JEPIN
(Jurnal Edukasi dan Penelitian Informatika)
ISSN(e): 2548-9364 / ISSN(p) : 2460-0741

Vol. 5
No. 2
Agustus
2019

Pembandingan Tiga Nada Vokal /e/ untuk Animasi Gerak Bibir

Anung Rachman [#1], Risanuri Hidayat [*2], Hanung Adi Nugroho [*3]

[#]*Institut Seni Indonesia (ISI) Surakarta*
Jl. Ring Road, Mojosongo, Jebres, Kota Surakarta, Jawa Tengah 57127
[1]anung.rachman@mail.ugm.ac.id

[*]*Departemen Teknik Elektro dan Teknologi Informasi, Universitas Gadjah Mada*
Jl. Grafika No.2, Yogyakarta, Senolowo, Sinduadi, Mlati, Kota Yogyakarta, Daerah Istimewa Yogyakarta 55281
[2]risanuri@ugm.ac.id
[3]adinugroho@ugm.ac.id

***Abstrak*— Saat ini teknologi animasi gerak bibir tengah berkembang secara signifikan seiring dengan perkembangan industri kreatif. Metode yang sering digunakan untuk membuat gerak bibir tersebut adalah peta fonem ke visem. Pembangunan peta fonem ke visem awalnya mengacu pada ketentuan baku susunan fonem yang sudah ada, namun kemudian susunan ini berkembang mengikuti kebutuhan. Fonem vokal mengambil peran terbesar karena energi percakapan terakumulasi padanya. Variasi pengucapan vokal yang sangat beragam menyebabkan susunan keberadaan fonem vokal pada peta juga beragam. Keragaman ini berimplikasi pada akurasi peta fonem ke visem yang juga berujung pada akurasi gerak bibir animasi. Paper ini membahas tentang signifikansi perbedaan tiga macam nada vokal /e/ Bahasa Indonesia baik dari ciri audio maupun dari ciri visual untuk menunjang susunan baku vokal pada peta fonem ke visem. Metode yang digunakan adalah filter LPC untuk mengekstraksi ciri frekuensi forman, Par-CLR untuk mengekstraksi ciri visual, hingga uji statistik untuk mengetahui signifikansi perbedaan. Hasilnya menunjukkan sebagian nada tersebut memiliki perbedaan signifikan satu sama lain. Sehingga peta fonem ke visem akan lebih akurat jika menyertakan unsur ketiga /e/ tersebut.**

***Kata kunci*— animasi gerak bibir, peta fonem ke visem, ciri audio, ciri visual**

I. PENDAHULUAN

Animasi merupakan salah satu sektor industri kreatif yang sangat berkembang pesat. Ada dua kiblat animasi di dunia yaitu Jepang dan Amerika. Animasi Jepang memiliki ciri khas berupa kekuatan karakter, sedangkan animasi Amerika sangat rinci dalam setiap pergerakannya. Pergerakan yang sangat rinci juga tampak pada dialog khususnya pada bagian mulut. Pergerakan ini memerlukan ketepatan dalam sinkronisasi bibir atau *lipsync*. Kekuatan gerak bibir merupakan bagian dari kualitas sebuah klip animasi. Untuk menghasilkannya, pembuatan gerak bibir selama ini dilakukan secara manual. Data rekaman suara yang sudah tersedia, selanjutnya disinkronisasi dengan pergerakan mulut, *frame by frame*.

Upaya sinkronisasi dengan cara manual tentu saja memerlukan waktu yang lama dan butuh konsentrasi dari pembuatnya agar terlihat sangat rinci. Teknologi gerak bibir otomatis diciptakan untuk membantu pekerjaan tersebut. Tentu saja perkembangan teknologi ini belum dapat menggantikan pekerjaan sinkronisasi secara penuh agar terlihat bagus. Namun munculnya teknologi gerak bibir otomatis sangat membantu dalam pekerjaan dalam bidang animasi yang lebih ringan misalnya untuk klip berdurasi pendek.

Gambar 1. Bagian suara pada teknologi animasi gerak bibir [1]

Gambar 1 menunjukkan teknologi animasi gerak bibir otomatis, pada bagian visual (menggerakkan bibir animasi), sebuah peta fonem ke visem digunakan sebagai referensi visual untuk menghasilkan animasi gerak bibir. Fonem adalah satuan bunyi yang dapat membedakan pengucapan baik pada kata yang sama atau berbeda [2]. Sedangkan visem adalah visual fonem, sama seperti fonem tetapi dalam bentuk visual. Pada teknologi gerak bibir otomatis, ciri suara input diekstrak, lalu berdasarkan peta, ciri fonem yang terdeteksi pada bagian klasifikasi dikonversi menjadi visem.

