AUTO LIP-SYNC PADA KARAKTER *VIRTUAL* 3 DIMENSI MENGGUNAKAN *BLENDSHAPE*

Matahari Bhakti Nendya[1] dan Syahri Mu'min[2]
[1] Program Studi Animasi, Jurusan Televisi
Fakultas Seni Media Rekam, ISI Yogyakarta
No. *HP.* 085643507654, *E-mail*: dida.nendya@gmail.com
[2] Jurusan Teknik Elektro, Fakultas Teknologi Industri
Institut Teknologi Sepuluh Nopember, Surabaya
E-mail: syahri88@gmail.com

ABSTRAK

Proses pembuatan karakter virtual 3D yang dapat berbicara seperti manusia merupakan tantangan tersendiri bagi animator. Problematika yang muncul adalah dibutuhkan waktu lama dalam proses pengerjaan serta kompleksitas dari berbagai macam fonem penyusun kalimat. Teknik *auto lip-sync* digunakan untuk melakukan pembentukan karakter virtual 3D yang dapat berbicara seperti manusia pada umumnya. *Preston blair phoneme series* dijadikan acuan sebagai pembentukan *viseme* dalam karakter. Proses pemecahan fonem dan sinkronisasi *audio* dalam *software* 3D menjadi tahapan akhir dalam proses pembentukan auto *lip-sync* dalam karakter virtual 3D.

Kata kunci: *lip-sync, blendshapes*, karakter virtual 3D, fonem, *viseme*, animasi

ABSTRACT

Auto Lip-Sync on 3D Virtual Character Using Blendshape. *Process of making a 3D virtual character who can speak like humans is a challenge for the animators. The problem that arise is that it takes a long time in the process as well as the complexity of the various phonemes making up sentences. Auto lip-sync technique is used to make the formation of a 3D virtual character who can speak like humans in general. Preston Blair phoneme series used as the reference in forming viseme in character. The phonemes solving process and audio synchronization in 3D software becomes the final stage in the process of auto lip-sync in a 3D virtual character.*
Keywords: lip-sync, blendshapes, 3D virtual character, phoneme, viseme, animation

PENDAHULUAN

Lip-sync atau *lip synchronization* merupakan istilah teknis dari pencocokan gerakan bibir dengan *audio* atau pengucapan vokal yang telah direkam sebelumnya. Dalam ranah animasi, *lip-sync* dapat disebut sebagi seni dalam membuat animasi karakter yang dapat berbicara berdasarkan *track* rekaman atau dialog secara tepat. Teknik *lip-sync* dalam animasi pertama kali diperkenalkan oleh Max Fleischer pada tahun 1926 dalam *My Old Kentucky Home* dan penggunaanya terus berlanjut hingga sekarang. Beberapa film animasi dan acara televisi seperi *Sherk, Lilo & Stich,* dan *The Simpsons* menggunakan teknik *lip-sync* untuk membuat karakter buatan mereka berbicara. Teknik *lip-sync* juga digunakan dalam komedi seperti *This Hour Has 22 Minutes* dan satir politik lainnya dengan melakukan penggantian secara keseluruhan atau sebagian dari *track* dialog yang telah direkam. *Lip-sync* juga digunakan untuk melakukan terjemahan film dari satu bahasa ke bahasa lainnya misalnya dalam *Spirited Away*. *Lip-*

sync dapat menjadi masalah yang sulit dalam melakukan terjemahan karya asing untuk rilis domestik. Terjemahan yang sederhana sering memunculkan *overrun* dan *underrun* dalam dialog tinggi terhadap pergerakan mulut.

Dalam produksi film, *lip-sync* dimasukkan dalam bagian *post-production*. *Dubbing* film dengan bahasa asing dan membuat animasi karakter yang dapat berbicara menggunakan teknik *lip-sync*. Beberapa *video game* menggunakan *lip-sync* dalam karakter untuk memberikan efek lingkungan yang *immersive*.

