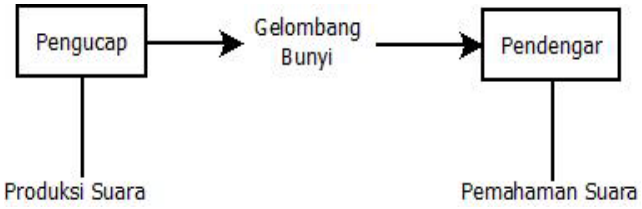
BAB II

DASAR TEORI

2.1 Suara (*Speaker*)

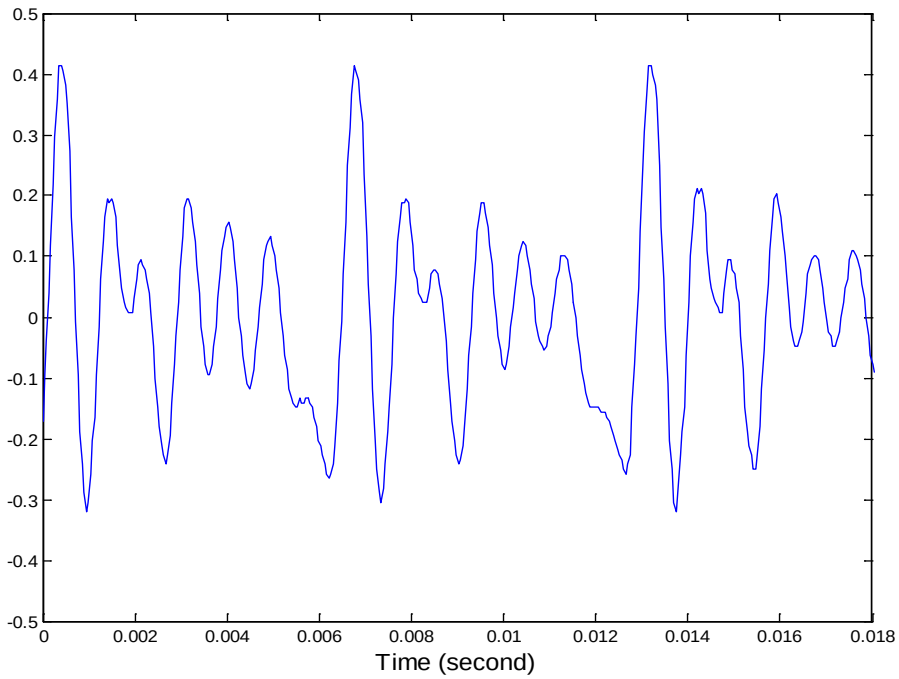
Suara adalah sinyal atau gelombang yang merambat dengan frekuensi dan amplitudo tertentu melalui media perantara yang dihantarkannya seperti media air, udara maupun benda padat. Manusia dapat berkomunikasi dengan manusia lainnya dengan suara. Pembangkitan ucapan manusia dimulai dengan awal konsep dari gagasan yang ingin disampaikan pada pendengar. Pengucap mengubah gagasan tadi dalam struktur linguistic dengan memilih kata atau frasa yang secara tepat dapat mewakili dan membawakannya dengan tata bahasa yang dimengerti antara pengucap dan pendengar. Ucapan yang diucapkan memiliki tujuan tertentu dengan asumsi bahwa ucapan tersebut diucapkan secara benar, dapat diterima, dan dipahami oleh pendengar yang dituju.

Pembangkitan ucapan pada hakekatnya berhubungan dengan kemampuan mendengar. Sinyal ucapan dibangkitkan oleh organ vokal dan ditransmisikan melalui udara menuju telinga pendengar. Pada Gambar 2.1 diperlihatkan proses antara pengucap dengan pendengar serta mekanisme dalam produksi suara dan pemahaman suara oleh manusia [1].



Gambar 2.1 Lingkaran komunikasi Suara

Sinyal suara terjadi secara perlahan waktu variasi sinyal (disebut sebagai kuasi stasioner). Contoh dari sinyal suara yang ditunjukkan pada Gambar 2.2 dibawah. Ketika diperiksa selama periode yang cukup singkat (5 sampai 100 msec), karakteristiknya cukup stasioner. Namun, selama jangka waktu yang lama (diurutan 1/5 detik atau lebih) sinyal karakteristik dapat mengubah pantulan berbicara berbeda dengan suara yang diucapkan. Oleh karena itu, waktu singkat spectral analisis adalah cara yang paling umum untuk mengkarakteristik sinyal suara.



