VPS VIBRATIONS IN PHYSICAL SYSTEMS

DOI: 10.21008/j.0860-6897.2024.1.12
Vibrations in Physical Systems, 2024, 35(1), 2024112

Cloning the voice and speech of Piotr Fronczewski for Polish speech synthesis

Krzysztof SZKLANNY

Multimedia Department, Polish-Japanese Academy of Information Technology, 02-008 Warsaw, Poland, Koszykowa 86

Corresponding author: Krzysztof Szklanny, email: kszklanny@pjwstk.edu.pl

Abstract The quality of synthetically generated speech has improved significantly in recent years, largely due to the technological development of speech synthesis systems, in particular those based on deep neural networks (DNN). However, the problem of emotion in speech synthesis still remains a challenge. Most of the existing speech synthesis systems do not convey the pervasive emotional contexts in human-human interaction. The lack of expression limits the emotional intelligence of current speech synthesis systems. This work aimed to develop a recording method for preparing a balanced corpus of emotional recordings in the Polish language for use in speech synthesis based on artificial intelligence (AI) algorithms. An essential aspect of the work was the selection of a voice-over artist who would allow the recording of the spectrum of an actor's voice, emphasizing the actor's interpretations and emotions derived from the content. Outstanding actor Piotr Fronczewski was chosen for the role.

Keywords: voice, emotions, corpus, recordings, speech synthesis, Piotr Fronczewski.

1. Introduction

Speech synthesis, commonly called Text-to-speech (TTS), is a problem relying on sophisticated technology encompassing various fields, including acoustics, linguistics, digital signal processing, and statistics. Its primary objective is to transform textual input into comprehensible acoustic speech output, however, many of the recent challenges involve meeting other, more subjective criteria, like naturalness, fluency, and emotion.

There are several types of speech synthesis, starting from formant synthesis (popular in the 1980s), concatenation synthesis (popular in the 1990s), unit-selection speech synthesis (still used today because of its speed and naturalness of speech), statistical-parametric synthesis (e.g. based on Hidden Markov Models), which is still being developed today and finally systems using deep learning. Especially the last two deserve attention. Deep neural networks for generating synthetic speech from text are trained using a large amount of recorded speech and associated labels or input labels in a text-to-speech conversion system.

The Tacotron 2 model proposed by Google combines the WaveNet vocoder with the revised Tacotron architecture to perform end-to-end speech synthesis. It starts by generating mel-spectrograms autoregressively from text and then synthesizes speech from the generated mel-spectrograms using a separately trained vocoder [1, 2]. Typically, the main challenge is slow inference and problems with stability (such as skipping or repeating words), but Tacotron2 can generate high-quality speech approaching the quality of a human voice. In recent times, there has been a growth in the development of non-autoregressive Text-to-Speech (TTS) models aimed at resolving these concerns. FastSpeech 2 is one of the most notable and accomplished models [1]. According to Ren et al. [3], the quality of the Mean Opinion Score test (MOS) results in Tacotron 2 of 3.86 ± 0.09, FastSpeech 3.84 ± 0.08, Transformer TTS 3.88 ± 0.08 where the audio quality was judged for 4.41 ± 0.08. In the MOS test, each audio was listened to by at least 20 testers, who are all native English speakers. The FastSpeech model can speed up the mel-spectrogram generation by 270x and the end-to-end speech synthesis by 38x, almost eliminating the problem of word skipping and repeating. It can also adjust voice speed (0.5x – 1.5x) smoothly, unlike Transformer TTS.

Since then, the above methods have become the leading research topic because many researchers worldwide have noticed the power of end-to-end speech synthesizers [4, 5].

Polish speech synthesis is offered by many companies, for example, the Edinburgh-based CereProc, which is a company that offers custom synthetic voices for individual customers [6], using unit selection speech synthesis. The building of custom voices involves adapting an acoustic model based on approximately four hours of recorded speech. A female voice (Pola) is available for the Polish language, but

it is not possible to adjust the synthesizer to simulate one's own voice. Acapela is another company producing custom-made synthetic voices. Microsoft Azure offers Custom Neural Voice service, a set of online tools for creating voice for brands. In the Custom Neural Voice Pro version, 300–2000 utterances are required [7]. Here, the Polish language is available. Narakeet [8] and Speechify [9] companies also offer Polish TTS .

