Studi Format Audio dan Teks Untuk Modul *Speech to Text*

Elizabeth Irenne Yuwono[1], Tony Antonio[2]

***Abstrak*— *Automatic Speech Recognition* (ASR) atau *speech to text* adalah bidang teknologi identifikasi ucapan manusia dalam bentuk teks transkripsi. Penelitian ini merupakan studi pada format masukan dan luaran *speech to text,* yaitu audio (ucapan) dan teks. Studi berfokus pada karakteristik dan format sinyal audio, pemrosesan sinyal audio secara digital dan relasinya dengan modul *speech to text,* pengetahuan linguistik, karakteristik dan format teks, serta isu pengembangan modul *speech to text.* Sinyal audio untuk ucapan memiliki beberapa karakteristik unik yang membedakannya dengan sinyal audio lain. Karakter-karakter ini merupakan fitur yang digunakan untuk identifikasi ucapan dalam sinyal audio masukan. Dalam modul *speech to text* sinyal digital mengalami beberapa proses sebelum identifikasi ucapan dilakukan. Proses sinyal digital ini dilakukan untuk memperoleh sinyal ucapan dengan tingkat kebisingan terendah dan hasil akurasi tinggi. Beberapa proses tersebut antara lain: *sampling,* kuantisasi, Fourier transform, sistem waktu diskrit, dan *digital filter*. Modul perlu memiliki pengetahuan linguistik untuk dapat mengetahui tata bahasa ucapan. Hasil identifikasi disimpan dalam bentuk teks transkripsi sesuai dengan karakter huruf bahasa tersebut. Melalui studi ini, diperoleh beberapa isu yang dapat dijadikan pertimbangan untuk penelitian selanjutnya terkait rancang-bangun modul *speech to text,* antara lain: pengaruh sumber dan format sinyal audio pada akurasi modul, kompleksitas tata bahasa dan pengucapan dan hubungannya dengan proses sinyal digital, pengaturan format karakter teks untuk luaran modul dan potensi pengembangan penelitian pada bidang lain.**

Kata Kunci: **sinyal audio, sinyal ucapan, format audio, format teks, *speech to text***

***Abstract*— Automatic Speech Recognition (ASR) is technology field focusing on identifying human speech in the form of transcription text. This research is a study of the input and output of Speech to Text, they are audio (speech) and text. The study concentrates on speech to text module, linguistic knowledge, characteristic and format text, and the researcg issue on Speech to Text. In Speech to Text module, digital signal is processed in several phases before speech is obtained. They are: sampling, quantization, Fourier transform, discrete time system, and digital filter. In order to understanding the speech grammar, Speech to Text module is required to have linguistic knowledge. The result is saved in text format suitable for the spoken language character. There are some research issues**

learned for next related researches; the relation between audio signal source and signal to module accuration, the complexity of grammar and speech including their relation to digital signal processing, setting for text character format as module output and the potential for future researches.

Keywords: audio signal, speech signal, audio format, text format, speech to text

I. PENDAHULUAN

Automatic Speech Recognition (ASR) adalah bidang teknologi yang mampu mengidentifikasi audio ucapan manusia dalam bentuk teks transkripsi. ASR telah banyak diterapkan dalam kehidupan sehari-hari, contohnya adalah operasi pada *smartphone* yang dapat dijalankan melalui perintah otomatis *(voice recognition)*. Perkembangan teknologi ASR saat ini telah sampai pada tahap pemahaman ucapan berkesinambungan dari pembicara berbeda dengan akurasi 99% untuk waktu training dibawah 10 menit. Kosakata yang dapat disimpan oleh sistem mencapai 250.000 istilah dan 160.000 kata aktif. Tingkat kompleksitas perkembangan ASR terus meningkat sejalan dengan kompleksitas ucapan dan bahasa manusia. Penelitian ini merupakan studi pada format audio sebagai masukan dan teks sebagai luaran dari modul *speech to text.*

Jurnal ini terbagi menjadi enam bagian; bagian pertama membahas mengenai sinyal audio dan formatnya serta karakteristik sinyal ucapan. Bagian kedua membahas pemrosesan sinyal digital yang digunakan untuk memperoleh informasi sinyal audio. Hasil luaran dari pemrosesan sinyal digital digunakan untuk identifikasi ucapan pada modul *speech to text*. Bagian selanjutnya membahas pengetahuan linguistik. Pengetahuan linguistik merupakan komponen modul *speech to text* yang digunakan untuk pengaturan tata bahasa teks ucapan. Kemudian hasil luaran modul diatur sesuai format teks yang sesuai dengan bahasa pengucapan. Bagian selanjutnya membahas modul *speech to text* dan isu-isu pengembangan modul tersebut dengan relasinya dengan komponen-komponen modul yang telah dibahas sebelumnya. Bagian akhir dari jurnal ini membahas isu-

[1] *Jurusan Teknik Informatika Fakultas Industri Kreatif Universitas Ciputra, UC Town, Surabaya 60219 INDONESIA (telp: 031-745 1699; e-mail: eirene@ciputra.ac.id)*

[2] *Jurusan Teknik Informatika Fakultas Industri Kreatif Universitas Ciputra, UC Town, Surabaya 60219 INDONESIA (telp: 031-745 1699; e-mail: tonyantonio@ciputra.ac.id)*

isu penelitian pada *speech to text* sebagai hasil studi penelitian ini.

II. SINYAL AUDIO

Perubahan pada tekanan udara yang diterima oleh pendengaran kita adalah gelombang suara. Sinyal audio adalah gelombang suara dalam arah longitudinal. Frekuensi untuk setiap sinyal audio berbeda-beda dan merepresentasikan seberapa cepat perubahan sinyal audio per detik. Frekuensi gelombang suara yang dapat diterima oleh indra pendengaran manusia adalah antara 20 Hz sampai 20 kHz. Gelombang suara pada area frekuensi ini yang dianggap sebagai sinyal audio (Umapathy, 2010). Area intensitas pendengaran sinyal audio adalah 120 dB yang merepresentasikan area. Pemrosesan sinyal Terdapat properti lain dari sinyal audio selain frekuensi, yaitu periode dan amplitudo. Periode adalah waktu yang dibutuhkan selam satu siklus gelombang dalam satuan sekon. Amplitudo adalah ukuran volum sinyal pada waktu tertentu.

(Gerhard, 2003) Sinyal audio dapat diklasifikasikan menjadi dua jenis, yaitu suara yang terdengar oleh manusia dan suara yang tidak dapat didengarkan. Suara yang dapat didengarkan terdiri dari: suara alami, buatan, ucapan, musik, dan kebisingan. Suara alami adalah semua suara yang berasal dari alam dan tidak menerima modifikasi apapun. Suara buatan adalah suara yang dibuat atau dimodifikasi oleh manusia kecuali ucapan dan musik. Musik adalah suara yang berasal dari instrument musik dengan melodi yang harmonis. Kebisingan adalah sinyal acak yang dapat diklasifikasikan berdasarkan distribusi energinya pada spektrum sinyal. Khusus untuk kebisingan, suatu sinyal audio dapat dianggap sebagai kebisingan dari beberapa aspek. Untuk identifikasi sinyal audio pada modul *speech to text,* sinyal kebisingan adalah sinyal selain sinyal ucapan manusia, termasuk suara alami, buatan, dan musik.