Penelitian ini mengusulkan pembentukan karakter virtual yang dapat melakukan pembicaraan dengan menggunakan teknik *lip-sync* berbasis *blendshape*. Pembobotan *blendshape* digunakan untuk membentuk *viseme* dari karakter virtual yang kemudian dilakukan sinkronisasi dengan *file audio* dan teks supaya karakter virtual tersebut dapat berbicara.

TINJAUAN PUSTAKA

Blendshape

Blendshape atau interpolasi *shape* merupakan salah satu metode yang sering dipakai dalam pembentukan animasi wajah. Model *blendshape* melakukan penyederhanaan penjumlahan total bobot linier dari *shape* yang memiliki bentuk sama dengan bentuk *shape* asli. Pendekatan matematis untuk melakukan penjumlahan bobot dapat dilihat dalam model matematika sebagai berikut:

$$v_j = \sum w_k b_{kj}$$

Diketahui v_j merupakan titik vertek ke-j hasil animasi model, w_k merupakan nilai bobot percampuran, dan b_{kj} merupakan nilai vertek ke-j dari *blendshape* ke-k. Penjumlahan total bobot tersebut dapat bentuk dan dilakukan model simulasi ke dalam model *polygonal* atau ke dalam model titik kendali *spline*. Bobot dikendalikan oleh animator dalam bentuk *slider* perbobotnya atau dapat diperoleh secara otomatis dengan penerapan sebuah algoritma (Deng, Chiang, Fox & Neuman, 2006:43-48). Metode ini sampai sekarang masih banyak digunakan dalam beberapa proyek film animasi, seperti: *Stuart Little, Star Wars*, dan *Lord of the Rings*, bahkan beberapa perangkat lunak komersial animasi seperti Maya dan 3D Studio Max mengadopsi metode ini. Contoh sederhana adalah interpolasi antara dua *keyframes* saat posisi ekstrem di interval waktu tertentu.

Interpolasi linier sering dipakai untuk penyederhanaan (Bergeron & Lachapelle, 1985:1-19); (Pighin, Hecker, Lischinski, Szeliski, & Salesin, 1988:75-84), namun sebuah fungsi interpolasi *cosinus* (Waters & Levergood, 1993:12) atau variasi lainnya seperti *spline* yang mampu menunjukkan terjadinya efek percepatan dan perlambatan saat pada awal ataupun pada akhir animasi. Ketika ada empat *keyframes* yang terlibat, bukannya dua, interpolasi bilinier menghasilkan banyak variasi ekspresi wajah daripada interpolasi linier (Parke, 1974:5). Interpolasi bilinier jika digabungkan dengan perubahan *morphing* citra secara simultan akan menghasilkan cakupan perubahan ekspresi wajah yang luas (Arai, Kurihara, & Anjyo, 1996:105-116).

Gambar 1. Interpolasi linier dilakukan pada *blendshapes*. Kiri: pose netral, kanan: pose dengan bentuk mulut "A", dan tengah: hasil interpolasinya.

Citra hasil interpolasi didapat dari pengubahan parameter fungsi interpolasi. Interpolasi geometri secara langsung mengubah posisi titik 2D atau 3D dari *mesh* wajah ketika parameter fungsi kendali interpolasi tidak secara langsung memindahkan titik. Contoh, Sera et al. (1996:207-212) menggunakan interpolasi linier parameter kontraksi otot, bukannya posisi titik, untuk mendapatkan animasi mulut dalam berbicara.

Pengembangan terkini dari metode *blendshape* mencoba meningkatkan efisiensi produksi gerak otot berbasis animasi *blend shape* (Choe & Ko, 2001:12-19); (Sifakis, Neverov, & Fedkiw, 2005:417-425). Metode *pose space deformation* (*PSD*) yang dikenalkan oleh Lewis et al. (Lewis, Cordner, & Fong, 2000:165-172) mampu menghasilkan kerangka kerja umum *example-based interpolation* yang dapat digunakan untuk animasi wajah metode *blendshape*. Dalam penelitiannya, deformasi permukaan wajah diperlakukan sebagai sebuah fungsi himpunan parameter abstrak, seperti: {tersenyum, alis naik,...}, dan sebuah permukaan baru yang dihasilkan oleh interpolasi data yang tersebar.

