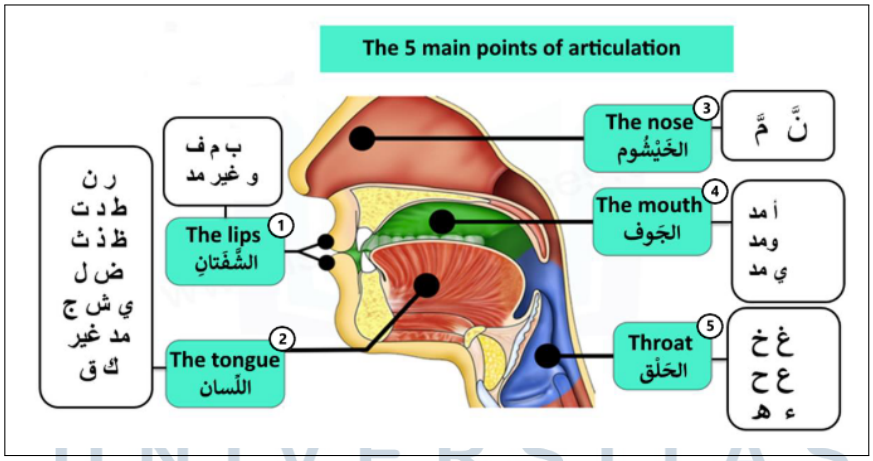
BAB 2
LANDASAN TEORI

2.1 Makhraj Fathah

Makhraj merupakan istilah dalam ilmu tajwid yang merujuk pada tempat keluarnya huruf hijaiyah saat dilafalkan. Konsep ini menjadi landasan utama dalam pembelajaran Al-Qur'an karena ketepatan makhraj berpengaruh langsung terhadap kejelasan bunyi serta makna bacaan. Secara linguistik, makhraj didefinisikan sebagai titik artikulasi pada rongga mulut atau tenggorokan yang menghasilkan suatu fonem tertentu. Dalam konteks pengajaran Al-Qur'an, pemahaman makhraj sangat penting agar setiap huruf dilafalkan sesuai dengan karakteristik fonetiknya.

Penelitian "*Identifying the Correct Articulation Point of Quranic Letters of the Throat (Al-Halqu) Makhraj*" menjelaskan bahwa makhraj adalah titik artikulasi huruf Arab yang dapat berasal dari beberapa bagian organ ucap, seperti dua bibir (1. As-Syafatan), lidah (2. Al-Lisan), rongga hidung (3. Al-Khaisyum), rongga mulut dan tenggorokan (4. Al-Jauf), dan tenggorokan (5. Al-Halqu) [8] yang dapat dilihat pada Gambar 2.1.



Gambar 2.1. Titik artikulasi bahasa Arab

Sumber: [8]

Meskipun terdapat perbedaan pendapat di kalangan ulama mengenai jumlah rinci makhraj, secara umum lima kategori utama tersebut menjadi acuan dalam pembelajaran tajwid. Setiap huruf hijaiyah memiliki lokasi artikulasi yang spesifik, sehingga perubahan kecil pada posisi lidah, bibir, atau tenggorokan dapat menghasilkan perbedaan bunyi yang signifikan.

Dalam konteks pembelajar non-penutur asli bahasa Arab, kesalahan pengucapan makhraj sering terjadi akibat pengaruh bahasa ibu dan perbedaan sistem fonetik. Penelitian "*The Classification of Actual Pronunciation of Quranic Alphabets for Non-Arab Speaker*" menunjukkan bahwa variasi aksen dan latar belakang bahasa menyebabkan ketidaktepatan artikulasi huruf-huruf Al-Qur'an. Studi tersebut juga menegaskan bahwa hingga saat ini belum banyak penelitian yang secara komprehensif mengidentifikasi ketepatan pelafalan seluruh 28 huruf hijaiyah pada pembelajar non-native [14].

Salah satu aspek penting dalam pelafalan huruf hijaiyah adalah penggunaan harakat, termasuk fathah. Fathah merupakan vokal pendek dengan karakteristik bunyi menyerupai fonem /a/ dalam bahasa Arab. Berbeda dengan huruf konsonan yang memiliki titik artikulasi spesifik, fathah sangat dipengaruhi oleh posisi makhraj huruf yang menyertainya. Artinya, kualitas bunyi fathah akan berbeda tergantung pada huruf dasar yang dilafalkan. Misalnya, fathah pada huruf yang berasal dari makhraj tenggorokan (Al-Halqu) akan memiliki resonansi yang berbeda dibandingkan fathah pada huruf yang berasal dari makhraj lidah (Al-Lisan). Selain itu, fathah memiliki durasi yang relatif singkat sehingga sensitif terhadap variasi temporal dan perubahan kecil pada posisi artikulator.

Karakteristik tersebut menjadikan evaluasi akurasi makhraj fathah sebagai tantangan tersendiri dalam analisis fonetik berbasis sinyal suara. Perubahan kecil pada titik artikulasi dapat memengaruhi spektrum frekuensi serta pola temporal sinyal audio yang dihasilkan. Oleh karena itu, pendekatan berbasis pengolahan sinyal dan pembelajaran mesin diperlukan untuk mengidentifikasi pola akustik yang merepresentasikan ketepatan artikulasi secara objektif dan terukur.

Dengan demikian, pemahaman terhadap konsep makhraj dan karakteristik fonetik huruf berharakat fathah menjadi landasan teoritis utama dalam penelitian ini. Evaluasi yang dilakukan tidak berfokus pada pengenalan kata atau transkripsi teks seperti pada sistem Automatic Speech Recognition (ASR), melainkan pada analisis ketepatan artikulasi berdasarkan karakteristik akustik yang dihasilkan. Oleh karena itu, pendekatan berbasis pengolahan sinyal suara dan pembelajaran mesin digunakan untuk mengidentifikasi pola spektral dan temporal yang merepresentasikan akurasi makhraj secara objektif dan terukur.

2.2 Konsep Evaluasi Artikulasi dalam Pembelajaran Bahasa

Evaluasi artikulasi merupakan proses penilaian terhadap ketepatan pelafalan bunyi bahasa berdasarkan aspek fonetik tertentu, seperti *place of articulation*, *manner of articulation*, serta karakteristik segmental lainnya. Dalam konteks pembelajaran bahasa, khususnya pembelajaran Al-Qur'an, evaluasi artikulasi makhraj menjadi komponen penting untuk memastikan ketepatan pelafalan huruf sesuai dengan kaidah tajwid.

Dalam literatur *Computer-Aided Pronunciation Training* (CAPT), sistem evaluasi pelafalan otomatis dikembangkan untuk memberikan umpan balik yang konsisten dan objektif kepada pembelajar bahasa. Berdasarkan kajian *Automatic Pronunciation Assessment – A Review*, CAPT memiliki dua fungsi utama, yaitu:

1. *Pronunciation teaching*, yaitu membimbing perbaikan pelafalan.
2. *Pronunciation assessment*, yaitu menilai dan mendeteksi kesalahan pelafalan.

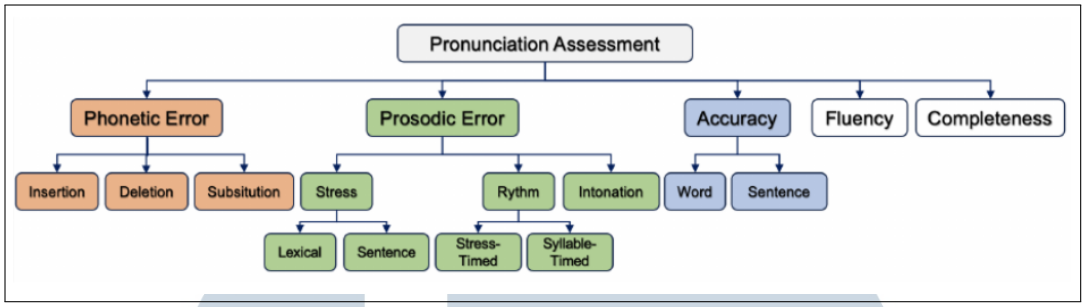
Penelitian tersebut menekankan bahwa penilaian pelafalan otomatis lebih kompleks dibandingkan dengan *Automatic Speech Recognition* (ASR) konvensional. ASR bertujuan mengenali kata tanpa terlalu memperhatikan kesalahan fonetik, sedangkan *pronunciation assessment* harus mampu mendeteksi perbedaan fonetik yang terkadang sangat halus [15].

Secara konseptual, tidak terdapat definisi absolut benar atau salah dalam pelafalan. Pelafalan berada dalam spektrum mulai dari tidak dapat dipahami hingga menyerupai penutur asli. Oleh karena itu, evaluasi pelafalan dapat diklasifikasikan menjadi dua pendekatan utama, yaitu evaluasi objektif dan evaluasi subjektif.

Evaluasi objektif melibatkan analisis berbasis parameter fonetik yang dapat diukur secara akustik, meliputi:

1. Segmental (bunyi per huruf atau fonem),
2. Suprasegmental (intonasi dan ritme),
3. Sub-segmental (tempat dan cara artikulasi).

