DATASET SUARA DAN TEKS BERBAHASA INDONESIA PADA REKAMAN *PODCAST* DAN *TALK SHOW*

Martin Novela[1)], T. Basaruddin[2)]
[1]Ilmu Komputer, Fakultas Ilmu Komputer, Universitas Indonesia
email: martin.novela81@ui.ac.id
[2]Ilmu Komputer, Fakultas Ilmu Komputer, Universitas Indonesia
email: chan@cs.ui.ac.id

Abstract

One of the success factors of a learning model in machine learning or deep learning is the dataset used. This paper presents voice datasets from recorded podcasts and talk shows along with their Indonesian transcriptions. This dataset is presented because there is no publicly accessible dataset in Indonesian to be used for modeling Text-to-Speech or Audio Speech Recognition models. The dataset consists of 3270 records that are processed to obtain transcription in the form of text or sentences in Indonesian. Several stages were carried out such as preprocessing, translation stage, first validation stage, and second validation stage. The dataset is created in a format that follows the format of LJSpeech dataset to facilitate dataset processing when used in a model as an input. This dataset is expected to improve the quality of modeling for Text-to-Speech processing such as in the Tacotron2 model or in Audio Speech Recognition processing for Indonesian.

Keywords*: Dataset, Indonesian, Text-to-Speech, Audio Speech Recognition.*

Abstrak

Salah satu faktor keberhasilan suatu model pembelajaran dalam *machine learning* atau *deep learning* adalah *dataset* yang digunakan. Pada tulisan ini menyajikan *dataset* suara dari rekaman *podcast* dan *talk show* beserta transkripsi berbahasa Indonesia. *Dataset* ini disajikan karena belum adanya ketersediaan *dataset* berbahasa Indonesia yang dapat diakses secara publik untuk digunakan pada pembelajaran model *Text-to-Speech* ataupun *Audio Speech Recognition*. *Dataset* terdiri dari 3270 rekaman yang diproses untuk mendapatkan transkripsi berupa teks atau kalimat berbahasa Indonesia. Dalam pembuatan *dataset* ini dilakukan beberapa tahapan seperti pra-pemrosesan, tahapan translasi, tahapan validasi pertama dan tahapan validasi kedua. *Dataset* dibuat dengan format yang mengikuti format dari *dataset* LJSpeech untuk memudahkan pemrosesan *dataset* ketika digunakan dalam suatu model sebagai input. *Dataset* ini diharapkan dapat membantu meningkatkan kualitas pembelajaran untuk pemrosesan *Text-to-Speech* seperti pada model Tacotron2 ataupun pada pemrosesan *Audio Speech Recognition* untuk Bahasa Indonesia.

Keywords: *Dataset,* Bahasa Indonesia*, Text-to-Speech*, *Audio Speech Recognition*.

PENDAHULUAN

Perkembangan teknologi semakin pesat dalam membantu kegiatan manusia. Banyak teknologi dikembangkan untuk mempermudah penyelesaian masalah yang ditemukan dalam kehidupan sehari-hari yang semakin kompleks. Hal ini mendorong peneliti untuk mengembangkan teknologi yang menyerupai kecerdasan manusia dengan meniru cara kerja otak manusia. Salah satu metode yang digunakan dalam meniru cara kerja otak manusia saat ini adalah *deep learning* yang merupakan suatu algoritma pengembangan lebih lanjut dari *machine learning*. *Deep learning* adalah bagian dari kecerdasan buatan yang dikembangkan dari *neural network multiple layer* untuk melakukan tugas dengan tepat seperti pengenalan objek, pengenalan suara, penerjemahan bahasa, perubahan teks menjadi suara, dan sebagainya [1].