III. KARAKTERISTIK SINYAL AUDIO

Terdapat beberapa karakteristik pada sinyal audio yang dapat digunakan untuk menentukan perbedaan antara sinyal ucapan dan sinyal lain yang dianggap sebagai kebisingan. Karakteristik ini dikategorikan menjadi dua, yaitu karakteristik fisik dan persepsi. Karakteristik fisik meliputi energi, *zero-crossing rate*, fitur spektral, frekuensi dasar, lokasi forman, fitur berbasis waktu, dan modulasi (Gerhard, 2003). Karakteristik persepsi meliputi nada dan prosodi, bingkai suara dan tanpa suara, warna nada, dan ritme.

Karakteristik fisik adalah fitur-fitur unik dalam sinyal audio yang diperoleh dari perhitungan langsung dari amplitudo gelombang audio atau nilai spektral jangka pendek. *Zero-crossing rate* diterapkan untuk mengklasifikasikan suara perkusi. *Zero-crossing rate* (ZCR) mengukur perubahan sinyal saat melalui frekuensi nol dari positif ke negatif dalam satuan waktu dan menyimpannya dalam bentuk informasi sinyal spektral. Sinyal spektral adalah distribusi frekuensi dalam sinyal, analisa spektral berhubungan dengan pendengaran dan persepsi manusia untuk sinyal audio. Terdapat relasi antara ZCR dan frekuensi dasar (f_0); frekuensi dasar adalah frekuensi terendah dalam gelombang sinyal audio secara periodik. Dalam modul *speech to text,* frekuensi dasar digunakan sebagai nilai basis untuk menentukan yang mana yang merupakan frekuensi sinyal ucapan. Lokasi forman hanya terdapat dalam sinyal ucapan, sehingga fitur ini dapat digunakan untuk deteksi sinyal ucapan.

Karakteristik persepsi adalah karakteristik relative yang diperoleh dari informasi sinyal audio yaitu karakteristik fisik. Nada memiliki relasi dengan frekuensi dasar, perbedaan antara kedua fitur ini adalah; frekuensi dasar merupakan kuantitas numerik yang absolut sementara nada merupakan kuantitas relatif. Bingkai suara berasal dari nada atau frekuensi dasar. Jika suatu sinyal ucapan memiliki energi namun tidak terdeteksi nada pada tingkat nada normal maka sinyal tersebut dianggap memiliki bingkai tanpa suara. Bingkai suara cenderung harmonis dan memiliki *centroid* spektral lebih rendah dibandingkan bingkai tanpa suara. Warna nada membantu membedakan dua suara atau instrument yang memiliki tingkat kemiripan tinggi dengan mengkombinasikan informasi dari nada, intensitas, dan sebagainya. Ritme merepresentasikan tempo dari situasi sama yang berulang dalam sinyal audio. Fitur ini digunakan untuk deteksi pengulangan suara.

Jenis karakteristik sinyal audio dapat dibagi menjadi dua jenis (Rao, 2007), yaitu: karakteristik sementara dan spektral. Karakteristik sementara termasuk durasi dan modulasi amplitudo sinyal termasuk naik turun bentuk gelombang. Karakteristik spektral berhubungan dengan komponen frekuensi dan jumlahnya.

Gelombang audio dapat bersifat periodik dan aperiodik. Gelombang yang periodik memiliki nada kompleks yang terdiri dari sebuah frekuensi fundamental dan rangkaian tingkatan nada atau beberapa frekuensi fundamental. Warna nada dipengaruhi oleh amplitudo dan fase frekuensi. Sementara gelombang aperiodik terdiri dari nada-nada yang tidak harmonis dan memiliki frekuensi sinyal kebisingan.

IV. SINYAL AUDIO UCAPAN

Sinyal ucapan adalah sinyal audio yang berasal dari sistem pengucapan manusia dengan tujuan untuk komunikasi. Berbeda dengan sinyal audio lainnya, sinyal ucapan diasumsikan sebagai kumpulan *phone*. Sementara sinyal musik sebagai evolusi dari pola nada-nada. Terdapat dua jenis sinyal ucapan, yaitu sinyal ucapan bersuara dan tanpa suara. (Pollak, 2004) Sinyal ucapan bersuara berasal dari getaran pita suara, sementara ucapan tanpa suara memiliki sinyal bising seperti ‘s’, ‘sh’.

Frekuensi standar untuk sinyal ucapan manusia antara

100 sampai 3000 Hz, sedangkan frekuensi sinyal audio yang dapat diterima oleh pendengaran manusia antara 20 sampai 20.000 Hz.

V. KARAKTERISTIK SINYAL UCAPAN

Menurut Carey, karakteristik sinyal ucapan adalah *bandwidth* yang terbatas, area pergantian sinyal tanpa suara dan dengan suara, tingkat nada terbatas, durasi suku kata vocal, dan variasi energi antara level rendah dan tinggi (Carey, 1999). Terdapat beberapa karakteristik sinyal audio yang dapat digunakan untuk mengidentifikasi sinyal ucapan bersuara dan tanpa suara, antara lain:

A. Energi

Perbedaan antara energi pada sinyal ucapan bersuara dan tanpa suara pada area waktu tertentu dapat diamati dari amplitudonya. Amplitudo sinyal ucapan tanpa suara lebih tinggi dari sinyal ucapan tanpa suara. Berikut adalah perhitungan untuk memperoleh energi setiap blok sinyal ucapan dalam waktu 10-20 ms dengan asumsi energi sinyal ucapan bersuara selalu lebih besar daripada energi sinyal ucapan tanpa suara:

$$\boldsymbol{P_{av}} = \frac{\mathbf{1}}{\boldsymbol{L}}\sum_{\boldsymbol{n=1}}^{\boldsymbol{L}} \boldsymbol{x^2(n)}$$

dimana $\boldsymbol{P_{av,voiced}} > \boldsymbol{P_{av,unvoiced}}$ (1)

B. Zero-crossing Rate

Zero-crossing rate mengidentifikasi fluktuasi dalam sinyal ucapan. Sinyal ucapan tanpa suara mengalami pergerakan lebih cepat dibandingkan sinyal ucapan bersuara, sehingga ZCR pada sinyal tanpa suara lebih tinggi daripada sinyal bersuara.

$$x(n_0)\, x(n_0 + 1) < 0$$

dimana $x(n_0)_{unvoiced} > x(n_0)_{voiced}$ (2)

ZCR mengukur jumlah perubahan gelombang audio pada waktu tertentu. ZCR memberikan informasi spektral dengan energi rendah. Untuk pencarian ZCR pada sinusoid digunakan frekuensi fundamental. Pada sinyal yang lebih kompleks digunakan frekuensi rata-rata. Pada sinyal ucapan ZCR diperoleh dari nilai segmen suara dan tanpa suara yang mengalami fluktuasi cepat. Konsentrasi energi spektral dalam sinyal ucapan berbeda dengan konsentrasi energi spektral pada sinyal audio lain, pada sinyal ucapan terjadi persebaran sehingga ZCR diperoleh dari nilai persebaran tersebut.

