Eurographics/ ACM SIGGRAPH Symposium on Computer Animation (2006)
M.-P. Cani, J. O'Brien (Editors)

# Simulating Speech with a Physics-Based Facial Muscle Model

Eftychios Sifakis† Stanford University Intel Corporation
Andrew Selle† Stanford University Industrial Light+Magic
Avram Robinson-Mosher† Stanford University
Ronald Fedkiw† Stanford University Industrial Light+Magic

**Abstract**

*We present a physically based system for creating animations of novel words and phrases from text and audio input based on the analysis of motion captured speech examples. Leading image based techniques exhibit photo-real quality, yet lack versatility especially with regard to interactions with the environment. Data driven approaches that use motion capture to deform a three dimensional surface often lack any anatomical or physically based structure, limiting their accuracy and realism. In contrast, muscle driven physics-based facial animation systems can trivially integrate external interacting objects and have the potential to produce very realistic animations as long as the underlying model and simulation framework are faithful to the anatomy of the face and the physics of facial tissue deformation. We start with a high resolution, anatomically accurate flesh and muscle model built for a specific subject. Then we translate a motion captured training set of speech examples into muscle activation signals, and subsequently segment those into intervals corresponding to individual phonemes. Finally, these samples are used to synthesize novel words and phrases. The versatility of our approach is illustrated by combining this novel speech content with various facial expressions, as well as interactions with external objects.*

Categories and Subject Descriptors (according to ACM CCS): I.3.5 [Computer Graphics]: Computational Geometry and Object Modeling – Physically based modeling; I.3.7 [Computer Graphics]: Three-Dimensional Graphics and Realism – Animation

## 1. Introduction

Photorealistic facial animation is both difficult to achieve and in high demand, as illustrated by [PLB*05], which discussed some of the challenges faced in recent blockbuster films and high profile research efforts. Many computer graphics practitioners are interested in animating conversation (see e.g. [CPB*94, CVB01]), and this has led to enormous interest in the key ingredients of speech and expression. Moreover, as stressed by [CM93], visible speech plays a large role in the interpretation of auditory speech.

Along the lines of talking presidents in "Forest Gump," [BCS97] proposed a method that used existing video footage to create a new video of a person speaking novel words. [EP00] also proposed an image-based approach that relied on the morphing of the visemes associated with phonemes. These results were further improved using a multidimensional morphable model in [EGP02] and used for retargeting in [CE05]. Although image based techniques produce animations of photo-real quality, they lack the versatility of some other approaches, e.g. it would be difficult to use them when the face has to interact with elements from the environment.

The idea of driving a three-dimensional character from text and audio (as in voice puppetry [Bra99]) is quite compelling. Data driven approaches tend to use motion capture data (see e.g. [Wil90, GGW*98]) to drive a three dimensional surface mesh. [CFP03] took this approach using independent component analysis to separate speech from expression (see also [CFKP04]). Similarly, [KMT03] used PCA of marker data to determine facial movement parameters. [CB05] used a bilinear model that separates expression from speech in order to drive a three dimensional blend shape animation with video input. In a similar vein, [DLN05] constructed a speech co-articulation model that can be mixed with keyframing in a manner that preserves expressiveness. [VBPP05] used multilinear models to separate expressions, visemes and identity in a three dimensional data set, enabling video to drive a three dimensional textured face model. While these methods have enjoyed recent popularity, especially for speech and visemes, they lack any anatomical or physically based structure, limiting their potential for accuracy and realism.

Even though [WF95] advocated the use of muscles rather than surfaces for speech animation early on, physically based simulation methods have not enjoyed popularity for phoneme or viseme research (as pointed out e.g. in [RE01]).

† email: {sifakis|aselle|avir|fedkiw}@cs.stanford.edu

**Figure 1:** ***A synthesized utterance of the word "algorithm" using a physically-based facial muscle model***

This could be due to the high computational cost associated with the level of fidelity required to study speech. Although there is precedent for estimating muscle contraction parameters from video [TW90, TW93] (see also [EBDP96, EP97, MIT98]), [BOP98b, BOP98a] avoided the internal anatomy altogether using only a surface based finite element model while studying lip motion. In fact, recent work in this area includes [CLK01, CK01] which conclude that it might be better to estimate linear combinations of sculpted basis elements rather than muscle activations. However, [SNF05] argued that the limiting factor for fidelity was not the use of simulation per se, but rather the lack of *realistic* muscles with biomechanical nonlinearity and anatomical accuracy. Furthermore, they argue that a person's face *is* driven by muscle activations, therefore an anatomically faithful model with the control granularity of actual human facial muscle exactly spans the space of facial expression.

Both image-based animation methods and data-driven surface deformation techniques have traditionally been preferred over physics-based approaches for facial animation. Both approaches operate directly on sample data without requiring an intricate anatomical facial model or the overhead of simulation for either analysis or synthesis. Yet, interacting with the simulated character in ways that are not spanned by the recorded training data is recognized as a task that lies beyond the scope of either approach. Physics-based approaches provide the unique ability to interact with the character in any way that can be physically prescribed while respecting the fundamental characteristics of a performance, namely motion style, expression and verbal content.

Following the approach of [SNF05] we build a high resolution, anatomically and biomechanically accurate flesh and muscle model of a subject's face. Then we automatically determine muscle activations based on three-dimensional sparse motion capture marker data. In particular, we focus on the capture of speech, constructing a phoneme database parameterized in muscle activation space. Notably, each phoneme is stored with temporal extent. We demonstrate that physically based approaches can be used for speech analysis *and* synthesis by creating animations of novel words and phrases from text and audio input. Moreover, we capture muscle activations representative of expression and show that these can be mixed with the speech synthesis to independently drive speech and emotion. Finally, we illustrate the versatility of a physically driven three dimensional model via interaction with foreign objects.

