Artikel ini telah dipresentasikan dalam *Innovative and Creative Information Technology Conference* (ICITech) dengan tema "*E-Transaction and Power Play"* yang diselenggarakan oleh Program Studi Teknik Informatika, Fakultas Teknologi Informasi, Universitas Kristen Satya Wacana – Salatiga, pada 24 November 2016.

Membangun Sistem Text-to-Audiovisual Bahasa Indonesia Berdasarkan Database Suara Berbasis Suku Kata Untuk Mendukung Pembelajaran Pelafalan Bahasa Indonesia

Arifin[1], Surya Sumpeno[2], Mochamad Hariadi[3], Arry Maulana Syarif[4]

[2,3]*Institut Teknologi Sepuluh Nopember*, [1,4]*Dian Nuswantoro University*

[2,3]*Surabaya*, [1,4]*Semarang INDONESIA*

[1]arifin@dsn.dinus.ac.id, [2]surya@ee.its.ac.id, [3]mochar@ee.its.ac.id, [4]arry_m @dsn.dinus.ac.id

***Abstract*—** This paper aims to develop a system Text-to-Audio Visual Indonesian to support learning of Indonesian pronunciation based on speech database syllable-based. This system can visualize the pronunciation of the sentences Indonesian synchronized with speech signals. We conduct several research stages, namely forming the Indonesian viseme models, creating the speech database syllable-based, converting the text into syllables dan synchronizing. The synchronization process is a compilation the viseme models and the speech signal based on input text. This system was evaluated by involving 30 respondents who rate the system based on "lip-reading". Each respondent provides an assessment of the 10 Indonesian sentences about the level of compatibility between the visualization of syllable and speech spoken based on text input. The MOS methode (Mean Opinion Score) is used to calculate the average ratings of respondents. MOS calculation results is 4.24, It shows that the level of conformity visualization syllable pronunciation and spoken voice is good.

Keywords— Database Suara Berbasis Suku Kata, Kalimat-Kalimat Berbahasa Indonesia, Sistem Text-to-Audiovisual Bahasa Indonesia, Viseme (Visual Phoneme)

I. PENDAHULUAN

Pada umumnya manusia berkomunikasi melalui bahasa dengan cara menulis atau berbicara. Komunikasi yang dilakukan melalui berbicara, alat ucaplah yang memegang peranan penting untuk menghasilkan bunyi bahasa. Secara umum, proses pembentukan suara dapat dibagi menjadi tiga subproses, yaitu pembangkitan sumber, artikulasi dan radiasi [1]. Organ tubuh yang terlibat dalam proses produksi suara meliputi paru-paru, tenggorokan (trachea), laring (larinx), faring (pharynx), rongga hidung (nasal cavity), dan rongga mulut (oral cavity) serta mulut.

Dewasa ini, Bahasa Indonesia semakin diminati oleh orang asing. Hal ini dapat dilihat dengan semakin banyak dibukanya lembaga - lembaga yang mengajarkan Bahasa Indonesia sebagai bahasa asing di beberapa negara. Pemerintah melalui Biro Perencanaan dan Kerjasama Luar Negeri (BPKLN) Kementrian pendidikan dan kebudayaan menyelenggrakan program 'Darmasiswa' sejak tahun 2005. Program ini menyelenggarakan pembelajaran Bahasa Indonesia bagi penutur asing yang diikuti oleh 110 negara dari lima benua, yaitu Asia, Amerika, Australia, Eropa dan Afrika. Sedangkan, Di dalam negeri terdapat 45 perguruan tinggi yang menyelenggarakan Program Darmasiswa ini.

Walaupun penggunaan bahasa Indonesia sudah sedemikian luas, tetapi masih banyak dijumpai masalah-masalah yang berkaitan dengan pelafalannya. Kaidah

pelafalan suatu bunyi bahasa akan berbeda tergantung bahasa yang digunakan. Contoh permasalahan yang sering dijumpai pada pelafalan bahasa Indonesia, antara lain adalah kata ‘teknik’ dilafalkan dengan ‘tehnik’ yang seharusnya dilafalkan ‘teknik’. Kata ‘energi’ dilafalkan dengan ‘enerhi’, ‘enersi’, ‘enerji’ yang seharusnya dilafalkan ‘energi’. Contoh lain adalah pelafalan fonem ‘E’, akan berbeda maknanya pada kata ‘tEras’ yang artinya halaman rumah dan ‘teras’ yang artinya pejabat. Dan masih banyak dijumpai permasalahan-permasalahan pelafalan lainnya.

Animasi wajah yang natural, hidup dan realistis merupakan bidang penelitian yang menantang saat ini [2]. Beberapa aplikasi yang dapat dikembangan adalah animasi karakter wajah untuk film animasi, terapi wicara bagi tuna rungu, dan merancang sistem Interaksi Manusia-Komputer yang baru. Animasi berbicara yang realistis merupakan salah satu bagian penting dalam animasi karakter wajah. Pada umumnya, animasi berbicara dibangun menggunakan viseme yang disinkronkan dengan suara dan fonem yang diucapkan. Beberapa fonem yang berbeda divisualisasikan oleh viseme yang sama, seperti fonem ‘m’, ‘p’, dan ‘b’. Fonem-fonem seperti itu dapat dikelompokan kedalam satu kelas. Faktanya bahwa visualisasi sebuah fonem dapat menjadi berbeda-beda tergantung koartikulasi yang mengikuti artikulasi tersebut. Sebagai contoh, visualisasi fonem ‘b’ akan berbeda pada kata ‘buku’ dan ‘baca’. Hal ini disebabkan oleh koartikulasi yang mengikuti fonem tersebut. Konsep visualisasi sebuah fonem yang berbeda-beda tergantung pada koartikulasi yang menyertai disebut viseme dinamis. Pengembangan animasi berbicara yang realistik memerlukan pembahasan viseme dinamis ini.