Pengembangan metode pemetaan fonem ke visem tidak mudah. Terlalu banyak karakteristik yang menyertai sehingga peta yang dihasilkan diantara peneliti bisa berbeda signifikan. Karakteristik tersebut antara lain rumitnya otot wajah (bagian mulut), hingga standar fonem yang bisa berkembang sesuai dengan ragam suku kata yang bisa diucapkan.

Standar fonem Bahasa Indonesia memuat vokal sebanyak 5 buah yaitu /a/./i/,/u/,/e/, dan /o/. Namun jika dilihat lebih lanjut, vokal /e/, dan vokal /o/ memiliki lebih dari satu nada pengucapan. Vokal /e/ untuk kata 'pesta' berbeda nada pengucapan dengan /e/ pada kata 'peristiwa', dan /e/ pada kata 'nenek'. Hal yang sama juga terjadi pada vokal /o/. Kata 'koperasi' memiliki nada /o/ yang berbeda pada kata 'motor'. Jika fonem tersebut diterapkan untuk animasi gerak bibir, terjadi kemungkinan visem menjadi kurang akurat.

Paper ini membahas tentang signifikansi perbedaan antara fonem vokal /e/ pada Bahasa Indonesia untuk tiga macam nada. Jika berbeda signifikan, maka sudah seharusya peta fonem ke visem juga memuat ketiga nada /e/ tersebut agar animasi gerak bibir terlihat lebih rinci.

II. Peta Fonem ke Visem

Pada animasi gerak bibir, beberapa visual mulut memiliki bentuk yang mirip untuk suara yang berbeda. Agar tidak terjadi tumpang tindih, maka beberapa fonem dikelompokkan menjadi satu kelas berdasarkan visem yang mirip. Proses pengelompokkan tersebut dikenal sebagai pemetaan fonem ke visem. Terdapat dua macam pendekatan untuk mengelompokkan visem, yaitu pendekatan melalui ilmu linguistik (termasuk intuisi) dan pendekatan berdasarkan ciri [3].

Peta fonem ke visem telah banyak dihasilkan oleh para peneliti. Metode pemetaan yang digunakan juga bermacam-macam. Menggunakan ilmu linguistik [4] [5], menggunakan metode *decision trees* yang dipadu dengan ilmu linguistik [6], menggunakan *bottom-up clustering* berdasarkan kemiripan model Gaussian pada titik tengah frame visual segmen fonem melalui ciri PCA [7], hingga menggunakan PCA yang dikombinasi dengan LDA untuk mengekstraksi ciri bentuk bibir yang kemudian dikelompokkan melalui algoritme *K-Means* [8]. Hasil peta dari para peneliti tersebut dapat dilihat pada dua tabel berikut, Tabel I adalah pemetaan visem pada Bahasa Inggris, dan Tabel II adalah pemetaan visem pada Bahasa Indonesia.

TABEL I
PETA FONEM KE VISEM BAHASA INGGRIS

Peta	Kelompok Visem
Neti dkk [6]	[sil,sp], [ao,ah,aa,er,oy,aw,hh], [uw,uh,ow], [ae,eh,ey,ay], [ih,iy,ax], [l,el,r,y], [s,z], [t,d,n,en], [sh,zh,ch,jh], [p,b,m], [th,dh], [f,v], [ng,k,g,w]
Lee dan Yook [9]	[b,p,m], [d,t,s,z,th,dh], [g,k,n,ng,l,y,hh], [jh,ch,sh,zh], [f,v], [r,w], [iy,ih], [eh,ey,ae], [aa,aw,ay,ah], [ao,oy,ow], [uh,uw], [er], [sil]
Hazen dkk [7]	[sil], [ax,ih,iy,dx], [ah,aa], [ae,eh,ay,ey,hh], [aw,uh,uw,ow,ao,w,oy], [el,l], [er,axr,r], [y], [b,p], [bcl,pcl,m,em], [s,z,epi,tcl,dcl,n,en], [ch,jh,sh,zh], [t,d,th,dh,g,k], [f,v], [gcl,kcl,ng]
Bozkurt dkk [4]	[sil], [ay,ah], [ey,eh,ae], [er], [ix,iy,ih,ax,axr,y], [uw,uh,w], [ao,aa,oy,ow], [aw], [g,hh,k,ng], [r], [l,d,n,en,el,t], [s,z], [ch,sh,jh,zh], [th,dh], [f,v], [m,em,b,p]

