VOCAL: Vowel and Consonant Layering for Expressive Animator-Centric Singing Animation

Yifang Pan
University of Toronto
Canada
evan.pan@dgp.toronto.edu

Chris Landreth
University of Toronto
Canada
chrisl@dgp.toronto.edu

Eugene Fiume
Simon Fraser University
Canada
elf@dgp.toronto.edu

Karan Singh
University of Toronto
Canada
karan@dgp.toronto.edu

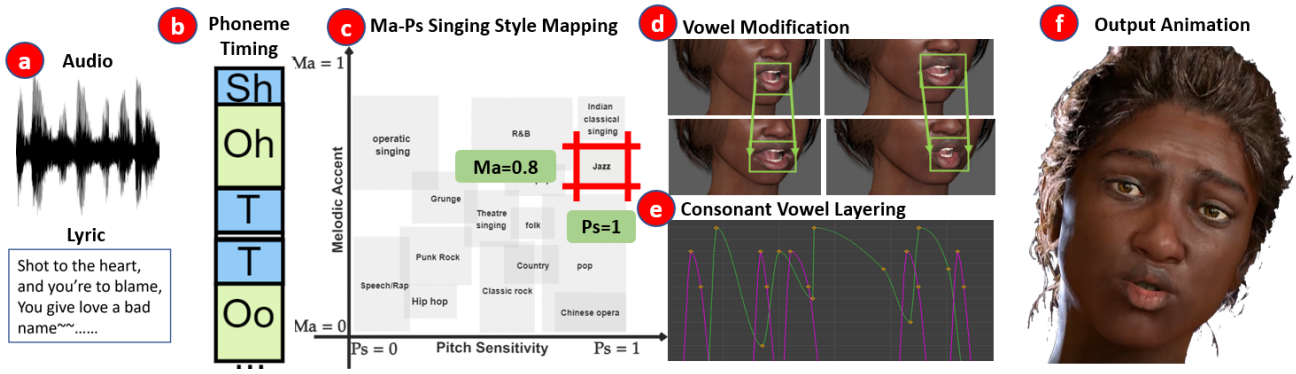

Figure 1: VOCAL is a vowel-consonant layered approach to expressive singing animation: Input audio and lyrics (a) are processed to produce a phonetic alignment (b). We define Melodic accent *Ma* and Pitch sensitivity *Ps* parameters, that can be configured to capture a range of singing styles (c). We detect and modify vowels that are sung differently to their transcription (d) and generate vowel animation curves that carry the melody, layered with consonant curves for lyrical clarity and rhythmic emphasis (e). Our output is an audio-driven, lower face animation (f).

ABSTRACT

Singing and speaking are two fundamental forms of human communication. From a modeling perspective however, speaking can be seen as a subset of singing. We present VOCAL, a system that automatically generates expressive, animator-centric lower face animation from singing audio input. Articulatory phonetics and voice instruction ascribe additional roles to vowels (projecting melody and volume) and consonants (lyrical clarity and rhythmic emphasis) in song. Our approach directly uses these insights to define axes for Melodic-accent and Pitch-sensitivity (*Ma-Ps*), which together provide an abstract space to visually represent various singing styles. In our system. vowels are processed first. A lyrical vowel is often sung tonally as one or more different vowels. We perform any such vowel modifications using a neural network trained on input audio. These vowels are then dilated from their spoken behaviour to bleed into each other based on Melodic-accent (*Ma*), with Pitch-sensitivity (*Ps*) modeling visual vibrato. Consonant animation curves are then layered in, with viseme intensity modeling rhythmic emphasis (inverse to *Ma*). Our evaluation is fourfold: we show the impact of our design parameters; we compare our results to ground truth and prior art; we present compelling results on a variety of voices and singing styles; and we validate these results with professional singers and animators.

CCS CONCEPTS

• **Computing methodologies** → **Procedural animation**;

KEYWORDS

facial animation, lip-sync, music, singing

ACM Reference Format:
Yifang Pan, Chris Landreth, Eugene Fiume, and Karan Singh. 2022. VOCAL: Vowel and Consonant Layering for Expressive Animator-Centric Singing Animation. In *SIGGRAPH Asia 2022 Conference Papers (SA '22 Conference Papers), December 6–9, 2022, Daegu, Republic of Korea.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3550469.3555408

SA '22 Conference Papers, December 6–9, 2022, Daegu, Republic of Korea

ACM ISBN 978-1-4503-9470-3/22/12…$15.00
https://doi.org/10.1145/3550469.3555408

1 INTRODUCTION

The recent explosion of interest in digital avatars and 3D facial animation has redoubled the need for research on representations and synthesis of all forms of expressive facial communication. Singing, as much as speaking, is a primeval and essential form of human communication. Singing characters appear in most animated films, from early Disney content, to blockbuster films like *Shrek*, *Frozen*, and *Coco*. Recent work on audio-driven 3D facial animation [Edwards

et al. 2016, 2020] has shown the disruptive potential of transforming vocal performances into visual performances. However, lip-sync timing and mouth shapes, or visemes, designed for speech visualization are usually ill-suited to singing, particularly when the performance moves from spoken lyrics to dramatic singing styles.

