**PENERAPAN SPEECH RECOGNITION MENGGUNAKAN METODE LONG SHORT-TERM MEMORY (LSTM) UNTUK PRESENTASI DINAMIS**

**LAPORAN TUGAS AKHIR**



**SATRIYA ADHITAMA**

**5200411545**

**PROGRAM STUDI INFORMATIKA**

**FAKULTAS SAINS & TEKNOLOGI**

**UNIVERSITAS TEKNOLOGI YOGYAKARTA**

**YOGYAKARTA**

**2022**

# KATA PENGANTAR

Puji syukur dipanjatkan atas kehadirat Allah SWT, karena dengan limpahan karunia-Nya penulis dapat menyelesaikan Tugas Akhir dengan judul ***“Penerapan Speech Recognition Menggunakan Metode Long Short-Term Memory (LSTM) untuk Presentasi Dinamis”.***

Penyusunan Tugas Akhir diajukan sebagai salah satu syarat untuk memperoleh gelar sarjana pada Program Studi Informatika Fakultas Sains & Teknologi Universitas Teknologi Yogyakarta.

Tugas Akhir ini dapat diselesaikan tidak lepas dari segala bantuan, bimbingan, dorongan dan doa dari berbagai pihak, yang pada kesempatan ini penulis ingin menyampaikan ucapan terima kasih kepada:

1. Bapak Bambang Moertono S., Dr., MM.,Akt.,CA.
2. Ibu Endy Marlina, Dr.,MT.
3. Ibu Enny Itje Sela, Dr.,S.Si.,M.Kom
4. Bapak Donny Avianto, S.T.,M.T.

Akhir kata, penulis menyadari bahwa sepenuhnya akan terbatasnya pengetahuan penyusun, sehingga tidak menutup kemungkinan jika ada kesalahan serta kekurangan dalam penyusunan Tugas Akhir, untuk itu sumbang saran dari pembaca sangat diharapkan sebagai bahan pelajaran berharga dimasa yang akan datang.

Yogyakarta, 3 Oktober 2022

Penulis

# DAFTAR ISI

[KATA PENGANTAR ii](#_Toc134933521)

[DAFTAR ISI iii](#_Toc134933522)

[DAFTAR GAMBAR iv](#_Toc134933523)

[DAFTAR TABEL v](#_Toc134933524)

[BAB I PENDAHULUAN 1](#_Toc134933525)

[1.1 Latar Belakang 1](#_Toc134933526)

[1.2 Rumusan Masalah 2](#_Toc134933527)

[1.3 Batasan Masalah 3](#_Toc134933528)

[1.4 Tujuan penelitian 3](#_Toc134933529)

[1.5 Manfaat Penelitian 3](#_Toc134933530)

[1.6 Sistematika Penulisan 3](#_Toc134933531)

[BAB II KAJIAN HASIL PENELITIAN DAN LANDASAN TEORI 5](#_Toc134933532)

[2.1 Kajian Hasil Penelitian 5](#_Toc134933533)

[2.2 Landasan Teori 8](#_Toc134933534)

[2.2.1 Automatic Speech Recognition 8](#_Toc134933535)

[2.2.2 Long Short-Term Memory (LSTM) 10](#_Toc134933536)

[BAB III METODE PENELITIAN 12](#_Toc134933537)

[3.1 Bahan/Data 12](#_Toc134933538)

[3.1.1 Data yang diperoleh 12](#_Toc134933539)

[3.1.2 Prosedur pengumpulan data 14](#_Toc134933540)

[3.2 Aturan Bisnis 14](#_Toc134933541)

[BAB IV ANALISIS DAN PERANCANGAN SISTEM 15](#_Toc134933542)

[4.1 Analisis Sistem yang Berjalan 15](#_Toc134933543)

[4.2 Analisis Sistem yang Diusulkan 16](#_Toc134933544)

[4.3 Desain Sistem 16](#_Toc134933545)

[BAB V Penutup 17](#_Toc134933546)

[5.1 Simpulan 17](#_Toc134933547)

[5.2 Saran 17](#_Toc134933548)

[DAFTAR PUSTAKA 18](#_Toc134933549)

# DAFTAR GAMBAR

[Gambar 2. 1 Arsitektur ASR 9](#_Toc134933562)

[Gambar 2. 2 Arsitektur LSTM 10](#_Toc134933563)

[Gambar 3. 1 Kerangka Penelitian 12](#_Toc124192515)

[Gambar 3. 2 Arsitektur sistem saat ini 15](#_Toc124192516)

[Gambar 3. 3 Arsitektur sistem usulan 16](#_Toc124192517)

[Gambar 3. 4 Flowchart sistem 19](#_Toc124192518)

[Gambar 3. 5 Flowchart *Keyword Spotting* 20](#_Toc124192519)

