Segmentasi Motion Data untuk Model *Viseme* Dinamis Bahasa Indonesia

Nurul fadillah
Fakultas Teknologi Industri
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
email: nurul479@gmail.com

Arifin
Fakultas Teknologi Industri
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
arifin@dsn.dinus.ac.id

Surya Sumpeno
Fakultas Teknologi Industri
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
surya@ee.its.ac.id

Mauridhi Hery Purnomo
Fakultas Teknologi Industri
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
hery@ee.its.ac.id

***Abstract*— Animasi bicara yang natural sangat dibutuhkan bagi Industri animasi. Penelitian animasi berbicara Bahasa Indonesia masih sangat jarang dilakukan, sehingga kami melakukan penelitian bidang ini. Animasi bicara yang natural sangat ditentukan oleh kesesuaian antara pengucapan dan *viseme* (*visual phoneme*) tersebut. *Viseme* adalah bentuk bibir ketika mengucapkan suatu fonem atau bunyi bahasa. Penelitian ini bertujuan untuk melakukan segmentasi data *motion capture (mocap)* sehingga diperoleh data fitur setiap suku kata dari kalimat bahasa Indonesia yang diucapkan oleh seorang model. Data yang kami rekam adalah wajah seorang model yang telah dipasang 37 penanda aktif diwajahnya dengan mengucapkan 5 kalimat Bahasa Indonesia. Teknologi yang digunakan untuk merekam adalah teknologi *motion capture* (*mocap*). Data fitur yang diperoleh digunakan sebagai dasar pada proses klasterisasi, sehingga dihasilkan kelas-kelas *viseme* dinamis Bahasa Indonesia. Penelitian ini menjelaskan beberapa kegiatan yaitu perekaman data mocap, konversi data mocap menjadi sistem koordinat dunia, proses normalisasi posisi 3D, proses segmentasi, dan visualisasi. Hasil penelitian menunjukan bahwa data fitur hasil proses segmentasi dapat diterapkan pada proses klaterisasi dengan kualitas klaster yang baik.**

I. PENDAHULUAN

Di bidang animasi tuntutan penyajian animasi yang realitis dan pantas serta menarik semakin tinggi. Animasi harus dapat menampilkan karakter yang sangat mirip dengan di dunia nyata. Ada banyak produk animasi di Indonesia. Salah satunya film yang menarik perhatian kami adalah 'Meraih Mimpi', yang merupakan film animasi Indonesia yang diproduksi oleh *Infinite Frameworks* (IFW)[7]. Film meraih mimpi merupakan film pertama animasi 3D yang ditayangkan di bioskop. Sayangnya, animasi bibir dalam film ini tidak baik. Bibir animasi tidak terlihat realistis karena *viseme* tidak melakukan sinkronisasi dengan fonem yang diucapkan pada saat berbicara[7]. Oleh karena itu, penting untuk menentukan artikulasi *viseme* Indonesia. Hingga saat ini di Indonesia belum ada yang menyelenggarakan standar *viseme* Bahasa Indonesia. Pada penelitian ini kami bertujuan untuk melakukan segmentasi data *motion capture*. Segmentasi data *motion capture* merupakan masalah penting dan sering diteliti di bidang visi komputer[5].