Gambar 2.2 Contoh sinyal suara

Pada dasarnya banyak macam kemungkinan parameter yang mewakili sinyal suara untuk melakukan pengenalan pembicara, seperti Linear Prediksi Coding (LPC), *Mel Frequency Cepstrum Coefficients* (MFCC), dan lain –lain. MFCC mungkin yang paling dikenal dan paling popular, dan akan dijelaskan dalam tulisan ini.

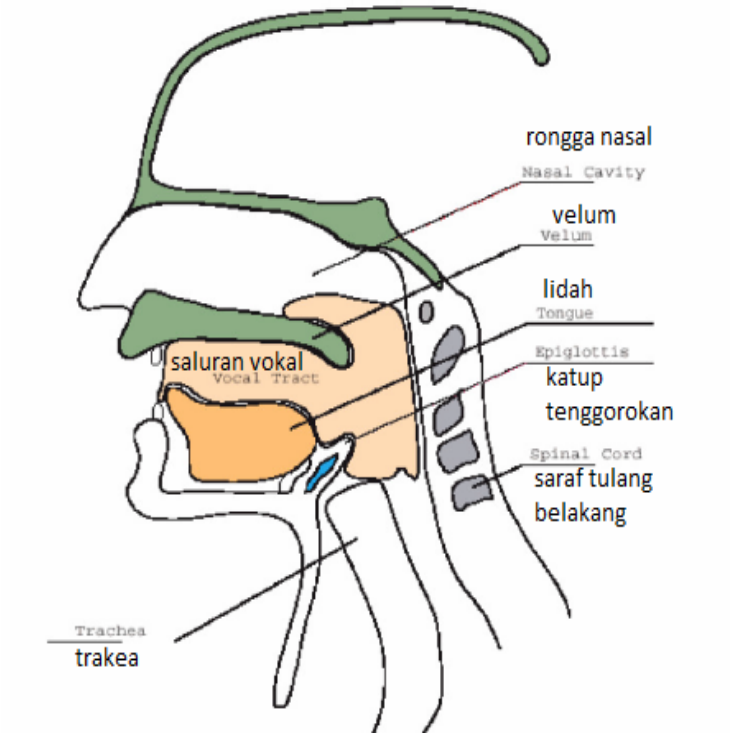
MFCC (*mel frequency cepstrum* coefficients) yang didasarkan pada variasi Bandwidth yang dikenali telinga manusia dengan frekuensi, filter spasi linear pada frekuensi rendah dan logaritmik pada frekuensi tinggi telah digunakan untuk menangkap karakteristik penting dari pembicara. Hal ini dinyatakan dalam skala *mel frequency*, yang merupakan frekuensi linier berada dibawah 1000 Hz dan logaritmik diatas 1000 Hz [2].

2.2 Pengolahan suara

Pengolahan suara adalah suatu perkembangan teknik dan sistem yang memungkinkan komputer suatu perangkat untuk mengenali dan memahami kata – kata yang diucapkan dengan cara digitalisasi kata dan mencocokkan sinyal digital tersebut dengan suatu pola tertentu yang tersimpan dalam suatu perangkat. Kata - kata diucapkan diubah bentuknya menjadi sinyal digital dengan cara mengubah gelombang suara menjadi sekumpulan angka yang kemudian disesuaikan dengan kode – kode tertentu untuk mengidentifikasikan kata – kata tersebut, hasil dari identifikasi kata yang diucapkan dapat ditampilkan dalam bentuk tulisan atau dapat dibaca oleh perangkat teknologi sebagai sebuah komando untuk melakukan suatu pekerjaan [3].

2.2.1 Produksi Pengolahan Ucapan

Untuk dapat memahami bagaimana produksi ucapan dilakukan, maka kita perlu mengetahui bagaimana Mekanisme vocal manusia dibangun. Pada Gambar 2.3 bagian yang paling penting dari mekanisme vocal manusia adalah saluran vocal bersama dan rongga nasal, yang dimulai pada velum. Velum merupakan sebuah mekanisme seperti pintu jebakan yang digunakan untuk merumuskan bunyi nasal saat diperlukan. Ketika velum diturunkan, rongga nasal digabungkan bersama-sama dengan saluran vocal untuk merumuskan sinyal ucapan yang diinginkan. Daerah *crossectional* dari saluran vocal dibatasi oleh lidah, bibir, rahang dan velum dan bervariasi 0-20 cm2 [4].



