SHIFT

MetamorphoSis of cultural Heritage
Into augmented hypermedia assets
For enhanced accessibiliTy
and inclusion

Funded by
the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

DOCUMENT INFO

DOCUMENT ID:	D3.6 - Text & Video to Affective Speech Synthesis final version
VERSION DATE:	
TOTAL NUMBER OF PAGES:	27
ABSTRACT:	This deliverable D3.6 finalizes the Text & Video to Affective Speech Production (TTS) tool started in deliverable D3.2 by incorporating non-English-language neural networks for TTS, expanding the systems support of languages beyond English. Additionally, enhances the naturalness of TTS voices through the integration of Voice Cloning and Affective Expansion of TTS voices, improving the emotional and affective expressiveness of SHIFT TTS. For this we develop a novel Speech Emotion Recognition (SER) system. Deliverable D3.6 also includes an immersive audio soundscape experience, going beyond speech-only TTS. Finally, this report describes how to use the Text & video to Affective Speech Production Tool with various application example-use cases for the CH partners, incorporating also the plethora of SHIFT tools as auxiliary steps in producing Video / Audio content for CH institutions. This deliverable is dedicated to task T3.3.
KEYWORDS	Affective Text-to-Speech Synthesis, Soundscapes Generation, Auditory Storytelling, StyleTTS2, Valence / Arousal / Dominance dimensions emotion space, wav2small

AUTHORS

NAME	ORGANISATION	ROLE
Dionyssos Kounadis-Bastian		Researcher

REVIEWERS

NAME	ORGANISATION	ROLE

VERSION HISTORY

VERSION	DESCRIPTION	DATE
0.0.0	First Draft	24/02/2025

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

EXECUTIVE SUMMARY

SHIFT aims to revolutionise cultural heritage (CH) institutions by introducing a suite of interconnected technologies that foster inclusivity. This inspiring initiative harnesses the latest advancements in artificial intelligence (AI), machine learning (ML), digital content transformation, linguistic analysis, and haptic interfaces to transform how we engage with cultural heritage.

This report is dedicated to Task T3.3 and presents Deliverable D3.6: “Text and Video to Affective Speech Synthesis System - Final Version” documenting the progress of T3.3 until month 30 (M30). T3.3 focuses on creating the Affective Speech Synthesis tool able to vocalise: Silent images / videos using textual descriptions and/or text-description of environmental soundscapes.

This deliverable (D3.6) presents the final version of the system, named SHIFT TTS, that incorporates Affective TTS with Multilingual support, for the CH partners languages: Namely Hungarian, Romanian, Serbian, German, Greek, English. Furthermore for English it provides 134 native-accent voices as well as 195 non-native accent voices (foreign accents of English) as well as one affective voice for the other 6 (non-english) languages. Even further we equip the system with a text-to-audio soundscape generation algorithm, namely AudioGen [1] that fulminates a visitor's auditory sensory experience.

For the development of the System, audEERING introduces a novel Speech Emotion Recognition (SER) system presented in publication [2]. This SER system drove the discovery of the affective TTS voices of SHIFT TTS system.

The role of D3.6 TTS system within the totality of SHIFT's pool of loosely connected AI tools is to be provided by textual descriptions from human CH curators or generated by D3.5 / D4.2 / D4.3 as well as videos / still images, as those build from D2.2, and subsequently generate Affective TTS audio interwoven with audio soundscapes. The present report serves as a comprehensive manual of the functionalities of D3.6 the SHIFT TTS system. Today we have implemented a Flask API command line interface to provide user access to the SHIFT TTS tool and its various options and functionalities.

The SHIFT TTS implements the following use cases: 1) Video Dubbing of a native voice by TTS or (video or silent video) through synthesis of speech from text / time-stamp subtitles. 2) AudioBook synthesis via .docx file. 3) Image-to-Speech in this case a still image serves as background of video accompanied by TTS narration of some .txt. 4) From a text description or its translation to contemporary or foreign languages, the SHIFT TTS tool creates a video / audio-track of the translated text via the appropriate TTS voice.

All functionalities include an optional argument (few word text description) of the background soundscape which the SHIFT TTS tool can synthesize through a customized AudioGen and overlay along to its TTS speech. The full SHIFT TTS system requires a Unix system with CUDA support and a GPU with at least 16 GB RAM.

The first deliverable D3.2 for T3.3 focused on technical description of the algorithmic components of the SHIFT TTS tool that still lay part of the foundation of the SHIFT TTS tool as they are proven strong across the TTS community. However D3.6 focuses on benchmarking / crafting affective TTS voices and unifying soundscape generation / video processing / non English TTS languages.

Funded by the European Union

This project has received funding from the European Union’s Horizon Europe research and innovation programme under grant agreement no 101060660.

CONTENTS

1 INTRODUCTION..... 6
1.1 SCOPE AND OBJECTIVES..... 7
1.2 STRUCTURE OF THIS DELIVERABLE..... 7
2 TEXT & VIDEO TO AFFECTIVE SPEECH PRODUCTION TOOL..... 8
2.1 AFFECTIVE TTS WITH MULTILINGUAL SUPPORT..... 8
2.1.1 AFFECTIVE EXPANSION OF TTS VOICES..... 9
2.1.2 FALLBACK MULTILINGUAL TTS ENGINE..... 13
2.2 ADVANCING THE STATE OF THE ART OF SPEECH EMOTION RECOGNITION.. 13
2.3 FLASK API..... 16
2.4 VIDEO FUNCTIONALITY..... 19
2.4.1 VIDEO DUBBING via TIMESTAMPS / SUBTITLES..... 19
2.4.2 AUDIO ONLY TTS (.wav)..... 20
2.4.3 NATIVE VOICE TO TTS (via Voice Cloning)..... 20
2.4.4 IMAGE TO SPEECH (via TXT)..... 21
2.4.5 AUDIOBOOK..... 22
3 SOUNDSCAPES SYNTHESIS..... 23
3.1 SOUNDSCAPE GENERATIONS VIA AUDIOGEN..... 23
3.2 FLASK API FOR SOUNDSCAPES..... 24
4 CONCLUSION AND FUTURE STEPS..... 27
REFERENCES..... 28

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

ABBREVIATIONS / ACRONYMS

ABBREVIATION / ACRONYM	DESCRIPTION
AI	Artificial Intelligence
A/D/V	Arousal / Dominance / Valence Emotional State Space
ANBPR	National Association of Librarians and Public Libraries of Romania
BMN	Balkan Museum Network
CER	Character Error Rate
CH	Cultural Heritage
DBSV	Deutscher Blinden- und Sehbehindertenverband
NMOS	Naturalness Mean Opinion Score
SER	Speech Emotion Recognition
SIMAVI	Software Imagination & Vision
SMB	Staatliche Museen zu Berlin
SOM	Semmelweis Orvostörténeti Múzeum
TTS	Text To Speech
VITS	Variational Inference Text (to) Speech

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

1 INTRODUCTION

SHIFT aims to deliver a set of loosely coupled technologies to increase inclusion at CH institutions. This is done by embracing current state-of-the-art innovations in artificial intelligence (AI), machine learning (ML), multi-modal data processing, digital content transformation, semantic representation, linguistic analysis, and haptic interfaces. This report presents Deliverable D3.6 and finalises the SHIFT's "Text and video to Affective speech production tool", abbreviated "SHIFT TTS tool".

The first version of SHIFT TTS was presented in the associated Deliverable D3.2 on M15. The present report, D3.6, finalises the tool of D3.2 and reflects the efforts allocated to Task T3.3 until M30.

Describing CH assets using Text-to-Speech (TTS) technology is important in today's world as it follows the pulse towards augmented reality and visitor engagement in CH exhibitions. TTS provides a vast horizon of benefits for CH exhibitions as it makes information accessible to people with visual impairments or reading difficulties, also diverse linguistic backgrounds, promoting inclusivity. It also helps preserve fragile vocal records by transferring them into audio, ensuring that valuable historical accents are not lost.

SHIFT TTS can synthesize speech from subtitles or translations of original text and blend speech to videos, vocalising CH assets from various languages. D3.6 is delegating the text preparation to the other SHIFT tools: Namely the tool of WP2 for visual content / silent video - motion-sequence, also the tool of T3.1 / T3.2 / WP4 for text creation/translation/preparation: for special audiences such as elderly, children, or visually-impaired people.

D3.6 together with the plethora of SHIFT tools brings historical artefacts, audiobooks, and video narrations to life with precisely tuned captivating affective TTS voices and also provides audio visual content for SHIFT's haptic interface tool. This report delves into the advancements of the SHIFT TTS tool, exploring its capabilities in affective voice synthesis, video dubbing, audiobook generation, and landscape to soundscape background sounds.

The SHIFT TTS system is built by advancing state-of-the-art voice cloning/synthesis neural networks, namely StyleTTS2 [3]. By utilizing StyleTTS2 auxiliary voice template input, we fine-tune the affective expressiveness of its synthesized speech, and build >200 affective voices for SHIFT TTS tool. The voices are defined as short .wav files audibly altered to access hidden-state representations within StyleTTS2 vocoder, in par with [4], to prompt it to deliver natural affective speech. The >200 voices training has been made by initial raw recordings of free open source TTS datasets followed by speed-pitch-tuning through a 2nd TTS system [5] that helps for environmental artifact limiting. Our pre-built affective voices not only assure voice naturalness but also convey emotional depth. This approach allows the system to enhance prosody, tonality and captivation for longer duration narrations, making it a powerful tool for CH and beyond.