TABEL II
PETA FONEM KE VISEM BAHASA INDONESIA

Peta	Kelompok Visem
Arifin dkk [8]	[sil], [a,h], [p,b,m], [d,t,n,l,r], [o,au,u,w], [k,g,kh], [c,j,s,i,z,sy,ny], [E,y,oi,ai], [f,v],[ng,e]
Setyati dkk [5]	[b,p,m], [f,v,w,ph], [d,dh,dl,dz,l,n,t,th], [r], [c,j,s,z,ts,ps,ks,sh,sy,x,y,ny], [g,gh,h,k,kh,q,ng,ngg], [a], [i,I,oi], [∂, e, ε,ai], [o,O,au], [u,U], [sil]
Liyanthy dkk [10]	[a], [i], [u], [e], [o], [b,m,p], [c,j,s,x,z], [d], [f,v], [g,h,k,n,q,r,y], [l], [t], [w]

Vokal memiliki peran paling dominan pada pengenalan tutur maupun animasi gerak bibir, karena energi terbesar dan durasi waktu terpanjang pada sinyal tutur terakumulasi pada vokal [1]. Terlihat pada Tabel I dan Tabel II, keberadaan vokal maupun susunannya berbeda-beda. Perbedaan ini terjadi akibat kerumitan yang ada pada otot wajah sehingga membuat gaya pengucapan sangat bervariasi. Variasi pengucapan tersebut dikenal sebagai alofon [11]. Terlihat pada Tabel II, tiga peneliti menghasilkan peta dengan posisi vokal /e/ Bahasa Indonesia yang berbeda satu sama lain. Arifin dkk [8] menggunakan dua macam /e/ yang berbeda kelas, sedangkan peta Setyati dkk [5] memuat tiga macam /e/ dan menempatkannya dalam satu kelas bersama diftong /ai/. Dan pada peta fonem ke visem Liyanthy dkk [10] terdapat satu macam /e/ serta menempatkannya dalam satu kelas tersendiri.

Keberadaan vokal alofon tentu saja akan mempengaruhi kinerja peta fonem ke visem atau akurasi gerak bibir animasi. Perbedaan keberadaan dan susunan vokal /e/ mengkonfirmasi bahwa terdapat metode pemetaan yang menghasilkan peta kurang akurat. Untuk mengetahuinya, paper ini mengusulkan metode untuk memvalidasinya berupa perbedaan pada ciri audio dan ciri visual.

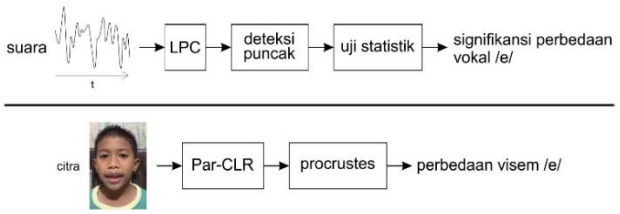
III. Metodologi

A. Ikhtisar

Perbandingan terhadap ketiga vokal /e/ dilakukan dalam dua macam yaitu perbandingan ciri audio dan ciri visual seperti yang terlihat pada Gambar 2. Pada perbandingan ciri audio, metode yang digunakan adalah melalui filter LPC [12] untuk menghasilkan ciri frekuensi forman. Dengan filter ini, akan diketahui sinyal informasi dari input audio vokal /e/. Selanjutnya pada setiap resonansi pada sinyal informasi akan dideteksi bagian puncaknya. Puncak sinyal setiap resonansi ini yang disebut sebagai frekuensi

forman. Ciri khas dari sebuah sinyal audio terletak pada forman pertama dan kedua (F1 dan F2) [13] dari banyak forman yang dideteksi.

Sedangkan perbandingan terhadap ciri visual dilakukan melalui algortima deteksi citra yaitu Par-CLR [14]. Algoritma ini akan menghasilkan ciri-ciri visual bagian mulut berupa titik koordinat yang disebut sebagai *landmark*. Untuk melihat perbedaan diantara *landmark* vokal /e/, langkah selanjutnya adalah mengukurnya melalui analisis *procrustes* [15].



Gambar 2. Ikhtisar pembandingan ciri suara vokal /e/ (atas), dan pembandingan ciri visual citra mulut vokal /e/ (bawah).

B. Data

Pada bagian suara, data yang digunakan adalah 10 rekaman suara untuk satu macam vokal /e/ di satu posisi sebuah kata. Vokal /e/ ditentukan ada pada dua posisi, yaitu bagian depan dan bagian belakang pada sebuah kata. Penentuan posisi ini dimaksudkan agar rekaman mengandung ciri artikulasi atau nada transisi dari suatu suku kata terhadap suku kata disebelahnya. Karena ditentukan dua posisi dalam sebuah kata, maka jumlah rekaman suara adalah 20 data untuk satu macam vokal /e/. Sedangkan vokal /e/ yang dicari perbedaannya adalah tiga macam yaitu /e1/,/e2/, dan /e3/. Sehingga jumlah total adalah sebanyak 60 data rekaman kata yang mengandung vokal /e/. Proses perekaman menggunakan *sample rate* 24KHz, kanal mono, kedalaman bit 32, dan menggunakan format MP3.

