

LAPORAN TUGAS BESAR

Pembangunan Model Pengenalan Ucapan Otomatis

Diajukan untuk memenuhi tugas

Mata Kuliah IF4071 Pemrosesan Suara



Oleh :

Ilham Prasetyo Wibowo 13520013

Mahesa Lizardy 13520116

Muhammad Fahmi Irfan 13520

PROGRAM STUDI TEKNIK INFORMATIKA
SEKOLAH TEKNIK ELEKTRO DAN INFORMATIKA
INSTITUT TEKNOLOGI BANDUNG
2023

A. Pembangunan Model

1. Algoritma *Training*

Algoritma yang digunakan adalah HMM-GMM (Hidden Markov Model-Gaussian Mixture Model). Ini adalah pendekatan yang umum digunakan dalam pemrosesan bahasa alami dan pengenalan ucapan. CMUSphinx merupakan salah satu tools yang memanfaatkan pendekatan HMM-GMM untuk tujuan pengenalan ucapan (Speech Recognition).

2. *Tools dan Framework*

a. Model Akustik

Model akustik dibangun menggunakan CMUSphinx toolkit yang secara langsung mendukung model berbasis GMM-HMM. Toolkit spesifik yang dipakai adalah pocketsphinx yang mendukung pelatihan model ringan dan *low footprint* mengingat sumber daya komputasi yang terbatas. Dokumentasi CMUSphinx dapat dilihat pada tautan ini [CMUSphinx Open Source Speech Recognition](#).

b. Model Bahasa

Model bahasa merupakan model n-gram yang dibangun dengan *toolkit* KenLM. Dokumentasi KenLM dapat dilihat pada tautan ini [kenlm .code .Kenneth Heafield \(kheafield.com\)](#).

3. Proses *Training*

Proses training dilakukan dengan cara melatih model menggunakan dataset yang sudah dibangun pada tahap sebelumnya. Tahap pertama yang dilakukan yaitu Preprocessing dataset. Dataset audio akan dilakukan normalisasi, denoising, dan VAD(Voice Activity Detection).

1. Model Bahasa

Pembangunan model bahasa menggunakan data dari internet dengan total kalimat sekitar 10ribu kalimat. Model bahasa yang dibangun merupakan model n-gram yang disimpan dalam format ARPA untuk digabungkan dengan model akustik. *Base* model

yang dibangun adalah model 3-gram. Namun, format ARPA memakan *space* yang cukup besar. Karena pocketsphinx mendukung format bin, hasil pembangunan model bahasa dikonversi menjadi format bin untuk meningkatkan performa.

2. Model Akustik

Tahap selanjutnya yaitu membangun *Acoustic Model*. Pembangunan akustik model terdiri dari beberapa tahap, tahap pertama yaitu menyiapkan dataset yang sudah dilakukan pre proses sebelumnya dan melakukan transkripsi pada setiap dataset audio tersebut. Setelah itu dibangun *phonetic dictionary* yang berisi kata beserta phonetic transkripsi nya dan *phoneset file* yang berisi list *phone* yang terdapat pada *phonetic dictionary*. *Phonetic dictionary* dibuat berdasarkan ARPABET, yaitu salah satu kode transkripsi fonetik yang memiliki representasi alfanumerik dari sistem alfabet fonetik IPA (International Phonetics Alphabet), dengan sedikit penyesuaian terhadap aksan Bahasa Indonesia untuk bahasa-bahasa serapan. Audio yang digunakan menggunakan sample rate 44100. Dataset tersebut akan dilatih untuk membangun model acoustic pada CMUSphinx. CMUSphinx sendiri memiliki beberapa konfigurasi yang dapat diubah seperti Sample Rate, jumlah filter, dll.

Pada proses pelatihan, terdapat 5 tahapan yang dilakukan. Tahap pertama melibatkan ekstraksi fitur dari berkas audio. Sistem tidak langsung memproses sinyal akustik, melainkan sinyal tersebut diubah menjadi urutan vektor fitur yang digunakan sebagai representasi pengganti dari sinyal akustik sebenarnya.

Tahap kedua melibatkan perhitungan urutan vektor berdimensi 13 (vektor fitur) pada setiap percakapan pelatihan yang terdiri dari Koefisien Cepstral Frekuensi Mel (MFCC). MFCC tersebut akan disimpan dalam direktori bernama feat untuk proses selanjutnya

Tahap ketiga, akan dilatih model Context-Independent (CI) untuk unit-unit sub-kata dalam kamus yang telah dibangun sebelumnya. Setelah itu, dilatih model untuk unit sub-kata Context-Dependent (trifon) dengan status yang tidak terikat (CD-untied). Model CD-untied diperlukan untuk membangun pohon keputusan guna mengikat status.

Tahap keempat, Setelah pohon keputusan untuk setiap status dari setiap unit sub-kata dibentuk, akan dilakukan *prune tree* untuk memangkas pohon keputusan dan mengikat status.

Tahap terakhir, akan dilatih model akhir untuk trifen dalam korpus pelatihan Anda yang disebut model CD-tied. Model CD-tied dilatih dalam beberapa tahap. Dimulai dari 1 Gaussian per state HMM, dilanjutkan dengan pelatihan 2 Gaussian per state HMM, dan seterusnya hingga jumlah Gaussian yang diinginkan per State telah terlatih.

4. Proses *Decoding*

Decoding dilakukan dengan menggunakan model yang sudah dilatih sebelumnya. Dapat menggunakan secara *live speech* maupun menggunakan input audio. Dengan konfigurasi Sample rate 44100, mono, dan 16-bit encoding.