Dalam konteks ini, Gambar 2.2 menjelaskan kesalahan pelafalan (*mispronunciation*) secara umum dibagi menjadi dua kategori utama:



Gambar 2.2. Tipe-tipe Error Pengucapan untuk Penilaian

Sumber: [15]

1. Phonetic (Segmental) Errors

Kesalahan segmental berkaitan dengan produksi bunyi individual seperti vokal dan konsonan. Kesalahan ini dapat berupa:

1. *Substitution* (penggantian bunyi),
2. *Insertion* (penambahan bunyi),
3. *Deletion* (penghilangan bunyi).

Kesalahan ini terjadi akibat perbedaan sistem fonologi, kesalahan konversi huruf-ke-bunyi, atau kesalahan artikulasi. Dalam pembelajaran tajwid, kesalahan segmental dapat berupa kesalahan makhraj, seperti pergeseran tempat keluarnya bunyi dari tenggorokan ke rongga mulut.

2. Prosodic (Suprasegmental) Errors

Kesalahan prosodik berkaitan dengan unsur yang memengaruhi pengucapan pada tingkat kata atau kalimat secara keseluruhan, seperti:

1. *Stress* (penekanan),
2. *Rhythm* (ritme),
3. *Intonation* (pola nada).

Fitur prosodik sangat berpengaruh terhadap intelligibility, terutama pada bahasa tonal. Namun, dalam penelitian ini, fokus utama berada pada aspek segmental dan sub-segmental, yaitu ketepatan tempat dan cara artikulasi makhraj huruf hijaiyah berharakat fathah.

Evaluasi subjektif dilakukan melalui penilaian manusia berdasarkan beberapa indikator berikut:

1. *Intelligibility*, yaitu sejauh mana ujaran dapat dipahami.
2. *Comprehensibility*, yaitu tingkat kemudahan ujaran untuk dipahami.
3. *Accentedness*, yaitu tingkat kemiripan pelafalan dengan standar atau penutur asli.

Dalam sistem evaluasi otomatis modern, penilaian pelafalan sering dikaitkan dengan beberapa parameter kuantitatif tambahan, yaitu:

1. **Accuracy**, yaitu tingkat ketepatan pengucapan terhadap referensi yang benar.
2. **Fluency**, yaitu kelancaran pengucapan yang berkaitan dengan tempo, jeda, dan kesinambungan ujaran.
3. **Completeness**, yaitu kelengkapan pengucapan tanpa adanya bunyi yang terlewat atau terpotong.

Ketiga parameter tersebut umum digunakan dalam sistem *automatic pronunciation assessment* untuk memberikan skor komprehensif terhadap performa pelafalan. Metode tradisional pembelajaran tajwid menggunakan pendekatan subjektif melalui talaqqi, yaitu interaksi langsung antara guru dan murid. Meskipun efektif, metode ini memiliki keterbatasan dalam hal aksesibilitas, skalabilitas, dan konsistensi penilaian. Perkembangan *Artificial Intelligence* (AI) memungkinkan sistem otomatis memberikan evaluasi pelafalan secara kuantitatif. Sistem berbasis AI mampu mengukur kesalahan segmental maupun prosodik melalui analisis fitur akustik dan model pembelajaran mesin [16]. Dengan pendekatan ini, evaluasi tidak hanya bersifat klasifikasi benar atau salah, tetapi juga menghasilkan skor numerik yang merepresentasikan tingkat kesesuaian pelafalan.

Terdapat beberapa tantangan utama dalam membangun sistem evaluasi artikulasi, yaitu:

1. **Variasi Temporal**
   Sinyal suara bersifat berurutan dan memiliki variasi durasi, tempo, serta jeda antar pembicara, sehingga memerlukan mekanisme penyesuaian waktu.
2. **Perbedaan Fonetik Halus**
   Perbedaan makhraj huruf hijaiyah terkadang sangat tipis secara akustik, sehingga membutuhkan metode ekstraksi fitur yang sensitif terhadap distribusi energi frekuensi.

3. **Keterbatasan Data Kesalahan**
   Data kesalahan pelafalan umumnya tidak seimbang dibandingkan data pelafalan benar.

Oleh karena itu, sistem evaluasi harus mampu menangkap pola spektral sekaligus mempertimbangkan aspek temporal dalam sinyal suara. Dalam konteks penelitian ini, sistem yang dikembangkan berperan sebagai *Decision Support System* (DSS) dalam evaluasi artikulasi pelafalan huruf hijaiyah. Fokus evaluasi berada pada aspek segmental dan sub-segmental, yaitu ketepatan produksi bunyi huruf serta kesesuaian tempat keluarnya bunyi (*makhraj*). Untuk merepresentasikan karakteristik akustik dari artikulasi tersebut, penelitian ini menggunakan fitur *Mel-Frequency Cepstral Coefficients* (MFCC) yang mampu menangkap distribusi energi frekuensi pada sinyal suara. Selanjutnya, *Convolutional Neural Network* (CNN) digunakan untuk mempelajari pola spektral yang merepresentasikan karakteristik artikulasi huruf. Selain itu, variasi durasi pelafalan antar pembicara juga menjadi faktor penting dalam evaluasi pengucapan. Oleh karena itu, metode *Dynamic Time Warping* (DTW) digunakan untuk melakukan penyesuaian temporal (*temporal alignment*) antara sinyal suara pengguna dan sinyal referensi. Dengan pendekatan ini, sistem tidak hanya melakukan klasifikasi ketepatan artikulasi, tetapi juga menghitung tingkat kemiripan temporal pelafalan terhadap referensi. Melalui kombinasi MFCC, CNN, dan DTW, sistem diharapkan mampu memberikan evaluasi artikulasi makhraj huruf hijaiyah berharakat fathah secara objektif, terukur, dan dapat digunakan sebagai alat bantu dalam proses pembelajaran tajwid.

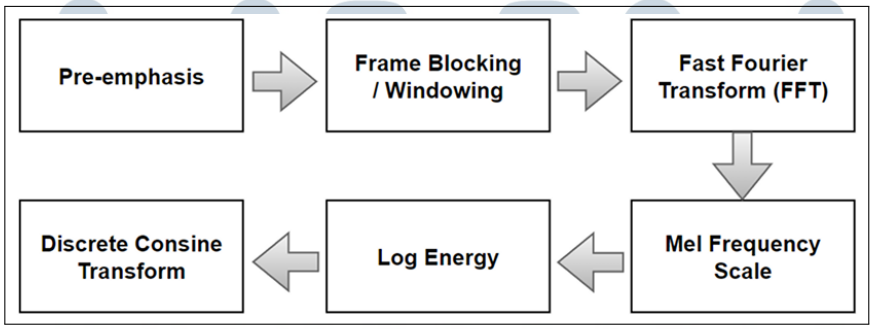
2.3 Mel-Frequency Cepstral Coefficients (MFCC)

Mel-Frequency Cepstral Coefficients (MFCC) merupakan metode ekstraksi fitur audio yang banyak digunakan dalam analisis sinyal suara untuk merepresentasikan karakteristik spektral ujaran secara numerik. Metode ini diperkenalkan oleh Steven B. Davis dan Paul Mermelstein pada tahun 1980 sebagai representasi parametrik sinyal suara yang mendekati karakteristik persepsi pendengaran manusia. MFCC bekerja dengan mengubah sinyal suara dari domain waktu ke domain frekuensi, kemudian memproyeksikannya ke dalam skala Mel, yaitu skala frekuensi yang bersifat linear di bawah 1000 Hz dan logaritmik di atas 1000 Hz.

Dalam berbagai penelitian modern, MFCC terbukti memberikan performa yang sangat baik sebagai fitur input pada model pembelajaran mesin. Studi

"Performance Analysis of Feature Mel Frequency Cepstral Coefficient and Short Time Fourier Transform Input for Lie Detection using Convolutional Neural Network" menunjukkan bahwa penggunaan MFCC sebagai metode ekstraksi fitur memberikan akurasi yang lebih tinggi dibandingkan Short Time Fourier Transform (STFT) ketika diproses menggunakan Convolutional Neural Network (CNN). Hasil penelitian tersebut menunjukkan akurasi sebesar 97,13% dengan nilai AUC 0,97 untuk MFCC, lebih tinggi dibandingkan STFT yang memperoleh akurasi 95,39% [17]. Temuan ini mengindikasikan bahwa MFCC lebih efektif dalam merepresentasikan karakteristik akustik yang relevan untuk proses klasifikasi berbasis CNN.