## 2. Previous Work

Early work on three dimensional facial animation includes [Par72, PB81, Wat87, MTPT88, KMMTT92] (see also [EF78]). [VBMH*96] relate skin deformation of a physics based model to oral tract deformation while [LMVB*97, LM99] use a physics based model driven by muscle-control signals acquired by AMG and compare surface deformation against the human subject. Based on scanned data, [LTW95] constructed an anatomically motivated, biomechanical facial model featuring a multilayer, deformable skin model with embedded muscle actuators. [KGB98] used finite elements to predict emotions on a post-surgical face (see also [KGC*96, RGTC98] for finite elements for facial surgery). [DMS98] used variational modeling and face anthropometry techniques to construct smooth face models, [PHL*98, PSS99] animated faces based on photographs and video, and [JTDP03] worked on automatic segmentation for blending. The face was also divided into subregions for the facial animation in [ZLGS03]. [BV99] proposed a vector space representation of shapes and textures for animation transfer [BBPV03] and face exchange in images [BSVS04]. [KHS01] built a muscle based facial model and considered morphing to other faces [KHYS02] and forensic analysis [KHS03]. [KP05] used a parametric muscle model with time varying visemes to extend the coarticulation algorithm of [CM93]. [BB02] added expressiveness to the MPEG-4 Facial Animation Parameters. A number of authors have worked on facial motion transfer [NN01, PKC*03, NJ04, SP04]. [CXH03] used tracking to drive animations from a motion capture database, [WHL*04] tracked facial motion with a multiresolution deformable mesh with the aim of learning expression style, and [ZSCS04] proposed a face inverse kinematics system.

## 3. Data Capture

### 3.1. Model Building

We constructed a high-resolution volumetric model of facial flesh and musculature for both our analysis of speech samples and the synthesis of new utterances. First, we obtained

**Figure 2:** ***Eight camera optical motion capture layout (left), 250 facial marker set (right).***

an MRI scan which provided an approximation of the tissue extent and the shape of the interface between soft tissue and bone. Then a life-mask cast of the subject was scanned at a resolution of 100 microns, producing a 10 million triangle model and a fully registered texture map. The detail from this high resolution surface scan was integrated into our volumetric flesh model. The facial flesh volume was discretized into a 1,870,000 element tetrahedral mesh, with 1,080,000 elements in the frontal facial volume that was used to simulate deformation under action of the facial muscles. Due to the limited resolution of the MRI scan, much of the internal tissue structure was manually adjusted to create a muscle set that conforms to the anatomical prototypes published in the medical literature. Our model includes 39 of the muscles that are predominantly involved in facial expressions and speech. Muscles that have no effect or only a subtle effect on facial motion were excluded, as their behavior would not be reliably captured with our surface motion capture marker set.

### 3.2. Motion capture

We take a data-driven approach to speech synthesis constructing a database of prototypical subject-specific utterances of speech primitives (sample phonemes within a context of words or phrases). The motion component of these utterances was recorded with a motion capture system consisting of 8 cameras with 4MP CCD sensors. 250 thin circular patches of retroreflective material with a diameter of 3mm were placed on the subject's face at an average distance of 8-10mm apart. A small subset of markers were specifically placed on predominantly rigid parts of the head to capture the rigid head motion. The performance was sampled at 120Hz. See Figure 2.

### 3.3. Inverse activations

Following [SNF05] we model the isotropic response of passive fatty tissue by a hyperelastic Mooney-Rivlin constitutive model for the deviatoric component, with an additional volumetric pressure component for quasi-incompressibility. The parameters of the Mooney-Rivlin model are spatially adapted to the heterogeneity of the simulated tissues, yielding different stiffness values for areas occupied by collagen, cartilage, and tendinous structures. Areas of the flesh that are occupied by contractile muscle tissue are further assigned

**Figure 3:** ***Estimated muscle activations of expressions from motion capture, left smile, right frown.***

an anisotropic strain response corresponding to the passive or active behavior of muscle tissue along the direction of its fiber field. The inverse activation estimation framework employs the quasistatic simulation method of [TSIF05]. This formulation uses fast conjugate gradients solvers to evolve constrained deformable objects to an equilibrium state, and provides robust handling of mesh degeneracies such as element inversion, as well as rigid body and self-collision handling. We should point out that this quasistatic assumption is preferred for the estimation process as it greatly simplifies the inverse control problem. While it can also be used for the forward simulation of slow speech, a fully dynamic simulation method is superior for the simulation of faster speech from muscle activation controls.

## 4. Phonemes and Visemes

### 4.1. A Muscle Activation Basis for Speech

The inverse activation framework described in section 3.3 allows us to translate our database of motion captured speech samples into temporal sequences of control parameters for our deformable face model (i.e. muscle activations and kinematic configuration of the bones). We subsequently use these controls as the parameterization of facial motion for analysis and synthesis tasks.

A defining property of visual speech synthesis techniques is the choice of the feature space used to describe facial motion. Common examples found in the literature (cited above) include image-based descriptions and surface shape bases. Our approach provides the versatility to edit the animated performance affecting the emotion and expression of the character, as well as allowing physical interaction of the face with objects from the environment. In this context, the relevant feature is not the appearance or the shape of the face per se, but rather the *action* of speech articulation. Therefore, we follow the formulation of [SNF05] using the activation signals that stimulate the facial muscles as our feature space, an approach that was pioneered in [TW90, TW93].

