SPEECH TO TEXT MENGGUNAKAN METODE *HIDDEN MARKOV MODEL*

SPEECH TO TEXT USING HIDDEN MARKOV MODEL

Muhammad Fariz Taswarul Afkar[1], Budhi Irawan[2], Surya Michrandi Nasution[3]
[1,3]Prodi S1 Teknik Komputer, Fakultas Teknik, Universitas Telkom
[1]**farizafkar@student.telkomuniversity.co.id,**
[2]**budhiirawan.staff.telkomuniversity.ac.id,** [3]**michrandi.staff.telkomuniversity.ac.id**

Abstrak

Aplikasi Speech to Text (STT) ini menggunakan metode Hidden Markov Models (HMM) Hybird dengan Gaussian Mixture Model (GMM). Tahap awal dari Hidden Markov Models adalah ketika ada suara, maka suara tersebut akan dikenali sebagai Speech Signal. Kemudian menggunakan Feature extraction yaitu Mel-frequency cepstral coefficients (MFCC) signal tersebut disimpan ke dalam frame-frame dan dicari nilai koefisien cepstral-nya. Selanjutnya tiap vector di kuantisasi yang menghasilkan output simbol observasi (codebook). Setiap kata yang tidak dikenal maka akan dimodelkan dengan HMM/GMM sehingga mendapatkan model kata. Untuk proses pengenalan kata maka akan dihitung probabilitas kemiripan pola dari tiap model HMM/GMM yang dimiliki dengan hasil dari observasi. Hasil probabilitas paling maksimum kemudian ditetapkan sebagai kata yang di kenali. Pengujian ini dilakukan dengan mengubah nilai feature MFCC dan nilai mixture GMM. Performansi sistem diukur berdasarkan akurasi yang didapat dari parameter WER(Word Error Rate). Setelah dilakukan pengujian terhadap sistem dengan beberapa skenario, diperoleh akurasi terbaik 100% dalam mengenali 10 kata. Akurasi ini deperoleh dari hasil pengujian dengan MFCC 13 Feature dan GMM 6 mixture.

Kata kunci : ***Speech to Text, Hidden Markov Model, Mel-frequency cepstral coefficient, Gaussian Mixture Model***

Abstract

This Speech to Text (STT) application uses the Hidden Markov Models (HMM) Hybird method with the Gaussian Mixture Model (GMM). The initial stage of Hidden Markov Models is when there is a sound, then the sound will be recognized as Speech Signal. Then using Feature extraction that is Mel-frequency cepstral coefficients (MFCC) the signal is stored in the frames and the cepstral coefficient value is searched. Furthermore, each vector is quantized to produce an observation symbol output (codebook). Every unknown word will be modeled with HMM / GMM so that it gets a word model. For the word recognition process, the probability of likelihood of each HMM / GMM model will be calculated with the results of the observation. The maximum probability results are then set as words that are recognized. This test is done by changing the value of MFCC features and GMM mixture values. System performance is measured based on the accuracy obtained from the WER (Word Error Rate) parameter. After testing the system with several scenarios, the best accuracy is 100% in recognizing 10 words. This accuracy was obtained from the test results with MFCC 13 Feature and GMM 6 mixture.

Keywords: ***Speech to Text, Hidden Markov Model, Mel-frequency cepstral coefficient, Gaussian Mixture Model***

1. Pendahuluan

Istilah era globalisasi terdiri dari dua kata yaitu era yang berati masa atau zaman dan globalisasi yang berati proses menggelobal. Salah satu efek globalisasi adalah adanya bahasa internasional yaitu bahasa inggris. Dengan ada bahasa internasional ini, menjadikan tuntunkan untuk menjadi Sumber Daya Manusia (SDM) yang lebih baik itu meningkat. Untuk menjadi SDM yang mampu bersaing di era globalisasi salah satu caranya adalah memperkenalkan pada generasi muda untuk belajar bahasa inggris sejak dini. Tidak hanya tentang teori tapi lebih tepat nya untuk berbicara menggunakan bahasa inggris. Pada acara launching program ‘English for Indonesia' di

Kedubes Inggris, Jakarta, rabu (3/10/2018). Menteri Ketenagakerjaan, M. Hanif Dhakiri mengungkapkan bahwa," Peningkatan kemampuan berbahasa asing dibutuhkan untuk meningkatkan kualitas pekerjaan. Apalagi saat ini dunia menjadi semacam desa global sehingga SDM Indonesia juga harus menguasai bahasa asing." [1].