Penelitian lain berjudul "MFCC as Features for Speaker Classification using Machine Learning" juga menunjukkan bahwa MFCC merupakan fitur yang sangat baik untuk klasifikasi suara, termasuk identifikasi gender dan identifikasi pembicara. Dengan menggunakan sinyal suara berdurasi pendek (window 20–40 ms) dan sampling rate 16 kHz, MFCC mampu menghasilkan akurasi hingga 99,6% pada model klasifikasi seperti SVM, KNN, dan Decision Tree [18]. Hal ini menegaskan bahwa MFCC efektif dalam menangkap perbedaan karakteristik suara antar individu, termasuk perbedaan frekuensi dasar (pitch) dan pola spektral. Secara umum, proses ekstraksi MFCC terdiri dari beberapa tahapan utama yang dapat dilihat pada Gambar 2.3 sebagai berikut:



Gambar 2.3. Tahapan-tahapan MFCC Feature Extraction

Sumber: [19]

1. **Pre-emphasis**

   Tahap ini bertujuan untuk memperkuat komponen frekuensi tinggi pada sinyal suara guna menyeimbangkan spektrum frekuensi serta mengurangi efek redaman alami pada sinyal ujaran. Proses ini umumnya dilakukan menggunakan filter linear orde pertama.

$$y[n] = x[n] - \alpha x[n-1] \tag{2.1}$$

Pada persamaan tersebut, $x[n]$ merupakan sinyal suara asli pada sampel ke-n, sedangkan $y[n]$ merupakan sinyal hasil proses *pre-emphasis*. Parameter α merupakan koefisien *pre-emphasis* yang biasanya bernilai antara 0.95 hingga 0.97 untuk mengontrol tingkat penguatan frekuensi tinggi.

2. **Frame Blocking**

   Sinyal suara dibagi menjadi beberapa frame berdurasi pendek, umumnya antara 20–40 milidetik, dengan kemungkinan adanya overlap antarframe. Pembagian ini dilakukan karena sinyal ujaran dapat diasumsikan bersifat stasioner dalam interval waktu yang sangat singkat.

   Jika panjang frame adalah N dan overlap antar frame sebesar M, maka sinyal dapat direpresentasikan sebagai:

$$x_i[n] = x[n + iM] \tag{2.2}$$

   di mana $x_i[n]$ merupakan frame ke-i, $x[n]$ adalah sinyal asli, dan M merupakan jumlah pergeseran sampel antar frame.

3. **Windowing**

   Setelah sinyal dibagi menjadi beberapa frame berdurasi pendek, setiap frame dikalikan dengan fungsi jendela untuk mengurangi diskontinuitas pada tepi frame yang dapat menyebabkan *spectral leakage*. Salah satu fungsi jendela yang paling umum digunakan adalah *Hamming window*. Rumus 2.3 menunjukkan fungsi matematis dari *Hamming window*.

$$w[n] = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right) \tag{2.3}$$

   Pada persamaan tersebut, $w[n]$ merupakan nilai fungsi jendela pada sampel ke-n, N merupakan jumlah sampel dalam satu frame, dan n adalah indeks sampel dengan rentang $0 \leq n \leq N-1$.

4. **Fast Fourier Transform (FFT)**

   Setelah proses *windowing*, sinyal kemudian ditransformasikan dari domain waktu ke domain frekuensi menggunakan *Fast Fourier Transform* (FFT). Transformasi ini bertujuan untuk memperoleh spektrum frekuensi yang menunjukkan distribusi energi pada setiap komponen frekuensi sinyal. Secara matematis, proses ini didasarkan pada *Discrete Fourier Transform* (DFT) seperti yang ditunjukkan pada Rumus 2.4.

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-j2\pi kn/N} \tag{2.4}$$

   Pada persamaan tersebut, $X[k]$ merupakan representasi spektrum frekuensi pada indeks frekuensi ke-k, $x[n]$ merupakan sinyal pada domain waktu, dan N merupakan jumlah sampel dalam satu frame.

5. **Mel Frequency Scale**

   Selanjutnya, spektrum frekuensi linier dikonversi ke dalam skala Mel menggunakan sejumlah *Mel filter bank*. Skala Mel dirancang untuk meniru persepsi frekuensi pada sistem pendengaran manusia yang lebih sensitif terhadap perubahan frekuensi rendah dibandingkan frekuensi tinggi. Hubungan antara frekuensi linier dan skala Mel ditunjukkan pada Rumus 2.5.

$$Mel(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \tag{2.5}$$

   Pada persamaan tersebut, f merupakan frekuensi dalam satuan Hertz (Hz), sedangkan $Mel(f)$ merupakan frekuensi yang telah dikonversi ke dalam skala Mel. Frekuensi pada skala Mel kemudian diproses menggunakan sejumlah filter berbentuk segitiga yang dikenal sebagai *Mel filter bank*. Setiap filter menghitung energi pada rentang frekuensi tertentu sehingga menghasilkan representasi spektral yang lebih sesuai dengan persepsi pendengaran manusia.

6. **Log Energy**

   Energi yang dihasilkan oleh setiap filter Mel kemudian dikompresi menggunakan fungsi logaritma untuk menyesuaikan representasi energi dengan karakteristik persepsi intensitas suara manusia yang bersifat logaritmik. Proses ini dinyatakan pada Rumus 2.6.

$$S[m] = \log\left(\sum_{k=0}^{N-1} |X[k]|^2 H_m[k]\right) \tag{2.6}$$

Pada persamaan tersebut, $S[m]$ merupakan energi log dari filter Mel ke-m, $X[k]$ merupakan spektrum frekuensi hasil transformasi Fourier, dan $H_m[k]$ merupakan respons filter Mel ke-m.

7. **Discrete Cosine Transform (DCT)**
   Tahap terakhir dalam proses ekstraksi MFCC adalah penerapan *Discrete Cosine Transform* (DCT). Transformasi ini digunakan untuk mengubah hasil log energi filter Mel menjadi koefisien cepstral yang bersifat lebih terdekorrelasi sehingga lebih efektif digunakan sebagai fitur pada proses klasifikasi. Rumus 2.7 menunjukkan perhitungan koefisien MFCC menggunakan DCT.

$$c[n] = \sum_{m=1}^{M} S[m] \cos\left(\frac{\pi n}{M}(m - \frac{1}{2})\right) \tag{2.7}$$

   Pada persamaan tersebut, $c[n]$ merupakan koefisien MFCC ke-n, $S[m]$ merupakan log energi dari filter Mel ke-m, dan M merupakan jumlah filter bank yang digunakan dalam proses ekstraksi fitur.

Sebelum digunakan sebagai input model pembelajaran mesin, koefisien MFCC yang dihasilkan umumnya dinormalisasi terlebih dahulu untuk menghilangkan perbedaan skala antar koefisien yang dapat mengganggu proses pelatihan [20]. Salah satu metode normalisasi yang umum diterapkan adalah normalisasi *z-score*, yang mengubah setiap koefisien agar memiliki rata-rata nol dan deviasi standar satu. Normalisasi ini diterapkan secara per-koefisien, artinya setiap dimensi koefisien dinormalisasi secara independen terhadap distribusinya sendiri di sepanjang seluruh frame [20]. Rumus 2.8 menunjukkan formula normalisasi *z-score* pada koefisien MFCC.

$$\hat{c}[k,t] = \frac{c[k,t] - \mu_k}{\sigma_k + \varepsilon} \tag{2.8}$$

di mana $c[k,t]$ adalah nilai koefisien MFCC ke-k pada frame ke-t, μ_k dan σ_k masing-masing merupakan rata-rata dan deviasi standar koefisien ke-k pada seluruh

frame, dan ε adalah konstanta kecil bernilai 10^{-8} untuk mencegah pembagian dengan nol. Normalisasi ini memastikan bahwa tidak ada satu koefisien pun yang mendominasi proses pembelajaran akibat perbedaan rentang nilai.

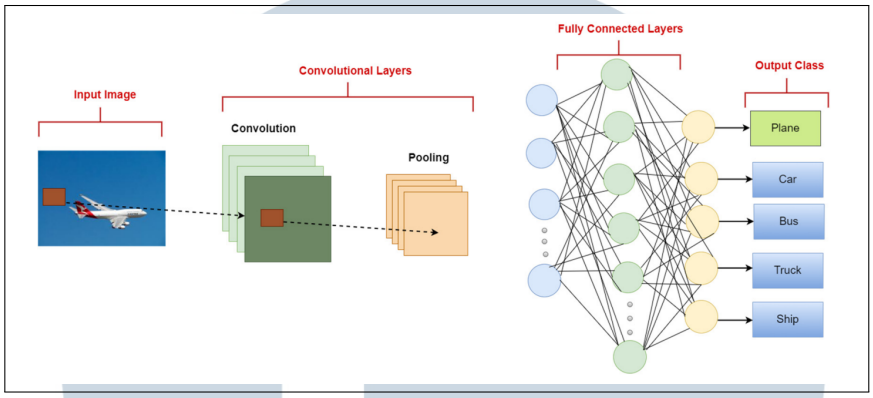
Hasil akhir dari proses MFCC berupa matriks koefisien (time × cepstral coefficients) yang sering divisualisasikan sebagai spektrogram 2D. Koefisien MFCC yang dihasilkan dari proses tersebut kemudian disusun dalam bentuk matriks dua dimensi (waktu × koefisien cepstral). Representasi ini dapat diperlakukan sebagai pola spasial yang menyerupai citra sehingga sesuai digunakan sebagai input pada model Convolutional Neural Network (CNN). Dalam penelitian ini, fitur MFCC digunakan sebagai representasi akustik dari pelafalan huruf hijaiyah berharakat fathah, yang selanjutnya dianalisis menggunakan CNN untuk klasifikasi pola artikulasi serta Dynamic Time Warping (DTW) untuk mengukur tingkat kemiripan temporal terhadap referensi pelafalan.

