Synteza mowy emocjonalnej

Postępujący w ostatnich latach w błyskawicznym tempie rozwój metod sztucznej inteligencji opartych o przetwarzanie języka naturalnego pozwala zakładać, że w najbliższej przyszłości systemy interakcji człowiek-komputer w coraz większej mierze bazować będą 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.

W odpowiedzi na problem niedostatecznej naturalności generowanego metodami sztucznej inteligencji głosu, badacze zajmujący się tym zagadnieniem proponują szereg rozwiązań mających na celu *uemocjonalnienie* syntezowanych próbek mowy. Podejmowane są m.in. próby przeprowadzenia procesu transferu stylu [1, 2], kontroli manualnej realizowanej poprzez wytrenowanie sieci warunkowej (Conditional GAN [3] lub RNN [4]) lub poprzez modyfikację przestrzeni ukrytej sieci typu enkoder-dekoder [5]. Mimo to, synteza mowy emocjonalnej pozostaje problemem dalekim od rozwiązania, chociażby z uwagi na wysoce różnorodną naturę występującej w głosie ekspresji emocji – może ona występować na poziomie całego zdania, ale również na poziomie pojedynczego słowa, a ta sama emocja może być wyrażana przez różne osoby za pomocą innych cech głosu, na przykład jego barwy, intensywności a także intonacji czy tempa wypowiedzi [6].

Celem projektu jest przygotowanie systemu opartego o głębokie sieci neuronowe, pozwalającego na generację oraz/lub 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 synteza głosu radosnego bądź przestraszonego). Jednocześnie, przekształcenia dokonywane w warstwie stylu nie powinny mieć wpływu na językową treść wypowiedzi, która powinna pozostać niezmienna (w przypadku dokonywania modyfikacji) lub być z góry określona (w przypadku syntezy na drodze text-to-speech).

Literatura:

1. R. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. Weiss, R. Clark, and R. A. Saurous. Towards End-to-End Prosody Transfer for Expressive Speech Synthesis with Tacotron. *arXiv:1803.09047,* 2018
2. T. Li, S. Yang, L. Xue, and L. Xie. Controllable emotion transfer for end-to-end speech synthesis, in *2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, IEEE, 2021.
3. N. Jia, C. Zheng and W. Sun. A Model of Emotional Speech Generation Based on Conditional Generative Adversarial Networks, *2019 11th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, Hangzhou, China, 2019.
4. Y. Lee, A. Rabiee, S. Lee. Emotional End-to-End Neural Speech Synthesizer. *arXiv:1711.05447*, 2017.

5. J. Gao, D. Chakraborty, H. Tembine, O. Olaleye. Nonparallel Emotional Speech Conversion. *arXiv:1811.01174*, 2018.
6. Y. Lei, S. Yang, X. Wang, L. Xie. MsEmoTTS: Multi-scale emotion transfer, prediction, and control for emotional speech synthesis. *arXiv:2201.06460,* 2022.