Plan Overview

A Data Management Plan created using DMPonline

Title: Iterative Models in Speech Synthesis

Creator: Xiaozhou Tan

Affiliation: The University of Sheffield

Template: Postgraduate Research DMP (The University of Sheffield)

Project abstract:

In text-to-speech area, the industry is focused on building large models capable of generating high-quality speech as fast as possible. However, in a scenario where computing resources are limited, an alternative method is needed. In particular, this project is focused on iterative models which refining generated speech several times. Iterative models seeks to produce speech comparable in quality to the state-of-the-art industry results by trading somewhat longer generation time for a smaller model size suitable in limited resource conditions.

ID: 181411

Start date: 01-03-2024

End date: 01-09-2027

Last modified: 04-07-2025

Iterative Models in Speech Synthesis

Defining your data

- What digital data (and physical data if applicable) will you collect or create during the project?
- How will the data be collected or created, and over what time period?
- What formats will your digital data be in? (E.g. .docx, .txt, .jpeg)
- Approximately how much digital data (in GB, MB, etc) will be generated during the project?
- Are you using pre-existing datasets? Give details if possible, including conditions of use.

1.What digital data (and physical data if applicable) will you collect or create during the project.

Participants' subjective evaluation about naturalness of speech will be collected. Their personal information(e.g. contact details) will be handled by data collection platform Prolific which will only be used after approval of IT Service. We will not store their personal information.

Around 40 participants are invited to listen to hundreds of speech audios.

The neural network will be created from some baselines obtained from publicly available repositories in github and will be modified by use. The code/parameters for neural networks will be installed in github or university cloud storage provided by the university.

2.How will the data be collected or created, and over what time period?

The audio data will be generated by our model, which may take several months. The participant metadata will be collected via evaluation platform Prolific which may take several weeks.

3.What formats will your digital data be in? (E.g. .docx, .txt, .jpeg)

Audio spectrogram will be stored in pdf/png format. Audio waveform will be stored in .wav or .mp3 format. participant metadata will be stored in .csv/.txt/.xlsx format.

4.Approximately how much digital data (in GB, MB, etc) will be generated during the project.

About 10 TB will be generated during the project, with around 2.5TB data generated per year

5.Are you using pre-existing datasets? Give details if possible, including conditions of use.

1) VCTK Corpus, Condition of use: It can be used freely—even commercially—as long as proper credit is given and any modifications is shared transparently. 2) LJ Speech, Condition of use: The LJ Speech dataset could be used in commercial products—provided that Linda Johnson is attributed, the license is linked, and any modifications are noted.

Looking after data during your research

- Where will you store digital data during the project to ensure it is secure and backed up regularly? (University research storage)
- How will you name and organise your data files? (An example filename can help to illustrate this)
- If you collect or create physical data, where will you store these securely?
- How will you make data easier to understand and use? (E.g. include file structure and methodology in a README file)
- Will you use extra security precautions for any of your digital or physical data? (E.g. for sensitive and/or personal data)

1.Where will you store digital data during the project to ensure it is secure and backed up regularly?

The data collected will be stored in electronic files on University Google Drive, university HPC cloud storage, our research team specific mimas cloud storage. All of them are secured and provided by the university

2.How will you name and organise your data files? (An example filename can help to illustrate this)

Files will keep their original name(if they have) with prefix implying generation method, suffix implying generation purpose

3.If you collect or create physical data, where will you store these securely?

No, we will not but lab notebooks can be used and they are stored in overleaf with limited access to public and cloud storage provided by the university.

4.How will you make data easier to understand and use? (E.g. include file structure and methodology in a README file.

Include file structure and methodology in a README file; Add documents in overleaf online

5.Will you use extra security precautions for any of your digital or physical data? (E.g. for sensitive and/or personal data)

Data collected from participants will be absolute anonymous and unidentifiable by avoid collecting contact details, ensuring that the platform doesn't collect IP addresses. Its security will be checked by our research team.

Storing data after your research

- Which parts of your data will be stored on a long-term basis after the end of the project?
- Where will the data be stored after the project? (E.g. University of Sheffield repositoryORDA, or a

subject-specific repository)
- How long will the data be stored for? (E.g. standard TUoS retention period of minimum 10 years after the project)
- Who will place the data in a repository or other long-term storage? (E.g. you, or your supervisor)
- If you plan to use long-term data storage other than a repository, who will be responsible for the data?

1.Which parts of your data will be stored on a long-term basis after the end of the project?

All digital data related to the project, including audio recordings, participant metadata, will be stored on a long-term basis.

2. Where will the data be stored after the project?

The data will be archived and stored at re3data.org or orda.shef.ac.uk, the Sheffield data repository .

3. How long will the data be stored for?

The data will be stored for a minimum period of 10 years after the end of the project, following the standard TUoS retention period.

4. Who will place the data in a repository or other long-term storage?

My supervisor, any related parties, and I will be responsible for placing the data in the designated repository or long-term storage facility.

5. If you plan to use long-term data storage other than a repository, who will be responsible for the data?

If alternative long-term data storage options are considered, the principle responsibility for data is on me.

Sharing data after your research

- How will you make data available outside of the research group after the project? (E.g. openly available through a repository, or on request through your department)
- Will you make all of your data available, or are there reasons you can't do this? (E.g. personal data, commercial or legal restrictions, very large datasets)
- If there are reasons you can't share all of your data, how might you make as much of it available as possible? (E.g. anonymisation, participant consent, sharing analysed data only)
- How will you make your data as widely accessible as possible? (E.g. include a data availability statement in publications, ensure published data has a DOI)
- What licence will you apply to your data to say how it can be reused and shared? (E.g. one of the Creative Commons licences)

1. How will you make data available outside of the research group after the project?

A DOI will be included in data availability statements in published outputs. Copies of the code will also be included in the same repository.

2. Will you make all of your data available, or are there reasons you can't do this?

The vast majority of the data will be made available. The handle of potential sensitive or personal data will comply with data protection regulations and ethical considerations.

3. If there are reasons you can't share all of your data, how might you make as much of it available as possible?

I will make the data as unidentifiable as possible then will only share it after consulting the University of Sheffield.

4. How will you make your data as widely accessible as possible?

I will provide detailed guidance for our data in the github page and promote it in related community. DOI will be used for published data.

5. What license will you apply to your data to say how it can be reused and shared?

Licenses like CC BY and CC BY-SA from Creative Commons will define the terms for data reuse and sharing, offering maximum flexibility while ensuring proper attribution and adherence to licensing conditions.

Putting your plan into practice

- Who is responsible for making sure your data management plan is followed? (E.g. you with the support of your supervisor)
- How often will your data management plan be reviewed and updated? (E.g. yearly and if the project changes)
- Are there any actions you need to take in order to put your data management plan into practice? (E.g. requesting University research storage via your supervisor.)

1. Who is responsible for making sure your data management plan is followed?

The project manager (i.e. myself), with the support of the supervisory team, is responsible for ensuring the adherence to the data management plan.

2. How often will your data management plan be reviewed and updated?

The data management plan will be reviewed and updated periodically, with a minimum frequency of yearly reviews and additional updates as needed due to project changes or evolving data management requirements.

3. Are there any actions you need to take to put your data management plan into practice?

Actions may include requesting university research storage, setting up access controls, developing documentation templates, and implementing encryption protocols for sensitive data. These actions will be coordinated by the project team in collaboration with relevant university departments and support services.