# 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**

**Mei 2019**

# ABSTRAK

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

Oleh
**Gugy Lucky Khamdani**
**NIM: 23517041**
**(Program Studi Magister Informatika)**

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.

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.

**Kata kunci**: katakunci1, katakunci2, katakunci3, katakunci4

# *ABSTRACT*

## *TRANSFORMER MODEL FOR AUDIOVISUAL SPEECH RECOGNITION IN BAHASA INDONESIA*

*By*
**Gugy Lucky Khamdani**
**NIM: 23517041**
***(Master's Program in Informatics/Computer Science)***

*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.*

*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.*

***Keywords**: keyword1, keyword2, keyword3, keyword4*

# JUDUL TESIS: PENGGUNAAN MODEL TRANSFORMER PADA *AUDIOVISUAL SPEECH RECOGNITION* UNTUK BAHASA INDONESIA

Oleh
**Gugy Lucky Khamdani**
**NIM: 23517041**
**(Program Studi Magister Informatika)**
Institut Teknologi Bandung

Menyetujui

Pembimbing

tanggal 21 Maret 2019.

| Pembimbing I, | Pembimbing II |
| --- | --- |
| Dr. Dessi Puji Lestari, ST., M.Eng. | Nugraha Priya Utama, S.T., M.A., Ph.D. |
| NIP. 197912012012122005 | NIP. 118110074 |

# PEDOMAN PENGGUNAAN TESIS

Gunakan bagian ini untuk memberikan ucapan terima kasih kepada semua pihak yang secara langsung atau tidak langsung membantu penyelesaian tugas akhir, termasuk pemberi beasiswa jika ada. Utamakan untuk memberikan ucapan terima kasih kepada tim pembimbing tugas akhir dan staf pengajar atau pihak program studi, bahkan sebelum mengucapkan terima kasih kepada keluarga. Ucapan terima kasih sebaiknya bukan hanya menyebutkan nama orang saja, tetapi juga memberikan penjelasan bagaimana bentuk bantuan/dukungan yang diberikan. Gunakan bahasa yang baik dan sopan serta memberikan kesan yang enak untuk dibaca. Sebagai contoh: "Tidak lupa saya ucapkan terima kasih kepada teman dekat saya, Tito, yang sejak satu tahun terakhir ini selalu memberikan semangat dan mengingatkan saya apabila lengah dalam mengerjakan Tugas Akhir ini. Tito juga banyak membantu mengoreksi format dan layout tulisan. Apresiasi saya sampaikan kepada pemberi beasiswa, Yayasan Beasiswa, yang telah memberikan bantuan dana kuliah dan biaya hidup selama dua tahun. Bantuan dana tersebut sangat membantu saya untuk dapat lebih fokus dalam menyelesaikan pendidikan saya. ....". Ucapan permintaan maaf karena kekurangsempurnaan hasil Tugas Akhir tidak perlu ditulis.

# KATA PENGANTAR

Gunakan bagian ini untuk memberikan ucapan terima kasih kepada semua pihak yang secara langsung atau tidak langsung membantu penyelesaian tugas akhir, termasuk pemberi beasiswa jika ada. Utamakan untuk memberikan ucapan terima kasih kepada tim pembimbing tugas akhir dan staf pengajar atau pihak program studi, bahkan sebelum mengucapkan terima kasih kepada keluarga. Ucapan terima kasih sebaiknya bukan hanya menyebutkan nama orang saja, tetapi juga memberikan penjelasan bagaimana bentuk bantuan/dukungan yang diberikan. Gunakan bahasa yang baik dan sopan serta memberikan kesan yang enak untuk dibaca. Sebagai contoh: "Tidak lupa saya ucapkan terima kasih kepada teman dekat saya, Tito, yang sejak satu tahun terakhir ini selalu memberikan semangat dan mengingatkan saya apabila lengah dalam mengerjakan Tugas Akhir ini. Tito juga banyak membantu mengoreksi format dan layout tulisan. Apresiasi saya sampaikan kepada pemberi beasiswa, Yayasan Beasiswa, yang telah memberikan bantuan dana kuliah dan biaya hidup selama dua tahun. Bantuan dana tersebut sangat membantu saya untuk dapat lebih fokus dalam menyelesaikan pendidikan saya. ....". Ucapan permintaan maaf karena kekurangsempurnaan hasil Tugas Akhir tidak perlu ditulis.

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.

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.

Bandung, 13 Maret 2019

Penulis

# DAFTAR ISI

# DAFTAR GAMBAR

# DAFTAR TABEL

# 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 penelitian Arifin dkk. (2013), berfokus pada pembangunan *viseme* dalam bahasa Indonesia dengan cara melakukan *clustering* menggunakan K-Means pada data berisi gambar *speech* visual. Hasil penelitian tersebut menunjukkan bahwa dalam bahasa Indonesia terdapat 10 kelas viseme.

Hingga saat ini, penelitian 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, penelitian-penelitian 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 penelitian Zhou dkk. (2014). Untuk yang berbahasa Indonesia terdapat penelitian 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 penelitian 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 penelitian, 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 dilengkapi 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

Penelitian mengenai pengenalan ucapan otomatis pada bahasa Indonesia sudah banyak dilakukan, akan tetapi kebanyakan masih memerlukan praproses untuk mereduksi *noise*. Penelitian 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 penelitian 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.

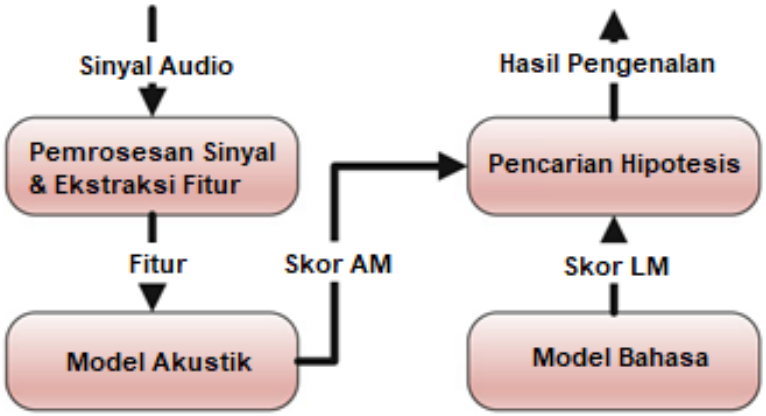
**Bab V Penutup** berisi simpulan yang mengandung ulasan ringkas ketercapaian tujuan penelitian berdasarkan eksperimen dan evaluasi yang dilakukan, dan saran pengembangan lebih lanjut dari penelitian ini.

# Bab II Tinjauan Pustaka

Pada bab ini dijelaskan mengenai deskripsi umum sistem pengenalan ucapan, sistem pengenalan gerak bibir, dan model *sequence-to-sequence*.

## II.1 Sistem Pengenalan Ucapan Otomatis

Sistem pengenalan ucapan otomatis atau disebut juga *automatic speech recognition* (ASR) adalah proses pengenalan atau penerjemahan bahasa lisan dalam bentuk sinyal audio menjadi teks secara otomatis oleh komputer. Pada umumnya, arsitektur dari sebuah system ASR bisa dibagi menjadi empat buah komponen utama, yaitu pemrosesan sinyal dan ekstraksi fitur, model akustik atau acoustic model (AM), model bahasa atau language model (LM), dan pencarian hipotesis (Yu dan Deng, 2014).




Gambar II.1: Arsitektur umum dari sistem ASR (Yu dan Deng, 2014)

### II.1.1 Pemrosesan Sinyal dan Ekstraksi Fitur

Komponen pemrosesan sinyal dan ekstraksi fitur mengambil masukan berupa sinyal audio mentah dan diproses agar *noise*-nya dihilangkan, mengubah sinyal dari domain waktu ke domain frekuensi, dan mengekstraksi fitur-fitur yang paling penting dari sinyal tersebut sehingga cocok dengan komponen model akustik.

### II.1.2 Model Akustik

Komponen model akustik tersebut mengintegrasikan pengetahuan mengenai akustik dan fonetik, dan dengan menggunakan fitur yang diekstraksi di komponen sebe-

lumnya lalu menghasilkan skor AM untuk fitur tersebut.