Sistem Text-To-Audiovisual merupakan gerakan aksi wajah terutama bentuk mulut ketika seseorang berkomunikasi dengan orang lain [3]. Sintesis ini mengambil input teks baku dan memberikan transkripsi fonetik untuk setiap kata. Teks dibagi menjadi unit prosodi, seperti frase, klausa, dan kalimat. Kombinasi dari transkrip fonetik dan informasi tentang unit prosodi dapat digunakan untuk membuat simbolis yang merepresentasikan linguistik, yang dianggap sebagai front-end dari sistem Text-to-Speech. Sintesiser menggunakan simbol-simbol ini untuk merepresentasikan linguistik dan mengkonversikannya menjadi suara.

II. KAJIAN PUSTAKA

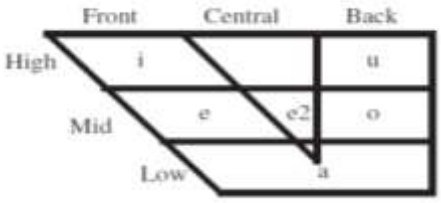
A. Karakteristik Bahasa Indonesia

Bahasa Indonesia telah dinyatakan seagai bahasa nasional sejak tahun 1928. Saat itu bahasa Indonesia menjadi bahasa pergaulan antar etnis (lingua franca) yang mampu merekatkan suku-suku di Indonesia. Mulanya bahasa Indonesia ditulis dengan tulisan Latin-Romawi mengikuti ejaan Belanda, hingga tahun 1972 diberlakukan Ejaan Yang Disempurnakan (EYD). Dalam bahasa Indonesia, dikenal satuan bahasa seperti frase dan kalimat. Kalimat merupakan satuan bahasa terkecil dalam wujud lisan atau tulisan yang mengungapkan pikiran yang utuh. Suatu kalimat dapat terdiri dari beberapa unsur seperti subyek, predikat, objek, pelengkap dan keterangan. Sebuah kalimat minimal memiliki unsur subyek dan predikat. Gabungan dari unsur-unsur kalimat tersebut akan membentuk kalimat yang mengandung arti.

B. Pengertian Dasar Fonem

Fonem adalah satuan bunyi bahasa terkecil yang dapat membedakan arti [9]. Sebagai contoh huruf ‘h’ pada kata ‘harus’ adalah fonem. Apabila huruf ‘h’ pada kata tersebut dihilangkan, maka akan menjadi ‘arus’. Kata ‘harus’ berbeda dengan ‘arus’, sehingga keberadaan huruf ‘h’ dapat membedakan arti. Fonem bahasa Indonesia berisi 33 simbol fonem yang terdiri 10 vokal (termasuk diftong), 22 konsonan dan 1 simbol diam. Pola artikulasi vokal Bahasa Indonesia yang menunjukan 2 resonansi pertama, F1 (height) dan F2 (backness), ditunjukan pada Gambar 1. Vokal adalah bunyi ujaran yang tidak mendapatkan rintangan saat dikeluarkan dari paru-paru. Vokal dibagi menjadi dua, yaitu vokal tunggal (monoftong) yang meliputi ‘a’, ‘i’, ‘u’, ‘e’, ‘o’ dan vokal rangkap (diftong), yang meliputi ‘ai’, ‘au’, ‘oi’. Konsonan adalah bunyi ujaran yang dihasilkan dari paru-paru dan mengalami rintangan saat keluarnya. Contoh konsonan antara lain ‘p’, ‘b’, ‘m’, ‘w’, ‘f’, ‘v’, ‘t’, ‘d’, ‘n’, ‘c’, ‘j’, ‘k’, ‘g’, ‘h’. Konsonan rangkap disebut kluster. Contoh kluster pada kata ‘drama’, ‘tradisi’, ‘film’, ‘modern’.

Grafem merupakan satuan unit terkecil sebagai pembeda dalam sebuah sistem aksara [10]. Grafem membahas mengenai huruf, sedangkan fonem membahas mengenai bunyi. Hubungan grafem dan fonem dapat terjadi hubungan one-to-one, seperti pada kata ‘kursi’ yang terdiri dari grafem ‘k’, ‘u’, ‘r’, ‘s’, ‘i’ dan pengucapannya juga terdiri dari 5 fonem ‘k’, ‘u’, ‘r’, ‘s’, ‘i’. Jenis hubungan grafem dan fonem yang lain adalah many-to-one, seperti pada kata ‘ladang’ yang terdiri dari grafem ‘l’, ‘a’, ‘d’, ‘a’, ‘n’, ‘g’. Sedangkan, pengucapannya terdiri dari fonem ‘l’, ‘a’, ‘d’, ‘a’, ‘ng’. Jadi, grafem ‘n’ dan ‘g’ direpresentasikan oleh satu fonem ‘ng’. Table I menampilkan huruf-huruf alfabet dan fonem-fonem bahasa Indonesia.