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

1.1 SCOPE AND OBJECTIVES

The objectives of T3.3:

- To build a Text to Speech Synthesis (TTS) system with appealing/affective voices, first for English and later for foreign languages.
- To enrich the TTS system voice's with affective attributes, such as control of a voice's emotion, or cloning the emotion of a real (user specified) voice, via emotion recognition models.
- To develop a novel Text & Video-to-Affective-Speech tool able to convert textual inputs, e.g. video subtitles or other textual descriptions / e.g. creations of T3.1 & T3.2, into speech and subsequently overlay this newly synthesized speech into the source video or an alternative visual stimuli, e.g. visually enhanced video by T2.3.

This report provides the technological background necessary for T3.3 implementation. It covers the implementation of Text-to-Speech (TTS) system with affective voices, enhancing videos with synthesised speech pronouncing synchronised subtitles or non-synchronous plain text and moves beyond TTS to fully immersive auditory experience via speech and background sound synthesis.

As there exist two dedicated deliverables, D3.2 and D3.6, addressing specifically T3.3. This report finalises the work of D3.2 for an Affective TTS & video tool.

1.2 STRUCTURE OF THIS DELIVERABLE

The structure of this deliverable is as follows:

1. Chapter 1: Introduction.
2. Chapter 2: Text & Video to Affective Speech Production Tool - Final Version, presents how we generate affective TTS voices, our novel Speech Emotion Recognition (SER) architecture that we developed to benchmark the Affectiveness of the SHIFT TTS tool. Overview of the tool, Overview of the >200 available affective voices and languages, Flask API interface description, and how-to usage by example.
3. Chapter 3: Soundscapes Synthesis: presents how we extended the SHIFT TTS tool to have, in addition to synthetic speech, the functionality to produce background sounds such as environmental landscape sound synthesis.
4. Chapter 4: Conclusion / Future possibilities.

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

2 TEXT & VIDEO TO AFFECTIVE SPEECH PRODUCTION TOOL

Truly appealing audio experience extends beyond just realism. The ability to imbue visual stimuli together with TTS of nuanced emotional expression, reflecting the affective character of the CH artifact represents a crucial frontier.

The SHIFT TTS tool addresses the necessity of developing on affective multilingual TTS system capable of generating pleasant / affective speech.

This Chapter is organised as follows. Section 2.1 Presents our novel method for raising the affectivity of SHIFT TTS voices by fine-tuning the voice-cloning algorithm of a SoTA TTS system, namely StyleTTS2. Also describes how we built 134 native-accent English affective voices as well as 196 Affective non-native (foreign) accent English voices that are now available for the English TTS. Afterwards, Section 2.1.2 describes how we introduced a Fallback system for non-English (foreign) languages TTS.

Noticing the need of a foreign TTS system arose as letters of Latin script has to be converted to different phonemes for English, or German or Serbian or Hungarian vocalisations. The fallback foreign language TTS component has one voice per language due to data scarcity for training. The voice is tuned for affectivity in the spirit of [4]. We also implement various character pronunciation tuning for Albanian, German, Greek, Hungarian, Romanian, Serbian. As those languages are needed from the CH SHIFT partners.

Section 2.2 presents publication [2] of our novel Speech Emotion Recognition (SER) system named wav2small that we develop for SHIFT and apply in Section 2.1.1 as evaluator of emotionality of SHIFT TTS. Wav2small enables quantitative evaluation of the SHIFT TTS system besides human listener evaluations. Section 2.3 describes the application interface Flask API of the SHIFT TTS tool. Finally, Section 2.4 presents the functionalities of SHIFT TTS.

2.1 AFFECTIVE TTS WITH MULTILINGUAL SUPPORT

In our previous deliverable, D3.2, we explored the state-of-the-art of TTS algorithms, whose speaker's voice is encapsulated / hard-coded as floating point weights within the neural network during its training phase. Thus a TTS voice could not be changed into a different voice/speaker/emotion and a full new neural network had to be designed for every TTS voice.

Now in D3.6 we advance the SHIFT TTS to be using a multi-voice (multi-speaker) neural network that allows changing / cloning / emotionalisation of its output voice through a short prompt .wav file (short recording <4s).

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

2.1.1 AFFECTIVE EXPANSION OF TTS VOICES

This Section presents a phenomenon and benchmarks of how we enhance a State of the Art TTS system namely StyleTTS2 to become more emotional. For core ingredient of the SHIFT TTS we chose StyleTTS2 as it is free open source and has paramount sound quality even on its non-affective open source original implementation - https://huggingface.co/spaces/TTS-AGI/TTS-Arena.

Let us start by demonstrating our Experimental evaluation of StyleTTS2 using audio prompts (speaker voices) generated by a second TTS system vs natural speech prompts of the EmoDB [6] dataset.

A **prompt** (.wav file) is interchangeably called also **style** or **voice** or **voice-cloning template** or **speaker reference**. A prompt has the ability to instruct emotions / speaker characteristics for StyleTTS2. In SHIFT TTS we employ Mimic-3 [5] 134 english voices (VITS neural networks) and 196 non-English voices to construct artificial prompts for StyleTTS2. We synthesize all 767 Harvard sentences [7] with styleTTS2 using Mimic-3 prompts and also using natural speech prompts [6][18] and produce Table 1.

Voice (Prompt .wav speech)			SHIFT TTS	
	nMOS	CER (%)	nMOS	CER (%)
Mimic-3 En.	2.9	0.92	3.6	0.72
Mimic-3 En. 4x	3.7	12.90	4.1	0.59
Mimic-3 Foreign	1.7	62.23	3.4	0.85
Mimic-3 Foreign 4x	2.7	82.67	4.0	1.13
EmoDB [6]	4.7	77.81	4.0	1.15

Table 1. Naturalness MOS and CER (%) of SHIFT TTS synthesising the Harvard sentences via human speech (EmoDB [6]) prompt vs via pre-generated synthetic voice prompt. Pre-generated voice prompt made by Mimic-3 of 1x or 4x speed.

One of the most audible differences between natural speech prompts and Mimic-3 prompts lies in the presence of inadvertent environmental noise and paralinguistic artifacts in natural speech prompts. If for example one uses a voice cloning prompt for StyleTTS2 and in this recording a dog barks in the background, then StyleTTS2 can ignore the speech and focus at cloning dog's voice characteristics!

To a vaster extent one can introduce purposeful distortions / audible artifacts on the voice prompt, such as pitch, or speed or timbre manipulation in order to drive StyleTTS2 to produce voice emotionalisation.

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

Let us define a Naturalness (n)MOS scale as in [3]: For nMOS=1 indicates a un-natural speech while nMOS=5 indicates perfectly natural speech. Natural speech recordings have nMOS = 4.7 (EmoDB prompts in Table 1). However such high, fully natural nMOS is non-transferable to the synthetic speech output because an internal HiFiGAN vocoder of StyleTTS2 only transfers up to a limited amount of voice characteristics as it preserves / balance voice clone fidelity vs speech intelligibility.

Altering TTS prosody can be done either through an audio file such as a voice cloning audio prompt or via a text prompt. Achieving high naturalness in speech synthesis using artificial (synthetic) voice prompts / speakers might seem counterintuitive: However, natural speech has environmental noises that adulterates voice cloning. Audio prompts allow for precise indication of prosody, on the other hand text prompting [8] is limited by non descriptiveness of non-verbal paralinguistics. SHIFT TTS focus is an in between zone: What if we use synthetic TTS voice prompts instead of natural speech prompts to drive the voice cloning of StyleTTS2.

Our intuition is that synthetic audio prompts are free of environmental artefacts. In specific we discover that synthetic voice prompts particularly when sped-up to 4x enhance the emotional valence and suppress unpleasant emotions [4], more than natural speech prompts. This audio-domain prompt engineering allows the SHIFT TTS tool to reach nMOS = 4.1 Table 1, audibly improving the baseline StyleTTS2 - nMOS = 4.0.

Our experiments consist of five setups for audio prompts for StyleTTS2 to tune its prosody. For benchmark text for synthesis we use Harvard Sentences [7]. As prompt input we use 5 experimental setups: **Mimic-3 English voices** of **1x speed** / **4x speed**, **Mimic-3 Foreign voices** of **1x** / **4x speed** also **natural speech** [6].

You can Listen Audios here https://huggingface.co/dkounadis/artificial-styletts2/discussions/2

Mimic-3 [5] is a TTS system of 134 English voices and 196 Foreign (Non-english) TTS voices. After testing various other TTS engines for voice cloning prompting We discovered that it provides interesting artifact free audio voice prompts that enable the internal vocoder of StyleTTS2 to improve its affectiveness. For benchmarking of SHIFT TTS we apply SER to observe the emotional profile of SHIFT TTS output over the duration of synthesizing the 720 Harvard sentences. We use Wav2small-Teacher [2] for Arousal, Dominance, and Valence (A/D/V) dimensional SER that we developed for SHIFT and present in Section 2.2. We use a WavLM model from categorical SER [9]. The categorical SER WavLM outputs probabilities for emotion categories: Happy, Anger, Sad, Fear, Disgust. As for A/D/V levels or dimensions of affect [10][17]. Arousal indicates voice excitement, dominance shows how imposing a voice is, and valence reveals perception pleasantness.