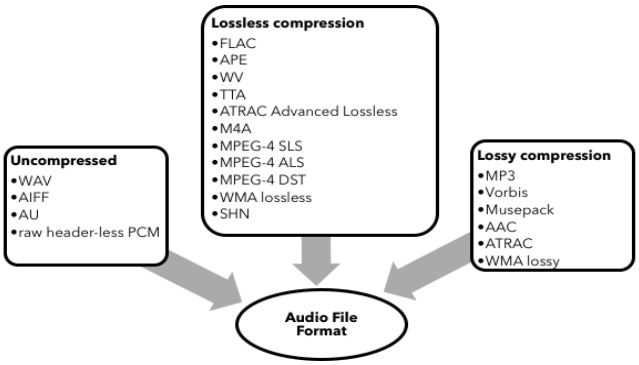
C. Deteksi Nada

Suara manusia memiliki tingkat nada terbatas. Suara pria memiliki tingkat nada antara 85 sampai 155 Hz, sementara suara wanita memiliki tingkat nada 165 sampai 255 Hz. Area ini tidak meliputi tingkat nada suara manusia saat bernyanyi yaitu antara 80 sampai 1100 Hz. Batas frekuensi yang telah diketahui ini digunakan sebagai tolak ukur saat deteksi nada.

VI. FORMAT FILE AUDIO

File audio digital adalah wadah penyimpanan informasi sinyal audio dalam format digital. File audio terdiri dari dua komponen yaitu kontainer dan *raw data*. Kontainer menyimpan *raw data* audio tanpa perlindungan keamanan. Sementara *raw data* sendiri adalah data audio sebenarnya yang belum mempunyai properti rating sampel, saluran audio, dan sebagainya. Terdapat tiga jenis format file audio berdasarkan sistem penyimpanannya, yaitu: *uncompressed*, *lossless compression*, dan *lossy compression* seperti ditunjukkan pada gambar 1.

Audio *uncompressed* disimpan sebagaimana adanya sesuai format *raw data*. *Lossless* dan *lossy compression* mengalami proses kompresi sebelum disimpan. Proses kompresi mengurangi area dinamis dari sinyal audio yaitu dengan menggunakan metode koding, identifikasi pola, dan prediksi linear untuk mengurangi kuantitas penyimpanan data. Perbedaan antara kompresi *lossless* dan *lossy* adalah pada hasil kompresinya; kompresi *lossless* menghasilkan duplikat dari alur audio asli dengan rasio kompresi antar 50 sampai 60 persen, kompresi *lossy* menghasilkan alur audio dengan rasio kompresi lebih tinggi namun dengan akurasi lebih rendah.



Gambar 1. Klasifikasi Format File Audio Berdasarkan Metode Kompresinya

Dalam ASR atau *speech to text* semakin tinggi akurasi audio, maka semakin tinggi pula akurasi hasil transkripsi. Oleh karena itu audio *uncompressed (raw data)* merupakan jenis masukan terbaik untuk modul *speech to text*. Format file audio *uncompressed* antara lain: WAV *(Waveform Audio File Format)*, AIFF *(Audio Interchange File Format)*, LPCM *(Linear Pulse Code Modulation)*. Format WAV adalah generasi file audio pertama. WAV memiliki kualitas audio terbaik karena belum mengalami alterasi digital namun memiliki ukuran file besar. AIFF memiliki kualitas dan prinsip penyimpanan yang sama dengan WAV namun ditujukan

untuk penggunaan pada sistem operasi Macintosh. LPCM menggunakan metode sandi modulasi kode pulsa linear untuk menyimpan sinyal audio digital. Dalam prinsip LPCM, gelombang audio direpresentasikan sebagai nilai amplitudo yang di rekam dalam rangkaian waktu tertentu. Modulasi kode pulsa adalah metode penyimpanan dan transmisi audio *uncompressed* digital. LPCM menggunakan kuantisasi linear untuk merepresentasikan nilai amplitude dalam skala linear.

VII. PEMROSESAN SINYAL DIGITAL

Digital signal processing atau pemrosesan sinyal digital adalah proses perubahan, penyaringan, dan estimasi sinyal audio menjadi data binary. Untuk memperoleh nilai dari setiap karakteristik sinyal audio diperlukan pemrosesan sinyal digital. Modul *speech to text* menerima data audio dalam bentuk data binari untuk dapat melakukan identifikasi ucapan. Terdapat beberapa proses sinyal digital yang sering digunakan untuk tujuan identifikasi ucapan, antara lain:

A. Sampling

Terdapat dua jenis representasi sinyal audio yaitu: sinyal kontinu dan diskrit. Sejak awal sinyal audio merupakan sinyal kontinu yang cukup kompleks untuk sistem, oleh karena itu *sampling* berfungsi untuk menyederhanakan sinyal tersebut menjadi sinyal diskrit. Sinyal diskrit menyimpan rangkaian sinyal sampel dalam interval waktu tertentu. Sampel adalah jumlah periode gelombang pada waktu tertentu.

Terdapat dua kondisi untuk menentukan interval sampel berdasarkan teorema *sampling* (Orfanidis, 2010), yaitu:

1) Sinyal *x(t)* harus tersimpan dalam bentuk berkas sinyal, spektrum frekuensinya harus kurang dari atau sama dengan frekuensi maksimal f_{max}.
2) Rating sampel f_s minimal dua kali dari frekuensi maksimal.

$$\boldsymbol{f_s \geq 2.f_{max}} \quad (3)$$

Rating sampel minimal disebut rating Nyquist, sedangkan kuantitas $f_s/2$ disebut sebagai frekuensi Nyquist. Batas interval frekuensi Nyquist adalah . Frekuensi maksimal f_{max} sinyal ucapan adalah 4 kHz dengan rating Nyquist 8 kHz, sementara secara umum sinyal audio memiliki frekuensi maksimal 20 kHz dan rating Nyquist 40 kHz.