Voice cloning is a powerful technology that allows the creating a digital copy of someone's voice. It has been developed to improve human life, such as restoring voices for people who have lost it. In Szklanny et al. [10] a speech synthesizer was implemented for a person with laryngeal cancer. The recordings were made just before surgery. The representative corpus of the Polish language was used for the recordings. The voice quality in the recordings is significantly altered, confirmed by the RBH perceptual scale scores and the AVQI parameter (The Acoustic Voice Quality Index). The speech synthesizer model was trained using the Merlin library. Twenty-five experts took part in MUSHRA listening tests and rated the synthesized voice at 69.4% on a scale of 0 to 100, which is a very good score.

The discussed technology also finds its application in creating audiobooks, personalized digital assistants, or speech translation services. For the Polish language, voice cloning technology as a paid service is offered by companies Resemble [11], Elevenlabs [12], and Beyondwords [13]. Voice cloning services are often associated with additional capabilities to create avatars for e-learning content, such as syntesia.io [14].

The challenge for researchers today is to obtain a synthetic voice containing emotions. In the paper [15], the topic was addressed for English. For the Polish language, it remains an ongoing challenge.

The first goal of the work is to digitize speech and to prepare a database of emotions of the outstanding actor, Piotr Fronczewski to clone his voice and synthesize his acting interpretations and emotions derived directly from the text content. Many people from the generation of today's 50-year-olds were raised on his fairy tales and radio plays. Older people remember his excellent roles - in films, TV series, on stage, in cabaret, and, above all, in the theatre. He is also known for his created vocal versions. His voice is positively received by children and adolescents, which can be applied in personalized therapies and coaching. The corpora prepared so far as recordings constitute a prototype for further work training the Polish speech synthesis model in Tacotron 2.

2. Corpus

Neural text-to-speech (TTS) methods typically require large amounts of high-quality speech data, making it challenging to obtain a dataset with emotion labels [16]. Therefore, corpus was used for the experiments, verified in the author's doctoral dissertation and other papers on speech synthesis for the Polish language [17-20].

The dataset used for the recordings comprised selected speeches from parliamentary sessions. Initially, it consisted of a 300 MB text file containing 5 778 460 sentences. All metadata was removed, and abbreviations, acronyms, and numbers were replaced with corresponding full words. The SAMPA phonetic alphabet, a computer-readable phonetic system, was employed to generate a phonetic transcription. SAMPA transcriptions are designed to be parsed without spaces between symbols.

The decision tree-based method for phonetic transcription was used. This method achieved higher accuracy in phonetic transcription for the Festival speech synthesis system [18]. To ensure balance in the corpus, a greedy algorithm was implemented. This approach best fulfilled the criteria, such as the number of phonemes, diphones, triphones per sentence, or segments in the final corpus. The CorpusCrt program, developed by Alberto Sesma Bailador in 1998 at the Polytechnic University of Catalonia, was utilized for corpus balancing and distributed as freeware [21].

The parliamentary speech corpus was divided into 12 sub-corpora, 20 MB each. This division was determined by the maximum corpus size the Corpus CRT program could handle.

To select the most representative and balanced sentences, the following criteria were applied:

- Each sentence should contain a minimum of 30 phonemes;
- Each sentence should contain a maximum of 80 phonemes;
- The output corpus should contain 2 500 sentences;
- Each phoneme should occur at least 40 times in the corpus;
- Each diphone should occur at least four times in the corpus;
- Each triphone should occur at least three times in the corpus (this particular criterion can only be met for the most frequently used triphones).

These assumptions were made based on [18, 22, 23].

After the first balancing process, 12 different sub-corpora, each containing 2 500 sentences, were created. The corpus was balanced three times to increase the number of diphones and triphones. Next, words containing rare phonemes were added.

In the last phase of constructing the corpus, manual correction was done to remove sentences that were either nonsensical or challenging to pronounce. As a result, the corpus consists of 2 150 sentences.

The finalized corpus was utilized in a doctoral dissertation focusing on optimizing the cost function in unit selection speech synthesis, specifically for the Polish language. More information about the corpus can be found in the doctoral thesis [18].

An example input sentence in our initial corpus is represented in (a) orthography, (b) phonemes, (c) diphones, and (d) triphones.

a. dowiedziawszy się o morderczych skłonnościach ciotek , zamierza powiększyć cmentarz w piwnicy.
b. # d o v j e dz' a f S I s' e~ o m o r d e r tS I x s k w o n n o s' ts' a x ts' o t e k z a m j e Z a p o v j e N k S I ts' ts m e n t a S v p i v n' i ts I #
c. #d do ov vj je edz' dz'a af fS SI Is' s'e~ e~o om mo or rd de er rtS tSI Ix xs sk kw wo on nn no os' s'ts' ts'a ax xts' ts'o ot te ek kz za am mj je eZ Za ap po ov vj je eN Nk kS SI Its' ts'ts tsm me en nt ta aS Sv vp pi iv vn' n'i its tsI I#
d. #do dov ovj vje jedz' edz'a dz'af afS fSI SIs' Is'e~ s'e~o e~om omo mor ord rde der ertS rtSI tSIx Ixs xsk skw kwo won onn nno nos' os'ts' s'ts'a ts'ax axts' xts'o ts'ot ote tek ekz kza zam amj mje jeZ eZa Zap apo pov ovj vje jeN eNk NkS kSI SIts' Its'ts ts'tsm tsme men ent nta taS aSv Svp vpi piv ivn' vn'i n'its itsI tsI#