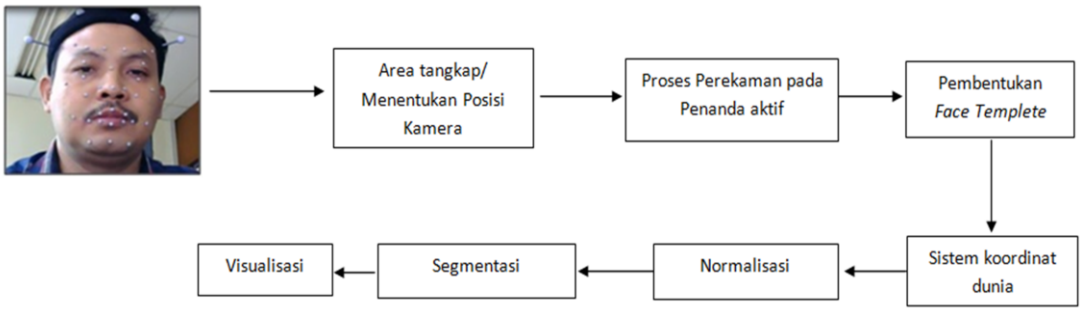
Viseme merupakan representasi visual dari fonetik wicara[6]. Data yang digunakan pada segmentasi *motion capture* merupakan data hasil dari *motion capture* yaitu sampling dan rekaman gerak manusia, hewan, benda mati sebagai data 3D[2]. Hasil dari rekaman *motion capture* tersebut berupa *file* C3D[2]. Data C3D ini yang akan di proses untuk segmentasi *motion capture*. Segmentasi *motion capture* merupakan salah satu langkah awal untuk mendapatkan proses klaterisasi, sehingga diperlukan langkah yang tepat untuk mencari lokalisasi bibir pada saat animasi bicara untuk mendapatkan suku kata dari kalimat yang diucapkan[7][8].

Kami mencari nilai dari gerakkan bibir pada saat animasi berbicara untuk mendapatkan suku kata (*syallabel*). Nilai yang didapatkan dari gerakkan bibir akan digunakan untuk data koordinat dunia dari *motion capture*. Data koordinat dunia yang didapat digunakan untuk proses normalisasi 3D. Proses normalisasi 3D merupakan proses data koordinat dunia yang datanya berubah-berubah pada gerakkan bibir dengan nilai yang tetap terhadap gerakkan kepala yang bertujuan untuk merubah dari data sistem koordinat dunia ke data sistem koordinat lokal. Setelah didapat dari proses normalisasi 3D dilakukan segmentasi *motion capture* yang merupakan proses dari hasil normalisasi yang akan digunakan untuk mencari *frame* pada awal dan akhir setiap pengucapan suku kata. Untuk lebih jelas akan diterangkan di metode diusulkan.

II. METODE DIUSULKAN

Metode penelitian ini secara garis besar digambarkan pada gambar 1. Ada beberapa proses yang akan dilakukan dalam penelitian ini. Pertama yaitu menentukan jenis kamera

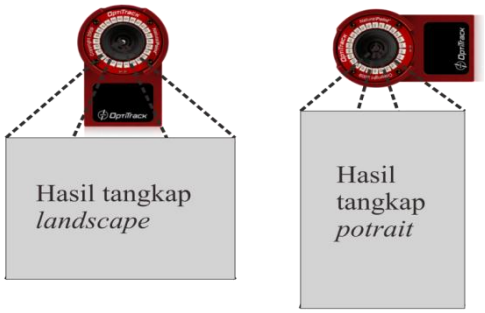
yang akan digunakan untuk pengambilan data. Kedua melakukan proses persiapan model dengan mengunakan 37 penanda aktif yang diletakan pada wajah model. Ketiga mengatur tata letak kamera yang akan digunakan untuk mendeteksi penanda aktif pada wajah model. Lalu, dilakukan proses perekaman penanda aktif pada wajah model yang akan menghasilkan *bone*. Hasil *bone* diproses untuk membentukan *face templete* yang digunakan untuk menangkap gerakkan pada wajah model. Selanjutnya proses mencari data koordinat dunia yang terdapat pada data *motion capture*. Data dari koordinat dunia yang didapat akan digunakan untuk proses normalisasi agar data yang besar menjadi yang lebih kecil. Data normalisasi digunakan untuk proses segmentasi data *motion capture* yang hasilnya terdiri dari beberapa suku kata dari animasi bicara.



Gambar 1. Perancangan sistem penelitian

A. Model Kamera

Kamera yang digunakan pada pengambilan data yang digunakan adalah kamera bertipe VR100:R2, kamera *motion capture OptiTrack*^{TM} memiliki resolusi sebesar 480 x 640 dan memiliki kecepatan tangkap sebesar 100 *frame per second* (*fps*)[1] , seperti gambar 2.