Speech to Text adalah bagian dari *Speech Recognition* yaitu bidang Ilmu Komputer dan Elektronik yang berurusan dengan Sinyal dan Sistem, Pengolahan Sinyal, Peningkatan Sinyal, dan lain-lain [2]. Dalam proses *Speech to Text* terdapat dua bagian besar yaitu proses ektrasi ciri dan proses pemodelan/pengenalan. Proses ekstrasi ciri atau yang dikenal dengan *feature extraction* adalah proses merubah sinya suara menjadi sebuah set *vectors*. Tujuan dari proses ini adalah untuk mendapatkan sebuah representasi baru yang lebih *compact*, *less redundant*, dan lebih cocok untuk statistik [3]. Meskipun banyak *feature extraction* yang diusulkan dalam literatur, *feature extraction* yang paling populer adalah *Mel Frequency Cepstral Coefficents* (*MFCC*), karena *MFCC* dapat merepresentasikan konsep dari pendengaran dan persepsi mannusia [4]. Proses pemodelan/pengenalan adalah proses dimana sinyal suara yang sudah di ekstrak fitur cirinya yang menghasilkan sebuah set *vectos*. Kumpulan dari set *vectors* tersebut nantinya akan diklasifikasi dan di modelkan dengan metode *Hidden Markov Model (HMM).*

HMM merupakan model statistik dari sebuah sistem yang diumpamakan sebagai sebuah "Proses *Markov*" dengan parameter yang tidak terdeteksi, dapat menentukan parameter yang tidak terdeteksi tersebut melalui parameter yang terdeteksi. *HMM* sudah banyak diterapkan di bidang *signal processing*, dan *speech processing* [5].

Pada tugas akhir ini penulis akan mengimplementasikan metode *MFCC* untuk *feature extraction* dan metode *HMM* untuk pemodelan/pengenalan kata dalam aplikasi *Speech to Text* (*STT*). Dengan adanya aplikasi *STT* ini, diharapkan pengguna dapat mengenal lebih banyak kosa kata bahasa inggris dan bagaimana cara pelafalannya. Sehingga dapat memperkuat fondasi untuk berani memulai pembicaraan dan melatih pengucapan berbahasa inggris.

2. Dasar Teori

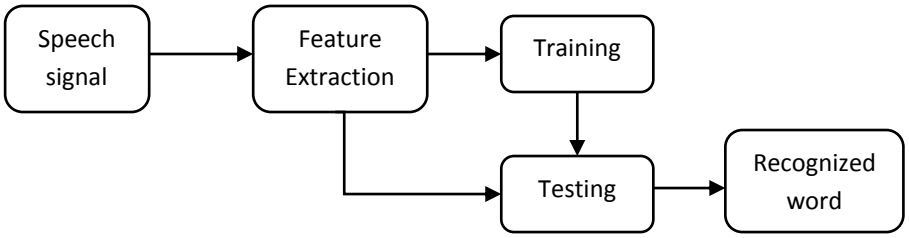
2.1 Speech Recognition

Speech Recognition adalah bidang ilmu interdisipliner linguistik komputasi yang mengembangkan metodologi dan teknologi yang memungkinkan pengenalan dan terjemahan bahasa lisan ke dalam teks dengan komputer dan dikenal juga sebagai *Automatic Speech Recognition* (*ASR*) atau *Speech to Text* (*STT*).

a) Algoritma untuk *Speech to Text* adalah sebagai berikut:
   1. Sinyal suara (*Speech Signal*) masuk melalui mikrofon.
   2. Fitur ekstrak ciri (*Feature extraction*) dari sinyal suara.
   3. *Training* dataset dari *Feature extraction.*
   4. *Testing* dataset dari *Feature extraction* .
   5. *Output* berupa teks dari sinyal yang dikenali.