3. Recordings

The recordings took place in the Prosound Studio with 2.0 and 5.1 TV mixing and mastering room and a 5.1, 7.1, and Dolby Atmos HE mixing stage. Neumann U87 microphone with a pop filter was used for the recordings and Preamp Focusrite, ISA Two. The audio signal was recorded in the WAV format, with a sampling frequency of 192 kHz and a resolution of 24 bits. The recording was done using the Avid S3 audio interface. So far, over 800 prompts have been recorded in ProTools Ultimate software, each in neutral and emotional versions. The version of the corpus recorded neutrally provides a reference point for the emotional version. It is read as far as possible in an emotionless but non-monotonic way. The emotional version is directed at the substantive content of the text, thus providing a good source for training speech synthesis models like creating emotional speech synthesis. The corpus has been recorded during three two-hour sessions. Also 8 poems by Konstanty Ildefons Gałczyński were recorded: „Umarł Stalin", „Dziwny kelner", „Dymiący piecyk", „Straszny koniec Spóźnialskich", „Telefon do Hermenegildy", „Lament twórcy", „W szponach kofeiny" and „Smutny, Koegzystencja, Sylwester".

W swoim wystąpieniu Giertych scharakteryzował poprzednie rządy. Według niego, rząd Tadeusza Mazowieckiego był poprzetykany komunistami i agentami, rząd Jana Olszewskiego był ospały, Waldemara Pawlaka – kostyczny, Hanny Suchockiej – wyprzedający majątek narodowy, Józefa Oleksego – obły, Włodzimierza Cimoszewicza – arogancki, Leszka Millera – groźny, Jerzego Buzka – aferalno-koleżeński.

The aforementioned text was recorded in several styles: informative, angry, sad, speech impaired, simulating a drunk person, hilarious, and frightened.

4. Corpus analysis

The fundamental frequency (pitch) is employed to define the source of speech, encompassing speech's tonal and rhythmic characteristics. In this research, Praat software was used for calculating these parameters [24]. The autocorrelation algorithm is utilised among various methods to determine the fundamental frequency. Assessing the behaviour of F0 solely based on its contour can pose challenges. Therefore, instead of relying on the pitch value itself, analysing global statistical features of the pitch contour across the entire utterance is widely accepted. Another informative factor is the speech signal's energy, which indicates the speech's volume or intensity. This energy can differentiate emotions, such as joy and anger, which exhibit higher energy levels than other emotional states [25].

Table 1 shows the analysis of Piotr Fronczewski's different speaking styles for F0, mean value, standard deviation, mean absolute slope, and mean absolute slope without octave jumps. Red indicates the smallest, while blue is the largest value for a given style. In the next stage, with the acquisition of more recordings, it is planned to expand the analysis of the speaker's speaking styles. However, the goal has been achieved at

this stage, and the following study makes it possible to distinguish speech styles. A similar study was made for recordings from the main corpus. Figure 1 shows two spectrograms, at the top for the informational style and at the bottom for the style characteristic of a drunk person. The blue color indicates the course of the F0 value, and the yellow color the sound intensity. For the sentence: „*Gdy Nora nie może spać w nocy, mąż robi jej jajecznicę*", it was obtained:

- Average F0 = 122 Hz (neutral sentence), vs 163 Hz (emotional sentence),
- STD 48 Hz vs 58 Hz,
- Mean absolute slope 326 Hz/s vs 568 Hz/s,
- Mean absolute slope without octave jumps 28 semitones/s vs 41 semitones/s.

In this case, the analysis also made it possible to distinguish speaking styles. In future, Geneva Minimalistic Acoustic Parameter Set (GeMAPS) toolkit [26] and Open-Source Audio Feature Extractor (openSMILE) will be used for voice research and affective computing [27].

Table 1. Analysis of Piotr Fronczewski's speech styles.