The Grey shadow of Fig. 1 / Fig. 2 shows the emotional profile of StyleTTS2 prompted by natural speech from EmoDB [6] and serves as baseline emotional profile. The (blue line) in Fig. 1 / Fig. 2 shows the emotional profile of StyleTTS2 prompted by our synthetic Mimic-3 engineered prompts. Fig. 2. shows also Anger probability of StyleTTS2 output (blue line) is always lower via Mimic-3 4x speed prompts than natural speech EmoDB prompts. Disgust probability is almost vanished using 4x speed prompts. Clearly showing the enhancement of affectiveness by our engineered voice prompts.

Different voices yield different duration that explains the slight mis-alignment of Grey and Blue lines. Interestingly, Happy emotion probability increases when 4x speed prompts are fed to StyleTTS2. Table

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

1. shows that StyleTTS2 achieves highest nMOS = 4.1 as well as improved Character Error Rate (CER) = 0.59% when using Mimic-3 English 4x speed TTS prompts. The high nMOS correlates with increase in valence and rise of Happy probability: Fig. 2. (blue line) vs (Grey line).

The voice prompts have low intelligibility, e.g. CER = 12.90% for Mimic-3 En. 4x speed, Table 1. However StyleTTS2 synthetic speech output is not affected by the intelligibility of voice prompt, as shown by achieving low CER = 1.15 % via foreign prompts, Table 1.

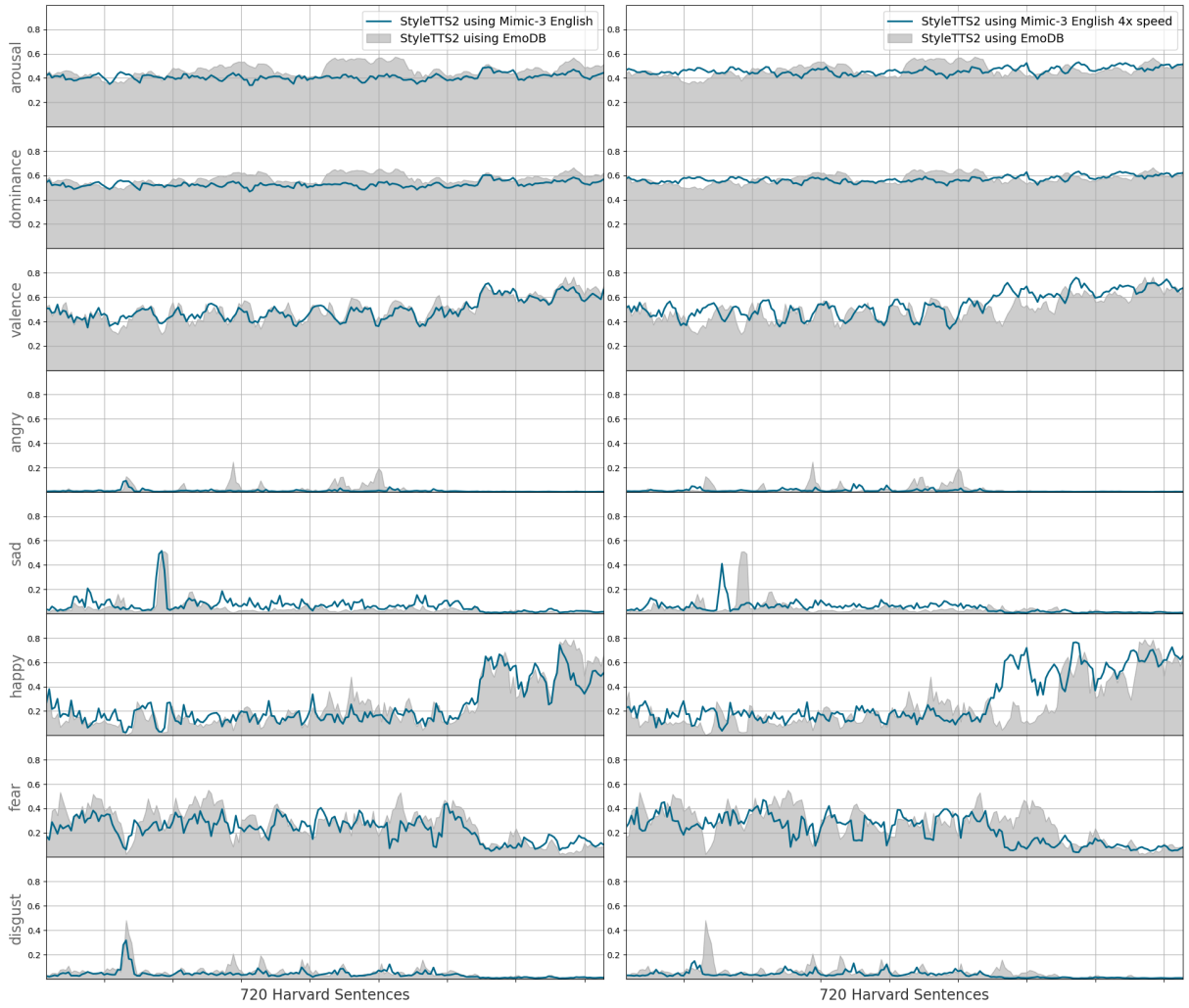


Fig 1. Emotion for the 720 Harvard Sentences synthesized by StyleTTS2 with style: BLUE LEFT: Mimic-3 (English voices) 1x speed. BLUE RIGHT: Mimic-3 (English voices) 4x speed. GREY: Human speech from EmoDB..

Audible backchannels are vocal cues that listeners use to show engagement at paying attention to a conversation. The probability of valence and Happiness in Fig. 1. / Fig. 2. shows that StyleTTS2 via Mimic-3 (blue line - 4x speed) is almost always higher than StyleTTS2 via EmoDB prompts (grey line), irrespective of the language of prompt. We discovered that the actual words in prompt-audio do not affect the intelligibility of StyleTTS2. However, extra punctuation placed in the text for prompt creation by Mimic-3 such as ``...!!!;" produces un-natural noises by Mimic-3 like ``hah/hiss/scratch``

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

sounds. Those sounds if observed by StyleTTS2 in its prompt, force the HiFiGAN to generate subtle emotional backchannels, such as sighs and breaths that are pleasant to hear. Hence we believe that voice cloning backchannels stylisation via artificial prompts is an open avenue for prosody in TTS.