Gambar. 2. Kamera V100:R2

B. Persiapan Model

Model untuk pengambilan data adalah model wajah manusia. Wajah model akan diletakan penanda aktif sebanyak 37 yang terdiri dari 33 + 4. 33 Penanda aktif akan diletakan di area wajah model sedangkan 4 penanda aktif diletakan diatas kepala. Ilustrasi tersebut dapat dilihat pada gambar 3b. Sebagai contoh peletakan penanda aktif pada model wajah manusia dapat dilihat pada model penanda aktif yang disediakan khusus oleh $OptiTrack^{TM}$ seperti pada gambar 3a

Pada $OptiTrack^{TM}$ terdapat konfigurasi untuk peletakan penanda aktif di wajah model. Konfigurasi $OptiTrack^{TM}$ untuk penanda aktif di wajah terdiri dari 3 peletakan penanda aktif yaitu:

1. 23 di wajah + 4 di atas kepala
2. 33 di wajah + 4 di atas kepala

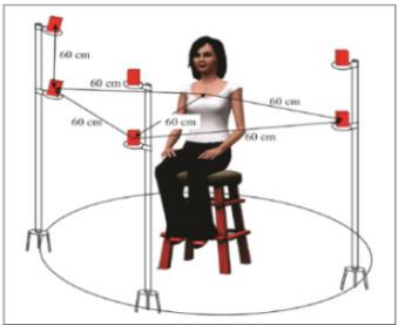
Kami menggunakan 33 + 4 penanda aktif dikarenakan hasil yang diperoleh untuk membentuk *templete facial motion* sangat baik dan hasil gerakkan bibir sesuai dengan model. Sedangkan untuk 23 + 4 hasil yang diperoleh untuk membentuk *templete facial motion* dan gerakkan bibir kurang baik, sehingga sulit untuk dilakukan proses segmentasi.

a b

Gambar. 3. Ilustrasi peletakan penanda aktif di wajah yang terdapat di *OptiTrack*, b peletakan penanda aktif pada Model

C. Area Tangkap/ Posisi Kamera

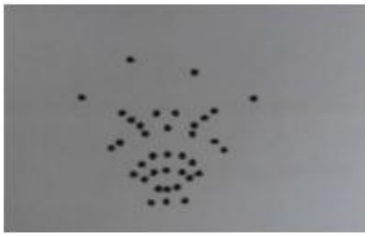
Kamera *motion capture OptiTrack*^{TM} yang berjumlah enam buah disusun menyerupai busur lingkaran dengan kisaran sudut 120^{O}. Masing-masing kamera dipasang secara orientasi dan landscape. Tiga kamera disusun diatas kepala dan tiga kamera disusun setinggi dada. Jarak kamera dengan model sepanjang 60 cm, seperti ilustrasi pada gambar 4.



Gambar 4. Ilustrasi tata letak kamera *OptiTrack*

D. Proses Perekaman penanda aktif di wajah Model

Proses perekaman penanda aktif bertujuan untuk membentuk letak penanda aktif yang terdapat pada wajah model. Proses ini dilakukan untuk menghasilkan tampilan *facial capture*. Sebelum proses perekaman dimulai. Terlebih dahulu menentukan waktu sebelum perekaman, Kemudian menetukan lama waktu yang digunakan untuk perekaman dan menyalakan kamera, sehingga kamera akan mulai merekam. Hasil perekaman penanda aktif seperti pada gambar 5 .



Gambar 5. Hasil *bone*

E. Proses Pembentukan *Face Templete*

Sebelum proses pembentukan *face Templete* dilakukan terlebih dahulu model duduk di depan sistem kamera seperti ilustrasi pada Gambar 4. Sehingga gerak wajah model dapat dilihat oleh beberapa kamera. Setiap *frame* gerakkan pada wajah model akan dilacak oleh *Optik Track.* Sehingga kamera mulai berkerja secara *realtime* melacak penanda aktif yang terdapat pada wajah model. *Software Optik Track* akan secara real time membentuk *face templete* [3].

