

ADVISORY REPORT

PSV DYNAMIC AUDIO

CONTENTS

Introduction.....	3
Executive solution	4
Research	5
Requirements.....	5
Investigated voice cloning tools.....	5
Resemble.ai	5
Tacotron 2.....	6
Synthesia.....	6
Descript.....	6
Respeecher.....	6
Microsoft Azure	6
NVIDIA Nemo.....	7
Acapela	7
Alternative solution(s).....	8
Audio morphing	8
Conclusion	9
Our advice	9
Appendix.....	10

INTRODUCTION

Since little was known about dynamic audio and the available tools and techniques at the beginning of this project, we have done research to gather more information about this topic. This is important, so we make the right decisions during the development of the project in order to achieve the best end result.

In this advisory report, the main research results from the research phase of the PSV Dynamic audio project are discussed, together with possible alternatives. At the end of the document you will find the advice on how best to continue the project according to the developers.

If you have any questions regarding this document, you can contact the creator via:



Aman Sharma (Developer)
+91 94663 53530
aman.sharma@greenhousegroup.com



Mathijs Fox
+31 6 12725740
Mathijs.fox@greenhousegroup.com



Niek van de Vondervoort (Developer)
+31 6 23771745
niek.van.de.vondervoort@greenhousegroup.com

EXECUTIVE SOLUTION

To send out personalized audio messages to PSV's newsletter subscribers, we need a tool that will generate these messages. Since no one within the project team has ever worked with dynamic audio, it is important to investigate which dynamic audio tools / methods works best in our project for PSV. The main question is: "How can we best apply the dynamic audio technique to personalize PSV newsletter messages?".

To find out which dynamic audio tools and techniques works best in our case, we have made a list with requirements that the tools and/or techniques need to meet. So that when we start working with a specific tool or technique, we are able to critically assess whether it meets our requirements. Ultimately, we started working with every dynamic audio tool / method we could find, and we made an overview of all the advantages and disadvantages of every tools and method.

After a lot of trying and experimenting with dynamic audio, we came to the conclusion that the Voice Cloning method (digitizing a voice with the help of AI, so that you can generate audio messages with the help of TTS) does not sound realistic enough yet to be use in production. If you still want to do this, we recommend doing this in a context where a realistic sounding voice is not necessary. When it comes to shorter messages of which the output consists of 80% the same text each time (for example: the same sentence with varying parameters), it is recommended to use audio morphing. This way you can generate realistic sounding dynamic audio messages. The disadvantage of this is that you have to enter a studio every time you want to create a new dynamic message.

RESEARCH

During the research phase of the project the developer's team of the dynamic audio project experimented with multiple dynamic audio techniques. The main technique we have focused on was voice cloning. With voice cloning you create a digital clone of somebody's voice by using AI. So, when you give the clone some text input, it will convert it to an audio file with the voice of the recorded person.

The benefits of using voice cloning is:

- Once the voice is cloned you can re-use the voice for other campaigns.
- You do not have to manually record all different parameters for a message, so all people can be reached.


Requirements






The requirements for these tools are:



- The output must sound realistic.
 - o Not robotic.
 - o With intonation to convey emotion.
 - o Must sound like the voice of the recorded player.
- The tool must be low cost.
- The output must be available in Dutch language.
- The tool does not require a lot of input data.
 - o We cannot bother the players of PSV for a long period of time, since they must focus on their training, instead of recording samples of their voice.
 - o The maximum recording time is 1 hour.

Investigated Voice Cloning tools

The tools we have investigated are listed below.

Resemble.ai <i>Pro's</i> <ul style="list-style-type: none">- <i>Easy to use.</i>- <i>Needs only 50 input sentences</i>- <i>Has an API available.</i> <i>Con's</i> <ul style="list-style-type: none">- <i>Output doesn't sound realistic.Paid.</i>	
---	---