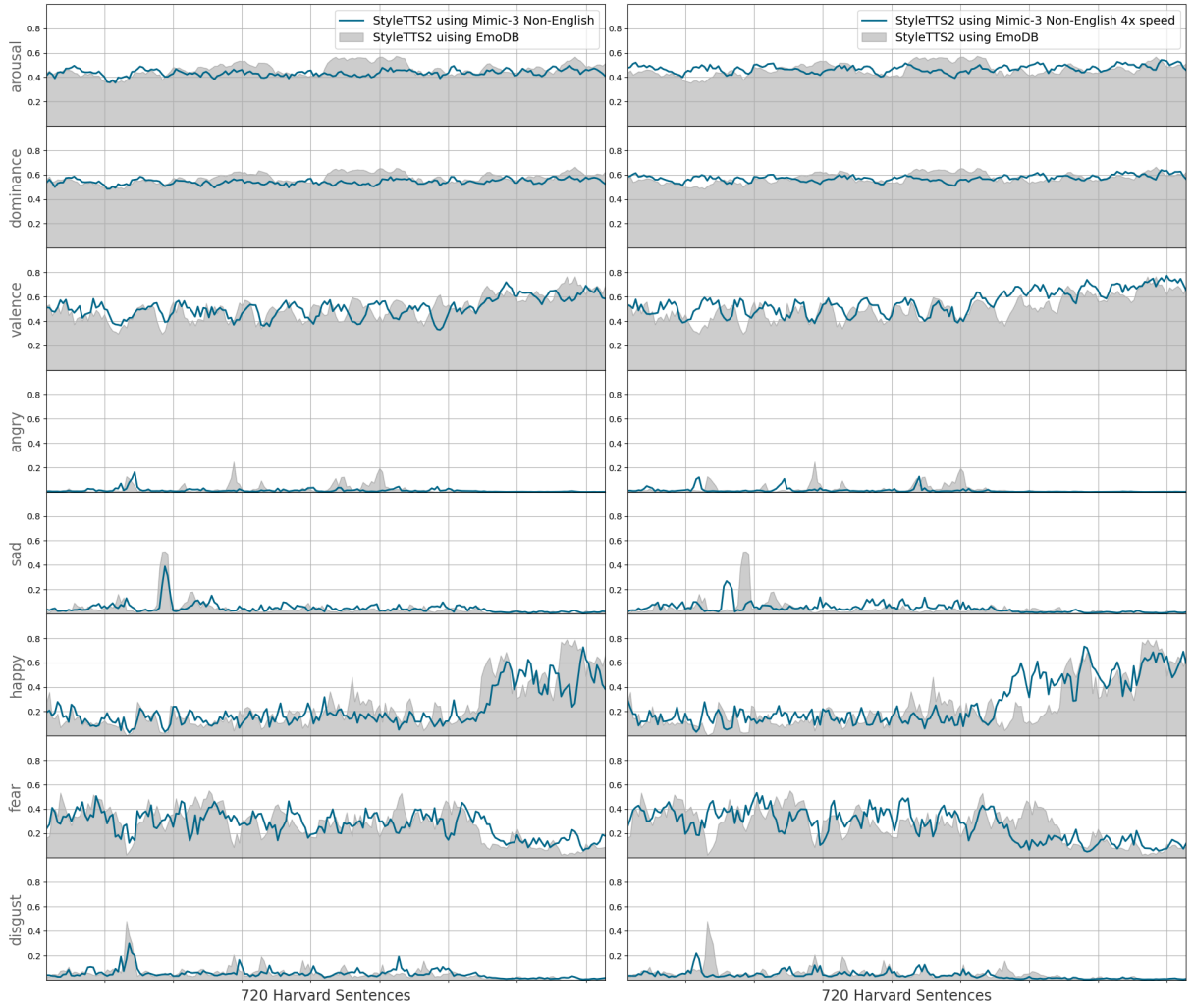


Fig 2. Emotional Valence and Happy probabilities of StyleTTS2 rise by using 4x speed artificial TTS styles (style/voice is only the prompt reference wav that stylises the final speech. A voice acts as indicator of pitch/speed/speaker/emotion internally in the neural network. The neural network however has been trained to acquire viable acoustic characteristics from this voice template although always transforming those characteristics to a natural speech output, thus not fully preserving the voice template/identity.

After benchmarking the affectiveness of StyleTTS2 for various voices / prompts. We discovered that synthetic voices pre-build by Mimic-3 amplify positive valence and happy emotion in the output of StyleTTS2 beyond natural speech prompts. In [4] an initial link between emotion and pitch/speed has already been discovered. For D3.6 we brought this knowledge internally to HiFiGAN by accelerating (4x speed) synthetically crafted voice prompts to drive StyleTTS2 for prosody enhancement.

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

2.1.2 FALLBACK MULTILINGUAL TTS ENGINE

In Sec. 2.1.1 we presented StyleTTS2 being a component of SHIFT TTS for English TTS where we crafted >200 affective voices for a CH users experience.

Here we explain how we incorporate a component for Non-English text TTS functionality for the SHIFT TTS system. in an effort to cover use-cases for the Non-english CH partner's languages and non English speaking visitors. CH partners of SHIFT are from 9 different native languages, namely Albanian, Bosnian, German, Hungarian, North Macedonian, Romanian, Serbian, Greek. Our aim is that SHIFT TTS supports those languages.

As balkan region languages have very low amounts of data available freely to train a new TTS model. We opt for using MMS TTS [12] that provides open source VITS TTS models of single speaker / voice per language and enables TTS for the 9 aforementioned languages.

We had to finetune MMS TTS to be able to pre-process input text / CH description to be brought in a phonetically correct vocalisation form. That includes for example re-writing of numbers e.g. `123' as `сто двадесет три' or `sto dvadeset tri'.

Despite the inherent challenges of low-resource language TTS such as Balkan languages, SHIFT TTS now successfully provides a framework for affective narration as well as video functionalities across diverse linguistic landscapes.

Users can effortlessly choose their preferred language or affective voice via a unified API argument (**--voice**) through the SHIFT TTS Flask API interface, detailed in Section 2.3.

MMS TTS provides TTS synthesis for Serbian-Carpathian group of language / phonemes. We have enhanced the text pre-processing in the SHIFT TTS tool and opt for Serbian accent after starting from the official text processor of MMS TTS for phoneme selections, although our system still has edge cases where the final speech output perhaps sounds as a blend of another Balkan region neighbor language such as Bosnian / Herzegovinian or Croatian. For other CH partner's native languages we discovered less pronunciation blending as Albanian, Hungarian, Greek, Romanian, German are far dissimilar phonetically.

Nonetheless we believe even for this difficult case of scarce speech training audio data from Balkan region Languages we provide a first functionality tuned for affective narration.

2.2 ADVANCING THE STATE OF THE ART OF SPEECH EMOTION RECOGNITION

Speech Emotion Recognition needs high computational resources to overcome the challenge of substantial annotator disagreement. Today, SER is shifting towards dimensional annotations of arousal, dominance, and valence (A/D/V) [10]. Instance-level measures as the L2 distance prove unsuitable for evaluating A/D/V accuracy due to volatile / non-converging consensus of annotator opinions. Concordance Correlation Coefficient (CCC) arose as an alternative measure where A/D/V predictions are evaluated to match a whole dataset's annotations in terms of CCC rather than

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

focusing in exact prediction of A/D/V annotations of individual audios. Recent studies have shown that Wav2Vec2 [11] / WavLM [9] architectures achieve today's highest CCC. However the Wav2Vec2 / WavLM family has a high computational footprint.

For the SHIFT project we focused on research of a novel SER A/D/V model that we used for building Affective TTS voices in Sec 2.1.1. Our novel SER A/D/V model is presented in our paper [2].

In [2] we developed a State of the Art A/D/V model that as of February 2025 holds the current Top-1 position in https://paperswithcode.com/task/speech-emotion-recognition A/D/V Leaderboard.

See **wav2small-teacher** in Fig. 4. Our paper also delves on a parallel research for low resource A/D/V SER models for Mobile devices / low resource hardware.

Our novel A/D/V model set a new high for the benchmarks of CCC in Arousal Dominance and Valence on the MSP Podcast dataset [13]. The MSPPodcast dataset is chosen as our focus for A/D/V development as it is one of few A/D/V public datasets that is collected in the wild from podcasts, paving away from controlled / acted emotion audio recordings.

The conceptualization of human emotion in a 3D space, assuming Arousal, Dominance, Valence as dimensions [10] provides a continuous space for algorithmic recognition of emotions that resolves limitations of finite set of emotional categories, e.g. Angry, Happy, ..

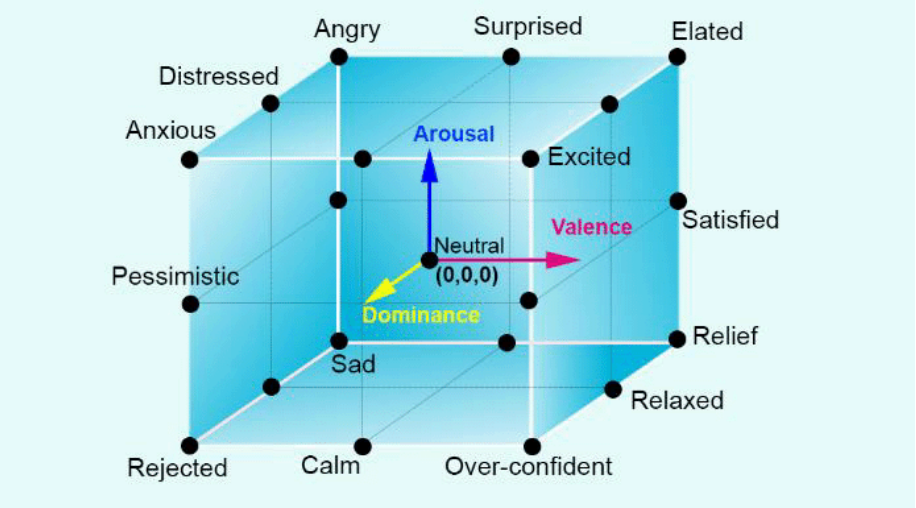


Fig 3. Representing emotions in the three dimensions of arousal, valence, and dominance [10].

For the assessment of uncertainty of annotation agreement for training / benchmarking of SER audio labels we have developed two additional investigations during the course of SHIFT in publications [14][15].

Funded by the European Union

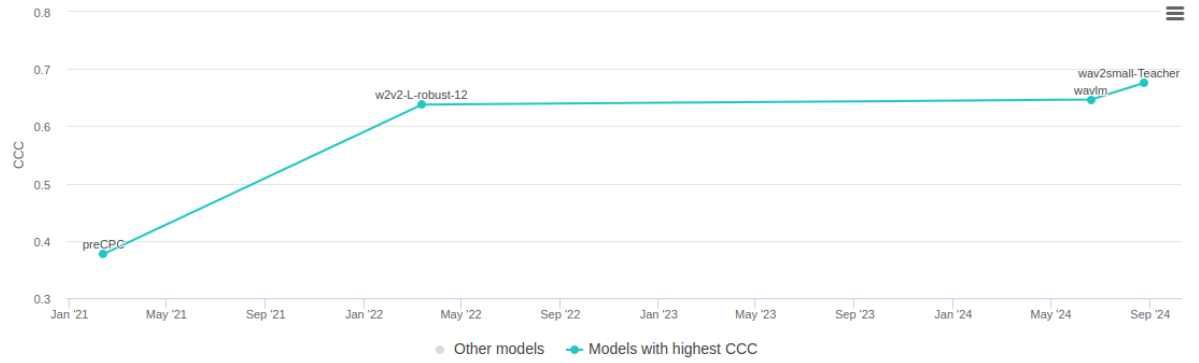
This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

Speech Emotion Recognition

Speech Emotion Recognition on MSP-Podcast (Valence)

Leaderboard Dataset

View CCC by Date for All models



Filter: untagged

Edit Leaderboard

Rank	Model	CCC	Extra Training Data	Paper	Code	Result	Year	Tags
1	**wav2small-Teacher**	0.676	×	Wav2Small: Distilling Wav2Vec2 to 72K parameters for Low-Resource Speech emotion recognition			2024	
2	**wavlm**	0.6466753	×	Odyssey 2024 - Speech Emotion Recognition Challenge: Dataset, Baseline Framework, and Results			2024	
3	**w2v2-L-robust-12**	0.638	×	Dawn of the transformer era in speech emotion recognition: closing the valence gap			2022	
4	**preCPC**	0.377	✓	Contrastive Unsupervised Learning for Speech Emotion Recognition			2021	

Fig 4. SER results for [2] from https://paperswithcode.com/task/speech-emotion-recognition Valence Leaderboard.