Penelitian Han dkk. (2018) menunjukkan bahwa model akustik dengan *word error rate* (WER) terbaik untuk korpus bahasa Inggris yaitu sebesar 5.0% pada dataset Switchboard dan 9.1% pada dataset CallHome, didapatkan dengan menggunakan dense TDNN-LSTM. Fitur yang digunakan adalah MFCC dengan dimensi sebesar 39, sama seperti yang dilakukan oleh Xiong dkk. (2017) akan tetapi model akustik yang digunakan adalah CNN-BLSTM.

Selain itu, terdapat penelitian lain yang menggunakan pendekatan *sequence-to-sequence*, seperti Chan dkk. (2015) yang menggunakan *Bidirectional* LSTM (BLSTM) dengan struktur piramid untuk mendapatkan representasi tingkat tinggi dari sinyal audio. Struktur piramid digunakan untuk mengurangi *time-step* yang dibutuhkan untuk mendapatkan representasi dari keseluruhan sinyal audio, yang bisa mencapai ribuan pada sistem ASR. BLSTM struktur piramid tersebut pada pendekatan *sequence-to-sequence* disebut sebagai *encoder*, menghasilkan context vectors yang kemudian digunakan pada *decoder* untuk menghasilkan rangkaian kata-kata sebagai hasil pengenalan. Model ini berhasil mencapai WER 14.1% tanpa menggunakan kamus dan model bahasa, dan berhasil mencapai WER 10.3% jika menggunakan model bahasa. Penelitian lain yang menggunakan pendekatan *sequence-to-sequence* pada data bahasa Inggris adalah penelitian Chung dkk. (2017), yang sama seperti penelitian sebelumnya hanya saja pada *decoder*-nya tidak menerima masukan berupa sinyal audio mentah, tetapi sudah diekstraksi fitur terlebih dahulu menjadi fitur MFCC berdimensi 13 dengan urutan waktu terbalik. Model ini berhasil mencapai WER 17.7% pada dataset LRS.

Penelitian mengenai ASR pada bahasa Indonesia salah satunya adalah oleh Yuwan (2018), yang menggunakan model DNN-HMM dan dibandingkan dengan model tolak ukur GMM-HMM. Pengujian model dilakukan pada skema tertutup dan terbuka, dan berhasil mencapai penurunan WER dari ASR berbasis GMM-HMM ke ASR berbasis DNN-HMM sebesar 2,53% pada skema tertutup dan 3,89% pada skema terbuka.

### II.1.3 Model Bahasa

Komponen model bahasa mengestimasi probabilitas dari rangkaian kata-kata hipotesis dengan cara mempelajari korelasi antar kata-kata tersebut berdasarkan korpus data latih. Estimasi probabilitas tersebut disebut sebagai skor LM.

### II.1.4 Pencarian Hipotesis

Dengan menggunakan skor AM dan LM yang didapatkan dari vektor fitur dan kata-kata kandidat yang mungkin, komponen pencarian hipotesis akan menghasilkan keluaran berupa rangkaian kata-kata dengan skor AM dan LM tertinggi sebagai hasil pengenalan.

## II.2 Pemodelan *Sequence*

Pemodelan *sequence* dengan *deep learning* berkembang dengan pesat, terutama menggunakan *recurrent neural network* (RNN), yang telah menunjukkan hasil yang baik pada banyak permasalahan pembelajaran mesin, khususnya untuk permasalahan yang memiliki masukan dan/atau keluaran yang memiliki panjang yang bervariasi. Penelitian Sutskever dkk. (2014) dan Bahdanau dkk. (2015) menunjukkan bahwa RNN dapat melakukan pemodelan *sequence* dengan baik sama seperti sistem-sistem yang sudah ada untuk permasalahan sulit seperti mesin translasi.

RNN merupakan perluasan dari jaringan saraf tiruan *feedforward* konvensional yang dapat menangani masukan dengan panjang bervariasi. RNN menanganinya dengan menggunakan *recurrent hidden* state yang pengaktifannya bergantung pada *hidden state* pada waktu yang sebelumnya. Secara formal, RNN memperbarui *recurrent hidden state* ht-nya sebagai berikut (Chung dkk., 2014):

$$h_t = \begin{cases} 0 & jika \quad t = 0 \\ \varphi(h_{t-1} - x_t) & jika \quad otherwise \end{cases}$$

yang pada hal ini $x$ adalah rangkaian $x = (x_1, x_2 \ldots, x_T)$ dan $\varphi$ merupakan fungsi non-linier seperti fungsi *logistic sigmoid*. RNN juga bisa memiliki keluaran $y = (y_1, y_2 \ldots, y_T)$ yang juga memiliki panjang yang bervariasi.

Bobot atau parameter yang terdapat pada RNN secara tradisional diperbarui sebagai

berikut (Chung dkk., 2014):

$$h_t = g(Wx_t + Uh_{t-1})$$

yang dalam hal ini $g$ merupakan fungsi seperti *logistic sigmoid* atau *hyperbolic tangent*.

Akan tetapi, pelatihan RNN sulit dilakukan untuk menangkap dependensi yang jaraknya jauh karena nilai gradien pada RNN cenderung menghilang atau meningkat drastis. Hal ini membuat optimasi berbasis gradien menjadi sulit untuk dilakukan. Pendekatan yang bisa dilakukan untuk menangani masalah ini bisa dengan melakukan optimasi yang lain yang tidak berbasis gradien, seperti menggunakan *stocahstic gradient desecent*. Pendekatan lainnya yang bisa dilakukan adalah menggunakan arsitektur RNN yang lain yang menggunakan fungsi aktivasi yang lebih mutakhir, seperti *long short-term memory* (LSTM) (Hochreiter dan Urgen Schmidhuber, 1997) dan *gated recurrent unit* (GRU) (Cho dkk., 2014b).

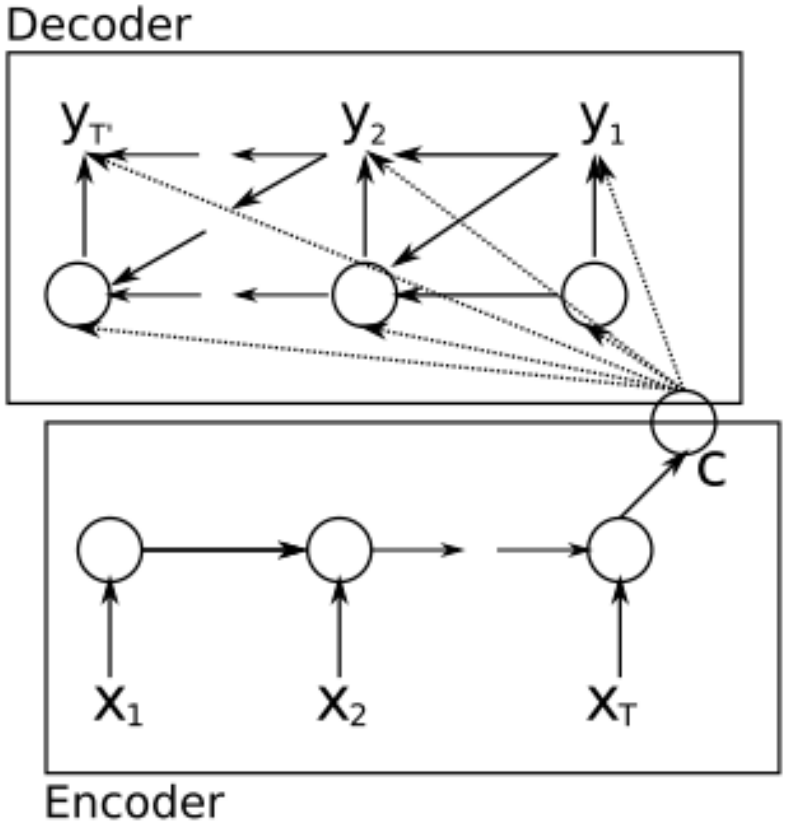
Pemodelan *sequence* yang paling banyak digunakan pada penelitian-penelitian terakhir ini adalah menggunakan model *sequence-to-sequence* yang memetakan sebuah rangkaian masukan dengan panjang $n$ menjadi rangkaian keluaran dengan panjang $m$, dipopulerkan dalam penggunaanya untuk permasalahan terjemahan mesin.

Terdapat dua pendekatan umum untuk model *sequence-to-sequence*, yaitu model *encoder-decoder* berbasis rekurens dan model *encoder-decoder* berbasis *attention*.