[Gambar 3. 6 Flowchart pembuatan model LSTM 21](#_Toc124192520)

[Gambar 3. 7 DFD Level 0 22](#_Toc124192521)

[Gambar 3. 8 DFD Level 1 23](#_Toc124192522)

[Gambar 3. 9 Diagram HIPO 24](#_Toc124192523)

[Gambar 3. 10 *Entity Relationship Diagram* (ERD) 25](#_Toc124192524)

[Gambar 3. 11 Tampilan menu utama 27](#_Toc124192525)

[Gambar 3. 12 Tampilan *real-time speech recognition* 27](#_Toc124192526)

[Gambar 3. 13 Tampilan *voice recognition stream* tidak terdeteksi kata perintah 28](#_Toc124192527)

[Gambar 3. 14 Tampilan *voice recognition stream* terdeteksi kata perintah 28](#_Toc124192528)

[Gambar 3. 15 Tampilan input file suara 29](#_Toc124192529)

[Gambar 3. 16 Tampilan spectogram dari file suara 29](#_Toc124192530)

[Gambar 3. 17 Tampilan spectogram dengan deteksi dari file suara 30](#_Toc124192531)

[Gambar 4. 1 Tahapan persiapan data 32](#_Toc124254256)

# DAFTAR TABEL

[Tabel 1. 1 Sistematika Penulisan 4](#_Toc124192565)

[Tabel 2. 1 Penelitian Terdahulu 7](#_Toc124192570)

[Tabel 3. 1 Dataset *speech\_commands\_v0.02* 13](#_Toc124192585)

[Tabel 4. 1 Daftar kata perintah sistem 31](#_Toc124254236)

[Tabel 4. 2 Fungsi kata perintah sistem 31](#_Toc124254237)

[Tabel 4. 3 Pembagian data latih, uji, dan validasi 32](#_Toc124254238)

[Tabel 4. 4 Hasil data metrik data latih 34](#_Toc124254239)

# PENDAHULUAN

## Latar Belakang

Perkembangan teknologi yang semakin maju menjadikan berbagai aktivitas untuk lebih mudah dan efektif. Banyak pekerjaan ringan hingga berat dapat disederhanakan dari segi proses persiapan sampai penggunaan. Teknologi ini menjawab berbagai kebutuhan untuk segala aspek dan bidang kehidupan. Salah satu bidang yang mendapatkan manfaat besar dari kehadiran teknologi adalah komunikasi.

Presentasi merupakan metode komunikasi dengan menyampaikan informasi, ide, dan gagasan oleh seseorang kepada sekelompok orang/audiens yang banyak. Power point menjadi salah satu media yang dapat mendukung pembicara untuk dapat menyampaikan informasi, ide, atau gagasan yang diinginkan. Power point dapat berupa informasi visual dan tulisan yang berupa inti dari ide atau gagasan yang akan disampaikan. Tulisan-tulisan dan gambar ini akan ditampilkan pada layar untuk dapat dilihat oleh audiens secara jelas. Kemudian, pembicara akan memaparkan apa yang menjadi penjelasan dari tulisan dan/atau gambar yang ditampilkan.

Power point dapat disusun secara dinamis. Dinamis yang dimaksud adalah pembicara dapat melakukan interaksi dengan leluasa terhadap bagian-bagian tertentu yang terdapat dalam power point yang disusun. Pembicara dapat memajukan atau memundurkan slide presentasi, menuju ke bagian tertentu, menjalankan audio atau video, menampilkan bagian yang disembunyikan, dan lain sebagainya. Pembicara akan berinteraksi secara dinamis terhadap power point agar informasi yang ingin disampaikan dapat tepat sasaran dan audiens lebih tertarik untuk mendengar pembicara dan melihat layar presentasi.

Namun, terkadang materi yang disusun secara dinamis pada software pendukung presentasi akan mengakibatkan pembicara kesusahan untuk berinteraksi. Pembicara harus selalu mengarahkan pointer ke bagian yang dinginkan atau membutuhkan asisten untuk dapat membantu melakukan interaksi dengan software presentasi. Pembicara memerlukan alat tambahan untuk dapat mengoperasikan software presentasi seperti keyboard, mouse, atau remote control.

Hal ini menimbulkan distraksi bagi pembicara karena harus berurusan dengan hal teknis presentasi daripada fokus dengan informasi apa yang harus disampaikan. Teknis presentasi untuk memilih konten yang ingin ditampilkan selama presentasi berlangsung akan menjadi gangguan tersendiri bagi presenter. Perlu adanya teknologi yang bisa membantu pembicara untuk melakukan presentasi secara dinamis dengan mudah dan minim distraksi teknis. Maka dari itu, disusunlah penelitian ini untuk merancang sistem yang mampu merekognisi suara kemudian mengubahnya menjadi perintah (command) guna mengoperasikan perangkat lunak pembantu presentasi.

