



Speech Breathing Synthesis for Empathic Virtual Agents

The weekly updates presentation.

Nicolò Loddo



Week 2 - Speech Breathing Empathy Project

Nicolò Loddo

What happened

- I started experimenting with Flowtron. I installed it and used the inference module to synthesize simple text to speech.
- Resolved problems with WaveGlow (mel-spectrogram to audio) to achieve inference.
- Started looking into Style Transfer for Flowtron through a colab notebook that explains how to. It didn't go through for now because of an update of PyTorch. It seems quite easy to solve accessing Flowtron's code
- Found a library that features various TTS models that I have seen in the literature in a convenient framework: <https://github.com/coqui-ai/TTS> and that features a simple model implementation framework
- Read further into spontaneous speech synthesis on the KTH Speech Synthesis demo page which features various studies with interesting results (<https://www.speech.kth.se/tts-demos/>)

- Started some sample observation of the INTERSPEECH dataset
- Extended my literature readings with other examples of spontaneous speech applied to VAs

Next week to-dos

- Try other models on coqui.ai page
- Do a systematic exploration of the INTERSPEECH dataset
- Try to set up the training of a model, maybe using coqui.ai and/or with Flowtron itself, using the INTERSPEECH dataset
- Look more into training spontaneous speech models papers
- Extend literature review
- Make schedule also for phase 2 how to respond to RQ and all the design process

Examples



01

Spontaneous1

<https://www.speech.kth.se/tts-demos/inter-speech2022/>



02

Spontaneous2

https://hfkm.github.io/pc_nhmm_tts/



03

Spontaneous3

<https://www.speech.kth.se/tts-demos/LREC22/>

04

Flowtron

<https://nv-adlr.github.io/Flowtron>



Flowtron

05

Microsoft

<https://azure.microsoft.com/en-us/products/cognitive-services/text-to-speech/#features>