### II.2.1 Model *Encoder-Decoder* berbasis rekurens

Model dasar *sequence-to-sequence* diperkenalkan oleh Cho dkk. (2014a), yang terdiri dari dua buah *recurrent neural network* (RNN), yaitu *encoder* yang bertugas untuk merepresentasikan rangkaian masukan menjadi sebuah *fixed-length vector*, dan *decoder* yang bertugas untuk menghasilkan rangkaian keluaran berdasarkan vektor yang didapatkan dari encoder tadi.

Selain itu ada penelitian dari Sutskever dkk. (2014) yang sama menggunakan arsitektur *encoder-decoder* akan tetapi berbeda dengan arsitektur sebelumnya yang

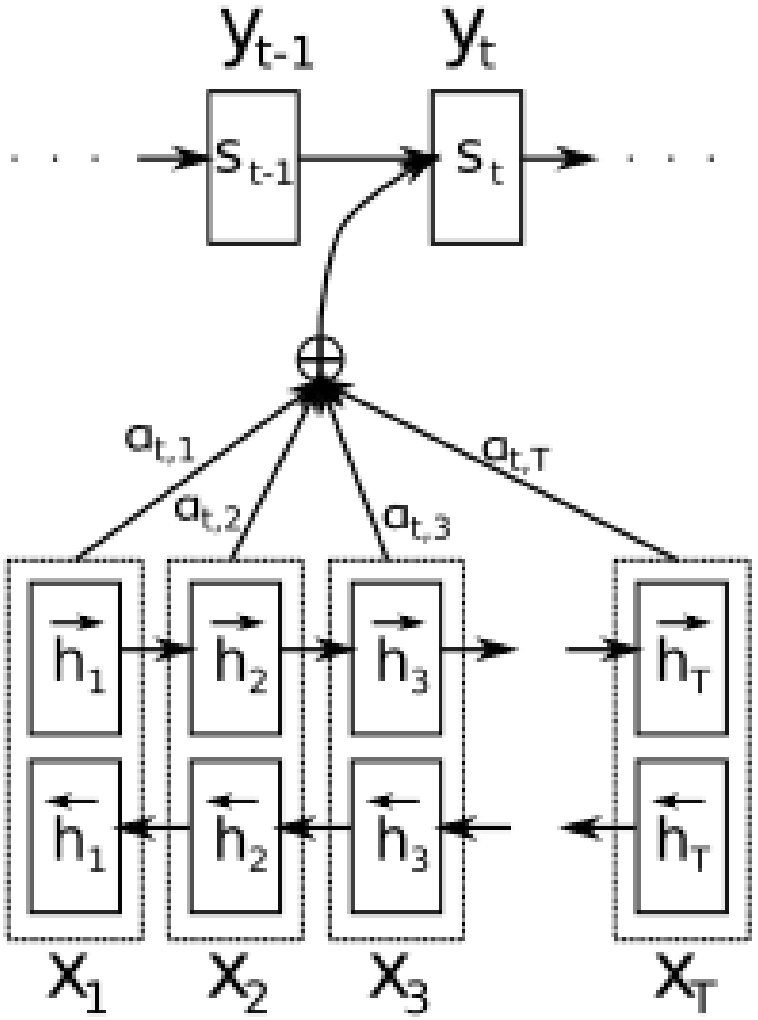



Gambar II.2: Ilustrasi dari arsitektur *encoder-decoder* (Cho dkk., 2014a)

hanya mengintegrasikan jaringan saraf tiruannya ke dalam sistem *statistical machine translation* (SMT), model ini merupakan model yang sepenuhnya menggunakan RNN dan proses pelatihannya dilakukan secara *end-to-end*. Hanya saja model *encoder-decoder* ini harus dapat memampatkan informasi dari satu kalimat utuh menjadi sebuah *fixed-length vector*. Hal ini bisa jadi menyulitkan RNN untuk merepresentasikan kalimat yang panjang, terutama jika panjangnya lebih panjang dari yang terdapat pada data latih di corpus. Cho dkk. (2014b) menunjukkan bahwa memang performa dari model *encoder-decoder*-nya semakin memburuk seiring dengan semakin meningkatnya panjang dari kalimat masukan.

Untuk menangani masalah tersebut, Bahdanau dkk. (2015) mengaplikasikan mekanisme *attention* pada *decoder*. Dengan mekanisme *attention* ini *decoder* dapat menentukan bagian mana dari kalimat masukan yang harus mendapatkan perhatian lebih untuk selanjutnya menjadi masukan pada fungsi aktivasi pada *decoder* untuk

menentukan keluaran apa yang akan dihasilkan. Dengan membiarkan *decoder* menentukan masukan mana yang penting untuk menghasilkan keluaran selanjutnya, *decoder* menjadi tidak perlu untuk merepresentasikan semua informasi yang ada dalam sebuah kalimat ke dalam sebuah vektor. Dengan ini informasi pada sebuah kalimat bisa direpresentasikan secara tersebar ke seluruh rangkaian kata-kata pada kalimat, yang selanjutnya bisa dipilih oleh *decoder* dengan mekanisme *attention*.




Gambar II.3: Ilurstrasi mekanisme attention (Bahdanau dkk., 2015)

Seperti yang bisa dilihat pada gambar II.3, tingkat kepentingan dari masukan dipilih oleh mekanisme *attention* berdasarkan nilai dari bobot $\alpha_{i,j}$ dari setiap masukan $h_j$, yang dihitung dengan fungsi *softmax* (Bahdanau dkk., 2015):

$$\alpha_{ij} = \frac{exp(e_{ij})}{\sum_{k=1}^{Tx} exp(e_{ik})}$$

yang dalam hal ini,

$$e_{ij} = a(s_{i-1}, h_j)$$

yang merupakan model *alignment* yang menilai kecocokan antara masukan pada sekitaran posisi $j$ dan keluaran pada posisi $i$. Parameter pada model *alignment* dilatih sebagai *feedforward neural network* yang kemudian dilatih secara bersama-sama dengan komponen-komponen lain pada sistem yang diusulkan.

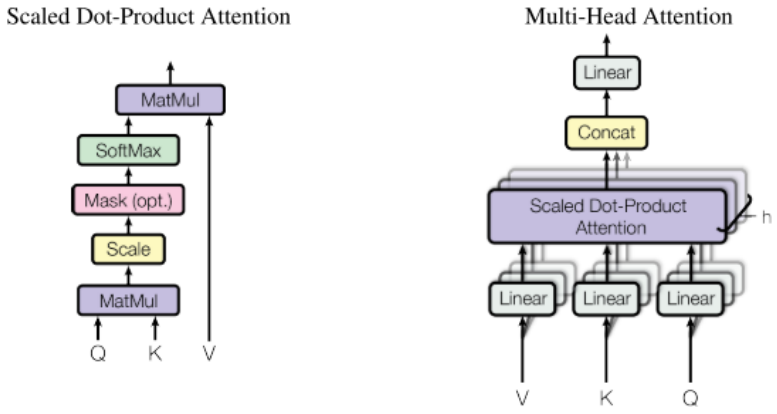
### II.2.2 Model *Encoder-Decoder* berbasis *attention*

Mekanisme *attention* sudah diaplikasikan pada berbagai permasalahan selain permasalahan mesin translasi dan menjadi bagian penting dalam pemodelan rangkaian dan model transduksi, sehingga memungkinkan model untuk memodelkan dependensi tanpa harus melihat jarak antara masukan dan keluaran. 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 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 rekurens.

Arsitektur dari model transformer secara keseluruhan mengikuti model *encoder-decoder*, hanya saja komponen penyusunnya tidak menggunakan RNN tapi menggunakan *stacked self-attention*, dan *point-wise fully connected layer*. Selain itu untuk model transfromer juga tidak menggunakan RNN untuk meng-*encode* rangkaian, tetapi menggunakan layer *positional encodings* yang kemudian diikuti oleh

fungsi *attention*.

Fungsi *attention* bisa dideskripsikan sebagai pemetaan *query* dan sekumpulan pasangan *key-value* menjadi sebuah keluaran, yang pada hal ini *query*, pasangan *key-value*, dan keluaran semuanya berbentuk vektor. Keluaran tersebut dihitung sebagai jumlah tertimbang. Jenis fungsi *attention* yang digunakan pada penelitian ini disebut sebagai *Scaled Dot-Product Attention*, yang kemudian fungsi *attention* tersebut bisa dihitung secara paralel, yang kemudian disebut sebagai *Multi-Head Attention*.




