PENGGUNAAN MODEL TRANSFORMER PADA *AUDIOVISUAL SPEECH RECOGNITION* UNTUK BAHASA INDONESIA

TESIS

Karya tulis sebagai salah satu syarat
untuk memperoleh gelar Magister dari
Institut Teknologi Bandung

Oleh

GUGY LUCKY KHAMDANI
NIM: 23517041
(Program Studi Magister Informatika)

PROGRAM STUDI MAGISTER INFORMATIKA

SEKOLAH TEKNIK ELEKTRO DAN INFORMATIKA

INSTITUT TEKNOLOGI BANDUNG

Februari 2019

Bab I Pendahuluan

Bab Pendahuluan menjelaskan latar belakang penelitian, rumusan masalah, tujuan, batasan, ruang lingkup, dan metodologi yang diterapkan pada penelitian ini.

I.1 Latar Belakang

Automatic speech recognition (ASR) adalah proses pengenalan atau penerjemahan bahasa lisan dalam bentuk sinyal audio menjadi teks secara otomatis oleh komputer. Salah satu permasalahan pada ASR adalah pengenalan menjadi sulit jika dilakukan di lingkungan yang bising, terutama apabila pengenalan dilakukan hanya dengan berbasis audio. Sedangkan, manusia memanfaatkan informasi suara dan informasi visual berupa gerakan bibir dalam melakukan pengenalan ucapan (Calvert dkk., 2004). Oleh sebab itu, penambahan informasi visual dalam sistem pengenalan ucapan diharapkan bisa dilakukan untuk meningkatkan akurasi pengenalan ucapan secara umum. Selain itu, informasi visual ini bisa diaplikasikan menjadi sebuah sistem pengenalan gerak bibir dan digunakan untuk memberikan instruksi atau pesan kepada komputer di lingkungan yang bising (Garg dkk., 2016), mentranskripsikan kata-kata yang diucapkan pada film-film bisu atau video tanpa audio, menyelesaikan permasalahan pengenalan suara pada pembicara lebih dari satu, dan juga dapat meningkatkan performa dari sistem pengenalan suara secara umum (Chung dkk., 2017).

Ada dua jenis pendekatan yang paling banyak dilakukan saat ini dalam melakukan pengenalan ucapan melalui gerak bibir, yaitu pendekatan yang memodelkan kata-kata (Wand dkk., 2016) dan pendekatan yang memodelkan *viseme* (Chung dkk., 2017). *Viseme* merupakan satuan terkecil dalam sebuah bahasa yang masih bisa menunjukkan perbedaan kata pada suatu video. Jika fonem merupakan satuan terkecil dalam bentuk bunyi, maka *viseme* setara dengan bentuk visualnya. Dalam riset Arifin dkk. (2013), berfokus pada pembangunan *viseme* dalam bahasa Indonesia dengan cara melakukan *clustering* menggunakan K-Means pada data berisi gambar *speech* visual. Hasil riset tersebut menunjukkan bahwa dalam bahasa Indonesia terdapat 10 kelas viseme.

Hingga saat ini, riset mengenai pengenalan gerak bibir untuk bahasa Indonesia masih terbilang sedikit dibandingkan dengan bahasa-bahasa lain seperti bahasa Inggris, dan untuk bahasa tersebut pun masih sedikit yang menggunakan *deep learning*. Oleh sebab itu, riset-riset tersebut membutuhkan pra-proses yang cukup banyak untuk mengekstraksi fitur dari gambar frame-frame di video, dan juga pra-proses secara temporal menggunakan *optical flow* atau deteksi gerakan untuk mengekstraksi fitur video, atau menggunakan metode berbasis aturan (*rule-based*) lainnya, seperti yang dijelaskan lebih mendalam pada riset Zhou dkk. (2014). Untuk yang berbahasa Indonesia terdapat riset Achmad and Fadillah (2015) yang menggunakan Hidden Markov Model berdimensi satu untuk modul pengenalan polanya, tetapi masih belum tergeneralisasi dengan baik karena hasilnya masih berpengaruh pada kondisi bibir pembicara, yang dalam hal ini pembicara wanita dengan bibir yang menggunakan lipstik memiliki koefisien korelasi yang tinggi sedangkan untuk yang bibir berwarna pucat dan bibir yang memiliki kumis di atasnya memiliki koefisien korelasi yang rendah. Data yang digunakan berjumlah 25 video data yang masing-masing berisi data pembicara yang berbeda dan data tersebut dibuat khusus untuk riset ini. Penggunaan *deep learning* membuat data yang diperlukan menjadi sangat besar, akan tetapi sejauh ini belum ditemukan dataset untuk bahasa Indonesia yang berukuran besar yang seragam digunakan untuk lebih dari satu riset, sehingga timbul keperluan untuk membangun dataset dari awal dengan ukuran besar.