F. Proses Sistem Koordinat Dunia

Proses sistem koordinat dunia merupakan proses yang di hasilkan dari *facial motion capture* yang bersifat relatif terhadap gerakkan kepala. Untuk mendapatkan data dari gerakkan mulut pada saat model mengucapkan beberapa kalimat yang akan menghasilkan suku kata.

G. Proses Normalisasi Posisi 3D

Sistem koordinat yang dihasilkan dari *facial motion capture* adalah sistem koordinat dunia yang bersifat relatif terhadap gerakkan kepala. Data-data koordinat tiap *frame* akan mudah berubah seiring dengan gerakkan kepala. Oleh karena itu, diperlukan transformasi dari sistem koordinat dunia ke sistem koordinat lokal.

Pada proses transformasi ini diperlukan sebuah bidang yang digunakan sebagai acuan terhadap data-data koordinat dari penanda aktif yang lain. Bidang ini disusun dari titik-titik penanda aktif yang mempunyai sifat relatif tetap terhadap gerakkan kepala.

Kami memilih tiga titik penanda aktif, yaitu titik penanda aktif yang akan digunakan terdiri dari *head*_1, *head*_2 dan *head*_4 (lihat gambar 9(b)) yang masing-masing disebut sebagai p_1, p_2 dan p_3 sehingga sebuah bidang seperti terlihat pada gambar 8. Sumbu Z tegak lurus terhadap bidang $P_1P_2P_3$, maka :

$$V_Z = \frac{\overrightarrow{p_1p_2} x \overrightarrow{p_1p_3}}{|\overrightarrow{p_1p_2} x \overrightarrow{p_1p_3}|} = (Z_i, Z_j, Z_k) \quad (1)$$

$$V_X = \frac{\overrightarrow{P_1P_2}}{|\overrightarrow{P_1P_2}|} = (X_i, X_j, X_k) \quad (2)$$

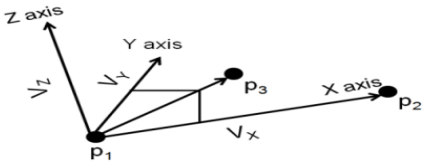
$$V_Y = V_Z x V_X = (Y_1, Y_2 . Y_3) \quad (3)$$

Sehingga terbentuk matriks M :

$$M = \begin{bmatrix} X_1 & Y_1 & Z_1 & p_{1_1} \\ X_2 & Y_2 & Z_2 & p_{1_2} \\ X_3 & Y_3 & Z_3 & p_{1_3} \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (4)$$

$$Mi = inv(M) \quad (5)$$

Selanjutnya, sistem koordinat dari seluruh titik penanda aktif di area mulut (lihat gambar 9(a)) dikalikan dengan dengan matriks *Mi*. Sistem koordinat yang dihasilkan ini yang digunakan pada tahap selanjutnya.

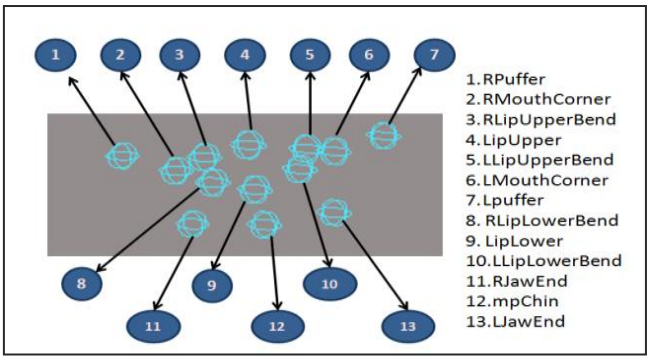


Gambar 8. Bentuk Bidang Sebagai Acuan

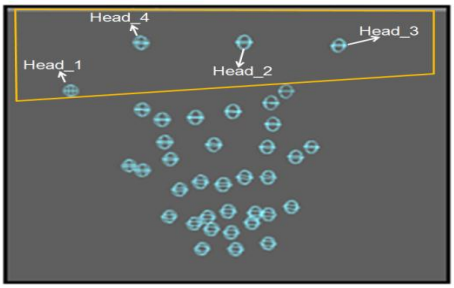
Hasilnya hanya diambil pada bagian area mulut saja untuk digunakan pada proses segmentasi