Gambar II.4: Ilustrasi *scaled dot-product attention* dan *multi-head attention*. (Vaswani dkk., 2017)

*Scaled dot-product attention*, mengambil input berupa beberapa *query* (Q) dan *key* (K) yang berdimensi $d_k$, dan *value* (V) yang berdimensi $d_v$, dan dapat diformulasikan sebagai berikut:

$$Attention(Q,K,V) = \frac{QK^T}{\sqrt{d_k}}V$$

Pada dasarnya *scaled dot-product attention* sama seperti *dot-product attention*, perbedaannya terletak pada penambahan faktor penskala $\frac{1}{d_k}$ pada perhitungannya. faktor penskala tersebut digunakan untuk menangani masalah ketika nilai $d_k$ terlalu besar sehingga hasil *dot-product*nya tumbuh secara besar, sehingga hasil dari fungsi softmaxnya memiliki gradien yang sangat kecil dan selanjutnya menimbulkan masalah ketika melakukan propagasi balik.

*Multi-head attention* merupakan pengembangan dari *scaled dot-product attention*. Menurut Vaswani dkk. (2017), daripada menghitung satu kali *attention* dari *query*, *key*, dan *value* yang sebanyak $d_{model}$, akan lebih baik jika *query*, *key*, dan *value*nya diproyeksikan secara linier terlebih dahulu sebanyak $h$ kali, dengan proyeksi liner ke dimensi $d_q$, $d_k$, $d_v$ dan yang berbeda-beda dan dipelajari dari data. Masing-masing *query*, *key*, dan *value* yang sudah diproyeksikan tersebut kemudian dihitung *attention*nya secara paralel, dan menghasilkan output berdimensi $d_v$ yang kemudian dikonkatenasi dan diproyeksikan lagi, yang merupakan hasil akhir dari perhitungan *multi-head attention*. *Multi-head attention* dapat diformulasikan sebagai berikut:

$$MultiHead(Q,K,V) = Concat(head_1,\ldots,head_h)W^O$$
$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$$

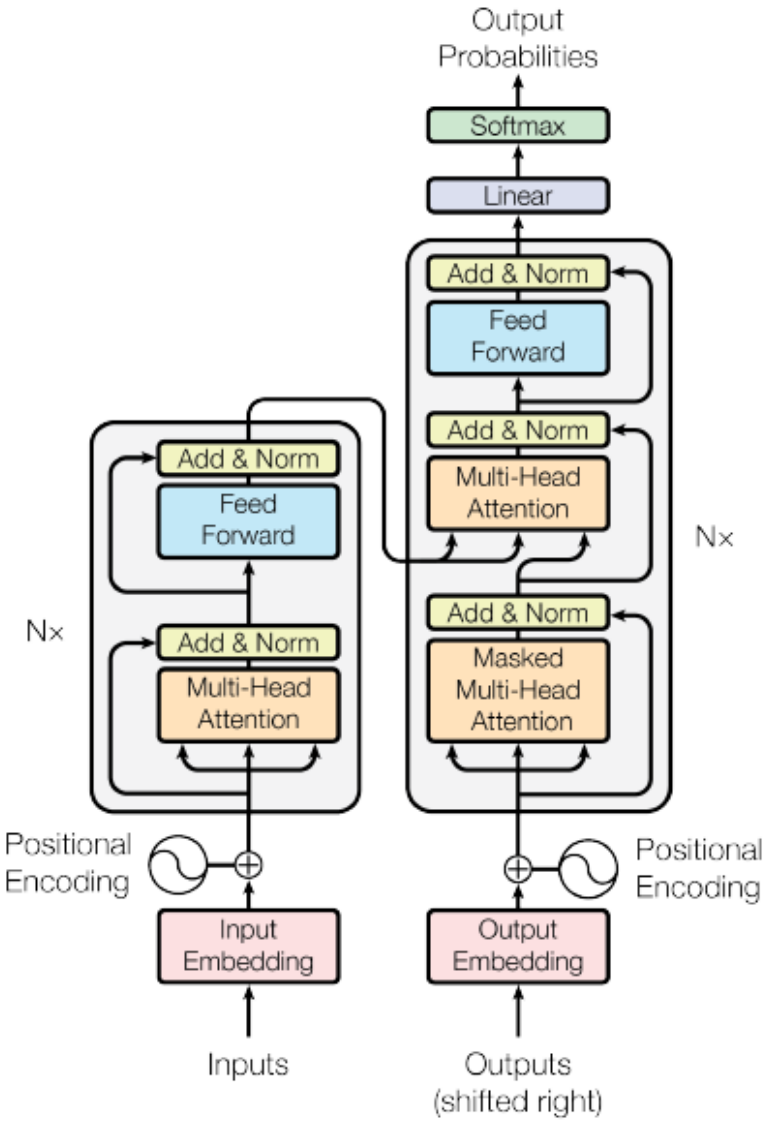
yang dalam hal ini $QW_i^Q$, $KW_i^K$, dan $VW_i^V$ merupakan matrix parameter untuk proyeksi, yang dipelajari dari data.

Jenis *attention* ini memungkinkan model untuk mempelajari informasi mana yang paling perlu diperhatikan dari berbagai representasi pada ruang pencarian dengan posisi yang berbeda-beda.

Selain menggunakan fungsi *attention*, setiap layer pada *encoder* dan *decoder* berisi *fully connected feedforward network*, yang terdiri dari dua transformasi linier dengan fungsi aktivasi ReLU diantaranya. Di layer masukan terdapat layer *embedding* dan di layer keluaran digunakan layer softmax. Karena modelnya tidak mempunyai komponen rekurens atau konvolusi, supaya modelnya mampu mempelajari urutan dari rangkaiannya, maka modelnya harus dimasukkan informasi tambahan mengenai posisi relatif dan posisi absolut dari token di dalam sebuah kalimat. Untuk itu digunakan *positional embedding* setelah melalui *embedding* di layer masukan tadi.

### II.3 Pemodelan Gambar dan Video

Penelitian yang paling banyak mengenai pemodelan gambar dan video adalah mengenai pembangkitan keterangan otomatis (*automatic caption generation*). Pem-




Gambar II.5: Arsitektur model Transformer (Vaswani dkk., 2017)

bangkitan keterangan otomatis merupakan permasalahan mendasar pada kecerdasan buatan yang menggabungkan *computer vision* dan pemrosesan bahasa alami. Penelitian mengenai pembangkitan keterangan otomatis akhir-akhir ini bisa dikategorisasikan menjadi dua, yaitu pembangkitan keterangan otomatis dari citra dan pembangkitan keterangan otomatis dari video. Jika proses pembangkitan dilakukan dari citra maka informasi penting yang harus diperhatikan hanya informasi posisi saja sedangkan jika dilakukan dari video maka selain itu harus juga diperhatikan informasi temporal antar frame videonya. Permasalahan utamanya adalah sebuah deskripsi dihasilkan harus bisa menangkap semua objek yang terdapat di dalam ci-

tra tersebut, dan juga harus bisa mengekspresikan keterkaitan antara objek dan juga atribut-atribut apa saja yang menjelaskan objek tersebut dan aktivitas apa yang melibatkan objek tersebut. Terlebih lagi, deskripsi tersebut harus dideskripsikan dengan menggunakan bahasa yang sealami mungkin.

