Rule-Based TTS System For Georgian Language

Nika Iakobashvili, Mariam Luarsabishvili, Nino Kasabishvili, Nikoloz Tsiskadze

e-mail: nika.iakobashvili099@ens.tsu.edu.ge

Department of Computer Science, Faculty of Exact
and Natural Sciences, Tbilisi State University

Language is one of the most important pillars of cultural identity and its integration into the technological domain is vital to its preservation and development. Today's TTS systems for Georgian language require improvement both because of the language's particular characteristics and the scarcity of available resources. In the face of this deficit, no large-scale study has yet examined the specific challenges and requirements involved in deploying TTS systems for Georgian. The goals of our project are to: (1) Investigate the primary barriers to the development of Georgian TTS systems. (2) Design and implement a rule-based TTS prototype. (3) Define future perspectives for advancing TTS systems. Although rule-based approaches represent one of the simplest TTS models, our project will establish a solid foundation for future research and technological innovation. Moreover, the closed source nature and functional limitations of existing Georgian TTS systems further underscore the need for an open, transparent, and easily controllable solution. Consequently, our system has the potential to deliver significant benefits in education, meet the needs of users with disabilities, and increase access to Georgian culture.

We will present the project as a Windows application featuring a text input window where users enter their text. The application will automatically normalize the input, converting numbers to their written form, expanding abbreviations, and so on. For each word, it will identify the set of phonemes required for correct pronunciation, then "glue" (i.e. concatenate) them, and then add appropriate pauses and rhythm. Finally, the system will generate an audio recording that can be played directly within the application or downloaded in WAV format. The implementation will use Python and its standard libraries. Our rule-based TTS model will map each syllable to a pre-recorded phoneme. We will build and manage our own phoneme database, recording samples with Audacity and storing them in SQLite.

References

[1] T. Dutoit, High-quality text-to-speech synthesis: An overview, in Proc. Conf. on Text-to-Speech Synthesis, 2004. [Online].

[2] ა. შანიძე, ქართული ენის გრამატიკის საფუძვლები, III ტომი, თბილისის უნივერსიტეტის გამომცემლობა, თბილისი, 1980. [Online]. Available: https://archive.org/details/shanidze

[3] J. Ure, Lexical density and register differentiation, in Applications of Linguistics, G. Perren and J. L. M. Trim, Eds. Cambridge: Cambridge University Press, 1971, pp. 443–452.

[4] P. Taylor, Text-to-Speech Synthesis, Cambridge University Press, 2009.

[5] A. Hunt and A. Black, Unit selection in a concatenative speech synthesis system using a large speech database, in Proc. IEEE Int. Conf. Acoust. Speech Process., Munchen, Germany, vol. 1, pp. 373–376, 1996.

[6] A. Black and P. Taylor, Automatically clustering similar units for unit selection in speech synthesis, in Proc. Eurospeech'97, pp. 601–604, 1997.

[7] D. Chazan, R. Hoory, Z. Kons, A. Sagi, S. Shechtman, and A. Sorin, Small footprint concatenative text-to-speech synthesis using complex envelope modeling, in Proc. Interspeech'05, Lisbon, Portugal, pp. 2569–2572, 2005.

[8] A. Vadapalli, P. Bhaskararao, and K. Prahallad, Significance of word-terminal syllables for prediction of phrase breaks in Text-to-Speech systems for Indian languages, in 8th ISCA Tutorial and Research Workshop on Speech Synthesis, 2013.

[9] X. Wang, J. Lorenzo-Trueba, S. Takaki, L. Juvela, and J. Yamagishi, A comparison of recent waveform generation and acoustic modeling methods for neural-network-based speech synthesis, in 43rd ICASSP, 2018.

[10] Z. Kons, S. Shechtman, A. Sorin, C. Rabinovitz, and R. Hoory, High quality, lightweight and adaptable TTS using LPCNet, arXiv preprint arXiv:1905.00590, 2019.

[11] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, FastSpeech: Fast, Robust and Controllable Text to Speech, in Advances in Neural Information Processing Systems, pp. 3165–3174, 2019.

[12] A. Gutkin, Uniform multilingual multi-speaker acoustic model for statistical parametric speech synthesis of low-resourced languages, 2017.

[13] T. Tu, Y.-J. Chen, C.-c. Yeh, and H.-y. Lee, End-to-end text-to-speech for low-resource languages by cross-lingual transfer learning, arXiv preprint arXiv:1904.06508, 2019.

[14] A. Van Den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, WaveNet: A generative model for raw audio, arXiv preprint arXiv:1609.03499, 2016.

[15] A. Rendel, R. Fernandez, R. Hoory, and B. Ramabhadran, Using continuous lexical embeddings to improve symbolic-prosody prediction in a text-to-speech front-end, 2016.

[16] J. Zhao, G. Gao, F. D. Bao, and P. Mermelstein, Research on HMM-based Mongolian speech synthesis, Computer Science, no. 41, pp. 80–104, 2014.

[17] R. Liu, F. Bao, G. Gao, and Y. Wang, Mongolian text-to-speech system based on deep neural network, in Man-Machine Speech Communication. NCMMSC 2017. Communications in Computer and Information Science, Springer, vol. 807, 2018.

[18] J. Xu, X. Tan, Y. Ren, T. Qin, J. Li, S. Zhao, and T.-Y. Liu, LRSpeech: Extremely Low-Resource Speech Synthesis and Recognition, in Proceedings of the 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '20), August 23–27, 2020, Virtual Event, CA, USA, ACM, New York, NY, USA, pp. 11, 2020.

[19] E. Cooper, E. Li, and J. Hirschberg, Characteristics of Text-to-Speech and Other Corpora, Proceedings of Speech Prosody 2018, 2018.

[20] E. L. Cooper, Text-to-speech synthesis using found data for low-resource languages, Ph.D. Dissertation, Columbia University, 2019.

[21] N. Li, S. Liu, Y. Liu, S. Zhao, M. Liu, and M. Zhou, Neural Speech Synthesis with Transformer Network, AAAI, 2019.

[22] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, Deep Voice 3: 2000-Speaker Neural Text-to-Speech, in International Conference on Learning Representations, 2018.

[23] Praat, http://www.fon.hum.uva.nl/praat/.

[24] SPTK, http://sp-tk.sourceforge.net/.

[25] P. C. Loizou, Speech Quality Assessment, in Multimedia Analysis, Processing & Communications, pp. 623–654, Springer, Heidelberg, 2011.

[26] Y. Xiaofei, F. Bao, H. Wang, et al., A Novel Approach to Improve the Mongolian Language Model Using Intermediate Characters, in 15th China National Conference on Chinese Computational Linguistics, pp. 103–113, 2016.

[27] B. Feilong, G. Guanglai, and Y. Xueliang, Research on grapheme to phoneme conversion for Mongolian, Application Research of Computers, vol. 30, pp. 1696–1700, 2013.

[28] L. Rui, F. Bao, G. Gao, and H. Zhang, Approach to Prediction Mongolian Prosody Phrase Based on CRF Model, in 13th National Conference on Man-Machine Speech Communication, Tianjin, 2015.

[29] F. Alias and X. Llora, Evolutionary weight tuning based on diphone pairs for unit selection speech synthesis, in Proceedings of Eurospeech 2003, 2003.

[30] E. L. Amdtatham, Word and syllable concatenation in text-to-speech synthesis, in Proceedings of the European Conference on Speech.