Meskipun interpolasi sangat cepat dan mampu menghasilkan animasi wajah, kemampuannya untuk membentuk konfigurasi wajah realistik yang bercakupan luas tidak dapt dilakukan. Kombinasi gerak wajah yang bebas sulit dihasilkan dan *non-orthogonal blend shapes* saling memengaruhi masing-masing sehingga animator harus kembali dan memperbaiki nilai bobot *blend shapes*. Lewis et al. (2005:25-29) mempresentasikan teknik antarmuka pengguna untuk otomatisasi pengurangan pengaruh *blendshape*. Deng et al. (2006:43-46) mempresentasikan teknik otomatisasi pemetaan data *motion capture* yang langka untuk perancangan awal model wajah *blendshape* 3D dengan terlebih dahulu dilakukan pembelajaran *radial basis function* berbasis regresi.

Phoneme dan *Viseme* (*Visual Phoneme*)

Phoneme (fonem) merupakan bagian linguistik dari sistem wicara secara akustik. Fonem mewakili suara konstrastif dari bahasa sehingga dapat digunakan secara tegas untuk menuliskan ucapan dalam sistem wicara. Secara visual bentuk fonem tidak didefinisikan secara pasti. Fisher pada tahun 1986 mendefinisikan bentuk visual dari fonem yang disebut dengan viseme (*visual phoneme*) (Fisher, 1986:796-800). Model identifikasi fonem dilakukan dengan mengelompokkan fonem berdasarkan artikulasi visual yang sama kemudian dikelompokkan dalam bentuk *viseme* tunggal. Pengelompokan fonem berdasarkan bentuk visual dilakukan secara subjektif (Fisher, 1968:796 - 800; Walden, Prosek, Montgomery, Scherr, & C. J. Jones, 1977:130-145) dan objektif (Goldschen, Garcia, & Petajan, 1994:572-577; Martino, Magalhaes, & Violaro, 2006:971-980) berdasarkan *range* bentuk mulut yang berbeda, stimuli, dan bentuk pengenalan secara visual. Akan tetapi, tidak ada bentuk pemetaan yang jelas dari bentuk fonem dalam bentuk *viseme*. Hal ini dikarenakan tidak adanya pemetaan sederhana dari banyak ke satu model fonem secara visual. Viseme statis tidak memperhitungkan koartikulasi visual yang merupakan pengaruh tinggi rendahnya pengucapan suara dari fonem. Koartikulasi menyebabkan bibir berpose untuk suara yang sama untuk menunjukkan bentuk visual yang berbeda tergantung dalam pembuatan *viseme* dan pada waktu artikulasi beberapa suara mungkin tidak terlihat sama sekali.

Gambar 2. Beberapa *frame* selama artikulasi dari /t/ yang menampilkan variasi dari pose *articulator* yang menyebabkan koartikulasi.

Model yang lebih realistis dari model percakapan visual adalam *viseme* dinamis (Taylor, Mahler, Theobald, & Matthews, 2012:275-284). *Viseme* dinamis merupakan model pergerakan dari percakapan yang bukan pose statis dan dikelompokkan berdasarkan bentuk visual percakapan secara independen dari fonem dasar. Beberapa data *training video* yang berisi percakapan dengan artikulator yang terlihat dan terlacak kemudian dimasukkan dalam parameter di ruang dimensi rendah. Parameterisasi ini kemudian secara otomatis tersegmentasi dengan melakukan identifikasi poin penting untuk membedakannya suara tidak tumpang tindih. Poin penting ini secara visual intuitif diberikan berdasarkan lokasi tempat artikulator berubah, misalnya sebagai bibir dekat selama bilabial, atau puncak pembukaan bibir selama vokal. Pola pergerakan percakapan inilah yang dikemudian dilakukan segmentasi untuk membentuk kelompok *viseme* dinamis.