Week 3 - Speech Breathing Empathy Project

Nicolò Loddo

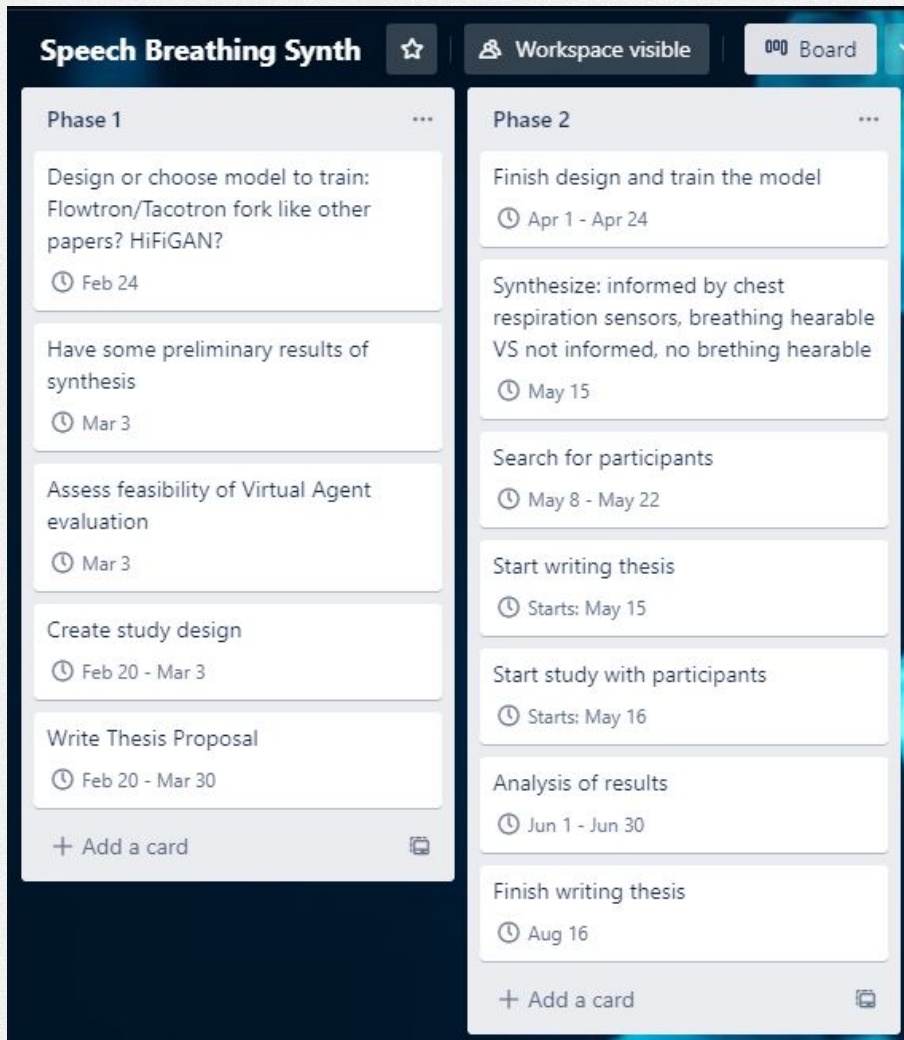
What happened

- Fixed the Style Transfer for Flowtron and applied it with the RAVDNESS Emotional dataset.
- Small exploration of the INTERSPEECH to understand how to exploit the data
- Wrote scripts for transcription of audio files through IBM Watson Speech to Text recognition (API)
- Wrote scripts to use the Gentle Aligner: insertion of Filled Pauses, accurate time alignment and phonemes recognition
- Applied the above said scripts on the INTERSPEECH dataset: I managed to transcribe it all, with filled pauses and phonemes, and to have it aligned with timestamps per phoneme.
- I qualitatively checked consistency of the timestamps with the breathing timestamps, it seems to check out

- Tried Flowtron Style Transfer with the transcribed INTERSPEECH: memory problems due to the length of the audios
- Made schedule
- Found other interesting spontaneous speech models, in particular AdaSpeech seems really good

Next week to-dos

- Search for emotional dataset in english
- Try other models, maybe on coqui.ai
- Adapt a model design for the training of our data. The breath levels are what makes our data special
- Try to set up the training of the forked model
- Read better the newly found spontaneous speech papers
- Extend literature review



HiFiGAN Demo Audio:

<https://jik876.github.io/hifi-gan-demo/>

Week 4 - Speech Breathing Empathy Project

Nicolò Loddo

What happened

- Made structure of thesis proposal
- Examined other models and approaches to speech
- Extended speech synthesis literature
- Wrote scripts for segmenting audio to feed to flowtron, then changed idea on it
- Wrote scripts to use Montreal Forced Aligner
- Wrote script to use AssemblyAI Speech to text API
- Evaluated results the Montreal Forced Aligner against the previously used Gentle
- Evaluated results of AssemblyAI against the previously used IBM Watson
- Wrote script that systematically identifies breathing pauses in the audio from the aligned transcriptions

- Looked more into HiFi-GAN which turned out to be a vocoder, not a full text to speech model. Still it can be useful for later processes
- Searched a bit for a spontaneous speech emotional dataset in english, with not great results

Next week to-dos

- Search better for emotional datasets in english
- Start putting papers into each section of the thesis proposal
- Read and comment better the speech synthesis literature with its evaluation methods
- Better exploration of INTERSPEECH
- Make script to label the breathing pauses on the dataset and possibly make a go at style transfer with INTERSPEECH (finally!)
- Report the evaluations on the aligners, transcriptions and whatever I try

Week 5 - Speech Breathing Empathy Project

Nicolò Loddo

What happened

- Full exploration of INTERSPEECH
- Development of a library to analyse voice and breathing datasets with visual plots
- Continued the evaluation and made decisions on aligner and transcriptor
- Reported the comments on the aligners and transcriptors on overleaf
- Started thesis proposal
- Enhanced breath labeling script
- Fixed the project directories, splitted the exploration jupyter notebook

Next week to-dos

- Search better for emotional datasets in english
- Put papers into each section of the thesis proposal
- Read and comment better the speech synthesis literature with its evaluation methods
- Make a go at style transfer with INTERSPEECH
- Start the writing of the literature review

Week 6 - Speech Breathing Empathy Project

Nicolò Loddo

What happened

- Decided to not choose a specific aligner but rather merge two of them to get the best parts of both.
- Developed a script that merges Gentle and Montreal aligners putting priority on Gentle's alignment, and that manages inconsistencies in timing of close words.
- Generalized and polished all the scripts to be used on datasets of audio even without the breathing signal
- Finished the breath labeling script, added parameters to be used from command line
- Qualitative and quantitative evaluation of the parameters to use for the breath detection and labeling
- Ran the breath detection and labeling: I now have transcriptions with breathing tags and their alignments
- Wrote script to segment big audios into smaller sentences with breathing instances separating them. It features upper and lower limits of time that cuts the audio even if there is no breathing.