Value/Style	Informative	Drunken	With anger	Anxiety, fear	Mockingly	Sadly	Speech dysfunction
Minimum [Hz]	74.99	74.99	76.45	75.04	77.17	76.01	75.11
Maximum [Hz]	531.3	524.2	414.6	514.2	510.3	686.8	1008.1
Average [Hz]	121.2	146.7	163.8	156.5	170.6	127.7	153.2
Standard deviation [Hz]	31.4	44.9	36.7	53.8	53.6	75.6	98.3
Mean absolute slope [Hz/s]	122.1	179.8	314.5	176.7	343	236.1	258.7
Mean absolute slope without octave jumps [semitones/s]	12.4	15.8	26.5	12.3	25.2	13.1	14.0

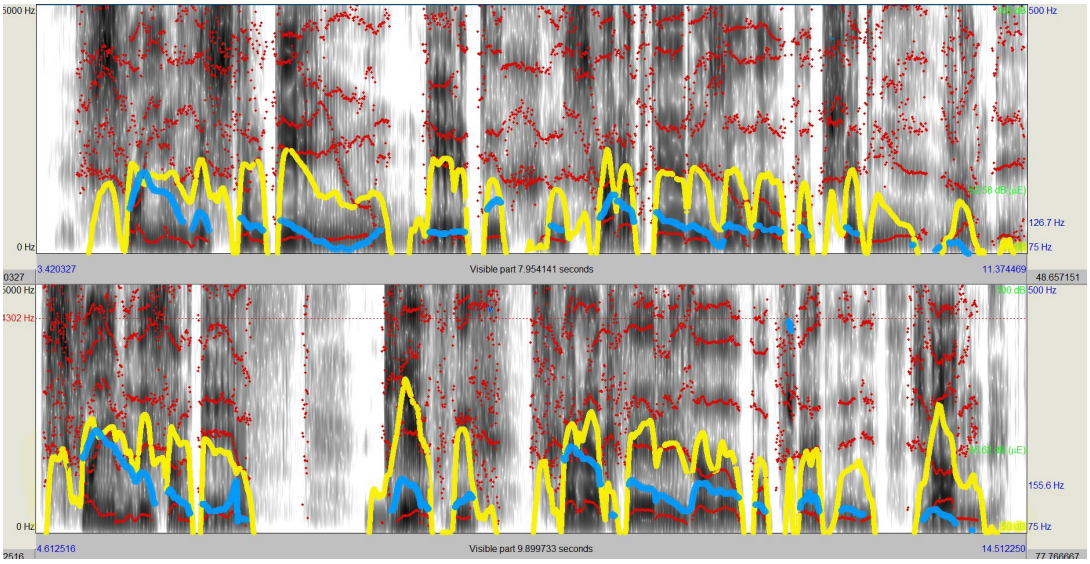

Figure 1. Comparison of sound intensity and F0 waveforms for two styles: informative (top) styled as a drunk person (bottom).

Figure 2. Sound director room.

Figure 3. Piotr Fronczewski during the recordings. One of the rooms of the recording studio. Microphone settings.

The corpus in its emotional form will be annotated in the Annotation Pro [28] software. This version of the corpus has time-varying emotions depending on the semantic layer of the text, so it is possible to obtain several styles in one sentence.

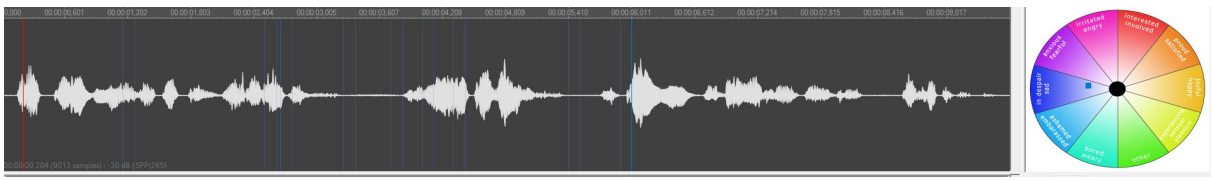

Figure 4. Screenshot of the emotion annotation feature of the Annotation Pro program.

4.1. Corpus for emotional recordings based on the Plutchik model

Automatic recognition and speech synthesis use emotional speech databases [29-33]. Regarding synthesis, it might be sufficient to focus on studying a single speaker, precisely their unique way of expressing emotions. Even though there exists quite a significant number of emotional speech datasets collected in the

world, most work has been done on Germanic languages [30]. The coverage for other language groups is relatively sparse [31]. There is a need to create an emotional database for the Polish language [34].

Emotional databases are created in one of three ways. The speech is recorded by well-trained performers (simulated database), artificial emotional situations are simulated (induced database), or natural cases are recorded (natural database) [35].