Phoneme to *Viseme Mapping*

Model fonem Preston Blair merupakan satu set model dari *viseme* yang umum digunakan sebagai animasi wajah dalam film kartun. Model fonem Preston Blair terdiri dari 10 model *viseme* yang digunakan untuk memetakan semua kemungkinan fonem yang ada.

Berikut ini adalah model *viseme* yang digunakan berdasarkan model fonem dari Preston Blair dan penerapannya dalam model *English Vowel* (suara berdasarkan huruf vokal) dalam kalimat.

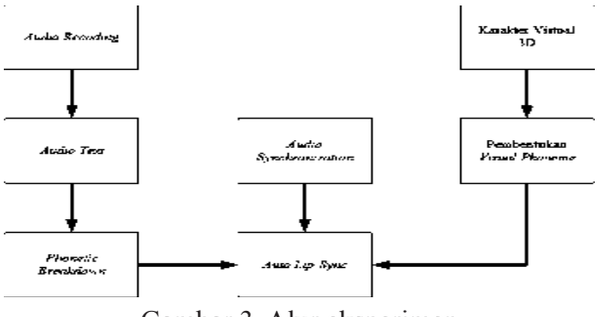
Tabel 1. Penggunaan English Vowel dan Model Fonem Preston Blair

No.	Fonem	Contoh dalam kalimat	Model Fonem Preston Blair
1.	AI	apple, day, hat, happy, rat, act, plait, dive, aisle.	
2.	E	egg, free, peach, dream, tree.	
3.	O	honk, hot, off, odd, fetlock, exotic, goat.	
4.	U	fund, universe, you runner, jump, fudge, treasure.	
5.	C D G K N R S Y Z	sit, expend, act, pig, sacked, bang, key, band, buzz, dig, sing.	
6.	F V	forest, daft, life, fear, very, endeavour.	
7.	L	election, alone, elicit, elm, leg, pull.	

8.	M B P	embark, bear, best, put, plan, imagine, mad, mine.	
9.	W Q	cower, quick, wish, skewer, how.	
10.	rest	Rest, Merupakan keadaan berhenti antara kalimat	

METODE EKSPERIMEN

Secara umum penelitian ini dilakukan dengan metode eksplorasi eksperimentatif dengan tahapan seperti dalam Gambar 3 dan karakter virtual 3D yang digunakan seperti dalam Gambar 4.



Gambar 3. Alur eksperimen

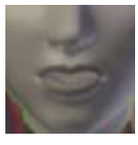
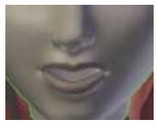
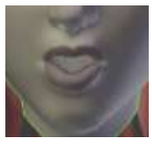
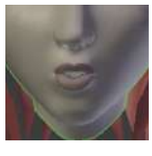
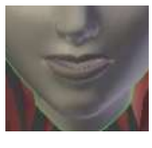
Gambar 4. Karakter virtual 3D yang digunakan

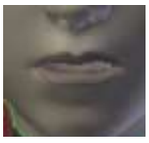
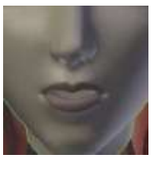
PEMBAHASAN

Viseme dalam Karakter Virtual

Pembentukan *viseme* dalam karakter virtual berdasarkan bentuk visual dari mulut karakter yang mengacu pada model fonem dari Preston Blair. Bentuk *viseme* dapat dibentuk dengan mengombinasikan perubahan *blendshape* berdasarkan beberapa masukan dengan parameter yang telah ditentukan.