An inherent reason for this failing is that all phonemes play largely the same role in speech audio: phonemes all contribute to the listener's comprehension of the language spoken. Fundamentally, in addition to any lyrical content, singing must communicate melody and rhythm. Bio-acoustically, the sustained open-mouthed sound of vowels are much better suited than consonants to carry the volume and pitch variation of a melody [Bozeman 2017]. Sung vowels are also often modified *aggiustamento* for tonal quality or sustained resonance. The craft of voice instruction explicitly teaches the principal importance of vowels in singing [Bozeman 2013].

The role of consonants in contrast is to preserve lyrical comprehension, and punctuate the melody, emphasizing beat and rhythm. We note, of course that singing can span a stylistic spectrum from spoken lyric *sprechgesang* and rap, to legato *bel canto*, vocalese, and ultimately to the consonant-free drone of an Indian classical raga (Figure 1). We represent this dominance of vowels in accentuating melody, using a dynamically varying *Ma* (Melodic-accent) parameter $\in [0, 1]$ that captures the continuum from regular speech (*Ma*=0) to a consonant-free transition of vowels into each other (*Ma*=1).

We also note that the communicative efficiency of speech lends itself to viseme animation curves that are monotonically represented using an attack, sustain and decay behavior. Beyond melodic pitch variations of sustained vowels, singing is often enriched with musical ornaments, the best known of which are vibrato and trills, which introduce small periodic variations of pitch around notes. Such ornaments, although often appearing as minute, transient facial motions, are important perceptually in co-relating the vocal and visual performance. We capture this aspect of singing using a *Ps* (Pitch-sensitivity) parameter $\in [0, 1]$ that varies from the monotonic rise and fall in intensity of vowels in regular speech, to the quivering vowel mouth animation of a strong and deep vibrato. Together with the established *Jaw* and *Lip* speech style parameters [Edwards et al. 2016], we thus induce a *Ma-Ps-Ja-Li* 4D space that encompasses a wide range of stylistic spoken and sung behaviors.

After a review of related work on audio-driven speech and computational singing (Section 2), we describe our *Ma-Ps* singing model to define an overall *Ma-Ps-Ja-Li* 4D representation of visual song. The viseme animation curves are then computed based on aligned lyrics and the 4D vocal space (Section 3). Vowels are processed first, modified as necessary from the lyric to the sung vowel using a neural model acting on input audio. The sung vowels are dilated in time to bleed into each other based on *Ma*, with periodic intensity variation based on *Ps*. Consonant curves are then layered in, with intensity weighted inversely to *Ma* (Section 4). We evaluate the impact of our algorithmic parameters, compare against prior art and ground truth on sung performance, and provide professional singer/animator critique for a gallery of VOCAL generated singing animations (accompanying Video and Section 5).

Our principal contribution is what we believe to be the first computational method to the visual representation of a wide range of singing styles in an animator-centric fashion. While we show compelling results of our approach on an animated lower face, we observe that paralingual expressive behavior of the upper face, head and neck, is more important for singing than it is for speech, in terms of its correlation to beat/rhythm and emotion in the song. We conclude with a discussion of exciting directions for future work in animating a singing face (Section 6).

2 RELATED WORK

Visual singing pertains to lip-sync animation, broadly divided into 3 categories: performance-capture, data-driven, and procedural.

Performance-capture. Performance-capture methods map captured facial motion data of human actors to a digital facial model [Williams 1990], yielding natural and high-quality visual speech. Though earlier works employing this approach often required the use of physical markers [Blanz et al. 2003; Guenter et al. 1998], with advances in camera technology, motion-capture equipment, and 3D reconstruction algorithms, performance-capture approaches have become much more accessible. Through the use of stereo and depth cameras, high-fidelity motion capture can be done without markers [Bradley et al. 2010; Weise et al. 2009]. Using deep learning to learn a shape-prior also gave rise to approaches that made use of mono-cameras [Hu 2017; Olszewski et al. 2016]. Due to demand for high-quality performance in the entertainment industry, this approach has been applied widely. Products such as Faceware are often used in film production, and interactive character animation systems such as the Adobe Character Animator and Vroid Studio enable anyone to create a speaking avatar in real-time. However, the disadvantage of this approach is that the quality of performance depends on the ability of the actor, and the rigidity of captured motion data often removes creative control from an animator who may wish to edit or tune the animation.

Data Driven. Data-driven methods make use of large motion datasets to generate animation based on input speech audio. Prior to deep learning, most data-driven methods produced animation by performing a search within a corpus of visual speech clips, often minimizing a cost function that traded off similarity between candidate clips and phonemic context and smoothness [Bregler et al. 1997; Cao et al. 2005; Cosatto and Graf 2000]. Active Appearance models (AAM) [Anderson et al. 2013] and Hidden Markov Models (HMM) [Wang et al. 2012] can be employed to model speech dynamics and improve effectiveness of search.