Gambar 1. Pola Artikulasi Vokal-Vokal Bahasa Indonesia

TABEL I
HURUF ALFABET DAN FONEM BAHASA INDONESIA

Jenis	Konsonan	Vokal Tunggal (Monoftong	Vokal Rangkap (Diftong)
Huruf Alfabet	‘b’, ‘c’, ‘d’, ‘f’, ‘g’, ‘h’, ‘j’, ‘k’, ‘l’, ‘m’, ‘n’, ‘p’, ‘q’, ‘r’, ‘s’, ‘t’, ‘v’, ‘w’, ‘x’, ‘y’, ‘z’	‘a’, ‘e’, ‘i’, ‘u’, ‘o’	
Fonem	‘p’, ‘b’, ‘t’, ‘k’, ‘d’, ‘g’, ‘c’, ‘j’, ‘f’, ‘z’, ‘h’, ‘m’, ‘kh’, ‘sy’, ‘n’, ‘ng’, ‘r’, ‘w’, ‘y’, ‘v, ny, ‘diam’	‘a’, ‘e’, ‘E’, ‘i’, ‘o’, ‘u’	‘ao’, ‘au’, ‘au’

C. Pola-Pola Suku Kata Bahasa Indonesia

Suku kata adalah bagian kata yang diucapkan dalam satu hembusan nafas dan umumnya terdiri dari beberapa fonem. Setiap suku kata dalam bahasa Indonesia ditandai oleh sebuah vokal yang dapat diikuti maupun didahului oleh konsonan. Pola-pola suku kata dalam bahasa Indonesia seperti yang terlihat di Tabel 2 [13]. Vokal dinotasikan oleh V, sedangkan konsonan dinotasikan oleh K.

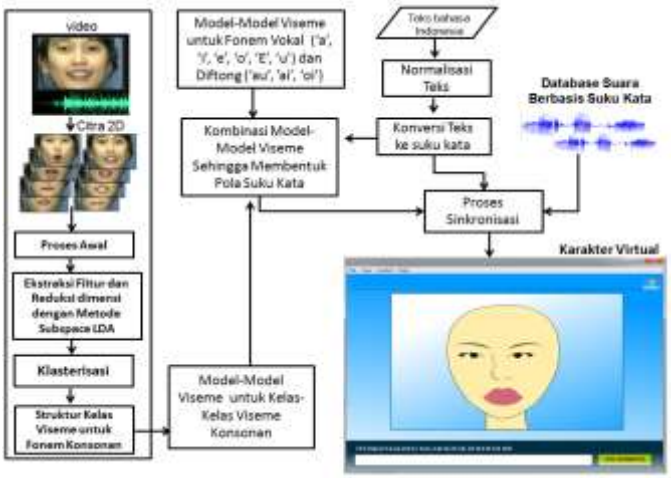
TABEL II
POLA-POLA SUKU KATA DALAM BAHASA INDONESIA

Pola Suku Kata	Contoh
V	a-dik, i-tu
VK	ka-in, am-dal
KV	la-ma, ba-ru
KVK	ne-nek, ma-buk
KKV	pra-ha-ra, dra-ma
KKVK	pres-ta-si
VKK	eks-pan-si
KVKK	teks, pers
KKVKK	kom-pleks, tri-pleks
KKKV	stra-ta, in-stru-men
KKKVK	struk-tur

III. METODE PENELITIAN

D. Metode yang Diusulkan

Dalam penelitian ini, ada beberapa tahapan yang dilakukan yaitu membuat database suara, melakukan klasterisasi terhadap dataset citra 2D untuk memperoleh kelas-kelas viseme konsonan, membangun model-model viseme untuk kelas-kelas viseme konsonan, viseme vokal dan diftong, dan proses sinkronisasi suara dan model-model viseme berdasarkan teks yang diinputkan. Secara keseluruhan tahapan-tahapan penelitian tersebut dapat dilihat seperti Gambar 2.



Gambar 2. Metode yang Diusulkan

Dataset yang digunakan dalam proses klasterisasi merupakan citra 2 dimensi yang dihasilkan dari proses ekstraksi terhadap video yang bersisi adegan orang yang sedang mengucapkan kalimat-kalimat berbahasa Indonesia yang berjumlah 200 kalimat. Kalimat-kalimat berbahasa Indonesia yang digunakan dalam perekaman ini sudah melalui analisis bahwa seluruh kalimat yang digunakan sudah mencakup seluruh pola suku kata dalam bahasa Indonesia. Fokus dalam pembuatan video ini adalah perekaman gerakan wajah dan mulut saat mengucapkan kalimat-kalimat berbahasa Indonesia

E. Pembentukan Model-Model Viseme Untuk Fonem Konsonan

Pembentukan model-model viseme untuk fonem-fonem konsonan didasarkan pada hasil proses klasterisasi terhadap dataset citra 2D. Kami telah membuat video berdurasi 6 menit 36 detik yang berisi gerakan wajah orang yang sedang mengucapkan 200 kalimat-kalimat berbahasa Indonesia. Video ini diekstraksi sehingga diperoleh sekitar 10.000 frame-frame citra 2D. Kami memilih frame-frame yang mewakili viseme untuk fonem-fonem konsonan sehingga diperoleh 315 frame Langkah selanjutnya adalah preprocessing terhadap frame-frame tersebut, yaitu mengubah format warna citra, melakukan croping citra 2D di daerah mulut dan mengubah ukuran seluruh citra agar mempunyai ukuran yang sama.

