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

Maret 2019

Bab III Analisis Masalah dan Perancangan Solusi

Bab ini memaparkan skema penelitian yang dilakukan untuk menjawab persoalan yang dibahas pada Bab I.

III.1 Analisis Permasalahan

Ada beberapa permasalahan yang ditemukan pada permasalahan pengenalan ucapan, dan salah satunya dan yang menjadi fokus pada penelitian ini adalah kesulitan sistem pengenal ucapan dalam mengenali suara pada lingkungan yang bising. Beberapa penelitian-penelitian terkait mengenai permasalahan tersebut mencoba untuk menyelesaikannya dengan menambahkan informasi visual pada proses pengenalan suaranya, seperti pada penelitian Chung dkk. (2017), Chung and Zisserman (2016), dan Assael dkk. (2016).

Untuk penelitian-penelitian mengenai pengenalan ucapan untuk bahasa Indonesia sejauh ini sudah banyak yang menggunakan pendekatan *deep learning*, seperti pada penelitian Yuwan (2018) yang membangun model akustik ucapan spontan bahasa Indonesia berbasis DNN-HMM, akan tetapi belum ada penelitian yang menggabungkan fitur akustik dan fitur visual, baik yang menggunakan fitur *handcrafted* dan model akustik berbasis statistik maupun *deep learning*.

Fitur visual yang digunakan dalam mengenali ucapan adalah berupa informasi gerak bibir, seperti yang biasa dilakukan oleh manusia. Penelitian-penelitian terkait pembacaan gerak bibir untuk bahasa Indonesia masih terbilang sedikit jika dibandingkan dengan penelitian-penelitian terkait untuk bahasa lain, terutama bahasa Inggris. Penelitian mengenai pembacaan gerak bibir yang menggunakan dataset bahasa Indonesia kebanyakan masih belum menggunakan *deep learning*, seperti pada penelitian Achmad and Fadillah (2015) yang menggunakan HMM menunjukkan bahwa hasil pengenalan 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.

Selain penelitian tersebut, pada saat penulisan, hanya ada satu penelitian mengenai pembacaan gerak bibir dalam bahasa Indonesia yang menggunakan pendekatan *deep learning*, yaitu oleh Maulana and Fanany (2017), yang menggunakan *spatio-temporal* CNN untuk menangkap struktur *spatiotemporal* dari video, dan menggunakan *bidirectional* Gated Recurrent Unit (GRU) untuk memodelkan keseluruhan rangkaian frame video dari dua arah, baik dengan urutan frame normal maupun dengan urutan terbalik. Kinerjanya sudah sangat baik, dengan WER 13.3% dan BLEU 90.4% untuk model yang dilatih menggunakan pembicara-pembicara yang tidak ada di data uji, dan WER 8.0% dan BLEU 94.7% untuk model yang dilatih menggunakan pembicara-pembicara yang ada di data uji. Akan tetapi data yang digunakan adalah data AVID, yang merupakan versi bahasa Indonesia dari dataset GRID. Dataset GRID itu sendiri terdiri atas kalimat-kalimat berstruktur *command* + *color* + *preposition* + *letter* + *digit* + *adverb* dan kosakata yang terbatas, dibandingkan dengan dataset LRW yang memiliki kosakata terbuka. Jika dibandingkan dengan penelitian Chung and Zisserman (2016) untuk pengenalan dalam bahasa Inggris, model *sequence-to-sequence*nya berhasil mencapai kinerja WER 3.0% untuk dataset GRID, tetapi untuk dataset LRW hanya mencapai WER 23.8%, sehingga masih ada ruang untuk perbaikan, baik untuk bahasa Indonesia maupun bahasa Inggris.

Salah satu permasalahan yang ditemui dalam penggunaan model *sequence-to-sequence* untuk mentransduksi sebuah rangkaian menjadi rangkaian lain adalah kebanyakan model ini menggunakan RNN atau variannya sehingga prosesnya tidak bisa diparalelisasi karena sifatnya yang rekurens. Oleh sebab itu, proses pelatihan model membutuhkan waktu yang lama hingga model akhirnya konvergen. Pada penelitian Vaswani dkk. (2017) diusulkan model yang disebut sebagai model transformer, sebuah arsitektur model yang menghindari penggunaan rekurens dan bergantung sepenuhnya 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,

dan juga berhasil mengungguli model *encoder-decoder* berbasis RNN dalam mentransduksi rangkaian.