The advance of deep learning propelled the development of high-quality generative models. By viewing visual speech generation as a sequence-to-sequence mapping problem, neural networks have been used to generate visual speech in both 2D [Suwajanakorn et al. 2017; Thies et al. 2020; Vougioukas et al. 2018; Zakharov et al. 2019; Zhou et al. 2019, 2020] and 3D [Cudeiro et al. 2019; Fan et al. 2022; Karras et al. 2017; Liu et al. 2015; Richard et al. 2021; Taylor et al. 2017] media. While these data-driven methods can produce plausible human speech animation, a fundamental difference between them and systems like ours (or JALI [2016]), is that we produce compact, animator-centric, animation curves, with meaningful parameters to edit animations in space, time and style (see Video 3:16).

Procedural. Procedural systems segment speech audio into a series of phonemes, then use look-up-tables, rules, models, or simulation to determine visemes (mouth configurations) for each phoneme,

which are then keyframed and interpolated into animation[Fisher 1968]. However, visemes alone do not produce realistic visual speech, as sequences of visemes are often co-articulated by human speakers. Various procedural methods can be classified by how they model co-articulation. Dominance models determine viseme activation based on dominance functions of adjacent phonemes [Cosi et al. 2002; King and Parent 2005; Massaro et al. 2001]. Bigram and trigram models use hand-crafted transitions to model short sequences of visemes together at run-time [Neumann et al. 2006; Xu et al. 2013]. Rule-based systems use explicit, often extensible co-articulation rules to determine how adjacent viseme are activated together [Bevacqua and Pelachaud 2004; Edwards et al. 2016; Wang et al. 2007]. Although procedural systems had lost favor due to advances in deep learning algorithms, they are lightweight, explainable, extendable, configurable using deep learning [Zhou et al. 2018], and generate compact motion curves that animators can easily refine. We thus develop a procedural representation for visual singing as a physiological manifestation of the acoustic signal.

Visual Singing. Comparing to visual speech synthesis, there is a much smaller corpus of work on the topic of visual singing. King and Parent applied their procedural speech model to generate visual singing [King and Parent 2004], but they found that the viseme model that worked well for speech vowels fell short when animating the much longer and more expressive singing vowels. More recently, [Kim and Park 2020] uses a two blendshape system (mouth open/closed), and uses the total spectrum energy of the audio to control the mouth opening/closing to generate low fidelity animation, and [Iwase et al. 2020; Yu et al. 2019] use deep learning to generate animation. Of relevance to singing is also simulation research on the modeling of breath [Zordan et al. 2004] and audio-driven simulation of laughter [DiLorenzo et al. 2008].

Our work instead, is based on visual singing insights from articulatory phonetics [Gick et al. 2012], singing pedagogy [Bozeman 2013], and physiological research on the relation between mouth configuration and acoustic qualities [Austin 2007; Lindblom and Sundberg 1971; Sundberg 1970; Titze 2011]. These insights form the basis of an animator-friendly visual singing system VOCAL, that significantly outperforms prior art on visual speech [Edwards et al. 2016; Fan et al. 2022], and singing [Iwase et al. 2020] (we only compare against animator-centric or singing-focused systems).

3 MAPS MODEL DESIGN

Based on empirical observation, literature review, and insights from singing coaches, we conclude that the credible visual depiction of singing requires considering both the physiology of phonation, and style of performance. Physiologically, our framework introduces vowel modification and larynx movements to reflect timbre and pitch changes. Stylistically, we build on the Jaw and Lip *JaLi* parameterization proposed for speech [Edwards et al. 2016]. We propose an additional *MaPs* field: two independent axes that embed various singing styles, and provide animators with stylistically meaningful control. Lastly, we propose a layering of *Ma* modulated consonants over the vowel dominant animation curves.

Note that while traditional visemes have a fixed spatio-temporal mouth shape [Taylor et al. 2012], JALI [2016] visemes have a *Ja*w and *Li*p parameterized, and contextually varying mouth shape. We further decouple and layer the spatio-temporal contribution of vowels and consonants with *Ma* and *Ps* parameters, to better handle the complex co-articulations and vibrato of singing.

3.1 Physiological Considerations

The acoustic quality of a voice is affected by the configuration of the larynx and upper vocal tract. The larynx affects the vibration frequency of the vocal folds, perceived as *pitch*; the jaw, tongue, pharynx (throat), and lips affects sound resonance, perceived as *timbre* [Bozeman 2013]. Animated realism declines if these visible physiological structures remain static during changes in voice acoustics. We thus introduce larynx movement and vowel modification to reflect pitch and timbre change, respectively.

Larynx Movement. The larynx is an internal structure on most rigs, visible only as a protrusion that moves up and down (superior-anterior) on the neck. Raising the larynx decreases the length of the upper vocal tract, increasing the frequency of formants and perceived vocal brightness [Bozeman 2013]. In practice, singers often use the larynx to sing at a higher melodic pitch. We thus raise the larynx when vowels are phonated, with the amplitude of movement determined by pitch.