Tabel 2. Bentuk *viseme* dalam karakter virtual

No.	Kelas *Viseme*	Bentuk mulut dari Preston Blair Phoneme Series	Bentuk mulut dari Karakter Virtual	Fonem
1.	*Viseme* 1			A dan I
2.	*Viseme* 2			E
3.	*Viseme* 3			O
4.	*Viseme* 4			U
5.	*Viseme* 5			C, D, G, K, N, R, S, Th, Y dan Z (dimasuk-kan dalam model etc)

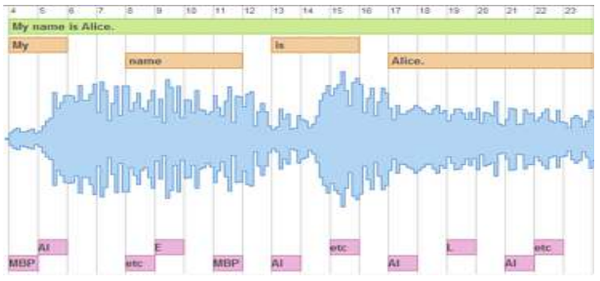
6.	*Viseme* 6	 	F dan V
7.	*Viseme* 7	 	L
8.	*Viseme* 8	 	M, B dan P
9.	*Viseme* 9	 	W dan Q
10.	*Viseme* 10	 	Rest

Auto Lip-Sync

Proses ujicoba *viseme* dalam virtual karakter virtual 3D dilakukan dengan bantuan *software Papagayo* untuk melakukan pemecahan (*breakdown*) fonem dalam bahasa Inggris. *Papagayo* melakukan pemecahan fonem berdasarkan model dari Preston Blair Phoneme series. Sampel dari uji coba *viseme*, karakter akan mencoba mengucapkan kata:

My Name is Alice

Proses pemecahan fomen dari kalimat "My Name is Alice" dihasilkan bentuk fonetik sebagai berikut:



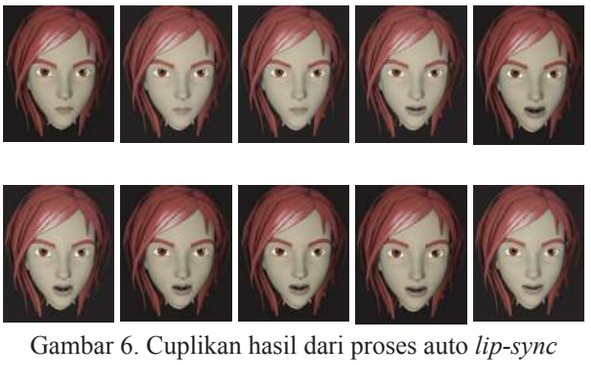
Gambar 5. Pemecahan fomen

Pemecahan fonem mengacu pada bentuk Preston Blair Fonem Series, kalimat yang digunakan sebagai sampel akan dibagi ke dalam beberapa bentuk model *viseme*. Model *viseme* inilah yang kemudian digunakan sebagai acuan pengucapan kalimat dalam karakter virtual 3D. Hasil dari pemecahan berdasarkan fonem menghasilkan data sebagai berikut.

Tabel 3. Hasil pemecahan fonem

No.	*Viseme*	Jumlah
1	MBP	2
2	AI	4
3	Rest	9
4	Etc	3
5	E	1
6	L	1

Dengan bantuan *software* 3D *blender*, hasil dari proses pemecahan fonem kemudian disinkronisasikan dengan *audio* untuk menghasilkan auto *lip-sync* dalam karakter. Berikut beberapa hasil animasi yang didapatkan dari proses auto *lip-sync*.



Gambar 6. Cuplikan hasil dari proses auto *lip-sync* dalam karakter virtual

SIMPULAN

Pembentukan auto *lip-sync* dalam karakter virtual berbasis *blendshape* sangat bergantung pada pembentukan *viseme* dan pemberian bobot *blendshape* dari karakter virtual. Pembentukan *viseme* dijadikan acuan penting untuk menjamin ketepatan bentuk dari *lip-sync* yang dibangun. Semakin baik dan detail bentuk *viseme* yang dibentuk, maka hasil yang didapatkan semakin baik.

