Narzędzie do transferu emocji w próbkach mowy

Postępujący w ostatnich latach w błyskawicznym tempie rozwój metod sztucznej inteligencji opartych na przetwarzaniu języka naturalnego pozwala zakładać, że w najbliższej przyszłości systemy interakcji człowiek-komputer w coraz większej mierze będą bazować na naturalności oraz prostocie użycia. Jednym z elementów takiego systemu powinien być syntezator mowy, cechujący się możliwie jak najlepszym odwzorowaniem wszelkich atrybutów oraz wyróżników mowy ludzkiej. Komunikacja prowadzona z systemem komputerowym w języku naturalnym może w dużej mierze zostać pozbawiona swobody poprzez występujące często ograniczenia modułów text-to-speech, które niejednokrotnie syntezują głosy brzmiące „płasko" bądź „robotycznie", innymi słowy, pozbawione emocji.

Z uwagi na te niedostatki, obecnie kładzie się duży nacisk na rozwój metod syntezy mowy opartych na głębokich sieciach neuronowych, pozwalających na niezależną kontrolę nad różnymi brzmieniowymi właściwościami mowy, takimi jak prozodia czy barwa, które powinny być w procesie syntezy kontrolowalne w oderwaniu od głoskowego bądź sylabicznego ciągu wypowiedzi [1]. Posiadanie takiego stopnia kontroli nad syntezowaną treścią pozwala na realizowanie różnorodnych zadań, takich jak edycja wyrazowego ciągu wypowiedzi [2], transfer głosu (Speech-to-Speech) [3], synteza oparta na tekście (Text-to-Speech) [4] bądź edycja zabarwienia emocjonalnego syntezowanej bądź rekonstruowanej próbki mowy [5].

Celem projektu jest przygotowanie systemu opartego na głębokich sieciach neuronowych, pozwalającego na modyfikację próbek mowy, tak aby cechowały się one występowaniem określonego zabarwienia emocjonalnego, przy równoczesnym zachowaniu naturalności i wiarygodności. Kontrola nad syntezowanym głosem może odbywać się zarówno na poziomie indywidualnych cech (tempo, intensywność, ton głosu), jak i w ujęciu bardziej ogólnym (na przykład zmiana głosu na radosny bądź przestraszony). Jednocześnie, przekształcenia dokonywane w warstwie stylu nie powinny mieć wpływu na językową treść wypowiedzi, która powinna pozostać niezmienna.

Literatura:

1. *Z. Ju i in., „NaturalSpeech 3: Zero-Shot Speech Synthesis with Factorized Codec and Diffusion Models". arXiv, 5 marzec 2024. Dostęp: 25 marzec 2024. [Online]. Dostępne na: http://arxiv.org/abs/2403.03100*
2. P. Peng, P.-Y. Huang, D. Li, A. Mohamed, i D. Harwath, „VoiceCraft: Zero-Shot Speech Editing and Text-to-Speech in the Wild". arXiv, 25 marzec 2024. Dostęp: 27 marzec 2024. [Online]. Dostępne na: http://arxiv.org/abs/2403.16973
3. Z. Borsos *i in.*, „AudioLM: a Language Modeling Approach to Audio Generation". arXiv, 25 lipiec 2023. Dostęp: 27 marzec 2024. [Online]. Dostępne na: http://arxiv.org/abs/2209.03143
4. C. Wang *i in.*, „Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers". arXiv, 5 styczeń 2023. Dostęp: 27 marzec 2024. [Online]. Dostępne na: http://arxiv.org/abs/2301.02111
5. Y. Lei, S. Yang, X. Wang, i L. Xie, „MsEmoTTS: Multi-scale emotion transfer, prediction, and control for emotional speech synthesis". arXiv, 17 styczeń 2022. Dostęp: 27 marzec 2024. [Online]. Dostępne na: http://arxiv.org/abs/2201.06460