JALI: An Animator-Centric Viseme Model for Expressive Lip Synchronization

Pif Edwards, Chris Landreth, Eugene Fiume, Karan Singh*
Dynamic Graphics Project
Department of Computer Science
University of Toronto

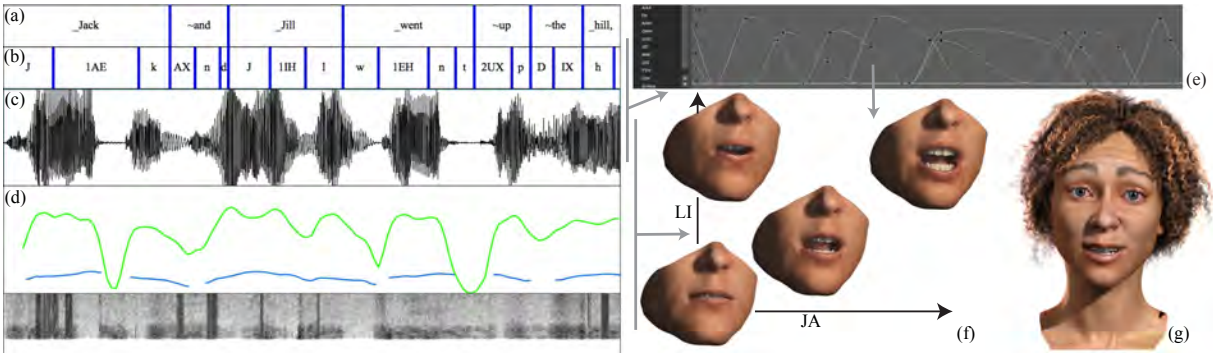


Figure 1: *JALI workflow: given a speech transcript (a) and audio (c), we compute a phonetic alignment (b). (a) and (b) are used to procedurally animate the phonemes of a JALI rig (e). Audio signal features like volume (green), pitch (blue) and formants (d) are used to animate the JALI viseme field (f). The JALI values in (f) modulate the animation curves in (e) to animate a 3D character (g).*

Abstract

The rich signals we extract from facial expressions imposes high expectations for the science and art of facial animation. While the advent of high-resolution performance capture has greatly improved realism, the utility of procedural animation warrants a prominent place in facial animation workflow. We present a system that, given an input audio soundtrack and speech transcript, automatically generates expressive lip-synchronized facial animation that is amenable to further artistic refinement, and that is comparable with both performance capture and professional animator output. Because of the diversity of ways we produce sound, the mapping from phonemes to visual depictions as visemes is many-valued. We draw from psycholinguistics to capture this variation using two visually distinct anatomical actions: **Ja**w and **Li**p, where sound is primarily controlled by jaw articulation and lower-face muscles, respectively. We describe the construction of a transferable template JALI 3D facial rig, built upon the popular facial muscle action unit representation FACS. We show that acoustic properties in a speech signal map naturally to the dynamic degree of *jaw* and *lip* in visual speech. We provide an array of compelling animation clips, compare against performance capture and existing procedural animation, and report on a brief user study.

Keywords: facial animation, procedural animation, lip synchronization, speech synchronization, audio-visual speech.

Concepts: •**Computing methodologies** → **Procedural animation; Natural language generation; Simulation by animation;** •**Applied computing** → **Performing arts;**

*{pif, chrisl, elf, karan}@dgp.toronto.edu

SIGGRAPH '16 Technical Paper,, July 24 - 28, 2016, Anaheim, CA,

1 Introduction

The proportion of our brains involved in facial processing vastly outweighs the processing of other categories of objects [Rossion et al. 2012]. The evolutionary advantage gained by this ability comes with a high developmental cost. Indeed, the years of human acquisition that allow such apparently effortless expression of feeling through our faces is in fact a highly complex phenomenon for actors and animators to realize. Facial animation tools in industrial practice have remained remarkably static, typically using animation software like MAYA to animate a 3D facial rig, often with a simple interpolation between an array of target *blend shapes*. More principled rigs are anatomically inspired with skeletally animated jaw and target shapes representing various facial muscle action units (FACS) [Ekman and Friesen 1978], but the onus of authoring the detail and complexity necessary for human nuance and physical plausibility remain tediously in the hands of the animator.

While professional animators may have the ability, budget and time to bring faces to life with a laborious workflow, the results produced by novices using these tools, or existing procedural or rule-based animation techniques, are generally less flattering. Procedural approaches to automate aspects of facial animation such as lip-synchronization, despite showing promise in the early 1990s, have not kept pace in quality with the complexity of the modern facial models. On the other hand, facial performance capture has achieved such a level of quality that it is a viable alternative to production facial animation. As with all performance capture, however, it has three shortcomings: the animation is limited by the capabilities of the human performer, whether physical, technical or emotional; subsequent artistic refinement is difficult; and partly-hidden anatomical structures that play a part in the animation, such as the tongue, have to be animated separately.

The challenge is thus to produce animator-centric procedural animation tools that are comparable in quality to performance capture, and that are easy to edit and refine. Note that while preserving the ability of expert animators to tune final results to their liking, other

ISBN: 978-1-4503-4279-7/16/07
DOI: http://dx.doi.org/10.1145/2897824.2925984

non-artistic adjustments are often necessary in speech synchronization to deal with, for example, prosody, mispronunciation of text, and speech affectations such as slurring and accents.

In the context of speech synchronization, we define the problem as follows: *given an input audio soundtrack and speech transcript, generate a realistic, expressive animation of a face with lip, tongue and jaw movements that synchronize with the audio soundtrack*. Beyond producing realistic output, we impose further requirements: our workflow must integrate with the traditional animation pipeline, including the use of motion capture, blend shapes and key-framing; we must allow full animator editing of the output; the system must respond to editing of the speech transcript to account for speech anomalies; and it must be possible to target facial animation to a variety of face rigs. We do not require the system to perform synthetic speech generation. We thus position our system to work alongside traditional animation tools. Because the signal processing step that transforms inputs to our internal representation is fast, it would be feasible to extend our work to interactive dialogue, as long as the transcript and audio signal are streamed in time for processing.

The JALI model It is in principle possible to synthesize facial animation by developing control solutions that directly model the neuro-muscular facial function. Such a fine grained model awaits a deeper understanding of face physiology and neural activation sequencing; real-time physical simulation of the face may have to wait even longer. We instead observe that for the task of speech synchronization, we can aggregate its attendant facial motions into two independent categories: functions related to jaw motion, and those related to lip motion (see **Figure 2**) [Badin et al. 2002]. These two dimensions, which are the basis of our JALI model, capture a wide range of the speech phenomenology and permit interactive exploration of an expressive face space.

A facial animation is created in JALI as a sequence of signal and text processing steps as follows.

1. An input speech transcript and corresponding audio soundtrack is taken as input.
2. *Forced alignment* is employed to align utterances in the soundtrack to the text, giving an output time series containing a sequence of phonemes [Brugnara et al. 1993].
3. Audio, text and alignment information are combined to give text/phoneme and phoneme/audio correspondences.
4. Lip-synchronization viseme action units are computed first by extracting jaw and lip motions for individual phonemes. Humans, however, do not articulate each phoneme separately. We thus blend the corresponding visemes into *co-articulated* action units that more accurately track real human speech [Deng et al. 2006; Jurafsky and Martin 2008].

Contributions This paper makes several contributions to the problem of expressive speech synchronization:

- The jaw and lip action model, its validation, and the construction of FACS-based JALI rigs that provide a wide range of idiomatic facial speech expression.
- That procedural facial animation techniques for speech synchronization can produce expressive results that appear to surpass existing procedural techniques in both scope and quality, to the point that they approach the results produced by performance capture or expert key-framed approaches.
- An end-to-end automated solution that is amenable to further manual refinement, and that can be combined with other facial animation methods.
- The ability to animate automatically anatomical features such as the tongue that are only occasionally and partially visible.

We follow a survey of related work (Section 2) by a detailed development and validation of the JALI Viseme model (Section 3). We then show how the JALI Viseme model can be constructed over a typical FACS-based 3D facial rig and transferred across such rigs (Section 4). Section 5 provides system implementation details for our automated lip-synchronization approach. We show a variety of compelling examples produced by our system, and by way of evaluation provide comparisons with performance capture, professional hand animation, and state of the art in procedural lip-synchronization techniques. Section 6 concludes with other applications of the JALI model and directions for future work.

