are sufficient, we can even explore the mechanism of the difference in sound from the perspective of anatomy, specifically explaining what kind of people are more likely to have such differences, how the sounds they hear and the actual sounds are mathematically different. The specific solution depends on the budget of the project, so this project has better flexibility.

Finally, let's explore the application scenario of the project. This project not only enables people like me to make their own voices, realizes the music dreams of many people, but also develops "personal-specific voice changers" and other products to create economic benefits through cooperation with music software. Even after deep understanding of the mechanism of this difference, we can apply this project more widely in interdisciplinary research, for example, from the perspective of human physiology, to explain the changes of Chinese pronunciation from ancient times to the present.

My sharing is over, thank you!

演讲稿（中文原始版本）

尊敬的评委老师们、在座的朋友们：

大家好！

我是来自山东大学药学院的孙睿康，我今天与大家分享我的破壁项目：破解自己听到的声音与实际发出声音的差异。

我们知道，由于人体头骨的构造，我们实际发出的声音与自己听上去的声音是不同的：由于颅骨的骨传导作用，人听到自己的声音，比实际发出的声音更加低沉。就拿我来说吧，我听自己的歌声，认为我是一个歌唱家，但我的朋友们听上去，却认为我像一个唐老鸭。当然，不仅仅我有这样的困扰，还有许多人面对着同样的问题。那么我们为什么不能基于神经网络，纠正这种差异呢？

我们既可以用一个一维数组来表示某一时刻的声音，也可以用一个二维张量来表示一段时间内的声音。诚然，使用后者我们将获得更好的效果，但采用哪一种方案取决于项目的经费与落地应用场景的计算力大小。我们征集一批志愿者，让他们调整麦克风记录的他们的声音，使其更像他们发出声音时自己听上去的声音。我们将这样的声音数据转化为一维数组，输入一个全连接神经网络，经过训练后，便可以得到我们所希望的声音输出了。

如果说简单的神经网络仍然满足不了我们的需求，我们可以将数字化的音频套用在现有更高级的迁移学习框架中。例如，现在十分热门的图像迁移学习框架DcGAN、Pix2Pix等，我们可以将人发出的声音转化为刚刚那样的语谱图，也就是一个二维张量，以适应现有框架，再进行进一步的处理。如果经费充足，我们甚至可以从解剖学的角度探索发出声音差异的机理，具体解释什么样的人更容易出现这种差异、自己听到的声音与实际发出的声音具体有怎样数学差异等问题。具体的解决方案取决于项目的预算，因此这个项目具有较好的灵活性。

最后，我们来探讨一下该项目的应用场景。这个项目不仅能让像我这样五音不全的人发出自己想发出的声音，实现很多人的音乐梦想，还能通过与音乐软件的合作，开发出“个人专属变声器”等产品，创造经济效益。 甚至，在深入了解这种差异的机理后，我们可以将这个项目更广泛地应用于跨学科科研中，例如从人类体质学角度，阐释汉语发音从古至今的变迁等等。

我的分享到此结束，谢谢大家！