Untuk mendukung metode *deep learning* dalam menghasilkan suatu tugas dengan tepat, maka dibutuhkan *dataset* sebagai acuan pembelajaran dari model yang akan dibuat terhadap

keputusan yang akan diambil. *Dataset* merupakan bagian yang sangat penting dalam pembuatan suatu model *deep learning* [2]. Tanpa adanya *dataset*, tentu saja pembelajaran tidak bisa dilakukan. Di samping itu, data yang digunakan juga harus mewakili secara umum terhadap semesta dari data yang ada agar tingkat hasil pembelajaran dapat menghasilkan keputusan yang lebih tepat.

Salah satu pembelajaran yang banyak digunakan saat ini yaitu model *deep learning* yang membantu translasi teks menjadi suara ataupun suara menjadi teks. Kedua model tersebut saat ini sedang menjadi penelitian yang populer di kalangan peneliti. Pada pembuatan model *deep learning*, dalam memproses teks menjadi suara ataupun suara menjadi teks, dibutuhkan suatu *dataset* yang secara spesifik dapat mewakili suatu bahasa tertentu. Hal tersebut dikarenakan setiap bahasa memiliki karakteristik masing-masing yang menjadi ciri khasnya. Oleh karena itu, untuk membuat model dengan spesifik suatu bahasa, maka harus digunakan juga *dataset* secara spesifik menggunakan bahasa tersebut. Saat ini salah satu *dataset* yang sudah memiliki banyak sumber untuk dilakukan penelitian adalah *dataset* berbahasa Inggris. Hasil dari pelatihan terhadap model dengan *dataset* berbahasa Inggris juga sudah terbukti memiliki hasil yang baik. Salah satu *dataset* Bahasa Inggris yaitu LJSpeech yang merupakan *dataset* yang digunakan oleh model Tacotron2 dalam memproses teks menjadi suara sintesis [3]. *Dataset* tersebut terbukti mendukung model Tacotron2 menjadi *state-of-the-art* saat ini dalam pemrosesan teks menjadi suara dalam Bahasa Inggris [4].

Pada kenyataannya, model Tacotron2 tersebut tidak bisa langsung diaplikasikan ke bahasa lain untuk memproses teks menjadi suara karena perbedaan karakteristik bahasa yang dimiliki. Maka dari itu, dibutuhkan *dataset* baru yang lebih spesifik untuk bahasa yang diinginkan dalam melakukan pelatihan pada model *deep learning*. Salah satu contoh spesifik yaitu *dataset* dengan Bahasa Indonesia. Saat ini belum tersedia secara publik *dataset* yang berupa suara dan teks dalam Bahasa Indonesia. Oleh karena itu, pada tulisan ini akan dijelaskan tahapan pembentukan *dataset* yang merupakan hasil rekaman *podcast* dan *talk show* berbahasa Indonesia beserta transkripsi rekaman suara tersebut. Dengan tersedianya *dataset* tersebut, diharapkan dapat digunakan untuk melakukan pelatihan terhadap model *deep learning* yang lebih spesifik untuk Bahasa Indonesia. Kemudian model yang telah dilatih tersebut akan digunakan oleh model Tacotron2 untuk melakukan pemrosesan teks menjadi suara yang memiliki kualitas konversi yang lebih baik untuk Bahasa Indonesia.

DATASET SUARA DAN TEKS BAHASA INDONESIA

Dataset merupakan merupakan kumpulan dari suatu data yang menggambarkan suatu topik tertentu. Tulisan ini membahas *dataset* secara spesifik yaitu *dataset* suara dan teks dalam Bahasa Indonesia. Pembuatan *dataset* ini ditujukan karena kurangnya sumber untuk mendapatkan *dataset* secara spesifik yang publik dalam Bahasa Indonesia. Dalam pembuatannya, *dataset* ini diharapkan dapat membantu pembuatan model *deep learning* dengan Bahasa Indonesia yang diperuntukkan sebagai model yang memproses teks menjadi suara ataupun teks menjadi suara. Hal tersebut mengacu pada kondisi yang mana setiap bahasa memiliki keunikan dan perbedaan masing-masing, sehingga *dataset* ini dibutuhkan dalam pembentukan suatu model *deep learning* untuk memproses teks menjadi suara.