MSP-Podcast (Valence)	wav2small-Teacher			See all
MSP-Podcast (Activation)	wav2small-Teacher			See all
MSP-Podcast (Dominance)	wav2small-Teacher			See all

Fig 5. Our Proposed wav2small-teacher [2] SER model developed for SHIFT wins in Arousal/Dominance/Valence benchmarks https://paperswithcode.com/task/speech-emotion-recognition.

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

2.3 FLASK API

The SHIFT TTS tool is accessible through a Flask API for integration with various applications. The Flask API provides a command line interface (CLI) for generating speech from text, and/or generating videos. This API enables multiple possibilities exposed to the user via 7 arguments:

Flask API input arguments : **--image, --text, --video, --voice, --soundscape, --affective, --native**

Those arguments allows a user / CH professional to select Affective voice / language by specifying the `--voice' argument with possibilities shown in Fig. 5. The Flask API serves as an interface between HTTP Requests and the SHIFT TTS engine. The SHIFT TTS engine runs on a Linux system for having access to libraries as ffmpeg & CUDA. The API uses HTTP requests, making it compatible with web demo / applications. The client interface `**tts.py**' delegates the computation to the TTS engine is presented via examples in Section 2.4.

As shown in Fig. 6. the first setup starts the SHIFT TTS engine (**api.py)** on the server machine. The client machine does not require ffmpeg / CUDA library / GPU, only needs Python 3.10 to call the client interface.

A request to API involves sending a POST request with a JSON payload containing the `**--text**' to be synthesized, the desired `**--voice**' (selects voice / language / accent as shown in Fig.5) and optional parameters `**--imag**' specifying input image for video background, or `**--video**' for video dubbing / audio track replacement. If any of `**--imag'** or `**--video**' is specified then the output is **.mp4** video otherwise the output is a **.wav** file of TTS. The argument `**--soundscape**' is described in Chapter 3.

The SHIFT TTS tool also has an experimental parameter `**--native'** where user can provide a .wav file with a speech audio to be cloned as speaker's voice in the TTS output. This argument overrides the `--voice' that selects one of the build in affective voice of the system. This feature is experimental as it may capture background environmental noises to convey in TTS instead of the user's speech signature.ently we planned for two distinct ways to help us detect the intended emotion, and we have implemented the 1st. The `--native' arg If provided acts as voice template. The native .wav can be specified by the user or is taken by the video if it has an audio track (and is not only silent motion sequence).

The SHIFT TTS tool is available in https://github.com/audeering/shift/. The has also an experimental option `--native' where users can provide a .wav file with a speech audio to be cloned as speaker's voice in the TTS output. This argument overrides the `--voice' option that choses one of the >200 pre-built affective TTS voices. This feature is experimental as it may capture background environmental noises to convey in TTS instead of the user's speech/emotion signature. If a video is also passed as input via the (--video) option then the audio track of the video is used instead (if not silent motion sequence) as --native .wav voice template. The SHIFTs Affective TTS tool is available in https://github.com/audeering/shift/. Fig. 6. depicts a sample of the available affective TTS voices along with their emotion volatility.

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

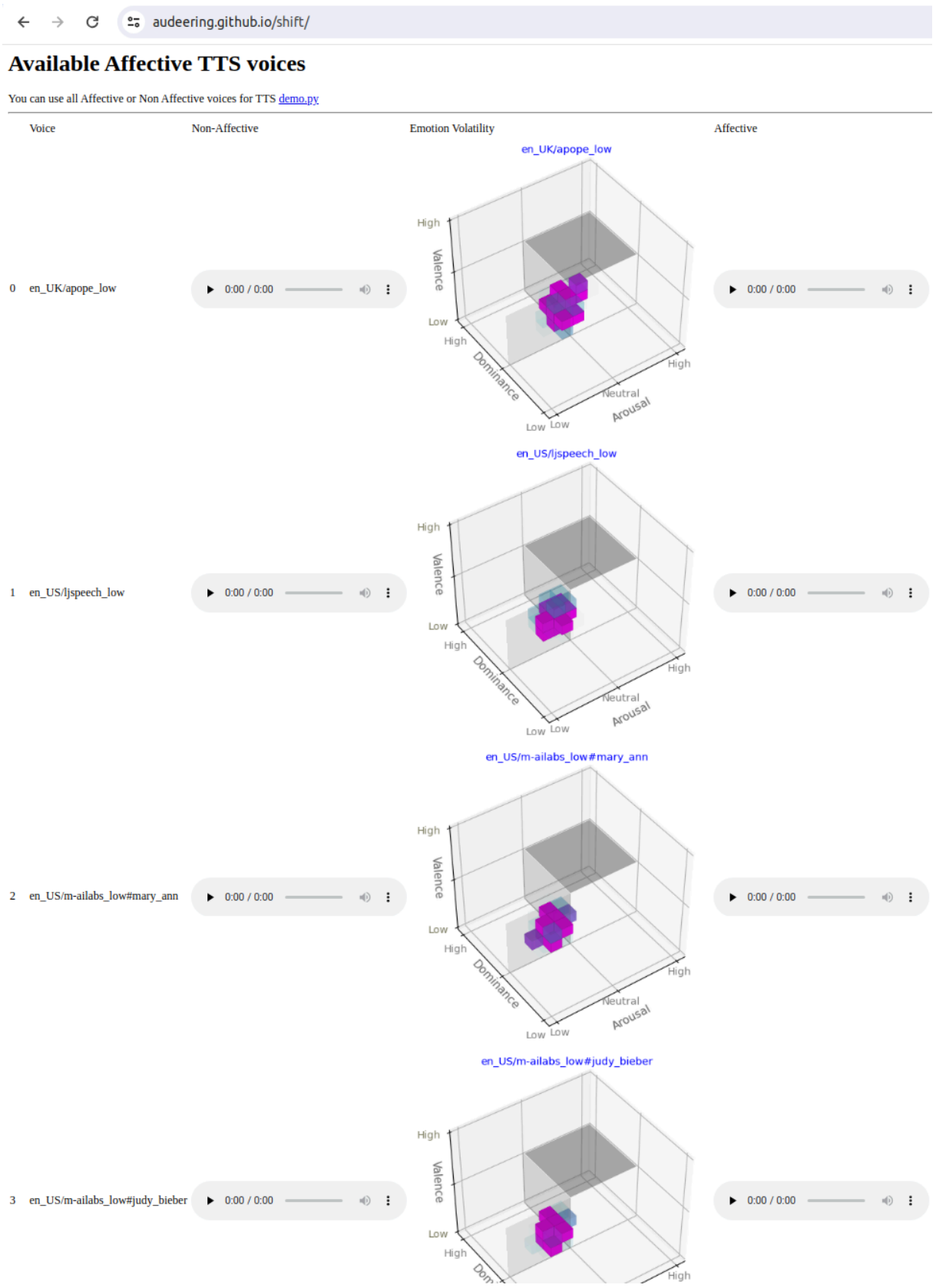
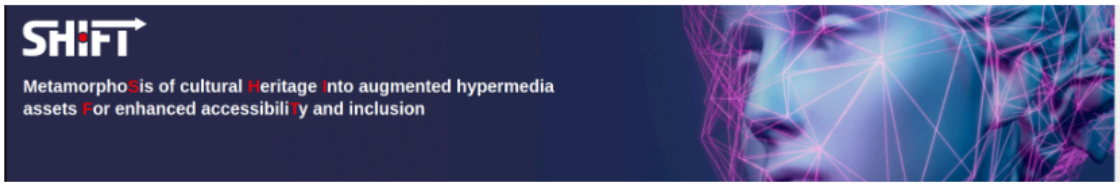


Fig 6. Available Affective TTS voices along with emotional span - https://github.com/audeering/shift

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

README



AFFECTIVE TTS using this phenomenon. Synthesize speech from `.txt` or `.srt` and overlay it to videos / picture.

- Has 134 affective voices for English tuned for StyleTTS2. Supports single-voice foreign languages TTS via MMS.
- A Beta Version of this tool for TTS & audio soundscape is build here

Available Voices

Listen to available voices!

Install

```
virtualenv --python=python3 ~/.envs/.my_env
source ~/.envs/.my_env/bin/activate
cd shift/
pip install -r requirements.txt
```

Demo. Result saved as `demo_affect.wav`

```
CUDA_DEVICE_ORDER=PCI_BUS_ID HF_HOME=./hf_home CUDA_VISIBLE_DEVICES=0 python demo.py
```

API

Flask `api.py` on a `tmux-session`

```
CUDA_DEVICE_ORDER=PCI_BUS_ID HF_HOME=./hf_home CUDA_VISIBLE_DEVICES=0 python api.py
```

Inference

Following examples need `api.py` to be running. Set this IP to the IP shown when starting `api.py`.