2 Related Work

Computer facial animation has been an active area of research since the early 1970s [Parke 1972; Parke and Waters 1996]. The large corpus of research on audiovisual speech animation [Bailly et al. 2012] can further be broadly classified as *procedural*, *data-driven*, or *performance-capture*.

Procedural speech animation segments speech into a string of phonemes, which are then mapped by rules or look-up tables to visemes – typically many-to-one, (e.g., / m b p / all map to the viseme *MMM* in **Figure 4**). **Viseme** is short for *visible phoneme* [Fisher 1968] and refers to the shape of the mouth at the apex of a given phoneme. This would be a simple affair were it not for the human habit of *co-articulation*. When humans speak, our visemes overlap and crowd each other out in subtle ways that have baffled speech scientists for decades. Thus, the crux of any procedural model is its co-articulation scheme. A popular model is Cohen and Massaro's *dominance model* [1993; 2012]. It uses dominance functions that overlap, giving values indicating how close a given viseme reaches its target shape given its neighbourhood of phonemes. A common weakness of this model is the failure to ensure lip closure of bilabials (/m b p/) [Mattheyses and Verhelst 2015]. Several variants/improvements of the model have been developed over the years, such as [Cosi et al. 2002; King and Parent 2005]. *Rule-based co-articulation* models use explicit rules to dictate the co-articulation under explicit circumstances [Kent and Minifie 1977; Pelachaud et al. 1996; Bevacqua and Pelachaud 2004; Wang et al. 2007]. Diphone co-articulation [Xu et al. 2013] defines a specific animation curve for every pair of phonemes used in a given language. These are then concatenated to generate speech animation. This approach has also been explored for tri-phone co-articulation [Deng et al. 2006]. Several off-the-shelf tools are available to produce procedural lip-sync such as FACEFX (*www.facefx.com*).

Procedural animation techniques generally produce compact animation curves amenable to refinement by animators, but over time have lost ground in terms of expressive realism to data-driven and performance-capture methods [Orvalho et al. 2012]. We hope to rejuvenate animator-centric procedural facial animation by demonstrating results comparable with current production techniques. As noted by Taylor et al. [2012], phonemes map one-to-many to visemes (e.g., the phoneme /t/ has the lip shape of /OW/ in the word 'stow' but the lip shape of /y/ in the word 'sty'), based on phonetic context. This can be handled by a categorical co-articulation model such as ours, or by using data-driven dynamic visemes [Taylor et al. 2012]. None of the above methods explicitly model speech styles, namely the continuum of viseme shapes manifested by intentional variations in speech, which is what our JALI viseme field models.

Data-driven methods smoothly stitch pieces of facial animation data from a large corpus, to match an input speech track [Bregler et al. 1997; Cao et al. 2005; Ma et al. 2006]. Multi-dimensional morphable models [Ezzat et al. 2002], hidden Markov models

[Brand 1999; Wang et al. 2012], and active appearance models (AAM) [Anderson et al. 2013; Bailly et al. 2009] and have been used to capture facial dynamics. For example, AAM-based, Dynamic Visemes [Taylor et al. 2012] uses cluster sets of related visemes, gathered through analysis of the TIMIT corpus. Data-driven methods have also been used to drive a physically-based model [Sifakis et al. 2006]. A statistical model for evaluating the quality of data-driven speech is available [Ma and Deng 2012]. However, the quality of data-driven approaches is often limited by the data available; many statistical models drive the face directly, taking ultimate control away from the animator.

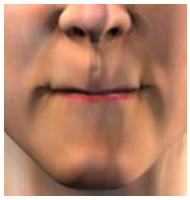
Performance-capture based speech animation transfers acquired motion data from a human performer onto a digital face model [Williams 1990]. Performance capture has gained in popularity in recent years with the widespread adoption of cameras, depth sensors and other motion capture equipment. Approaches such as Face/Off [Weise et al. 2009] and other work based on real-time performance-based facial animation [Weise et al. 2011; Li et al. 2013], while not specifically focused on speech, are able to create high-quality facial animation. Speech analysis can complement and improve these approaches [Weise et al. 2011]. A hybrid capture/data-driven system by [Liu et al. 2015] uses a precaptured database to correct performance capture with a deep neural network trained to extract phoneme probabilities from audio input in real-time using an Kinect sensor. Various commercial products such as FACESHIFT (*faceshift.com*) and FACEWARE (*faceware.com*), also provide facial performance capture using commodity hardware. The disadvantage of performance capture is that is limited by the actor's abilities and is difficult for an animator to refine.

We show that our results compare favourably with procedural [Massaro et al. 2012], data-driven [Taylor et al. 2012] and performance-capture FACEWARE approaches in **Figure 9**.

3 JALI Viseme Model

Our JALI viseme model is motivated by the directly observable bio-acoustics of sound production using a mixture of diaphragm, jaw, and lip. Experimental and empirical validation for our visual model is provided by [Badin et al. 2002], which, through a variety of data capture processes, determined that the majority of variation in visual speech is accounted for by jaw, lip and tongue motion. While trained ventriloquists are able to speak entirely using their diaphragm with little observable facial motion [Metzner et al. 2006], we typically speak using a mix of independently controllable **JA**W and **LI**P facial action. JALI simulates visible speech as a linear mix of jaw-tongue (with minimal face muscle) action and face-muscle action values. The absence of any JA and LI action is not a static face but one perceived as poor-ventriloquy or mumbling, and the other extreme is hyper-articulated screaming (**Figure 2**). A powerful feature of the JALI model is thus the ability to capture a broad variety of visually expressive speaking styles.

3.1 JALI Motivation



Traditional animation of human speech [Blair 1947], is based on a mapping from *phonemes* to a *visemes*, such as the two labio-dental phonemes /f v/ mapping to a single FFF viseme shown inset, where the lower lip is pressed against the upper teeth. Animators create linearly superposed blend-shapes to represent these visemes and animate speech by keyframing these blend-shapes over time [Osipa 2010]. This traditional approach overlooks the fact that phonemes in speech can be expressed by a continuum of viseme shapes based on phonetic

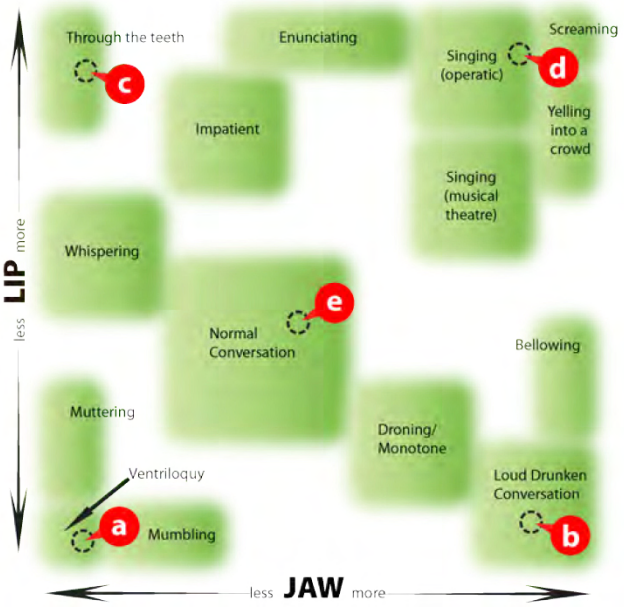
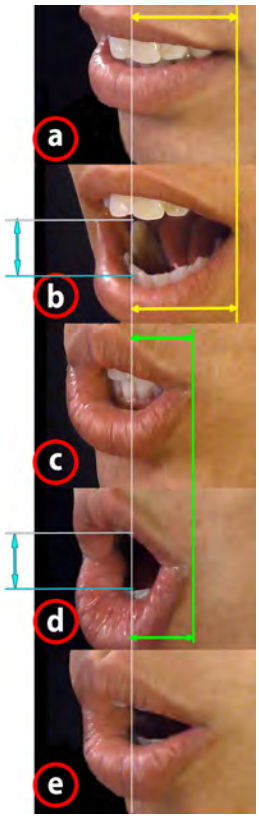


Figure 2: *Speaking styles captured by the* JALI *viseme field.*

context and speech style. When we *hyper-articulate* (i.e., over-enunciate), we form visemes primarily with lip motion using facial muscles, with little or no jaw movement. Conversely, when we *hypo-articulate* (i.e., speak in a drone), we use primarily jaw/tongue motion with little or no lip action. In normal conversation, we use varying combinations of lip and jaw/tongue formations of visemes arbitrarily. We can thus map each phoneme to a 2D viseme field along nearly independent jaw and lip axes that captures a wide range of expressive speech (**Figure 2**).