Pada umumnya, jumlah koefisien MFCC yang digunakan berkisar antara 12 hingga 13 koefisien utama pada setiap frame, tidak termasuk koefisien energi tambahan [21]. Nilai tersebut dianggap cukup untuk merepresentasikan karakteristik spektral utama dari sinyal suara. Dalam konteks penelitian ini, MFCC digunakan untuk merepresentasikan pola resonansi akustik yang dihasilkan oleh artikulasi makhraj huruf hijaiyah berharakat fathah. Karena setiap makhraj menghasilkan distribusi energi frekuensi yang berbeda akibat posisi artikulator (lidah, tenggorokan, bibir), maka MFCC mampu menangkap perbedaan tersebut dalam bentuk koefisien numerik. Representasi ini menjadi dasar analisis lebih lanjut oleh model CNN dan DTW dalam mengevaluasi tingkat akurasi artikulasi secara objektif.

2.4 Convolutional Neural Network (CNN)

Convolutional Neural Network (CNN) merupakan arsitektur *deep learning* yang dirancang untuk mengekstraksi pola spasial dari data berbentuk matriks, termasuk representasi dua dimensi dari sinyal suara seperti MFCC. Dalam penelitian ini, CNN digunakan bukan untuk melakukan transkripsi ujaran, melainkan untuk mengklasifikasikan tingkat akurasi artikulasi makhraj berdasarkan pola spektral yang dihasilkan. CNN awalnya banyak digunakan dalam klasifikasi citra, namun dalam beberapa tahun terakhir telah berkembang secara luas dalam bidang *speech recognition* karena kemampuannya menangkap pola lokal pada sinyal akustik [22].

Secara umum, CNN terdiri dari beberapa lapisan utama yang dapat dilihat pada Gambar 2.4, yaitu:



Gambar 2.4. Arsitektur Standar CNN

Sumber: [22]

1. **Convolution Layer**

Lapisan konvolusi merupakan komponen inti pada *Convolutional Neural Network* (CNN) yang berfungsi untuk mengekstraksi fitur lokal dari data masukan. Operasi konvolusi dilakukan dengan menggeser kernel (filter) pada matriks input sehingga menghasilkan *feature map*. Pada data audio yang telah direpresentasikan dalam bentuk spektrogram atau matriks MFCC, lapisan ini mampu mendeteksi pola frekuensi tertentu, seperti transisi formant, distribusi energi, atau karakteristik resonansi yang berkaitan dengan artikulasi huruf.

Secara matematis, operasi konvolusi pada citra atau matriks input dapat dituliskan sebagai berikut:

$$y(i,j) = \sum_m \sum_n x(i+m, j+n) \cdot w(m,n) \tag{2.9}$$

di mana x merupakan matriks input, w merupakan kernel atau filter konvolusi, dan y merupakan *feature map* yang dihasilkan.

Setiap filter pada *convolution layer* akan mempelajari pola spesifik, misalnya pola frekuensi rendah, perubahan energi secara tiba-tiba, atau tekstur spektral tertentu. Semakin dalam arsitektur jaringan, fitur yang dipelajari akan semakin kompleks dan abstrak.

2. **Activation Layer**

Activation layer diletakkan setelah *convolution layer* dan juga setelah *fully connected layer* (kecuali pada layer output tertentu). Secara umum, urutan operasi pada arsitektur CNN adalah sebagai berikut:

$$Convolution \rightarrow Activation \rightarrow Pooling$$

Fungsi aktivasi bertujuan untuk memperkenalkan sifat non-linear ke dalam jaringan. Tanpa *activation layer*, seluruh jaringan hanya akan menjadi transformasi linear, sehingga tidak mampu mempelajari pola kompleks pada data suara.

Fungsi aktivasi yang paling umum digunakan adalah *Rectified Linear Unit* (ReLU). Rumus 2.10 menunjukkan definisi matematis dari fungsi ReLU.

$$f(x) = \max(0, x) \tag{2.10}$$

Berdasarkan Persamaan 2.10, fungsi ReLU akan mengubah seluruh nilai negatif menjadi nol dan mempertahankan nilai positif apa adanya.

Keunggulan penggunaan ReLU antara lain sebagai berikut:

(a) Mengurangi masalah *vanishing gradient*.

(b) Mempercepat proses pelatihan model.

(c) Meningkatkan efisiensi komputasi.

Dalam konteks penelitian ini, *activation layer* membantu jaringan dalam menonjolkan pola frekuensi penting yang berkaitan dengan artikulasi makhraj, sekaligus menekan sinyal yang kurang relevan.

3. **Pooling Layer**

Pooling layer berfungsi untuk mengurangi dimensi *feature map* sekaligus mempertahankan informasi penting yang telah diekstraksi. Proses ini membantu mengurangi kompleksitas komputasi serta meningkatkan kemampuan generalisasi model. Beberapa metode *pooling* yang umum digunakan antara lain:

(a) **Max Pooling**

Memilih nilai maksimum dalam setiap jendela *pooling*. Metode ini mempertahankan fitur yang paling dominan dan sering memberikan performa yang baik pada tugas klasifikasi karena menekankan respons aktivasi terkuat.

(b) **Average Pooling**

Menghitung nilai rata-rata dalam setiap jendela *pooling*. Pendekatan ini lebih halus dalam menangani noise dan membantu proses generalisasi.

(c) **Min Pooling**

Memilih nilai minimum dalam jendela *pooling*. Metode ini umumnya digunakan pada kasus tertentu seperti deteksi anomali.

4. **Fully Connected Layer (Hidden Layer)**

Setelah melalui beberapa tahap konvolusi dan *pooling*, *feature map* yang dihasilkan diratakan (*flatten*) dan diteruskan ke *fully connected layer*. Lapisan ini berfungsi sebagai pengklasifikasi yang menggabungkan seluruh fitur hasil ekstraksi untuk menghasilkan prediksi akhir.

5. **Output Layer**

Lapisan akhir biasanya menggunakan fungsi aktivasi seperti *softmax* untuk menghasilkan distribusi probabilitas pada setiap kelas. Fungsi *softmax* secara matematis dirumuskan sebagai berikut:

$$P(y_i) = \frac{e^{z_i}}{\sum_{j=1}^{k} e^{z_j}} \tag{2.11}$$

di mana z_i merupakan nilai output neuron ke-i dan k merupakan jumlah kelas. Probabilitas tertinggi kemudian digunakan sebagai dasar penentuan label hasil klasifikasi.

Menurut penelitian "*CNN Based Automatic Speech Recognition: A Comparative Study*", CNN sangat efektif dalam menangkap dependensi jangka pendek pada sinyal suara, seperti pengenalan kata pendek atau fonem. Dalam penelitian tersebut, data suara berdurasi satu detik diubah menjadi spektrogram dan

digunakan sebagai input CNN untuk sistem pengenalan perintah suara. Model yang diusulkan mencapai akurasi sebesar 94,60%, menunjukkan bahwa CNN mampu mengidentifikasi pola lokal dalam domain waktu–frekuensi secara efektif [23].

Keunggulan CNN dalam konteks *speech recognition* terletak pada kemampuannya mendeteksi pola spasial pada representasi dua dimensi seperti spektrogram atau MFCC. Karena MFCC menghasilkan matriks koefisien (waktu $\times$ frekuensi), maka data tersebut dapat diperlakukan sebagai citra dua dimensi sehingga CNN dapat digunakan untuk mengekstraksi pola artikulasi secara otomatis tanpa memerlukan perancangan fitur manual tambahan.

Dalam penelitian ini, fitur MFCC dari setiap sampel suara direpresentasikan dalam bentuk matriks dua dimensi yang kemudian diproses oleh CNN untuk mengekstraksi pola spektral khas dari setiap huruf hijaiyah berharakat fathah. Lapisan konvolusi berperan dalam mendeteksi pola frekuensi yang berkaitan dengan resonansi makhraj, sedangkan *pooling layer* membantu mereduksi dimensi data dengan tetap mempertahankan fitur dominan. Selanjutnya, *fully connected layer* melakukan proses klasifikasi untuk menentukan apakah artikulasi yang dihasilkan sesuai dengan referensi pelafalan yang benar atau tidak.

2.5 Pencarian Hyperparameter (*Grid Search*)

Hyperparameter merupakan parameter konfigurasi model yang nilainya ditetapkan sebelum proses pelatihan dimulai dan tidak dipelajari secara otomatis oleh model dari data. Berbeda dengan bobot jaringan saraf yang diperbarui melalui *backpropagation*, hyperparameter seperti *learning rate* dan *dropout rate* harus ditentukan secara eksplisit oleh peneliti. Pemilihan nilai hyperparameter yang tepat sangat berpengaruh terhadap kecepatan konvergensi model, kemampuan generalisasi, dan performa akhir pada data uji [23].

Grid Search merupakan metode pencarian hyperparameter yang sistematis dengan cara mencoba seluruh kombinasi dari sekumpulan nilai kandidat yang telah ditentukan sebelumnya. Setiap kombinasi dilatih dan dievaluasi, kemudian kombinasi dengan performa terbaik dipilih sebagai konfigurasi final. Pendekatan ini menjamin bahwa seluruh ruang pencarian yang didefinisikan dieksplorasi secara menyeluruh, sehingga hasil yang diperoleh bersifat reprodusibel dan dapat dipertanggungjawabkan secara ilmiah [23].