b) Komponen penyusun
   1. Sinyal suara.
   2. Dataset *training.*
   3. Dataset *testing.*

c) Diagram sistem

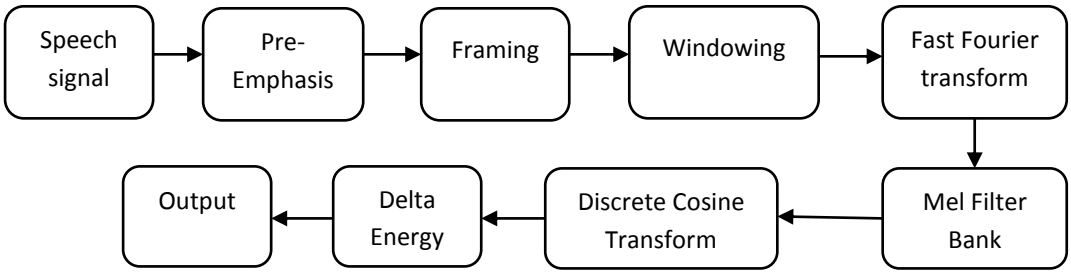


Gambar 1. Gambar konsep sistem *Speech Recognition* [6]

d) Cara kerja
   1. Fitur ekstrak ciri dari sinyal suara yang masuk menggunakan *MFCC.*
   2. *Training* adalah proses *clustering* dan pembuatan *codebook* dari sinyal *sample* menggunakan *K-means Clustering.*
   3. Testing adalah fitur ekstrak ciri dari sinyal test dan menghitung semua *likelihood* ke sinyal *sample* yang ada pada *codebook*.
   4. Didapat semua nilai likelihood, nilai yang paling mendekati ke probabilitas sinyal sample dijadikan keluaran berupa teks.

2.2 Mel Frequency Cepstral Coefficents (MFCC)

Bagian yang penting dari *Speech Recognition* adalah ektrasi ciri (feature extraction). Kegunaan dari ekstrasi ciri adalah mengurangi informasi-informasi yang tidak dibutuhkan dari sensor dan mengkonversi informasi-informasi yang penting dari signal untuk pengenalan pola agar menghasilkan format yang lebih sederhana dengan kelas-kelas yang jelas. MFCC merupakan cara yang paling sering digunakan, karena cara kerjanya didasarkan pada perbedaan frekuensi yang dapat ditangkap oleh telinga manusia [7].
MFCC memiliki beberapa tahap untu ekstrasi suara, berikut diagramnya :



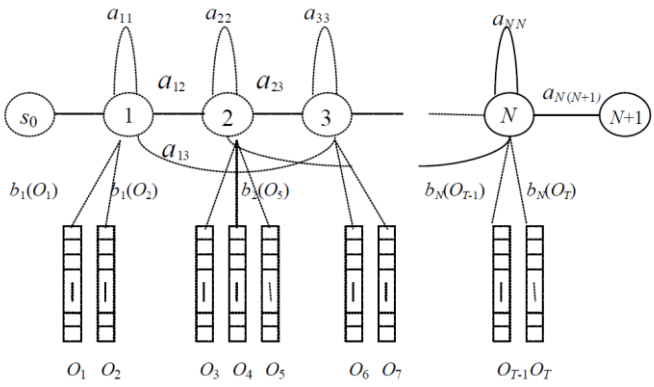
Gambar 2. blok diagram dari MFCC

2.3 Hidden Markov Model (HMM)

Hidden Markov model (HMM) adalah model statistik di mana sistem yang dimodelkan diasumsikan sebagai proses Markov dengan parameter yang tidak diketahui, dan tantangannya adalah menentukan parameter tersembunyi dari parameter yang dapat diamati. Parameter model yang diekstraksi kemudian dapat digunakan untuk melakukan analisis lebih lanjut, contohnya untuk aplikasi pengenalan pola / ucapan. Dalam *HMM*, *state* tidak langsung terlihat, tetapi variabel yang dipengaruhi oleh *state* terlihat. Setiap *state* bagian memiliki distribusi probabilitas terhadap kemungkinan keluaran token. Transisi *state* juga bersifat probabilistik. Oleh karena itu, urutan token yang dihasilkan oleh *HMM* memberikan beberapa informasi tentang urutan keadaan. Model *HMM* lengkap dilambangkan sebagai $\lambda = (A, B, \pi)$.