Visemes corresponding to five arbitrarily-chosen speaking styles for the phoneme /AO/ in 'thOUght' performed by an actor are shown inset. In all five articulations /AO/ is pronounced with equal clarity and volume, but with considerable viseme variation. From (a) to (e) (also marked on **Figure 2**), /AO/ is pronounced: like an amateur ventriloquist with minimal jaw and lip activity (a); with considerable jaw activity but little or no facial muscle activity, as in loud drunken conversation (b); with high face muscle activation but minimal jaw use, as though enunciating 'through her teeth' (c); with substantial activity in both jaw and lip, like singing operatically (d); and with moderate use of both lip and jaw, in normal conversation (e). Note that the lip width is consistent for (a) and (b) (both having minimal lip activation), and for (c) and (d) (maximal lip activation). Also note that jaw opening is consistent for (a) and (c) (both having minimal jaw activation) and for (b) and (d) (maximal jaw activation). The varying use of speaking styles is illustrated in two video clips available with this paper: "Phonemes in Three Speaking

Styles" as shown above, and "Five Moods of Little Bo Peep", where a male actor recites a nursery rhyme, transitioning through several distinct speaking styles.

As seen in the video "Five Jacks", the JALI viseme field provides an animator with an easy to control abstraction over expressive speech animation of the same phonetic content. We further show in Section 4.3 that the JALI field setting over time, for a given performance, can be extracted plausibly through analysis of the audio signal. A combination of our JALI viseme field with our improved procedural lip-synchronization algorithm is thus able to animate a character's face with considerable realism and accuracy (Section 6).

3.2 JALI-driven characters: Valley Girl and Boy

Figure 3: *JALI rigs: Valley Girl and Valley Boy*

We now describe the construction of an animatable facial rig compatible with the JALI viseme field (Figure 3). Valley Girl (named after *The Uncanny Valley* [Mori 1970]), is a fairly realistic facial model rigged in MAYA. Her face is controlled through a typical combination of blend-shapes (to animate her facial action units) and skeletal skinning (to animate her jaw and tongue). The rig controls are based on FACS but do not exhaustively include all AUs, nor is it limited to AUs defined in FACS.

A traditional facial rig sometimes has individual blend-shapes for each viseme (usually with a many-to-one mapping from phonemes to visemes, or many-to-many using dynamic visemes [Taylor et al. 2012]). A JALI-rigged character requires that such visemes be separated to capture sound production and shaping as mixed contribution of the jaw, tongue and facial muscles that control the lips. In other words, the *face* geometry is a composition of a neutral face *nface*, overlaid with skeletal jaw and tongue deformation jd, td, displaced by a linear blend of weighted blend-shape action unit displacements au, i.e., $face = nface + jd + td + au$.

To create a viseme within the 2D field defined by *JA* and *LI* for any given phoneme p, it seems natural to set the geometric $face(p)$ for any point *JA,LI* in the viseme field of p to be $face(p, JA, LI) = nface + JA * (jd(p) + td(p)) + LI * au(p)$, where $jd(p), td(p), au(p)$, represent an extreme configuration of the jaw, tongue and lip action units for the phoneme p. Suppressing both the *JA* and *LI* values here would result in a static neutral face, barely obtainable by the most skilled of ventriloquists. Natural speech without *JA*, *LI* activation is closer to a mumble or an amateur attempt at ventriloquy.

Open-Jaw Neutral Pose and 'Ventriloquist Singularity': We configure the neutral face of the JALI rig so that the character's jaw hangs open slightly (**Figure 5b**), and the lips are locked with a low-intensity use of the "lip-tightening" muscle (*orbicularis oris*, AU 23), as if pronouncing a bilabial phoneme such as /m/ (**Figure 5c**).

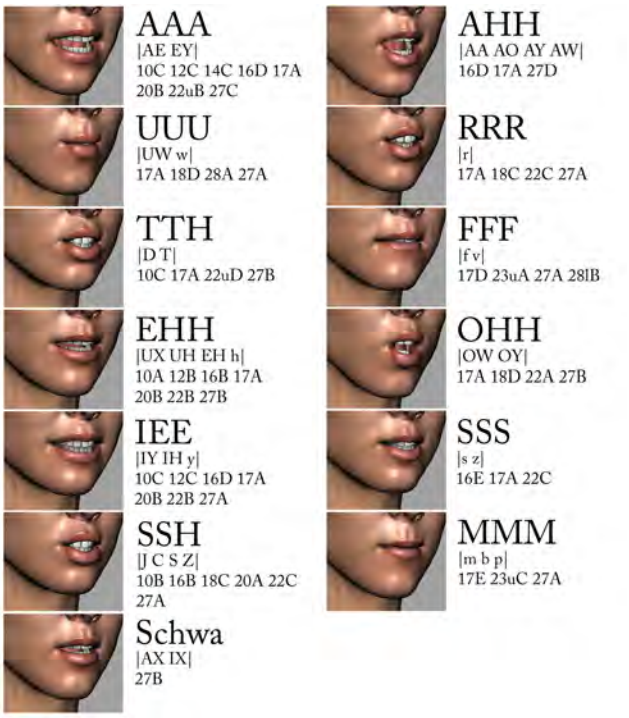


Figure 4: *Our phoneme-to-viseme mapping. To the right of each image we have the viseme name (top), phonemes (middle), action units (bottom). Only the AUs listed here are needed to replicate our model. Here, JA=0.333, LI=0.5*

This JALI neutral face is more faithful to a relaxed human face than the commonly used neutral face, with jaw clenched shut and no facial muscles activated (**Figure 5a**).

The JALI neutral face is thus better suited to produce 'ventriloquist' visemes (with zero (*JA*,*LI*) activation). We use three 'ventriloquist' visemes: the neutral face itself (for the bilabials /b m p/), the neutral face with the *orbicularis oris superior* muscle relaxed (for the labiodentals /f v/), and the neutral face with both *orbicularis oris superior* and *inferior* muscles relaxed, with lips thus slightly parted (for all other phonemes). This 'Ventriloquist Singularity' at the origin of the viseme field (i.e. (*JA*,*LI*) = (0,0)), represents the lowest-energy viseme state for any given phoneme.

For any given phoneme p, the geometric *face* for any point (p, *JA*, *LI*) is thus defined as $face(p, JA, LI) = nface + JA * jd(p) + (vtd(p) + JA * td(p)) + (vau(p) + LI * (au(p))$, where $vtd(p), vau(p)$ are the small tongue and muscle deformations necessary to pronounce the ventriloquist visemes.

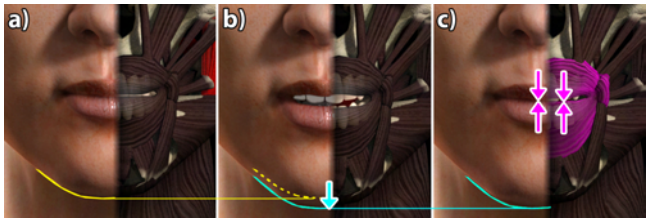


Figure 5: *Making a* JALI *neutral face: (a) neutral face of a conventional rig, with closed jaw and lips, (b) jaw hanging open from gravity, and (c) neutral face of a* JALI *rig, with open jaw and lips closed due to lip-tightening muscle (AU23).*

3.3 Animated Speech using JALI

The JALI model provides a layer of speech abstraction over the phonetic structure. A JALI rig can be phonetically controlled by traditional keyframing or automatic procedurally generated animation (Section 4.2). The JALI viseme field can be independently controlled by the animator over time, or automatically driven by the audio signal (Section 4.3). As shown in the video "Five Jacks", our character Valley Boy is animated to a voice track of the nursery rhyme "Jack and Jill" repeated to a click track, five times in varying speaking styles. In all five performances, we then use a single representative set of procedural animation curves for the actor's phonetic performance, and only the (*JA,LI*) controls are varied from one performance to the next.