Pada bagian visual, proses pengambilan data berawal dari perekaman video terhadap seseorang yang mengucapkan tiga macam vokal /e/ dengan sudut pengambilan di area wajah. Rekaman video kemudian diekstrak menjadi rangkaian citra. Setiap macam rangkaian vokal /e/ kemudian diambil satu sampel daerah tengah pengucapan. Resolusi citra adalah 720x1280 piksel sesuai resolusi rekaman video.

C. Ciri Audio Vokal /e/

Linear Predictive Coding (LPC) [12] merupakan filter resonansi yang meniru sistem pembentuk nada yang ada pada manusia ketika menghsilkan suara. Filter tersebut merupakan tiruan semacam bibir, lidah, hingga saluran suara pada mulut manusia, terhadap sumber suara yaitu pita. Setiap organ pada mulut pembentuk nada tersebut menghasilkan jeda suara yang terbentuk antara satu organ terhadap organ disebelahnya. Jika sinyal suara yang mengandung jeda ini adalah $c(n-t), t = 1 \dots p$, sinyal suara yang dihasilkan oleh mulut manusia adalah $\ddot{c}$,

$$\ddot{c}(n) = a_1 c(n-1) + a_2 c(n-2) + \cdots + a_p c(n-p) \quad (1)$$

$\ddot{c}$(n) adalah suara yang dihasilkan dari sumber suara c(n), dengan a sebagai koefisien.

Persamaan di atas dapat disederhanakan menjadi,

$$\ddot{c}(n) = \sum_{t=1}^{p} a_t \, c(n-t) \quad (2)$$

Untuk mencari persamaan filter, maka persamaan linier diubah dalam bentuk domain z. Filter $f(z)$ adalah,

$$f(z) = \sum_{t=1}^{p} a_t z^{-t} \quad (3)$$

Setiap prediksi linier, selalu saja menghasilkan kesalahan prediksi yaitu selisih dari sinyal prediksi terhadap sinyal input. Jika kesalahan prediksi adalah e, maka,

$$e(n) = c(n) - \ddot{c}(n) \quad (4)$$

dalam domain z,

$$e(z) = a(z)c(z) \quad (5)$$

dengan $a(z)$ adalah inverse dari $f(z)$, sehingga,

$$a(z) = 1 - f(z) \quad (6)$$

dan filter LPC, $H(z)$, adalah,

$$H(z) = \frac{1}{a(z)} = \frac{1}{1 - f(z)} = \frac{1}{1 - \sum_{t=1}^{p} a_t z^{-t}} \quad (7)$$

Filter $H(z)$ menghasilkan sinyal informasi berbentuk resonansi. Frekuensi forman terletak pada setiap puncak sinyal resonansi tersebut, dari f1, f2, hingga f seterusnya. Ciri-ciri audio yang membedakan banyak terletak pada dua frekuensi forman pertama (f1 dan f2) [13].

D. Ciri Visem Vokal /e/

Untuk membandingkan visem, ciri citra yang digunakan adalah *landmark*, yaitu titik-titik koordinat yang menunjukkan area mulut. *Landmark* dihasilkan melalui metode *Parallel Cascade of Linear Regression* (Par-CLR) [14]. Metode ini merupakan pengembangan dari metode Seq-CLR yang perberdaaanya terletak pada proses menghasilkan titik koordinat [16].

Cara kerja Par-CLR adalah sebagai berikut. Sebuah fungsi f akan mendeteksi gambar wajah j_k untuk menghasilkan vektor ciri q sebagai parameter bentuk wajah, x_1^k. Setiap j_k yang telah mengandung vektor ciri dinotasikan sebagai x_*^k. Fungsi yang dapat digunakan

untuk mendeteksi gambar wajah $f(x^k)$ antara lain SIFT [17] atau *Histogram of Oriented Gradient* (HOG) [18]. Titik *landmark* akan dihasilkan melalui algoritme Seq-CLR sebagai berikut,

$$\underset{C_m, b_m}{\arg min} \sum_{J_k} \sum_{x_m^k} \|x_*^k - x_m^k - C_m f(x_m^k) - b_m\|^2 \quad (8)$$

C_m adalah ciri *landmark*, sedangkan b adalah bias.