I. Proses Segmentasi

Segmentasi *motion capture* merupakan proses yang berdasarkan pencarian pada suku kata. Kami mencatat jumlah *frame* pada awal dan akhir setiap pengucapan kalimat untuk mendapatkan suku kata. Data koordinat x,y,z masing-masing penanda aktif dari kumpulan *frame* dihitung nilai rata-ratanya. Nilai rata-rata ini yang selanjutnya digunakan sebagai data fitur untuk masing-masing penanda aktif. Nilai rata-rata tersebut menjadi hasil dari segmentasi *motion capture*.



(a)



(b)

Gambar 9. Penanda aktif yang digunakan untuk normalisasi dimana titik penanda aktif pada area mulut (a) titik penanda aktif pada area kepala (b)

III. HASIL EKSPERIMEN

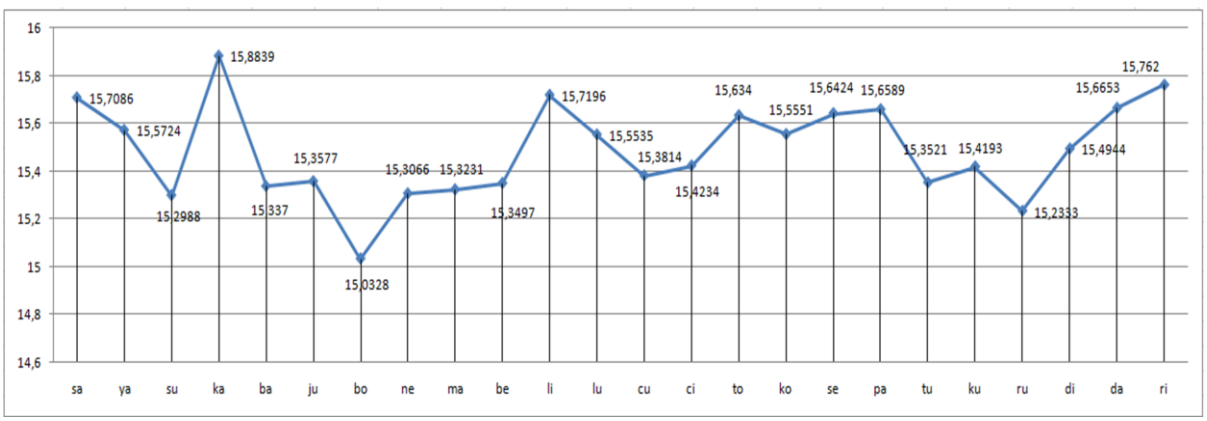
Grafik lintasan gerak seperti gambar 10 menjelaskan perubahan gerakkan bibir animasi, sumbu y menunjukan nilai perubahan gerak bibir animasi pada saat membaca 5 kalimat sedangkan pada sumbu x menunjukan setiap suku kata dari hasil proses segmentasi. Sebagai contoh perubahan gerak bibir animasi, kami mengambil koordinat y untuk Lintasan gerak perubahan bibir animasi, dikarenakan koordinat y lebih unik dari koordinat x dan z, karena nilai pada koordinat y berubah-rubah, sehingga nilai pada koordinat y diambil dari nilai data dibagian area mulut tertentu saja yaitu *LipLower*. Di bagian *LipLower* lebih baik nilai yang didapatkan pada saat model membaca kalimat tersebut. Kalimat yang dibacakan itu terdapat lima kalimat yaitu "Saya suka baju boneka", "mama beli boneka lucu sekali", "mama cuci baju saya", "sepatu baruku dibeli sama papa", "baju baruku dari mama".