III.2 Analisis Solusi

Berdasarkan ulasan masalah yang dipaparkan pada subbab III.1, penelitian ini akan membangun solusi pemodelan akustik dan visual menggunakan model transformer. Agar diketahui peningkatan kinerja AVSR yang menggunakan model transformer, diimplementasikan juga model *sequence-to-sequence* dari Chung dkk. (2017) sebagai model tolok ukur. Selain itu dibangun juga solusi pemodelan dengan berbasis akustik saja, dan juga pemodelan dengan berbasis visual saja, untuk melihat peningkatan kinerja ASR setelah penambahan fitur visual.

III.2.1 Persiapan Korpus Video

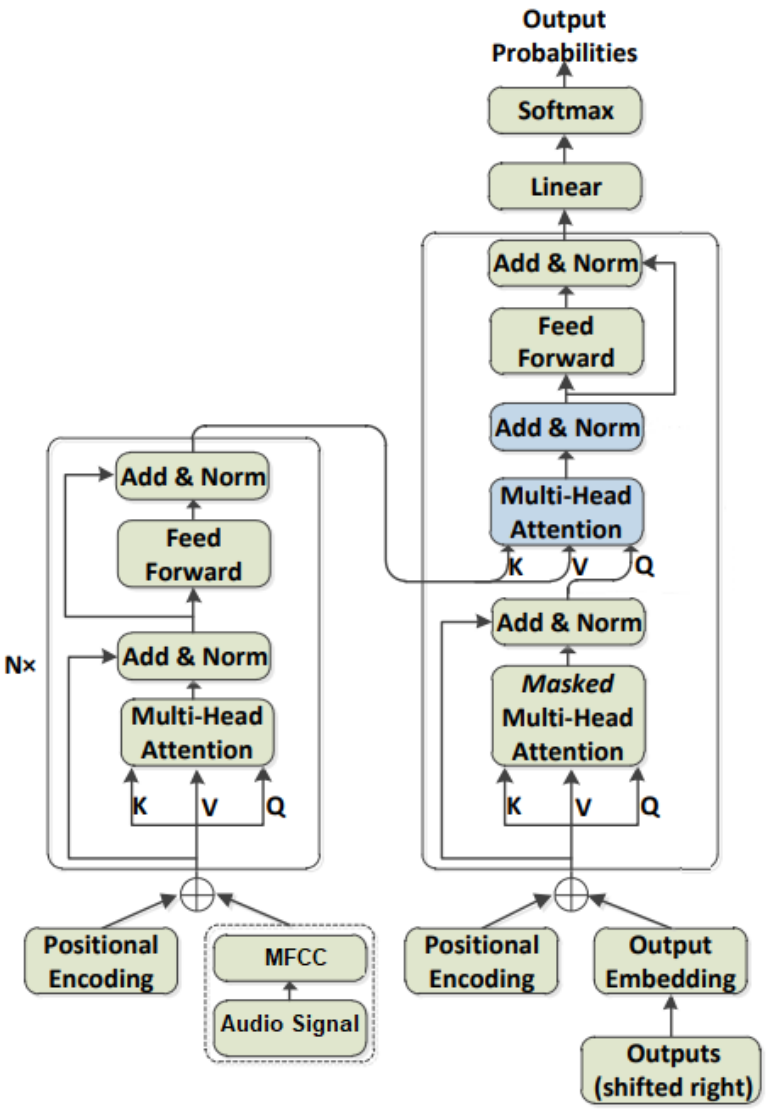
Korpus video diperoleh dari data yang digunakan pada penelitian sebelumnya Chung dkk. (2017). Korpus video tersebut terdiri atas tiga jenis korpus, yaitu LRW, LRS2, dan LRS3. Untuk korpus LRW dan LRS2 bersumber dari televisi BBC UK sehingga untuk perolehan korpus tersebut sedikit dipersulit sehingga hanya digunakan korpus LRS3 saja yang bersumber dari TED. Ketiga korpus tersebut merupakan korpus dalam bahasa Inggris dan digunakan untuk pelatihan model awal. Selanjutnya dilakukan *transfer learning* dan model selanjutnya dilatih menggunakan korpus bahasa Indonesia. korpus bahasa Indonesia ini dikumpulkan dari situs YouTube berisi pembicara-pembicara di TEDxJakarta dan BukaTalks, dengan total durasi video adalah 10 jam. Untuk korpus LRS3 berdurasi sekitar 800 jam dan memuat 150 ribu kalimat atau sekitar 50 ribu kosakata.