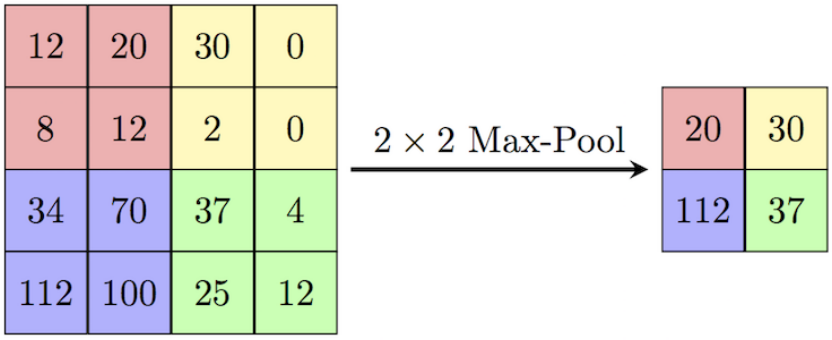
### II.3.1 Pembangkitan Keterangan Otomatis dari Citra

Penelitian Vinyals dkk. (2014) mengemukakan sebuah model yang diinspirasi oleh kemajuan-kemajuan terbaru pada permasalahan mesin translasi yang menggunakan model *encoder-decoder* berbasis RNN. Perbedaannya pada penelitian ini adalah sebagai ganti dari RNN, modelnya menggunakan *convolutional neural network* (CNN) sebagai encoder. Dalam beberapa tahun terakhir penggunaan CNN telah menunjukkan hasil yang baik dalam merepresentasikan sebuah citra menjadi sebuah *fixed-length vector*, sehingga selanjutnya bisa digunakan sebagai masukan dari *decoder* RNN untuk menghasilkan keluaran berupa kalimat deskripsi. /bigskip

CNN merupakan bagian dari *deep feedforward artificial neural networks* yang pada umumnya digunakan untuk menganalisis citra. CNN merupakan variasi dari *multilayer perceptron* dan didesain supaya memerlukan praproses seminimal mungkin. CNN terdiri dari banyak lapisan tersembunyi yang melakukan konvolusi dan *pooling* terhadap masukan yang pada umumnya berbentuk citra. Lapisan-lapisan konvolusi dan *pooling* ini memiliki nilai-nilai parameter yang dipelajari dari data sehingga Lapisan-lapisan tersebut secara otomatis menyesuaikan untuk bisa mengekstraksi informasi yang paling penting. Pada CNN, lapisan tersembunyinya biasanya terdiri atas lapisan konvolusi, lapisan *pooling*, *fully connected layers* dan lapisan normalisasi.

Lapisan konvolusi menggunakan operasi konvolusi kepada masukan, dan meneruskan hasilnya ke lapisan selanjutnya. Setiap lapisan konvolusi memproses data hanya pada *receptive field*nya. *Receptive field* adalah sebagian area dari keseluruhan data yang akan diproses. Area tersebut biasanya berbentuk persegi. *Receptive field* digunakan untuk mengurangi jumlah parameter jika dibandingkan dengan menggunakan *fully connected layer*, karena ukuran dari sebuah citra yang biasanya berukuran besar, dan setiap pixelnya merupakan input yang relevan.

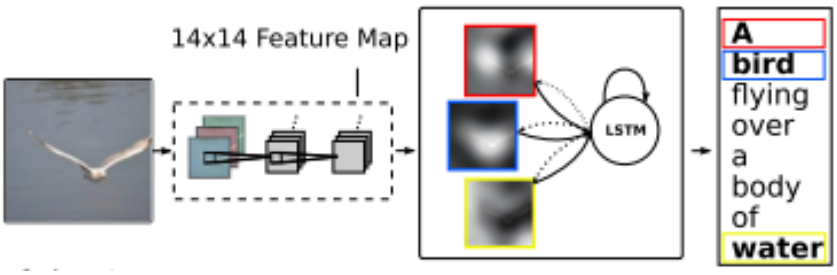
Lapisan *pooling* pada CNN bertugas untuk menggabungkan kluster neuron dari lapisan keluaran sebelumnya menjadi satu neuron pada layer selanjutnya. Misalnya, pada lapisan m*max-pooling* diambil nilai maksimum dari setiap kluster di lapisan sebelumnya. Contoh lainnya adalah lapisan *average-pooling* yang mengambil nilai rata-rata dari setiap kluster di lapisan sebelumnya.




Gambar II.6: Ilustrasi lapisan *max-pooling*

*Fully connected layer* cara kerjanya sama seperti *multilayer perceptron*, menghubungkan setiap neuron pada lapisan sebelumnya ke setiap neuron yang ada pada lapisan setelah lapisan tersebut.

Diinspirasi lebih lanjut dari permasalahan mesin translasi yang menggunakan mekanisme *attention*, penelitian Xu dkk. (2015) mengembangkan lebih lanjut penelitian Vinyals dkk. (2014) dengan memberikan mekanisme *attention* pada saat *decoder* melakukan pembangkitan deskripsi dari citra.

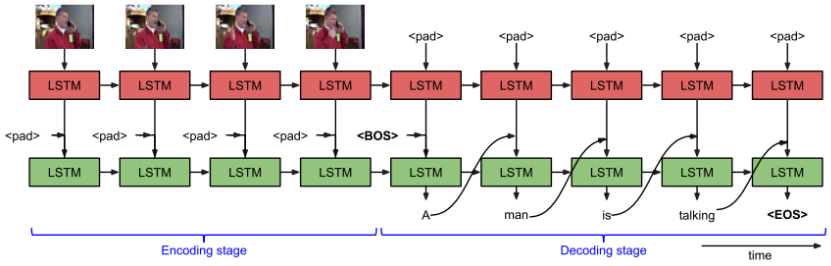



Gambar II.7: Ilustrasi model *encoder-decoder* dengan menggunakan *attention* pada *image captioning* (Xu dkk., 2015).

### II.3.2 Pembangkitan Keterangan Otomatis dari Video

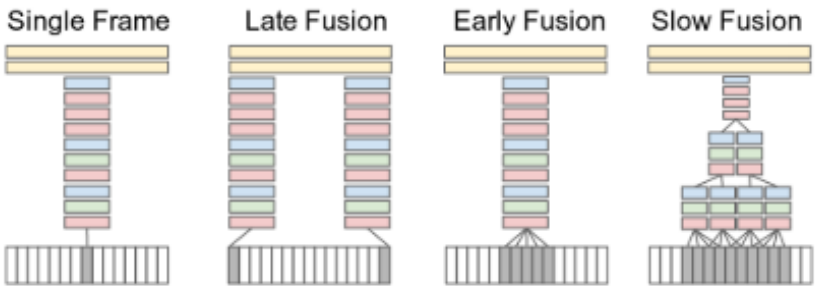
Penelitian lain yang juga terinspirasi dari perkembangan terkini dalam mesin translasi adalah penelitian mengenai pembangkitan keterangan otomatis dari video yang

dilakukan oleh Venugopalan dkk. (2015). Model yang dikemukakan menggunakan model *seqence-to-sequence* dengan *recurrent network* yang digunakan sebagai *decoder* dan *encoder*-nya adalah LSTM bertumpuk. Masukan dari LSTM bertumpuk tersebut berupa citra RGB dan informasi *optical flow* yang kemudian keduanya direpresentasikan menjadi sebuah *embedding* dengan menggunakan CNN yang parameternya dilatih bersama-sama dengan parameter yang terdapat pada LSTM bertumpuk.




Gambar II.8: Ilustrasi model *video-to-text* (Chung dkk., 2017).

Selain itu ada penelitian lain yang tidak menggunakan *recurrent network* tetapi menggunakan CNN yang telah dimodifikasi untuk dapat memanfaatkan informasi spasial dan temporal pada video, sehingga diberi nama *spatio-temporal convolutional neural network* (STCNN) (Karpathy dkk., 2014). Cara kerjanya sama seperti CNN hanya saja keterhubungan jaringannya diperluas pada dimensi waktunya sehingga model dapat mempelajari fitur spatio-temporal. Perluasannya tersebut dibagi menjadi tiga, yaitu *early fusion*, *late fusion*, dan *slow fusion*.




Gambar II.9: Ilustrasi STCNN (Karpathy dkk., 2014).

Pada *early fusion*, modelnya menggabungkan informasi temporal di seluruh frame dengan *window* yang telah ditentukan, langsung pada tingkat pixel. Sedangkan pa-

da *late fusion*, dari *window* yang telah ditentukan diambil dua buah frame, frame awal dan frame akhir, lalu masing-masing frame melalui berbagai lapisan konvolusi, lalu menggabungkan kedua representasi yang didapatkan dengan menggunakan *fully connected layer*. Untuk *slow fusion*, merupakan model yang menggabungkan dua pendekatan tadi (*early fusion* dan *late fusion*) dengan cara menggabungkan informasi temporal dari semua frame pada *window* secara perlahan-lahan dan hirarkis.

