DAFTAR PUSTAKA

About FFmpeg (no date). Available at: https://ffmpeg.org/about.html (Accessed: 20 May 2025).

Alharthi, D. *et al.* (2023) 'Evaluating Speech Synthesis by Training Recognizers on Synthetic Speech'. Available at: https://arxiv.org/pdf/2310.00706 (Accessed: 31 May 2025).

Azizah, K., Adriani, M. and Jatmiko, W. (2020) 'Hierarchical transfer learning for multilingual, multi-speaker, and style transfer DNN-based TTS on low-resource languages', *IEEE Access*, 8, pp. 179798–179812. Available at: https://doi.org/10.1109/ACCESS.2020.3027619.

Bahdanau, D., Cho, K. and Bengio, Y. (2014) 'Neural Machine Translation by Jointly Learning to Align and Translate', *arXiv preprint arXiv:1409.0473* [Preprint].

Bernard, M. and Titeux, H. (2021) 'Phonemizer: Text to Phones Transcription for Multiple Languages in Python', *Journal of Open Source Software*, 6(68), p. 3958. Available at: https://doi.org/10.21105/JOSS.03958.

Chorowski, J.K. *et al.* (2015) 'Attention-Based Models for Speech Recognition', *Advances in Neural Information Processing Systems*, 28.

Dhiaulhaq, M. *et al.* (2022) 'GAN-Based End to End Text-to-Speech System for Indonesian Language', *inacl.id* [Preprint]. Available at: http://inacl.id/journal/index.php/jlk/article/view/115 (Accessed: 2 November 2024).

Foote, J. (1999) 'Visualizing Music and Audio using Self-Similarity', *MULTIMEDIA 1999 - Proceedings of the 7th ACM International Conference on Multimedia (Part 1)*, 1, pp. 77–80. Available at: https://doi.org/10.1145/319463.319472.

Goodfellow, I. (2016) *Deep learning*. MIT Press.

Gravetter, F.J. and Wallnau, L.B. (2015) *Statistics for the Behavioral Sciences*. 10th edn. Boston: Cengage Learning. Available at: www.cengage.com/highered.

Hasanabadi, M.R. (2023) 'An overview of text-to-speech systems and media applications', *ABU Technical Review journal 2023/6* [Preprint]. Available at: https://doi.org/https://doi.org/10.48550/arXiv.2310.14301.

Hedderich, M.A. *et al.* (2021) 'A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios', pp. 2545–2568. Available at: https://en.wikipedia.org/wiki/List_ (Accessed: 30 October 2024).

ITU-T (1996) 'Methods for subjective determination of transmission quality (ITU-T Recommendation P.800)', *International Telecommunication Union* [Preprint].

Kaur, N. and Singh, P. (2022) 'Conventional and contemporary approaches used in text to speech synthesis: a review', *Artificial Intelligence Review 2022 56:7*, 56(7), pp. 5837–5880. Available at: https://doi.org/10.1007/S10462-022-10315-0.

Kingma, D.P. and Ba, J.L. (2014) 'Adam: A Method for Stochastic Optimization', *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings* [Preprint]. Available at: https://arxiv.org/abs/1412.6980v9 (Accessed: 14 November 2024).

Kong, J., Kim, J. and Bae, J. (2020) 'HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis', *Advances in Neural Information Processing Systems*, 2020-December. Available at: https://arxiv.org/abs/2010.05646v2 (Accessed: 2 November 2024).

Kurniawati, A. and Jatmiko, W. (2021) 'Transfer Learning, Style Control, and Speaker Reconstruction Loss for Zero-Shot Multilingual Multi-Speaker Text-to-Speech on Low-Resource Languages'. Available at: https://doi.org/10.1109/ACCESS.2022.3141200.

Lim, D., Jung, S. and Kim, E. (2022) 'JETS: Jointly Training FastSpeech2 and HiFi-GAN for End to End Text to Speech', in *Interspeech 2022*. ISCA: ISCA, pp. 21–25. Available at: https://doi.org/10.21437/Interspeech.2022-10294.

Narasimhan, N. (2021) 'Digital Accessibility in the Asia-Pacific Region', *Accessible Technology and the Developing World*, pp. 106–127. Available at: https://doi.org/10.1093/OSO/9780198846413.003.0006.

Novela, M. and Basaruddin, T. (2021) 'Dataset Suara dan Teks Berbahasa Indonesia Pada Rekaman Podcast dan Talk show', *Jurnal FASILKOM*, pp. 61–66. Available at: https://doi.org/https://doi.org/10.37859/jf.v11i2.2628.

Nur, T. *et al.* (2023) 'Perbedaan Fonologi Bahasa Indonesia Dan Bahasa Inggris', *Pragmatik : Jurnal Rumpun Ilmu Bahasa dan Pendidikan* , 1(3), pp. 01–09. Available at: https://doi.org/10.61132/PRAGMATIK.V1I3.201.

Oord, A. van den *et al.* (2016a) 'WaveNet: A Generative Model for Raw Audio'. Available at: https://arxiv.org/abs/1609.03499v2 (Accessed: 8 November 2024).

Oord, A. van den *et al.* (2016b) 'WaveNet: A Generative Model for Raw Audio'. Available at: https://arxiv.org/abs/1609.03499v2 (Accessed: 2 November 2024).

Oyucu, S. (2023) 'A Novel End-to-End Turkish Text-to-Speech (TTS) System via Deep Learning', *Electronics 2023, Vol. 12, Page 1900*, 12(8), p. 1900. Available at: https://doi.org/10.3390/ELECTRONICS12081900.

Radford, A. *et al.* (2022) 'Robust Speech Recognition via Large-Scale Weak Supervision'. Available at: https://github.com/openai/ (Accessed: 20 May 2025).

Ravanelli, M. *et al.* (2024) 'Open-Source Conversational AI with SpeechBrain 1.0'. Available at: https://arxiv.org/abs/2407.00463v5 (Accessed: 30 October 2024).

Ren, Y. *et al.* (2019) 'FastSpeech: Fast, Robust and Controllable Text to Speech', *Advances in Neural Information Processing Systems*, 32. Available at: https://arxiv.org/abs/1905.09263v5 (Accessed: 8 November 2024).

Resemble AI (2023) *Resemblyzer*, *https://github.com/resemble-ai/Resemblyzer*.

Schuster, M. and Paliwal, K.K. (1997) 'Bidirectional recurrent neural networks', *IEEE Transactions on Signal Processing*, 45(11), pp. 2673–2681. Available at: https://doi.org/10.1109/78.650093.

Shen, J. *et al.* (2017) 'Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions'. Available at: http://arxiv.org/abs/1712.05884.

Siddique, Z.Q. (2022) 'Child Speech Synthesis using Deep Learning', *Doctoral dissertation, Dublin, National College of Ireland* [Preprint].

Tieleman, T. and H.G. (2012) *Lecture 6.5 - RMSProp, COURSERA: Neural Networks for Machine Learning*.

Wang, Y. *et al.* (2017) 'Tacotron: Towards End-to-End Speech Synthesis', *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2017-August, pp. 4006–4010. Available at: https://doi.org/10.21437/Interspeech.2017-1452.

Zalkow, F. *et al.* (2023) 'Evaluating Speech-Phoneme Alignment and its Impact on Neural Text-To-Speech Synthesis', *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings* [Preprint]. Available at: https://doi.org/10.1109/ICASSP49357.2023.10097248.