Bahasa merupakan ekspresi dari pikiran dan perasaan manusia yang menggunakan suara dalam penyampaiannya [5]. Setiap negara memiliki perbedaan bahasa dengan karakteristiknya masing-masing. Tidak hanya antar negara yang memiliki perbedaan bahasa, bahkan di dalam satu negara itu sendiri, di setiap daerah terkadang memiliki karakteristik masing-masing. Di Indonesia, dengan keberagaman bahasa yang berbeda-beda di setiap pulau ataupun daerahnya seperti Bahasa Jawa, Bahasa Sunda, Bahasa Batak, Bahasa Madura, dan lain-lain. Namun keberagaman tersebut dipersatukan dengan bahasa nasional yaitu Bahasa Indonesia itu sendiri [5]. Maka dari itu, untuk membuat model *deep learning* dalam pemrosesan teks menjadi suara yang mewakili bahasa dari negara Indonesia, digunakan langsung bahasa nasional dari Indonesia yaitu Bahasa Indonesia.

2.1. Properti Data

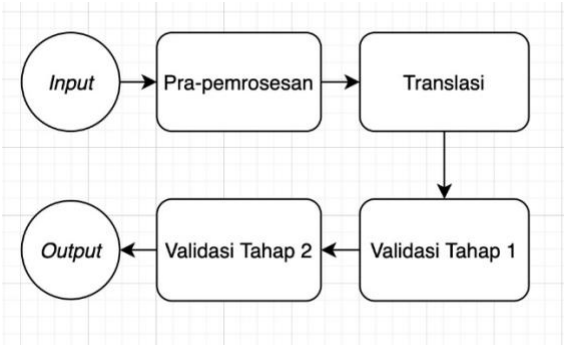
Dataset ini diambil dari rekaman *podcast* dan *talk show* berbahasa Indonesia. *Dataset*

rekaman suara dari *podcast* dan *talk show* dipilih karena beberapa pertimbangan seperti akses untuk mendapatkan rekaman suara yang mudah karena rekaman tersebut dapat di akses secara publik, kemudian isi dari rekaman suara yang terdapat pada *podcast* dan talk show merupakan pembicaraan natural yang terarah. Alasan lainnya pemilihan rekaman dari *podcast* dan *talk show* adalah dikarenakan *podcast* dan *talk show* saat ini merupakan *platform* yang populer dan banyak didengarkan, sehingga setiap pembahasan dapat diikuti dengan baik oleh pendengar.

Data rekaman terdiri dari 3270 rekaman, yang mana terdapat 2116 rekaman dari *podcast* dan 1154 rekaman dari *talk show*. Panjang rekaman dikelompokkan menjadi 2, yaitu rekaman dengan durasi 0-3 detik serta rekaman dengan durasi waktu 3-6 detik. Jumlah durasi dari rekaman adalah selama 3 jam 14 menit. *Dataset* ini merupakan *dataset* dengan pembicaraan berdasarkan kondisi yang sesungguhnya. Kondisi yang dimaksud merupakan hasil dari komunikasi antara pembicara dengan suatu topik pembahasan spesifik yang menghasilkan pembicaraan untuk saling bertukar pikiran atau pendapat secara langsung. Pembicaraan pada rekaman tersebut berjalan secara natural tanpa membuat suatu skenario tertentu untuk dibicarakan.

2.2. Prosedur Pembuatan

Dalam prosedur pembuatan transkrip *dataset* dari rekaman suara *podcast* dan *talk show* dilakukan beberapa tahapan. Tahapan tersebut seperti pra-pemrosesan, pembuatan transkrip dari mendengarkan rekaman suara secara langsung, validasi pertama terhadap hasil transkrip, validasi kedua sebagai pemeriksaan terakhir. Urutan dari tahapan-tahapan tersebut dapat dilihat pada Gambar 1 dibawah ini.