Secara umum, proses *Grid Search* dapat dirumuskan sebagai pencarian nilai optimal dari fungsi performa model f terhadap kombinasi hyperparameter θ pada

himpunan kandidat Θ:

$$\theta^* = \arg\max_{\theta \in \Theta} f(\theta) \tag{2.12}$$

di mana θ^* adalah kombinasi hyperparameter terbaik, Θ adalah ruang pencarian yang terdiri dari seluruh kombinasi nilai kandidat, dan $f(\theta)$ adalah fungsi evaluasi performa model, dalam penelitian ini menggunakan akurasi validasi (*val_accuracy*).

2.6 Dynamic Time Warping (DTW)

Dynamic Time Warping (DTW) merupakan algoritma yang digunakan untuk mengukur tingkat kemiripan antara dua urutan data yang memiliki variasi dimensi waktu. Dalam konteks evaluasi artikulasi, DTW digunakan untuk menyelaraskan pola temporal antara sinyal suara uji dan sinyal referensi sehingga perbedaan tempo pengucapan tidak langsung memengaruhi hasil penilaian.

Menurut penelitian "Application of Dynamic Time Warping Optimization Algorithm in Speech Recognition of Machine Translation", DTW digunakan sebagai metode pencocokan pola dalam sistem pengenalan suara dengan tingkat akurasi yang tinggi. Dalam lingkungan tenang, sistem berbasis DTW mampu mencapai tingkat pengenalan di atas 93,85%, sedangkan dalam kondisi bising tetap mempertahankan akurasi di atas 91,4%. Selain tingkat pengenalan yang tinggi, DTW juga memiliki kecepatan komputasi yang relatif baik untuk kosakata terbatas [24]. Hal ini menunjukkan bahwa DTW efektif untuk sistem pengenalan suara berbasis template yang tidak memerlukan dataset besar.

Secara umum, cara kerja DTW dalam pengenalan suara melibatkan beberapa tahapan utama:

1. **Preprocessing dan Ekstraksi Fitur**

   Sinyal suara terlebih dahulu melalui tahap prapemrosesan seperti proses *sampling*, *framing*, dan normalisasi. Selanjutnya dilakukan ekstraksi fitur, misalnya menggunakan *Mel-Frequency Cepstral Coefficients* (MFCC), untuk menghasilkan representasi numerik yang lebih ringkas dan informatif. Representasi ini memungkinkan perbandingan antar sinyal dilakukan secara lebih efektif.

2. **Pembuatan Matriks Jarak**

Setiap elemen atau vektor fitur dari sinyal uji dibandingkan dengan elemen pada sinyal referensi untuk membentuk matriks jarak (*distance matrix*). Perhitungan jarak umumnya menggunakan metode jarak Euclidean. Secara matematis, jarak antara dua vektor fitur x_i dan y_j dapat dinyatakan sebagai berikut:

$$d(x_i, y_j) = \sqrt{\sum_{k=1}^{n} (x_{ik} - y_{jk})^2} \tag{2.13}$$

di mana x_i dan y_j merupakan vektor fitur pada frame ke-i dan ke-j, serta n adalah jumlah dimensi fitur.

3. **Penentuan *Warping Path* Optimal**

Algoritma DTW mencari jalur dengan total jarak minimum melalui matriks jarak menggunakan pendekatan *dynamic programming*. Jalur ini disebut sebagai *optimal warping path*, yang menggambarkan penyelarasan terbaik antara dua sinyal dengan mempertimbangkan kemungkinan perbedaan durasi atau kecepatan pengucapan.

Perhitungan akumulasi jarak minimum pada DTW dirumuskan sebagai berikut:

$$D(i, j) = d(i, j) + \min \begin{cases} D(i-1, j) \\ D(i, j-1) \\ D(i-1, j-1) \end{cases} \tag{2.14}$$

di mana $D(i, j)$ merupakan jarak kumulatif minimum hingga titik (i, j) pada matriks jarak, sedangkan $d(i, j)$ adalah jarak lokal antara dua frame fitur.

4. **Perhitungan Nilai Jarak Minimum**

Total akumulasi jarak dari jalur optimal digunakan sebagai ukuran kemiripan antara dua sinyal. Semakin kecil nilai jarak yang diperoleh, semakin tinggi

tingkat kemiripan antara sinyal uji dan sinyal referensi. Nilai ini kemudian dapat digunakan sebagai dasar pengambilan keputusan dalam proses evaluasi atau klasifikasi.

Keunggulan utama DTW adalah kemampuannya untuk "meregangkan" atau "memampatkan" dimensi waktu pada sinyal uji agar selaras dengan sinyal referensi. Dengan demikian, variasi tempo atau durasi pengucapan tidak langsung menyebabkan kesalahan pengenalan.

Penelitian "Implementasi Speech Recognition pada Smart-Home yang menggunakan algoritma Dynamic Time Warping dan Deep Neural Network" menunjukkan bahwa DTW dapat digunakan sebagai metode pencocokan awal, namun performanya meningkat secara signifikan ketika dikombinasikan dengan model pembelajaran mendalam seperti Deep Neural Network (DNN). Dalam penelitian tersebut, penggunaan DTW saja menghasilkan akurasi sebesar 59,26%, tetapi setelah dikombinasikan dengan DNN, akurasi meningkat menjadi 96,92% [25]. Temuan ini menunjukkan bahwa DTW sangat efektif sebagai mekanisme pembanding pola temporal, terutama ketika dikombinasikan dengan model klasifikasi berbasis pembelajaran mesin.

Dalam konteks penelitian ini, DTW tidak digunakan sebagai *classifier* utama, melainkan sebagai mekanisme evaluasi kesesuaian pola temporal antara suara uji dan suara referensi makhraj fathah yang benar. Mengingat fathah merupakan vokal pendek dengan durasi singkat dan sensitif terhadap variasi tempo pengucapan, pendekatan berbasis klasifikasi seperti CNN saja belum tentu mampu menangkap pergeseran temporal secara eksplisit.

Oleh karena itu, dalam penelitian ini CNN (Metode A) dan DTW (Metode B) dijalankan secara **terpisah dan independen** sebagai dua metode evaluasi yang dibandingkan. MFCC dan CNN berfungsi untuk mengklasifikasikan ketepatan artikulasi berdasarkan pola spektral, sedangkan DTW menghitung jarak kemiripan antara representasi fitur suara dan template referensi untuk mengevaluasi kesesuaian pola temporal. Dengan membandingkan kedua metode secara mandiri, sistem dapat memberikan perspektif evaluasi yang saling melengkapi: CNN mengukur apakah pola frekuensi huruf yang diucapkan sesuai dengan karakteristik spektral kelas yang benar, sementara DTW mengukur seberapa dekat pola dinamis pelafalan terhadap template referensi makhraj.

2.7 Metrik Evaluasi Klasifikasi

Evaluasi performa sistem klasifikasi dilakukan menggunakan metrik kuantitatif yang mengukur seberapa baik model memprediksi kelas yang benar dibandingkan dengan label sebenarnya. Metrik-metrik ini diturunkan dari *confusion matrix*, yaitu tabel yang merangkum hasil prediksi model terhadap data uji. Dalam *confusion matrix*, terdapat empat elemen utama yang menjadi dasar perhitungan seluruh metrik, yaitu *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN).

Pada konteks klasifikasi multi-kelas seperti dalam penelitian ini yang melibatkan 28 huruf hijaiyah, metrik dihitung secara per-kelas menggunakan pendekatan *one-vs-rest*, di mana setiap huruf diperlakukan sebagai kelas positif dan seluruh huruf lainnya sebagai kelas negatif. Empat metrik utama yang digunakan adalah sebagai berikut.

Akurasi merupakan metrik paling dasar yang mengukur proporsi prediksi yang benar terhadap keseluruhan sampel. Rumus 2.15 menunjukkan formula perhitungan akurasi.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.15}$$

Precision mengukur seberapa tepat prediksi positif yang dihasilkan oleh model, yaitu proporsi prediksi kelas tertentu yang benar-benar merupakan kelas tersebut. Nilai *precision* yang tinggi menunjukkan bahwa model jarang mengklasifikasikan kelas lain sebagai kelas yang dimaksud. Rumus 2.16 menunjukkan formula perhitungan *precision*.

$$Precision = \frac{TP}{TP + FP} \tag{2.16}$$

Recall mengukur kemampuan model dalam menemukan seluruh sampel yang sebenarnya termasuk kelas tertentu, yaitu proporsi sampel positif yang berhasil diidentifikasi dengan benar. Nilai *recall* yang tinggi menunjukkan bahwa model jarang melewatkan sampel dari kelas yang dimaksud. Rumus 2.17 menunjukkan formula perhitungan *recall*.

$$Recall = \frac{TP}{TP + FN} \tag{2.17}$$

F1-Score merupakan rata-rata harmonik dari *precision* dan *recall* yang memberikan keseimbangan antara keduanya. Metrik ini sangat berguna pada kasus klasifikasi dengan distribusi kelas yang tidak seimbang, di mana mengoptimalkan salah satu metrik saja dapat mengorbankan metrik yang lain. Rumus 2.18 menunjukkan formula perhitungan *F1-Score*.

$$F1\text{-}Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{2.18}$$

di mana TP adalah *True Positive* (sampel kelas tertentu yang diprediksi benar), FP adalah *False Positive* (sampel kelas lain yang salah diprediksi sebagai kelas ini), dan FN adalah *False Negative* (sampel kelas ini yang salah diprediksi sebagai kelas lain).