Implementasi pengenalan ucapan (speech recognition) dilakuakan menggunakan model machine learning sequential Long Short-Term Memory (LSTM). LSTM yang merupakan turunan dari RNN telah mencapai peningkatan performa untuk Automatic Speech Recognition (Oruh dkk., 2022). LSTM dapat menangkap informasi time series yang cocok digunakan untuk pengenalan ucapan (speech recognition) (Xiang dkk., 2019). LSTM akan dapat mengenali kata tertentu yang muncul pada sinyal suara (keyword spotting). Dengan demikian, LSTM menjadi pilihan yang tepat untuk diimplementasikan pada sistem pengenalan ucapan software presentasi.

## Rumusan Masalah

Berdasarkan latar belakang tersebut, penelitian ini memiliki rumusan masalah yaitu distraksi pengoperasian *software* presentasi bagi pembicara. *Software* presentasi hanya dapat dioperasikan menggunakan perangkat pendukung lain seperti *keyboard, mouse,* atau *remote control*. Hal ini tentu akan mengganggu jalannya presentasi pembicara karena diperlukannya aktivitas tambahan selain harus fokus terhadap materi yang akan dikomunikasikan. Dengan demikian, diperlukan sistem *speech-to-command* guna membantu memudahkan teknis presentasi.

## Batasan Masalah

Penelitian “Penerapan Speech Recognition Menggunakan Metode *Long Short-Term Memory* (LSTM) untuk Presentasi Dinamis”, yang mencakup berbagai hal, sebagai berikut:

1. Perancangan sistem untuk dapat melakukan kontrol presentasi *power point* desktop dengan perintah suara;
2. Perancangan dilakukan dalam bahasa pemrograman python;
3. Speech recognition diterapkan menggunakan metode *long short-term memory* (LSTM)

## Tujuan penelitian

Penelitian ini bertujuan untuk mengimplementasi dan menganalisis sistem speech recognition menggunakan metode *Long Short-Term Memory* (LSTM) yang dapat digunakan sebagai penunjang bagi pembicara untuk dapat melakukan presentasi secara dinamis hanya dengan perintah suara.

## Manfaat Penelitian

Implementasi *speech recognition* untuk presentasi dapat memudahkan *presenter*/pembicara untuk dapat mengoperasikan *software* presentasi tanpa menggunakan perangkat tambahan pembantu lain seperti *mouse*, *keyboard*, atau *remote control*. Dengan demikian, distraksi yang muncul dari penggunaan perangkat tambahan pembantu ini dapat dikurangi, sehingga pembicara dapat lebih fokus untuk menyampaikan materi presentasi dengan baik ke audiens.

## Sistematika Penulisan

Tugas akhir ini disusun dengan sistematika yang merupakan kerangka dan pedoman penulisan untuk memudahkan mengetahui dan memahami pembahasan yang terkandung secara menyeluruh. Berikut adalah sistematika penulisan yang digunakan:

Tabel 1. 1 Sistematika Penulisan

|  |  |
| --- | --- |
| **BAB I** | **PENDAHULUAN**  Bab I berisi penjelasan mengenai masalah yang melatarbalakangi mengapa penelitian dilakukan. Pada bab ini disajikan rincian tentang masalah dari umum ke khusus. Adapun susunan dari bab ini terdiri atas latar belakang, rumusan masalah, manfaat penelitian, dan sistematika penulisan |
| **BAB II** | **KAJIAN HASIL PENELITIAN DAN LANDASAN TEORI**  Bab II meliputi penelitian-penelitian terdahulu yang relevan terhadap pembahasan yang akan disajikan pada penelitian sekarang, yaitu berkaitan tentang *speech recognition* dan metode LSTM. Kemudian, dasar-dasar teori yang dibutuhkan pada penelitian ini akan dijelaskan secara rinci, seperti  *automatic speech recognition* (ASR)*, Long Short-Term Memory* (LSTM). |
| **BAB III** | **METODE PENELITIAN**  Bab ini menjelaskan tentang metode penelitian yang dilakukan penulis dalam pengembangan rancangan implementasi sistem yang meliputi desain penelitian, penentuan objek penelitian, fokus penelitian, sumber data, teknik pengumpulan data, dan teknik analisis data. |
| **BAB IV** | **IMPLEMENTASI DAN HASIL SERTA PEMBAHASAN**  Bab ini terdiri dari hasil implementasi dan analisis penelitian. Implementasi dari sistem pengenalan ucapan (*speech* recognition) menggunakan metode LSTM akan diujikan berdasarkan parameter yang telah ditentukan sehingga dapat diketahui bangaimana tingkat keberhasilan penelitian untuk dapat menjawab masalah yang ada. |
| **BAB VI** | **PENUTUP**  Bab ini berisi kesimpulan dan saran dari penelitian yang telah dilakukan. Kesimpulan menjelaskan penyelesaian dari rumusan masalah yang dilakukan secara objektif. Saran berisi jalan keluar untuk mengatasi segala masalah yang ada pada penelitian. |