Our approach is subject to a number of limitations. The quality of our parameterization and the fidelity of the resulting simulations are only as good as the detail and accuracy of our muscle-driven model as well as the physical consistency of the simulation method used. This highlights the need for detailed, nonlinear, volumetric finite element models of the

**Figure 4:** ***Frames corresponding to estimated muscle activations of phonemes from motion capture. Top left - er, top right - r, lower left - sh, and lower right - iy.***

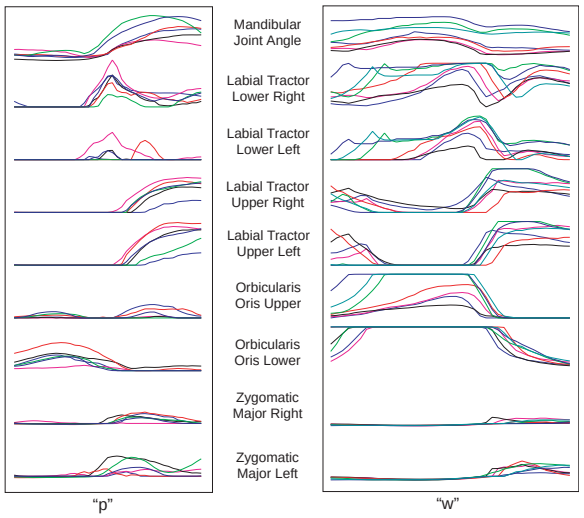


**Figure 5:** ***Comparison of the physeme samples in our database associated with the phonemes "p" and "w", illustrating the waveforms of the major muscle activations and kinematic parameters involved. Note that the samples for each phoneme are highly correlated within the muscle activation space showing the effectiveness of our basis. The "p" phoneme is dynamic and exhibits three distinct activation phases while the "w" phoneme has a uniform activation pattern.***

anatomical components of the face. Additionally, our adoption of a quasistatic simulation scheme for analysis leads to a deviation from the true dynamic behavior expected of a physical system. However, for our training set of short words spoken at a casual pace, inputting the estimated muscle activation sequences into a forward quasistatic simulation produced a very close match to the original capture respecting almost all nuances of individual utterances. This supports our use of quasistatics for the estimation of muscle activations, although a full dynamic simulation would be superior for synthesis (especially for faster speech).

### 4.2. Primitives of Speech Simulation (Physemes)

We collected a database of motion capture data for the phoneme sets suggested by [Ann06] and [EP00] where phonemes are presented within the context of sample words. We recorded 4-5 distinct captures of each phoneme set and used the inverse activations estimation process to convert each captured word into a short sequence of muscle activation signals and mandible articulation parameters. An examination of the signals corresponding to various phonemes revealed several important patterns. First, we observed a high degree of correlation between segments of words that contained the same phoneme, as illustrated in Figure 5 for samples of the phonemes **p** and **w** (we adopt the phoneme codes used in [BTC99]). The temporal extent of this correlation varied with the particular phoneme being considered and its phonetic context. Phonemes with matching context (the phonemes immediately before and after the one in question) tended to correlate over a much longer time segment. In addition, several phonemes (such as the consonants **sh**, **v**, **z**) typically reached a steady state in activation space, surrounded by the transitions from the previous and to the next phoneme. Other phonemes (such as the diphthongs **ay**, **ey**) exhibited a characteristic *dynamic* pattern over their temporal extent, often marked by a distinctive transition. In Figure 5 we classify **p** as a dynamic phoneme (we can identify 3 distinct stages of mouth closure, lip retraction and mouth opening) while the static **w** appears to achieve a steady state in between transitions. We note that muscles in the oral region typically exhibited higher degree of correlation across utterances of the same phoneme than peripheral muscles.

These observations support the hypothesis that *time-varying sequences of muscle activations* capture a large amount of information about phonemes and phoneme transitions. In particular, by recording the muscle activation signals over a time interval that extends beyond the duration of each phoneme and into its neighboring ones we capture the effect that the utterance of each phoneme has and receives from its context, formally known as *coarticulation*. Therefore, we associate each of these extended intervals of muscle activations and bone kinematics with its corresponding phoneme and use them as the primitives of our physics-based visual speech synthesis, labeling them as *physemes* (in analogy to phonemes and visemes).

To create a database of physemes we use the audio track to identify each phoneme, using the Festival Speech Synthesis System [BTC99] to segment utterances into individual phrases and then into individual phonemes. The label-

ings were not completely accurate and sometimes manual annotation was also required. Once every word had been partitioned into time segments corresponding to different phonemes, physeme samples were collected by selecting the estimated muscle activation signals corresponding to the time range of each phoneme (see Figure 6) and padding the signal on each side of the time segment with enough data from its context in order to capture the coarticulatory effects.

## 5. Synthesis

### 5.1. Physeme-based Speech Synthesis

Our physeme database captures the motion signatures of phonemes and the transitional effects between them in the physically motivated space of muscle activation. We present a first approach to using this physeme basis directly for synthesis of visual speech. The input to our systems consists of an audio recording along with a transcript of its verbal content. We again use Festival [BTC99] to segment this novel audio track into time intervals corresponding to distinct phonemes. The result is typically satisfactory for utterances of individual words or slow speech. However, sometimes with longer and faster speech passages it was necessary to manually adjust the phoneme annotation.

After we determine the constituent phonemes of the text to be synthesized, we assemble a matching temporal arrangement of physemes from our database. Each physeme contains muscle activation signals that extend beyond the duration of its associated phoneme, namely it starts with a *lead-in* from the previous phoneme, followed by the *body* corresponding to its base phoneme and a *lead-out* from the following phoneme. We place physemes in arrangements with their bodies contiguous and their lead-in and lead-out overlapping into the body of the adjacent physeme (see Figure 7). Silent intervals are modeled with a special arbitrary length "pause" physeme, with muscle activations corresponding to the neutral pose of the face. In general, the length of each phoneme in an audio recording will not match the length of the corresponding physeme in our database, so the muscle activation signal of the physeme is time-scaled to the appropriate length. A single, uniform time scaling is applied to the body as well as the lead-in/out of the physeme.

For each physeme inserted in an arrangement we use a blending curve that yields constant weights equal to unity throughout the body of the physeme and decays to zero at the outer endpoints of the lead-in and lead-out following a $C^1$ continuous sigmoid curve. We extract a single muscle activation signal from the complete physeme arrangement by performing weighted averaging of the signals overlapping at any instance in time using their corresponding blending weights, as illustrated in Figure 7.