B. Fourier Transform

Fourier transform melakukan analisa spektral pada sinyal audio. Fourier menyederhanakan domain frekuensi pada sinyal audio dengan menyimpan informasi gelombang sebagai jumlah fungsi sinusoid dari frekuensi-frekuensi berbeda. Fungsi sinusoid merupakan hasil pemecahan sinyal berdasarkan domain frekuensinya. Nilai fungsi sinusoid merepresentasikan nilai karakteristik sinyal audio tersebut. Suatu fungsi sinusoid dikatakan memiliki nilai benar apabila kumpulan nilai tersebut (Fourier *series*) dapat disusun menjadi sinyal audio yang sama lagi. Fungsi dasar yang digunakan untuk representasi sinusoid adalah fungsi sinus dan cosinus. Hasil dari Fourier transform adalah kumpulan domain frekuensi sinyal dalam jumlah kecil. Terdapat empat jenis Fourier transform antara lain: Fourier transform, Fourier *series*, Fourier transform diskrit, dan Fourier transform waktu diskrit. Perbedaan antara keempat jenis Fourier transform ditunjukkan pada tabel 1 (Smith, 1997). Fourier *series* adalah jumlah nilai fungsi basis (sinus dan cosinus) frekuensi. Untuk pemrosesan sinyal audio jenis Fourier transform yang digunakan adalah Fourier transform diskrit. Fourier transform ini bekerja sesuai periode waktu dan mengukur setiap siklus frekuensi $X_k.e^{i2\pi kn/N}$. . Dalam setiap siklus frekuensi, nilai, deviasi, dan kecepatan rotasi diperhitungkan (Azad, n.d.). x_n adalah nilai sinyal pada waktu-n hasil penjumlahan seluruh siklus sebanyak N periode.

$$x_n = \frac{1}{N}\sum_{k=0}^{N-1} X_k.e^{i2\pi kn/N} \quad (4)$$

TABEL I
JENIS-JENIS FOURIER TRANSFORM

Jenis Transform	Perbedaan
Fourier transform	Sinyal kontinu dan periode tidak beraturan
Fourier *series*	Sinyal kontinu dan periode beraturan
Fourier transform diskrit	Sinyal diskrit dan periode beraturan
Fourier transform waktu diskrit	Sinyal diskrit dan periode tidak beraturan

C. Kuantisasi

Kuantisasi adalah pemetaan kumpulan nilai masukan dalam jumlah besar menjadi kumpulan kecil. Jika *sampling* merupakan digitalisasi domain frekuensi, maka kuantisasi adalah digitalisasi area frekuensi.

D. Sistem Waktu Diskrit

Sistem waktu diskrit adalah prose penambahan rangkaian masukan *x[n]* menjadi rangkaian luaran *y[n]* dengan properti tambahan sesuai kebutuhan.

E. Digital Filter

Digital filter mengurangi atau meningkatkan aspek tertentu dalam sinyal waktu diskrit. Terdapat dua kategori *digital filter* dalam pemrosesan sinyal yaitu: representasi domain waktu dan domain frekuensi. Kategori domain waktu dalam *digital filter* adalah *filtering* sinyal berdasarkan koefisien formula perbedaan dan respon impuls. Sementara kategori domain frekuensi adalah *filtering* sinyal berdasarkan fungsi transfer dan respon frekuensi.

1) Koefisien Formula Perbedaan

Formula perbedaan adalah formula matematika untuk menghitung sampel luaran dalam *digital filter*

berdasarkan sampel masukan seperti ditunjukkan pada persamaan di bawah ini:

$$y(n) = b_0x(n) + b_1x(n-1) + \cdots + b_Mx(n-M) \\ -a_1y(n-1) - \cdots - a_Ny(n-N) \quad (5)$$

x adalah sinyal masukan dan *y* adalah sinyal luaran. *a* dan *b* adalah koefisien formula perbedaan atau koefisien *filter*.

2) Respon Impuls

Impuls adalah sinyal masukan yang menerima respon dari *digital filter*. Impuls didefinisikan oleh:

$$\delta(n) = \begin{cases} 0, n \neq 0 \\ 1, n = 0 \end{cases} \quad (6)$$

Sementara respon dari impuls tersebut didefinisikan oleh:

$$y(n) = \sum_{m=-\infty}^{\infty} h(m)x(n-m) \quad (7)$$

Terdapat dua jenis *digital filter* yang menggunakan respon impuls yaitu: FIR *(Finite Impulse Response)* dan IIR *(Infinite Impulse Response) filter*. Dalam IIR *filter*, respon impuls memiliki durasi terbatas sementara FIR *filter* tidak terbatas atau bekerja secara rekursif.

3) Fungsi Transfer

Fungsi transfer adalah *z transform* dari respon impuls yang didefinisikan oleh persamaan 8. *Z transform* dari respon impuls *H(z)* adalah rasio antara luaran *filter Y(z)* dengan transformasi masukan *filter X(z).*

$$H(z) = \frac{Y(z)}{X(z)} \quad (8)$$

4) Respon Frekuensi

Respon frekuensi adalah fungsi transfer yang berasal dari persamaan 6. Fungsi ini dievaluasi dalam unit lingkaran sebagai berikut:

$$H\left(e^{j\hat{w}}\right) = \sum_{k=0}^{M} b_k e^{-j\hat{w}k} \quad (9)$$

VIII. APLIKASI PEMROSESAN SINYAL DIGITAL

Pemrosesan sinyal digital dapat diaplikasikan dalam beberapa bentuk berbeda sesuai dengan fungsinya, Bab ini membahas beberapa aplikasi pemrosesan sinyal digital yang berhubungan dengan modul *speech to text*.

A. Penyaringan Kebisingan (Noise Reduction Filters)

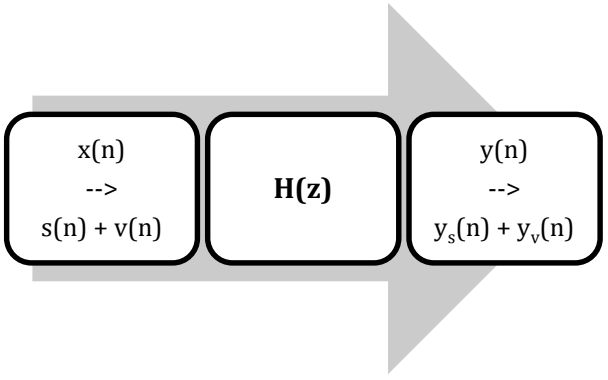
Penyaringan kebisingan adalah proses mengurangi tingkat kebisingan pada sinyal audio dengan melakukan perubahan nilai pada frekuensi dan karakteristik sinyal audio lain. Secara matematika dapat dituliskan sebagai berikut:

$$x(n) = s(n) + v(n) \quad (10)$$

x(n) adalah sinyal audio sumber, sementara *v(n)* adalah sinyal kebisingan yang ingin disaring, *s(n)* adalah hasil sinyal yang telah mengalami proses *filtering*. Terdapat beberapa macam sinyal kebisingan, antara lain:

1) Sinyal kebisingan putih, adalah sinyal latar belakang yang diperoleh dari proses pengukuran pada umumnya.
2) Sinyal gangguan periodikal, adalah sinyal frekuensi yang selalu muncul selalu periodikal, contoh: *60 Hz power-frequency pickup.*
3) Sinyal kebisingan dengan frekuensi rendah, contoh: kekacauan pada sinyal radar.
4) Lain-lain, sinyal kebisingan yang termasuk kategori ini tidak memiliki kriteria yang spesifik, suatu sinyal audio termasuk sinyal kebisingan berdasarkan kebutuhannya. Pada modul *speech to text* sinyal kebisingan adalah semua sinyal audio selain sinyal ucapan.