Text To Speech

```
# Basic TTS - See Available Voices Above - saves .wav in ./out
python tts.py --text assets/LLM_description.txt --voice "en_US/m-ailabs_low#mary_ann"
```

Listen to Various Generations

Fig 7. Install Guide for - https://github.com/audeering/shift .

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

2.4 VIDEO FUNCTIONALITY

The Flask API functionality expects users to provide a **.txt** or **.srt** file, that serves as text for TTS. It may be plain (**.txt**) or time-stamped (**.srt**). The SHIFT TTS tool looks at `--**voice'** argument that selects language and/or voice for TTS from the >200 possible voices. Additionally the system enables optional arguments. For example if a user provides also an image `**--imag pic.jpg**' the system creates a **.mp4** video with TTS and the image as background. Further options available is `**--video vid.mp4**' specify to use **vid.mp4** as input and dub its audio-track by TTS, then if the user passed a synchronous `**.srt**' text then TTS is synchronised accordingly. If non synchronous plain text `**.txt**' is provided then the TTS sentences follow one-another without special silence synchronisation.

2.4.1 VIDEO DUBBING via TIMESTAMPS / SUBTITLES

We can transfer the textual content of a video to another language / TTS voice by calling the tool of SHIFT deliverable D3.5 for translating the original text to the desired language, e.g. Romanian to English, then call SHIFT TTS tool to build a video with TTS voice dubbing as follows:

Fig 8. From Native Romanian video of ANBPR - https://www.youtube.com/watch?v=tmo2UbKYAqc - we replace the native Romanian voice with English TTS voice - https://www.youtube.com/watch?v=geI1Vqn4QpY - preserving emotion.

After having started the server **api.py** as described in Sec. 2.3. From client side run this command :

python tts.py --voice "en_US/m-ailabs_low#mary_ann" --video assets/anbpr.webm --text assets/anbpr.en.srt

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

This will generate the video in Fig 7. All example files, e.g. **assets/anbpr.webm** or other files are downloaded by git clone https://github.com/audeering/shift.

Fig. 7 is an example use case of videos only available in Romanian with auto generated English text description that is here used by the SHIFT TTS tool to synthesize English speech and overlay to the original video.

The above functionality is applicable also for overlay TTS audiotrack on a silent / motion-sequence video, such as those produced by WP2 tools, with or without synchronous text / subtitles.

2.4.2 AUDIO ONLY TTS (.wav)

A basic however useful functionality of SHIFT TTS is to generate .wav audio files. Those may be used for example in the haptic Interface. For this functionality no image / video input is provided.

Text To Speech

```
# Basic TTS - See Available Voices Above - saves .wav in ./out
python tts.py --text assets/LLM_description.txt --voice "en_US/m-ailabs_low#mary_ann"
```

Fig 9. Synthesize output.wav 24 kHz TTS from llm_description.txt

2.4.3 NATIVE VOICE TO TTS (via Voice Cloning)

An optional functionality of SHIFT TTS is transferring the speaker style through voice cloning of the original/native voice of video or native audio .wav to the TTS (for a new text). This is an alternate option to using our >200 pre-build Affectively tuned voices of Section 2.1.

Text 2 Speech

```
# Basic TTS - See Available Voices
python tts.py --text sample.txt --voice "en_US/m-ailabs_low#mary_ann" --affective

# voice cloning
python tts.py --text sample.txt --native assets/native_voice.wav
```

Fig 10. Using the optional argument –native for voice cloning and –affective that choose a high-clarity variant of the selected voice to enhance intelligibility.

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

2.4.4 IMAGE TO SPEECH (via TXT)

One of the core use cases of SHIFT is automating the appealing exposition of visual content to people with visual impairments. For a provided image, as in Fig. 10., by our ANBPR CH partner. There exist a multiverse of options of possible text descriptions. SHIFT has allocated T3.1 T3.2 and WP4 for curation/generation of text descriptions of visual content. Here we implement the final step where from a provided image and a text description we generate a video with TTS speech. We apply the tools of T3.1 T3.2 WP4 to generate text descriptions for this image using various LLM instructions, such as for different language or age groups.

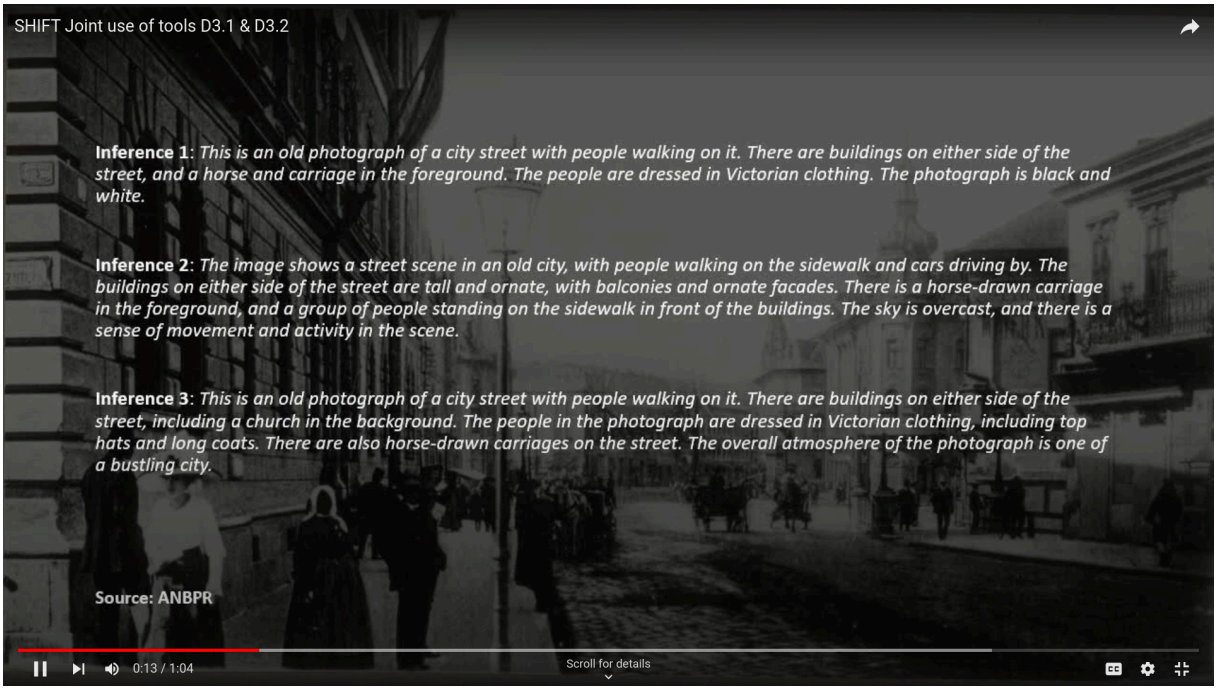


Fig 11. For this ANBPR image we call D3.1 tool to generate text captions (3 different inferences) and call D3.2 tool to synthesise Affective speech and build a video - https://youtu.be/wWC8DpOKVvQ

The user runs the following CLI command to produce the video in Fig.10 narrating an image :

python tts.py --text assets/LLM_description.txt --image assets/image_from_T31.jpg --voice "en_US/cmu-arctic_low#jmk"

Notice that --voice argument can be chosen from the many available affective voices & languages available in https://github.com/audeering/shift/.

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

2.4.5 AUDIOBOOK

During the discussions with the CH partners an interesting use case arose: How can we have audiobook vocalisation functionalities in SHIFT TTS tool. From technical side the difficulty was to parse/filter the textual input, e.g. book.docx as how to split a long form text into sentences / text-chunks that a TTS can synthesize on consumer grade hardware and seamlessly concatenate together to form a long audio file having natural sounding continuation between the multiple short TTS segments.

This functionality is now available in the SHIFT TTS tool by calling this python script:

AudioBook

Audiobook `.wav` file will be saved in `./tts_audiobooks/`. Listen to it in YouTube!

```
# Narrate assets/INCLUSION_IN_MUSEUMS_audiobook.docx
python audiobook.py
```

A generated audiobook is now available online on the SHIFT webpage / YouTube shown in Fig. 18. However a user can call audiobook.py to easily choose another voice / language / .docx file.