Pengenalan gerak bibir merupakan permasalahan yang sulit karena membutuhkan ekstraksi fitur spatiotemporal dari video, karena posisi dan gerakannya merupakan informasi yang penting. Dengan adanya perkembangan dalam *deep learning*, pada beberapa tahun terakhir ada beberapa upaya dalam mengaplikasikan *deep learning* ke permasalahan pengenalan gerak bibir *lipreading*) ini, seperti oleh Noda dkk. (2014) yang mempelajari fitur visual dengan menggunakan *convolutional neural network* yang kemudian digunakan GMM-HMM untuk mengklasifikasikan fonem.

Diinspirasi dari perkembangan terkini pada permasalahan mesin translasi dalam memodelkan *sequence-to-sequence* menggunakan model *encoder-decoder* yang di-

lengkapi dengan mekanisme *attention* (Bahdanau dkk., 2015), model *encoder-decoder* ini kemudian sudah diaplikasikan ke berbagai macam permasalahan lain seperti *speech recognition* (Chan dkk., 2015), *automatic image captioning* (Vinyals dkk., 2014) (Xu dkk., 2015), dan pengenalan gerak bibir (Chung dkk., 2017). Model ini mengambil masukan berupa rangkaian S dengan panjang m yang kemudian dipetakan menjadi rangkaian T dengan panjang n. Rangkaian T dihasilkan dari *hidden state* h_t yang merupakan fungsi dengan masukan h_{t-1} dan rangkaian S untuk *time-step* ke t. Karena sifatnya yang sekuensial, membuat paralelisasi pada saat proses pelatihan model menjadi tidak bisa dilakukan, sehingga prosesnya menjadi sangat lama terutama pada data latih yang memiliki rangkaian yang sangat panjang, juga dikarenakan terbatasnya ukuran memori jika dilakukan proses pelatihan dengan mode batch.

Mekanisme *attention* sudah diaplikasikan pada berbagai permasalahan yang menggunakan model *encoder-decoder*, dan telah menjadi bagian penting dalam pemodelan rangkaian dan model transduksi. Mekanisme *attention* ini memungkinkan bagian *decoder* untuk dapat melihat keseluruhan rangkaian masukan dan menilai seberapa penting bagian dari rangkaian masukan tersebut. Akan tetapi, kebanyakan pengaplikasian mekanisme *attention* ini hanya sebatas digunakan sebagai pelengkap untuk jaringan saraf rekuren. Oleh sebab itu Vaswani dkk. (2017) mengusulkan model yang dinamakan transformer, sebuah arsitektur model yang menghindari penggunaan rekurens dan bergantung secara penuh pada mekanisme *attention* untuk menggambarkan dependensi global antara masukan dan keluaran. Selain itu model transformer ini memungkinkan dilakukannya paralelisasi sehingga dapat mempercepat proses pelatihan model.

Hal tersebut menjadi latar belakang dari tesis ini. Secara umum, tesis ini akan mencoba untuk mengaplikasikan model transformer pada permasalahan pengenalan gerak bibir untuk meningkatkan performa *speech recognition* dalam bahasa Indonesia. Selain pengaplikasian model, tesis ini juga berfokus pada pengumpulan data untuk pengenalan ucapan yang dilengkapi dengan gerak bibir dalam bahasa Indonesia.