Metode untuk penyaringan kebisingan adalah dengan merancang suatu *filter H(z)* yang akan menghilangkan sinyal kebisingan *s(n)* dari *x(n).* Gambar 2 menunjukkan bagan proses penyaringan kebisingan standar. *Filter H(z)* melakukan pemrosesan digital pada sinyal *x(n)* dan diasumsikan dalam proses $y_s(n) = s(n)$ dan $y_v(n) = 0$.



Gambar 2. Bagan Proses *Noise Filtering* Standar

Dalam prakteknya persamaan ini tidak dapat diaplikasikan untuk semua kondisi, oleh karena itu digunakan domain frekuensi untuk analisa spektral antara sinyal sumber dan kebisingan yang tidak tumpang tindih. Berikut adalah

kondisi komponen sinyal pada *filter H(z)*.

$$Y_s(\omega) = H(\omega)S(\omega) = S(\omega)$$
$$Y_v(\omega) = H(\omega)S(\omega) = 0 \quad (11)$$

Tingkat kebisingan sinyal yang telah diproses oleh *filter H(z)* dapat diukur dengan *noise reduction ratio* (NRR). Namun NRR hanya berlaku untuk penyaringan sinyal kebisingan putih. Untuk sinyal kebisingan jenis lain, NRR tidak selalu dapat mengukur tingkat kebisingannya.

$$NRR = \frac{\sigma_{yv}^2}{\sigma_v^2} = \int_{-\pi}^{\pi} |H(\omega)|^2 \frac{d\omega}{2\pi} = \sum_n h_n^2 \quad (12)$$

Penyaringan kebisingan dan optimalisasi sinyal dapat diformulasikan dalam bentuk rasio sinyal ke *noise*. Minimalisasi NRR sama dengan maksimalisasi rasio sinyal ke *noise* (SNR) pada hasil *filter* $H(\omega)$.

B. Pemerataan Sinyal (Signal Averaging)

Pemerataan sinyal adalah proses meningkatkan rasio sinyal ke *noise* (SNR). Pada proses ini *noise* atau kebisingan pada sinyal tidak dihilangkan namun hanya dimanipulasi melalui pemerataan sinyal sehingga SNR sinyal lebih tinggi. Proses ini termasuk dalam konversi rating sampel untuk memperoleh sinyal diskrit dengan SNR lebih tinggi. Pemerataan sinyal melalui interpolasi dan desimalisasi menggunakan *finite impulse response* (FIR) *filter* disebut dengan *cascaded integrator-comb* (CIC) *filter*. CIC *filter* terdiri dari satu atau lebih pasangan integrator dan *comb filter*. *Comb filter* membuat versi sinyal berbeda dengan menambahkan jeda, respon frekuensinya menyerupai rangkaian gerigi seperti sisir.

Terdapat dua kondisi dalam pemerataan sinyal menggunakan CIC *filter*, yaitu: desimalisasi dan interpolasi (Miller, 2010). Untuk kondisi desimalisasi, sinyal masuk melalui alur integrator kemudian mengalami *downsampling,* proses ini dilakukan untuk setiap area *comb* sinyal. Dalam desimalisasi jumlah area *comb* sama dengan jumlah integrator. Desimator CIC diimplementasikan sebagai alur dari semua integrator, faktor-N *downsampler,* dan terusan dari pembeda setiap segmen *comb*. Untuk kondisi interpolasi, sinyal diproses dengan urutan sebaliknya dan *downsampling* diganti dengan *upsampling (zero-stuffer).* Pada interpolasi faktor-N *upsampler* diposisikan sebelum integrator dan setelah pembeda *comb*.

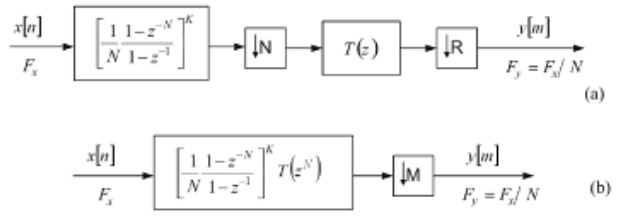
Terdapat dua parameter dalam *comb filter,* yaitu: *passband droop* dan faktor selektivitas. *Passband droop* (d_C^K) merupakan aturan maksimal untuk batas *bandwidth* yang dibandingkan dengan *low pass filter* ideal. Faktor selektivitas (ϕ_C^K) adalah rasio antara nilai respon tingkat *filter* yang diperoleh pada frekuensi batas *passband* F_m dan pada frekuensi batas bawah *aliasing band* pertama $\left({}^{F_x}/_{N-F_m}\right)$. *Passband droop* dan faktor selektivitas untuk frekuensi digital diperoleh dari persamaan berikut:

$$d_C^K = \left|\frac{\sin(\pi N f_m/2)}{N\sin(\pi f_m/2)}\right|^K$$
$$\phi_C^K = \left|\frac{\sin(\pi(1/N - f_x/2))}{\sin(\pi f_x/2)}\right|^K \quad (13)$$

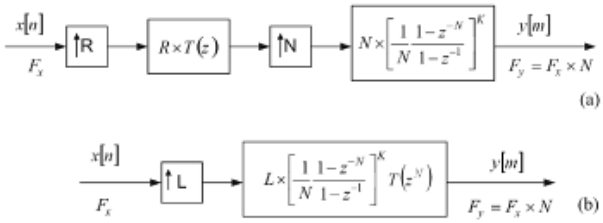
Selain CIC, terdapat FIR *(finite impulse response filter)*, FIR banyak digunakan untuk pemerataan sinyal menggunakan metode *comb filtering.* Pemerataan sinyal menggunakan FIR *filter* diperoleh melalui:

$$H(z) = \frac{1}{N}\,\frac{1-Z^{-N}}{1-Z^{-1}} \quad (14)$$

CIC dan FIR *filter* dapat diintegrasikan membentuk desimator dan interpolator dua langkah. Gambar 3 menunjukkan bagan proses desimator dua langkah antara CIC dan FIR *filter*. Gambar 3a adalah implementasi alur CIC dan gambar 3b adalah langkah penyetaraan satu tahap. Sementara gambar 4 menunjukkan bagan proses interpolasi yang terdiri dari dua tahap. Gambar 4a menunjukkan proses implementasi alur CIC dan gambar 4b menunjukkan penyetaraan satu tahap. Perbedaan antar proses desimalisasi dan interpolasi pada CIC dan FIR *filter* adalah pada alur prosesnya. Sama seperti pada CIC standar, alur proses interpolasi terbalik dengan alur proses desimalisasi pada tahap pertama *filtering*.