Dalam penelitian ini, keempat metrik tersebut digunakan untuk mengukur performa model CNN pada setiap kelas huruf hijaiyah secara individual, sehingga dapat diidentifikasi huruf-huruf mana yang lebih sulit diklasifikasikan oleh sistem. Analisis per-kelas ini penting mengingat setiap huruf hijaiyah memiliki karakteristik akustik yang berbeda-beda dengan tingkat kesulitan yang tidak seragam.

2.8 Penelitian Terdahulu

Penelitian terkait pengenalan suara berbasis MFCC, CNN, dan DTW telah banyak dilakukan pada berbagai domain. Bagian ini membahas 15 penelitian terdahulu yang relevan dengan penelitian yang dilakukan, mencakup sistem evaluasi tajwid, pengenalan suara berbasis pembelajaran mendalam, pencocokan pola temporal, serta penilaian pelafalan otomatis.

2.8.1 *Identifying the Correct Articulation Point of a Quranic Letter of the Throat (Al-Halqu) Makhraj*

Penelitian oleh [8] membahas pengembangan sistem berbasis *Artificial Intelligence* untuk mendeteksi ketepatan makhraj huruf Al-Qur'an pada area tenggorokan (*Al-Halqu*), khususnya huruf-huruf yang termasuk dalam izhar halqi. Sistem dikembangkan menggunakan MFCC sebagai metode ekstraksi fitur dan *Convolutional Neural Network* (CNN) sebagai model klasifikasi. Dataset

dikumpulkan melalui platform SanaAlTajweed yang berisi rekaman pelafalan huruf hijaiyah dari berbagai pembicara. Sebelum pelatihan, data melalui tahap augmentasi dan preprocessing untuk meningkatkan keragaman dan kualitas fitur. Evaluasi dilakukan menggunakan metrik akurasi dan hasil penelitian menunjukkan bahwa model CNN mampu mencapai akurasi sebesar 89,39%. Penelitian ini membuktikan bahwa CNN efektif dalam mengekstraksi pola fonetik dari representasi MFCC dua dimensi untuk keperluan klasifikasi makhraj. Namun demikian, penelitian ini memiliki keterbatasan pada cakupan huruf yang dievaluasi, yaitu hanya huruf-huruf dari makhraj tenggorokan tanpa mencakup seluruh 28 huruf hijaiyah. Selain itu, sistem hanya menggunakan klasifikasi berbasis CNN tanpa mempertimbangkan evaluasi kesesuaian temporal secara eksplisit. Variasi durasi dan tempo pelafalan antar pembicara belum ditangani melalui mekanisme *alignment* temporal seperti DTW.

2.8.2 *Implementation of Audio Recognition Using MFCC and DTW in Wirama Praharsini*

Penelitian oleh [10] mengimplementasikan MFCC dan DTW untuk melakukan pencocokan suara pada wirama praharsini, yaitu karya seni suara klasik Bali yang dilantunkan dalam upacara tradisional. Sistem dirancang untuk mengevaluasi ketepatan pelafalan pengguna terhadap rekaman referensi yang dibuat oleh ahli seni tradisional. Proses kerja sistem dimulai dengan ekstraksi fitur MFCC dari sinyal audio pengguna dan audio referensi. Selanjutnya, DTW digunakan untuk mengukur kemiripan pola temporal antara kedua sinyal tersebut. Pengujian dilakukan terhadap berbagai rekaman dari pengguna dengan tingkat kemahiran yang berbeda-beda. Hasil penelitian menunjukkan akurasi sebesar 88,89%, dengan DTW terbukti mampu menangani variasi tempo antar pembicara secara efektif. Keterbatasan utama penelitian ini adalah tidak digunakannya model *deep learning* seperti CNN, sehingga kemampuan sistem dalam mempelajari pola spektral kompleks masih terbatas. Sistem sepenuhnya bergantung pada template matching berbasis jarak DTW, yang kurang adaptif terhadap variasi distribusi frekuensi yang tidak tertangkap oleh template.

2.8.3 *End-to-end Elevator Voice Command Recognition System*

Penelitian oleh [12] mengembangkan sistem pengenalan perintah suara untuk pengoperasian lift menggunakan arsitektur *deep neural network* berbasis CNN yang dirancang secara *end-to-end*. Motivasi penelitian ini adalah kebutuhan pengoperasian lift tanpa sentuhan (*touchless*) pascapandemi untuk meminimalisasi risiko penularan penyakit melalui permukaan. Sistem menggunakan korpus suara yang dibangun secara mandiri dan mengimplementasikan algoritma pemotongan segmen suara (*speech segment interception*) untuk mendapatkan segmen audio dari aliran suara waktu nyata. Segmen tersebut langsung diproses oleh model CNN untuk inferensi tanpa tahap preprocessing terpisah. Hasil penelitian menunjukkan peningkatan akurasi sekitar 10% dibandingkan model MFCC-CNN konvensional dan sekitar 25% dibandingkan metode DTW tradisional. Penelitian ini menunjukkan bahwa arsitektur CNN yang dioptimalkan dapat mengungguli DTW pada tugas pengenalan suara berbasis perintah pendek. Keterbatasan penelitian ini terletak pada domain aplikasinya yang spesifik pada pengenalan perintah suara dalam lingkungan tertutup, bukan pada evaluasi ketepatan artikulasi fonetik. Sistem tidak dirancang untuk menganalisis ketepatan makhraj atau mendeteksi perbedaan fonetik yang halus.

2.8.4 *Estimasi Pitch Menggunakan CNN dan DTW pada Rekaman Suara Penyanyi*

Penelitian oleh [13] mengombinasikan CNN satu dimensi dan DTW untuk mengevaluasi ketepatan intonasi pada rekaman suara penyanyi. CNN digunakan untuk melakukan estimasi *pitch* dari sinyal audio, kemudian hasil estimasi berupa kontur nada dibandingkan dengan template referensi menggunakan DTW untuk menghasilkan skor kemiripan intonasi. Kontribusi utama penelitian ini adalah demonstrasi bahwa kombinasi CNN dan DTW dapat saling melengkapi: CNN menangkap representasi fitur akustik melalui pembelajaran, sementara DTW menyediakan mekanisme perbandingan pola temporal yang fleksibel. Sistem mampu menghasilkan rata-rata akurasi tertinggi sebesar 97,425% dengan nilai threshold DTW optimal pada rentang distance 1000–1500. Penelitian ini menjadi referensi konseptual utama bagi penelitian yang sedang dilakukan, karena secara arsitektural memiliki kesamaan dalam penggunaan CNN untuk ekstraksi fitur dan DTW untuk evaluasi kesesuaian pola. Perbedaan mendasar dengan penelitian

yang sedang dilakukan terletak pada fokus evaluasi: penelitian Pratama et al. berfokus pada aspek *suprasegmental* berupa kesesuaian tinggi-rendah nada (*pitch*), sedangkan penelitian ini berfokus pada aspek *segmental* berupa ketepatan tempat artikulasi (*makhraj*) yang tercermin pada distribusi energi spektral.

2.8.5 *MFCC as Features for Speaker Classification using Machine Learning*

Penelitian oleh [18] mengembangkan sistem klasifikasi pembicara menggunakan fitur MFCC yang diproses dengan beberapa algoritma *machine learning* konvensional, yaitu *Support Vector Machine* (SVM), *K-Nearest Neighbor* (KNN), dan *Decision Tree*. Penelitian ini secara sistematis mengevaluasi efektivitas MFCC dalam merepresentasikan karakteristik suara individu. Dengan menggunakan sinyal suara berdurasi pendek (window 20–40 ms) dan *sampling rate* 16 kHz, MFCC mampu menghasilkan akurasi hingga 99,6% untuk tugas identifikasi gender dan pembicara. Hasil ini membuktikan bahwa MFCC sangat efektif dalam menangkap perbedaan karakteristik akustik antar individu, termasuk perbedaan frekuensi dasar (*pitch*) dan pola formant. Penelitian ini memperkuat dasar pemilihan MFCC sebagai metode ekstraksi fitur utama dalam penelitian yang sedang dilakukan. Keterbatasan utama penelitian ini adalah penggunaan algoritma klasifikasi konvensional (SVM, KNN, Decision Tree) yang tidak mampu mengekstraksi fitur hierarkis seperti CNN, sehingga potensi representasi MFCC belum dimanfaatkan secara maksimal.

