1. Pendahuluan

Salah satu tempat dimana text-to-speech sering digunakan adalah pada media live streaming seperti Twitch dan Youtube. Streamer menggunakan TTS untuk membacakan pesan dari donasi yang mereka terima dan sebagai salah satu cara untuk streamer dan penonton berkomunikasi. Menggunakan suara yang berbeda-beda ataupun menggunakan TTS dengan bahasa yang berbeda merupakan hal yang dilakukan oleh streamer untuk menambahkan hiburan pada stream mereka.

Pada paper [12] mengajukan sistem TTS yang mampu mengkloning suara dan mengucapkan teks dengan fasih litas bahasa [12]. TTS yang digunakan untuk hiburan, membuat kesan suara seperti orang asing berbicara Bahasa Indonesia memiliki pesona yang berbeda dengan TTS yang mampu mengucapkan dengan fasih. Kesan orang asing berbicara Bahasa Indonesia bisa didapatkan dengan cara pengucapan dan aksen yang tidak cocok [12]. Pengucapan yang tidak sesuai dapat berpengaruh pada kemampuan pendengar untuk memahami teks yang diucapkan sistem. Untuk mendapatkan aksen orang asing dan pendengar mampu memahami teks yang diucapkan oleh sistem, penulis mengajukan untuk mengganti silabel berdasarkan fonetiknya. Dengan begitu penyebutan masih mirip untuk pendengar dapat menangkap pesan yang diucapkan sistem serta aksen orang asingnya masih bertahan.

Soundex merupakan sistem yang dibangun untuk pencocokan fonetik, digunakan dalam pencarian nama yang memiliki penyebutan yang mirip [1]. Nama pada databe kemungkinkan ada yang ejaannya salah atau ejaan yang tidak sesuai sehingga pencarian nama jika berdasarkan ejaan yang tepat tidak memberikan hasil. Soundex dapat mencari kata dengan penyebutan yang mirip sehingga penulis menggunakan Soundex untuk mengatasi masalah silabel yang tidak dikenali oleh sistem TTS. Silabel yang tidak dikenali oleh sistem tidak akan dibunyikan seperti silabel ”ril” sehingga diberikan silabel pengganti ”reel” karena pengucapannya mirip.

Sampel suara yang akan dikloning sistem dapat bersumber dari cuplikan film, audiobook, dan sumber lainnya. Setiap sampel suara yang akan digunakan memiliki sample rate yang berbeda tergantung dari sumbernya seperti rekaman suara dengan format Audio CD menggunakan sample rate 44100 Hz [2] dan Sample rate untuk suara yang dilatih pada sistem menggunakan sample rate 16000 Hz. TTS yang digunakan pada [5] dan [12] dilatih menggunakan korpus librispeech dengan sample rate 16 kHz. Penggunaan sample rate yand rendah memiliki keuntungan dalam storage untuk menyimpan korpus dan juga memudahkan pada speech processing [7]. Kekurangan menggunakan sample rate yang rendah adalah tidak dapat menangkap fitur khas dari suara dengan baik seperti kompenen dengan frekuensi yang tinggi [8]. Dari hasil pada [4] menunjukkan bahwa semakin tinggi sample rate yang digunakan, maka hasil hasil rekognisi suara semakin lebih bagus. Penggunaan sample rate yang tinggi untuk kloning suara pada TTS hasilnya bisa berbeda dengan penggunaan pada speech recognition.

Latar Belakang

Berdasarkan paper [12] aksen dapat ditahan karena adanya ketidak sesuaian pengucapan namun hasilnya MOS menjadi lebih kecil. Kefasihan dalam pengucapan akan menghilangkan aksen namun MOS yang didapat tinggi. Untuk mempertahankan aksen dengan MOS yang tinggi, pengucapan tidak harus fasih namun setidaknya sangat mirip atau mendekati. Soundex dapat mencari silabel yang memiliki penyebutan yang mirip [1] sehingga dapat digunakan untuk meraih TTS yang mampu mempertahankan aksen dan MOS yang tinggi.

Kloning suara pada TTS memerlukan sampel suara sebagai input. Sampel suara yang digunakan dapat memiliki sample rate yang berbeda tergantung sumbernya. Sample rate yang rendah tidak mampu menangkap frekuensi tinggi [8]. Frekuensi tinggi tersebut bisa saja merupakan fitur khas pada suara dan itu dapat berakibat suara yang dikloning tidak mirip dengan sampel. Penulisan ini melihat pengaruh sample rate terhadap kemiripan suara dan pengucapan.

Topik dan Batasannya

Penulisan ini melihfat bagaimana pengaruh pada sample rate suara sampel dengan kemiripan suara pada hasil dan daya tangkap pendengar terdahap hasil output dari sistem. Sistem yang dibangun berbasis [5] dan menerapkan Soundex untuk lintas bahasa. Sistem yang dibangun berusaha mengucapkan teks mirip dengan bunyi yang seharusnya untuk membertahankan aksen dan pesan yang diucap dapat ditangkap pendengar. Jumlah silabel yang diganti dapat berpengaruh terhadap pesan yang ditangkap oleh pendengar menjadi salah. Silabel pada Bahasa Indonesia dan Bahasa Inggris banyak perbedaan dan memiliki perbedaan pengucapan. Penggantian silabel diperlukan untuk menghindari silabel yang tidak dikenali oleh sistem karena sistem jika ditemukan sistem tidak menghasilkan suara dan silabel dilewati. Adanya voice cloning pada TTS memberikan opsi untuk menggunakan suara siapa saja dan dari mana saja. Keterbukaan itu dapat membuat pengguna menggunakan sample rate yang berbeda beda.

Tujuan

Penulisan ini bertujuan untuk melihat pengaruh perbedaan sample rate yang digunakan pada sampel suara untuk dikloning terhadap kemiripan suara hasil dan untuk mengetahui apakah pendengar mampu menangkap pesan yang diucapkan oleh sistem. Penulisan ini juga bertujuan untuk melihat apakah jumlah kata berpengaruh terhadap pendengar dalam menangkap pesan yang diucapkan oleh sistem.