Gambar 2.3 Mekanisme vocal manusia

2.2.2 Sifat ucapan manusia

Salah satu tolak ukur yang paling penting dari ucapan adalah frekuensi ucapan itu sendiri. Ucapan dapat dibedakan satu sama lain dengan bantuan frekuensi. Ketika frekuensi ucapan meningkat, nada ucapan menjadi tinggi dan menyakitkan. Ketika frekuensi ucapan berkurang, ucapan akan lebih dalam. Gelombang ucapan adalah gelombang yang terjadi dari getaran materi ucapan. Nilai tertinggi dari frekuensi yang manusia dapat hasilkan sekitar 10 kHz. Dan nilai terendah adalah sekitar 70 Hz.

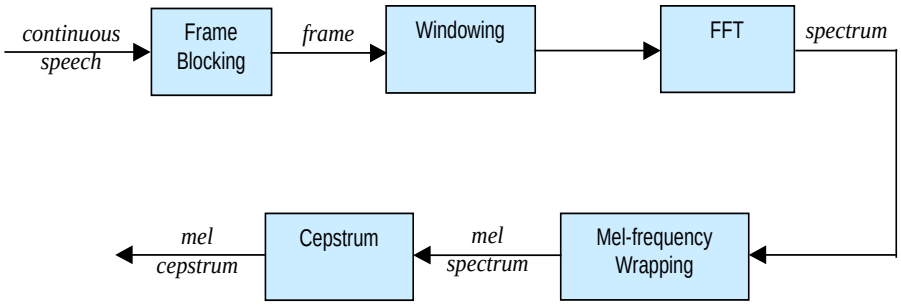
Ini adalah nilai – nilai maksimum dan minimum. Interval frekuensi ini berubah untuk setiap orang. Dan besarnya ucapan dinyatakan dalam decibel (dB). Ucapan manusia normal memiliki Interval frekuensi 100 Hz – 3200 Hz dan besarmya antara 16 Hz dan 20 kHz. Dan 0,5 % perubahan frekuensi adalah kepekaan telinga manusia [4].

Karakteristik Pembicara :

a) Berdasarkan perbedaan panjang saluran vocal, laki-laki, perempuan, dan ucapan anak-anak yang berbeda.
b) Aksen daerah adalah perbedaan frekuensi resonansi, jangka waktu, dan nada.
c) Individu memiliki pola frekuensi resonansi dan pola durasi yang unik (memungkinkan kita untuk mengidentifikasi pembicara).

2.3 *Mel Frequency Cepstrum Coefficients* (MFCC)

Mel Frequency Cepstrum Coefficients (MFCC) merupakan satu metode yang banyak dipakai dalam bidang *speech recognition.* Metode ini digunakan untuk melakukan *feature extraction*, sebuah proses yang mengkonversikan sinyal suara menjadi beberapa parameter. Masukan suara biasanya direkam pada sampling rate diatas 10000 Hz. Frekuensi sampling ini dipilih untuk meminimalkan atau mengkonversi efek aliasing dari analog ke digital. Sinyal-sinyal ini dapat menangkap semua frekuensi sampai dengan 5 Hz, yang meliputi sebagian besar energi suara yang dihasilkan oleh manusia. Seperti yang telah dibahas sebelumnya, tujuan utama dari proses MFCC adalah untuk mengikuti perilaku telinga manusia. Lihat Gambar 2.4 [2].



Gambar 2.4 Block diagram proses MFCC

Keunggulan dari metode MFCC ini adalah :

a. Mampu menangkap karakteristik suara yang sangat penting bagi pengenalan suara atau dengan kata lain mampu menangkap informasi-informasi yang terkandung dalam sinyal suara.

b. Menghasilkan data seminimal mungkin tanpa menghilangkan informasi-informasi penting yang ada.
c. Mereplikasi organ pendengaran manusia dalam melakukan persepsi sinyal suara.

2.4 *Frame Blocking*

Frame Blocking adalah pembagian sinyal *audio* menjadi beberapa *frame* yang nantinya dapat memudahkan dalam perhitungan dan analisa sinyal, suatu *frame* terdiri dari beberapa sampel tergantung tiap berapa detik suara akan disampel dan berapa frekuensi *samplingnya*. Pada proses ini dilakukan pemotongan sinyal dalam slot-slot tertentu agar memenuhi syarat yaitu *linear* dan *timeinvariant.*

Dalam langkah ini sinyal suara yang kontinyu diblock menjadi *frame sampel* N, dengan *frame* yang berdekatan dipisahkan oleh M (M<N). *Frame* pertama terdiri dari N sampel, *Frame* kedua dimulai sampel M setelah *frame* yang pertama, dan melawati dari sampel N-M dan seterusnya. Proses ini berlanjut sampai semua suara dicatat dalam satu *frame* atau lebih. Nilai-nilai untuk N dan M akan berubah-rubah sesuai dengan pengujian yang akan dilakukan [5].