B. Hasil Eksperimen

Eksperimen pertama dilakukan dengan parameter-parameter berikut.

Sampling rate : 44100

Jumlah filter : 40

Filter (High) : 6800

Filter (Low) : 130

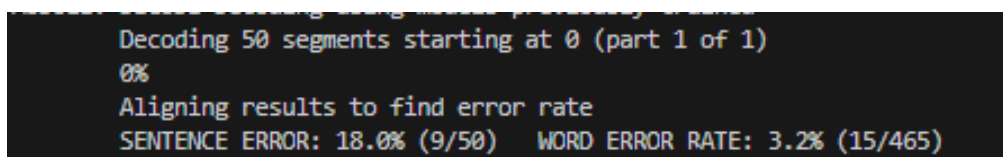
Jumlah koefisien MFCC : 13

Language model : 3-gram

Dataset tanpa Male 16.

Hasil yang diperoleh adalah sebagai berikut.

Eksperimen dengan sebagian data train :



```
Decoding 50 segments starting at 0 (part 1 of 1)
0%
Aligning results to find error rate
SENTENCE ERROR: 18.0% (9/50)  WORD ERROR RATE: 3.2% (15/465)
```

Eksperimen dengan data test :

```
Decoding 50 segments starting at 0 (part 1 of 1)
0%
Aligning results to find error rate
SENTENCE ERROR: 50.0% (25/50)  WORD ERROR RATE: 13.8% (66/478)
```

Eksperimen kedua dilakukan dengan parameter-parameter berikut.

Sampling rate : 16000

Jumlah filter : 25

Filter (High) : 6800

Filter (Low) : 130

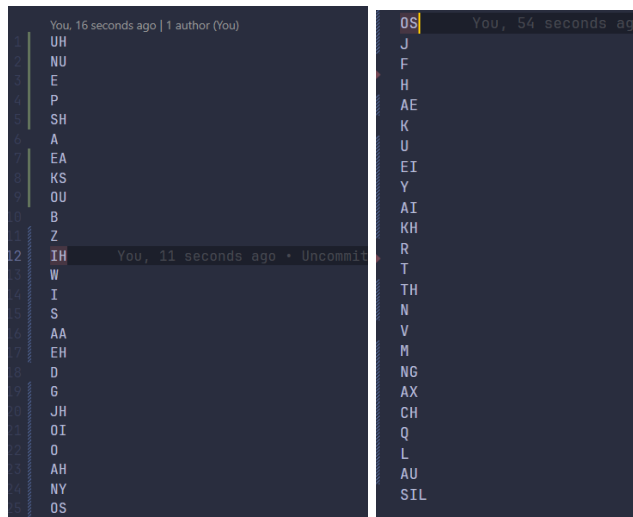
Jumlah koefisien MFCC : 13

Language model : 3-gram

Hasil yang diperoleh adalah sebagai berikut.

```
TOTAL Words: 1194 Correct: 1151 Errors: 47
TOTAL Percent correct = 96.40% Error = 3.94% Accuracy = 96.06%
TOTAL Insertions: 4 Deletions: 8 Substitutions: 35
```

Eksperimen ketiga dilakukan dengan menggunakan phone dictionary yang berbeda sebagai berikut. Percobaan menggunakan phone yang lebih kompleks. Phone sebelumnya hanya menggunakan alphabet A-Z. Sedangkan phone baru adalah sebagai berikut.



Eksperimen ketiga dilakukan dengan parameter-parameter berikut.

Sampling rate : 16000

Jumlah filter : 25

Filter (High) : 6800

Filter (Low) : 130

Jumlah koefisien MFCC : 13

Language model : 3-gram

Hasil yang diperoleh adalah sebagai berikut.

Hasil yang diperoleh adalah sebagai berikut.

```
MODULE: DECODE Decoding using models previously trained
Decoding 123 segments starting at 0 (part 1 of 1)
0%
Aligning results to find error rate
SENTENCE ERROR: 8.1% (10/123)  WORD ERROR RATE: 2.5% (29/1194)
```

C. Analisis Hasil dan Rekomendasi

Dalam membangun model, dataset yang digunakan memegang peranan krusial. Namun, terdapat beberapa kekurangan pada dataset yang perlu diperbaiki. Sebagian percakapan dalam dataset kurang jelas, dan terdapat kebisingan yang signifikan meskipun telah dilakukan pra-pemrosesan. Idealnya, pengumpulan dataset sebaiknya dilakukan menggunakan mikrofon dan pengaturan ruangan yang seragam guna memastikan kualitas data yang optimal.

Dokumentasi CMUSpeech merekomendasikan penggunaan sample rate 16000 sebagai pilihan yang lebih baik. Sedangkan dataset audio yang akan di train memiliki sample rate 44100. Meskipun memungkinkan untuk menggunakan sample rate lain seperti 44100, percobaan telah menunjukkan bahwa menggunakan sample rate 16000 menghasilkan kualitas yang lebih baik.

Berdasarkan hasil percobaan yang dilakukan, model yang optimal ditemukan pada Model 3 dengan penggunaan sample rate 16000 dan penggunaan phone yang lebih kompleks. Phone yang lebih kompleks memiliki kemampuan untuk mendeteksi fitur-fitur yang lebih

detail dalam data suara, sehingga memberikan hasil yang lebih baik dalam pengolahan informasi suara.