Metode Subspace LDA digunakan untuk melakukan ekstraksi fitur dan reduksi dimensi. Metode ini merupakan kombinasi dari metode PCA (Principal Componen Analysis) dan LDA (Linear Discriminant Analysis). Penggunaan metode PCA bertujuan untuk memproyeksikan data pada arah yang memiliki variasi terbesar, yang ditunjukan oleh vektor eigen yang bersesuaian dengan nilai eigen terbesar dari matriks kovarian [10]. Secara spesifik, tugas PCA adalah mereduksi dimensi dengan melakukan transformasi linier dari suatu ruang berdimensi tinggi ke dalam ruang berdimensi rendah. Sedangkan, metode LDA bertujuan menemukan sub ruang linier yang memaksimalkan jarak matriks sebaran antar kelas (between-class) dinotasikan dengan SB dan meminimalkan jarak matriks sebaran dalam kelas (within-class) dinotasikan dengan SW. Hasil dari proses ini adalah kelas-kelas saling terpisah secara linier.

Proses ekstraksi fitur bertujuan untuk mengambil ciri pada data citra 2D. Ciri yang diperoleh digunakan sebagai dataset dalam proses klasterisasi. Dimisalkan terdapat himpunan sebanyak M data citra dari dataset citra bentuk bibir (Aj), dimana Aj = [A1,A2,..,AM], (j = 1, 2, ..., M) dengan dimensi citra baris x kolom pixels yang diproyeksikan ke dalam matrik dua dimensi (T) adalah :

$$T = \begin{bmatrix} x_{11} & x_{21} & \cdots & x_{m1} \\ x_{12} & x_{22} & \cdots & x_{m2} \\ \cdots & \cdots & \cdots & \cdots \\ x_{1n} & x_{2n} & \cdots & x_{mn} \end{bmatrix} \qquad (1)$$

dimana x adalah nilai tiap pixels matrik citra A_j.

Dari matriks T tersebut, metode PCA diterapkan untuk melakukan reduksi dimensi dengan tahapan-tahapan sebagai berikut :

1. Menghitung matriks rata-rata baris dari matrik T dengan menggunakan persamaan (2) :

$$\bar{A}_{im} = \frac{1}{M_i} \sum_{J=1}^{M_i} X_{jm} \quad (2)$$

dimana M_i adalah jumlah data baris ke *i* dan X_{jm} adalah data-data pada baris ke *i* .

2. Menghitung matrik *ATrain* yang merupakan selisih dari data citra *T* dan matriks rata-rata baris.

$$ATrain = T - \bar{A} \quad (3)$$

dimana $\bar{A}$ adalah nilai rata-rata baris $\bar{A}_{im}$.

3. Menghitung matrik kovarian S_T (total matriks Scatter S_T) yang didefinisikan dengan menggunakan persamaan (4).

$$S_T = ATrain \; x \; ATrain' \quad (4)$$

Dari matriks kovarian S_T, dihitung *eigenvalue(D)* dan *eigenvektor(V)*. *Eigenvalue* merupakan nilai karakteristik dari suatu matrik bujursangkar, sedangkan *eigenvector* merupakan nilai yang diambil berdasarkan nilai eigen yang lebih besar dari 0. Dalam penelitian ini, nilai *eigenvalue (D)* dan *eigenvector (V)* dicari dengan menggunakan fungsi Matlab.

4. Menghitung nilai eigenfaces yang merupakan ciri data citra. Persamaan yang digunakan untuk menghitung nilai eigenfaces adalah :

Eigenfaces = ATrain x V (5)

5. Menghitung matriks proyeksi PCA.

PCA_Train = Eigenfaces' x T (6)

Matriks proyeksi *PCA_Train* merupakan hasil proses PCA yang selanjutnya digunakan untuk proyeksi LDA. *Data sets PCA_train* yang diperoleh dari proses PCA akan digunakan untuk proses proyeksi LDA. Matrik scatter dalam kelas *(S_W)*, dan matrik scatter antar kelas *(S_B)* didefinisikan sebagai berikut :

$$S_W = \sum_{i=1}^{c} \sum_{a_j \in A_i} (A_j - \overline{A}_i)(A_j - \overline{A}_i)^T \quad (7)$$

$$S_B = \sum_{i=1}^{c} N_i (\overline{A}_i - \overline{A})(\overline{A}_i - \overline{A})^T \quad (8)$$

dimana c adalah jumlah kelas dan N_i adalah jumlah data pada kelas A_i. Sedangkan $\overline{A}_i$ adalah nilai rata-rata per kelas dan A_j adalah *PCA_train* yang diambil per kelas.

Hasil proses reduksi dimensi dengan metode LDA adalah matriks proyeksi *LDA_train* yang selanjutnya digunaka sebagai dataset dalam proses klasterisasi dengan menggunakan *K-Means*. Algoritma *K-Means* merupakan algoritma untuk klaster n data berdasarkan atribut tertentu menjadi k partisi, dimana k < n [11]. Tahapan-tahapan dalam proses klasterisasi dengan algoritma *K-Means* adalah sebagai berikut :