4 Procedural Lip-Synchronization

Our method has three main phases: *Input*, *Animation* and *Output* (**Figure 6**). The *Input Phase* produces an alignment of the input audio recording of speech and its transcript, by parsing the transcript into phonemes and then aligning the phonemes with the audio using any off-the-shelf forced-alignment tool. In the *Animation Phase*, the aligned phonemes are mapped to visemes, viseme amplitudes are set (articulation), then re-processed for co-articulation to produce viseme timings, and resulting animation curves for the visemes (in our case, a Maya *MEL* script of sparsely keyframed visemes). The *Output Phase* drives the animated viseme values on a viseme compatible rig such as shown in **Figure 4**. For JALI compatible rigs, we can further compute and control JALI values from the an analysis of the recording as described in Section 4.3.

Figure 6: *Procedural Lip-Synchronization schematic.*

4.1 Input Phase

Accurate speech transcript is critical to procedural lip-synchronization, as extra, missing, or mispronounced words and punctuation can result in poor alignment and cause cascading errors in the animated speech. While we have experimented with automatic transcription tools, which would be essential for real-time speech anmiation, we found manual transcription from the speech recording to be easy and suitable for the purpose of this paper. Many transcript text-to-phoneme conversion for various languages are freely available. We use speech libraries built into Mac OS X, to convert English text into a phonemic representation.

Forced Alignment is then used to align the speech audio to its phonemic transcript. Unlike the creation of speech text transcript, this task requires automation, and is typically done by training a Hidden Markov Model (HMM) on speech data annotated with the beginning, middle, and end of each phoneme, and then aligning phonemes to the speech features. Several tools exist for this task, including HTK [Young and Young 1993], SPHINX [Carnegie Mellon University 2014], and the FESTIVAL system [Black et al. 2001] which was used by [Cao et al. 2005] and [Deng et al. 2006]. In our tests, alignment misses are within 15 ms of the actual timings. For our purposes, due to co-articulation, anticipation and perceptual factors, this level of accuracy is less than ideal, but adequate.

4.2 Animation Phase

We animate a facial rig by producing sparse animation keyframes for visemes. The viseme to be keyframed is determined by our co-articulation model, its timing is determined by forced alignment after it has been processed by through the co-articulation model, and the amplitude is determined by lexical and word stresses returned by the phonemic parser. We build our visemes on Action Units (AU), and can thus drive any facial rig (simulated muscle, blend-shape, or bone-based) that has a Facial Action Coding System (FACS) [Ekman and Friesen 1978] or MPEG-4 FA [Pandzic and Forchheimer 2002] based control system.

Amplitude We set the amplitude of the viseme based on two inputs: *Lexical Stress* and *Word Prominence*. These are retrieved as part of the phonemic parsing. Lexical Stress indicates which vowel sound in a word is emphasized by convention. For example, the word 'water' stresses the 'a' not the 'e' by convention. One can certainly say 'watER' but usually people say 'WAter'. Word Prominence is the de-emphasis of a given word by convention. For example, the 'of' in 'out *of* work' has less word prominence than its neighbours. If a vowel is lexically stressed, the amplitude of that viseme is set to high (e.g., 9 out of 10). If a word is de-stressed, then all visemes in the word are lowered (e.g., 3), if a de-stressed word has a stressed phoneme or it is an un-stressed phoneme in a stressed word, then the viseme is set to normal (e.g., 6).

Co-articulation Our timing is based on the alignment returned by the forced alignment and the results of the co-articulation model. Given the amplitude, the phoneme-to-viseme conversion must be processed through a co-articulation model or else the lips, tongue and jaw will distinctly pronounce each phoneme, which is neither realistic nor expressive. Severe mumbling or ventriloquism makes it clear that coherent audible speech can often be produced with very little visible facial motion, making co-articulation essential.

In Linguistics: "*Co-articulation* is the movement of articulators to anticipate the next sound or preserving movement from the last sound" [Jurafsky and Martin 2008]. Fortunately the graphics of speech has a few simplifying agents. First, many phonemes map to a single viseme (e.g., the phonemes: /AO/ (c**au**ght), /AX/ (**a**bout), AY/ (**i**te), and /AA/ (f**a**ther) all map to the viseme AHH. See **Figure 4** for details). Second, all motion of the tongue is nearly completely hidden; only glimpses of motion are necessary to convince the viewer the tongue is playing its part.

This general model for audio-visual synchronized speech is based on three anatomical dimensions of visible movements: *Tongue*, *Lips* and *Jaw*. Each affect speech and co-articulation in particular ways. The rules for visual speech production are for the most part based on linguistic categorization and divided into *constraints*, *conventions* and *habits*.

There are four **constraints of articulation** that must be met:

1. *Bilabials* (m b p) must close the lips (e.g., 'm' in **m**ove);
2. *Labiodentals* (f v) must touch bottom-lip to top-teeth or cover top-teeth completely (e.g., 'v' in mo**v**e);
3. *Sibilants* (s z J C S Z) narrow the jaw greatly (e.g., 'C' and 's' in '**Ch**e**ss**' both bring the teeth close together);
4. *Non-Nasal* phonemes must open the lips at some point when uttered (e.g., 'n' does not need open lips).

These visual rules are easily observable and – for all but a trained ventriloquist – necessary, to physically produce these phonemes.

These are three **speech conventions** which influence articulation:

1. *Lexically-stressed vowels* usually produce strongly-articulated corresponding visemes (e.g., 'a' in w**a**ter);
2. *De-stressed words* usually get weakly-articulated visemes for the length of the word (e.g., 'and' in 'cats **and** dogs'.);
3. *Pauses* (, . ! ? ; : aspiration) usually leave the mouth open.

It takes conscious effort to break these rules and we have not seen a common visual speaking style that is not influenced by them.

There are nine **co-articulation habits** that shape neighbouring visemes:

1. *Duplicated visemes* are considered one viseme (e.g., /p/ and /m/ in 'po**p** **m**an' are co-articulated into one long MMM viseme.);
2. *Lip-heavy visemes* (UW OW OY w S Z J C) start early (anticipation) and end late (hysteresis);
3. *Lip-heavy visemes* replace the lip shape of neighbours that are not labiodentals and bilabials;
4. *Lip-heavy visemes* are simultaneously articulated with the lip shape of neighbours that are labiodentals and bilabials;
5. *Tongue-only visemes* (l n t d g k N) have no influence on the lips: the lips always take the shape of the visemes that surround them;
6. *Obstruents and Nasals* (D T d t g k f v p b m n N) with no similar neighbours, that are less than one frame in length, have no effect on jaw (excluding Sibilants);
7. *Obstruents and Nasals* of length greater than one frame, narrow the jaw as per their viseme rig definition;
8. Targets for co-articulation look into the word for their shape, always anticipating, except that the last phoneme in a word looks back (e.g., both /d/ and /k/ in 'duke' take their lip-shape from the 'u'.);
9. *Articulate* the viseme (its tongue, jaw and lips) without co-articulation effects, if none of the above rules affect it.

Speech Motion Curves The goal of speech motion is to optimize both *simplicity* (for benefit of the editing animator) and *plausibility* (for the benefit of the unedited performance).

In general, speech onset begins 120ms before the apex of the viseme: the apex coincides with the beginning of the sound [Bailly 1997]. The apex is sustained in an arc to the point where 75% of the phoneme is complete, then it takes another 120 ms to decay to zero [Ito et al. 2004]. However, there is evidence [Schwartz and Savariaux 2014] that there is a variance in onset times for different classes of phonemes and phoneme combinations. Empirical measurements have been made [Chandrasekaran et al. 2009] of specific phonemes /m p b f/ in two different states: after a pause (mean range: 137-240ms) and after a vowel (mean range: 127-188ms). We use these context-specific, phoneme-specific mean-time offsets in our model, and ours is the first procedural model to do so. Phoneme onsets are parameterized in our system, so new empirical measurements of phonemes onsets can be quickly assimilated.

The keen observer will note that if phoneme durations are very short, then visemes will have a wide influence beyond its direct neighbours. This is intentional. Visemes are known to influence mouth shape up to five phonemes away [Kent and Minifie 1977], specifically lip-protrusion [Mattheyses and Verhelst 2015]. In our implementation, each mouth shape is actually influenced by both direct neighbours (since the start of one is the end of another and both are keyed at the point). The second-order neighbours are likely also involved since each viseme starts at least 120ms before it is heard and ends 120ms after; in the case of lip-protrusion it is extended to 150ms onset and offset.

The **Arc** is one of Lasseter's Principles of Animation [Lasseter 1987] and in accordance, we fatten and retain our action in one smooth motion arc over duplicated visemes. All the phoneme articulations have an exaggerated quality in line with the principle of **Exaggeration**. This is due to the clean curves, the sharp rise and fall of each phoneme, each simplified and each slightly more distinct from its neighbouring visemes than in real-world speech.