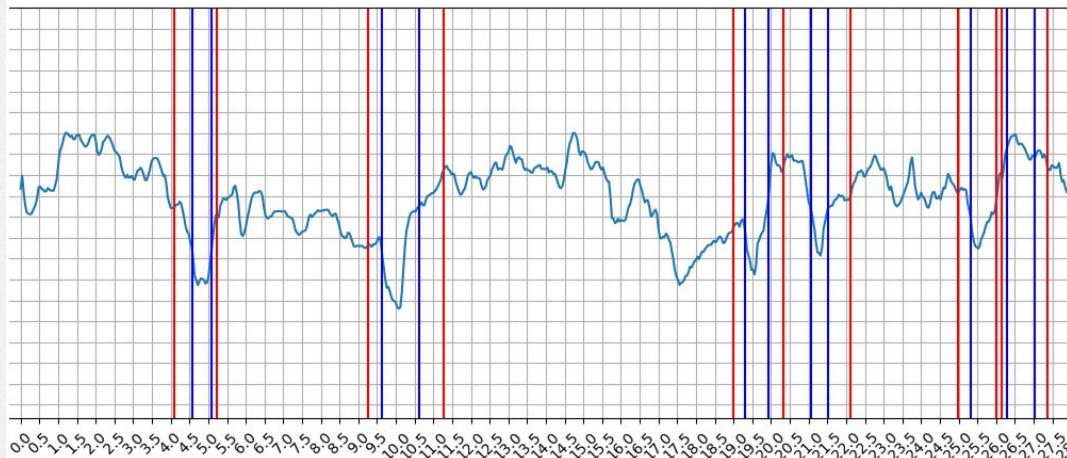
- Qualitative evaluation of the audio-segmenting script: everything seems fine. The preprocessing pipeline is done
- Wrote script that encapsulates the preprocessing pipeline

Next week to-dos

- Report the evaluations on overleaf
- Run the segmentation script
- Make a go at style transfer with INTERSPEECH
- Make a go at training some model with the data
- Put papers into each section of the thesis proposal
- Start the writing of the literature review
- Read and comment better the speech synthesis literature with its evaluation methods
- Search for more emotional datasets in english


```
ndt.search_breath_analysis('devel_00.wav', gentle_mfa_json, nonstop = True, printonlyfinal = True)
```

Final total breath sections: 62



Breath detection with
visual plots on the
breathing signal

This is the detection output of the script, with statistics on the breaths of all files:

FINAL STATS:

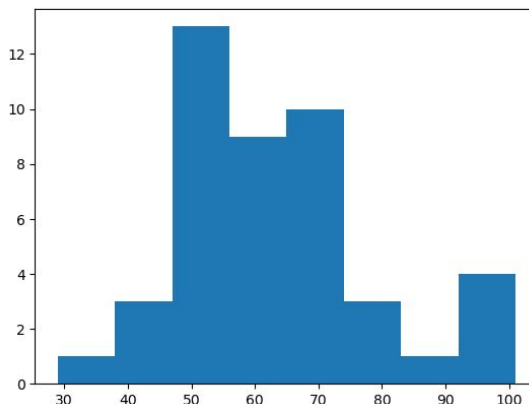
Average number of breaths: 62.43181818181818
Max breaths in a file: 101
Min breaths in a file: 29
Average breath duration: 0.5777072092602468
Max breath duration: 4.88
Min breath duration: 0.27

Since all the recordings have a length of 4 minutes we have the following respirations per minutes (rpm) stats:

Average rpm: 15.5
Max rpm: 25
Min rpm: 7

Which perfectly checks out with the known average rpm of a resting person (between 12 and 16)

mean: 62.43
std: 14.78



Breath labeling stats
(on the left); number
of breaths histogram
(on the right)

Week 7 - Speech Breathing Empathy Project

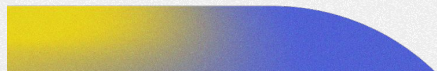
Nicolò Loddo

What happened

- Reported evaluations of breath detection and aligners on overleaf
- Further evaluated the breath detection's script varying its parameters
- Started thesis proposal document, made front page and restructured the sections
- Searched literature on breath statistics and wrote it on the thesis proposal document
- Divided papers in the sections of the thesis proposal
- Started writing Introduction, Methodology and Literature review
- Developed script to translate transcription and alignment into <breath>, <disfluency> or <word> tokens

Next week to-dos

- Finish Introduction
- Finish the firsts chapters of literature review
- Run the segmentation script
- Make a go at style transfer or training with the INTERSPEECH dataset
- Search for more emotional datasets in english?