Based on the analysis of the recorded corpus, a corpus was prepared to record several speaking styles. The corpus contains 100 sentences and has been balanced for Polish like the main corpus. The corpus will be recorded based on Plutchik's model of emotions. Plutchik proposed a psychoevolutionary classification approach for general emotional responses [36]. He considered eight primary emotions: anger, fear, sadness, disgust, surprise, anticipation, trust, and joy. He suggested eight primary bipolar emotions: joy versus sadness; anger versus fear; trust versus disgust; and surprise versus anticipation. The vertical dimension represents intensity. The opposite emotions on the circle are the opposite emotions. So, each sentence will be spoken in 8 different ways corresponding to the emotions in Plutchik's model.

All other emotions can be regarded as mixed or derivative states of these primary emotions. According to the emotion wheel theory, the intensity changes could produce the diverse amount of emotions we can feel. Besides, the adding up of primary emotions could produce new emotion types. For example, delight can be produced by combining joy and surprise [37].

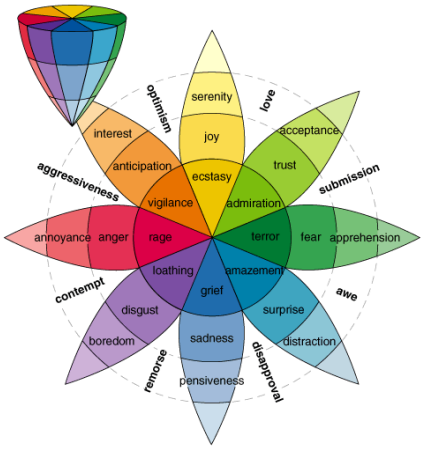

Figure 5. Plutchik's Wheel of Emotions.

4.2. Electroglottography

Emotions can be described with acoustic features and are also related to voice quality. The relationship between intensity and pitch is often associated with activation, meaning that higher pitch tends to be accompanied by increased intensity, while lower pitch is associated with decreased intensity [32]. Several factors influence the mapping from acoustic variables to emotional expression, including whether the speaker is performing, significant speaker variations, and the individual's mood or personality [35].

As mentioned, emotions are related to voice quality (Table 2). This is the reason why it was decided to record an electroglottographic signal. Previously, Szklanny et al. proved that electroglottography and advanced acoustic analysis can be used in rare genetic diseases in assessing voice in children with vocal nodules and singers [38-42].

Electroglottography is a noninvasive method that monitors the vibration of the vocal folds by measuring the electrical impedance between them. It was initially introduced by Fabre [43] and further developed with influential contributions by Baken and Orlikoff [44]. Compared to modal phonations, it is possible to differentiate between a breathy voice, a creaky voice, and a tense voice.

Electroglottographic recordings contain about 100 prompts. To record a high-quality signal, it is important to position the electrodes around the vocal folds; for this purpose, velcro neck straps are used, which makes phonation difficult. The recorded signal will enable implementation two or three-mass model of Piotr Fronczewski's vocal folds [45].

Table 2. Summarized form of acoustic variations observed based on emotions [35].

Emotions	Pitch	Intensity	Speaking rate	Voice quality
Anger	abrupt on stress	much higher	marginally faster	breathy, chest
Disgust	wide, downward inflections	lower	very much faster	grumble chest tone
Fear	wide, normal	lower	much faster	irregular voicing
Happiness	much wider, upward inflections	higher	faster/slower	breathy, blaring tone
Joy	high mean, wide range	higher	faster	breathy; blaring timbre
Sadness	slightly narrower	downward inflections	lower	resonant

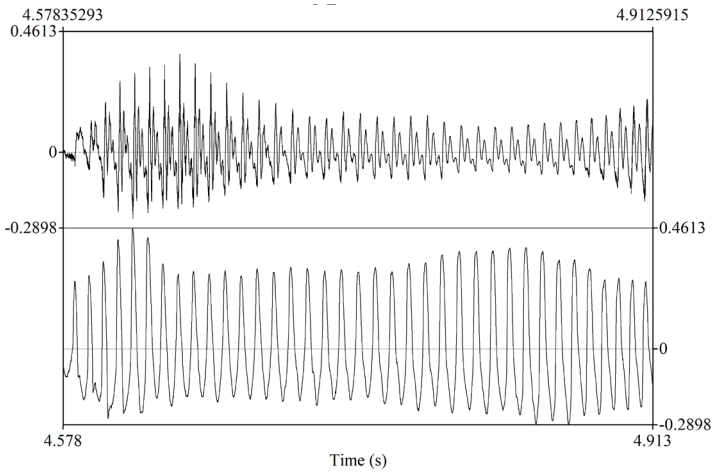

Figure 7. Signal from the microphone (top) and EGG signal (bottom).