### II.4 *Visual Speech Recognition*

Penelitian Chung dkk. (2017) melakukan task pembangkitan keterangan otomatis dari video yang cakupannya lebih kecil, yaitu melakukan pengenalan gerak bibir untuk *visual speech recognition*. Model yang digunakan dapat dibagi menjadi tiga modul, yaitu modul *Watch*, modul *Listen*, modul *Attend*, dan modul *Spell*, dan bisa diformalisasikan sebagai berikut (Chung dkk., 2017):

$$s^v, o^v = Watch(x^v)$$
$$s^a, o^a = Listen(x^a)$$
$$P(y|x^v, x^a) = Spell(s^v, s^a, o^v, o^a)$$

Modul *Watch*, merupakan *encoder* citra yang terdiri atas modul konvolusi yang menghasilkan fitur citra $f_i^v$ untuk setiap masukan di *time-step* $x_i^v$, ditambah dengan sebuah modul rekuren yang menghasilkan fixed-length vector $s^v$ dan sekumpulan vektor output $o^v$.

Modul *Listen*, merupakan *encoder* yang berisi sama seperti modul *Watch* tetapi tanpa memiliki modul konvolusi. Modul rekurennya menerima masukan berupa hasil ekstraksi fitur menggunakan MFCC berdimensi 13, lalu dari modul rekuren tersebut dihasilkan vektor berukuran tetap $s^a$ dan sekumpulan vektor keluaran $o^a$.

Modul *Spell*, berbasis pada transduser LSTM dan menghasilkan keluaran berupa rangkaian token pada level karakter. Modul ini bisa diformulasikan sebagai berikut

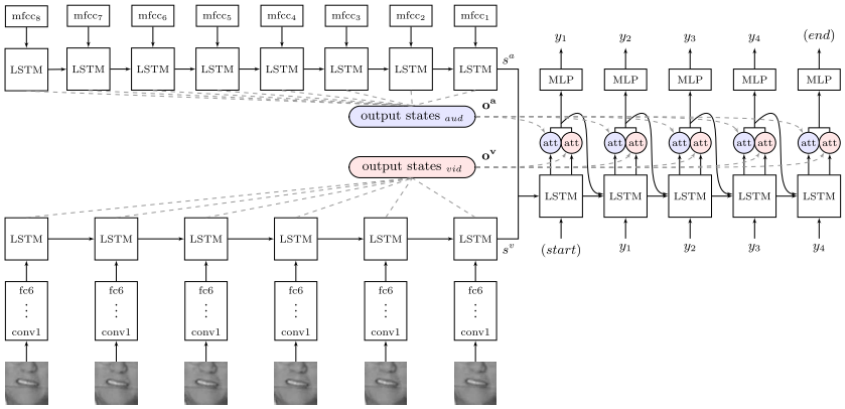
(Chung dkk., 2017).

$$h_k^d, h_k^d = LSTM(h_{k-1}^d, y_{k-1}, c_{k-1}^v, c_{k-1}^a)$$

$$c_k^v = o^v \cdot Attention^v(h_k^d, o^v)$$

$$c_k^a = o^a \cdot Attention^a(h_k^d, o^a)$$

$$P(y_i|x^v, x^a, y_{<i}) = softmax(MLP(o_k^d, c_k^v, c_k^a))$$




Gambar II.10: Arsitektur model *Watch*, *Listen*, *Attend*, dan *Spell* (Chung dkk., 2017).

# 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.

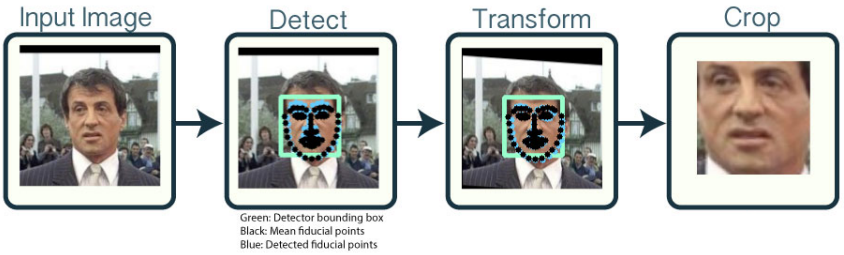
### III.2.1 Persiapan Korpus Video

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

Untuk bagian video, sebelum masuk ke dalam model transformer, setiap framenya akan melalui tahap pengenalan wajah terlebih dahulu untuk diketahui apakah frame tersebut terdapat wajah manusia atau tidak, lalu kemudian jika ditemukan wajah maka akan dicari *landmark* pada wajah tersebut, sehingga dapat diketahui letak bibir pada wajah tersebut. Tahap pengenalan wajah ini dilakukan dengan menggunakan *library* OpenFace. OpenFace merupakan *library* Python dan Torch dalam mengimplementasikan modul pengenalan wajah yang menggunakan algoritma DNN (*deep neural network*). Torch mendukung agar OpenFace dapat diekseku-

si dengan menggunakan CPU maupun dengan menggunakan GPU (dalam hal ini menggunakan CUDA). OpenFace menggunakan beberapa *library* pendukung yakni dlib dan OpenCV.




Gambar III.1: Alur pengenalan wajah pada OpenFace.

### III.2.3 Pembangunan Kamus Pelafalan

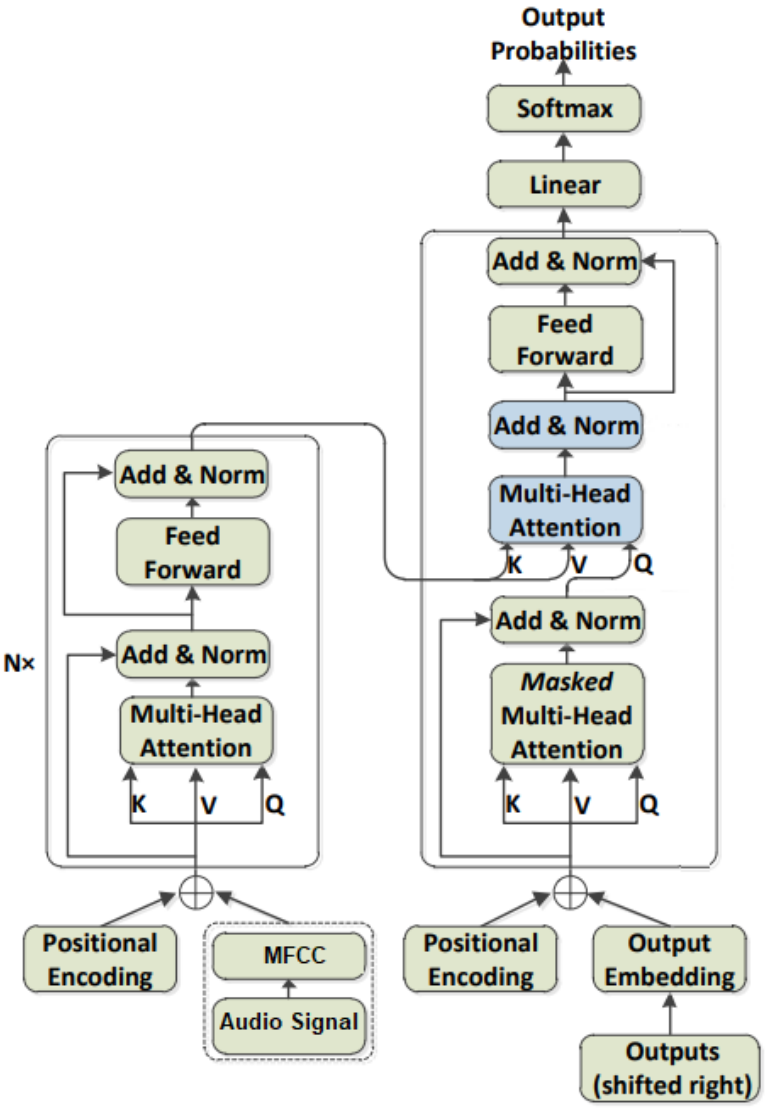
Kamus pelafalan dibangun dengan cara memetakan setiap kata yang ada pada corpus ke dalam fonem bahasa Indonesia. Fonem bahasa Indonesia tersebut seluruhnya didaftarkan sebagai fonem penyusun leksikon. Pada penelitian ini kamus pelafalan dibangun dengan menggunakan kakas Corpus Management Tools, yang dikembangkan oleh Hoesen (2015). Kakas bekerja dengan cara mengolah teks masukan kemudian membangun skema pelafalan kata bahasa Indonesia berdasarkan aspek fonetis.

### III.2.4 Rancangan Arsitektur

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.2) 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 menggunakan MFCC. Selanjutnya model ini akan disebut sebagai model transASR.