Phase 2 Week 1+2 - Speech Breathing Empathy Project

Nicolò Loddo

What happened

- Examined the eINTERFACE'05 Dataset: I found that even though it was supposedly containing spontaneous speech, the recordings I examined qualitatively seem to be acted and do not sound spontaneous at all
- Studied the DisfluencyPredictor module of AdaSpeech3
- Studied Szekely methods of training to introduce breathing and disfluencies
- Forked VITS model
- Fixed VITS to work with disfluencies and breath labels
- Fixed VITS to work on my PC (kinda...)
- Downloaded IEMOCAP and analysed it qualitatively

Next week to-dos

- Think about Experiment
- Improve Design Study
- Finish fixing VITS
- Check out IEMOCAP



Phase 2 Week 3 - Speech Breathing Empathy Project

Nicolò Loddo

What happened: thoughts on the Study Design

- It would be fun to make it a guessing game with the possibility of continuing to guess
- It is also important to assess the quality of the synthesized speech with a MOS evaluation

Therefore the idea is:

- Demographics: age; sex; native language (maybe) -> 3 questions
- MOS evaluation on the quality -> 1 question

GUESS THE EMOTION QUIZ:

- 4 multichoice questions (4 choices)
- Possibility of continuing

REMOVED EVALUATIONS:

- No MOS on emotionality because emotionality will be assessed in the quiz; No evaluation of linguistic content.

P.S.: I started considering an Open Source implementation of FastSpeech2 as model.

The Quiz is described in the next slide.



Emotional Conditions: [possibly providing additional statistically significant results]

(as defined by James Russell's Circumplex Model through Arousal and Valence parameters; adjective placement from Nagel et al.'s study:

www.researchgate.net/publication/45189833_Worms_in_Emotion_Visualizing_Powerful_Emotional_Music


- High arousal, negative valence (Annoyed)
- High arousal, positive valence (Delighted/Excited)
- Low arousal, negative valence (Sad)
- Low arousal, positive valence (Relaxed/Serene)

Speech Features Conditions: [actually analysed in RQ]

- Without breathing
- Without filled pauses
- Without pitch contour
- Full features

Quiz parameters:

- Randomly extracted Sentences (with same linguistic emotional content, as recognized by ER model), never the same sentence, to not provide a comparison baseline
- Always randomly extracted Emotion Condition, can happen to be the same
- 1 question per Feature Condition
- If the subject continues with the quiz: randomly extracted Feature Condition (max 4 more)



Number of conditions: $4 \times 4 = 16$. The conditions at issue in RQ are though only 4: the Speech Features Conditions. For our maximum 8 questions I need 8 chosen sentences with same linguistic emotional content. I have to synthesize them across all conditions, for a total of: 128 recordings.

Phase 2 Week 5 - Speech Breathing Empathy Project

Nicolò Loddo

What happened

- Built a python class that wraps around our good Datasets to be more efficient in the preprocessing
- Quantitative exploration of IEMOCAP
- Did quantitative test on IEMOCAP emotion-discriminating coordinates: valence, activation and dominance, found statistical significance
- Tried OpenAI's Whisper to transcribe IEMOCAP better because the transcriptions and alignments are not satisfying
- Almost finished with a script that adds [HEAVY_BREATHING] and [LAUGHTER] tags to Whisper's transcriptions: those tags are present in the manually labeled IEMOCAP transcriptions
- Run my breathing labeling script on one Whisper alignment of IEMOCAP: good results

- Found out that IEMOCAP + Whisper is really promising and has all the features we need. Only problem is the breathing is not so clear as in INTERSPEECH, therefore a good idea would be to merge the two.
- Thought more about possible study designs: what about a modified Milgram experiment but on compassion and empathy towards conscious AI (instead of authority)?

Next week to-dos

- Finish the script that adds the tags from IEMOCAP transcriptions
- Run the transcription script, breathing labeler and tag adder on the whole IEMOCAP
- Merge with INTERSPEECH and train
- Synthesize