E. Procrustes

Analisis *Procrustes* [15] merupakan metode statistika yang digunakan untuk menganalisa distribusi pada sekumpulan obyek bentuk. Melalui dua obyek bentuk, *Procrustes* membandingkan satu sama lain melalui transformasi berupa translasi, rotasi dan skala.

1) Translasi: Translasi adalah memindahkan dua obyek bentuk yang diperbandingkan ke titik tengah (*centroid* / titik nol), sehingga kedua obyek tersebut berada pada titik tengah yang sama. Jika terdapat obyek bentuk dengan titik *landmark* sejumlah k,

$$\big((x_1, y_1), (x_2, y_2), \dots, (x_k, y_k)\big) \quad (9)$$

Maka rerata dari kedua obyek bentuk tersebut adalah,

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_k}{k}, \overline{y} = \frac{y_1 + y_2 + \cdots + y_k}{k} \quad (10)$$

$(x, y) \rightarrow (x - \overline{x}, y - \overline{y})$ menghasilkan titik $(x_1 - \overline{x}, y_1 - \overline{y})$,...

2) Skala: Setelah ditranslasi, terhadap ke dua obyek selanjutnya dilakukan penskalaan untuk membuat keduanya dalam besaran bentuk yang sama. Penskalaan (s) melalui metode *Root Mean Square Distance* (RMSD) sehingga menghasilkan nilai 1 terhadap keduanya.

$$s = \sqrt{\frac{(x_1 - \overline{x})^2 + (y_1 - \overline{y})^2 + \cdots}{k}} \quad (11)$$

Nilai akan menjadi 1 ketika titik koordinat dibagi dengan nilai skala (s),

$$\left(\frac{(x_1 - \overline{x})}{s}, \frac{(y_1 - \overline{y})}{s}\right) = 1 \quad (12)$$

3) Rotasi: Jika terdapat titik koordinat dua obyek bentuk adalah $\big((x_1, y_1), \dots\big), ((w_1, z_1), \dots)$, maka salah satu bentuk dapat digunakan sebagai referensi orientasi perputaran hingga mencapai sudut putar yang optimal mengacu pada tingkat ketidakmiripan, *Sum of the Squared Distances* (SSD) yang paling minimal. Sebuah rotasi dengan sudut θ adalah sebagai berikut,

$$(u_1, v_1) = (cos\theta w_1 - sin\theta z_1, sin\theta w_1 + cos\theta z_1) \quad (13)$$

dengan (u,v) adalah koordinat titik rotasi.

Jika $(u_1 - x_1)^2 + (v_1 - y_1)^2 + \cdots = 0$ maka,

$$\theta = tan^{-1}\left(\frac{\sum_{i=1}^{k}(w_i y_i - z_i x_i)}{\sum_{i=1}^{k}(w_i x_i - z_i y_i)}\right) \quad (14)$$

4) Perbandingan bentuk (kemiripan): Setelah melalui serangkaian superimpose terhadap dua obyek bentuk berupa translasi, skala, dan rotasi, maka tingkat kemiripan (SSD) yang dikenal sebagai *Procrustes Distance* dapat diketahui melalui persamaan,

$$d = \sqrt{(u_1 - x_1)^2 + (v_1 - y_1)^2 + \cdots} \quad (15)$$

Pengukuran kemiripan d menghasilkan nilai antara 0 dan 1 yang menggambarkan perbedaan antara suatu obyek bentuk yang menjadi acuan dan obyek bentuk hasil transformasi. Nilai mendekati 0 diartikan dengan mirip, dan jika mendekati 1 diartikan tidak mirip.

IV. HASIL DAN PEMBAHASAN

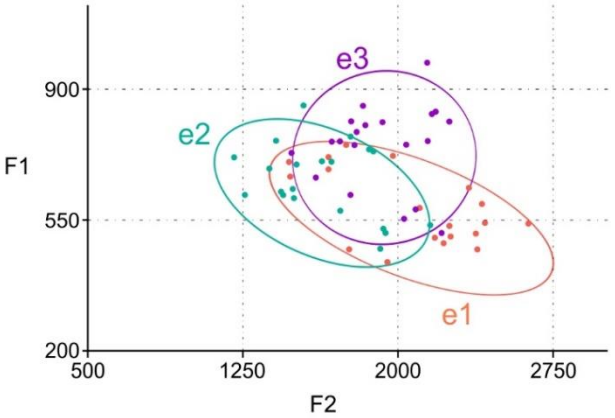
A. Ciri Audio Vokal /e/

Hasil ekstraksi ciri audio pada vokal /e/ terlihat seperti ditunjukkan pada Tabel III. Tabel tersebut terdiri dari vokal /e1/, /e2/, dan /e3/ pada sebuah kata baik pada posisi depan maupun belakang. Sebelum ciri vokal diekstrak, setiap rekaman kata diobservasi untuk mengetahui letak vokal /e/ pada sebuah spektogram, lalu diseleksi sesuai posisi yang ditentukan, dan selanjutnya ciri diekstrak pada bagian tengah vokal yang sudah diseleksi. Frekuensi forman diambil sebagai data hanya pada resonansi pertama hingga ketiga, yaitu F1, F2, dan F3.