4.3 Computing JALI values from Audio

We animate the JA and LI parameters of the JALI-based character by examining the pitch and intensity of each phoneme and comparing it to all other phonemes of the same class uttered in a given performance. We look at three classes of phonemes: vowels, plosives and fricatives. Each of these classes requires a slightly different method of analysis to animate the Lip parameter. *Fricatives* (s z f v S Z D T) create friction by pushing air past the teeth with either the lips or the tongue. This creates intensity at high frequencies, and thus they have markedly increased mean frequencies in their spectral footprints compared to those of conversational speech [Maniwa et al. 2009]. If we detect greater intensity at a high frequency for a given fricative, then we know that it was spoken forcefully and heavily-articulated. Likewise, with *Plosives* (p b d t g k), the air stoppage by lip or tongue builds pressure and the sudden release creates similarly high frequency intensity: the greater the intensity the greater the articulation.

Unlike fricatives and plosives, vowels are always voiced. This fact allows us to measure the pitch and volume of the glottis with some precision. Simultaneous increases in pitch and volume are associated with emphasis. High mean formant F_0 and high mean intensity are correlated with high arousal [Banse and Scherer 1996; Bachorowski 1999; Albrecht et al. 2005] (panic, rage, excitement, joy) which are associated with bearing teeth and greater articulation, and exaggerated speech [Hill et al. 2005]. Likewise simultaneous decreases are associated with low arousal [Banse and Scherer 1996] (shame, sadness, boredom).

A very important factor is that vowels are only considered if they are *lexically stressed* and fricatives/plosives are only considered if they arise before/after a lexically stressed vowel. This chooses our candidates carefully and keeps our animation from being too erratic. Specifically, lexically stressed sounds will be hit hardest by the intention to articulate, yell, speak strongly or emphasize a word in speech. Likewise the failure to do so will be most indicative of a mutter, mumble or an intention not to be clearly heard, due for example to fear, shame, or timidity.

There are other advantages to this method. The friction of air through lips and teeth make high frequency sounds which impair comparison between fricative/plosives and vowel sounds on both the pitch and intensity dimension, so they must be separated from vowels for coherent/accurate statistical analysis. We are comparing these three phoneme types separately because of the unique characteristics of the sound produced (these phoneme-types are categorically different). We do this in the way that best identifies changes specific to each given phoneme type. As future work we intend to add other phoneme-types and use a method of analysis best suited to detect articulation of that type.

Pitch and intensity of the audio is analyzed with PRAAT [Boersma and Weenink 2014]. Voice pitch is measured spectrally in *Hz* and retrieved from the *fundamental frequency*. The fundamental frequency of the voice is the rate of vibration of the glottis and abbreviated as **F0** [Jurafsky and Martin 2008]. Voice intensity is measured in dB and retrieved from the *power* of the signal. The significance of these two signals is that they are perceptual correlates. Intensity is power normalized to the threshold of human hearing and pitch

is linear between 100-1000Hz, corresponding to the common range of the human voice, and non-linear (logarithmic) above 1000Hz. In our implementation, high-frequency intensity is calculated by measuring the intensity of the signal in the 8-20kHz range.

For vocal performances of a character that is shouting throughout, automatic modulation of the JA parameter is not needed. The jaw value can simply be set to a high value for the entire performance. However, when a performer fluctuates between shouting and mumbling an automatic method becomes useful. Our method gathers statistics, mean/max/min and standard deviation for each, intensity and pitch and high frequency intensity. **Table 1** shows how jaw values are set for vowels. **Table 2** shows how lip values are set for vowels. **Table 3** shows how lip values are set for fricatives and plosives.

Intensity of vowel vs. Global mean intensity	Rig Setting
vowel_intensity $\leq$ mean – stdev	Jaw(0.1-0.2)
vowel_intensity $\approx$ mean	Jaw(0.3-0.6)
vowel_intensity $\geq$ mean + stdev	Jaw(0.7-0.9)

Table 1: *Jaw triggers and rig settings for a given vowel. The 'vowel_intensity' is of the current vowel, 'mean' is the global mean intensity of all vowels in the audio clip.*

Intensity/pitch of vowel vs. Global means	Rig Setting
intensity/pitch $\leq$ mean – stdev	Lip(0.1-0.2)
intensity/pitch $\approx$ mean	Lip(0.3-0.6)
intensity/pitch $\geq$ mean + stdev	Lip(0.7-0.9)

Table 2: *Lip triggers and rig settings for a given vowel. The 'intensity/pitch' is of the current vowel, 'mean' is the respective global mean intensity/pitch of all vowels in the audio clip.*

HF Intensity fricative/plosive vs. Global means	Rig Setting
intensity $\leq$ mean – stdev	Lip(0.1-0.2)
intensity $\approx$ mean	Lip(0.3-0.6)
intensity $\geq$ mean + stdev	Lip(0.7-0.9)

Table 3: *Lip triggers and rig settings for a given fricative or plosive. The 'intensity' is the high frequency intensity of the current fricative or plosive, 'mean' is the respective global mean high frequency intensity of all fricatives/plosives in the audio clip.*

5 Additional Implementation Details

Given two input files representing speech audio and text transcript, we use *applescript* and *praatscript* to produce a phonemic breakdown and forced alignment using the Apple utility REPEAT AFTER ME. This phonemic alignment is then used by PRAAT to produce pitch and intensity mean/min/max for each phoneme. We then run through the phonemes to create animated viseme curves by setting articulation and co-articulation keyframes of visemes, as well as animated JALI parameters, as a *MEL* script. This script is able to drive the animation of any JALI rigged character in MAYA.

A 10-second audio clip, processed on a 3.06 GHZ Intel Core i3 with 4 GB 1333 MHZ DDR3 memory and a ATI Radeon 4670 with 256 MB of VRAM (i.e., a mid-2010 iMac), takes a total of 6.487 seconds of processing time to run the various animation scripts. While this running time is sufficient for interactive animation work, with further optimization, our process will run in near real-time.

6 Results and Evaluation

6.1 Low-dimensionality

In this section, we demonstrate an important, attractive feature of the signals produced by our system: low-dimensionality. Our intent is to match the dimensionality of our output to the human communication signal. That is, people perceive phonemes and visemes, not arbitrary positions of part of the face. For example, the procedural result of saying the word "water" in **Figure 8** is more comprehensible and more amenable to animator editing than the motion capture result shown in **Figure 7** (created with FACESHIFT).

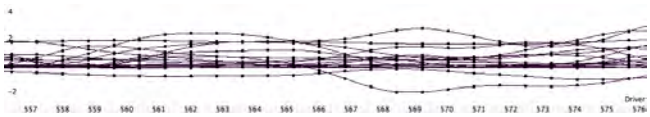


Figure 7: *648 points recorded for the performance capture of the word 'water' at 30 fps. Compare with **Figure 8**.*

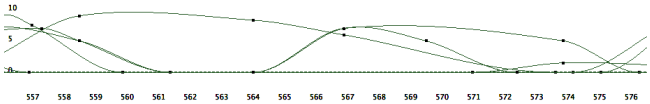


Figure 8: *20 points calculated for the word 'water' as output by our system. When compared with **Figure 7** we see performance capture requires 32.4x as many points as our method to represent the same word. Notice the long regular construction and arc shape in each animation curve; animators prefer to see and edit curves with this shape.*

6.2 Examples

Some vocal tracks were sourced from Librivox [LibriVox 2014], a website of free, public-domain readings of famous, public-domain texts performed by volunteers. The 'Quality of Mercy' video was created from this sound library.

6.3 Comparisons

The success of any realistic procedural animation model can be evaluated by comparing that animation to 'ground truth', i.e., a live-action source. Using live-action footage provided by [Taylor et al. 2012], we evaluate our JALI model by comparing it not only to this footage, but to the speech animation output from Dynamic Visemes [Taylor et al. 2012], and the Dominance model [Massaro et al. 2012], (see Related Work for details).

In this comparison, we utilize a facial motion capture tool, ANALYZER from Faceware Technologies, to track the face of the live-action actor from this footage, as well as the animated faces output from the aforementioned methods. We then use Faceware's RETARGETER to apply these tracking data to animate ValleyBoy, allowing us to evaluate these disparate models on a single facial rig. By comparing JALI, Dynamic Visemes and the Dominance model to the 'ground truth' of the motion-captured live-action footage, we can determine the relative success of each method. The included video, "Comparison of Procedural Speech Methods" demonstrates this evaluation using heatmaps of the displacement errors of each method with respect to the live-action footage.