Gambar 3. Bagan proses desimalisasi pada pemerataan sinyal menggunakan CIC dan FIR *filter*



Gambar 4. Bagan proses interpolasi pada pemerataan sinyal menggunakan CIC dan FIR *filter*

C. Savitzly-Golay Smoothing Filter

Smoothing filter ini berbasis pada publikasi mengenai tabel koefisien konvolusi untuk ragam polynomial dan kelompok data oleh Abraham Savitzky dan Marcel J.E. Golay pada 1964. Savitzky-Golay *smoothing filter* adalah *filter* digital yang berfungsi meningkatkan SNR *(signal to*

noise ratio) tanpa merubah sinyal itu sendiri. Proses ini disebut penghalusan sinyal.

Savitzky-Golay *smoothing filter* melakukan konvolusi dengan menyesuaikan kelompok poin data yang sejajar dengan polynomial derajat rendah menggunakan metode *linear least square*. Metode ini digunakan ketika poin-poin data terpisah dalam jarak yang sama terdeteksi. Persamaan *least square* menghasilkan kelompok koefisien konvolusi. Koefisien-koefisien ini dapat diimplementasikan pada semua kelompok data untuk memperoleh estimasi sinyal yang telah dihaluskan pada setiap titik tengah kelompok data.

IX. PENGETAHUAN LINGUISTIK

Pengetahuan linguistik dibutuhkan dalam modul *speech to text* sebagai sumber pemahaman sistem mengenai pengucapan, tata bahasa, dan pengetahuan linguistic lain. Terdapat enam kategori pengetahuan linguistic (Reddy, 1976), yaitu:

1) *Fonologi*
   Fonologi adalah ilmu pengucapan linguistik. Fonetik membuat model pengucapan kata dalam unit *phone.*

2) *Morfologi*
   Morfologi adalah studi identifikasi komponen penting dalam sebuah kata. Morfologi menyimpan informasi dari bentuk dan perilaku kata.

3) *Sintaksis (ilmu kalimat)*
   Sintaksis adalah studi relasi struktural antar kata. Sintaksis melakukan pengurutan dan pengelompokkan kata-kata dalam sebuah kalimat.

4) *Semantik*
   Semantik adalah ilmu makna linguistik. Terdapat dua jenis semantik, yaitu semantik leksikal dan semantik komposisi. Semantik leksikal menentukan makna dari kata, sementara semantik komposisi mempelajari makna komponen-komponen kata yang ada untuk menentukan makna besar dari komponen-komponen tersebut.

5) *Pragmatik*
   Pragmatik menentukan bentuk formal dan informal bahasa.

6) *Konvensi wacana*
   Wacana adalah ilmu linguistik mengenai unit yang lebih kompleks daripada kata, wacana dapat berupa percakapan. Konvensi wacana memperbaiki struktur kalimat agar sesuai dengan maksud pembicara.

Setiap kategori dapat diasumsikan sebagai solusi ambiguitas dari kategori sebelumnya. Ambiguitas ini dapat diselesaikan menggunakan model atau algoritma tertentu. Model atau algoritma yang paling sering digunakan berkaitan dengan linguistik *speech to text* antara lain: *part of speech tagging,* disambiguasi makna kata, parsing probabilitas, dan interpretasi tutur kata. *Part of speech tagging* menentukan apakah sebuah kata termasuk kata benda, kata kerja, atau yang lain. Disambiguasi makna kata menyelesaikan ambiguitas makna kata. Parsing probabilitas menentukan apakah dua ucapan yang mirip termasuk entitas kata yang sama. Interpretasi tutur kata menentukan apakah sebuah kalimat adalah pernyataan, pertanyaan, atau perintah.

Terdapat dua format umum dalam fonologi komputasi, yaitu: IPA *(International Phonetic Alphabet)* dan ARPAbet. IPA dan ARPAbet sebagian besar menggunakan karakter huruf latin sebagai simbol, namun ARPAbet lebih memiliki huruf yang familiar karena menggunakan simbol ASCII dibandingkan IPA.

X. FORMAT TEKS

Dalam modul *speech to text,* luaran dari modul adalah teks transkripsi dalam format teks sederhana tanpa diperlukan informasi ragam teks. Format karakter yang digunakan dalam teks transkripsi bergantung pada karakter bahasa percakapan dalam audio masukan.

Berdasarkan data dari Ethonolouge, dari 7.106 bahasa yang ada di dunia, 3.570 telah memiliki sistem penulisan. Dari 3.570 bahasa ini sebagian besar menggunakan sistem penulisan yang sama. Sistem penulisan yang paling banyak digunakan oleh bahasa di dunia adalah sistem penulisan alfabet. Tabel 2 menunjukkan daftar sistem penulisan yang ada saat ini menurut Omniglot, Ensiklopedia Sistem Penulisan dan Bahasa Dunia (Ager, 2015).

TABEL 2
DAFTAR SISTEM PENULISAN BAHASA DI DUNIA

Sistem Penulisan	Bahasa
Abjad / Alfabet Konsonan	Arab, Ibrani
Alfabet	Inggris, Indonesia, Afrika, Rusia
Alfabet Silabel / Abugidas	Korea (Hangul), Hindi, Nepali
Sistem Suku Kata	Jepang (Hiragana)
Semanto-fonetik	Mandarin
Lain-lain	Eropa lama, Rohonc Codex (Hungaria)

A. Alfabet Konsonan

Sistem penulisan ini menganggap semua huruf harus memiliki konsonan. Huruf konsonan dapat menjadi huruf independen, sementara huruf vocal ditemani satu atau lebih huruf konsonan. Contoh bahasa yang menggunakan sistem penulisan ini adalah bahasa Arab dan Ibrani. Gambar 5 menunjukkan contoh penulisan alfabet konsonan pada bahasa Arab.

B. Alfabet

Sistem penulisan alfabet adalah sistem penulisan umum

yang biasa digunakan pada bahasa Indonesia, Inggris, dan bahasa lainnya. Pada sistem penulisan ini setiap huruf vokal dan konsonan adalah independen.

C. Alfabet Silabel (Abugidas)

Sistem penulisan alfabet silabel adalah sistem penulisan dengan komponen utamanya adalah suku kata. Setiap karakter tersusun dari 2 hingga 4 huruf membentuk satu suku kata. Contoh penulisan alfabet silabel pada bahasa Korea (Hangul) ditunjukkan pada gambar 6.