TABEL III
CIRI FREKUENSI FORMAN VOKAL /E/ PADA SAMPEL KATA

Vokal /e/ pada kata	Forman (Hz)		
	F1	F2	F3
H/**e1**/ran	655	1663	2254
M/**e1**/dan	674	1469	2455
B/**e1**/cak	439	2382	2758
Sat/**e1**/	482	2374	2668
Jah/**e1**/	603	2342	2759
Kaf/**e1**/	473	2251	2683
S/**e2**/kolah	592	1434	1935
P/**e2**/lajar	686	1207	2432
K/**e2**/marin	601	1491	1708
Put/**e2**/r	706	1863	2475
Samb/**e2**/l	587	1443	3141
Ap/**e2**/s	666	1506	2964
N/**e3**/nek	698	1480	2354
B/**e3**/bek	483	2206	2500
H/**e3**/rbal	728	1717	2263
Cap/**e3**/k	780	1924	2708
Sem/**e3**/n	807	2179	2826
Polr/**e3**/s	720	2038	2310

Terlihat nilai forman bisa berbeda-beda hingga selisihnya lebih dari 200 (misalnya pada F1 antara vokal /e1/ pada kata "heran" dan /e1/ pada kata "becak"), padahal sama-sama dari sinyal suara vokal /e/, faktor artikulasi rupanya sangat mempengaruhi perbedaan ini. Plot distribusi nilai forman F1 dan F2 vokal /e/ pada Gambar 3 dapat digunakan untuk melihat tingkat perbedaan lebih jauh.



Gambar 3. Plot distribusi frekuensi forman F1 dan F2 untuk vokal /e1/, /e2/, dan /e3/

Gambar 3 menunjukkan letak F1 dan F2 untuk ketiga vokal /e/. Distribusi yang berdekatan bahkan saling tumpang tindih mengindikasikan kemiripan diantara ketiganya. Meskipun demikian, ketiganya memiliki ciri yang cenderung sedikit berbeda yaitu, forman vokal /e1/ yang cenderung pada frekuensi F2 tinggi namun F1 rendah, /e2/ yang cenderung tinggi untuk F1 namun rendah pada F2, dan /e3/ yang memiliki kecenderungan tinggi untuk F1 dan F2.

Untuk mengetahui signifikansi perbedaan ciri audio, maka dilakukan uji statistik terhadap frekuensi forman ketiga vokal /e/. Model yang digunakan adalah *one way Anova* dengan uji HSD (*Honestly Significantly Different*) *Tukey* pada rerata frekuensi forman.

TABEL IV
NILAI PERBEDAAN TIGA VOKAL /E/ PADA F1

Kontras	Perbedaan	Standar perbedaan	Pr > Diff	Signifikan
e3 vs e1	158,792	4,801	< **0,0001**	Ya
e3 vs e2	92,850	2,807	**0,018**	Ya
e2 vs e1	65,943	1,994	0,123	Tidak

TABEL V
NILAI PERBEDAAN TIGA VOKAL /E/ PADA F2

Kontras	Perbedaan	Standar perbedaan	Pr > Diff	Signifikan
e1 vs e2	429,101	4,891	< **0,0001**	Ya
e1 vs e3	140,188	1,598	0,255	Tidak
e3 vs e2	288,913	3,293	**0,005**	Ya

TABEL VI
NILAI PERBEDAAN TIGA VOKAL /E/ PADA F3

Kontras	Perbedaan	Standar perbedaan	Pr > Diff	Signifikan
e2 vs e3	198,313	1,868	0,157	Tidak
e2 vs e1	40,116	0,378	0,924	Tidak
e1 vs e3	158,197	1,490	0,303	Tidak

Tabel IV, V, dan VI adalah hasil uji statistik *Tukey* yang menampilkan nilai perbedaan diantara ketiga vokal /e/. Batas signifikansi 5% digunakan untuk menentukan nilai kritis q yang dibandingkan terhadap standar perbedaan diantara rerata (Pr>Diff). Ketiga tabel tersebut jika disimpulkan akan tampak seperti pada Tabel VII di bawah.