2.8.6 *Implementasi Speech Recognition pada Smart-Home Menggunakan DTW dan DNN*

Penelitian oleh [25] mengembangkan sistem pengenalan suara berbasis *Raspberry Pi* untuk kontrol perangkat rumah pintar dengan menggabungkan DTW dan *Deep Neural Network* (DNN). Penelitian ini secara eksplisit membandingkan performa DTW murni versus kombinasi DTW dan DNN. Metodologi penelitian mencakup dua tahap: pertama, pengujian sistem menggunakan DTW saja sebagai *baseline*; kedua, pengembangan sistem hibrida yang menggabungkan DTW untuk *alignment* temporal dan DNN untuk klasifikasi pola. Hasil menunjukkan kontras yang signifikan: DTW saja menghasilkan akurasi 59,26%, sedangkan kombinasi DTW dan DNN meningkat menjadi 96,92%. Temuan ini secara empiris membuktikan bahwa DTW memerlukan komponen pembelajaran mesin untuk

mencapai performa yang kompetitif, dan bahwa kombinasi keduanya memberikan hasil yang jauh lebih baik dibandingkan masing-masing metode secara tunggal. Meski demikian, penelitian ini berfokus pada pengenalan perintah suara dalam bahasa yang tidak spesifik pada fonetik Arab, bukan pada evaluasi ketepatan artikulasi makhraj.

2.8.7 *Pronunciation Error Detection Using MFCC and Modified VGG-16*

Penelitian ini mengembangkan sistem deteksi kesalahan pelafalan huruf *'ain* (ε) menggunakan kombinasi MFCC sebagai ekstraktor fitur dan CNN berbasis arsitektur *Modified VGG-16* sebagai model klasifikasi. Arsitektur VGG-16 yang dimodifikasi dipilih karena kemampuannya dalam mengekstraksi fitur hierarkis yang dalam (*deep features*) dari representasi dua dimensi. Penelitian ini secara khusus mengatasi tantangan yang ada pada huruf *'ain* yang memiliki karakteristik fonetik unik dan sering menjadi sumber kesalahan bagi penutur non-Arab. Dataset mencakup variasi pelafalan dari berbagai pembicara dengan harakat yang berbeda-beda. Hasil penelitian menunjukkan akurasi pelatihan hingga 100% dan akurasi validasi hingga 96%, membuktikan bahwa CNN dengan arsitektur yang memadai sangat efektif dalam membedakan variasi pelafalan huruf hijaiyah. Keterbatasan penelitian ini adalah cakupannya yang hanya pada satu huruf tertentu, sehingga belum dapat merepresentasikan sistem evaluasi yang komprehensif untuk seluruh 28 huruf hijaiyah. Selain itu, analisis temporal berbasis DTW belum diintegrasikan.

2.8.8 *Development of Acoustic Scene Classification Using CNN on Reduced DCASE Dataset*

Penelitian oleh [11] mengembangkan model klasifikasi akustik menggunakan MFCC dan CNN pada dataset DCASE (*Detection and Classification of Acoustic Scenes and Events*) yang direduksi. Penelitian ini mengeksplorasi efektivitas CNN dalam mengklasifikasikan berbagai jenis lingkungan akustik seperti bandara, pusat perbelanjaan, dan transportasi publik. Sistem menggunakan MFCC sebagai representasi fitur akustik utama yang kemudian diproses oleh model CNN untuk klasifikasi. Penelitian menunjukkan bahwa CNN mampu meningkatkan akurasi klasifikasi secara signifikan dibandingkan model *baseline* yang menggunakan metode klasifikasi konvensional, meski dilatih pada dataset yang lebih kecil. Kontribusi utama penelitian ini adalah demonstrasi kemampuan

generalisasi CNN pada tugas klasifikasi sinyal audio dengan keragaman kelas yang tinggi, yang relevan dengan tantangan mengklasifikasikan 28 kelas huruf hijaiyah dalam penelitian ini.

2.8.9 *CNN Based Automatic Speech Recognition: A Comparative Study*

Penelitian oleh [23] melakukan studi komparatif terhadap berbagai varian arsitektur CNN untuk tugas pengenalan suara otomatis. Penelitian ini secara sistematis membandingkan konfigurasi arsitektur yang berbeda, termasuk variasi jumlah lapisan konvolusi, ukuran filter, dan mekanisme regularisasi. Data suara berdurasi satu detik dikonversi menjadi spektrogram dan digunakan sebagai input CNN. Studi ini juga menganalisis pengaruh hyperparameter seperti *learning rate* dan *dropout rate* terhadap konvergensi dan akurasi akhir model. Model yang diusulkan mencapai akurasi sebesar 94,60%. Lebih relevan lagi, penelitian ini menyimpulkan bahwa *Grid Search* pada hyperparameter kunci dapat meningkatkan performa model secara signifikan dibandingkan pemilihan hyperparameter secara intuitif. Temuan ini menjadi landasan metodologis bagi penerapan *Grid Search* dalam penelitian yang sedang dilakukan.

2.8.10 *Application of DTW Optimization Algorithm in Speech Recognition of Machine Translation*

Penelitian oleh [24] mengkaji implementasi DTW yang dioptimalkan untuk sistem pengenalan suara dalam konteks penerjemahan mesin. Penelitian ini secara khusus mengevaluasi performa DTW pada kondisi lingkungan yang bervariasi, dari kondisi tenang hingga lingkungan bising. Sistem menggunakan MFCC sebagai representasi fitur dan DTW sebagai algoritma pencocokan pola. Hasil penelitian menunjukkan bahwa dalam lingkungan tenang, DTW mampu mencapai tingkat pengenalan di atas 93,85%, sedangkan dalam kondisi bising tetap mempertahankan akurasi di atas 91,4%. Penelitian ini juga memperkenalkan optimasi DTW berbasis *Sakoe-Chiba band* untuk membatasi jalur warping dan meningkatkan efisiensi komputasi. Ketangguhan DTW dalam berbagai kondisi lingkungan ini memperkuat dasar penggunaannya sebagai komponen evaluasi temporal dalam penelitian yang sedang dilakukan.

2.8.11 *Performance Analysis of MFCC and STFT for Lie Detection Using CNN*

Penelitian oleh [17] melakukan analisis komparatif antara MFCC dan *Short-Time Fourier Transform* (STFT) sebagai metode ekstraksi fitur untuk deteksi kebohongan berbasis sinyal suara menggunakan CNN. Metodologi penelitian mencakup pengujian sistematis kedua metode ekstrasi fitur pada arsitektur CNN yang identik, sehingga perbedaan performa dapat diatribusikan secara langsung pada perbedaan kualitas representasi fitur. Hasil penelitian menunjukkan MFCC menghasilkan akurasi 97,13% dengan AUC 0,97, lebih tinggi dibandingkan STFT yang memperoleh akurasi 95,39%. Perbedaan ini dikaitkan dengan kemampuan MFCC dalam merepresentasikan karakteristik perseptual suara manusia secara lebih akurat melalui skala Mel. Temuan ini memperkuat justifikasi pemilihan MFCC sebagai metode ekstraksi fitur utama dalam penelitian ini.

2.8.12 *Automatic Pronunciation Assessment – A Review*

Penelitian oleh [15] merupakan survei komprehensif terhadap sistem penilaian pelafalan otomatis (*Automatic Pronunciation Assessment* / APA) berbasis kecerdasan buatan. Penelitian ini mengkaji berbagai pendekatan yang telah dikembangkan dalam dekade terakhir, mulai dari sistem berbasis *Hidden Markov Model* (HMM) hingga sistem berbasis *deep learning* terkini. Survei ini mengidentifikasi dua fungsi utama sistem APA, yaitu *pronunciation teaching* (membimbing perbaikan pelafalan) dan *pronunciation assessment* (menilai dan mendeteksi kesalahan). Penelitian ini juga menekankan bahwa APA lebih kompleks dari sekadar *Automatic Speech Recognition* (ASR), karena harus mampu mendeteksi perbedaan fonetik yang sangat halus. Konsep *score-level fusion* yang dibahas dalam survei ini menjadi landasan teori bagi mekanisme kombinasi skor CNN dan DTW dalam penelitian yang sedang dilakukan. Survei ini juga menyoroti kesenjangan penelitian pada bahasa selain bahasa Inggris, khususnya bahasa Arab dan sistem tajwid.

2.8.13 *AI-Driven Pronunciation Assessment: The Impact of SpeechAce on EFL Learners*

Penelitian oleh [16] mengkaji dampak sistem evaluasi pelafalan berbasis AI (*SpeechAce*) terhadap kemampuan pelafalan pelajar bahasa Inggris sebagai bahasa asing (EFL). Penelitian ini menggunakan desain *quasi-experimental*

dengan kelompok kontrol dan kelompok eksperimen untuk mengukur efektivitas umpan balik otomatis dari sistem AI. Hasil penelitian menunjukkan bahwa pelajar yang menggunakan SpeechAce mengalami peningkatan signifikan pada tiga dimensi pelafalan: *accuracy*, *fluency*, dan *completeness*. Penelitian ini juga mendokumentasikan bahwa umpan balik berbasis AI yang diberikan secara konsisten dan objektif lebih efektif dalam memotivasi pelajar dibandingkan umpan balik subjektif dari pengajar manusia. Temuan ini memberikan dukungan empiris terhadap nilai sistem evaluasi pelafalan otomatis yang dikembangkan dalam penelitian ini, khususnya untuk konteks pembelajaran mandiri.