In **Figure 9** we see successes and failures of all three procedural methods. In (a), we see a timing error with the Dynamic Viseme

model, in that the lips fail to anticipate the leading phoneme just prior to the first spoken sentence. In (b), the Dominance method shows a lack of lip closing in the /F/ phoneme "to Fit"–the result of excessive co-articulation with adjacent vowel phonemes. In (c), the JALI method shows error in the lower lip, as it over-enunciates /AA/ ("dArkness").

In (d) we see the accumulated error for the 7-second duration of the actor's speech. The Dynamic Viseme and JALI models fare significantly better than the Dominance model in animating this vocal track. In general, Dominance incurs excessive co-articulation of lip-heavy phonemes such as /F/ with adjacent phonemes. The Dynamic Viseme model appears to under-articulate certain jaw-heavy vowels such as /AA/, and to blur each phoneme over its duration. JALI appears to over-articulate these same vowels at times.

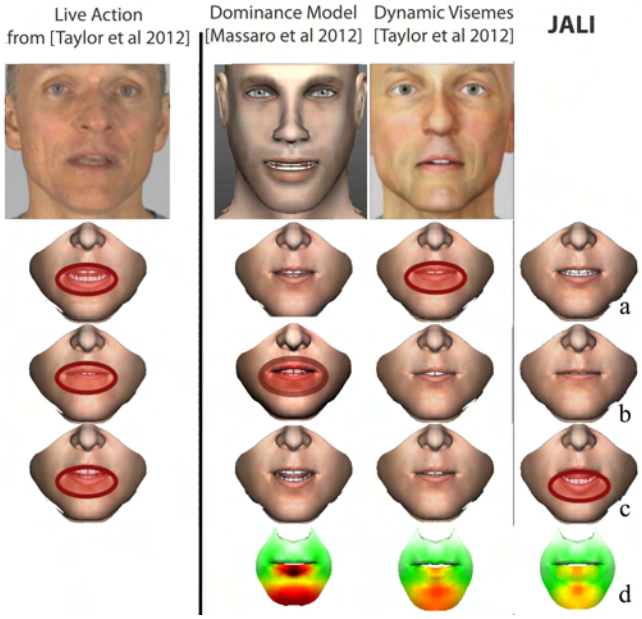


Figure 9: *Cumulative heat-map errors of three techniques.*

6.4 Technical Evaluation

The goal of lip-sync is for the viewer to remain undistracted while following the performance. Since the mouth is one of many features of the face, this is an easier goal than one may foresee at first blush. Nonetheless, we have identified several sources of error.

1. The forced alignment HMM may give incorrect timings in the range of 1-30ms.
2. The forced alignment HMM may improperly categorize a phoneme placing it too early, dropping out a phoneme and cascading the error through to the end of the word, phrase or simply matching the wrong viseme at the moment when the sound produced.
3. The transcript may not match the recording. Extra/missing words and punctuation, or a severely mispronounced word, can create cascading errors through the word or phrase.

Many of these errors can occur regularly with little perceptual impact on the performance. This is due to four factors: 1) people are generally bad lip-readers; 2) co-articulation blurs the boundaries between phonemes; 3) 40 phonemes map to only 14 visemes, so an improperly identified phoneme may still match the right viseme; 4) the mouth may not be the central focus when watching an animation.

6.5 User Evaluation

A short, informal pilot study was performed to get reaction from animators. We recruited three professional animators and one student animator; two animators were males and two were female. They completed three editing tasks: 1) adding a missing viseme, 2) fixing non-trivial out-of-sync phrase and 3) exaggerating a speech performance. Each of these tasks were completed three times: once with hand-generated data, once with mocap-generated data and once with JALI-generated data.

All participants disliked editing motion capture data and unanimously rated it lowest for ease-of-use, ability to reach expectations and quality of the final edited result for all tasks. They described working with the mocap data as "*Nightmarish*", "*...any progress I did and did well, felt like luck*", "*...pretty much better to have a blank slate*", "*an overload of information*", "*way more information than was needed and to make small changes was next-to-impossible*", "*tonnes of clean-up involved and it's hard to do*". Overall, editing with JALI was preferred 77% of the time, but hand-crafted animation was largely seen as equivalent. Given the informality of the study and the small sample size mostly consisting of experienced animators, we are confident in claiming that JALI-generated animation is at least as easy-to-edit as hand-generated animation and was unanimously preferred over mocap-generated animation. We would like in future to perform a more extensive study consisting of inexperienced animators.

In describing working with JALI participants said: "*best of both worlds... it's procedurally generated but you still get the control*", "*The quality could have been done with [hand-generated] ... but I think* JALI *was allowing me to get there faster which is what I liked about it*", "*say it saved me 10%, that is 10% more quality I could put into [the performance]*".

7 Discussion and Conclusions

The effective generation of expressive lip synchronized animation is both an art and science. We have demonstrated a procedural workflow based on a novel jaw/lip model that allows the automatic creation of believable speech-synchronized animation sequences using only text and audio as input. Unlike many data-driven or performance capture methods, our output is animator-centric, and amenable to further editing for more idiosyncratic animation. It is also easy to combine both our technique and its output with other animation workflows, as the accompanying video demonstrates. For example, our lip and jaw animation curves are easily combined with head motion obtained from performance-capture. That said, we believe procedural approaches have untapped potential: an integrated procedural facial animation system that would include head, neck and eye motions, together with higher quality rendering to facilitate more realistic skin and hair, is ongoing work.

Our approach admits a wide range of use cases:

1. In conjunction with **body motion capture**. Often the face and body are captured separately. One could capture the body and record the voice, then use this tool to automatically produce face animation that is quickly synchronized to the body animation *via* the voice recording. This is particularly useful in an VR/AR setting where facial motion capture is complicated by the presence of head mounted display devices.
2. For **video games**, specifically in RPGs where animating many lines of dialogue is prohibitively time-consuming.
3. For **crowds** and **secondary characters** in film, audiences attention is not focused on these characters nor is the voice track forward in the mix.

4. For **animatics** or pre-viz, to settle questions of layout.
5. For **animating main characters** since the animation produced is designed to be edited by a skilled animator.
6. For use in **novice** facial animation.

We look forward to seeing our solution applied to all of these cases.

Limitations and Future work
We do not consider articulation changes [Bevacqua and Pelachaud 2004] for fast *versus* slow speech. We focused on an animator-centric workflow and are pleased with the quality of the results without attention to the rate of speech. However, research shows [Taylor et al. 2014] that rate of speech affects co-articulation patterns, and undoubtedly JALI parameters. This and other audio features like jitter, and shimmer, may all be useful for shaping our JALI model as well as the attack, apex and decay of our animation curves, and this is subject to future work. Emotional speech styles that have been extensively researched, were explicitly excluded in our work. The introduction of an emotional model to ease the creation of subtle-yet-expressive animations – a typically difficult task even for expert animators – is an obvious extension to our work; we are working on an emotional model in our workflow, based on human physiology and guided by human perceptual studies.

In summary, we have presented a novel animator-centric workflow for the automatic creation of lip-synchronized animation: our approach produces results that are comparable or better than prior art in performance-capture and data-driven speech, encapsulating a range of expressive speaking styles that is easy-to-edit and refine by animators. We hope our work will inspire further research in tools like ours that support the creative process of animation.

Acknowledgements

Special thanks are due to our actors Patrice Goodman and Adrien Yearwood. We have benefited considerably from discussions with Gérard Bailly and Dominic Massaro. The financial support of the Natural Sciences and Engineering Research Council of Canada, the Canada Foundation for Innovation, and the Ontario Research Fund, is gratefully acknowledged.

References

ALBRECHT, I., SCHRÖDER, M., HABER, J., AND SEIDEL, H.-P. 2005. Mixed feelings: expression of non-basic emotions in a muscle-based talking head. *Virtual Reality 8*, 4 (Aug.), 201–212.

ANDERSON, R., STENGER, B., WAN, V., AND CIPOLLA, R. 2013. Expressive Visual Text-to-Speech Using Active Appearance Models. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 3382–3389.

BACHOROWSKI, J.-A. 1999. Vocal Expression and Perception of Emotion. *Current Directions in Psychological Science 8*, 2, 53–57.