Proess pembentukan auto *lip-sync* masih mengacu dalam pembentukan *viseme* dalam bahasa Inggris. Penelitian lanjutan dilakukan untuk melakukan pembentukan *viseme* dalam bahasa lain, misalnya bahasa Indonesia. Hal ini dilakukan supaya karakter dapat mengucapkan berbagai macam bahasa secara natural.

KEPUSTAKAAN

Arai, K., T. Kurihara, & K. Anjyo. 1996. "Bilinear interpolation for facial expression and metamorphosis in real-time animation". *The Visual Computer,* Vol. 12, 105–116.

Bergeron, P., & P. Lachapelle. 1985. *Controlling facial expression and body movements in the computer generated short "tony de peltrie*".

Cassell, J., S. Prevost, J. Sullivan, & E. Churchill. 2000. *Embodied Conversational Agents*. Cambridge, MA: MIT Press.

Choe, B., & H. Ko. 2001. "Analysis and synthesis of facial expressions with hand-generated muscle actuation basis". *IEEE Computer Animation Conference*, (pp. 12-19).

Deng, Z., P. Chiang, P. Fox, & U. Neumann. 2006. "Animating blendshape faces by cross-mapping motion capture data". Proceedings of the 2006 symposium on Interactive 3D graphics and games, (pp. 43-48).

Fisher, C. 1968. "Confusions among visually perceived consonants". *Journal of Speech and Hearing Research (JSHR)*, 796–800.

Goldschen, A. J., O.N. Garcia, & E. Petajan. 1994. "Continuous optical automatic speech recognition by lipreading". In Proceedings of the 28th Asilomar Conference on Signals, Systems, and Computers, (pp. 572–577).

Lewis, J., M. Cordner, & N. Fong. 2000. "Pose space deformation". Proceedings of the 27th annual conference on Computer graphics and interactive techniques (pp. 165-172). SIGGRAPH.

Lewis, J., J. Mooser, Z. Deng, & U. Neumann. 2005. "Reducing blendshape interference by selected motion attenuation". Proceedings of ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games (I3DG), (pp. 25-29).

Martino, J. M., L.P. Magalhaes, & F. Violaro. 2006. "Facial animation based on context-dependent visemes". *Journal of Computers and Graphics*, Vol. 30, No. 6, 971 – 980.

Parke, F. 1974. "A Parametric Model for Human Faces". Utah: Ph.D. Thesis, University of Utah.

Pighin, F., J. Hecker, D. Lischinski, R. Szeliski, & D. Salesin. 1998. "Synthesizing realistic facial expressions from photographs". SIGGRAPH Proceedings, (pp. 75-84).

Sera, H., S. Morishma, & D. Terzopoulos. 1996. "Physics-based muscle model for mouth shape control". IEEE International Workshop on Robot and Human Communication, 207-212.

Serenko, A., N. Bontis, & B. Detlor. 2007. "End-user adoption of animated interface agents in everyday work application". *Behaviour and Information Technology*, 119-123.

Sifakis, E., I. Neverov, & R. Fedkiw. 2005. "Automatic determination of facial muscle activations from sparse motion capture marker data". *ACM Trans. Graph* 24(3), (pp. 417–425).

Taylor, S., M. Mahler, B. Theobald, & I. Matthews. 2012. "Dynamic units of visual speech". In ACM/ Eurographics Symposium on Computer Animation (SCA), 275–284.

Walden, B. E., R.A. Prosek, A.A. Montgomery, C.K. Scherr, & C. J. Jones. 1977. "Effects of training on the visual recognition of consonants". *Journal of Speech, Language and Hearing Research (JSLHR)*, Vol. 20, No. 1, 130–145.

Waters, K., & T. Levergood, T. 1993. *Decface: An automatic lip-synchronization*.