5. Discussion

The development of deep learning technologies made obtaining synthetic speech of a quality close to natural speech possible. Expressive speech synthesis, however, remains a challenge. The same text can be spoken in many ways depending on the context, the emotional tone, and a speaker's dialectic and habitual speaking patterns of a speaker [15]. Emotional databases are necessary to create a TTS system for the Polish language. According to Khalil et al. [35], five free emotional databases are available; Berlin emotional Database (German), Danish emotional database, Interactive emotional Dyadic Motion Capture (English), Interface05 (English, Spanish, French, Slovenia), LDC Emotional English Speech and Transcripts) [35]. For the Polish language, a few databases were created by Staroniewicz et al. [34], Kamińska et al. [46], Igras et al. [47], Sapiński et al. [48], Piątek et al. [49], Janicki et al. [50], Cichosz [51], and Demenko et al. [52]. Still, an acoustic database containing eight emotions based on the Plutchik model is missing.

There is a need to create databases of emotions, especially recorded by a prominent actor. The corpus is in the process of implementation, and so far, it has been recorded:

- Over 800 prompts from the main corpus, in two versions, a neutral version and the other targeting emotions arising from the semantic layer of the text;
- Eight poems, with emotions arising directly from the text;
- Several sentences with different speaking styles;
- Approximately one hour of electroglottographic signal.

The methods that have proven successful in previous speech synthesis system implementations have been used to create corpus and database recordings. The corpus based on Plutchik's emotion wheel will be recorded in the following stages. In the future deep neural speech synthesis system will be trained. It is expected to achieve possibilities to create all kinds of audiobooks, personalized messages (including educational, psychological and emotional support), and dubbing capabilities (continuations in sequels). Optional multilingual capabilities, for example, Fronczewski speaking Chinese, will be possible. Interactive narratives in games (today, thousands of versions of messages are recorded, and thanks to speech synthesis, they could be created in real-time). Also, audio guides and possible vocal options could be possible.

6. Conclusions

The process involved in creating emotional acoustic database is very labour-intensive, as it includes many steps such as creating a text corpus, searching for suitable voice talent, conducting the recordings, taking into account many emotions and ensuring that the recordings are annotated correctly. The primary challenge in establishing a large and reliable database is ensuring the authenticity of emotions expressed in the recordings. The expression of emotions by actors and speakers raises concerns regarding the credibility of the research and its findings [49]. This is the reason why the outstanding Piotr Fronczewski voice was selected.

So far, the author has recorded four balanced corpora for speech synthesis purposes. Two with professional speakers, one semi-professional, and one by a laryngeal cancer patient. The implementation of the speech synthesis model has been started. It is planned to train model for professional speaker and, later on, for Piotr Fronczewski. This order is related to additional data prepared for mentioned corpora, like segmentation, annotation etc. Different versions of mentioned corpora allow us to conduct perceptual tests for speech synthesis models trained with Tacotron 2 or Fastspeech 2. In the future, emotional speech synthesis will be developed.

Acknowledgments

The project cooperates with Cineo Studio, owned by Kamil Przełęcki – a film producer and Ph.D. D.Sc. Jacek Hamela, associate professor of the University of Silesia. I want to thank Kamil Przełęcki for coordinating the project, Katarzyna Dzida-Hamela and Jacek Hamela for realizing the recordings in the sound studio.

Additional information

The author declare: no competing financial interests and that all material taken from other sources (including their own published works) is clearly cited and that appropriate permits are obtained.