Hasil segmentasi *motion data* yang didapatkan adalah hasil dari gerakkan mulut pada saat mengucapkan 5 kalimat yang terdiri dari "Saya suka baju boneka, mama beli boneka lucu sekali", "mama cuci baju saya", "sepatu baruku dibeli sama papa"," baju baruku dari mama", sehingga yang hanya diambil dari awal kata dan akhir untuk proses segmentasi yang akan menghasilkan suku kata seperti yang ditunjukan tabel 1.

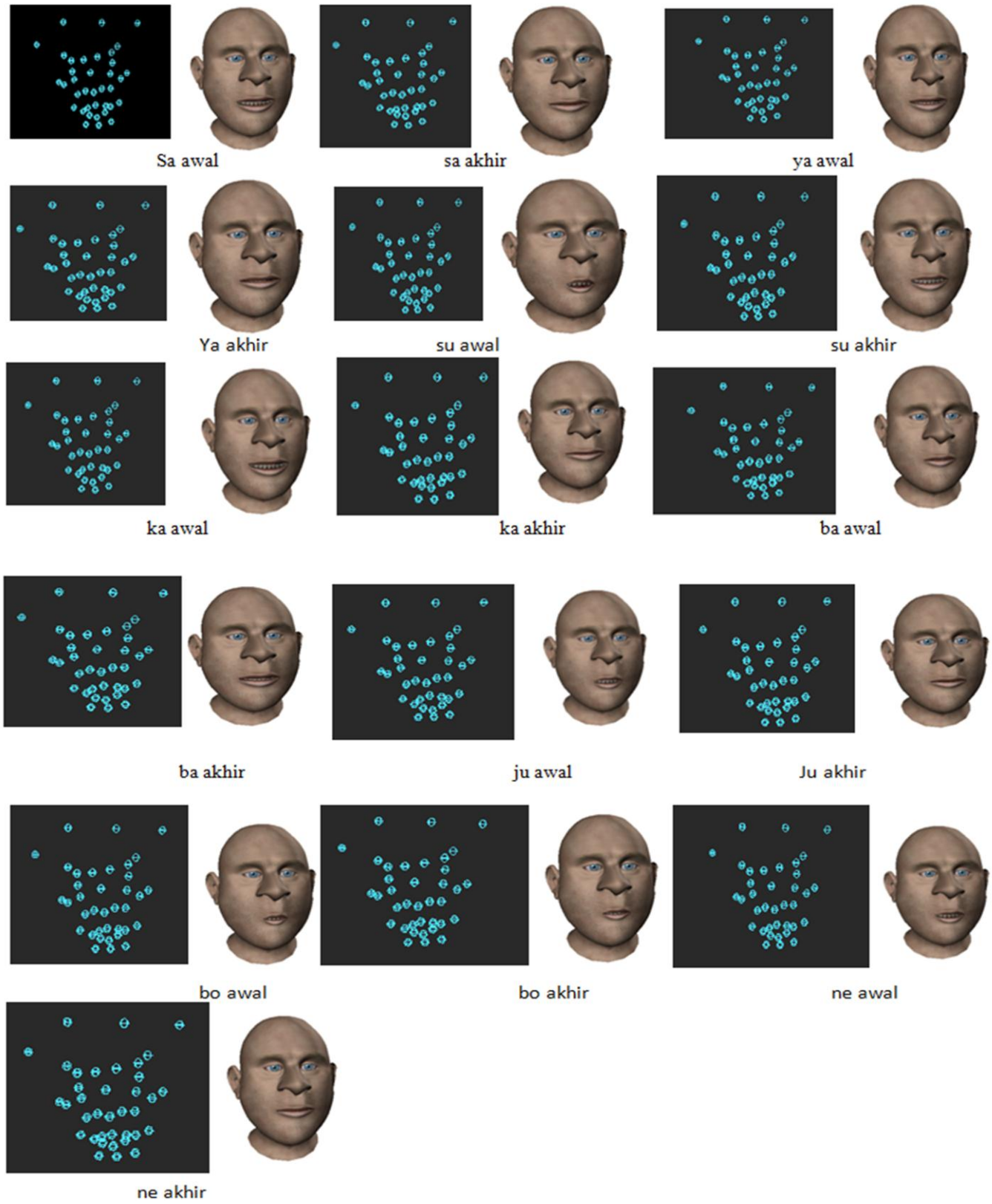
Serta visualisai pada wajah animasi pada saat mengucapkan 5 kalimat yang diperankan oleh model, hasil dari rekaman *motion* data berupa *file* C3D yang akan dikoneksikan dengan wajah animasi, sehingga didapatkan gerakkan bibir animasi yang sama dengan model yang diperankan oleh manusia di dunia nyata pada saat mengucapkan 5 kalimat. Seperti gambar 10.