Gambar 1. Diagram Proses Pembuatan *Dataset*

2.2.1. Pra-pemrosesan

Hal pertama yang dilakukan dalam pembuatan *dataset* suara dan transkrip suara untuk kebutuhan pelatihan terhadap model yang memproses teks menjadi suara dengan Bahasa Indonesia adalah pra-pemrosesan. Pra-pemrosesan yang dilakukan adalah dengan mengeliminasi beberapa rekaman yang dinilai kurang jelas atau tidak menggambarkan *dataset* yang diinginkan. Salah satu pertimbangan yang diambil sebagai alasan terhadap dilakukannya tahap pra-pemrosesan yaitu jikalau rekaman tersebut tetap diikutsertakan maka akan membuat tingkat kebenaran dari *dataset* berkurang atau bahkan dapat menghasilkan penilaian yang salah. Berikut beberapa kriteria yang di eliminasi sebelum proses pembuatan transkrip dari setiap rekaman suara:

1. Rekaman Suara yang Mengandung Bahasa Inggris
   Kategori ini dipilih karena fokus untuk *dataset* yang akan dibuat adalah *dataset* dengan Bahasa Indonesia. Maka dari itu, jika terdapat rekaman suara dan transkrip suara yang mengandung kalimat atau kata berbahasa Inggris, akan membuat data menjadi tidak spesifik dan membuat data menjadi bias. Contoh dari kasus tersebut seperti kalimat "*why people doing that*", "*work* betul-betul", atau "*you must assume that you are sick*".
2. Rekaman yang Mengandung Suara Tidak Jelas
   Kategori ini dipilih karena dalam memproses rekaman suara menjadi transkripsi suara tentu akan ditemukan beberapa kata yang tidak terdengar dengan jelas ataupun suara yang sangat kecil dari rekaman tersebut sehingga peneliti tidak bisa mendeskripsikan suara tersebut menjadi suatu transkrip dari suara yang didengarkan.
3. Rekaman Suara yang Mengandung Kata yang Sulit untuk Dituliskan Transkripnya
   Kategori ini dipilih karena dalam suatu rekaman terdapat *non-word* yang disebutkan, dimana kata tersebut terdengar hanya satu huruf namun memiliki durasi penyebutan yang kadang panjang dan kadang pendek. Selain karena sulit dituliskan, alasan yang juga menjadikan kategori ini dapat mengeliminasi suatu rekaman adalah karena kesulitan dalam memproses suara

tersebut menjadi transkrip. Hal tersebut akan membuat *dataset* menjadi tidak konsisten untuk dijadikan acuan dalam pelatihan. Contoh kasus yang termasuk dalam kategori ini seperti penyebutan "eeeeee eeeee", "aaa aaaaaa", atau "taa teteeetapi".

Dari ketiga kategori di atas, jika ditemukan suatu rekaman mengandung kategori tersebut maka rekaman tidak akan diikutsertakan untuk menjadi *dataset*. Dari tahapan pra-pemrosesan ini menghasilkan eliminasi data, dimana pada sebelum pra-pemrosesan terdapat 3270 rekaman dan setelah pra-pemrosesan terdapat 2560 rekaman yang tersisa untuk diproses. Selain dari kategori-kategori tersebut, sebelum membuat transkrip dari rekaman yang ada, dibuat beberapa aturan yang mengacu dari aturan-aturan yang ada pada data LJSpeech untuk proses translasi suara menjadi transkrip dari suara tersebut [6]. Seperti penyebutan angka yang direpresentasikan dalam bentuk kata dari angka tersebut, misalnya angka 19 akan dituliskan menjadi "sembilan belas". Hal tersebut akan mempermudah proses penghilangan karakter-karakter khusus (ASCII) sehingga dapat dihasilkan data yang lebih bersih. Kemudian dalam penulisan transkrip tidak terdapat tulisan yang disingkat, seperti "yg", atau "dg", dan lain-lain. Semua hasil dari translasi dituliskan langsung sesuai dengan suara yang didengar pada rekaman.