TABEL VII
SIGNIFIKANSI PERBEDAAN CIRI AUDIO TIGA VOKAL /E/

Vokal /e/		Beda signifikan		
		F1	F2	F3
e1	e2	Tidak	Ya	Tidak
e2	e3	Ya	Ya	Tidak
e3	e1	Ya	Tidak	Tidak

Hasil pengujian pada Tabel VII memperlihatkan bahwa forman F3, tidak berbeda signifikan untuk tiga macam vokal /e/. Hal ini sekaligus juga mengkorfimasi bahwa ciri forman hanya terletak pada dua frekuensi pertama [13].

Kondisi ideal terjadi ketika F1 dan F2 terjadi perbedaan signifikan seperti yang terjadi antara vokal /e2/ dan /e3/. Meskipun demikian, perbandingan vokal /e/ yang lain terdapat perbedaan signifikan pada salah satu frekuensi forman sehingga diantara ketiga vokal /e/ tersebut sudah seharusnya dimasukkan menjadi bagian dari peta fonem ke visem.

B. Ciri Visem Vokal /e/

Ciri visual berupa *landmark* diekstrak melalui metode Par-CLR. *Landmark* ditentukan khusus pada bagian mulut sebanyak 18 yang terdiri dari 12 titik lingkar luar dan sisanya berada di lingkar dalam mulut seperti terlihat pada Gambar 4.

Gambar 4. *Landmark* pada bagian mulut

Landmark berupa titik koordinat (x,y) untuk tiga visem /e/ terlihat pada Tabel VIII.

TABEL VIII
LANDMARK VISEM /E/

/e1/		/e2/		/e3/	
x	**y**	**x**	**y**	**x**	**y**
276,55	767,21	277,16	757,78	278,96	760,91
302,57	736,14	306,86	731,14	303,76	730,72
335,73	719,65	340,68	717,77	338,72	715,81
369,29	726,18	373,44	724,88	372,62	720,76
402,12	719,59	403,55	716,84	405,76	714,33
434,83	735,92	436,03	730,76	438,31	729,44
465,09	762,62	467,51	755,37	466,98	756,99
436,61	797,15	439,70	785,34	442,29	794,30
405,82	817,93	408,60	802,59	410,90	816,66
370,78	824,12	374,50	807,50	374,27	823,05
337,24	817,73	339,86	802,40	338,99	816,44
305,08	799,33	307,40	785,86	305,62	795,61
336,36	756,01	340,34	753,91	338,77	750,69
369,53	757,22	373,52	755,39	373,13	750,86
404,14	755,85	405,31	752,89	408,64	749,74
405,26	776,12	406,21	760,31	409,82	774,53
370,79	781,88	373,74	765,59	374,09	780,18
338,08	777,63	340,36	760,41	340,33	774,80

Untuk mengetahui tingkat kemiripan diantara ketiga visem /e/, langkah selanjutnya adalah mengukur menggunakan analisis *procrustes*, dan hasilnya tampak seperti pada Tabel IX berikut ini.

TABEL IX
NILAI PROCRUSTES VISEM /E/

	/e1/	**/e2/**	**/e3/**
/e1/	0	0,008925069	0,000898236
/e2/	0,008925069	0	0,013067376
/e3/	0,000898236	0,013067376	0

Procrustes menghitung transformasi terbaik dari suatu bentuk terhadap bentuk lainnya menggunakan jarak *Euclidean*. Nilai 0 mengindikasikan bahwa dua bentuk yang dibandingkan sama persis (tidak ada jarak), dan nilai 1 adalah sebaliknya.

Rerata nilai procrustes vokal /e/ adalah 0,00763. Pada nilai Tabel IX, jika menggunakan batas rerata ini maka perbedaan ada pada /e1/ dan /e2/ serta /e2/ dan /e3/. Jika perbedaan ciri audio dan ciri visual vokal /e/ dibandingkan sesuai Tabel VII dan Tabel IX, hasilnya terlihat seperti Tabel X.

TABEL X
SIGNIFIKANSI PERBEDAAN CIRI TIGA VOKAL /E/

Vokal /e/		**Beda signifikan**	
		Ciri Audio	**Ciri Visual**
e1	e2	Ya	Ya
e2	e3	Ya	Ya
e3	e1	Ya	Tidak

Tabel X memperlihatkan perbedaan signifikan diantara ketiga vokal /e/ untuk ciri audio dan ciri visual. Pada ciri audio, F1 dan F2 adalah satu paket, sehingga jika salah satu F atau kedua F berbeda signifikan, maka vokal /e/ dianggap berbeda. Perbedaan ciri audio pada tabel tersebut dapat diartikan bahwa ketiga vokal /e/ seharusnya ada pada peta fonem ke visem. Sedangkan untuk kelas visem, vokal /e1/ dan /e3/ dapat dikelompokkan menjadi satu kelas. Hasil ini jika dibandingkan dengan peta fonem ke visem yang sudah ada seperti pada Tabel II, maka peta Arifin dkk [8] paling mendekati, meskipun hanya memuat /e1/ dan /e2/, namun kedua /e/ tersebut terletak pada kelas visem yang berbeda.