Parameter Model *HMM*

1. A, distribusi probabilitas keadaan, A = {aij}.
2. B, densitas probabilitas simbol pengamatan, B = {bj (k)}.
3. π, distribusi keadaan awal, π = {πi}.

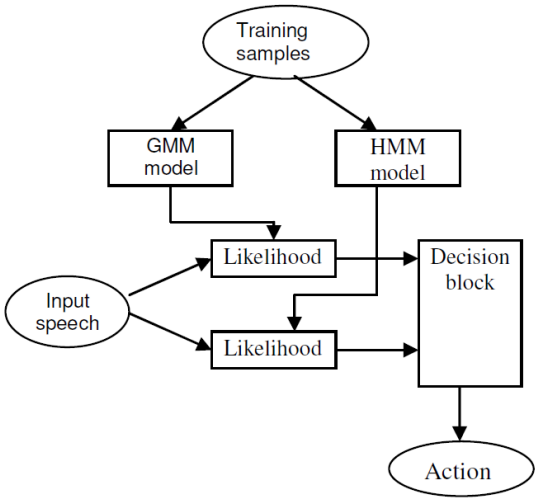


Gambar 3. Arsitektur dari HMM [10].

Prosedur pelatihan *HMM* membutuhkan *codebook* untuk memperkirakan parameter model. Dalam *codebook*, sejumlah besar vektor pengamatan dari data pelatihan dikelompokkan ke dalam *cluste*r vektor pengamatan M menggunakan *K-means* prosedur berulang. Berdasarkan pada vektor-vektor pengamatan yang dikelompokkan ini, perkiraan parameter model adalah dihasilkan selama pelatihan *HMM*. Prosedur pelatihan *HMM* mencoba memperkirakan nilai probabilitas state distribusi (A), densitas probabilitas simbol pengamatan (B), dan distribusi keadaan awal (π). Observasional vektor urutan pelatihan disegmentasi untuk masing - masing state N, perkiraan kemungkinan maksimum dari himpunan pengamatan yang terjadi dalam setiap state j masing-masing vektor pengamatan dalam suatu state diberi kode menggunakan M - kode - kata codebook. Sebelum estimasi yang tepat dari parameter model yang diinisialisasi ke baik perkiraan awal yang penting untuk konvergensi yang cepat dan tepat dari rumus estimasi ulang. Itu parameter diestimasikan ulang menggunakan Algoritma *Viterbi*, Algoritma *Forward-Backward* dan Algoritma *Baum / Welch* [9].

2.4 Gaussian Mixture Model/Hidden Markov Model (HMM/GMM)

GMM dapat dilihat sebagai model *hybrid* antara model kepadatan parametrik dan non-parametrik seperti yang ditunjukkan pada gambar di bawah ini



Gambar 4 Arsitektur dari *Gaussian Mixture Model*

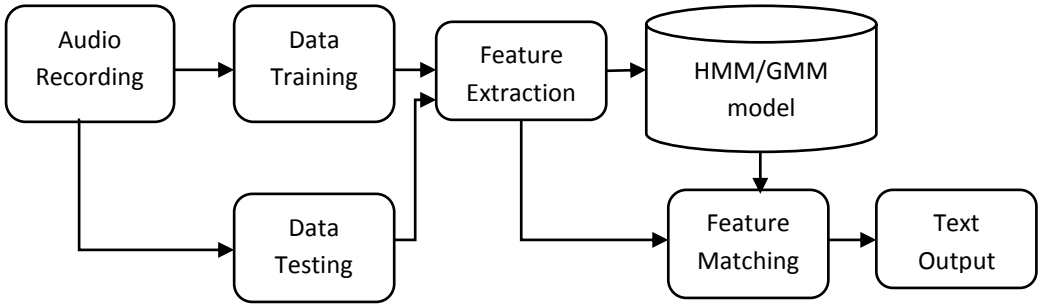
. Seperti model parametrik, ia memiliki struktur dan parameter yang mengontrol perilaku kepadatan diketahui cara. Seperti model non-parametrik ia memiliki banyak derajat kebebasan untuk memungkinkan pemodelan kepadatan yang sewenang-wenang.