2.2.2. Proses Translasi Suara Menjadi Transkrip

Untuk membuat transkrip dari rekaman suara dilakukan beberapa tahapan untuk memastikan proses translasi memiliki tingkat kebenaran yang baik. Berikut tahapan-tahapan yang dilakukan dalam proses translasi rekaman suara menjadi suatu transkrip:

1. Tahapan Translasi
   Tahapan pertama ini memproses *input* berdasarkan pemilihan data setelah melalui pra-pemrosesan. Sehingga data yang masuk sudah bersih dari parameter-parameter yang telah diterapkan. Tahapan ini dilakukan dengan mendengarkan rekaman secara seksama untuk kemudian langsung menuliskan hasil dari suara yang didengar menjadi transkripsi yang sesuai.
2. Tahapan Validasi Pertama
   Tahapan validasi pertama ini sebagai *filter* utama dari hasil transkripsi. Tahapan ini dikerjakan oleh orang yang berbeda dari orang yang melakukan translasi sebelumnya. Pada tahapan ini dilakukan pemeriksaan dengan cara mendengarkan ulang setiap rekaman suara dan menuliskan transkrip yang baru. Hasil tersebut akan dibandingkan dengan hasil dari transkrip pada proses translasi di tahapan pertama. Untuk menyatakan valid atau tidaknya suatu transkrip dapat diperoleh berdasarkan hasil perbandingan yang dilakukan. Jika transkrip pada proses translasi di tahapan pertama sama dengan transkrip yang dihasilkan pada tahapan validasi ini, maka transkrip yang dibuat dapat disimpulkan sebagai suatu transkrip yang valid. Namun apabila terdapat perbedaan dari perbandingan hasil transkrip, maka rekaman suara akan diputar ulang untuk memastikan kembali transkrip yang paling benar. Hasil pemeriksaan ulang tersebut akan dijadikan acuan untuk memperbaiki hasil transkrip yang sebelumnya agar menjadi transkrip yang valid.
3. Tahapan Validasi Kedua
   Pada tahapan ini dilakukan validasi akhir oleh penulis untuk memastikan kebenaran hasil transkrip dari proses sebelumnya. Validasi yang dimaksud dilakukan dengan cara memeriksa penulisan dari hasil transkrip. Beberapa kategori yang digunakan seperti melihat apakah terdapat huruf yang tertinggal dalam penulisan suatu kata, atau apakah ada huruf yang tertukar posisinya dalam penulisan suatu kata, serta apakah terdapat suatu kata yang secara KBBI merupakan kata yang salah. Jika pada suatu transkrip tidak ditemukan kategori pemeriksaan yang dimaksud, maka transkrip tersebut dapat disebut sebagai transkrip yang valid. Namun jika pada suatu transkrip ditemukan kategori yang dimaksud, maka rekaman dari suara akan diputar kembali, kemudian transkrip akan diperbaiki penulisannya untuk memastikan kebenaran hasil transkrip menjadi lebih valid dari sisi penulisan.

Tahapan tersebut dilakukan untuk membuat hasil dari *dataset* memiliki tingkat kebenaran yang baik agar dapat dijadikan acuan yang benar dalam proses pelatihan model pada *deep*

learning dalam memproses teks menjadi suara pada Bahasa Indonesia.

HASIL DAN PEMBAHASAN

Dataset pada tulisan ini terdiri dari dua domain yang berbeda yaitu domain suara dan domain teks dalam bentuk transkripsi dari rekaman suara. Penulisan transkrip dilakukan dengan mengikuti format dari *dataset* LJSpeech, dimana pada setiap hasil transkrip diikutsertakan suatu kata kunci sebagai penunjuk kepada alamat dari rekaman yang sesuai. Hal tersebut akan lebih memudahkan pemrosesan *dataset* untuk dijadikan *input* pada suatu pembelajaran. Hasil transkrip yang dimaksud ditulis dalam dua bagian yaitu bagian alamat dari *file* rekaman suara (*path*) dan bagian teks dari transkrip itu sendiri. Kedua bagian tersebut dipisahkan oleh karakter "|". Maka dari itu, dalam pemrosesan untuk pelatihan suatu model yang menggunakan *dataset* ini harus mempertimbangkan format dari hasil transkrip yang ada. Berikut beberapa contoh dari hasil transkripsi:

1. pc/0/netral/train/fix0 -pc_2.312.wav|berpindah dari satu ke yang lainnya
2. pc/0/netral/train/fix0-pc_3_f.27.wav|itu tadi pola hidup bersih dan sehat
3. ts/0/marah/train/fix0-ts_1_f.159.wav|komisi pemberantasan korupsi

Dari contoh diatas dapat dilihat bahwa seperti pada contoh pertama yaitu bagian "pc/0/netral/train/fix0 -pc_2.312.wav" mewakili alamat dari *file* rekaman yang dipilih. Kemudian pada bagian setelah karakter "|" yaitu "berpindah dari satu ke yang lainnya" Merupakan hasil transkrip dari rekaman suara pada alamat *file* yang ditunjuk. Kemudian hal lain yang dilakukan adalah mengubah format audio rekaman *dataset* dari *stereo* menjadi *monostereo* sesuai dengan format dari LJSpeech menggunakan *library sox* [7]. Selain dapat digunakan untuk pemrosesan oleh model Tacotron2 yang spesifik untuk pemrosesan teks menjadi suara, *dataset* ini juga dapat digunakan untuk pelatihan model-model pemrosesan lain seperti pemrosesan suara menjadi teks. Hal tersebut dikarenakan format data yang disusun dalam *dataset* ini cukup umum dan dapat dimodifikasi sesuai dengan kebutuhan model-model lain yang ingin dilatih menggunakan *dataset* ini. *Dataset* dapat diakses pada link berikut bit.ly/dataSuaraDanTeks. Dalam penggunaan *dataset* untuk dijadikan sebagai *input*, dapat dilakukan dengan mengunduh data dan menyesuaikan ulang alamat dari data rekaman pada transkrip agar bisa diakses oleh kode yang akan memproses *input* tersebut. Setiap rekaman disimpan dalam format WAV dan transkrip disimpan pada dokumen teks.

Dataset teks dan suara berbahasa Indonesia dari rekaman *podcast* dan *talk show* ini telah digunakan untuk pelatihan pembentukan model Tacotron2 dalam memproses suara pada Bahasa Indonesia [8]. Pelatihan dilakukan menggunakan arsitektur Tacotron2 yang sama dan sesuai dengan penelitian sebelumnya [4,9]. Hal yang berbeda hanya terdapat pada *dataset* yang digunakan. Dalam pembentukan model Tacotron2 berdasarkan *dataset* Bahasa Indonesia, dibutuhkan durasi waktu 3 jam 41 menit 30 detik. Setelah model terbentuk, dilakukan evaluasi menggunakan metode *Mean Opinion Score* (MOS) terhadap suara yang dihasilkan. Evaluasi dengan MOS dipilih karena belum adanya *ground truth* dalam menguji atau memberikan penilaian terhadap suatu suara yang mana yang paling baik [10]. Evaluasi dilakukan dengan menggunakan kuesioner yang melibatkan 25 responden. Dari evaluasi tersebut diperoleh nilai rata-rata MOS sebesar 4.01. Dengan nilai MOS tersebut, menunjukkan bahwa suara yang dibentuk oleh model Tacotron2 berdasarkan pelatihan menggunakan *dataset* teks dan suara berbahasa Indonesia dari rekaman *podcast* dan *talk show* cukup signifikan. Hal tersebut membuktikan bahwa *dataset* yang dibentuk dapat digunakan pada pemrosesan teks menjadi suara dan menghasilkan kualitas suara yang baik.