2.5 *Windowing*

Dalam melakukan pemrosesan sinyal, maka dari input yang dimasukkan akan terbentuk sinyal yang magnitudenya bervariasi pada awal maupun akhir *frame.* Hal tersebut menghambat pemrosesan sinyal dan menghasilkan keluaran

yang kurang akurat. Untuk itu perlu diaplikasikan suatu *window* penghalus pada setiap *frame* dengan melakukan *overlapping* antara satu *frame dengan* frame yang lain, sehingga dapat dibangkitkan suatu *feature* yang lebih halus sepanjang durasi waktu tersebut. Dalam proyek ini akan digunakan metode *Hamming*. Digunakan *Hamming window* karena *Hamming window* memiliki *side lobe* yang paling kecil dan *Main lobe* yang paling besar sehingga hasil *windowing* akan lebih dalam menghasilkan efek diskontinuitas. Konsep disini adalah untuk meminimalkan distorsi spectral dengan menggunakan *window* untuk sinyal ke nol pada awal dan akhir disetiap *frame.* Jika kita mendefenisikan *window seperti* ini, dimana N adalah jumlah sampel disetiap *frame,* maka hasil *windowing* adalah sinyal [2].

Sebuah fungsi *window* yang baik harus menyempit pada bagian *main lobe* dan melebar pada bagian *side lobe*-nya.

Berikut ini adalah representasi dari fungsi *window* terhadap *signal* suara yang diinputkan :

$$y_1(n) = x_1(n)w(n), \qquad 0 \leq n \leq N-1 \tag{2.1}$$

Dimana :

$x(n) = x_1(n)w(n)$ n = 0,1,….,N-1

$x(n)$ = nilai sampel *signal* hasil *windowing*

$x_1(n)$ = nilai sampel dari *frame* signal ke i

$w(n)$ = fungsi *window*

N = *frame size*, merupakan kelipatan 2

Windowing Hamming biasa digunakan sebagai berikut :

$$w(n) = 0.54 - 0.46.\cos\left(\frac{2.\pi.n}{N-1}\right), 0 \leq n \leq N-1 \quad (2.2)$$

Dimana :

w (n) = *windowing*

N = jumlah data dari sinyal

n = waktu diskrit

2.6 *Fast Fourier Transform* (FFT)

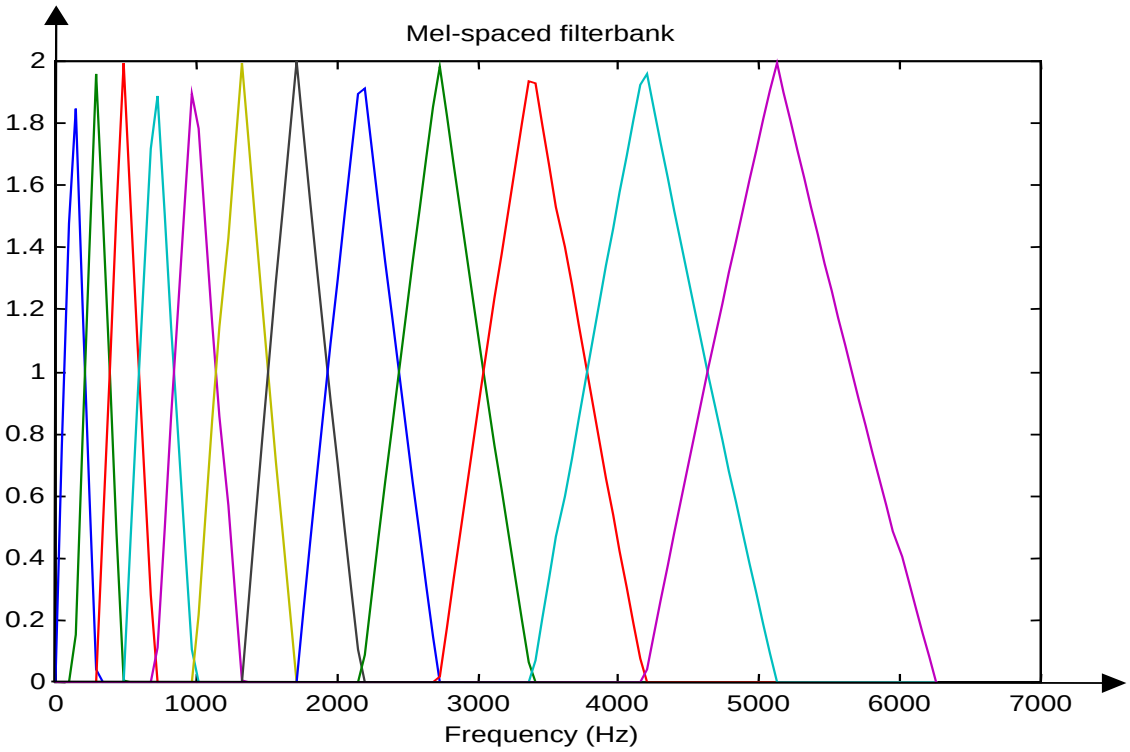
Langkah pengolahan selanjutnya adalah *Fast Fourier Transform* (FFT), yang mengubah setiap *frame* sampel N dari domain waktu ke domain frekuensi. FFT adalah algoritma cepat untuk mengimplementasikan *Discrete Fourier Transform* (DFT), yang didefenisikan pada himpunan N sampel {xn} sebagai berikut :

