FINAL PROJECT: ASMR is all you need
Spoken and Written Language Processing (POE-GCED)

Víctor Adell, Jordi Aguilar, Pau Autrand i Miquel Escobar

April 3, 2020

1 Introduction

ASMR is the sensation experienced by some people in response to specific sights and sounds, described as a warm, tingling and pleasant sensation starting at the crown of the head and spreading down the body. The 'tingles' are typically accompanied by feelings of calm and relaxation.
There are more than 13 million ASMR videos on YouTube which people watch to relax, relieve stress or sleep better. ASMR includes a wide range of sounds like tapping, scratching, water drops, rubbing and buzzing among others. It can be done on any kind of object such as metallic, plastic, hollow, flexible, rough...

2 Objectives

Our main goal is to create a neuronal network able to generate ASMR sounds. It should learn the main differences between the sounds and generate unlimited sequences of them.
Given the fact that there is a lot of variety of sounds, we could begin with a pre-selected set of them.

3 Database

Our plan is to collect the data from YouTube videos. The amount of ASMR videos in YouTube is vast, and through its API, we could easily access this content. To crop and tag the specific sounds reproduced in a video, we propose to extract this information from either the description of the video or the comments, where we usually find the timestamps of each trigger. Some preprocessing could be applied, using various signal processing techniques, in order to clean and prepare more minitiously the extracted data.

4 Work Plan

The expected work plan would consist of the following steps:

1. Extract the data mainly from Youtube videos, and label it.
2. Preprocess the data in order to ensure it is in an optimal form for being applied to our models.
3. Load the data into a database system compliant with our goals.

4. Think, design and implement multiple models, based on already existing ones such as GAN, cGAN and wavenet.

5. Train and test the models.

6. Enjoy the results! Optional: create a Youtube channel that uploads videos containing a set of produced sounds.

Finally, if time constraints allow it, we might explore to complement the produced sounds with whispered speech. In this sense, the following paper, Voice Conversion for Whispered Speech Synthesis, presents an approach to synthesize whisper by applying a handcrafted signal processing recipe and Voice Conversion (VC) techniques to convert normally phonated speech to whispered speech. It is relevant to note that this paper has been trained on the publicly available wTIMIT corpus.

5 Reference paper and baseline

- WaveNet: A generative model for raw audio, with the corresponding blog post from Deepmind
- AutoVC: Zero-Shot Voice Style Transfer with Only Autoencoder Loss

6 Evaluation

Our project will succeed if the generated output is close enough to the sounds we are intending to generate. If not evaluated by the neuronal network itself, we could apply a similarity function between our output and the desired sounds. In this case, a human evaluation of the generated audio, analyzing if it produces the expected relaxing sensations would also be valid.