D. Sistem Suku Kata

Sistem suku kata adalah sistem penulisan fonetik yang membaca setiap simbol sebagai satu suku kata. Gambar 7 adalah contoh penulisan sistem suku kata pada bahasa Jepang (Hiragana).

E. Semanto-fonetik

Sistem penulisan semanto-fonetik adalah sistem penulisan yang membaca simbol sebagai suatu ucapan (vokal, konsonan, suku kata) dan masing-masing memiliki arti. Terdapat tiga jenis simbol semanto-fonetik, antara lain: *pictogram, ideogram,* dan karakter majemuk. *Pictogram* adalah simbol yang merepresentasikan benda sebenarnya. Simbol jenis ini biasa digunakan pada *Hieroglyphic* Mesir kuno. *Ideogram* adalah simbol yang merepresentasikan ide abstrak, simbol jenis ini biasa digunakan untuk penomoran karakter Mandarin. Simbol jenis karakter majemuk banyak digunakan pada karakter Mandarin. Penulisan karakter majemuk merepresentasikan makna dan bunyi pengucapan karakter tersebut. Namun terkadang beberapa simbol hanya digunakan untuk bunyi pengucapannya tanpa peduli maknanya.

Enkripsi teks sederhana dapat menggunakan beberapa unit kode, antara lain: ASCII, UTF-8, EBCDIC, UTF-16, dan UTF-32. ASCII *(American Standard Code for Information Interchange)* merupakan enkripsi karakter berdasarkan alfabet untuk sistem penulisan bahasa Inggris. ASCII hanya menyimpan karakter huruf latin dan tidak bisa digunakan untuk bahasa dengan sistem penulisan berbeda dengan bahasa Inggris, seperti: bahasa Arab, Mandarin, dan Eropa. EBCDIC *(Extended Binary Coded Decimal Interchange Code)* adalah sistem penulisan karakter 8-bit yang menggunakan basis sistem operasi dan perangkat IBM. UTF-8, UTF-16, UTF-32 termasuk dalam sistem *Unicode* standar yang dikembangkan menggunakan basis *World Wide Web*. Secara umum sistem *Unicode* didesain untuk penulisan karakter huruf latin, namun seluruh sistem *Unicode* memiliki referensi karakter untuk simbol selain huruf latin seperti karakter Hangul (bahasa Korea), Hiragana (bahasa Jepang), Arab, dan sebagainya. Perbedaan antara ketiga jenis sistem *Unicode* tersebut adalah pada jumlah karakter yang digunakan saat *encoding*, yaitu transformasi 8-bit, 16-bit, dan 32-bit. UTF-8 merupakan sistem penulisan yang paling banyak digunakan saat ini dibandingkan unit kode lainnya. UTF-8 didesain untuk mengurangi kompleksitas pada format ASCII dalam pengaturan karakter bukan huruf latin. Kapasitas enkripsi UTF-8 mampu meliputi hampir semua jenis karakter dalam berbagai macam bahasa.

XI. MODUL *SPEECH TO TEXT*

Speech to text adalah bidang teknologi yang berfokus pada identifikasi ucapan manusia dalam bentuk teks transkripsi. *Speech to text* dikembangkan dengan tujuan untuk memperoleh informasi dari audio, dan pengembangan sistem komputer pintar (kecedasan buatan) yang mampu memahami bahasa manusia. Menurut Waibel, sistem *speech to text* dapat dibagi menjadi beberapa jenis berdasarkan tingkat kesulitan intrinsik dan dimensinya (Waibel, 1990):

A. Word Recognition-Isolated (WR)

Sistem WR adalah sistem *speech to text* dengan tipe masukan berupa ucapan kata-kata terisolasi. Setiap kata diberi jeda saat diucapkan, sehingga penggunaannya terbatas. Ukuran kosakata WR berkisar antara 10 hingga 300 kata.

B. Connected Speech Recognition-restricted (CSR)

Sistem CSR adalah sistem *speech to text* dengan tipe masukan berupa ucapan kata bersambung tanpa jeda. Sistem jenis ini dapat menyimpan kosakata antara 30 hingga 500 kata namun hanya dapat menerima bahasa perintah terbatas, bukan semua bahasa. Saat perekaman audio masukan, lingkungan perekaman harus hening dengan tingkat kebisingan rendah serta ucapan pembicara harus jelas. Tingkat akurasi sistem CSR dipengaruhi oleh kejelasan audio masukan.

C. Speech Understanding-restricted (SU)

Sistem SU merupakan jenis sistem *speech to text* yang bertujuan untuk memahami ucapan masukan. Oleh karena itu sistem SU dapat menerima ucapan kata bersambung dan dapat digunakan secara lebih bebas dibandingkan sistem WR dan CSR. Bahasa yang dapat dipahami oleh sistem SU adalah bahasa sesuai data pengetahuan linguistik (misal: bahasa Inggris) namun dengan tata bahasa terbatas. Keterbatasan tata bahasa SU dipengaruhi oleh kapasitas penyimpanan kosa kata sistem SU yaitu antara 100 hingga 2000 kata. Namun pada SU, sistem tidak lagi bergantung sepenuhnya pada siapa pembicaranya namun sistem sudah lebih pintar untuk dapat memahami ucapan berbagai pembicara sebagai satu kata yang sama. Tingkat pemahaman sistem *speech to text* bergantung pada sumber pengetahuan linguistik, semakin lengkap pengetahuan linguistik sistem maka tingkat pemahaman SU akan semakin kompleks.

D. Dictation Machine-restricted (DM)

Dictation machine adalah sistem *speech to text* yang merekam ucapan dan menyimpan hasil transkripsi teksnya untuk penggunaan selanjutnya. DM telah banyak

digunakan untuk keperluan personal maupun legal dan medis. DM memerlukan kosakata yang lebih besar daripada sistem *speech to text* lainnya yaitu antara 1000 hingga 10.000 kata dan pengguna DM harus merekam suaranya dengan jelas dalam lingkungan hening. Tingkat pemahaman DM bergantung pada kompleksitas pengetahuan linguistiknya dan kejelasan audio masukan.

E. Unrestricted Speech Understanding (USU)

USU adalah sistem *speech to text* yang membutuhkan kapasitas kosakata tidak terbatas untuk dapat menerima masukan ucapan kata bersambung dengan menggunakan pengetahuan linguistik untuk pemahaman teks transkripsi.

F. Unrestricted Connected Speech Recognition

Unrestricted Connected SR adalah sistem *speech to text* yang memiliki kapasitas kosakata dan jenis masukan sama dengan USU namun sistem tidak mengasumsikan jenis informasi apa yang disampaikan oleh pembicara sehingga pemahaman sistem lebih kompleks dari segi pengetahuan linguistik.

Dari seluruh jenis sistem *speech to text* tersebut, akurasi sistem bergantung dari jenis informasi yang ingin diperoleh. Jenis informasi yang ingin diperoleh menggunakan *speech to text* akan mempengaruhi waktu respon sistem dalam memproses audio masukan dan relasinya dengan pengetahuan linguistik. Akurasi dan waktu respon sistem dapat diatur, semakin tinggi akurasi sistem maka waktu respon semakin lama, hal ini juga berlaku sebaliknya.