1. Langkah awal adalah menentukan jumlah cluster.
2. Menentukan nilai centroid. Pada awal iterasi, nilai-nilai *centroid* ditentukan secara acak. Dan tahap iterasi berikutnya, nilai *centroid* ditentukan dengan mengitung nilai rata-rata tiap cluster dengan rumus :

$$\bar{V}_{ij} = \frac{1}{N_i} \sum_{k=0}^{N_i} X_{kj} \quad (9)$$

   dimana $\bar{V}_{ij}$ adalah *centroid* cluster ke-i untuk variabel ke-j. N_i adalah jumlah data dalam cluster ke-i, sedangkan X_{kj} adalah data ke-k untuk variabel ke-j.
3. Menghitung jarak antara *centroid* dengan tiap data. Untuk menghitung jarak tersebut digunakan *Euclidean Distance*, yaitu :

$$D_e = \sqrt{(x_i - s_i)^2 + (y_i - t_i)^2} \quad (10)$$

   dimana D_e adalah *Euclidean Distance* dan *i* adalah banyaknya data, sedangkan *(x,y)* merupakan koordinat data dan *(s,t)* merupakan koordinat centroid.
4. Mengelompokan data berdasarkan *Euclidean Distance* yang paling minimum.
5. Kembali ke langkah 2, lakukan perulangan hingga nilai *centroid* yang dihasilkan tetap dan anggota klaster tidak berpindah ke klaster yang lain.

Setiap klaster beranggotakan data-data yang lebih mirip satu sama lain dalam klaster itu dibanding dengan data-data dari anggota klaster yang lain. Salah satu cara agar klaster terdefinisi dengan baik, maka perlu menggunakan fungsi kriteria yang mengukur kualitas hasil proses klasterisasi. Salah satu cara yang sering digunakan adalah dengan menghitung jumlah dari kesalahan kuadrat (*Sum of Squared Error, SSE*). Semakin kecil nilai SSE menunjukan kualitas hasil proses klasterisasi semakin baik [12].

$$SSE = \sum_{i=1}^{k} \sum_{p \in C_i} d\,(p, m_i)^2 \quad (11)$$

Dimana *k* adalah jumlah klaster, *p* adalah titik data anggota masing-masing klaster C_i dan $d(p, m_i)$ adalah jarak masing-masing titik data *p* ke *centroid m* untuk klaster ke *i*.

Kualitas Klaster juga dapat dilihat dari perbandingan variasi data antar klaster (*between-class variation, BCV*) dengan variasi data dalam klaster (*within-class variation, WCV*). BCV merupakan rata-rata jarak antar centroid dan WCV merupakan *Sum of Square Error*[12]. Semakin besar nilai perbandingan BCV dan WCV menunjukan kualitas hasil proses klasterisasi semakin baik. Perbandingan antara BCV dengan WCV dihitung dengan menggunakan persamaan (12).

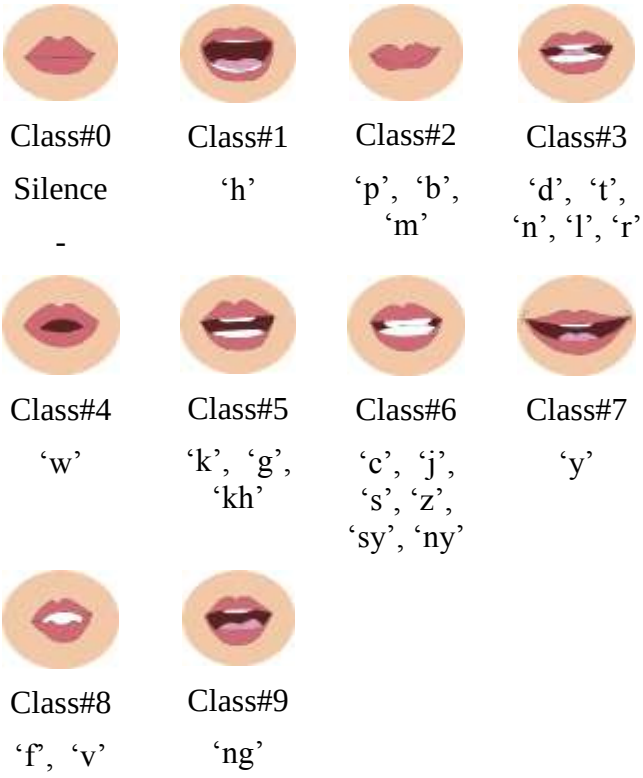
$$\frac{BCV}{WCV} = \frac{\frac{1}{n_k} \sum_{i=1}^{k} d\,(m_i, m_i)}{SSE} \quad (12)$$

dimana $\frac{1}{n_k}\sum_{i=1}^{k} d(m_i, m_i)$ adalah rata-rata jarak antar *centroid*.

Dalam penelitian ini, kami telah melakukan beberapa eksperimen dengan memasukan nilai k yang bervariasi [8]. Hasil perhitungan SSE dan rasio BCV dan WCV dari masing-masing eksperimen dapat dilihat seperti Tabel III. Kualitas hasil proses klasterisasi yang terbaik diperoleh pada k=9, dengan nilai SSE paling kecil dan rasio BCV dan WCV paling besar. Kelas-kelas yang terbentuk dari hasil proses klasterisasi ini digunakan sebagai dasar dalam pembentukan model-model kelas viseme untuk fonem-fonem konsonan seperti terlihat di Gambar 3.