BADIN, P., BAILLY, G., REVRET, L., BACIU, M., SEGEBARTH, C., AND SAVARIAUX, C. 2002. Three-dimensional linear articulatory modeling of tongue, lips and face, based on {MRI} and video images. *Journal of Phonetics 30*, 3, 533 – 553.

BAILLY, G., GOVOKHINA, O., ELISEI, F., AND BRETON, G. 2009. Lip-Synching Using Speaker-Specific Articulation, Shape and Appearance Models. *EURASIP Journal on Audio, Speech, and Music Processing 2009*, 1, 1–11.

BAILLY, G., PERRIER, P., AND VATIKIOTIS-BATESON, E., Eds. 2012. *Audiovisual Speech Processing*. Cambridge University Press. Cambridge Books Online.

BAILLY, G. 1997. Learning to Speak. Sensori-Motor Control of Speech Movements. *Speech Communication 22*, 2-3 (Aug.), 251–267.

BANSE, R., AND SCHERER, K. R. 1996. Acoustic Profiles in Vocal Emotion Expression. *Journal of Personality and Social Psychology 70*, 3 (Mar.), 614–636.

BEVACQUA, E., AND PELACHAUD, C. 2004. Expressive Audio-Visual Speech. *Computer Animation and Virtual Worlds 15*, 3-4, 297–304.

BLACK, A. W., TAYLOR, P., AND CALEY, R. 2001. *The Festival Speech Synthesis System: System Documentation Festival version 1.4*, 1.4.2 ed.

BLAIR, P. 1947. *Advanced Animation: Learn how to draw animated cartoons*. Walter T. Foster.

BOERSMA, P., AND WEENINK, D., 2014. Praat: doing phonetics by computer [Computer Program]. Version 5.4.04, retrieved 28 December 2014 from http://www.praat.org/.

BRAND, M. 1999. Voice Puppetry. In *SIGGRAPH '99: Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, ACM Press, Los Angeles, 21–28.

BREGLER, C., COVELL, M., AND SLANEY, M. 1997. Video rewrite: Driving visual speech with audio. In *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques*, ACM Press/Addison-Wesley Publishing Co., New York, NY, USA, SIGGRAPH '97, 353–360.

BRUGNARA, F., FALAVIGNA, D., AND OMOLOGO, M. 1993. Automatic segmentation and labeling of speech based on hidden markov models. *Speech Commun. 12*, 4 (Aug.), 357–370.

CAO, Y., TIEN, W. C., FALOUTSOS, P., AND PIGHIN, F. 2005. Expressive Speech-Driven Facial Animation. *ACM Transactions on Graphics (TOG) 24*, 4, 1283–1302.

CARNEGIE MELLON UNIVERSITY, 2014. CMU Sphinx: Open Source Toolkit for Speech Recognition [Computer Program]. Version 4, retrieved 28 December 2014 from http://cmusphinx.sourceforge.net/.

CHANDRASEKARAN, C., TRUBANOVA, A., STILLITTANO, S., CAPLIER, A., AND GHAZANFAR, A. A. 2009. The Natural Statistics of Audiovisual Speech. *PLoS Computational Biology 5*, 7 (July), 1–18.

COHEN, M. M., AND MASSARO, D. W. 1993. Modeling Coarticulation in Synthetic Visual Speech. *Models and Techniques in Computer Animation*, 139–156.

COSI, P., CALDOGNETTO, E. M., PERIN, G., AND ZMARICH, C. 2002. Labial Coarticulation Modeling for Realistic Facial Animation. In *ICMI'02: IEEE International Conference on Multimodal Interfaces*, IEEE Computer Society, 505–510.

DENG, Z., NEUMANN, U., LEWIS, J. P., KIM, T.-Y., BULUT, M., AND NARAYANAN, S. 2006. Expressive Facial Animation Synthesis by Learning Speech Coarticulation and Expression Spaces. *IEEE Transactions on Visualization and Computer Graphics 12*, 6 (Nov.), 1523–1534.

EKMAN, P., AND FRIESEN, W. V. 1978. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*, 1 ed. Consulting Psychologists Press, Palo Alto, California, Aug.

EZZAT, T., GEIGER, G., AND POGGIO, T. 2002. Trainable videorealistic speech animation. In *Proceedings of the 29th Annual Conference on Computer Graphics and Interactive Techniques*, ACM, New York, NY, USA, SIGGRAPH '02, 388–398.

FISHER, C. G. 1968. Confusions among visually perceived consonants. *Journal of Speech, Language, and Hearing Research 11*, 4, 796–804.

HILL, H. C. H., TROJE, N. F., AND JOHNSTON, A. 2005. Range- and Domain-Specific Exaggeration of Facial Speech. *Journal of Vision 5*, 10 (Nov.), 4–4.

ITO, T., MURANO, E. Z., AND GOMI, H. 2004. Fast Force-Generation Dynamics of Human Articulatory Muscles. *Journal of Applied Physiology 96*, 6 (June), 2318–2324.

JURAFSKY, D., AND MARTIN, J. H. 2008. *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition*, 2 ed. Prentice Hall.

KENT, R. D., AND MINIFIE, F. D. 1977. Coarticulation in Recent Speech Production Models. *Journal of Phonetics 5*, 2, 115–133.

KING, S. A., AND PARENT, R. E. 2005. Creating Speech-Synchronized Animation. *IEEE Transactions on Visualization and Computer Graphics 11*, 3 (May), 341–352.

LASSETER, J. 1987. Principles of Traditional Animation Applied to 3D Computer Animation. *SIGGRAPH Computer Graphics 21*, 4, 35–44.

LI, H., YU, J., YE, Y., AND BREGLER, C. 2013. Realtime Facial Animation with on-the-Fly Correctives. *ACM Transactions on Graphics (TOG) 32*, 4, 42.

LIBRIVOX, 2014. LibriVox — free public domain audiobooks. Retrieved 28 December 2014 from https://librivox.org/.

LIU, Y., XU, F., CHAI, J., TONG, X., WANG, L., AND HUO, Q. 2015. Video-Audio Driven Real-Time Facial Animation. *ACM Transactions on Graphics (TOG) 34*, 6 (Nov.), 182.

MA, X., AND DENG, Z. 2012. A Statistical Quality Model for Data-Driven Speech Animation. *IEEE Transactions on Visualization and Computer Graphics 18*, 11, 1915–1927.

MA, J., COLE, R., PELLOM, B., WARD, W., AND WISE, B. 2006. Accurate visible speech synthesis based on concatenating variable length motion capture data. *Visualization and Computer Graphics, IEEE Transactions on 12*, 2 (March), 266–276.

MANIWA, K., JONGMAN, A., AND WADE, T. 2009. Acoustic Characteristics of Clearly Spoken English Fricatives. *Journal of the Acoustical Society of America 125*, 6, 3962.

MASSARO, D. W., COHEN, M. M., TABAIN, M., BESKOW, J., AND CLARK, R. 2012. Animated speech: research progress and applications . In *Audiovisual Speech Processing*, G. Bailly, P. Perrier, and E. Vatikiotis-Bateson, Eds. Cambridge University Press, Cambridge, 309–345.

MATTHEYSES, W., AND VERHELST, W. 2015. Audiovisual Speech Synthesis: An Overview of the State-of-the-Art. *Speech Communication 66*, C (Feb.), 182–217.

METZNER, J., SCHMITTFULL, M., AND SCHNELL, K. 2006. Substitute sounds for ventriloquism and speech disorders. In *INTERSPEECH 2006 - ICSLP, Ninth International Conference on Spoken Language Processing, Pittsburgh, PA, USA, September 17-21, 2006*.

MORI, M. 1970. The Uncanny Valley (aka. 'Bukimi no tani'). *Energy 7*, 4, 33–35.

ORVALHO, V., BASTOS, P., PARKE, F. I., OLIVEIRA, B., AND ALVAREZ, X. 2012. A Facial Rigging Survey. *Eurographics 2012 - STAR – State of The Art Report*, 183–204.

OSIPA, J. 2010. *Stop staring: facial modeling and animation done right*. John Wiley & Sons.

PANDZIC, I. S., AND FORCHHEIMER, R., Eds. 2002. *MPEG-4 Facial Animation*, 1 ed. The Standard, Implementation and Applications. John Wiley & Sons, West Sussex.

PARKE, F. I., AND WATERS, K. 1996. *Computer Facial Animation*. A. K. Peters.