Tabel 1. Hasil Segmentasi *Motion capture*

No.	Suku kata	Data *LipLower* pada koordinat y
1	sa	15,7086
2	ya	15,5724
3	su	15,2988
4	ka	15,8839
5	ba	15,3370
6	ju	15,3577
7	bo	15,0328
8	ne	15,3066
9	ma	15,3231
10	be	15,3497
11	li	15,7196
12	lu	15,5535
13	cu	15,3814
14	ci	15,4234
15	to	15,6340
16	ko	15,5551
17	se	15,6424
18	pa	15,6589
19	tu	15,3521
20	ku	15,4193
21	ru	15,2333
22	di	15,4944
23	da	15,6653
24	ri	15,7620



Gambar 10. Hasil lintasan gerak segmentasi *facial motion capture* dari 5 kalimat yang dibacakan aktor menjadi 24 kosa kata



Gambar 11 a. Hasil C3D dan *Viseme* animasi saat mengucapkan awal dan akhir suku kata

IV. VALIDASI HASIL SEGMENTASI

Setelah hasil diperoleh dari proses segmentasi berupa data fitur suku kata seperti tabel 1, Kami selanjutnya melakukan uji coba hasil segmentasi. Hasil Segmentasi tersebut kami proses klasterisasi dengan menggunakan Algoritma K-mean. Sehingga hasil segmentasi akan di kelompokan dalam berapa *cluster*.

Dimana langkah klasterisasi untuk melakukan uji coba hasil penelitian menggunakan Algoritma K-mean adalah berikut:

a. Menentukan nilai k secara acak.
b. Menentukan nilai pusat massa. Pada awal iterasi, nilai-nilai centroid yang ditentukan secara acak. Pada langkah iterasi berikutnya, nilai massa ditentukan dengan menghitung rata-rata setiap *cluster*.
c. Menghitung jarak centroid dan masing-masing yang memiliki data .
d. Pengelompokan data berdasarkan minimum *Euclidean Distance*.
e. Selanjutnya akan kembali ke langkah b, mengulangi langkah-langkah sampai nilai *centroid* tetap dan anggota *cluster* tidak pindah ke *cluster* lain. Salah satu metode untuk menentukan *cluster* terdefinisi dengan baik adalah dengan menggunakan kriteria fungsi yang mengukur kualitas Klasterisasi. Ada metode yang digunakan secara luas, yaitu *Sum of Squared* Kesalahan (SSE)

Sehingga didapatkan struktur kelas suku kata *viseme* Seperti tabel 2

Tabel 2 . Hasil diskripsi untuk klasterisasi

Klasterisasi	Suku kata	Klasterisasi	Suku kata
cluster1	su	cluster2	sa
	ju		ya
	bo		ka
	lu		ma
	cu		ba
	to		pa
	ko		da
	tu		
	ku		
	ru		
Klasterisasi	Suku kata		
cluster3	ne		
	be		
	li		
	ci		
	se		
	di		
	ri		

Struktur kelas *viseme* Seperti tabel 2 yang terdiri dari beberapa kelas hasil klasterisasi merupakan struktur kelas viseme Indonesia untuk suku kata yang mencakup fonem Indonesia. Struktur kelas *viseme* Indonesia dalam penelitian ini terbentuk melalui proses pengelompokan untuk menemukan pengelompokan suku kata. Oleh karena itu, di masa depan dapat digunakan sebagai referensi ke sebuah struktur kelas *viseme* Indonesia yang ditentukan berdasarkan pengetahuan linguistik.