III.2.2 Pengenalan Wajah

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet

aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

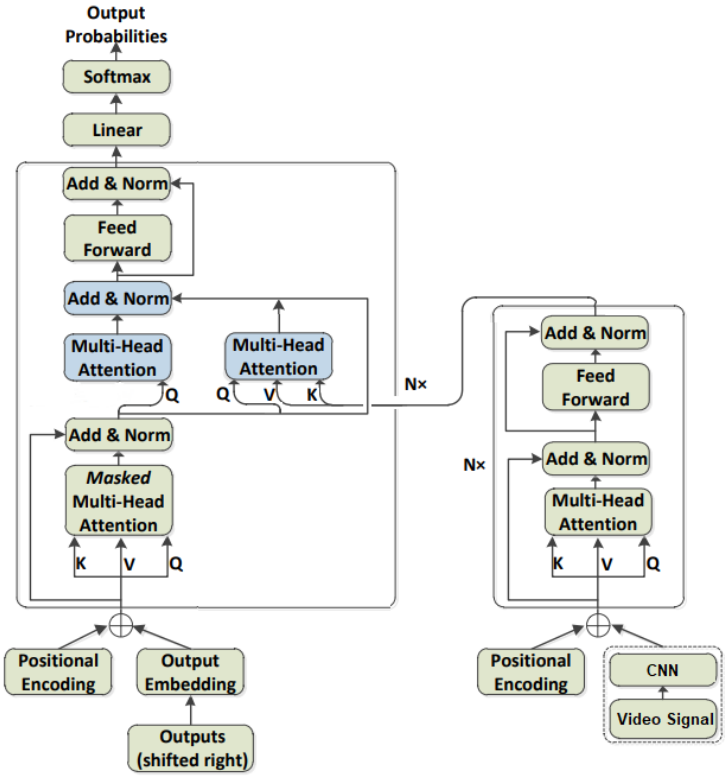
III.2.3 Rancangan Arsitektur



Gambar III.1: Rancangan Arsitektur ASR.

Pada bagian ini dijelaskan rancangan arsitektur untuk proses pengenalan ucapan. Rancangan model arsitektur yang diusulkan dibagi menjadi tiga model, yaitu:

1. model yang menggunakan modal akustik saja. Model ini menggunakan arsitektur transformer (Gambar III.1) dan dilakukan pelatihan menggunakan dataset yang sudah dijelaskan pada upabab III.2.1. Masukan dari dataset tersebut berupa *raw speech* yang kemudian diekstraksi fiturnya dengan menggu-



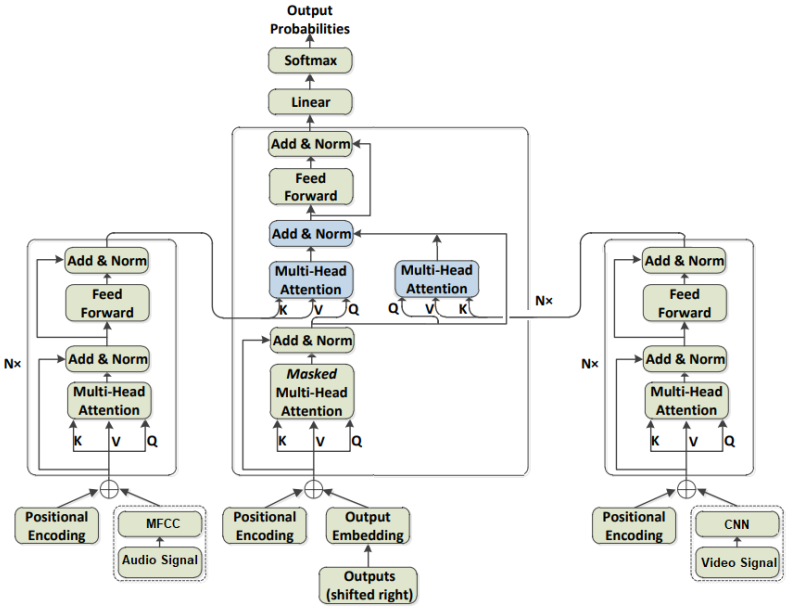
Gambar III.2: Rancangan Arsitektur VSR.