Vowel Modification. When phonating vowels, singers often adjust the timbre of vowels for melody, resonance or artistic effect, known as "aggiustamento" or *vowel modification* [Bozeman 2013]. For example, Whitney Houston sings an iconic "I" from the chorus of "I will always love you" as a triphthong: she starts singing the "I" with /ai/, transitions to /i/ then back to /a/ (see accompanying Video). Since resonance is largely determined by mouth shape, if the vowel modification were not reflected by the lip-sync, the animation would lose realism. We build our vowel modification based on the five pure Italian vowels (A, E, I, O, and U) commonly used in vocal exercises. Each vowel has a distinct timbre and can only be produced by certain jaw and lip configurations. In our framework, we propose using a neural network to identify these vowels from the audio signal and modify the lyrical vowel with the one(s) sung.

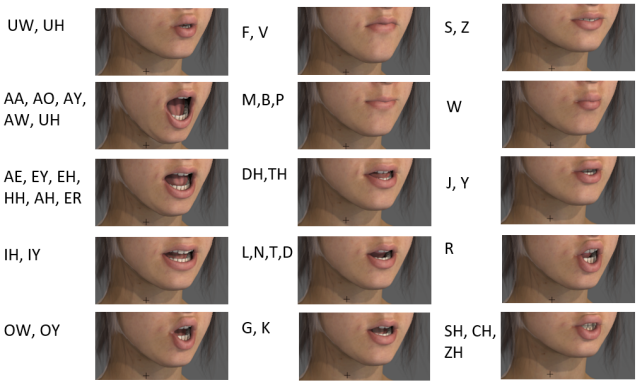

Figure 2: Table mapping phonemes (CMU notation) to visemes used in VoCAL.

3.2 MaPs Field

Traditional frameworks for procedural lip-sync animation are based on mapping phonemes to visemes. For example, the phoneme /æ/ would be mapped to the viseme /Eh/, which looks like opened jaw with stretched lips (see Figure 2 for a full list of mappings). With a

suite of hand-crafted blendshapes that resembles the visemes, animations can be created by key-framing and interpolating between the blendshapes over time.

However, the traditional methods cannot account for the various styles of phonation present in singing, where sung vowels and consonants play different roles: consonants emphasize rhythm, and vowels carry the melody. In our study, we discovered that a considerable proportion of singing-style variation can be mapped to a spectrum of different contribution of vowels and consonants. For example, rap songs that are rhythm-heavy would have clearly articulated consonants and speech-like vowels, while operatic bel-canto and Indian classical singers barely articulate consonants, focusing on melodic vowel transitions. We model the spectrum of singing styles using *Melodic accent* (Ma), and *Pitch sensitivity*(Ps). A rough *Ma-Ps* illustration of singing styles is shown in Figure 1(c).

Melodic-Accent. denotes the importance of melody relative to the phonetic parity of spoken lyrics, and the continuum between separately sung notes (staccato) and notes tied together (legato). Since vowels predominantly convey melody in song, increasing melodic accent shifts the visual performance from speech-like to one with diminished consonants and greater co-articulation between adjacent vowel visemes, where the mouth remains somewhat open in continuous phonation.

Pitch Sensitivity. Sung vowels can be phonated as either syllabic or melismatic. In syllabic singing (also known as speech singing), each vowel carries only one note, while in melismatic singing, rapid or considerable pitch changes can occur during the phonation of the same vowel. Though pitch change is largely an internal process, it may manifest externally in different ways. Amateur and pop singers habitually open their jaws wider to reach higher notes, and tremble their jaws to sing a vibrato [Bozeman 2013]. On the other hand, trained opera singers can perform pitch change and vibrato with minimal mouth movement. Our model parameterizes this using the notion of pitch sensitivity.

3.3 Vowel and Consonant Animation Layering (VOCAL)

In speech, the ratio of vowel-to-consonant duration is roughly 5:1. For singing, this ratio can rise to 200:1 [Nix 2015]. This can be attributed to the biomechanics of consonant phonation. While vowels are produced by vocal cord vibration actuated by constant airflow, consonants are produced by narrowing parts of the upper vocal tract and disrupting the airflow [Bozeman 2013]. To establish a more stable melody, singers may sacrifice the intelligibility of consonants. This layering of consonants over vowels in song closely parallels the instruction methodology of many vocal coaches [Tamplin 2016].

Rather than treating vowels and consonants as being in the same class of visemes, we consider vowel visemes as having both jaw and lip contributions, while most consonants only contribute to lip motion, with the exception of sibilants and labial dental consonants. In this formulation, since consonants occur at the boundary of vowels, the corresponding jaw movement is completely determined by the co-articulation of vowels. As a result, the *Ma* parameter also determines the apparent degree of consonant contribution. With a low value of *Ma*, consonants at a boundary would have a higher contribution, as they temporally overlap with narrowing the jaw between vowel visemes. Conversely, a high value for *Ma* would reduce the perceived contribution of the consonant.