References

1. J. Shen, R. Pang, R.J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, RJ. Skerry-Ryan, R.A. Saurous, Y. Agiomyrgiannakis, Y. Wu; Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions; https://arxiv.org/abs/1712.05884
2. N. Kaur, P. Singh; Conventional and contemporary approaches used in text to speech synthesis: A review; Artificial Intelligence Review, 2023, 56, 5837–5880
3. Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, T. Y. Liu; Fastspeech 2: Fast and high-quality end-to-end text to speech; arXiv preprint, 2020; https://arxiv.org/abs/2006.04558
4. W. Hu, X. Zhu; A real-time voice cloning system with multiple algorithms for speech quality improvement; Plos one, 2023, 18(4), e0283440
5. Y. Lei, S. Yang, X. Wang, L. Xie; Msemotts: Multi-scale emotion transfer, prediction, and control for emotional speech synthesis; IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2022, 30, 853–864
6. CereProc Company Website; http://www.cereproc.com/ (accessed on 2023.07.17)
7. Train Your Voice Model; https://docs.microsoft.com/en-us/azure/cognitive-services/speech-service/how-to-custom-voice-create-voice (accessed on 2023.07.17)
8. Narakeet Company Website; https://www.narakeet.com/ (accessed on 2023.07.28)
9. Speechify Company Website; https://speechify.com/text-to-speech-online/ (accessed on 2023.07.28)
10. K. Szklanny, J. Lachowicz; Implementing a Statistical Parametric Speech Synthesis System for a Patient with Laryngeal Cancer; Sensors, 2022, 22(9), 3188
11. Resemble Company Website; https://www.resemble.ai/cloned/ (accessed on 2023.07.17)
12. Elevenlabs Company Website; https://beta.elevenlabs.io/voice-lab (accessed on 2023.07.17)
13. Beyondwords Company Website; https://beyondwords.io/ai-voice-ethics/ (accessed on 2023.07.17)
14. Synthesia Company Website; https://www.synthesia.io/ (accessed on 2023.07.17)
15. Y.A. Li, C. Han, N. Mesgarani; Styletts: A style-based generative model for natural and diverse text-to-speech synthesis; arXiv preprint, 2022; https://arxiv.org/abs/2205.15439
16. X. Cai, D. Dai, Z. Wu, X. Li, J. Li, H. Meng; Emotion controllable speech synthesis using emotion-unlabeled dataset with the assistance of cross-domain speech emotion recognition; In: ICASSP 2021 – 2021 IEEE International Conference on Acoustics, Speech and Signal Processing, 2021, 5734–5738

17. K. Szklanny, S. Koszuta; Implementation and verification of speech database for unit selection speech synthesis; In: Proceedings of the 2017 Federated Conference on Computer Science and Information Systems (FedCSIS), Prague, Czech Republic, 3–6 September 2017
18. K. Szklanny; Optymalizacja funkcji kosztu w korpusowej syntezie mowy polskiej; PhD thesis, Polsko-Japońska Wyższa Szkoła Technik Komputerowych Warszawa, Warszawa, Poland, 2009
19. D. Oliver, K. Szklanny; Creation and analysis of a Polish speech database for use in unit selection synthesis; In: Proceedings of the Fifth International Conference on Language Resources and Evaluation, Genoa, Italy, 24–26 May 2006
20. K. Szklanny; Multimodal Speech Synthesis for Polish Language; In: Man-Machine Interactions 3. Advances in Intelligent Systems and Computing ; D. Gruca, T. Czachórski, S. Kozielski, Eds.; Springer, 2014, 242, 325–333
21. A.S. Bailador; CorpusCrt; Technical report, Polytechnic University of Catalonia (UPC), 1998
22. B. Bozkurt, O. Ozturk, T. Dutoit; Text design for TTS speech corpus building using a modified greedy selection; In: Proceedings of the Eighth European Conference on Speech Communication and Technology, Geneva, Switzerland, 1–4 September 2003
23. R.A. Clark, K. Richmond, S. King; Multisyn: Open-domain unit selection for the Festival speech synthesis system; Speech Commun., 2007, 49, 317–330
24. P. Boersma; Praat, a system for doing phonetics by computer; Glot. Int., 2001, 5(9), 341–345
25. D. Kamińska, T. Sapiński; Polish emotional speech recognition based on the committee of classifiers; Przegląd Elektrotechniczny, 2017, 93, 101–105
26. F. Eyben, K. Scherer, B. Schuller, J. Sundberg, E. André, C. Busso, K. Truong; The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing; IEEE transactions on affective computing, 2015, 7(2), 190–202
27. F. Eyben, M. Wöllmer, B. Schuller; Opensmile: the munich versatile and fast open-source audio feature extractor; In: Proceedings of the 18th ACM international conference on Multimedia, 2010, 1459–1462
28. K. Klessa, M. Karpiński, A. Wagner; Annotation Pro – a new software tool for annotation of linguistic and paralinguistic features; In: Proceedings of the Tools and Resources for the Analysis of Speech Prosody (TRASP) Workshop; D. Hirst, B. Bigi, Eds.; Aix en Provence, 2013, 51–54
29. F. Burkhard, A.,Paeschkhe, M. Rolfes, W. Sendlmeier, B. Weiss; A Database of German Emotional Speech; In: Proc. of Interspeech 2005, Lissabon, Portugal, 2005
30. E. Douglas-Cowie, N. Campbell, R. Cowie, P. Roach; Emotional speech: Towards a new generation of databases; Speech Communication, 2003, 40, 33–60
31. S.J. Jovcic, Z. Kasic, M. Dordevic, M. Rajkovic; Serbian emotional speech database: design, processing and evaluation; In: Proc. SPECOM 2004, St. Petersburg, Russia, 2004
32. D. Ververdis, C. Kotropoulos; A State of the Art on Emotional Speech Databases; In: Proc. of 1st Richmedia Conf. Laussane, Switzerland, October 2003, 109–119
33. P. Staroniewicz; Polish emotional speech database – design; In: Proc. of 55th Open Seminar on Acoustics, Wroclaw, Poland, 2008, 373–378
34. P. Staroniewicz, W. Majewski; Polish emotional speech database – recording and preliminary validation; In: Cross-Modal Analysis of Speech, Gestures, Gaze and Facial Expressions: COST Action 2102 International Conference Prague, Czech Republic, 15–18 October 2008; Revised Selected and Invited Papers; Springer, 2009, 42–49
35. R.A. Khalil, E. Jones, M.I. Babar, T. Jan, M.H. Zafar, T. Alhussain; Speech emotion recognition using deep learning techniques: A review; IEEE Access, 2019, 7, 117327–117345
36. R. Plutchik, H. Kellerman; Emotion, theory, research, and experience: theory, research and experience; Academic Press, 1980
37. K. Zhou, B. Sisman, R. Rana, B.W. Schuller, H. Li; Speech synthesis with mixed emotions; IEEE Transactions on Affective Computing, 2022, 14(4), 3120–3134
38. K. Szklanny, R. Gubrynowicz, K. Iwanicka-Pronicka, A. Tylki-Szymańska; Analysis of voice quality in patients with late-onset Pompe disease; Orphanet Journal of Rare Diseases, 2016, 11(1), 1–9
39. K. Szklanny, A. Tylki-Szymańska; Follow-up analysis of voice quality in patients with late-onset Pompe disease; Orphanet Journal of Rare Diseases, 2018, 13(1), 1–7
40. K. Szklanny, R. Gubrynowicz, A. Tylki-Szymańska; Voice alterations in patients with Morquio A syndrome; Journal of applied genetics, 2018, 59, 73–80