KESIMPULAN DAN PENELITIAN SELANJUTNYA

Proses segmentasi *motion capture* digunakan utuk mencari setiap suku kata yang dihasilkan dari gerakkan bibir animasi pada saat mengucapkan 5 kalimat yang terdiri dari “Saya suka baju boneka”, “mama beli boneka lucu sekali”, “mama cuci baju saya”, “sepatu baruku dibeli sama papa”, “baju baruku dari mama” menjadi 24 suku kata yang terdiri dari ,’sa’, ‘ya’, ‘su’, ‘ka’, ‘ba’, ‘ju’, ‘bo’, ‘ne’, ‘ma’,’be’, ‘li’,’lu’, ‘cu’, ‘ci’, ‘to’, ‘ko’, ‘se’, ‘pa’, ‘tu’, ‘ku’,’ ru’, ‘di’, ‘da’,’ ri’. Data yang digunakan pada penilitian ini berupa data *file* C3D yang akan dikoneksi ke animasi agar dapat berbicara secara natural sehingga mirip dengan model yang berada di dunia nyata. Nilai rata-rata yang didapatkan dari hasil segmentasi akan digunakan sebagai data fitur untuk masing-masing penanda aktif, kemudian data fitur dari hasil segmentasi yang akan disiapkan untuk proses klasterisasi suku kata. Penelitian selanjutnya adalah Klasterisasi yang merupakan proses pengelompokan suku kata kedalam beberapa kelas.

REFERENSI

[1] Aang P. Dyaksa, Surya Sumpeno, dan Muhtadin, “Analisis Penangkapan Gerak dari Gerakkan Dasar Manusia Menggunakan Optical *Motion capture*”,Jurusan Teknik Elektro, Fakultas Teknologi Industri, Institut Teknologi Sepuluh Nopember (ITS) 2012.

[2] Midori Kitagawa, Brian Windsor, ”Mocap for Artists” Publish by Elsevier Inc, United States of America, pp. 181, 2008.

[3] François Rocca, Thierry Ravet, Joëlle Tilmanne, “Huma*Face* : Human to Machine Facial Animation”, Laboratorie de Theorie des circuits et Traitement du Signal(TCTS),Universite de Mons(UMONS), Belgique , QPSR , March 2012.

[4] Hui Zhao and Chaojing Tang, “Visual Speech Synthesis based on Chinese Dynamic *Visemes*”, IEEE International Conference on Information and Automation, Zhangjiajie, China, 2008.

[5] Stanisław Badur, et al, “*Viseme* Segmentation by LDA Hysteresis”, Warsaw University of Tehnologi, Faculty of Eletronics and Information Technology, Politecnico di Milano, Dipartimento di Elettronica e Infomazione, Fraunhofer Institute for Telecommunication Heinrich-Hertz-Institute, Image Processing Departement , 2005.

[6] Gleason, H.A., “Introduction to Descriptive Linguistics", New York: Rinehart and Winston, 1970.

[7] Arifin, Mulyono, Surya Sumpeno, Mochamad Hariadi, “Towards Building Indonesian *Viseme* : A Clustering-Based Approach”, CYBERNETICSCOM 2013 IEEE International Conference on Computational Intelegence and Cybernetics, Yogyakarta, December 2013

[8] Sarah L. Taylor, Moshe Mahler, Barry-John Theobald and Ianin Matthews,”Dynamic Units of Visual Speech”, ACM SIGGRAPH Symposium on computer Animation, 2012.