I.2 Rumusan Masalah

Riset mengenai pengenalan ucapan otomatis pada bahasa Indonesia sudah banyak dilakukan, akan tetapi kebanyakan masih memerlukan praproses untuk mereduksi *noise*. Riset mengenai pengenalan gerak bibir pada bahasa Indonesia juga sudah dilakukan meski masih memiliki akurasi pengenalan yang belum baik jika dibandingkan dengan bahasa yang sudah banyak diriset, seperti bahasa Inggris. Hal ini disebabkan oleh keterbatasan sumber daya dan penggunaan teknik pengenalan yang kurang optimum. Selain itu riset mengenai penggabungan fitur akustik dan fitur visual berupa gerak bibir dalam mengenali ucapan pada bahasa Indonesia belum ada yang melakukan. Oleh karena itu, pada tesis ini diusulkan solusi berupa pembangunan sistem pengenalan ucapan dengan menggabungkan fitur akustik dan fitur visual berupa gereak bibir dengan menggunakan pendekatan *deep learning* seperti model *sequence-to-sequence* dan berbagai macam variannya. Penggunaan pendekatan yang lebih baik dan penambahan fitur visual ini diharapkan memberikan hasil pengenalan yang lebih baik dan meningkatkan akurasi pengenalan.

I.3 Tujuan

Tujuan utama dari tesis ini dirincikan sebagai berikut,

1. Membangun sistem pengenalan suara dengan menggunakan fitur akustik dan fitur visual berupa pengenalan gerak bibir pada kalimat bahasa Indonesia dengan menggunakan model transformer. Selain itu juga membangun sistem pengenalan suara yang sama tapi hanya menggunakan fitur akustik yang selanjutnya digunakan sebagai model *baseline*.

2. Melakukan perbandingan kinerja sistem pengenalan suara yang menggunakan fitur akustik dan fitur visual dengan model *baseline* yang hanya menggunakan fitur akustik saja.

3. Mengumpulkan atau membuat data pengenalan ucapan yang dilengkapi dengan gerak bibir dalam bahasa Indonesia baku.

I.4 Batasan Masalah

Penelitian ini hanya berfokus pada pengenalan ucapan pada kalimat-kalimat bahasa Indonesia baku.

I.5 Metodologi

Metodologi yang diterapkan pada pengerjaan penelitian ini adalah:

1. Analisis permasalahan. Pada tahap ini akan dilakukan analisis berdasarkan studi literatur untuk menentukan masalah-masalah pada penggunaan model transformer dan juga mengidentifikasi permasalahan-permasalahan yang terdapat pada *speech recognition*.

2. Perancangan solusi. Pada tahap ini akan dilakukan penentuan arsitektur yang tepat dan model-model yang akan dijadikan *baseline* untuk memetakan rangkaian frame video menjadi rangkaian kata.

3. Pengumpulan dataset untuk pelatihan model, dalam bentuk video yang berisi gambar dan suara orang yang mengucapkan kalimat dalam bahasa Indonesia baku.

4. Pembangunan model textitbaseline dan gabungan model pengenalan ucapan berbasis akustik dan model pengenalan ucapan berbasis visual berupa gerak bibir, lalu melakukan pelatihan model serta perbandingan antara model tersebut.

5. Tahapan akhir dari penelitian ini adalah melakukan analisis hasil dan membuat kesimpulan dari hasil eksperimen.

I.6 Sistematika Pembahasan

Laporan tesis ini disusun berdasarkan sistematika berikut:

Bab I Pendahuluan berisi latar belakang, rumusan masalah, tujuan, dan batasan yang diterapkan pada penelitian, serta metodologi pengerjaan dan sistematika pembahasan penelitian yang disajikan dalam laporan tesis ini.

Bab II Tinjauan Pustaka berisi penjelasan mengenai konsep dan dasar teori dari pengenalan ucapan, baik berbasis akustik maupun berbasis visual, Diberikan juga penjelasan mengenai arsitektur jaringan saraf tiruan yang digunakan, yang termasuk di dalamnya yaitu model *sequence-to-sequence* dan transformer.

Bab III Analisis Masalah dan Rancangan Solusi memberikan analisis awal terhadap kondisi data, gambaran umum skema eksperimen, dan pertimbangan solusi dalam mengatasi masalah yang diangkat dalam penelitian.

Bab IV Eksperimen dan Evaluasi menjelaskan skema dan konfigurasi pemodelan, jenis data yang digunakan serta hasil eksperimen pemodelan pada penelitian ini. Evaluasi berdasarkan hasil eksperimen yang dilakukan juga tercantum di dalam bab ini.

Bab V Penutup berisi simpulan yang mengandung ulasan ringkas ketercapaian tujuan penelitian berdasarkan eksperimen dan evaluasi yang dilakukan, dan saran pengembangan lebih lanjut dari penelitian ini.