3.4 Control Rig

For our prototype, we use the commercially available JALI Village facial rig from JALI Research [Edwards et al. 2016]. The rig is equipped with a suite of JALI visemes shown in Figure 2, as well as a set of action units (AU) from the Facial Action Coding System (FACS) [Ekman and Rosenberg 1997], each with a blend weight $\alpha \in [0, 1]$. The JALI visemes are parameterized with *Ja* and *Li* parameters, which we use to control viseme enunciation. For vowel modification, we additionally use the Action Units for lip-rounding and lip-stretching.

In VOCAL, *Ja-Li* parameters manipulate the spatial appearance of visemes (tongue-jaw configurations and lip shape) and *Ma-Ps* modulate the temporal behavior of the visemes (the extent and shape of the viseme animation curves). While *Ja-Li* parameters typically suffice for speech, *Ma-Ps* are essential to represent the different roles of vowels and consonants and temporal dynamics in singing. The *Ja-Li-Ma-Ps* parameters can be independently controlled by an animator to edit singing style, with *Ma=0,Ps=0* producing JALI speech animation.

4 VOCAL ALGORITHM

Our system creates lip-sync animation from audio in a two-phase process: tagging, and animation curve generation. In phase one, we perform forced alignment to temporally align phonemes to audio, and then tag the audio to identify intervals of constant pitch and vibrato to be used in subsequent phases. In phase two, we use the phoneme timing to generate four sets of viseme curves with different singing styles, as well as the animation curves for vowel modification and larynx movements. The animation curves are then blended into a final output using the MaPs Field, and visualized using Autodesk *Maya*.

4.1 Audio Tagging

Our tagging system requires a roughly acapella audio as input, which can be obtained by using a free vocal isolation tool such as *vocalremover.org*.

Phoneme Alignment. The first step of tagging is to generate timings for all phonemes present in the song, which we automate using forced alignment. This process employs a trained language model that maps input audio and a phonetic transcript, to phoneme timings [Schulze-Forster et al. 2021]. Phonetic transcripts can be automatically generated from song lyrics, using a pronunciation dictionary. We use the CMU LOGIOS Lexicon Tool [Boersma and Weenink 2001].

We also detect other acoustic events apart from phonemes, including intervals of vibrato and constant pitch. To detect these events, we make use of pitch estimation $f_0(t)$, which we obtain using Praat [Boersma and Weenink 2001]. To prevent the high frequency consonants from skewing the pitch estimate, we only perform this computation during the phonation of vowels.

Vibrato. With proper technique, the periodic pitch variations of a vibrato need not be visibly manifest on the face. However, in

accordance with the principle of exaggeration for animation, we found it necessary to include some physical movement in the jaw and neck to avoid the character looking static when holding a long vibrato. To detect vibrato, we find intervals with periodic oscillation in the pitch signal $f_0(t)$. We use finite differences to compute $f_0'(t)$, from which we obtain a list of zero-crossing points representing peaks in $f_0(t)$. We then iterate over the zero-crossing points to determine intervals at which the points are a similar distance apart. To filter out noise we uses the following constraints.

(1) The vibrato interval must have more than one period.
(2) The standard deviation must be less than 1 semi-tone.
(3) The vibrato period $\in [1/5, 1/8]$ seconds [Pecoraro et al. 2013].

Constant pitch intervals. To model melismatic singing, for each vowel phoneme, we also consider the pitch signal $f_0(t)$, where $t \in [t_0, t_N]$. Since raw pitch is too noisy to be used directly, we fit a series of linear segments to approximate the pitch signal. We view the relatively flat segments as notes, and steeper segments as note transitions. We perform piecewise-linear fitting for an interval of $f_0(t)$ with N points, using a dynamic programming approach to jointly minimize the number of linear segments and overall fitting error [McCrae and Singh 2009]. We populate a matrix $M \in \Re^{N \times N}$ (N =# pitch samples at 100fps in a vowel), as follows:

$$M(a,\ b)\ =\ \min_{a<x<b}\{M(a,k)\ +\ M(k,b),\ E_{fit}(a,b)\ +\ E_{cost}\}. \quad (1)$$

Here $M(a,b)$ denotes the minimal cost of connecting a to b using a series of linear segments. Since $a < b$, M is strictly upper triangular. E_{cost} is a constant penalty of adding additional line segments, which we empirically set as $E_{cost} = \frac{(f_{min}+f_{max})}{2}$, where f_{min}, f_{max} are minimum and maximum pitch within the interval. $E_{fit}(\cdot,\cdot)$ denotes the fitting error, which can be computed as follows:

$$E_{fit}(a,b)\ =\ \sum_{t=a}^{b} |f_0(t) - (slope_t \times t\ +\ yint_t)|, \quad (2)$$

where $slope_t$ and $yint_t$ are slope and y-intercept of the t^{th} interval, respectively. A bottom-up computation, yields a series of connected linear segments approximating the pitch signal, denoted as $f_{lin}(t)$. The segments in $f_{lin}(t)$ with a slope less than a threshold (empirically set as 50Hz) are considered as constant notes.