$$X_k = \sum_{n=0}^{N-1} x_n e^{-j2\pi kn/N}, \qquad k = 0,1,2,\ldots, N-1 \quad (2.3)$$

Dalam X_k*'s* adalah bilangan kompleks dan hanya mempertimbang kan nilai tersebut (besaran frekuensi). Urutan yang dihasilkan $\{X_k\}$ ditafsirkan sebagai berikut : frekuensi positif $0 \leq f < F_s / 2$ sesuai dengan nilai-nilai $0 \leq n \leq N/2 - 1$, sedangkan frekuensi negative $-F_s / 2 < 0$ sesuai dengan $N/2+1 \leq n \leq N-1$. Dimana Fs menunjukkan frekuensi sampling [2].

2.7 *Mel Frequency Wrapping*

Studi psikofisik telah menunjukkan bahwa persepsi manusia tentang frekuensi suara untuk sinyal ucapan tidak mengikuti skala linear. Jadi, untuk setiap suara dengan frekuensi seseungguhnya f, dalam Hz, sebuah pola diukur dalam sebuah skala yang disebut "*mel*". Skala "*mel frequency*" adalah skala frekuensi linear dibawah 1000 Hz dan skala logaritmik diatas 1000 Hz. Salah satu pendekatan untuk simulasi spectrum subjektif adalah dengan menggunakan filterbank, jarak pada mel skala (lihat Gambar 2.5). Artinya Filter bank memiliki respon frekuensi Bandpass segitiga, dan jarak bandwidth ditentukan oleh interval frekuensi mel konstan. Jumlah koefisien spectrum mel, K, biasanya dipilih sebagai 20.



Gambar 2.5 Contoh Mel - spasi filterbank

Filterbank ini dapat diterapkan dalam domain frekuensi, sehingga hanya sebesar yang diterapkan dijendela segitiga, bentuk seperti pada gambar diatas sampai spectrum. Sebuah cara yang digunakan tentang Filter bank *mel frequency* ini adalah untuk melihat setiap filter sebagai histogram bin (dimana bins memiliki kemampuan) dalam domain frekuensi [2].

Skala ini didefenisikan oleh Stanley Smith, John Volkman dan Edwin Newman sebagai :

$$mel(f) = 2595^{*}\log_{10}(1 + \frac{f}{700}) \qquad (2.4)$$

Dalam *mel frequency wrapping*, sinyal hasil FFT dikelompokkan kedalam berkas filter triangular ini. Maksud pengelompokan disini adalah setiap nilai FFT dikalikan terhadap *gain filter* yang bersesuaian dan hasilnya dijumlahkan.

2.8 Cepsturm

Cepstrum adalah sebutan kabalikan untuk *spectrum*. *Cepstrum* biasa digunakan untuk mendapatkan informasi dari suatu sinyal suara yang diucapkan oleh manusia. Pada langkah terakhir ini, spectrum log *mel* dikonversikan menjadi *cepstrum* menggunakan *Discrete Cosine Transform* (DCT). Oleh karena itu jika kita menunjukkan tersebut koefisien spectrum daya mel yang merupakan hasil dari langkah terakhir $\tilde{S}_0$, $k = 0,2,..., K-1$, kita dapat menghitung MFCC seperti :

$$\tilde{c}_n = \sum_{k=1}^{K} (\log \tilde{S}_k) \cos\left[n\left(k - \frac{1}{2} \right) \frac{\pi}{K} \right], \qquad n = 0,1,..., K\text{-}1 \qquad (2.5)$$

Perhatikan bahwa kita mengecualikan komponen pertama, dari DCT karena merupakan nilai rata-rata dari sinyal input, yang dilakukan speaker informasi spesifik [2].