SIMPULAN DAN SARAN

Pada tulisan ini dipublikasikan *dataset* teks dan suara berbahasa Indonesia berupa rekaman suara beserta transkrip dari suara tersebut. *Dataset* rekaman diperoleh dari *podcast* dan *talk show* yang kemudian dibuatkan transkrip dari rekaman suara tersebut. Terdapat beberapa tahapan dalam proses pembuatan transkrip, seperti tahapan pra-pemrosesan, tahapan translasi, dan dua tahapan validasi untuk memastikan tingkat kebenaran dari transkrip lebih baik. Setelah *dataset* berhasil dibentuk,

dataset diuji dengan cara digunakan pada pelatihan untuk pembentukan model Tacotron2 dalam pemrosesan teks menjadi suara pada Bahasa Indonesia. Dari pelatihan tersebut menghasilkan kualitas suara yang cukup baik dengan nilai MOS sebesar 4.01. Dengan nilai MOS tersebut dapat disimpulkan bahwa *dataset* yang dibentuk cukup signifikan, karena berhasil memproses teks menjadi suara pada Bahasa Indonesia dengan kualitas yang baik pada model Tacotron2.

Dalam proses yang dilakukan untuk membuat transkrip dari rekaman suara, penulis menemukan beberapa kesulitan seperti pelafalan dari suatu kata yang kurang jelas serta terdapat kata yang disebutkan dengan terbata-bata. Selain itu, juga terdapat keberagaman pembicara seperti ada pembicara pria dan wanita kemudian pembicara dari berbagai daerah yang berbeda sehingga membuat aksen dari suatu kata berubah-ubah. Kemudian juga terdapat beberapa rekaman yang mengandung kata Bahasa Inggris sehingga membuat rekaman tersebut harus dieliminasi dari *dataset*. Pada awalnya data suara memiliki durasi selama 3 jam 14 menit dan memiliki rekaman sejumlah 3270 rekaman. Karena adanya tahapan pra-pemrosesan, maka beberapa rekaman juga dihilangkan dari *dataset* yang menyebabkan berkurangnya jumlah dari rekaman yang terdapat dalam *dataset*.

DAFTAR PUSTAKA

[1] Goodfellow, I., Bengio, Y. and Courville, A., 2016. *Deep learning*. MIT press.
[2] Ferdiana, R., Jatmiko, F., Purwanti, D.D., Ayu, A.S.T. and Dicka, W.F., 2019. *Dataset* Indonesia untuk Analisis Sentimen. Jurnal Nasional Teknik Elektro dan Teknologi Informasi (JNTETI), 8(4), pp.334-339
[3] Kuligowska, K, Kisielewicz, P. and Wlodarz, A. (2018) Speech synthesis systems: disadvantages and limitations, International Journal of Engineering & Technology, [S.l.], v. 7, n. 2.28, p. 234–239.
[4] Shen, J., Pang, R., Weiss, R.J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R. and Saurous, R.A., 2018, April. *Natural tts synthesis by conditioning wavenet on mel spectrogram predictions*. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 4779-4783). IEEE.
[5] Cahyaningtyas, E. and Arifianto, D., 2017, December. *Development of under-resourced Bahasa Indonesia speech corpus*. In 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC) (pp. 1097-1101). IEEE=
[6] Keith Ito and Linda Johnson. The lj speech dataset.https://keithito.com/LJ-Speech-Dataset/, 2017
[7] Bagwell, C. (2015) SoX. Available at: http://sox.sourceforge.net/sox.html (Accessed: 14 July 2021).
[8] Novela, M. (2021). Pemrosesan Teks menjadi Suara menggunakan Model Tacotron2 berdasarkan *Dataset* Rekaman dari *Podcast* dan *Talk Show* Berbahasa Indonesia. Tesis. Program Magister Universitas Indonesia. Depok
[9] Oord, A. V. D., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., ... & Kavukcuoglu, K. (2016). Wavenet: A generative model for raw audio. arXiv preprint arXiv:1609.03499.
[10] Haryanto, H., & Sumpeno, S. (2018). *A Realistic Visual Speech Synthesis for Indonesian Using a Combination of Morphing Viseme and Syllable Concatenation Approach to Support Pronunciation Learning*. International Journal of Emerging Technologies in Learning, 13(8).