PARKE, F. I. 1972. Computer generated animation of faces. In *Proceedings of the ACM Annual Conference - Volume 1*, ACM, New York, NY, USA, ACM '72, 451–457.

PELACHAUD, C., BADLER, N. I., AND STEEDMAN, M. 1996. Generating Facial Expressions for Speech. *Cognitive Science 20*, 1, 1–46.

ROSSION, B., HANSEEUW, B., AND DRICOT, L. 2012. Defining face perception areas in the human brain: A large-scale factorial fmri face localizer analysis. *Brain and Cognition 79*, 2, 138 – 157.

SCHWARTZ, J.-L., AND SAVARIAUX, C. 2014. No, There Is No 150 ms Lead of Visual Speech on Auditory Speech, but a Range of Audiovisual Asynchronies Varying from Small Audio Lead to Large Audio Lag. *PLoS Computational Biology (PLOSCB) 10(7) 10*, 7, 1–10.

SIFAKIS, E., SELLE, A., ROBINSON-MOSHER, A., AND FEDKIW, R. 2006. Simulating Speech With A Physics-Based Facial Muscle Model. In *SCA '06: Proceedings of the 2006 ACM SIGGRAPH/Eurographics symposium on Computer animation*, Eurographics Association, Vienna, 261–270.

TAYLOR, S. L., MAHLER, M., THEOBALD, B.-J., AND MATTHEWS, I. 2012. Dynamic units of visual speech. In *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, Eurographics Association, Aire-la-Ville, Switzerland, Switzerland, SCA '12, 275–284.

TAYLOR, S. L., THEOBALD, B. J., AND MATTHEWS, I. 2014. The Effect of Speaking Rate on Audio and Visual Speech. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, IEEE, Disney Research, Pittsburgh, PA, 3037–3041.

WANG, A., EMMI, M., AND FALOUTSOS, P. 2007. Assembling an Expressive Facial Animation System. In *Sandbox '07: Proceedings of the 2007 ACM SIGGRAPH symposium on Video games*, ACM.

WANG, L., HAN, W., AND SOONG, F. K. 2012. High Quality Lip-Sync Animation for 3D Photo-Realistic Talking Head. In *ICASSP 2012: IEEE International Conference on Acoustics, Speech and Signal Processing*, 4529–4532.

WEISE, T., LI, H., VAN GOOL, L., AND PAULY, M. 2009. Face/Off: live facial puppetry. In *SCA '09: Proceedings of the 2009 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, ACM Request Permissions, 7–16.

WEISE, T., BOUAZIZ, S., LI, H., AND PAULY, M. 2011. Realtime performance-based facial animation. *SIGGRAPH '11: SIGGRAPH 2011 papers* (Aug.).

WILLIAMS, L. 1990. Performance-driven facial animation. In *Proceedings of the 17th Annual Conference on Computer Graphics and Interactive Techniques*, ACM, New York, NY, USA, SIGGRAPH '90, 235–242.

XU, Y., FENG, A. W., MARSELLA, S., AND SHAPIRO, A. 2013. A Practical and Configurable Lip Sync Method for Games. In *Proceedings - Motion in Games 2013, MIG 2013*, USC Institute for Creative Technologies, 109–118.

YOUNG, S. J., AND YOUNG, S. 1993. *The HTK Hidden Markov Model Toolkit: Design and Philosophy*. University of Cambridge, Department of Engineering.

Appendix

A Procedural Speech Details

We have included psuedocode in **Figure 10** of the rules in **Section 4.2**. We also take a walk through the generation of the visemes for a sample word.

Phonemes = list of phonemes in order of performance
Bilabials = { m b p }
Labiodental = { f v }
Sibilant = { s z J C S Z }
Obstruent = { D T d t g k f v p b }
Nasal = { m n NG }
Pause = { . , ! ? ; : aspiration }
Tongue-only = { l n t d g k NG }
Lip-heavy = { UW OW OY w S Z J C }-

LIP-SYNC *(Phonemes)* :
  for each Phoneme P_i in Phonemes *P*
   if (P_i **isa** lexically_stressed) power = high
  elsif (P_i **isa** destressed) power = low
  else power = normal
   if (P_i **isa** Pause) $P_i = P_{i-1}$
     if (P_{i-1} **isa** Pause) $P_i = P_{i+1}$
   elsif (P_i **isa** Tongue-only)
       ARTICULATE (P_i, power, start, end, onset(P_i), offset(P_i))
       $P_i = P_{i+1}$
       if (P_{i+1} **isa** Pause, Tongue-only) $P_i = P_{i-1}$
   if (viseme(P_i) == viseme(P_{i-1}))
     delete (P_{i-1})
     start = prev_start
   if (P_i **isa** Lip-heavy)
     if (P_{i-1} **isnota** Bilabial,Labiodental) delete (P_{i-1})
     if (P_{i+1} **isnota** Bilabial,Labiodental) delete (P_{i+1})
     start = prev_start
     end = next_end
   ARTICULATE (P_i, power, start, end, onset(P_i), offset(P_i))
  if (P_i **isa** Sibilant) close_jaw(P_i)
  elsif (P_i **isa** Obstruent,Nasal)
       if (P_{i-1},P_{i+1} **isa** Obstruent,Nasal **or** length(P_i) >frame) close_jaw(P_i)
   if (P_i **isa** Bilabial) ensure_lips_close
  elsif (P_i **isa** Labiodental) ensure_lowerlip_close
end

Figure 10: *Procedural Speech Algorithm*

A.1 Walkthrough Example: The word 'what'

Before animation begins, the speech audio track must first be aligned with the text. This happens in two stages: *phoneme parsing* then *forced alignment*. Initially, the word 'what' is parsed into the phonemes: `w 1UX t`; then, the forced alignment stage returns timing information: `w(2.49-2.54)`, `1UX(2.54-2.83)`, `t(2.83-3.01)`. This is all that is needed to animate this word.

Now the speech animation can be generated. First, `w` maps to a **Lip-Heavy** viseme thus begins early (start time would be replaced with the start time of the previous phoneme, if one exists), and ends late (the end time is replaced with the end time of the next phoneme): ARTICULATE ('w', 7, 2.49, 2.83, 150ms, 150ms). Next, the **Lexically-Stressed** `UX` (indicated by the '1' in front) is more strongly articulated; thus power is set to 10 (replacing the default value of 7): ARTICULATE ('UX', 10, 2.54, 2.83, 120ms, 120ms), Finally the `t` maps to a **Tongue-Only** viseme, thus articulates twice, 1) ARTICULATE ('t', 7, 2.83, 3.01, 120ms, 120ms); then it is replaced with the previous, which then counts as a duplicate and thus extends the previous, 2) ARTICULATE ('UX', 10, 2.54, 3.01, 120ms, 120ms).

A.2 Tongue-Only Visemes

Two **Tongue-only** visemes are mentioned in the text, though not fully described. In the interest of clarity we have included a graphic (**Figure 11**) and a description of them, here in the appendix.

Figure 11: *The tongue-only visemes cannot be depicted in FACS notation. These images and short description outlines their construction.* ***LNTD****: the tongue blocks airflow by sealing the upper palate starting at the front of the mouth.* ***GK****: blocks airflow by sealing the airway with the tongue at the back of the mouth.*

A.3 Phoneme Notation

The phonemic notation used throughout this document in shown in **Table 4** and is consistent with Apple libraries.

Symbol	Example	Symbol	Example
%	(silence)	@	(breath intake)
AE	bat	f	fin
EY	bait	g	gain
AO	caught	h	hat
AX	about	J	jump
IY	beet	k	kin
EH	bet	l	limb
IH	bit	m	mat
AY	bite	n	nap
IX	roses	N	tang
AA	father	p	pin
UW	boot	r	ran
UH	book	s	sin
UX	bud	S	shin
OW	boat	t	tin
AW	bout	T	thin
OY	boy	v	van
b	bin	w	wet
C	chin	y	yet
d	din	z	zoo
D	them	Z	measure

Table 4: *Phoneme notation, from https://developer.apple.com/library/mac/documentation/-UserExperience/Conceptual/SpeechSynthesisProgrammingGuide/Phonemes/Phonemes.html.*

A.4 JALI rig

For the purposes of evaluation and to aid construction of JALI on other facial rigs, we have made a simplified JALI-rig available for download: `www.dgp.toronto.edu/~elf/jali.html`.