Menurut Ethonologue, salah satu sumber tata bahasa dunia terbaik, pada tahun 2014 terdapat 7,106 bahasa dengan 6,2 milyar pembicara di seluruh dunia (Lewis, 2014). Jumlah ini menunjukkan betapa kompleks dan luasnya potensi bidang *speech to text.* Namun terdapat beberapa isu dalam pengembangan *speech to text* (Waibel, 1990), antara lain:

1) *Ucapan bersambung, terisolasi*
   Ucapan manusia memiliki berbagai keragaman meliputi gaya bahasa dan dialek, sebagian berbicara secara berkesinambungan tanpa jeda, sebagian berbicara dengan jeda untuk setiap kata. Karena perbedaan eksistensi jeda ini, maka modul *speech to text* harus mampu membedakan ucapan yang berbeda namun merupakan kata yang sama.

2) *Ukuran kosakata*
   Pada Desember 2010, sebuah studi oleh Harvard dan Google menemukan bahwa bahasa Inggris memiliki 1.022.000 kata dan terus bertambah dengan rating 8.500 kata per tahun (Injeeli, 2013). Berdasarkan data ini, modul *speech to text* harus mampu menyimpan setidaknya satu juta kata untuk setiap bahasa agar dapat mengidentifikasi ucapan secara akurat.

3) *Keterbatasan tata bahasa*
   Setiap bahasa memiliki tata bahasa yang berbeda, tata bahasa ini diterapkan sebagai pengetahuan linguistik.

4) *Ketergantungan sistem pada karakter suara pengguna*
   Modul *speech to text* memiliki dua jenis sistem identifikasi, yaitu ketergantungan pada suara pengguna dan sistem independen. Sistem yang bergantung pada suara pengguna memiliki keterbatasan identifikasi karena bersifat statis. Akurasi terbaik dapat diperoleh jika karakter suara pengguna sama atau mendekati data pelatihan. Sistem independen bersifat dinamis, setiap sistem mengidentifikasi suara baru, suara dimasukan ke data pelatihan untuk meningkatkan akurasi pada identifikasi selanjutnya.

5) *Ambiguitas akustik*
   Dalam beberapa kosakata bahasa, satu kata yang sama dapat memiliki arti yang berbeda tergantung pada konteksnya pada kalimat. Ambiguitas ini diselesaikan melalui aspek semantik pada pengetahuan linguistik.

6) *Kebisingan lingkungan*
   Untuk dapat memperoleh akurasi tinggi, sinyal ucapan masukan harus memiliki level kebisingan yang rendah. Namun tidak setiap masukan memiliki level kebisingan yang rendah, oleh karena itu bidang teknologi *noise filter* dikembangkan untuk mengurangi level kebisingan audio.

XII. ISU PENELITIAN

Berdasarkan studi ini terdapat beberapa isu penelitian yang perlu dipertimbangkan untuk penelitian selanjutnya terkait topik ini. Sinyal audio memiliki berbagai jenis konten tergantung lingkungan sumbernya. Beberapa sinyal audio dari sumber berbeda dapat memiliki nilai frekuensi yang sama, hal ini perlu diperhatikan dalam pengembangan modul *speech to text* agar modul mampu mengidentifikasi sumber audio sebenarnya.

Beberapa isu penting dalam *speech to text* yang telah disebutkan pada Bab 10 menunjukkan bahwa akurasi hasil *speech to text* dipengaruhi oleh keterbatasan pemrosesan audio pada format tertentu. Selain format audio, kompleksitas karakteristik ucapan manusia juga merupakan aspek penting yang harus diperhitungkan dalam pengembangan modul *speech to text.*

Isu penelitian *speech to text* pada aspek format teks berasal dari pemilihan format karakter teks untuk luaran modul. Modul *speech to text* harus dapat mencakup ragam karakter huruf dalam berbagai bahasa yang ada di dunia.
Isu-isu penelitian yang diperoleh dari penelitian ini menunjukkan luasnya potensi penelitian dalam bidang *speech to text*. Dalam hal ini penelitian selanjutnya tidak hanya meliputi bidang *speech to text* saja, namun juga bidang kecerdasan buatan, pemrosesan sinyal digital, matematika dan teori probabilitas, serta pengetahuan linguistik.

DAFTAR PUSTAKA

[1] Ager, S. (2015). *Omniglot: Types of Writing System*. Wales, United Kingdom: Omniglot Limited. Retrieved May 15, 2015 from www.omniglot.com

[2] Azad, K. (n.d.). Math, Better Explained. diakses dari http://betterexplained.com/articles/an-interactive-guide-to-the-fourier-transform/X

[3] Carey, MJ, Parris, ES, Lloyd-Thomas, H. (1999). A Comparison of Features for Speech, Music Discrimination. *Acoustics, Speech, and Signal Processing, 1999 IEEE International Conference, vol. 1*, pp. 149 – 152

[4] Gerhard, D. (2003). *Audio Signal Classification: History and Current Techniques.* Canada: University of Regina

[5] Greenberg, I, Bate, A. (1999). *The Future of Speech Recognition*. Atlanta: Emory University

[6] Orfanidis, SJ. (2010). *Introduction to Signal Processing*. USA : Rutgers University

[7] Injeeli, P. (2013). *Mind Your Words: Master The Art of Learning and Teaching Vocabulary.* Singapore: Trafford

[8] Lewis, M, Paul, Simons, GF, Fennig, CD. (2014). *Ethnolouge: Languages of The World, 17th Edition.* Dallas, Texas: SIL International. Retrieved January 10, 2015 from www.ethnologue.com

[9] Miller, FP, Vandome, AF, McBrewster, J. (2010). *Cascaded Integrator-Comb Filter.* Germany: VDM Publishing

[10] Pollak, I. (2004). *Digital Signal Processing with Applications – Speech Processing*. USA: Purdue University

[11] Rao, P. (2007). *Audio Signal Processing*. India: Indian Institute of Technology Bombay

[12] Reddy, DR. (1976). Speech Recognition by Machine: A Review. *1976 IEEE Proceedings, vol. 64, chapter 4*, pp. 502 – 531

[13] Smith, SW. (1997*). The Scientist and Engineer's Guide to Digital Signal Processing 1st edition*. California: California Technical Pub

[14] Umapathy, K, Ghoraani, B, Krishnan, S. (2010). Audio Signal Processing Using Time-Frequency Approaces: Coding, Classification, Fingerprinting, and Watermarking. *EURASIP Journal on Advances in Signal Processing 2010:451695*

[15] Waibel, A, Lee, KF. (1990). *Readings in Speech Recognition*. California: Morgan Kaufmann