TABEL III
HASIL PERHITUNGAN SSE DAN RASIO BCV DAN WCV

K value	Mean of Centroid Distance (BCV)	SSE (WCV)	BCV / WCV
k=5	0.88	66.58	0.0132
k=6	0.95	54.37	0.0175
k=7	0.88	50.28	0.0176
k=8	0.83	47.86	0.0174
k=9	0.79	42.38	0.0187
k=10	0.77	46.87	0.0165
k=11	0.75	43.59	0.0171
k=12	0.75	42.81	0.0175
k=13	0.70	42.73	0.0162

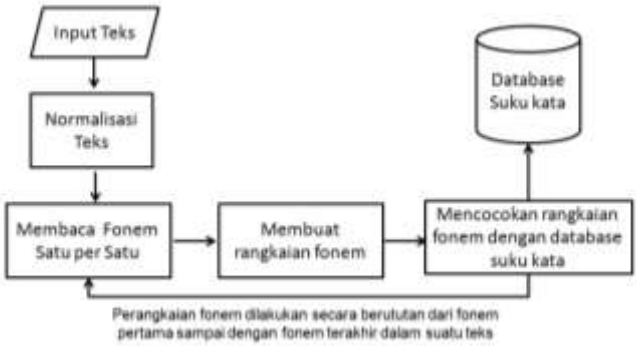


Gambar 3. Model-Model Kelas Viseme Untuk Fonem-Fonem Konsonan

Pembentukan kelas viseme ini bertujuan untuk mengurangi banyaknya variasi visualisasi bentuk mulut masing-masing fonem. Beberapa fonem yang berbeda ternyata dapat divisualisasikan oleh sebuah viseme yang sama, misalnya fonem ‘b’, ‘p’, dan ‘m’. Satu kelas viseme merupakan representasi visual dari sebuah fonem atau beberapa fonem yang berbeda.

F. Konversi Teks ke Suku Kata

Konversi teks ke suku kata merupakan proses pemisahan teks menjadi kata-kata dan pemisahan kata-kata menjadi suku kata – suku kata. Teks yang akan diproses, pada umumnya memiliki beberapa karakteristik diantaranya adalah memiliki dimensi tinggi, terdapat *noise*, dan juga struktur yang tidak baik. Oleh karena itu diperlukan tahap normalisasi teks yaitu *case folding*, menghilangkan karakter-karakter tanda baca dan mengubah angka menjadi rangkaian huruf [14]. Proses *case folding* adalah mengubah semua huruf dalam dokumen teks menjadi huruf kecil. Konversi kata menjadi suku kata dapat dilakukan dengan aturan konversi yang sederhana dengan mengimplementasikan tabel konversi yang berisi pola-pola suku kata yang dikenal dalam bahasa Indonesia. Algoritma penentuan suku kata dari suatu teks tertentu terlihat seperti Gambar 4.



Gambar 4. Algoritma Penentuan Suku Kata dari Suatu Teks

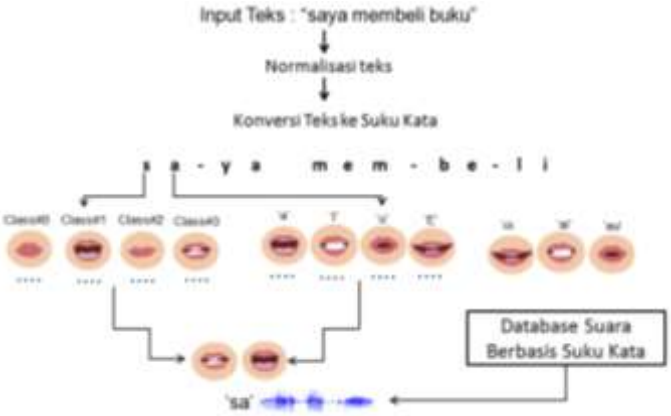
Membaca fonem satu persatu dari fonem pertama ke fonem selanjutnya dari teks bahasa Indonesia yang diinputkan. Masing-masing fonem dirangkai dan dicocokan dengan database suku kata. Rangkaian fonem yang dikenal dalam database suku kata akan dibentuk menjadi suatu suku kata. Langkah ini dilakukan sampai semua fonem dalam teks tersebut terbaca.

G. Membuat Database Suara Berbasis Suku Kata

Salah satu data yang digunakan dalam membangun sistem ini adalah database suara yang direkam berdasarkan bentuk suku kata yang mencakup seluruh pola suku kata dalam bahasa Indonesia. Kami telah merekam suara pengucapan suku kata-suku kata untuk seluruh pola suku kata yang ada dalam bahasa Indonesia. Jenis-jenis suku kata yang digunakan dalam proses perekaman mengacu pada suku kata dalam Kamus Besar Bahasa Indonesia (KBBI). Masing-masing suara untuk suku kata tertentu disimpan dalam file dengan format .wav.

H. Membentuk Visualisasi Pengucapan Suku Kata

Pada tahapan ini, dilakukan kombinasi model-model viseme untuk membentuk visualisasi pengucapan suku kata tertentu sebagai unit terkecil untuk dataset visual [13]. Model-model viseme yang dikombinasikan terdiri dari model-model viseme untuk fonem-fonem konsonan, model-model viseme untuk fonem-fonem vokal (‘a’, ‘i’, ‘u’, ‘e’, ‘o’, ‘E’) dan model-model viseme untuk fonem-fonem diftong (‘ai’, ‘au’, ‘oi’). Pembentukan visualisasi pengucapan suku kata ini didasarkan pada teks yang diinputkan ke sistem. Setelah teks tersebut dikonversikan menjadi suku kata - suku kata, maka masing-masing suku kata dibentuk visualisasinya dengan menggabungkan model-model viseme tersebut. Proses pembentukan visualisasi pengucapan suku kata dapat diilsutrasikan seperti Gambar 5.