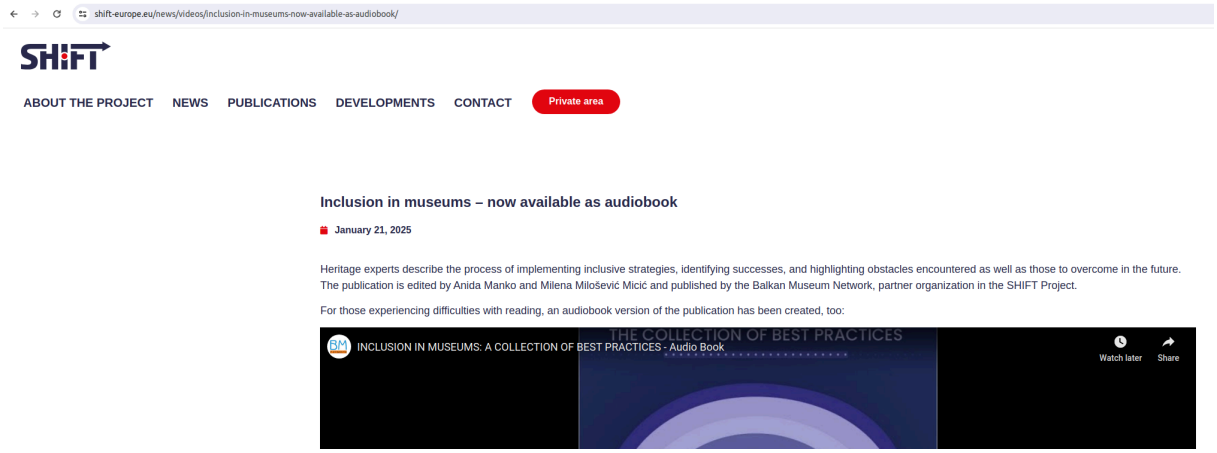


Fig 12. AudioBook file.docx converted to speech via the SHIFT TTS tool - https://www.youtube.com/watch?v=yy6fxV2DLC8

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

3 SOUNDSCAPES SYNTHESIS

Crafting Sonic Worlds together with Affective Multilingual TTS generation and also realistic soundscapes is a challenging endeavor. Recently, significant strides have been made in this domain, with AI driven models achieving remarkable feats in synthesizing complex audio environments. Among these cutting-edge tools, AudioGen [1] stands out as an open-source freely available state-of-the-art model, demonstrating a powerful capacity to translate text description to immersive soundscapes. For D3.6 by leveraging the SHIFT TTS system of Section 2 together with the pioneering implementation for soundscape synthesis of AudioGen we push the boundaries for truly captivating and emotionally engaging audio experiences.

3.1 SOUNDSCAPE GENERATIONS VIA AUDIOGEN

SHIFT CH partners realised early that TTS audio narrations could be enhanced by immersive sounds aiding landscape/painting description. For the final implementation of SHIFT TTS tool we added an experimental component for landscape (text description) to soundscape synthesis. AudioGen [1] is an open source auto-regressive language model that predicts sound tokens instead of typical text tokens in, e.g. chatGPT.

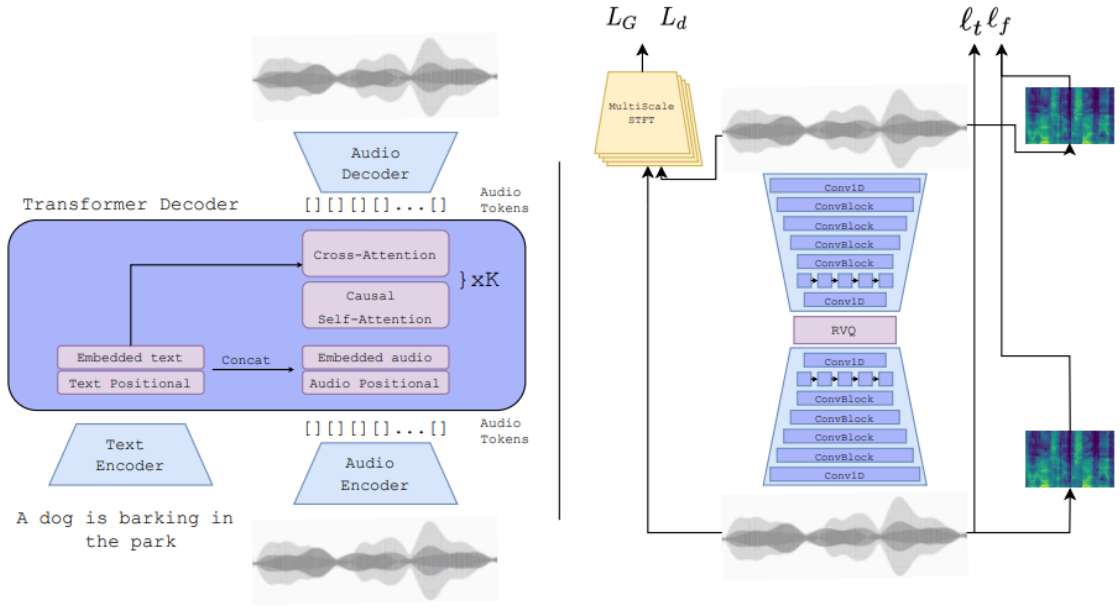


Fig 13. AudioGen architecture: From text description (or audio template) LEFT. AudioGen RVQ tokenizer RIGHT.

Fig. 10 shows the full AudioGen architecture where some text description input (or an audio excerpt input) conditions the sound token output probabilities via a cross-attention operation where input

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

condition, after encoding into a latent representation [11] acts as `key' and `value' for the attention while the output sound token logits act as `query'. The predicted sound tokens are finally interleaved with delays and inverted to time-domain soundscape waveform via a vocoder from [12].

For the finalisation of the SHIFT TTS tool we realised the need to have environmental background sounds to be played together with the TTS Speech descriptions to immerse CH visitor's experience. Beginning from [1] we crafted various accelerations to its python code for fast sound elongation. Due to the high computational resources required by AudioGen we have invented a few speed-up tricks for our SHIFT TTS implementation in order to accelerate the synthesis of long soundscapes to convey the full audio track of TTS speech. Our implementation of AudioGen employs cyclic/shift/cloning and sampling multiple best sound tokens from the output token probabilities of AudioGen.

The AudioGen pre-trained neural network has the capacity to generate various natural sounds far beyond the need for perfect description of the sounds. Its vast usefulness appears from its ability to understand misspellings for the text description as well as multilingual text descriptions for soundscapes and convey them in producing even more surprising sound variations than obtained from perfect orthographic text description.

3.2 FLASK API FOR SOUNDSCAPES

To have a simple user interface of the SHIFT TTS tool we enhance the Flask API presented in detail in Section 2.4 with one additional optional user input parameter **`--soundscape'**. With the support of the SMB CH partner we have been provided with paintings and textual descriptions. By starting the Flask API as described in Section 2.3. The user can now generate video with TTS and soundscape from an image and its text description by running

python tts.py --image example.jpg --text example.txt –soundscape "wind storm humid"

The above command will generate the following video Fig. 11.

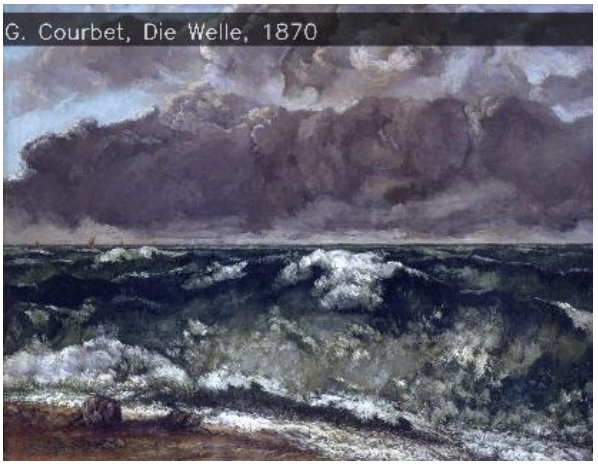


Fig 14. Video from an image / text description / soundscape - https://youtu.be/56MH7zOHrNQ)

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

For various examples and how to use the SHIFT TTS tool & soundscape argument please visit - ttps://github.com/audeering/shift/

Here we present some further generations from the SMB use case of the landscapes to soundscapes:

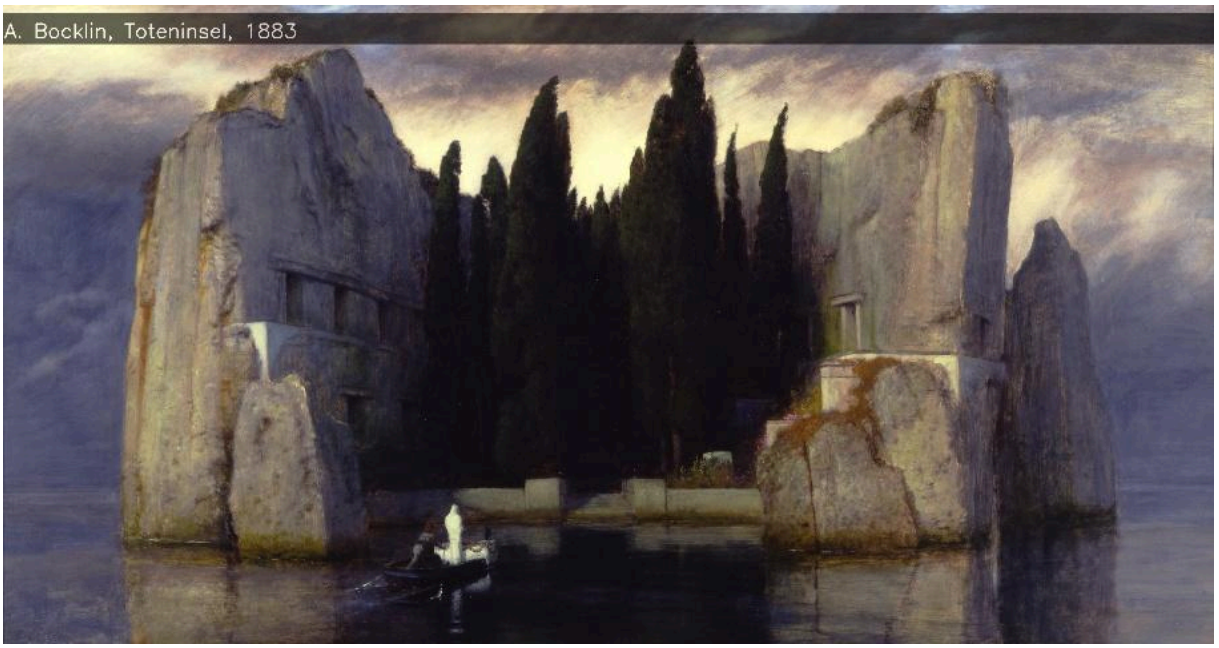


Fig 15. Generated video by SHIFT TTS from image and text and associated soundscape description - https://www.youtube.com/watch?v=Y8QyYUgLaCg

It is interesting for us to raise awareness for the non-determinism present in AudioGen generations. In fact repeated runs of AudioGen may produce slightly different soundscapes however all will be faithfull to the text description **`--soundscape'**.