2.8.14 *The Classification of Actual Pronunciation of Quranic Alphabets for Non-Arab Speaker*

Penelitian oleh [14] mengkaji dan mengklasifikasikan pola pelafalan huruf Al-Qur'an oleh penutur non-Arab dari berbagai latar belakang bahasa ibu. Penelitian ini mengidentifikasi pola kesalahan yang konsisten pada kelompok penutur tertentu berdasarkan kesamaan sistem fonetik bahasa ibu mereka. Penelitian ini mengembangkan taksonomi kesalahan pelafalan yang terstruktur, mengklasifikasikan kesalahan ke dalam kategori substitusi (*substitution*), penambahan (*insertion*), dan penghilangan (*deletion*) bunyi. Studi tersebut menegaskan bahwa hingga saat ini belum banyak penelitian yang secara komprehensif mengidentifikasi ketepatan pelafalan seluruh 28 huruf hijaiyah pada penutur non-*native*, sehingga masih terdapat kesenjangan signifikan yang perlu diatasi. Temuan ini memperkuat relevansi dan urgensi penelitian yang sedang dilakukan, yang secara spesifik mengevaluasi seluruh 28 huruf hijaiyah berharakat fathah.

2.8.15 *Penerapan CNN untuk Estimasi Pitch: Studi pada Domain Musik*

Penelitian dalam domain musik yang mengkaji estimasi *pitch* menggunakan CNN satu dimensi menunjukkan bahwa arsitektur CNN dapat diadaptasi untuk berbagai tugas analisis sinyal audio di luar klasifikasi suara konvensional. Dalam penelitian tersebut, CNN digunakan untuk mempelajari representasi fitur akustik dari rekaman vokal, kemudian output CNN berupa estimasi kontur *pitch* dibandingkan dengan template referensi menggunakan DTW. Relevansi penelitian ini terhadap penelitian yang sedang dilakukan terletak pada paradigma

arsitekturalnya: CNN sebagai komponen pembelajaran fitur dan DTW sebagai komponen evaluasi kesesuaian pola. Meskipun fokus penelitian tersebut adalah pada aspek *suprasegmental* (tinggi-rendah nada), prinsip kombinasi CNN-DTW yang dibuktikannya secara konseptual dapat diadaptasi untuk evaluasi aspek *segmental* berupa ketepatan makhraj, dengan perbedaan pada jenis fitur MFCC yang diekstraksi dan dimensi pola yang dibandingkan.

Tabel 2.1. Perbandingan penelitian terdahulu (No. 1–7)

No	Peneliti	Judul	Metode	Hasil	Fokus Penelitian	Perbedaan dengan Penelitian Ini
1	Ahmad et al. (2023) [8]	Identifying the Correct Articulation Point of a Quranic Letter of the Throat (Al-Halqu) Makhraj	MFCC + CNN	Acc 89,39%	Deteksi makhraj huruf halqi	Hanya 6 huruf, tidak ada evaluasi temporal DTW
2	Wibawa & Darmawan (2021) [10]	Implementation of Audio Recognition Using MFCC and DTW in Wirama Praharsini	MFCC + DTW	Acc 88,89%	Pencocokan suara wirama Bali	Tidak menggunakan *deep learning*, domain seni tradisional
3	Liu et al. (2024) [12]	End-to-end Elevator Voice Command Recognition System Based on Improved CNN	Improved CNN	+25% vs DTW	Pengenalan perintah suara lift	Tidak mengevaluasi artikulasi fonetik, bukan konteks tajwid
4	Pratama et al. (2021) [13]	Penerapan CNN untuk Melakukan Estimasi Pitch pada Rekaman Suara Penyanyi	CNN 1D + DTW	Acc 97,42%	Estimasi *pitch* rekaman penyanyi	Fokus *suprasegmental* (pitch), bukan makhraj
5	Mu & Min (2023) [18]	MFCC as Features for Speaker Classification using Machine Learning	MFCC + SVM/KNN/DT	Acc 99,6%	Identifikasi gender dan pembicara	Tidak mengevaluasi makhraj, bukan domain tajwid
6	Jurnal Smart Home (2025) [25]	Implementasi Speech Recognition pada Smart-Home Menggunakan DTW dan DNN	MFCC + DTW + DNN	Acc 96,92%	Perintah suara *smart-home*	Tidak fokus pada fonetik Arab atau tajwid
7	Makhraj *'Ain* VGG-16	Pronunciation Error Detection Using MFCC and Modified VGG-16	MFCC + CNN (VGG-16)	Val Acc 96%	Deteksi kesalahan huruf *'ain*	Hanya satu huruf, tidak ada komponen DTW

Tabel 2.2. Perbandingan penelitian terdahulu (No. 8–15)

No	Peneliti	Judul	Metode	Hasil	Fokus Penelitian	Perbedaan dengan Penelitian Ini
8	Singh et al. (2023) [11]	Development of Acoustic Scene Classification Model Using Neural Networks on Reduced DCASE Dataset	MFCC + CNN	Peningkatan vs *baseline*	Klasifikasi akustik lingkungan (DCASE)	Domain DCASE, bukan fonetik Arab
9	Ilgaz et al. (2024) [23]	CNN Based Automatic Speech Recognition: A Comparative Study	CNN komparatif	Acc 94,60%	Pengenalan suara otomatis komparatif	Pengenalan kata umum, bukan evaluasi makhraj
10	Jiang & Chen (2023) [24]	Application of DTW Optimization Algorithm in Speech Recognition of Machine Translation	DTW teroptimasi	>93,85%	Pengenalan suara untuk mesin terjemah	Hanya DTW, tanpa komponen CNN
11	Kusumawati et al. (2024) [17]	Performance Analysis of MFCC and STFT Input for Lie Detection Using CNN	MFCC vs STFT + CNN	MFCC: 97,13%	Deteksi kebohongan berbasis suara	Bukan domain fonetik atau tajwid
12	El Kheir et al. (2023) [15]	Automatic Pronunciation Assessment – A Review	Review APA berbasis AI	Review	Survei sistem penilaian pelafalan otomatis	Review umum, tidak khusus Arab/tajwid
13	Nguyen et al. (2025) [16]	AI-Driven Pronunciation Assessment: The Impact of SpeechAce on EFL Learners	SpeechAce AI	Signifikan	Penilaian pelafalan bahasa Inggris (EFL)	Bahasa Inggris, bukan huruf hijaiyah
14	Idris et al. (2021) [14]	The Classification of Actual Pronunciation of Quranic Alphabets for Non-Arab Speaker	Analisis fonetik	Kualitatif	Klasifikasi pelafalan Al-Qur'an non-Arab	Studi kualitatif, tidak ada sistem otomatis
15	**Penelitian Ini**	**Implementasi Algoritma CNN untuk Evaluasi Ketepatan Artikulasi Huruf Hijaiyah**	**MFCC + Grid Search + CNN + DTW**	**—**	**Evaluasi makhraj 28 huruf berharakat fathah**	**Integrasi Grid Search + CNN + DTW untuk seluruh 28 huruf**

Berdasarkan 14 penelitian terdahulu yang telah diulas, dapat diidentifikasi lima kesenjangan utama sebagai berikut:

1. **Tidak Menggabungkan Analisis Spektral dan Temporal secara Terpadu**

   Sebagian besar penelitian hanya menggunakan CNN untuk klasifikasi pola spektral (Ahmad et al., 2023; Makhraj 'Ain VGG-16) atau DTW untuk pencocokan temporal (Wibawa & Darmawan, 2021; Jiang & Chen, 2023) secara terpisah. Belum ada penelitian yang secara eksplisit membandingkan kedua metode tersebut pada domain makhraj huruf hijaiyah.

2. **Cakupan Huruf Hijaiyah yang Terbatas**

   Penelitian makhraj yang ada (Ahmad et al., 2023; Makhraj 'Ain VGG-16) hanya mencakup subset huruf tertentu (huruf halqi atau satu huruf saja), sehingga belum merepresentasikan evaluasi yang komprehensif untuk seluruh 28 huruf hijaiyah berharakat fathah.

3. **Tidak Adanya Pencarian Hyperparameter yang Sistematis**

   Sebagian besar penelitian menetapkan hyperparameter secara manual tanpa eksplorasi sistematis. Penelitian Ilgaz et al. (2024) menunjukkan bahwa *Grid Search* dapat meningkatkan performa secara signifikan, namun penerapannya pada sistem evaluasi makhraj belum dilakukan.

4. **Fokus Penelitian di Luar Domain Tajwid**

   Mayoritas penelitian berfokus pada domain perintah suara, identifikasi pembicara, atau domain musik, bukan pada evaluasi ketepatan artikulasi fonetik dalam konteks tajwid Al-Qur'an.

5. **Kurangnya Studi Perbandingan Skenario Berbeda Parameter**

   Belum ada penelitian yang secara eksplisit membandingkan performa model CNN dengan konfigurasi hyperparameter yang berbeda-beda pada domain makhraj, sehingga pengaruh pilihan hyperparameter terhadap akurasi klasifikasi makhraj belum terdokumentasi.

Posisi Penelitian Ini: Penelitian ini menjawab kelima kesenjangan tersebut dengan mengintegrasikan (1) MFCC sebagai representasi spektral artikulasi, (2) *Grid Search* untuk optimasi hyperparameter secara sistematis, (3) CNN untuk klasifikasi pola makhraj, dan (4) DTW untuk evaluasi kemiripan temporal, yang diterapkan secara komprehensif pada seluruh 28 huruf hijaiyah berharakat fathah.