2. model yang menggunakan modal visual saja. Model ini menggunakan arsitektur transformer (Gambar III.3) 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

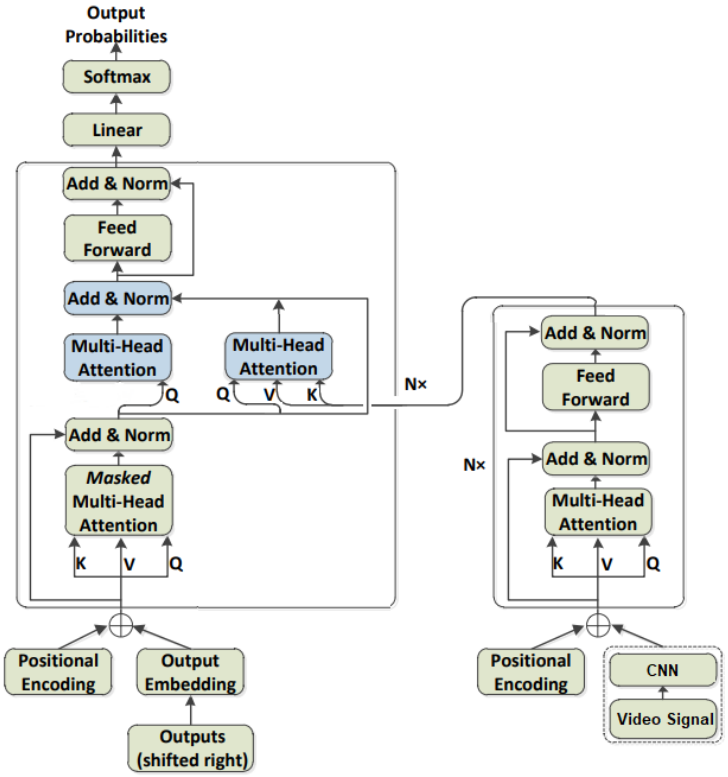



Gambar III.2: Rancangan Arsitektur ASR.

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.4) 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 VSR.

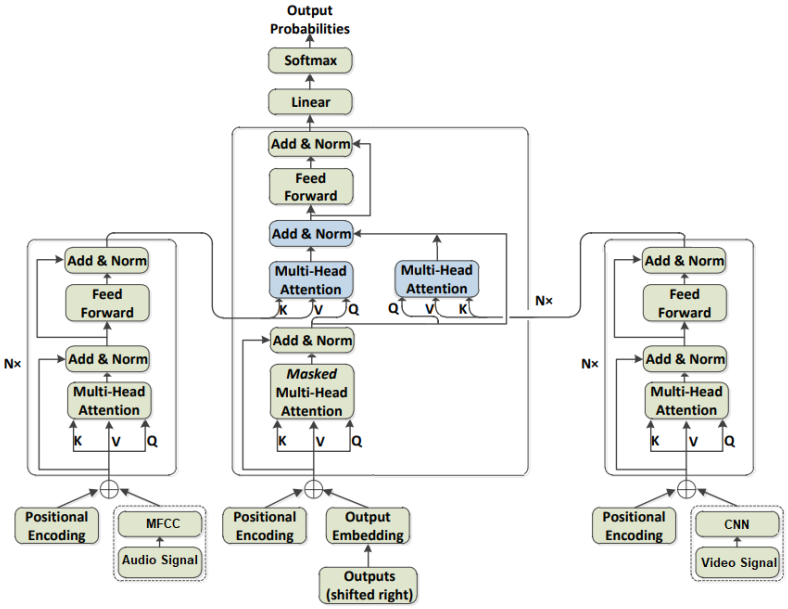
digunakan juga untuk masukan di time-step selanjutnya.

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 II.2.2.

### III.2.5 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:

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




Gambar III.4: Rancangan Arsitektur AVSR.

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.

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

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

$$BLEU = p \times e^{\sum_{n=1}^{N}(\frac{1}{N} \times logPn)}$$

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. Jenis BLEU yang akan digunakan dalam evaluasi ini adalah unigram BLEU.

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

# Bab IV Eksperimen dan Evaluasi

Bab ini menjelaskan tahapan implementasi pembangunan model, konfigurasi eksperimen, dan evaluasi kinerja model berdasarkan hasil eksperimen yang dilakukan.

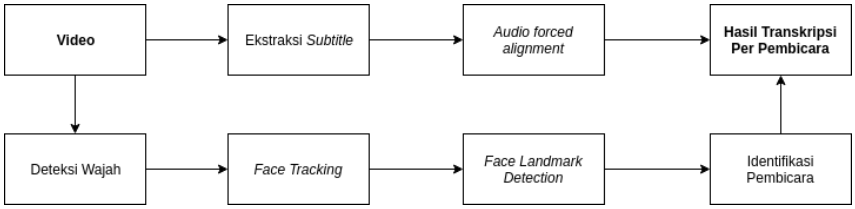
## IV.1 Tujuan Eksperimen

Eksperimen dalam penelitian ini dilakukan dalam penentuan model bahasa dan parameter model, dan juga untuk mengukur kinerja model ASR, VSR, dan AVSR yang dibangun. Kinerja yang diukur adalah hasil pembangkitan kalimat yang ditranskripsikan dari suara dan gambar pada video. Kinerja tersebut diukur dengan menggunakan metrik *word error rate* (WER). Kualitas pembangkitan transkripsi dinilai baik jika memilki WER yang rendah.

## IV.2 Pembangunan Model

Berdasarkan hasil eksperimen yang dilakukan, dibangun model ASR, VSR, dan AVSR yang sesuai untuk mengenali ucapan dalam bahasa Indonesia.

### IV.2.1 Persiapan dan Pembentukan Transkripsi

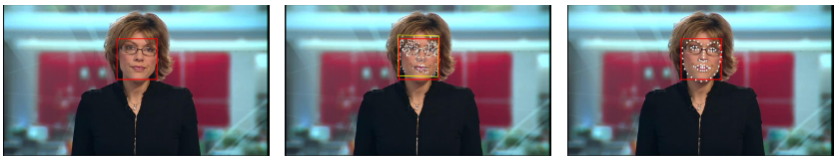



Gambar IV.1: Alur proses persiapan transkripsi.

Untuk melatih model diperlukan transkripsi dengan cap waktu yang selaras dengan suara yang diucapkan. Beberapa video di YouTube sudah ada yang menyediakan *subtitle* dan juga sudah selaras waktunya dengan kata-kata yang diucapkan di video. Akan tetapi subtitle tersebut hanya memiliki cap waktu dalam satuan kalimat, tidak kata per kata, sehingga perlu dilakukan penyelarasan menggunakan Penn Phonetics Lab Forced Aligner (dibangun berdasarkan kakas sumber terbuka HTK *Toolbox*). Teks yang akan diselaraskan merupakan teks dari *subtitle* yang kemudian dipetakan menjadi cara pelafalannya dengan kamus pelafalan yang sudah dibuat dengan menggunakan kakas Corpus Management Tools. Kakas penyelaras tersebut menggunakan algoritma Viterbi untuk menghitung *maximum likelihood alignment* antara

audio (yang dimodelkan menggunakan fitur PLP) dengan teks.

### IV.2.2 Persiapan Korpus Video



Gambar IV.2: **Kiri:** Pendeteksian wajah (*bounding box* merah). **Tengah:** *tracking* wajah menggunakan fitur KLT (*bounding box* kuning). **Kanan:** Pendeteksian *landmark* wajah.

Corpus video yang telah dikumpulkan dideteksi wajah-wajah yang terdapat pada setiap framenya menggunakan metode berbasis HOG. Setelah wajah berhasil dideteksi, masing-masing wajah tersebut di*tracking* menggunakan KLT *tracker*, yang berguna juga dalam mengurangi *false positive* pada saat tahap pendeteksian wajah. Kemudian dari wajah yang sudah terdeteksi tersebut dideteksi *landmark*nya, yang kemudian *landmark* tersebut bisa digunakan untuk menentukan posisi mulut pada wajah. Untuk menentukan siapa yang sedang berbicara pada video, ditentukan dengan cara menghitung jarak ternormalisasi antara bibir atas dan bibir bawah dari setiap wajah yang terdeteksi, dan juga dihitung frekuensi buka tutup dari bibirnya.