# KAJIAN HASIL PENELITIAN DAN LANDASAN TEORI

## Kajian Hasil Penelitian

Beberapa penelitian berikut merupakan hasil dari peneliti terdahulu yang memiliki bidang dan tema yang sama dengan penelitian yang akan dilakukan. Kajian penelitian ini diperlukan untuk dijadikan sebagai dasar alasan mengapa metode *long short-term memory* (LSTM) dapat diterapkan dengan baik untuk masalah pengenalan ucapan (*speech recognition*).

Pertama, penelitian oleh Cahuantzi, dkk (2021) membandingkan antara 2 metode untuk *learning symbolic sequences*, yaitu *long short-term memory* (LSTM) dan *gated recurrent unit* (GRU). Kompleksitas *seed strings* digunakan sebagai pembanding kedua metode tersebut untuk dapat mengetahui metode mana yang memiliki performa atau waktu *learning* yang terbaik. Kedua metode akan diuji pada kompleksitas *seed strings* yang berbeda. Pada kompleksitas rendah dilakukan pelatihan pada 3.600 RNN dengan total 37 *seed strings* berbeda, minimal 1.100 panjang karakter, kompleksitas LZW antara 2 hingga 12, jumlah unit pada *hidden layer* mencapai 25 hingga 250, serta 0,001 dan 0,01 untuk learning rate. Proses *training* menghasilkan median waktu latih 37,19 detik untuk LSTM dan 19,72 detik untuk GRU. Pada kompleksitas tinggi, digunakan 300 *seed strings* dengan 10, 33, atau 52 simbol, komleksitas LZW di antara 1.000 hingga 1.850, panjang karakter mencapai 2.400, dan total 4.500 RNN digunakan untuk *training*. Proses *training* untuk kompleksitas tinggi menghasilkan median waktu latih 12,53 untuk LSTM dan 22,84 detik untuk GTU. Hasil pengujian menunjukkan bahwa LSTM jauh mengungguli GRU dalam waktu latih untuk kompleksitas *sequences* yang tinggi. Peneliti juga menunjukkan bahwa LSTM dan GRU mencapai akurasi yang hampir sama berdasarkan perbandingan kompleksitas LZW dengan metrik jarak Damerau-Lavenshtein (DL) dan Jaro-Winkler (JW), namun LSTM dapat melatih lebih cepat dibanding GRU. Hal ini membuktikan bahwa LSTM dapat digunakan dengan baik untuk data yang memiliki kompleksitas *sequence* tinggi dengan mencapai perform metrics a latih yang cepat dan akurasi prediksi yang tidak jauh beda dengan metode pembandingnya (GRU).

Kedua, penelitian oleh Oruh dkk (2022) adalah tentang pengembangan model *hybrid* *long short-term memory-recurrent neural network* (LSTM-RNN) untuk *automatic speech recognition* (ASR). LSTM-RNN merupakan model gabungan dengan mengintegrasikan RNN sebagai *forget gate* pada arsitektur LSTM. Modifikasi ini menyebabkan perubahan pada peningkatan sumber komputasi. Penulis membandingkan performa LSTM-RNN terhadap model *deep learning* dan *sequential* lain*.* Pengujian dilakukan pada 2.400 file suara berbeda dengan format wav. File suara ini terdiri atas 15 penutur (pria & wanita) dengan masing-masing mengucapkan angka 0-9 sebanyak 16 kali. Penulis menetapkan sebanyak 90% (2.160 pengucapan) untuk data latih dan 10% (240 pengucapan) untuk data uji. Perbandingan LSTM-RNN dengan model *deep learning* lain menunjukkan bahwa model ini memiliki akurasi yang tinggi. LSTM-RNN menghasilkan 99,36%, 72,83% untuk ResNet-18, 74,17% untuk ResNet-34, 89,67% untuk DenseNet-121, 87,17% untuk DenseNet-169, dan 77,17% untuk VGG-16. Perbandingan LSTM-RNN dengan *sequential model* lain juga menunjukkan bahwa LSTM-RNN memiliki tingkat akurasi yang lebih tinggi. LSTM-RNN menghasilkan 99,36%, 87,11% untuk *simple LSTM*, 80,86% untuk *bidirectional LSTM*, 83,59% untuk *simple RNN*, dan 90,23% untuk GRU. Hasil penelitian tersebut membuktikan bahwa LSTM dengan model *hybrid* maupun *non-hybrid* memiliki performa yang baik untuk penerapan *automatic speech recognition* (ASR). Pada model *sequential* lain yang dibandingkan, variasi dari LSTM juga menunjukkan akurasi yang cukup baik, yaitu lebih dari 80%.