Among the kinematic parameters that are obtained through the inverse activation estimation process, the parameters that define the overall rigid body motion of the head (or

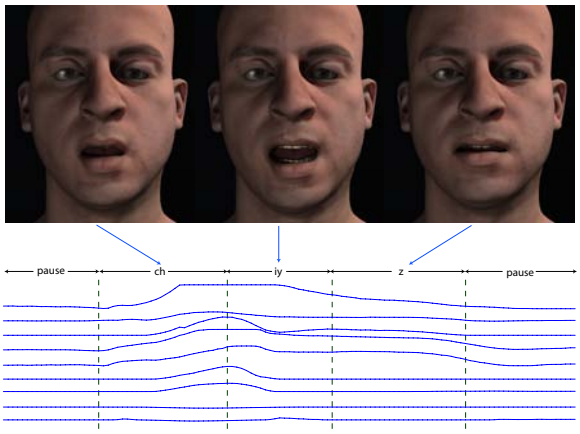


**Figure 6:** ***Segmentation of a time sequence of nine selected muscle activations for the captured word "cheese" into constituent physemes. For each physeme we store the time-varying signals of the phoneme segment and its context in the database. The rendered frames correspond to point samples of the physemes within each of the phonemes.***

the frame of reference of the cranium) receive special handling. Interpolating between different positions and orientations in such short time intervals as those corresponding to phoneme lengths would most likely incur sudden jumps and violent accelerations. Thus, instead of interpolating between frames of reference, we use the estimated rigid body motion for each individual physeme and use it to approximate the linear and angular velocity of the head at each frame. When blending phonemes we then proceed to blend linear and angular velocities instead of positions and orientations. The rigid body configuration is then obtained by integrating the linear and angular velocities forward in time. The resulting signals of muscle activations and rigid bone kinematics can be fed into a forward quasistatic or dynamic simulator to produce the final physically driven speech simulation.

### 5.2. Sequence Generation

We employ a semi-automatic interface for the creation of physeme-based simulations of speech including tools for the creation and refinement of physeme arrangements, preview of an approximate speech synthesis and final physics-based finite element simulation. Given the existence of tools such as Festival that simplify the segmentation of an audio speech signal into its constituent phonemes, we focus on the task of compiling a physeme arrangement to match a given, labeled phoneme sequence. The low dimensionality of our feature space (39 muscle activations and 3 kinematic parameters, a few tens of phonemes per sentence) makes optimization algorithms such as stochastic optimization attractive, i.e. since we avoid the overhead typically associated with them in higher dimensional spaces.

We adopted the constraints that the labels of the physemes have to match the labels of the phonemes that occupy the

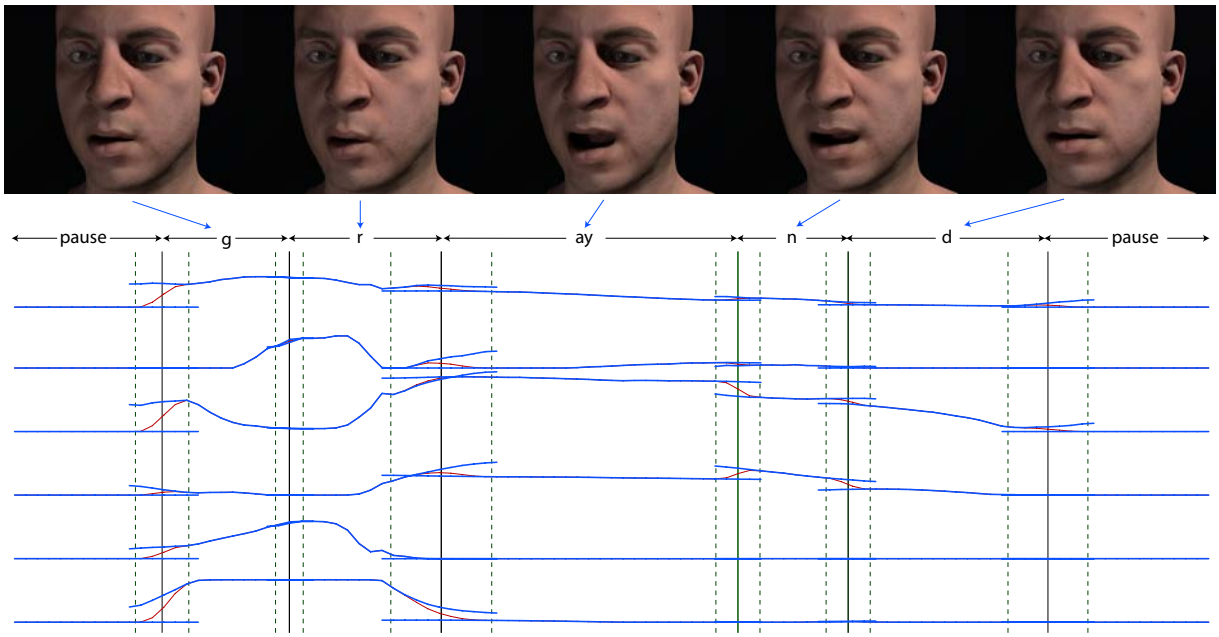


**Figure 7:** ***Arrangement of the word "grind" synthesized from the physemes in our database. The solid vertical lines represent physeme boundaries determined from the input audio signal while the dashed lines represent the blending region between physemes. The thick blue curves are the individual physeme samples used to build the arrangement while the thin red curve represents the blended signal.***

same time range in the audio recording, and we clamped the lead-in and lead-out of each physeme to 20% of the length of the body of the neighboring phoneme it overlaps with. Under these constraints, our free parameters are the choice of which of the 5-30 physemes from our database for each phoneme should be used to fill a particular time interval. We formulated a criterion for the quality of a particular phoneme arrangement and solved it using simulated annealing, using parameters that yield the global minimum with very high confidence. We obtain our quality criterion by computing the magnitude of the discrepancy between the muscle activation vectors for all frames where physeme extents would overlap and integrating over time. In order to prevent weak muscles or muscles that have little direct effect in the articulation of speech to dictate the quality of an arrangement, we scale the activation $a_i$ value by the average magnitude of the quasistatic shape Jacobian $\partial\mathbf{X}/\partial a_i$, computed over the entire range of motion captured visemes (as a by-product of the inverse activation estimation process). This biases the quality criterion towards accounting for muscles whose activation tends to have a more substantial effect on the shape of the face.