Model *hybrid HMM / GMM* memiliki kemampuan untuk menemukan probabilitas maksimum gabungan di antara semua referensi yang mungkin kata W diberi urutan observasi O. Dalam kasus nyata, kombinasi *GMM* dan *HMM* dengan a koefisien tertimbang mungkin merupakan skema yang baik karena perbedaan dalam metode pelatihan [11].

3. Perancangan

3.1 Desain Sistem

Berikut adalah gambaran umum dari perancangan sistem Speech to Text.



Gambar 5. Gambaran umum sistem *Speech to Text*

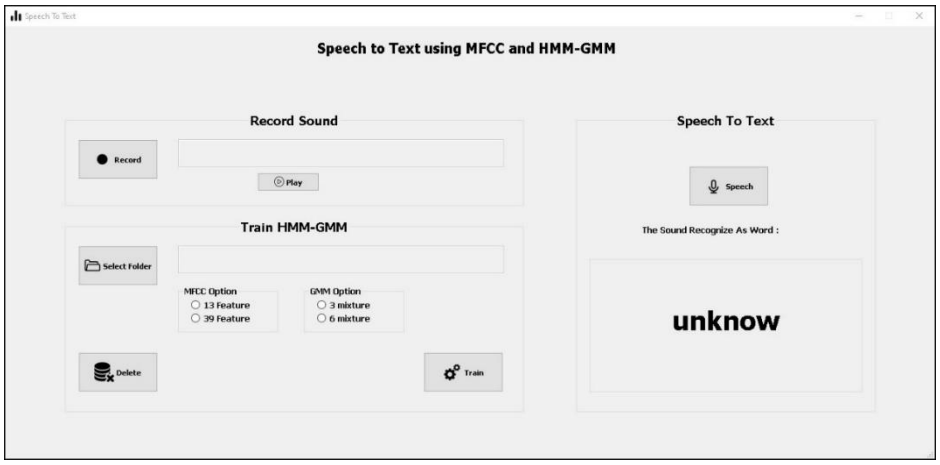
Berikut fungsi dari aplikasi *Speech to Text ini*:

1. Dapat melakukan *Training* yaitu dataset suara yang sudah ada di modelkan dengan metode *HMM/GMM* .
2. Dapat meyimpan dataset suara yang telah di *Training*, sebagai model *HMM/GMM*.
3. Dapat melakukan *Testing* dataset suara yang dikenali berdasarkan model *HMM/GMM* yang telah dibuat di proses *Training.*

Berikut beberapa fitur yang ada di aplikasi *Speech to Text ini*:

1. Fitur *record*, dimana pengguna dapat merecord suara untuk nantinya di lakukan proses *Training*.
2. Fitur *play*, dimana pengguna dapat mendengar hasil *record* suara yang sudah terekam.
3. Fitur *select folder*, dimana pengguna memilih *root folder* yang akan dijadikan data *Training.*
4. Fitur *MFCC Option*, dimana pengguna memilih feature dari *MFCC* yang akan digunakan 13 *Feature* atau 39 *Feature.*
5. Fitur *GMM Option*, dimana pengguna memilih mixture dari *GMM* yang akan digunakan 3 *mixture* atau 6 *mixture.*
6. Fitur *train*, dimana pengguna dapat memproses data training nya untuk di jadikan sebagai model *HMM/GMM.*
7. Fitur *Delete Model*, dimana pengguna dapat menghapus data model *HMM/GMM* yang telah ada.
8. Fitur *speech*, dimana pengguna dapat mencoba fitur Speech to Text secara langsung.