4.2 Animation Curve Generation

After obtaining the transcript with phonetic and acoustic features, the next step is to generate curves to drive the visemes and other FACS action units. The viseme curves and larynx motion curves are generated first, as the timing for vibrato and vowel modification are contingent on the timing of the viseme activation.

Viseme curves. We use sparse keyframes to sequentially activate visemes based on phonemic timing, acoustic information, and co-articulation rules [Edwards et al. 2016]. The profile of each viseme is specified by two types of keyframes: boundary keys and internal keys. The boundary keys demarcate lip movement before and after the phonation of the viseme, and internal keys control the lip movement during phonation (Figure 3). *Ma* and *Ps* values further control boundary and internal keyframes, as shown in Figure 3. Our viseme animation curves are a bilinear interpolation of four viseme curves generated with extreme values $\{0, 1\}$ for *Ma* and *Ps*. Note that no melodic accent and no pitch sensitivity (*Ma=Ps=0*), produces JALI lip-sync speech animation. We thus refer to the *Ma-Ps* extremes as speech and singing curves, respectively.

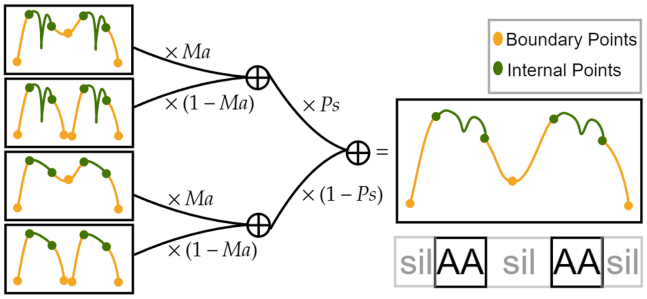

Figure 3: Viseme animation curves combine four curves based *Ma-Ps* values.

Generating a speech curve. We generate a speech curve in three passes. The first generates a four-key viseme curve for each phoneme, the second enforces vowel-consonant co-articulation rules, and the final pass corrects conflicting keyframes.

Pass 1 For each phoneme in the transcript, the viseme is selected as per the look-up table in Figure 2. The boundary frames for each viseme are timed 120ms before and after the phonation interval to reflect general speech physiology [Bailly 1997][Ito et al. 2004], and the internal frames are selected to reflect how the viseme would apex at the beginning at phonation and sustain until 75% of the sound is completed. The amplitude for the frame at the apex is chosen depends on the length and visemes types.
(1) Jaw and Lip closer consonants (B, P, M, F, V, S, SH) are fully articulated ($\alpha = 1$).
(2) Other consonants and short vowels (duration < 200ms) are articulated less prominently ($\alpha = 0.6$).
(3) Longer vowels (duration > 200ms) are more articulated ($\alpha = 0.8$).

Pass 2 Activating visemes in sequence is robotic and unrealistic. Indeed, it is important to consider co-articulation between neighboring phonemes. We use JALI's vowel-consonant co-articulation rules [Edwards et al. 2016].

Pass 3 Co-articulated, repeating vowels can minimally overlap at phonation boundaries, resulting in keyframes that are undesirably interpolated (Figure 4(left)). We resolve the conflict by combining (co-articulating) the two visemes as shown in Figure 4(right), replacing the overlapping keyframes by a single keyframe, inserted 120ms ahead of second viseme onset if possible, and mid-way between the two visemes otherwise. The amplitude of the new keyframe is chosen to reflect both the decay of the first viseme and the onset of the second viseme, controlled by a user-defined vowel co-articulation (VC) parameter (default $VC = 0.5$ for speech curves).

Generating a singing curve. The first of three passes, generates viseme curves with additional internal keys defined for different notes. The second pass first enforces consonant co-articulation,

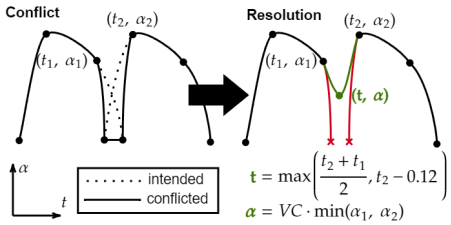

Figure 4: Co-articulated, repeating visemes can overlap (left). The conflicting keyframes are resolved by combining the viseme curves (right).

and subsequently modifies vowel boundary keys to reflect vowel-vowel co-articulation. The final pass resolves conflicting keyframes.