Our optimization process provided convincing physeme arrangements for simple examples (such as single, slowly spoken words) requiring little to no manual intervention. However, with more complicated examples or faster speech this result often required manual adjustment, such as fine-tuning the length of the lead-in/out segments or the precise placement of phoneme boundaries. We created a graphical interface that provided us with the functionality to alter all these parameters, as well as the individual choice of physemes used at each moment in time. As a by-product of our inverse estimation process, we possess a quasistatic face shape approximation for each frame in our captured training set. By blending these face shapes with the same weights used for physeme blending we obtain a fast preview for our edits without the need for simulation, which is run only when our adjustments are complete. On average, editing a medium-sized sentence would typically entail 2-3 hours of manual processing. We note that quasistatic simulation contributed substantially to that cost, since the lack of damping and inertia made the final result more sensitive to the muscle activation input signal than it would be with a full dynamic simulation.

## 6. Speech and Expression

The versatility of a physically-based muscle driven face model for speech synthesis is highlighted by the ability to augment the simulation with elements that are secondary to the process of speech articulation. Facial expression and emotion are characteristic examples of such elements. Although there exists a correlation between the emotional appearance and the verbal content of human speech, a human speaker may adjust his facial expression independent of the words spoken. We simulate this process by motion capturing facial expressions and using our inverse activation process to convert these expressions into characteristic muscle contractions. Subsequently, we blend these muscle activa-

**Figure 8:** ***The middle row shows a synthesized speech sequence of the word "seeping" with neutral expression. The other rows show the same activation sequence blended with a smile (top row) and frown (bottom row).***

tion values into our synthesized physeme sequences and use physics-based simulation to obtain the final animation.

As illustrated in Figure 8, the integration of expressions such as a frown or a smile can be performed in a very natural manner, through simulation, without compromising either the articulatory content or the emotional response elicited by the expression that was blended in. Such an augmentation is straightforward and requires no manual adjustment of the physeme arrangement. It should be emphasized that it is much more challenging and labor intensive to obtain such a result, in regard to *both* speech and expression, with a technique based on blending images or face shapes. The nonlocal effects of pronounced facial expressions, in conjunction with anatomical phenomena that arise from these expressions (such as bulging of skin, deepening of facial furrows, or changes in the contact pattern of the lips) become particularly difficult to capture convincingly without a physically-based approach.

## 7. Speech and Physics

Beyond the task of enriching the facial motion with an expression, the real power of physics-based approaches is revealed when the face is required to interact with the outside world. Physical simulation of a full volumetric facial musculature model allows us to produce effects that are difficult or impossible using image-based or data-driven surface deformation techniques. By keeping the muscle activation controls fixed and modifying the simulation environment, we can effortlessly produce a new facial deformation. In particular, once a speech database is created, reproducing such effects does not require additional motion capture data, analysis or modification of our synthesized physeme sequences. For example, we depict our virtual character speaking with a lollipop and candycane in his mouth in Figure 9, where the muscle activation signals were synthesized without regard to the object interaction.

## 8. Discussion

Using our quasistatic estimation framework, we processed approximately 10 minutes of speech at 120Hz in an average of 7 minutes per motion capture frame (including full analysis with full rigid body and self-collision handling) on a Xeon 3.8 Ghz CPU. The full processing of approximately 70,000 frames required the equivalent of a CPU year on clustered computer hardware. Although seemingly high in computational cost, this once-only process requires no human supervision and *all* the resulting muscle activation signals were of adequate quality for use in our database. Notably, our simulation model contained 1080K simulation elements, which is at least 3 times larger than typical high resolution finite element simulations in the computer graphics literature. Therefore, we expect our model to age much slower than the advance of computer hardware, making the computation affordable.

Once the facial model has been created and the physeme database has been assembled from the motion captured performance, the main labor-intensive effort is the manual adjustment of the physeme sequences to fine-tune the synthesized speech result. Currently at a cost of a few hours per sentence, the bulk of this effort is attributed to correcting mistakes of the speech analysis software (the Festival system) and adjustment of the transition intervals between successive phonemes. The latter would be substantially easier if a full dynamic framework was employed for the final

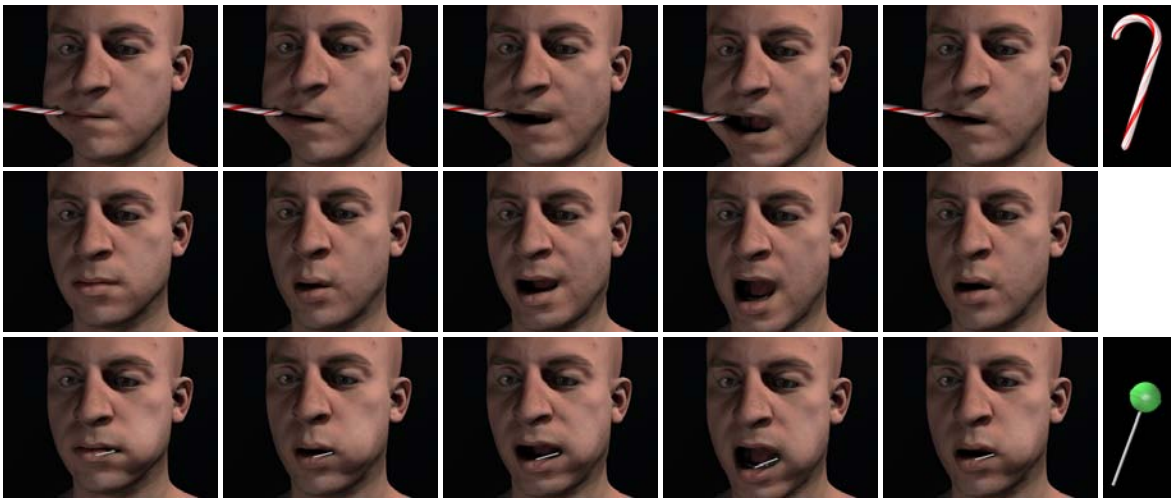

**Figure 9:** ***The middle row shows a selection from "we can also model" synthesized with our method. In the top and the bottom rows, we augment the simulation with interacting physical objects while using the same muscle activation controls.***

forward simulation, as the physics of facial motion would handle physeme transitions in a smooth way alleviating issues with blending intervals. While this editing cost might seem high compared to data-driven methods, especially if one is only concerned with casual speech, it becomes much more competitive when one requires the added versatility of the face interacting with its environment. Our method can achieve physically based environmental interactions with almost no additional cost, whereas the cost increases substantially for data-driven approaches even though the quality incurs a significant decrease.