3.2 User Interface

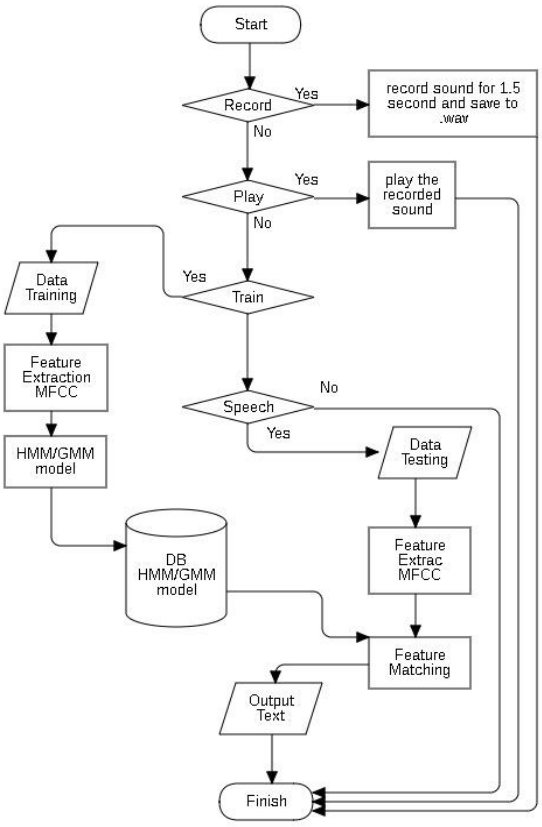


Berikut adalah rangcangan antar muka program *Speech to Text.*

Gambar 6. Gambar *User Interface*

3.3 Flow Chart

Berikut adalah *Flow Chart* dari aplikasi *Speech to Text*



Gambar 7 Gambar *Flow Chart*

4. Kesimpulan

4.1 Hasil Percobaan

Dataset *training* yang digunakan dalam pengerjaan tugas akhir ini adalah 17 data *audio/*suara, terdiri dari:

Tabel 1. Dataset *training*

No.	Nama berkas	Jumlah file tiap berkas	No	Nama berkas	Jumlah file tiap berkas
1	Apple	30	10	Orange	30
2	Banana	30	11	Peach	30
3	Eight	30	12	Pineapple	30
4	Five	30	13	Seven	30
5	Four	30	14	Six	30
6	Kiwi	30	15	Three	30
7	Lime	30	16	Two	30
8	Nine	30	17	Zero	30
9	One	30			
Jumlah		270	**Jumlah**		240
TOTAL					510

Dataset *testing* yang digunakan dalam pengejaan tugas akhir ini adalah 10 dataset suara yang terdiri dari:

Tabel 2.Dataset *testing*

No.	Nama berkas	Jumlah file tiap berkas	No	Nama berkas	Jumlah file tiap berkas
1	Apple	4	10	Kiwi	4
2	Banana	4	11	Lime	4
3	Eight	4	12	Nine	4
4	Five	4	13	One	4
5	Four	4	14	Orange	4

- Pengujian parameter A
  Dengan menggunakan *MFCC 13 Feature dan GMM 3 Mixture* didapatkan hasil:
  1. Kapasitas file yang dihasilkan sebesar 100 *Kb*.
  2. Waktu yang diperlukan untuk proses training 17 dataset suara 15,39 detik.
  3. Waktu yang diperlukan untuk proses testing 10 dataset suara rata-rata 0.33 detik dengan menggunakan mikrofon dan 0.30 detik dengan menggunakan *file*.
  4. Mengenali 9 dataset suara dari 10 dataset suara yang ditesting.

- Pengujian parameter B
  Dengan menggunakan *MFCC 39 Feature dan GMM 3 Mixture* didapatkan hasil:
  1. Kapasitas file yang dihasilkan sebesar 181 *Kb*.
  2. Waktu yang diperlukan untuk proses training 17 dataset suara 22,06 detik.
  3. Waktu yang diperlukan untuk proses testing 10 dataset suara rata-rata 0.47 detik dengan menggunakan mikrofon dan 0.34 detik dengan menggunakan *file*.
  4. Dari total 10 dataset yang ditesting, dikenali 9 dataset suara.

- Pengujian parameter C
  Dengan menggunakan *MFCC 13 Feature dan GMM 3 Mixture* didapatkan hasil:
  1. Kapasitas file yang dihasilkan sebesar 256*Kb*.
  2. Waktu yang diperlukan untuk proses training 17 dataset suara 19,08 detik.

3. Waktu yang diperlukan untuk proses testing 10 dataset suara rata-rata 0.41 detik dengan menggunakan mikrofon dan 0.30 detik dengan menggunakan *file*.
4. Dari total 10 dataset yang ditesting, dikenali 10 dataset suara.