Ketiga, penelitian oleh Atcheson dkk (2019) menggunakan proses Gaussian pada LSTM dan MLP untuk memprediksi *continuous-time* dan dimensi emosi pada ucapan ambigu. Penulis mengimplementasikan metode LSTM dan MLP dengan atau tanpa proses Gaussian untuk kasus *contiuous emotion recognition* (CER), sehingga dapat mengetahui nada emosi pada ucapan. Dengan demikian, dapat melengkapi pekerjaan *automatic speech recognition* (ASR), dengan mencoba melihat konten linguistik pada ucapan. Penelitian dilakukan dengan mengukur *condordance correlation coefficient (*CCC) antara rangkaian waktu yang terprediksi dan mean dari anotasi target. Selain itu juga dihitung *som of the log probabilities* (SLP), dengan semakin tinggi nilai maka semakin akurat prediksinya. Eksperimen yang telah dilakukan menujukkan bahwa *long short-term memory-gaussian process* (LSTM-GP) memiliki hasil prediksi terbaik dengan nilai CCC 0,78 (network 4-4-4) dan SLP -119,2 (network 2-2). Pelatihan dengan LSTM menghasilkan nilai CCC 0,71 (network 8-8) dan SLP –344,5 (network 4-4-4). Kedua variasi LSTM ini dibandingkan dengan *multi-layer perceptron* (MLP) yang menghasilkan nilai 0,63 CCC dan -356,4 SLP dengan proses gaussian, serta nilai 0,43 CCC dan -530,9 dengan tanpa proses gaussian. Penelitian yang telah dilakukan menunjukkan bahwa LSTM (tanpa atau dengan gaussian) memiliki performa yang mengungguli MLP dalam kasus CER. Hasil ini membuktikan bahwa LSTM dapat dengan baik digunakan dalam mendukung pengembangan *learning model* untuk ASR.

Beberapa hasil penelitian di atas digunakan sebagai dasar untuk pemilhan metode LSTM yang akan diimplementasikan pada kasus *speech recognition*. Hasil penelitian yang telah dipaparkan menunjukkan bahwa LSTM dengan berbagai variasi modelnya memiliki performa dan tingkat akurasi yang baik untuk digunakan pada data *sequential* yang kompleks. Dengan demikian, metode ini dapat dengan tepat digunakan untuk diterapkan untuk pengembangan sistem pengenalan ucapan (*speech recognition*).

Tabel 2. 1 Penelitian Terdahulu

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **No** | **Judul** | **Penulis** | **Metode** | **Hasil/ Kesimpulan** |
| 1 | *A Comparison of LSTM and GRU networks for learning symbolic sequences* | Roberto Cahuantzi, Xinye Chen, Steffan Güttel | LSTM dan GRU | LSTM memiliki performa yang jauh mengungguli GRU untuk kompleksitas data yang tinggi dengan perbandingan hasil waktu latih 12,53 detik dan 22,84 detik. Hasil juga menunjukkan prediksi akurasi dengan metrik jarak DL dan JW yang dihasilkan LSTM tidak jauh beda dengan GRU untuk semua kasus. |
| 2 | *Long Short-Term Memory Recurrent Neural Network for Automatic Speech Recognition* | Jane Oruh, Serestina Viriri, Adekanmi Adegun | LSTM-RNN | Perbandingan hasil dilakukan untuk beberapa model *deep learning* dan *sequential model* terhadap LSTM-RNN. Akurasi LSTM-RNN jauh unggul dibanding model *deep learning* dan *sequential model* lainmencapai 99,36%. |
| 3 | *Using Gaussian Processes with LSTM neural Network to predict continuous-time,dimensional emotion in ambiguous speech* | Mia Actheson,  Julien Epps, Vidhayasaharan | LSTM, gaussian LSTM, MLP, gaussian MLP | Penelitian ini membandingkan 2 metode neural network, yaitu MLP dan LSTM dengan perbedaan penggunaan proses gaussian (GP) terhadap masing-masing metode. Performa model tersebut menghasilkan *condordance correlation coefficient* (CCC) mencapai 0,78 untuk LSTM-GP, 0,71 untuk LSTM, 0,63 untuk MLP-GP, dan 0,43 untuk MLP |
| 4 | Penerapan Speech Recognition Menggunakan Metode *Long Short-Term Memory* (LSTM) untuk Presentasi Dinamis | Satriya Adhitama | LSTM |  |
| 5 |  |  |  |  |

## Landasan Teori

### Automatic Speech Recognition

*Automatic speech recognition* (ASR) merupakan salah satu tugasn dari *natural language processing* (NLP) yang melakukan transkripsi komputasi *real-time* dari bahasa yang diucapkan. ASR menjadi studi studi *human-computer interface* sejak 1950 dan berkembang pesat seiring waktu. Perkembangan dari ASR memberikan kemampuan untuk sistem dapat mengenali ucapan dengan kondisi yang beragam seperti adanya *noise* (kebisingan), aksen, diksi, dan nada (Kamath dkk., 2019).