Pass 1 For singing, consonant and vowel motion curves are generated separately. For each consonant, the four-key speech curve is re-used. For vowels, we utilize the notes detected from audio tagging (Section 4.1). If a vowel is syllabic (containing only one note), the viseme would apex when the note is reached, and decay as a speech vowel. For a melismatic vowel (containing multiple notes), we found that the viseme would often apex multiple times during phonation, with the apex coinciding with the start of each note and mildly decaying as the note finishes. To reflect these, we set internal keyframes at the start and end of each note, where the amplitude of the starting key depends on the pitch of the note: $\alpha_s = 0.4 \times (f_0(t_{start}) - f_{0,min})/(f_{0,max} - f_{0,min}) + 0.6$ and the amplitude of the end keyframe decays from the starting frame $\alpha_e = 0.95\alpha_s$. Last, to emphasize the transition between each note, we set an internal keyframe between each note at time $t = 0.5 * (t^{prev}_{end} + t^{next}_{start})$, with amplitude $\alpha = 0.9 * \min(\alpha^{prev}_{end}, \alpha^{next}_{start})$. Vibrato is animated over the detected interval (Section 4.1) by setting keys to oscillate the given *Ja* setting at 7Hz, with increasing amplitude upto ±0.6, as per an average vibrato [Pecoraro et al. 2013]

Pass 2 Similar to the speech curve, pass 2 enforces co-articulation. First, we use the JALI rules to ensure proper vowel-consonant co-articulation. Then, to model a strong melodic accent for vowels, we make closely spaced vowel visemes (phonation intervals < 300ms apart) blend into each other by extending the boundary keys of both visemes. Note that any consonants between such vowel visemes, would have little visual contribution.

Pass 3 Vowel-vowel and vowel-consonant co-articulation can introduce conflicting keyframes in the viseme and larynx motion curves. These are resolved as shown in the previous section, with $VC = 0.95$ to reflect greater melodic accent.

Given the speech curve ($Ma = 0$, $Ps = 0$) and singing curve ($Ma = 1$, $Ps = 1$), extreme curves ($Ma = 1$, $Ps = 0$) and ($Ma = 0$, $Ps = 1$) are generated by a mix-and-match of internal and boundary keyframes.

Computing Ma-Ps *values from audio.* While *Ma-Ps* values can be user-controlled for song-style, or learnt from a corpus of captured songs, we propose a psycho-acoustic heuristic to compute *Ma* and *Ps* from input audio. Strongly articulated fricative (S/Z/F/V/S/Sh/D/T) or plosive (P/B/D/T/G/K) consonants, produce turbulent airflow with localized high frequency (HF=8-20kHz) energy [Edwards et al. 2016]. We use the consonant's spectral HF energy ϵ, relative to the HF energy of consonants for the entire song, to determine *Ma*. *Ma*= 0.2 if $\epsilon \leq$ *mean-stdev*; *Ma*= 0.8 if $\epsilon \geq$ *mean+stdev*; else *Ma*= 0.5.

Pitch variation is common, during sustained vowels, where static lips seem unnatural. We use the duration of a vowel v, relative to the avg. length of a spoken vowel ($\tau = 0.2$s [Kuwabara 1996]) to determine *Ps*. We set *Ps*= 0.1 for $v \leq \tau$, else *Ps*= $\min(1, 0.1 + v - \tau)$.

While singing style can vary as frequently as every phoneme, simple neighboring averaging can produce a smoother *Ma-Ps* signal.

4.3 Vowel Modification

We detect timbre changes with a neural network to make viseme modifications for differently transcribed and sung vowels.

Vowel Modification Detection Network. The neural network maps an input audio feature vector, to a probability distribution for each of the five Italian vowels (and silence) at each timestamp. Our architecture, inspired by Visemenet [2018], consists of three LSTM layers, a ReLu activation layer, a single fully connected layer, and uses a *softmax* function to produce prediction probabilities.

Audio feature vector. Our feature vector is constructed in the same way as Visemenet, comprising 13 Mel Frequency Cepstral Coefficients (MFCCs), 26 raw Mel Filter Bank (MFB), and 26 Spectral Subband Centroid features, extracted every 10ms, with window size 25ms. We increase the network's receptive field by concatenat-captured iog sme with 12 prabsequent frames.

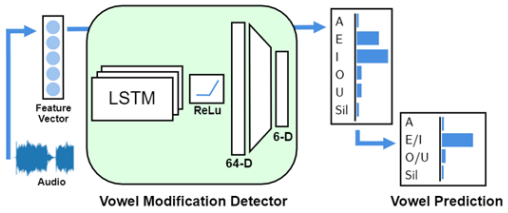

Figure 5: Vowel modification predicts vowel probabilities from input audio.

Training Data: Our network is trained on the VocalSet corpus [2018], comprising 10.1 hours of singing performed by 20 professional singers in a variety of styles for each of the five Italian vowels. Each audio file is labeled by the vowel and singing style used for that clip. For training, the audio tracks are split into 4-second clips, with each timestamp labeled by either the corresponding vowel of that clip or silence. The clips of 4 singers are reserved as the test set.

Network Training: The network is trained to minimize cross-entropy loss at each timestamp with the Adam optimizer [2017], with an initial learning rate of 0.0001 and a batch size of 512. Since the training data does not include transitions between vowels, we augmented training data by concatenated multiple clips. The model is trained on a Titan RTX 24GB GPU Card for five hours before being terminated by the early stop mechanism. Our model achieves a test accuracy of 70%. The error arises from confusion between lip-spreaders "E", "I" and between lip-rounders "O", "U". For this reason, we merge easily confused vowel predictions as shown in Figure 5, to achieve a test accuracy of 91%.