41. K. Szklanny, P. Wrzeciono; The application of a genetic algorithm in the noninvasive assessment of vocal nodules in children; IEEE Access, 2019, 7, 44966–44976
42. K. Szklanny; Acoustic Parameters in the Evaluation of Voice Quality of Choral Singers. Prototype of Mobile Application for Voice Quality Evaluation; Archives of Acoustics, 2019, 44(3), 439–446
43. M.P. Fabre; Un procede electrique percutane d'inscription de I'accolement glottique au cours de la phonationl; glottographie de haute frequence. Premiers resultats; Bull Acad Nat Med, 1957, 141, 66–69
44. R.J. Baken; Clinical measurement of speech and voice; College-Hill Press, 1987
45. L. Cveticanin; Review on mathematical and mechanical models of the vocal cord; Journal of Applied Mathematics, 2012, 928591
46. D. Kaminska, T. Sapinski, A. Pelikant; Polish Emotional Natural Speech Database; In: Proceedings of the Conference: Signal Processing Symposium, 2015
47. M. Igras, B. Ziółko; Baza danych nagrań mowy emocjonalnej; Studia Informatica, 2013, 34, 67–77
48. T. Sapiński, D. Kamińska, A. Pelikant, C. Ozcinar, E. Avots, G. Anbarjafari; Multimodal database of emotional speech, video and gestures; In: Pattern Recognition and Information Forensics: ICPR 2018 International Workshops, CVAUI, IWCF, and MIPPSNA, Beijing, China, 20–24 August 2018; Revised Selected Papers 24, Springer International Publishing, 2019, 153–163
49. Z. Piątek, M. Kłaczyński; Acoustic Methods in Identifying Symptoms of Emotional States; Archives of Acoustics, 2021, 46(2), 259–269
50. A. Janicki, M. Turkot; Rozpoznawanie stanu emocjonalnego mówcy z wykorzystaniem maszyny wektorów wspierających (svm); Krajowe Sympozjum Telekomunikacji i Teleinformatyki, Bydgoszcz, 2008
51. J. Cichosz; The use of selected speech signal features to recognize and model emotions for the Polish language [in Polish: Wykorzystanie wybranych cech sygnału mowy do rozpoznawania i modelowania emocji dla języka polskiego]; PhD thesis, Lodz University of Technology, Łódź, 2008
52. G. Demenko, M. Jastrzębska; Analiza stresu głosowego w rozmowach z telefonu alarmowego; XVIII Conference on Acoustic and Biomedical Engineering, Zakopane, Poland, 2011