ASR adalah metode untuk dapat mengubah ucapan kata menjadi teks dengan menggunakan komputer yang berbasis perangkat lunak. Sistem dirancang dengan teknik tertentu yang bertujuan untuk mengenali dan memproses suara manusia (Sen dkk., 2019). Prinsip dasar dari ASR yaitu seseorang berbicara mengeluarkan variasi tekanan suara pada *larynx* (pangkal tenggorokan), kemudian suara yang dihasilkan akan digitalisasi menggunakan *microphone* dan dikirimkan melalui sebuah perantara atau jaringan (Fendji dkk., 2022).

Mesin pengenal suara menerima *input* suara dalam bentuk yang telah didigitalisasi. Kemudian, *input* ini akan diubah menjadi *acoustic units* (*acoustic vectors*) melalui *acoustic model* (AM). Mesin pengenal menganalisis urutan (*sequence*) dari *acoustic vectors* dengan membandingkan kesesuaiannya pada data dan *acoustic model* untuk dapat dicari kandidat urutan yang paling mendekati. Sistem ASR mempu mengubah *input* suara dalam bentuk model matematis yang kemudian dapat dibuat menjadi teks untuk dapat mengenali ucapan.

Diagram

Description automatically generated

Gambar 2. 1 Arsitektur ASR

Gambar di atas merupakan arsitektur dari *automatic speech recognition* (ASR). Secara umum, sistem ASR memiliki arsitektur yang sama, meskipun kosakata yang digunakan terbatas, menengah, banyak, atau sangat banyak (Fendji dkk., 2022). ASR memiliki 5 (lima) komponen utama, yaitu *feature extraction*, *acoustic model* (model suara), *language model* (model bahasa), *pronounciation model* (model pengucapan), dan *decoder*.

### Long Short-Term Memory (LSTM)

*Long short-term memory* (LSTM) merupakan pengembangan metode dari *recurrent neural network* (RNN) yang dirancang dengan *memory cell* yang mampu merepresentasikan dependensi jangka panjang (*long-term dependency*) terhadap urutan waktu yang terjadi pada data (Sagheer & Kotb, 2019). Penanganan *vanishing gradient problem* pada RNN dapat diatasi menggunakan LSTM. LSTM secara khusus dirancang untuk menghindari masalah dependensi jangka panjang.

LSTM memiliki tiga gates atau gerbang yang masing-masing memiliki peran untuk melindungi dan mengontrol cell state. Cell state merupakan garis horizontal yang melewati bagian atas diagram sel LSTM yang memiliki kemampuan untuk menghapus atau menambahkan informasi baru yang masuk dalam waktu t dengan memanfaatkan struktur cermat yang disebut gerbang. Gates atau gerbang sendiri adalah sebuah cara yang digunakan oleh LSTM untuk melakukan seleksi terhadap informasi yang 10 masuk ke dalam sel

A diagram of a cell

Description automatically generated with low confidence

Gambar 2. Arsitektur LSTM

Forget gate merupakan gerbang pertama yang dioperasikan dalam sel LSTM. Forget gate ini menentukan informasi mana yang harus dipertahankan dan yang harus dibuang dari cell state. Gerbang ini menerima dua input, masing-masing dari ht-1 dan xt yang dimana ht-1 merupakan keluaran dari proses LSTM pada time step t-1 dan xt merupakan input pada time step t. Output dari gerbang ini adalah angka dengan rentang 0 hingga 1.

Setelah didapatkan nilai dengan rentang 0 hingga 1 pada forget gate, maka operasi selanjutnya adalah pada input gate. Input gate terdiri dari dua bagian, bagian pertama menggunakan fungsi sigmoid yang menentukan informasi mana yang ingin di-update. Sedangkan bagian kedua menggunakan fungsi tanh yang berfungsi untuk menentukan vektor yang akan ditambahkan pada nilai cell state (C’t). Kedua bagian tersebut berfungsi untuk menentukan informasi baru apa saja yang ingin disimpan pada cell state.