## 9. Conclusions and Future Work

We presented a physics-based approach to visual speech synthesis using an anatomically accurate, muscle actuated finite element model of the human face. We collected motion capture data for the utterances of words in a training set and converted them into the time varying muscle activation signals that give rise to the captured face motion. We segmented these muscle activation signals into time segments corresponding to different phonemes and assembled them into a database of *physemes*. We create sequences of physemes with smooth transitions between them to match the phonemes of audio recordings of new speech, and use the resulting muscle activation signals to drive a finite element simulation of facial motion. The animation is readily enriched by blending in expressive emotions or by introducing external objects that interact with the talking face.

Our adoption of a quasistatic simulation scheme was motivated by the tractability of the muscle activation estimation framework of [SNF05]. When our estimated activations were used in a quasistatic forward simulation, the results compared well to the original video visually validating our approach to estimation. However, for the reanimation of a novel synthesized physeme sequence, quasistatic simulation was only satisfactory for sequences of slow speech, where ballistic motion and pronounced inertial effects are not significant. When used with muscle activation signals created for faster speech, quasistatic simulation gave rise to rather abrupt and underdamped motion that lacked realism. Inertial effects have a profound effect on the motion of a real human face under such conditions, smoothing out phoneme transitions and smearing away the dynamics of individual utterances. We believe that a full dynamic simulation is the obvious choice for animations of such synthesized sequences of fast speech and discourage the use of quasistatics for forward simulation whenever possible. The practice of using quasistatics for the inverse problem and full dynamics for forward simulation is also a well established practice in the field of biomechanics, where it is well understood that even in cases where the estimated actuations are not smooth or even discontinuous, the simulated deformation using a full dynamic scheme is smooth and realistic due to the progressive fashion in which muscle activations translate to tissue deformation.

Our key objective for our future work is the illustration that a fully dynamic simulation of the synthesized muscle activation signals successfully treats demanding, fast-paced speech passages. Increased realism could be obtained by improving eyelid, eyebrow and forehead motion as well as modeling the effect of airflow in the oral cavity accommodating effects such as cheek puffing and improving the appearance of closed-mouth phonemes (such as *p* and *b*). Improved training sets will enable us to better capture the dynamics of speech in different contexts than that of short, slowly spoken phrases. The current work constitutes only a first step in using the muscle activation basis for speech analysis, and this compact, complete and physically motivated description provides the potential to improve several analysis techniques such as Principal Component Analysis

by allowing them to operate in a space that is much more tightly bound to the physical process of speech articulation rather than in a space of facial appearances or skin shapes.

**Acknowledgements**

Research supported in part by an ONR YIP award and a PECASE award (ONR N00014-01-1-0620), a Packard Foundation Fellowship, a Sloan Research Fellowship, ONR N00014-05-1-0479, ARO DAAD19-03-1-0331, NSF IIS-0326388, NSF ITR-0205671, NSF ITR-0121288, NSF ACI-0323866 and NIH U54-GM072970. E.S. was supported in part by a Stanford Graduate Fellowship. We would like to thank Jerry Talton for his contributions to our face renderer, Sergey Koltakov for editing and analyzing MoCap and audio data, XYZrgb Inc. for their invaluable assistance in creating our face model, Motion Analysis Corp. for their MoCap hardware and custom support, Garry Gold and Gary Glover for MRI imaging, Court Cutting and Aaron Oliker for providing anatomical expertise, Bill Dally and Christos Kozyrakis for computing resources, Dinesh Pai and Christophe Hery for feedback, and Hector Garcia-Molina for his expert photography.

## References

[Ann06] ANNOSOFT, LLC: Basic phoneset, 2006. Available at http://www.annosoft.com/phoneset.htm.

[BB02] BYUN M., BADLER N. I.: FacEMOTE: Qualitative parametric modifiers for facial animations. In *Proc. of ACM SIGGRAPH/Eurographics Symp. on Comput. Anim.* (2002), ACM Press, pp. 65–71.

[BBPV03] BLANZ V., BASSO C., POGGIO T., VETTER T.: Reanimating faces in images and video. In *Proc. of Eurographics* (2003), vol. 22.

[BCS97] BREGLER C., COVELL M., SLANEY M.: Video Rewrite: driving visual speech with audio. In *Proc. of ACM SIGGRAPH* (1997), pp. 353–360.

[BOP98a] BASU S., OLIVER N., PENTLAND A.: 3D lip shapes from video: a combined physical-statistical model. *Speech Communication 26* (1998), 131–148.

[BOP98b] BASU S., OLIVER N., PENTLAND A.: 3D modeling and tracking of human lip motions. IEEE Comput. Society, pp. 337–343.

[Bra99] BRAND M.: Voice puppetry. In *Proc. of ACM SIGGRAPH* (1999), pp. 21–28.

[BSVS04] BLANZ V., SCHERBAUM K., VETTER T., SEIDEL H. P.: Exchanging faces in images. In *Proc. of Eurographics* (2004), vol. 23.

[BTC99] BLACK A., TAYLOR P., CALEY R.: The festival speech synthesis system.

[BV99] BLANZ V., VETTER T.: A morphable model for the synthesis of 3D faces. In *Proc. of ACM SIGGRAPH* (1999), ACM Press, pp. 187–194.

[CB05] CHUANG E., BREGLER C.: Mood swings: expressive speech animation. *ACM Trans. Graph. 24*, 2 (2005), 331–347.

