

DAFTAR PUSTAKA

- Arisaputra, P., dan Zahra, A. (2023). Indonesian Automatic Speech Recognition with XLSR-53. *arXiv preprint arXiv:2308.11589*.
- Arum, E. R., dan Winarti, W. (2020). Penggunaan Linguistik Korpus Dalam Mempersiapkan Bahan Ajar English For Specific Purpose Di Bidang Radiologi: The Use of Corpus Linguistics in Preparing Teaching Materials of English for Specific Purposes in the Field of Radiology. *Jurnal Teras Kesehatan*, 3(1), 58-69.
- Bera, Mantu. (2021). Text to Speech Synthesis. Dept. of CSE, Bankura Unnayani Institute of Engineering, Bankura, India. *International Journal of Innovative Science and Research Technology*. ISSN No:-2456-2165.
- Bernard, M., & Titeux, H. (2021). Phonemizer: Text to phones transcription for multiple languages in python. *Journal of Open Source Software*, 6(68), 3958.
- Brownlee J. (2019). A Gentle Introduction to Generative Adversarial Networks (GANs). [online] Machine Learning Mastery. Available at: <https://machinelearningmastery.com/what-are-generative-adversarial-networks-gans/>.
- Chen, J dkk. (2020). Vflow: More expressive generative flows with variational data augmentation. In *International Conference on Machine Learning* (pp. 1660-1669). PMLR.
- Fiqqhi, M dkk. (2022). Pemanfaatan Google API Text-to-Speech Dalam Otomatisasi Pembacaan Pesan Masuk Perangkat Android.
- Furui, S. (2018). *Digital speech processing: synthesis, and recognition*. CRC Press.
- Girin, L., dkk. (2020). Dynamical variational autoencoders: A comprehensive review. *arXiv preprint arXiv:2008.12595*.

- Gomaa, W. H., dan Fahmy, A. A. (2013). A survey of text similarity approaches. *international journal of Computer Applications*, 68(13), 13-18.
- Goodfellow, I dkk. (2014). Generative adversarial nets. *Advances in neural information processing systems*, 27.
- Goyal, D., dan Pabla, B. S. (2015). Condition based maintenance of machine tools—A review. *CIRP Journal of Manufacturing Science and Technology*, 10, 24-35.
- Ho, J dkk. (2019). Flow++: Improving flow-based generative models with variational dequantization and architecture design. In *International Conference on Machine Learning* (pp. 2722-2730). PMLR.
- Hunston, Susan. (2022). *Corpora in applied linguistics*. Cambridge: Cambridge University Press.
- Inoue, K dkk. (2021). Model architectures to extrapolate emotional expressions in DNN-based text-to-speech. *Speech Communication*, 126, 35-43.
- International Telecommunication Union. Telecommunication Standardization Sector. (1996). *Methods for subjective determination of transmission quality*. International Telecommunication Union.
- Kim, J dkk. (2020). Glow-TTS: A generative flow for text-to-speech via monotonic alignment search. *Advances in Neural Information Processing Systems*, 33, 8067-8077.
- Kim, J dkk. (2021). Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *International Conference on Machine Learning* (pp. 5530-5540). PMLR.
- Kingma, D. P. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

- Kong, J dkk. (2020). HiFi-GAN: Generative adversarial networks for efficient and high-fidelity speech synthesis. *Advances in Neural Information Processing Systems*, 33, 17022-17033.
- Kong, J dkk. (2023). Vits2: Improving quality and efficiency of single-stage text-to-speech with adversarial learning and architecture design. *arXiv preprint arXiv:2307.16430*. www.itk.ac.id
- Kumar, K dkk. (2019). MelGAN: Generative adversarial networks for conditional waveform synthesis. *Advances in Neural Information Processing Systems*, 32.
- Łańcucki, A. (2021). Fastpitch: Parallel Text-To-Speech with Pitch Prediction. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6588-6592). IEEE.
- Lestari, D. P dkk. (2006). A large vocabulary continuous speech recognition system for Indonesian language. In *15th Indonesian Scientific Conference in Japan Proceedings* (pp. 17-22).
- Lestari, D. P dkk. (2011). Tokyo Institute of Technology Multilingual Speech Corpus-Indonesian. Available at: <http://ilps.science.uva.nl/resources/bahasa>.
- Li, N dkk. (2018). Close to human quality TTS with transformer. *arXiv preprint arXiv:1809.08895*, 2.
- Loshchilov, I. and Hutter, F. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*, 2019.
- McFee, B dkk. (2015). *librosa: Audio and music signal analysis in Python*. In *Proceedings of the 14th Python in Science Conference* (pp. 18–25). www.itk.ac.id
- MagicHub. (2021). *Indonesian Scripted Speech Corpus – Daily Use Sentence*. <https://magichub.com/datasets/indonesian-scripted-speech-corpus-daily-use-sentence/>

- Mahum, R dkk. (2023). Text to speech synthesis using deep learning. In *Intelligent Multimedia Signal Processing for Smart Ecosystems* (pp. 289-305). Cham: Springer International Publishing.
- Natasya, R. D. (2023). Implementasi Artificial Intelligence (AI) dalam Teknologi Modern. *KOMTEKS*, 2(1).
- Prenger, R dkk. (2019). WaveGlow: A flow-based generative network for speech synthesis. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 3617-3621). IEEE.
- Qian, S., dan Chen, D. (1999). Joint time-frequency analysis. *IEEE signal processing magazine*, 16(2), 52-67.
- Radford, A dkk. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434.
- Rezende, D., dan Mohamed, S. (2015). Variational inference with normalizing flows. In *International Conference on Machine Learning* (pp. 1530-1538). PMLR.
- Salehi, P dkk. (2020). Generative adversarial networks (GANs): An overview of theoretical model, evaluation metrics, and recent developments. arXiv preprint arXiv:2005.13178.
- Shaw, P dkk. (2018). Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155*.
- Streijl, R. C dkk. (2016). Mean opinion score (MOS) revisited: methods and applications, limitations and alternatives. *Multimedia Systems*, 22(2), 213-227.
- Suzié, S et al. (2019). DNN-based expressive text-to-speech with limited training data. In *2019 27th Telecommunications Forum (TELFOR)* (pp. 1-6). IEEE.

- Varin, C., dan Vidoni, P. (2005). A note on composite likelihood inference and model selection. *Biometrika*, 92(3), 519-528.
- Vaswani, A dkk. (2017). Attention is all you need. *Advances in neural information processing systems*, 30(2017).
- Verma, U., dan Dr. Padmanaban R. (2024). Speech Cloning: Text-To-Speech Using VITS. *Engineering And Technology Journal*, 9(5), 3951–3956. <https://doi.org/10.47191/etj/v9i05.10>
- Vincent, P dkk. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11(12).
- Wahyuni, R. T dkk. (2017). Penerapan algoritma cosine similarity dan pembobotan tf-idf pada sistem klasifikasi dokumen skripsi. *Jurnal Teknik Elektro*, 9(1), 18-23
- Wali, A dkk. (2022). Generative adversarial networks for speech processing: A review. *Computer Speech dan Language*, 72, 101308.
- Wei, R., dan Mahmood, A. (2020). Recent advances in variational autoencoders with representation learning for biomedical informatics: A survey. *IEEE Access*, 9, 4939-4956.
- Yang, X. (2017). Understanding the variational lower bound. *variational lower bound, ELBO, hard attention*, 22, 1-4.