nakan MFCC. Selanjutnya model ini akan disebut sebagai model transASR.

2. model yang menggunakan modal visual saja. Model ini menggunakan arsitektur transformer (Gambar III.2) dan dilakukan pelatihan menggunakan dataset yang sudah dijelaskan pada upabab III.2.1. Masukan berupa frame video yang direpresentasikan menjadi vektor berukuran tetap menggunakan arsitektur CNN yang belum ditentukan. Untuk selanjutnya model ini akan disebut sebagai model transVSR.

3. terakhir adalah model yang menggunakan modal akustik dan modal visual secara bersamaan. Model ini juga menggunakan arsitektur transformer (Gambar III.3) dan dilakukan pelatihan menggunakan dataset yang sudah dijelaskan pada upabab III.2.1. Masukan berupa *raw speech* dan frame video. Selanjutnya model ini akan disebut sebagai model transAVSR.

Selain frame video dan *raw speech*, terdapat juga masukan berupa keluaran-keluaran sebelumnya dari model, sehingga keluaran dari time-step saat ini akan



Gambar III.3: Rancangan Arsitektur AVSR.

digunakan juga untuk masukan di time-step selanjutnya.

Komponen-komponen pembentuk dari model transformer itu sendiri adalah komponen *positional encoding*, *multi-head attention*, dan *masked multi-head attention*, yang sudah dijelaskan sebelumnya pada upabab **??**.

III.2.4 Evaluasi Sistem

Kinerja dari sistem diukur dengan menggunakan tiga metrik pengujian yang umum digunakan pada kasus penganalan ucapan dan translasi, yaitu *character error rate* (CER), *word error rate* (WER), dan *bilingual evaluation understudy* (BLEU). WER dan CER dihitung dengan menggunakan rumus:

$$\begin{aligned} WER &= 100 \times \frac{S+D+I}{N} \\ &= 100 \times \frac{S+D+I}{S+D+C} \end{aligned}$$

yang dalam hal ini, S merupakan jumlah substitusi kata (untuk WER) atau karakter (untuk CER) yang dikenali, D merupakan jumlah penghapusan kata atau karakter, I merupakan jumlah penyisipan kata atau karakter, C merupakan jumlah kata atau

karakter yang benar, dan N adalah jumlah keseluruhan kata atau karakter yang ada pada transkripsi. Semakin kecil nilai WER dan CER maka semakin semakin baik kinerja model dalam mengenali ucapan.

BLEU merupakan presisi n-gram yang telah dimodifikasi dan digunakan untuk membandingkan kalimat yang dihasilkan sistem dengan kalimat yang ada di transkripsi. BLEU dihitung dengan menggunakan persamaan berikut.

$$P = \frac{m}{w_t}$$

$$p = \begin{cases} 1 & jika \quad c > r \\ e^{(1-\frac{r}{c})} & jika \quad c \leq r \end{cases}$$

$$BLEU = p \times e^{\Sigma_{n=1}^{N}}$$

yang dalam hal ini P merupakan presisi unigram, p merupakan *brevity penalty* atau penalti jika kalimat yang dihasilkan lebih pendek dari kalimat yang ada di transkripsi, r adalah panjang efektif dari kalimat di transkripsi, dan c merupakan total panjang keseluruhan kalimat yang dihasilkan sistem.

Pengujian dilakukan dengan membandingkan kegita model yang telah dibuat (transASR, transVSR, dan transAVSR).