I. Proses Sinkronisasi

Proses sinkronisasi bertujukan untuk mengkompilasi antara model-model viseme, suara dan suku kata serta menciptakan aliran visual dari rangkaian frame yang mengalami peralihan viseme [15]. Dasar proses sinkronisasi adalah kompilasi antara suku kata, visualisasi pengucapan suku kata dan suara yang diucapkan sesuai dengan teks yang diinputkan. Visualisasi pengucapan suku kata dan suara suku kata dirangkai sehingga membentuk pengucapan kalimat-kalimat berbahasa Indonesia yang realistik. Perangkaian suara-suara suku kata menggunakan teknik overloading sehingga terbentuk suara yang utuh tanpa terjadi suara yang terputus-putus antara satu suara suku kata dengan suara suku kata yang lain. Sedangkan untuk perangkaian visualisasi pengucapan suku kata, kami menggunakan dasar penentuan jumlah frame pada masing-masing fonem sehingga membentuk visualisasi suku kata yang realistik. Perhitungan jumlah frame masing-masing fonem dihitung berdasarkan nilai durasi saat awal dan akhir pengucapan fonem tertentu. Nilai durasi ini diperoleh berdasarkan sinyal suara yang disimpan dalam database suara untuk masing-masing suku kata.

IV. HASIL EKSPERIMEN DAN PEMBAHASAN

Kami melakukan pengujian terhadap sistem Text-to-Audiovisual bahasa Indonesia ini dengan memasukan 10 teks Bahasa Indonesia. Teks-teks bahasa Indonesia yang diinputkan dalam eksperimen ini seperti yang ditunjukan di Tabel 5. Didalam pengujian ini, kami melibatkan 30 responden yang akan mengamati

tingkat kesesuaian antara visualisasi pengucapan suku kata dan suara yang dihasilkan. Setiap responden memberikan penilaian tingkat kesesuaian tersebut sesuai dengan kriteria yang ada di Tabel 4. Sedangkan, hasil penilaian oleh responden direkapitulasi berdasarkan masing-masing tingkatan kriteria MOS untuk masing-masing kalimat seperti yang tersaji di Tabel 6. Metode yang digunakan untuk menghitung rata-rata penilaian oleh responden adalah MOS yang dirumuskan seperti persamaan (13).

\

TABEL IV
KRITERIA PENILAIAN MOS

MOS	Kualitas	Keterangan
5	Sangat Bagus	Suara dan visualisasi pengucapan suku kata sangat sesuai
4	Bagus	Suara dan visualisasi pengucapan suku kata sesuai
3	Cukup	Suara dan visualisasi pengucapan suku kata cukup sesuai
2	Jelek	Suara dan visualisasi pengucapan suku kata kurang sesuai
1	Sangat Jelek	Suara dan visualisasi pengucapan suku kata tidak sesuai sama sekali

TABEL V
TEKS BAHASA INDONESIA YANG DIGUNAKAN SEBAGAI INPUT SISTEM

No	Teks Bahasa Indonesia
1	sekarang ayah pergi ke kantor naik mobil dinas
2	paman yang memberiku buku tulis dan tas baru
3	sekolahnya di pinggir jalan raya yang sepi itu
4	nomer teleponnya tidak bisa dihubungi lagi sekarang
5	siang ini udaranya panas tidak mendung lagi
6	sekarang di negara kita sedang musim hujan
7	dia tidak masuk sekolah hari ini karena sakit
8	dia sakit malaria sejak dua hari yang lalu
9	biaya di rumah sakit internasional itu mahal
10	hari minggu kami sekeluarga akan pergi memancing

TABEL VI
REKAPITULASI PENILAIAN OLEH RESPONDEN

Teks	Tingkat Kesesuaian Visualisasi Pengucapan Suku Kata dan Suara yang Diucapkan				
	Buruk	Kurang	Cukup	Bagus	Sangat bagus
sekarang ayah pergi ke kantor naik mobil dinas	0	2	7	10	11
paman yang memberiku buku tulis dan tas baru	0	1	5	12	12
sekolahnya di pinggir jalan raya yang sepi itu	0	1	8	9	12
nomer teleponnya tidak bisa dihubungi lagi sekarang	0	1	5	9	15
siang ini udaranya panas tidak mendung lagi	0	0	7	10	13
sekarang di negara kita sedang musim hujan	0	2	5	7	16
dia tidak masuk sekolah hari ini karena sakit	0	0	8	9	13
dia sakit malaria sejak dua hari yang lalu	0	1	3	9	17
biaya di rumah sakit internasional itu mahal	0	2	3	7	18
hari minggu kami sekeluarga akan pergi memancing	0	0	1	12	17

$$MOS = \sum_{i=1}^{n} \frac{x(i).k}{N} \quad (13)$$

Dimana x(i) adalah nilai sampel ke i, k adalah jumlah bobot dan N adalah jumlah responden. Hasil perhitungan MOS adalah 4,24 dengan range nilai 1 s.d. 5. Ini menunjukan bahwa tingkat kesesuaian visualisasi pengucapan suku kata dan suara yang diucapkan pada sistem Text-to-Audiovisual Bahasa Indonesia adalah bagus.

V. KESIMPULAN DAN DISKUSI

Berdasarkan beberapa eksperimen yang telah dilakukan dan perhitungan hasil penilaian oleh responden dengan menggunakan metode MOS, maka dapat disimpulkan bahwa Sistem Text-to-Audiovisual Bahasa Indonesia yang dibangun berdasarkan database suara berbasis suku kata dapat menghasilkan visualisasi pengucapan suku kata yang bagus (mendekati realistik). Hal ini dapat dilihat dari nilai MOS yang diperoleh yaitu 4,24 dengan range nilai 1 s.d. 5.