V. KESIMPULAN

Vokal /e/ pada Bahasa Indonesia memiliki tiga macam nada yaitu /e1/, /e2/, dan /e3/. Ketiga vokal tersebut memiliki ciri audio yang berbeda dibuktikan dengan hasil uji yang menunjukkan perbedaan signifikan. Sedangkan pada ciri visual, visem /e1/ mirip terhadap /e3/, namun keduanya tidak mirip terhadap visem /e2/.

Perbedaan ini tentu saja harus diakomodasi ketika membangun teknologi gerak bibir otomatis. Ketiga ciri nada tersebut jika dimasukkan pada basis data referensi akan membuat hasil gerak bibir menjadi lebih akurat. Selain itu sistem juga seharusnya memasukkan basis data visem minimal dua citra untuk visem /e/. Penelitian berikutnya yang berkaitan dengan peta fonem ke visem Bahasa Indonesia dapat dikembangkan dengan memasukkan ketiga unsur vokal /e/.

REFERENSI

[1] S.-M. Hwang, H.-K. Yun, and B.-H. Song, "Korean Speech Recognition Using Phonemics for Lip-Sync Animation," in *Information Science, Electronics and Electrical Engineering (ISEEE)*, 2014, pp. 1011–1014.

[2] International Phonetic Association and I. P. Association, *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*. Cambridge University Press, 1999.

[3] L. Cappelletta and N. Harte, "Phoneme-to-Viseme Mapping for Visual Speech Recognition," in *1st International Conference on Pattern Recognition Applications and Methods (ICPRAM) Volume 2*, 2012, pp. 322–329.

[4] E. Bozkurt, Ç. E. Erdem, E. Erzin, T. Erdem, and M. Özkan, "Comparison of Phoneme and Viseme Based Acoustic Units for Speech Driven Realistic lip Animation," in *3DTV Conference*, 2007, pp. 1–4.

[5] E. Setyati, S. Sumpeno, M. H. Purnomo, K. Mikami, M. Kakimoto, and K. Kondo, "Phoneme-Viseme Mapping for Indonesian Language Based on Blend Shape Animation," *IAENG Int. J. Comput. Sci.*, vol. 42, no. 3, pp. 233–244, 2015.

[6] C. Neti *et al.*, "Audio-Visual Speech Recognition," *Work. Final Rep.*, p. 764, 2000.

[7] T. J. Hazen, K. Saenko, C.-H. La, and J. R. Glass, "A Segment-based Audio-visual Speech Recognizer: Data Collection, Development, and Initial Experiments," in *Proceedings of the 6th International Conference on Multimodal Interfaces*, 2004, pp. 235–242.

[8] Arifin, Muljono, S. Sumpeno, and M. Hariadi, "Towards Building Indonesian Viseme: A Clustering-Based Approach," in *IEEE International Conference on Computational Intelligence and Cybernetics (CYBERNETICSCOM)*, 2013, pp. 57–61.

[9] S. Lee and D. Yook, "Audio-to-Visual Conversion Using Hidden Markov Models," in *PRICAI 2002: Trends in Artificial Intelligence*, 2002, pp. 563–570.

[10] M. Liyanthy, H. Nugroho, and W. Maharani, "Realistic Facial Animation Of Speech Synchronization For Indonesian Language," in *2015 3rd International Conference on Information and Communication Technology (ICoICT)*, 2015, pp. 563–567.

[11] J. Xu, J. Pan, and Y. Yan, "Agglutinative language speech recognition using automatic allophone deriving," *Chinese J.*

Electron., vol. 25, no. 2, pp. 328–333, 2016.
[12] D. O'Shaughnessy, "Linear predictive coding," *IEEE Potentials*, vol. 7, no. 1, pp. 29–32, 1988.
[13] P. Ladefoged and K. Johnson, *A Course in Phonetics*, 7th ed. Stamford: Cengage Learning, 2014.
[14] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, "Incremental Face Alignment in the Wild," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 1859–1866.
[15] C. Goodall, "Procrustes Methods in the Statistical Analysis of Shape," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 53. WileyRoyal Statistical Society, pp. 285–339, 1991.
[16] X. Xiong and F. De la Torre, "Supervised Descent Method and Its Applications to Face Alignment," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 532–539.
[17] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.
[18] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," in *Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005, vol. 1, pp. 886–893.