Hasil yang didapatkan dari operasi pada forget gate dan input gate selanjutnya dioperasikan agar hasil dari operasi tersebut dapat digunakan untuk meng-update cell state Ct-1 .Pada operasi ini dilakukan perkalian pada state sel Ct-1 dengan hasil dari layer forget gate (ft). Lalu tambahkan hasil perkalian tersebut dengan it \* Ct. Setelah dua operasi tersebut (operasi perkalian Ct-1 dan ft lalu penambahan dengan it \* C’t ), langkah selanjutnya adalah dengan mengoperasikan output gate. Pada output gate, layer sigmoid akan dioperasikan terlebih dahulu untuk menentukan bagian dari cell state apa saja yang akan digunkan sebagai hasil keluaran. Setelah operasi sigmoid dilakukan, selanjutnya operasi tanh dijalankan pada cell state lalu dikalikan dengan keluaran dari layer sigmoid yang telah dihitung sebelumnya.

# METODE PENELITIAN

## Bahan/Data

### Data yang diperoleh

Sistem pengenalan ucapan yang akan dirancang membutuhkan data yang dapat digunakan sebagai perintah suara (*voice command*) untuk mengoperasikan *software* presentasi. Perintah dilakukan dengan mengucapkan kata-kata tertentu yang kemudian akan dikonversikan dalam bentuk teks, sehingga dapat dieksekusi sesuai dengan apa yang diucapkan pembicara. Ucapan dari pembicara akan dideteksi dengan model *machine learning* yang telah dibangun, kemudian hasilnya akan dicocokkan dengan daftar kata yang ada di dalam basis data.

Data yang diperoleh merupakan *dataset* yang berisi kumpulan *file* suara dengan format WAV. Masing-masing *file* memuat suara manusia untuk ucapan kata tertentu.. Kosa kata yang digunakan terbatas untuk memastikan bahwa proses menjadi ringan, namun tetap cukup beragam untuk pelatihan model pada data sehingga dapat berguna untuk berbagai pengaplikasian (Warden, 2018). Tujuan utama digunakannya dataset ini untuk menghasilkan model pelatihan mesin yang terbaik. Dataset memiliki 34 variasi kata berbahasa Inggris dengan jumlah 103.807 ucapan. Berikut adalah datanya:

Tabel 3. 1 Dataset *speech\_commands\_v0.02*

|  |  |  |
| --- | --- | --- |
| **No.** | **Kata** | **Jumlah Ucapan** |
| 1 | *Backward* | 1.664 |
| 2 | *Bed* | 2.014 |
| 3 | *Bird* | 2.064 |
| 4 | *Dog* | 2.128 |
| 5 | *Down* | 3.917 |
| 6 | *Eight* | 3.787 |
| 7 | *Five* | 4.052 |
| 8 | *Follow* | 1.579 |
| 9 | *Forward* | 1.557 |
| 10 | *Four* | 3.728 |
| 11 | *Go* | 3.880 |
| 12 | *Happy* | 2.054 |
| 13 | *House* | 2.113 |
| 14 | *Learn* | 1.575 |
| 15 | *Left* | 3.801 |
| 16 | *Marvin* | 2.100 |
| 17 | *Nine* | 3.934 |
| 18 | *No* | 3.941 |
| 19 | *Off* | 3.754 |
| 20 | *On* | 3.845 |
| 21 | *One* | 3.890 |
| 22 | *Right* | 3.778 |
| 23 | *Seven* | 3.998 |
| 24 | *Sheila* | 2.022 |
| 25 | *Six* | 3.860 |
| 26 | *Stop* | 3.872 |
| 27 | *Three* | 3.727 |
| 28 | *Tree* | 1.759 |
| 29 | *Two* | 3.880 |
| 30 | *Up* | 3.723 |
| 31 | *Visual* | 1.592 |
| 32 | *Wow* | 2.123 |
| 33 | *Yes* | 4.044 |
| 34 | *Zero* | 4.052 |

Pada unduhan dataset termuat *file* teks yang dinamakan *validation\_list.txt*, yang berisi kumpulan *file* untuk memvalidasi hasil selama pelatihan berlangsung, dan membantu menyesuaikan hyperparameters. *File testing\_list.txt* memuat nama dari klip suara yang digunakan untuk mengukur model yang dilatih. Data suara dan teks tersbut akan digunakan lebih lanjut untuk pelatihan model *machine learning* menggunakan LSTM.

### Prosedur pengumpulan data

Data yang dikumpulkan oleh Pete Warden ini dapat diunduh secara *online* karena sifatnya yang *open source*. Penulis mengunduh *speech\_command\_v0.02* pada <https://www.kaggle.com/datasets/mok0na/speech-commands-v002>. Data ucapan ini diambil melalui *microphone* telepon atau laptop yag dimiliki pengguna (Warden, 2018). Data yang dikumpulkan berupa ucapan berbahasa Inggris untuk membatasi cakupan pengumupulan data. Para subjek diminta untuk merekam suara pada ruangan tertutup sendirian dengan pintu tertutup untuk menghindari percakapan lain yang bersifat privasi. Subjek ini terdiri dari sejumlah orang yang berbeda-beda sehingga model dapat melakukan pelatihan dengan baik.