Kami juga mengamati visualisasi pengucapan suku kata untuk semua jenis pola suku kata dalam bahasa Indonesia. Hasil dari pengamatan menunjukan bahwa sistem text-to-audiovisual bahasa Indonesia ini dapat memvisualisasikan pengucapan kalimat-kalimat Bahasa Indonesia dengan sangat baik untuk kalimat-kalimat bahasa Indonesia yang mengandung pola suku kata V, VK, KV, KVK. Sistem text-to-audiovisual bahasa Indonesia ini juga dapat memvisualisasikan pengucapan kalimat-kalimat Bahasa Indonesia dengan cukup baik untuk kalimat-kalimat bahasa Indonesia yang mengandung pola suku kata KKV, KKVK, VKK, KVKK dan KKVKK. Sedangkan untuk kalimat-kalimat bahasa Indonesia yang mengandung pola suku kata KKKV dan KKKVK, sistem text-to-audiovisual bahasa Indonesia kurang baik dalam memvisualisasikan pengucapan suku kata tersebut. Ini disebabkan oleh beberapa konsonan yang berurutan. Oleh karena itu, perlu adanya pengaturan fonem konsonan yang harus ditonjolkan maupun disembunyikan. Pengaturan ini dapat dilakukan dengan menerapkan perhitungan jumlah frame untuk tiap fonem dalam pembuatan visualisasi. Untuk penelitian selanjutnya diperlukan formula yang menghitung jumlah frame tiap fonem berdasarkan nilai durasi masing-masing fonem.

UCAPAN TERIMA KASIH

Kami mengucapkan terim kasih secara khusus kepada Dirjen Dikti melalui program Penelitian Hibah Bersaingnya, sehingga kami mendapatkan dukungan dana untuk melakukan penelitian ini. Dengan program ini, kami mempunyai kesempatan lebih leluasa untuk meneliti dan mengembangan keilmuan dibidang Visi Komputer. Kami juga mengucapkan terima kasih kepada pengelola Laboratorium Human Centric System (HCS) Jurusan Teknik Multimedia dan Jaringan Institut Sepuluh Nopember (ITS) Surabaya, sehingga kami dapat belajar dan ikut menggunakan laboratorium tersebut untuk beberapa bulan.

DAFTAR PUSTAKA

[1] Furui, S., "Digital Speech Processing; Synthesis and Recognition", Marcel Dekker Inc., New York, 2001.

[2] Salil Deena, Shaobo Hou and Aphrodite Galata, "Visual Speech Synthesis by Modelling Coarticulation Dynamic using a Non-Parametric Switching State-Space Model", School of Computer Science, university of Mancester, UK, 2010.

[3] Hui Zhao and Chaojing Tang, "Visual Speech Synthesis Based on Chinese Dynamic Visemes", Proceeding of the 2008 IEEE International Conference on Information and Automation, June 20-23, Zhanjiajie, China, 2008.

[4] Arifin, Surya Sumpeno, Mochamad Hariadi, Hanny Haryanto, "A Text-to-Audiovisual Synthesizer for Indonesian by Morphing Viseme", International Review on Computers and Software (IRECOS), Vol. 10, N. 11, ISSN 1828-6003, pp. 1149-1156, November 2015.

[5] Johan Wouters, Michael W. Macon, "Control of Spectral Dynamics in Concatenative Speech Synthesis", IEEE Transaction on Speech and Audio Processing, Vol. 9, No. 1, pp. 30-38, January 2001.

[6] Sarah L. Taylor, Moshe Mahler, Barry-John Theobald and Lain Matthews, "Dynamic Units of Visual Speech", ACM SIGGRAPH Symposium on Computer Animation, 2012.

[7] Arifin, Surya Sumpeno, Mochamad Hariadi, Hanny Haryanto, "A Text-to-Audiovisual Synthesizer for Indonesian by Morphing Viseme", International Review on Computers and Software (IRECOS), Vol. 10, N. 11, ISSN 1828-6003, pp. 1149-1156, November 2015.

[8] 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.

[9] Chaer, Abdul, "Linguistik Umum", Jakarta: PT. Rineka Cipta, 2003.

[10] Turk MA and Pentland AP., "Face Recognation Using Eigenfaces", IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 586-591, 1991.

[11] K.A. Abdul Nazeer, M.P. Sebastian, "Improving the Accuracy and Efficiency of k-means Clustering Algorithm", Proceedings of the World Congress on Engineering, July 1 – 3, London, U.K., Vol I, ISBN : 978-988-17012-5-1, 2009.

[12] T. Larose, "Discovering Knowledge in Data", A John Wiley & Sons, Inc. Publication, USA, pp. 153–157, 2005.

[13] Subaryani D.H. Soedirdjo, Hasballah Zakaria, Richard Mengko, "Indonesian Text-to-Speech Syllable Concatenation for PC-based Low Vision Aid", 2011 International Conference on Electrical Engineering and Informatics, Bandung, Indonesia, 17-19 July 2011.

[14] K. Tokuda, T. Kobayashi, T. Masuko and S. Imai, "Melgeneralizedcepstral analysis — A unified approach to speech spectral estimation," Proc. ICSLP'94, pp.1043– 1046, Sep. 1994.

[15] T. Ezzat and T. Poggio, "Visual Speech Synthesis by Morphing Visemes", International Journal of Computer Vision, vol.38, no.1, pp.45-57, 2000.