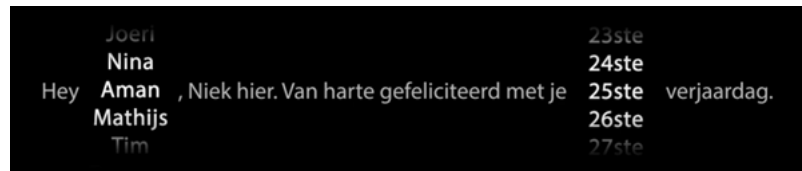
<p>Tacotron 2</p> <p><i>Pro's</i></p> <ul style="list-style-type: none"> - Needs only 5 seconds of input data. - Possibility for using a Dutch dataset <p><i>Con's</i></p> <ul style="list-style-type: none"> - Unrealistic sounding output - Hard to train a Dutch dataset 	
<p>Synthesia</p> <p><i>Pro's</i></p> <ul style="list-style-type: none"> - Has an API available. - Realistic sounding output - Has also an option for video messages <p><i>Con's</i></p> <ul style="list-style-type: none"> - High cost (€500/ month). - Maximum of 200 messages foreach month. 	
<p>Descript</p> <p><i>Pro's</i></p> <ul style="list-style-type: none"> - Realistic sounding output. - 10 minutes of input data. <p><i>Con's</i></p> <ul style="list-style-type: none"> - Part of a speech-to-text program, - Not possible to integrate this in our project - Paid (€24/ month) 	
<p>Respeecher</p> <p><i>Pro's</i></p> <ul style="list-style-type: none"> - Realistic sounding output. <p><i>Con's</i></p> <ul style="list-style-type: none"> - Cannot do text-to-speech, only convert input audio file from someone's voice to another person's voice. - Paid 	
<p>Microsoft Azure</p> <p><i>Pro's</i></p> <ul style="list-style-type: none"> - Has an API available. - Sounds realistic - Has a good and clear toolbox - Lots of documentation available <p><i>Con's</i></p> <ul style="list-style-type: none"> - Custom voice is not available in Dutch. 	

<p>NVIDIA NeMo</p> <p><i>Pro's</i></p> <ul style="list-style-type: none"> - <i>Free of charge</i> - <i>Good support from NVIDIA</i> - <i>Lots of documentation available</i> - <i>Needs 1-hour training data for voice fine-tuning</i> <p><i>Con's</i></p> <ul style="list-style-type: none"> - <i>Need a supercomputer with 4 GPU's to train a model</i> - <i>Need 24 hours of training data to train NeMo for the Dutch language.</i> 	
<p>Acapela</p> <p><i>Pro's</i></p> <ul style="list-style-type: none"> - <i>Has an API available</i> - <i>Available in Dutch</i> <p><i>Con's</i></p> <ul style="list-style-type: none"> - <i>Output does not sound realistic</i> - <i>High cost (€89.000 for 1 voice)</i> - <i>80 hours of input data</i> - <i>7-month delivery time</i> 	

ALTERNATIVE SOLUTION(S)

Audio Morphing

Another technique for dynamic audio is audio morphing. This is when you record a general message like *"Hey Aman, congratulations with your 24th birthday"*, and record a list of names and ages. With audio morphing you can cut those files in smaller parts. And change specific parameters of the general message and replace them with another name or age.



The big benefit of using this technique is that the output sounds as realistic as it can be, since it is the actual voice of a player. Not a digital clone of it. Since you have the person saying the names manually, the output will not have glitches or weird pitched words like you might have with voice cloning.

During the recording of the message and parameters, it is important to pay close attention to how you record it. If your general message is told with a lot of emotion and intonation, make sure you record the parameters in the same way. Otherwise the transition between the audio files will be very noticeable, and the output will not sound natural.

The downside of audio morphing is that every single name or age (or other parameter) needs to be recorded manually. Also trimming the audio files afterwards can take some time if you have many files. If you record for example 500 names, you will reach only approximately 62,55% of all people. With voice cloning you will not have this issue, since every name can be generated.

Top namen	Bereik*	Percentage
100	281645	30.79%
200	403208	44.08%
300	478982	52.36%
400	533396	58.31%
500	575159	62.88%
600	597206	65.29%
700	614319	67.16%
800	628015	68.65%
900	638591	69.81%
1000	651091	71.18%

*Totaal aantal personen in database: 914.736

CONCLUSION

Voice cloning is a very cool technology that can be of great added value for, for example, marketing campaigns, video games or films and animations. Despite the fact that there are many different voice cloning tools available at the moment, the result is not yet good enough to bring it into production. For that reason, many of the available tools are still under development.

Our advice

Do not use voice cloning yet if the generated voice needs to sound realistic. If you really need to use it for a project, use it in a way where it is logical that the generated voices sound less realistic. Like when using a phone, or a Google Home. In this way you can use non-realistic sounding voices. While maintaining the credibility of the messages.

If you really want to make with dynamic audio, and having it sound realistic is a must, try audio morphing. In this way you can create dynamic audio messages that sound realistic. If you use this method, try to keep messages short and try not to add more than 3 parameters since this will come unhandy when you need to record the messages. When you start recording messages, pay close attention to how you intonate messages. If you do not do this correctly, the messages will appear strange and unnatural when you stitch them together. Keep in mind that every time you need a new message, you have to go back to a studio to record the message, and to record all parameters.

APPENDIX

Item	Description	Link or file
A