- Pengujian parameter D

  Dengan menggunakan *MFCC 13 Feature dan GMM 3 Mixture* didapatkan hasil:
  1. Kapasitas file yang dihasilkan sebesar 492 *Kb*.
  2. Waktu yang diperlukan untuk proses training 17 dataset suara 24,66 detik.
  3. Waktu yang diperlukan untuk proses testing 10 dataset suara rata-rata 0.44 detik dengan menggunakan mikrofon dan 0.34 detik dengan menggunakan *file*.
  4. Dari total 10 dataset yang ditesting hanya dikenali 9 dataset suara.

4.2 Kesimpulan

Berdasrkan hasil dari 4 percobaan skenario di atas, kesimpulan yang dapat di ambil antara lain:

1. Akurasi yang didapat dari aplikasi Speech to Text yang di rancang dengan menggunakan *metode Feature extraction MFCC* dan *Hybird HMM/GMM* adalah 100% untuk pengenalan dengan 10 dataset suara.
2. Dengan Menggunakan *feature MFCC* sebesar 13 dan *mixture GMM* sebesar 6 didapatkan hasil yang bagus dalam performansi waktu, kapasitas *file* pemodelan yang kecil dan keakurasian yang tinggi.
3. Menggunakan metode *Hybird* akan memperbaiki tingkat performansi sistem. *GMM* mempermudah perhitungan probabilitas kemungkinan pemodelan akustik/fonem setiap *hidden state* pada *HMM* .

Daftar Pustaka:

[1] M. Idris, "news.detik.com," 03 Oktober 2018. [Online]. Available: https://news.detik.com/berita/d-4240335/menaker-ingatkan-pekerja-pentingnya-lancar-berbahasa-inggris. [Accessed 22 Juli 2019].

[2] N. S. Shirodkar, "KONKANI SPEECH TO TEXT RECOGNITION," researchgate.net, 2016.

[3] C. S. Yee and A. M. Ahmad, "Mel Frequency Cepstral Coefficients for Speaker Recognition Using Gaussian Mixture Model-Artificial Neural Network Model," University of Technology Malaysia.

[4] M. Ravanelli, "Deep Learning for Distant Speech Recognition," 2017.

[5] A. F. Aisyah and A. Noortjahja, "IMPLEMENTASI HIDDEN MARKOV MODELS (HMM) SEBAGAI FILTER UNTUK MEREDUKSINOISE PADA ESOPHAGEAL SPEECH," Jurnal Inovasi Fisika Indonesia, vol. 04, pp. 7-14, 2015.

[6] C. Ramaiah and V. S. Rao, "Speech samples recognition based on MFCC and vector Quantization," International Journal on Computer Science and Emerging Trends (IJCSET), vol. 01, no. 02, pp. 1-7, 2012.

[7] I. P. Permana and B. S. Negara, "Identifikasi Pembicara dengan Menggunakan Mel Frequency Cepstral Coefficient (MFCC) dan Self Organizing Map (SOM)," 2011.

[8] P. P. Singh and P. Rani, "An Approach to Extract Feature using MFCC," IOSR Journal of Engineering (IOSRJEN), vol. VI, no. 08, pp. 21-25, 2014.

[9] P. Bansal, A. Kant, S. Kumar, A. Sharda and S. Gupta, IMPROVED HYBRID MODEL OF HMM/GMM FOR SPEECH RECOGNITION, Varna, 2008.

[10] D. Dimov and I. Azmanov, "Experimental specifics of using HMM in isolated word speech recognition," 2005.

[11] I. M. M. El-emary, M. Fezari and H. Attoui, "Hidden Markov model/Gaussian mixture models (HMM/GMM) based voice command system: A way toimprove the control of remotely operated robot arm TR45," Scientific Research and Essays, vol. VI, pp. 341-350, 2011.

[12] B. A. Sonkamble and D. D. Doye, "Speech Recognition Using Vector Quantization through Modified K-meansLBG Algorithm," Computer Engineering and Intelligent Systems, vol. III, no. 7, pp. 137-145, 2012.