### IV.2.3 Ekstraksi Fitur

Terdapat dua ekstraksi fitur yang digunakan dalam penelitian ini, yaitu MFCC yang digunakan untuk mengektrasi fitur audio, dan CNN yang digunakan untuk mengektraksi fitur frame pada video. Konfigurasi parameter dari MFCC menggunakan 36-MFCC berdasarkan penelitian Yuwan (2018) dan untuk konfigurasi parameter CNN menggunakan konfigurasi pada penelitian Chung dkk. (2017).

### IV.2.4 Eksperimen Pemodelan Sekuens

Pada eksperimen pemodelan sekuens, sekuens akan memodelkan dari tiga jenis masukan, yaitu masukan akustik saja, masukan visual saja, dan masukan gabungan akustik dan visual. Akan tetapi arsitektur model dari sekuensnya itu sendiri tetap sama, sehingga parameter-parameter yang bisa diuji pun jumlahnya akan tetap sama. Pemodelan sekuens yang digunakan adalah model transformer dan diimplementasikan dengan menggunakan TensorFlow dan PyTorch, yang disediakan oleh

Dai dkk. (2019) pada laman GitHub [1].

Sebelum pemodelan sekuens dilakukan, perlu dilakukan proses ekstraksi fitur seperti yang sudah dijelaskan sebelumnya. Setelah itu, model transformer dibentuk dengan menggunakan perintah

```
bash run_enwik8_base.sh train --work_dir PATH_TO_WORK_DIR
```

untuk melakukan proses pelatihan dan menggunakan perintah

```
bash run_enwik8_base.sh eval --work_dir PATH_TO_WORK_DIR
```

untuk melakukan proses evaluasi. `PATH_TO_WORK_DIR` adalah direktori tempat hasil pemodelan disimpan. Selain itu ada juga opsi-opsi tambahan yang diuji seperti

- `--batch_chuck` untuk menukar performa kecepatan pelatihan dengan memori yang digunakan. Untuk `batch_chuck>1`, program akan membagi setiap data latih menjadi `batch_chuck` bagian dan melakukan pelatihan pada setiap *batch* secara berurutan, dan gradien yang terkumpul akan dibagi dengan jumlah `batch_chuck`.
- `--div_val` untuk mengurangi dimensi *embedding*.
- `--fp16` dan `--dynamic-loss-scale` untuk menjalankan pelatihan dengan menggunakan mode pseudo-fp16 dan *dynamic loss scaling*. Untuk penggunaan `--fp16` perlu dilakukan pemasangan *package* `apex`[2] terlebih dahulu.
- `mem_len=0` untuk melakukan pelatihan tanpa menggunakan mekanisme rekurens.
- `attn_type=2` untuk melakukan pelatihan dengan model transformer standar tanpa menggunakan *positional encoding* relatif.

### IV.3 Skenario Eksperimen

Proses pengujian dibagi menjadi tiga tahap. Pengujian pertama dilakukan terhadap pembangkitan transkripsi menggunakan model akustik. Pengujian kedua dilakukan terhadap pembangkitan transkripsi menggunakan model visual. Terakhir, pengujian

[1] `https://github.com/kimiyoung/transformer-xl`
[2] `https://github.com/NVIDIA/apex/`

ketiga dilakukan terhadap pembangkitan transkripsi menggunakan model akustik dan model visual.

## IV.4 Hasil Eksperimen dan Evaluasi

Upabab ini membahas tentang hasil eksperimen pemodelan sekuens sesuai dengan implementasi pada upabab IV.2

### IV.4.1 Parameter Model ASR

### IV.4.2 Parameter Model VSR

### IV.4.3 Parameter Model AVSR

# Bab V Penutup

## V.1 Kesimpulan

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.

## V.2 Saran

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.

# Daftar Pustaka


Achmad, B. dan L. Fadillah (2015). "Lip Motion Pattern Recognition for Indonesian Syllable Pronunciation Utilizing Hidden Markov Model Method". *TELKOMNIKA* 13.1.

Arifin, Muljono, S. Sumpeno, dan M. Hariadi (2013). "Towards building Indonesian viseme: A clustering-based approach". *2013 IEEE International Conference on Computational Intelligence and Cybernetics (CYBERNETICSCOM)*. IEEE, pp. 57–61.

Assael, Y. M., B. Shillingford, S. Whiteson, dan N. De Freitas (2016). "LIPNET: END-TO-END SENTENCE-LEVEL LIPREADING".

Bahdanau, D., K. Cho, dan Y. Bengio (2015). "Neural Machine Translation by Jointly Learning to Align and Translate". arXiv: 1409.0473.

Calvert, G. A., C. Spence, dan B. E. Stein (2004). *The Handbook of Multisensory Processes*.

Chan, W., N. Jaitly, Q. V. Le, dan O. Vinyals (2015). "Listen, Attend and Spell". arXiv: 1508.01211.

Cho, K., B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, dan Y. Bengio (2014a). "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation". arXiv: 1406.1078.

Cho, K., B. van Merrienboer, D. Bahdanau, dan Y. Bengio (2014b). "On the Properties of Neural Machine Translation: Encoder-Decoder Approaches". arXiv: 1409.1259.

Chung, J. S. dan A. Zisserman (2016). "Lip Reading in the Wild".

Chung, J. S., A. Senior, O. Vinyals, dan A. Zisserman (2017). "Lip Reading Sentences in the Wild". *IEEE Conference on Computer Vision and Pattern Recognition*.

Chung, J., C. Gulcehre, K. Cho, dan Y. Bengio (2014). "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling".

Dai, Z., Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, dan R. Salakhutdinov (2019). "Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context". arXiv: 1901.02860.

Garg, A., J. Noyola, dan S. Bagadia (2016). "Lip reading using CNN and LSTM".

Han, K. J., A. Chandrashekaran, J. Kim, dan I. Lane (2018). *The CAPIO 2017 Conversational Speech Recognition System*. Tech. rep. arXiv: arXiv : 1801 . 00059v2.

Hochreiter, S. dan J. Urgen Schmidhuber (1997). "Long Short-Term Memory". *Neural Computation* 9.8, pp. 1735–1780.

Karpathy, A., G. Toderici, S. Shetty, T. Leung, R. Sukthankar, dan L. Fei-Fei (2014). "Large-scale Video Classification with Convolutional Neural Networks".

Maulana, M. R. A. R. dan M. I. Fanany (2017). *Sentence-level Indonesian Lip Reading with Spatiotemporal CNN and Gated RNN*. Tech. rep.

Noda, K., Y. Yamaguchi, K. Nakadai, H. G. Okuno, dan T. Ogata (2014). "Lipreading using Convolutional Neural Network".

Sutskever, I., O. Vinyals, dan Q. V. Le (2014). "Sequence to Sequence Learning with Neural Networks". *Eletronic Proceedings of Neural Information Processing Systems 2014*.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, dan I. Polosukhin (2017). "Attention Is All You Need". *Eletronic Proceedings of Neural Information Processing Systems 2017*.

Venugopalan, S., M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, dan K. Saenko (2015). "Sequence to Sequence – Video to Text". *International Conference on Computer Vision 2015*.

Vinyals, O., A. Toshev, S. Bengio, dan D. Erhan (2014). "Show and Tell: A Neural Image Caption Generator". arXiv: 1411.4555.

Wand, M., J. Koutník, dan J. Schmidhuber (2016). "Lipreading with Long Short-Term Memory".

Xiong, W, L Wu, F Alleva, J Droppo, X Huang, dan A Stolcke (2017). *The Microsoft 2017 Conversational Speech Recognition System*. Tech. rep. arXiv: arXiv: 1708.06073v2.

Xu, K., J. L. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. S. Zemel, dan Y. Bengio (2015). “Show, Attend and Tell: Neural Image Caption Generation with Visual Attention”. *Proceedings of Machine Learning Research*.

Yu, D. dan L. Deng (2014). *Automatic Speech Recognition: A Deep Learning Approach*. Vol. 9. 2.

Yuwan, R. (2018). “Pemodelan Akustik Berbasis Deep Neural Network Pada Sistem Pengenal Ucapan Spontan Bahasa Indonesia Memanfaatkan Active Learning”. PhD thesis. Institut Teknologi Bandung, p. 51.

Zhou, Z., G. Zhao, X. Hong, dan M. Pietikäinen (2014). “A review of recent advances in visual speech decoding”.