Fig 16. Another example video created by the SHIFT TTS - https://youtu.be/BhMh02knkco

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

Notice the effect of TTS voice interwoven with soundscape giving rise to environmental background noises although without distortion of the TTS voice cloning, as tuned for in Section 2.1.

Fig 17. Video generation by the final SHIFT TTS tool - https://youtu.be/SSi3gUO4GtY

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

4 CONCLUSION AND FUTURE STEPS

The use of TTS technology in describing CH assets is crucial in today's digitalised world, aligning with trends in Virtual/Augmented Reality enhances visitor engagement in CH exhibitions.

The SHIFT Text and Video to Affective Speech production tool has now reached the potential of Affective multilingual TTS blended together with soundscape synthesis from text description inputs. It offers various functionalities including video/image dubbing with Multilingual Affective TTS, speech overlay, and landscapes vocalisation.

As the future unveils and unified neural networks move towards holistic universal audio visual language models [16] that soon will be capable of producing any sound & speech from any input modality such as text or vision. We hope that SHIFT TTS will still persist as a valuable, low computational demand, tool with apt soundscape quality and as well as TTS speech clarity thanks to its two path split design: Independently producing Affective TTS speech and soundscape sound.

By representing emotions through dimensions of valence, arousal, and dominance [10][17], we discovered and benchmarked all SHIFT TTS voices. This continuous domain of emotions especially through its valence dimension enabled disentangling the speech's emotion that arises from text phrases/words and from voice tonality [24].

All TTS voices available on our SHIFT TTS tool are artificially generated as presented in Section 2. The original audio used to train the SHIFT TTS underlying neural networks are publicly available datasets [19] [20] [21] [22] [23]. Therefore we believe SHIFT TTS to not infringe upon voice identity privacy.

This report has presented the final version of the Text and Video to Affective Speech system for SHIFT. This accomplishment stems from the integration of foundational elements from the TTS domain and the SER field. The synergetic fusion of these domains leverages scientific advancements, notably incorporating insights from audEERING (AUD) on a novel powerful Speech Emotion Recognition system [2]. As well as AUDs discoveries on linking emotions with prosodic elements [4], together with tuning voice templates to create >200 available TTS voices in SHIFT TTS tool. Lastly blending with audio soundscapes creates a fully immersive experience.

For as next step we are devoting the remaining months of SHIFT for pilot exhibitions for showcasing examples of audio visual content created by the SHIFT TTS tool.

The CH partners of SHIFT collected opinions from partially-sighted people regarding the ideal SHIFT TTS system. The participants comments told us they would like to have a proportion of sound effects together with text description that will transfer and present the general atmosphere of the images. To carefully select the appropriateness of audio details to support the narration as well as symbolic and/or factual "landscape" meaning of a specific scene or image. AudioGen addresses a need for factual soundscapes as well as symbolic sound effects through its latent encoding of the text input. Furthermore its ability to have Multilingual input and even some words from one language and other words from different language enable the SHIFT TTS tool user to tune the appropriate soundscape per use case. The SHIFT TTS is based on Adaptive vocoder via speaker embeddings that were fine tuned to imbue affective elements rather than speaker identity.

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

REFERENCES

[1] F Kreuk, G Synnaeve, A Polyak, U Singer, A Défossez, J Copet, D Parikh, Y Taigman, Y Adi, "AudioGen : Textually guided audio generation.", arXiv:2209.15352, 2022.

[2] D Kounadis Bastian, O Schrüfer, A Derington, H Wierstorf, F Eyben, F Burkhardt, B Schuller, "Wav2Small: Distilling Wav2Vec2 to 72K parameters for low-resource Speech emotion recognition.", arXiv:2408.13920, 2024.

[3] YA Li, C Han, V Raghavan, G Mischler, N Mesgarani, "Styletts 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models" Advances in Neural Information Processing Systems 36, 2024.

[4] F Burkhardt, U Reichel, F Eyben, B Schuller, "Going Retro: Astonishingly Simple Yet Effective Rule-based Prosody Modelling for Speech Synthesis Simulating Emotion Dimensions.", ESSV, 2023.

[5] https://github.com/MycroftAI/mimic3.

[6] F Burkhardt, A Paeschke, M Rolfes, WF Sendlmeier, and B Weiss, “A database of german emotional speech.,” Interspeech, 2005.

[7] Authors: IEEE, "Recommended practice for speech quality measurements.", IEEE Trans. on Audio and Electroacoustics. APPENDIX C 1965 Revised List of Phonetically Balanced Sentences (Harvard Sentences). DOI 10.1109/IEEESTD.1969.7405210, 1969.

[8] Z Du, Y Wang, Q Chen, X Shi, X Lv, T Zhao, Z Gao, Y Yang, C Gao, H Wang, F Yu, H Liu, Z Sheng, Y Gu, C Deng, W Wang, S Zhang, Z Yan, J Zhou, "Cosyvoice-2 : Scalable streaming speech synthesis with large language models.", arXiv preprint arXiv:2412.10117, 2024.

[9] L Goncalves, A N Salman, A R Naini, L M Velazquez, T Thebaud, L Paola Garcia, N Dehak, B Sisman, and C Busso, “Odyssey 2024- speech emotion recognition challenge: Dataset, baseline framework, and results,” in Odyssey 10 (9,290), 4-54, 2024.

[10] J R J Fontaine, K R Scherer, E B Roesch, P C Ellsworth, "The World of Emotions is not Two-Dimensional.", Journal of Psychol. Sci., 2007.

[11] J Wagner, A Triantafyllopoulos, H Wierstorf, M Schmitt, F Burkhardt, F Eyben, B W Schuller, “Dawn of the transformer era in speech emotion recognition: Closing the valence gap,” IEEE Trans. Pattern Anal. Mach. Intell., 45 (9), pp. 10745-10759, 2024.

[12] V Pratap, A Tjandra, B Shi, P Tomasello, A Babu, S Kundu, A Elkahky, Z Ni, A Vyas, M Fazel-Zarandi, A Baevski, Y Adi, X Zhang, W N Hsu, A Conneau, M Auli. "Scaling speech technology to 1,000+ languages", Journal of Machine Learning Research 25 (97), 1-52, 2023.

Funded by the European Union

This project has received funding from the European Union’s Horizon Europe research and innovation programme under grant agreement no 101060660.

[13] R Lotfian, C Busso, “Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings,” IEEE Transactions on Affective Computing, vol. 10, no. 4, pp. 471–483, 2019.

[14] O Schrüfer, M Milling, F Burkhardt, F Eyben, B Schuller, "Are you sure? Analysing Uncertainty Quantification Approaches for Real-world Speech Emotion Recognition", in Proc. INTERSPEECH 2024.

[15] F Burkhardt, BT Atmaja, A Derington, F Eyben, BW Schuller, "Check Your Audio Data: Nkululeko for Bias Detection", In Proc. of Conference of COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques, 2024.

[16] J Zhan, J Dai, J Ye, Y Zhou, D Zhang, Z Liu, X Zhang, R Yuan, G Zhang, L Li, H Yan, J Fu, T Gui, T Sun, "AnyGPT Unified multimodal LLM with discrete sequence modeling.", Y Jiang, X Q, arXiv preprint arXiv:2402.12226, 2024

[17] H Schlosberg, "Three dimensions of emotions". Psychological Review, 61, pp. 81–88, 1954.

[18] H Cao, DG Cooper, M. Keutmann, R.C. Gur, A. Nenkova, R Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset", IEEE Trans. Affective Computing, vol 5, n 4, pp. 377-390, 2014.

[19] https://paperswithcode.com/dataset/m-ailabs-speech-dataset, "M-AI Labs Dataset"

[20] https://datashare.ed.ac.uk/handle/10283/2950, "VCTK Dataset"

[21] https://www.openslr.org/32/ "Google NorthWestern University Dataset"

[22] https://github.com/google/language-resources "Google Language Res. Dataset"

[23] http://festvox.org/cmu_arctic/ "CMU Arctic Speech Synthesis Databases"

[24] M Amin, E Cambria, BW Schuller, "Will Affective Computing Emerge From Foundation Models and General Artificial Intelligence? A First Evaluation of ChatGPT" , IEEE Intelligent Sys. 38 (2), 2023.

Funded by the European Union

This project has received funding from the European Union’s Horizon Europe research and innovation programme under grant agreement no 101060660.