Pete juga mengumpulkan data dengan menghindari informasi yang dapat diidentifikasi dari kontributor dengan alasan privasi. Hal tersebut berarti *dataset* tidak termuat fitur seperti jenis kelamin, etnis, maupun ID pengguna. Data juga dibatasi untuk panjang durasi ucapan sekitar 1 (satu) detik. Data ucapan disimpan dalam satu *file* berbeda-beda, sesuai dengan kata yang telah ditentukan. Dengan demikian, tidak tersedia data yang berupa kalimat utuh agar lebih mudah untuk melakukan klasifikasi.

## Aturan Bisnis

# ANALISIS DAN PERANCANGAN SISTEM

## Analisis Sistem yang Berjalan

Penelitian ini menghasilkan sistem yang mampu mengenali kata tertentu pada ucapan yang panjang (*keyword spotting*). Dari dataset yang diperoleh, peneliti menetapkan batasan pemilihan kata yang akan digunakan untuk pembuatan model machine learning. Pembatasan kata ini bertujuan untuk memudahkan dan mempercepat proses pembuatan model *machine learning*, sehingga didapatkan hasil yang diinignkan. Sistem ini akan memanfaatkan 4 kata untuk digunakan sebagai perintah sederhana dalam pengoperasian *software* presentasi. Adapun kata-kata yang dipilih adalah *backward, forward, go* dan*, stop*.

Tabel 4. 1 Daftar kata perintah sistem

|  |  |  |
| --- | --- | --- |
| **No.** | **Kata** | **Jumlah Ucapan** |
| 1 | *Backward* | 1.664 |
| 2 | *Forward* | 1.557 |
| 3 | *Go* | 3.880 |
| 4 | *Stop* | 3.872 |

Sistem mampu menerima input berupa suara ucapan, kemudian akan diubah menjadi data digital berupa kumpulan angka. Dari input suara tersebut, sistem akan melakukan pemotongan (*slicing*) data untuk setiap 1 detik dengan *sample rate* 16000, sehingga dapat sesuai dengan data suara yang tersedia. *Slicing* dilakukan untuk dapat mendeteksi kata apa yang muncul setiap detiknya.

Daftar kata tersebut digunakan sebagai perintah suara (*voice command*) untuk mengoperasikan beberapa fungsionalitas dari *software* presentasi. Berikut adalah beberapa fungsi dari kata perintah yang telah ditentukan:

Tabel 4. 2 Fungsi kata perintah sistem

|  |  |  |
| --- | --- | --- |
| **No.** | **Kata** | **Perintah** |
| 1 | *Backward* | Kembali ke halaman (*slide*) sebelumnya |
| 2 | *Forward* | Maju ke halaman (*slide*) berikutnya |
| 3 | *Go* | Memainkan media player |
| 4 | *Stop* | Menghentikan media player |

## Analisis Sistem yang Diusulkan

## Desain Sistem

# Penutup

## Simpulan

## Saran

# DAFTAR PUSTAKA

Atcheson, M., Sethu, V., & Epps, J. (2019). Using Gaussian Processes with LSTM Neural Networks to Predict Continuous-Time, Dimensional Emotion in Ambiguous Speech. *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, 718–724. https://doi.org/10.1109/ACII.2019.8925450

Cahuantzi, R., Chen, X., & Güttel, S. (2021). *A comparison of LSTM and GRU networks for learning symbolic sequences*.

Fendji, J. L. K. E., Tala, D. C. M., Yenke, B. O., & Atemkeng, M. (2022). Automatic Speech Recognition Using Limited Vocabulary: A Survey. *Applied Artificial Intelligence*, *36*(1). https://doi.org/10.1080/08839514.2022.2095039

Kamath, U., Liu, J., & Whitaker, J. (2019). *Deep Learning for NLP and Speech Recognition*. Springer International Publishing. https://doi.org/10.1007/978-3-030-14596-5

Oruh, J., Viriri, S., & Adegun, A. (2022). Long Short-Term Memory Recurrent Neural Network for Automatic Speech Recognition. *IEEE Access*, *10*, 30069–30079. https://doi.org/10.1109/ACCESS.2022.3159339

Sagheer, A., & Kotb, M. (2019). Time series forecasting of petroleum production using deep LSTM recurrent networks. *Neurocomputing*, *323*, 203–213. https://doi.org/10.1016/j.neucom.2018.09.082

Sen, S., Dutta, A., & Dey, N. (2019). *Audio Processing and Speech Recognition*. Springer Singapore. https://doi.org/10.1007/978-981-13-6098-5

Warden, P. (2018). *Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition*.