[CE05] CHANG Y., EZZAT T.: Transferable videorealistic speech animation. *Eurographics/ACM SIGGRAPH Symp. on Comput. Anim.* (2005).

[CFKP04] CAO Y., FALOUTSOS P., KOHLER E., PIGHIN F.: Real-time speech motion synthesis from recorded motions. In *Proc. of 2003 ACM SIGGRAPH/Eurographics Symp. on Comput. Anim.* (2004), pp. 347–355.

[CFP03] CAO Y., FALOUTSOS P., PIGHIN F.: Unsupervised learning for speech motion editing. In *Proc. of the ACM SIGGRAPH/Eurographics Symp. on Comput. Anim.* (2003), pp. 225–231.

[CK01] CHOE B., KO H.-S.: Analysis and synthesis of facial expressions with hand-generated muscle actuation basis. In *Proc. of Comput. Anim.* (2001), pp. 12–19.

[CLK01] CHOE B., LEE H., KO H.-S.: Performance-driven muscle-based facial animation. *J. Vis. and Comput. Anim. 12* (2001), 67–79.

[CM93] COHEN M., MASSARO D.: Modeling coarticulation in synthetic visual speech. *Models and Techniques in Comput. Anim.* (1993).

[CPB*94] CASSELL J., PELACHAUD C., BADLER N., STEEDMAN M., ACHORN B., BECKET T., DOUBILLE B., PREVOST S., STONE M.: Animated conversation: Rule-based generation of facial expression, gesture and spoken intonation for multiple conversational agents. In *Proc. of ACM SIGGRAPH* (1994), ACM Press, pp. 413–420.

[CVB01] CASSELL J., VILHJÁLMSSON H. H., BICKMORE T.: BEAT: the Behavior Expression Animation Toolkit. In *Proc. of ACM SIGGRAPH* (2001), pp. 477–486.

[CXH03] CHAI J., XIAO J., HODGINS J.: Vision-based control of 3D facial animation. In *Proc. of ACM SIGGRAPH/Eurographics Symp. on Comput. Anim.* (2003), pp. 193–206.

[DLN05] DENG Z., LEWIS J., NEUMANN U.: Synthesizing speech animation by learning compact speech co-articulation models. *Comput. Graph. Int.* (2005), 19–25.

[DMS98] DECARLO D., METAXAS D., STONE M.: An anthropometric face model using variational techniques. In *Proc. of ACM SIGGRAPH* (1998), ACM Press, pp. 67–74.

[EBDP96] ESSA I., BASU S., DARRELL T., PENTLAND A.: Modeling, tracking and interactive animation of faces and heads using input from video. In *Proc. of Comput. Anim.* (1996), IEEE Comput. Society, pp. 68–79.

[EF78] EKMAN P., FRIESEN W. V.: *Facial Action Coding System*. Consulting Psychologist Press, Palo Alto, 1978.

[EGP02] EZZAT T., GEIGER G., POGGIO T.: Trainable videorealistic speech animation. In *ACM Trans. Graph.* (2002), vol. 21, ACM Press, pp. 388–398.

[EP97] ESSA I., PENTLAND A.: Coding, analysis, interpretation, and recognition of facial expressions. *IEEE Trans. on Pattern Analysis and Machine Intelligence 19*, 7 (1997), 757–763.

[EP00] EZZAT T., POGGIO T.: Visual speech synthesis by morphing visemes. In *Int. J. Comp. Vision* (2000), vol. 38, pp. 45–37.

[GGW*98] GUENTER B., GRIMM C., WOOD D., MALVAR H., PIGHIN F.: Making faces. In *Proc. ACM SIGGRAPH* (1998), ACM Press, pp. 55–66.

[JTDP03] JOSHI P., TIEN W. C., DESBRUN M., PIGHIN F.: Learning controls for blend shape based realistic facial animation. In *Proc. ACM SIGGRAPH/Eurographics Symp. on Comput. Anim.* (2003), pp. 365–373.

[KGB98] KOCH R., GROSS M., BOSSHARD A.: Emotion editing using finite elements. *Proc. of Eurographics 1998 17*, 3 (1998).

[KGC*96] KOCH R. M., GROSS M. H., CARLS F. R., VON BUREN D. F., FANKHAUSER G., PARISH Y. I. H.: Simulating facial surgery using finite element models. *Comput. Graph. 30*, Annual Conf. Series (1996), 421–428.

[KHS01] KAHLER K., HABER J., SEIDEL H.-P.: Geometry-based muscle modeling for facial animation. In *Proc. of Graph. Interface* (2001), pp. 37–46.

[KHS03] KAHLER K., HABER J., SEIDEL H.-P.: Reanimating the dead: Reconstruction of expressive faces from skull data. In *ACM Trans. Graph.* (2003), vol. 22, pp. 554–561.

[KHYS02] KAHLER K., HABER J., YAMAUCHI H., SEIDEL H.-P.: Head shop: Generating animated head models with anatomical structure. In *Proc. of ACM SIGGRAPH/Eurographics Symp. on Comput. Anim.* (2002), pp. 55–63.

[KMMTT92] KALRA P., MANGILI A., MAGNETAT-THALMANN N., THALMANN D.: Simulation of facial muscle actions based on rational free form deformations. In *Proc. of Eurographics* (1992), pp. 59–69.

[KMT03] KSHIRSAGAR S., MAGNENAT-THALMANN N.: Visyllable based speech animation. In *Proc. of Eurographics* (2003), vol. 22.

[KP05] KING S., PARENT R.: Creating speech-synchronized animation. *IEEE Transactions on Visualization and Computer Graphics 11*, 3 (2005), 341–352.

[LM99] LUCERO J., MUNHALL K.: A model of facial biomechanics for speech production. *Journal of the Accoustical Society of America 106*, 5 (1999), 2834–2842.

[LMVB*97] LUCERO J., MUNHALL K., VATIKIOTIS-BATESON E., GRACCO V., TERZOPOULOS D.: Muscle-based modeling of facial dynamics during speech production. *Journal of the Acoustical Society of America 101*, 5 (May 1997), 3175–3176.

[LTW95] LEE Y., TERZOPOULOS D., WATERS K.: Realistic modeling for facial animation. *Comput. Graph. (SIGGRAPH Proc.)* (1995), 55–62.

[MIT98] MORISHIMA S., ISHIKAWA T., TERZOPOULOS D.: Facial muscle parameter decision from 2D frontal image. In *Proc. of the Int. Conf. on Pattern Recognition* (1998), vol. 1, pp. 160–162.

[MTPT88] MAGNENAT-THALMANN N., PRIMEAU E., THALMANN D.: Abstract muscle action procedures for human face animation. *The Vis. Comput. 3*, 5 (1988), 290–297.

[NJ04] NA K., JUNG M.: Hierarchical retargetting of fine facial motions. In *Proc. of Eurographics* (2004), vol. 23.

[NN01] NOH J., NEUMANN U.: Expression cloning. In *Proc. of ACM SIGGRAPH* (2001), Fiume E., (Ed.), ACM Press, pp. 277–288.

[Par72] PARKE F. I.: Computer generated animation of faces. In *Proc. of ACM Conf.* (1972), ACM Press, pp. 451–457.

[PB81] PLATT S. M., BADLER N. I.: Animating facial expressions. *Comput. Graph. (SIGGRAPH Proc.)* (1981), 245–252.

[PHL*98] PIGHIN F., HECKER J., LISCHINSKI D., SZELISKI R., SALESIN D. H.: Synthesizing realistic facial expressions from photographs. In *Proc. of ACM SIGGRAPH* (1998), ACM Press, pp. 75–84.

[PKC*03] PYUN H., KIM Y., CHAE W., KANG H. W., SHIN S. Y.: An example-based approach for facial expression cloning. In *Proc. of ACM SIGGRAPH/Eurographics Symp. on Comput. Anim.* (2003), pp. 167–176.

[PLB*05] PIGHIN F., LEWIS J., BORSHUKOV G., BENNETT D., DEBEVEC P., HERY C., SULLIVAN S., WILLIAMS L., ZHANG L.: Digital face cloning. In *SIGGRAPH Course Notes* (2005), ACM.

[PSS99] PIGHIN F., SZELISKI R., SALESIN D.: Resynthesizing facial animation through 3D model-based tracking. In *Proc. of Int. Conf. on Comput. Vision* (1999), pp. 143–150.

[RE01] REVERET L., ESSA I.: Visual coding and tracking of speech related facial motion. In *Proc. of IEEE CVPR Int. Wrkshp. on Cues in Communication* (2001).

[RGTC98] ROTH S. H., GROSS M., TURELLO M. H., CARLS S.: A Bernstein-Bézier based approach to soft tissue simulation. In *Proc. of Eurographics* (1998), vol. 17, pp. 285–294.

[SNF05] SIFAKIS E., NEVEROV I., FEDKIW R.: Automatic determination of facial muscle activations from sparse motion capture marker data. *ACM Trans. Graph. (SIGGRAPH Proc.)* (2005).

[SP04] SUMNER R., POPOVIĆ J.: Deformation transfer for triangle meshes. In *ACM Transactions on Graphics (Proc. of ACM SIGGRAPH)* (2004), vol. 23, pp. 399 – 405.

[TSIF05] TERAN J., SIFAKIS E., IRVING G., FEDKIW R.: Robust quasistatic finite elements and flesh simulation. *Proc. of the 2005 ACM SIGGRAPH/Eurographics Symp. on Comput. Anim.* (2005).

[TW90] TERZOPOULOS D., WATERS K.: Physically-based facial modeling, analysis, and animation. *J. Vis. and Comput. Anim. 1* (1990), 73–80.

[TW93] TERZOPOULOS D., WATERS K.: Analysis and synthesis of facial image sequences using physical and anatomical models. *IEEE Trans. on Pattern Analysis and Machine Intelligence 15*, 6 (1993).

[VBMH*96] VATIKIOTIS-BATESON E., MUNHALL K., HIRAYAMA M., LEE Y., TERZOPOULOS D.: Dynamics of facial motion in speech: Kinematic and electromyographic studies of orofacial structures. In *Speechreading by Humans and Machines*, vol. 150 of *NATO ASI Series on Computer and System Sciences*. Springer-Verlag, March 1996, ch. 16, pp. 231–232.

[VBPP05] VLASIC D., BRAND M., PFISTER H., POPOVIĆ J.: Face transfer with multilinear models. In *ACM Transactions on Graphics (Proc. of ACM SIGGRAPH)* (2005), vol. 24, pp. 426–433.

[Wat87] WATERS K.: A muscle model for animating three-dimensional facial expressions. *Comput. Graph. (SIGGRAPH Proc.)* (1987), 17–24.

[WF95] WATERS K., FRISBIE J.: A coordinated muscle model for speech animation. In *Proc. of Graph. Interface* (May 1995), pp. 163–170.

[WHL*04] WANG Y., HUANG X., LEE C. S., ZHANG S., LI Z., SAMARAS D., METAXAS D., ELGAMMAL A., HUANG P.: High resolution acquisition, learning and transfer of dynamic 3-D facial expressions. In *Proc. of Eurographics* (2004), pp. 677–686.

[Wil90] WILLIAMS L.: Performance-driven facial animation. In *Comput. Graph. (Proc. of Int. Conf. on Comput. Graph. and Int. Techniques)* (1990), ACM Press, pp. 235–242.

[ZLGS03] ZHANG Q., LIU Z., GUO B., SHUM H.: Geometry-driven photorealistic facial expression synthesis. In *Proc. of ACM SIGGRAPH/Eurographics Symp. on Comput. Anim.* (2003), ACM Press, pp. 16–22.

[ZSCS04] ZHANG L., SNAVELY N., CURLESS B., SEITZ S.: Spacetime faces: High resolution capture for modeling and animation. In *ACM Transactions on Graphics (Proc. of ACM SIGGRAPH)* (2004), vol. 23, ACM Press, pp. 548–558.