Vowel Modification Curve Generation. For each vowel, the neural network is used to detect the likely sung vowel(s) from the audio. We avoid excessive modification by only modifying vowels with a prediction probability > 60% threshold. *lipSpread* and *lipRound* AUs are modulated (+/-) to modify these vowels as follows:

	NN Predicted vowel		
Transcript vowel	A	E or I	O or U
A	Nothing	+lipSpread	+lipRound
E or I	-lipSpread	Nothing	-lipSpread +lipRound
O or U	-lipRound	-lipRound +lipSpread	Nothing

We generate four-key motion curves (like JALI visemes), to modulate the desired (+/-) expression change. The apex amplitude is based on the prediction probability and the maximum amplitude of the AU ($\alpha = P(prediction) \cdot \alpha_{max}$). Lastly, the AU motion curves are co-articulated, and start and/or end just before the transcribed vowel.

5 EVALUATION

The results of our four-fold evaluation of VOCAL are best viewed in the accompanying Video. Note that VOCAL automatically generates lower face animation. Any head and upper face motion was keyframed or performance capture mimed using Faceware [2017].

(1) We show the impact of different design aspects of VOCAL on the output animation on an example song (Video 0:55-4:03).
(2) We compare VOCAL qualitatively, to prior art in visual singing and speech (Video 4:43-5:51; 13:41-16:43).
(3) We provide a quantitative comparison of both JALI and VOCAL to a performance captured ground truth (Video 5:59-6:43; 10:41-13:41).
(4) We present a variety of singing style clips, automatically generated with VOCAL (Video 7:21-9:20), along with professional critique.

5.1 Prior Art Comparison

We used the audio, from the Song2Face [2020] results on "Hey Jude", to generate singing animation using VOCAL, and two speech solutions, Faceformer [2022] and JALI [2016]. Figure 6 summarizes our comparison (Video 4:43-5:51). Song2face is consistently unable to produce plausible animation, a common weakness of deep-learning approaches that lack understanding of acoustics and human anatomy. Speech models Faceformer and JALI enunciate consonants well, but fail on vowels. Shorter vowels tend to over-articulate, robotically opening/closing the jaw completely for each vowel. Sustained vowels seem inexpressively monotonic, failing to show pitch change and vibrato. VOCAL's weakness in this clip is the inability to animate utterances, like heavy breathing, not tagged in the transcript.

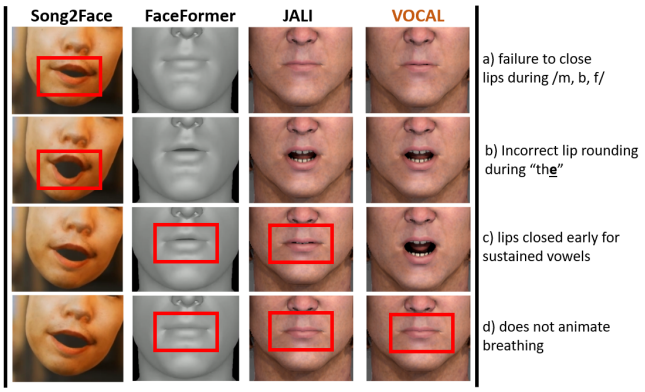

Figure 6: Failure cases for song2face, FaceFormer, JALI and our system.

5.2 Ground Truth Comparison

Physiologically, the mapping between sound production and facial appearance is not unique, especially when singing vowels (it is easy to sustain an "Ee" vowel while changing expression). Quantitative error of an animated face relative to a ground truth vocal performance alone, can thus be misleading. We do however, show that VOCAL has a lower cumulative error in both vertex position and velocity than JALI, in a quantitative comparison to ground truth Faceware captured vocal performances (see clip in Video 5:59-6:43 and 10:41-13:41).

5.3 Results and Animation Critique

We present the automated results of VOCAL on a range of singing clips (Video 7:21-9:20). We solicited feedback on these 6 clips from 4 professionals (2 voice instructors, 2 animators). We specifically asked them to focus on the lower face. The overall feedback on the visual performances were overwhelmingly positive with highlights being: "...the vibrato is very effective to my eyes" (ella); "...vowel adjustments on 'I' look good" (whitney); "...tonal variation during d'ee'p is very convincing" (adele); "...consonants are well done and not over-articulated" (james). On the critical front: "...cheeks are too relaxed especially during the 'A-I-A'" (whitney); "...'could' lips move improbably fast. 'O' of rolling seems natural but underemphasized" (adele); "...general feel of being slightly out of sync, unlike the other examples" (james);"...'no river' seems to lack effort" (rush).

We also ran a 31 lay-person, forced choice preference study between JALI and VOCAL animated output (presented randomly), for 10 clips (Video 13:41-16:43). Viewers strongly preferred VOCAL (> 70% votes) for 6/10 clips (Figure 8). The remaining 4 clips were speech-like (low *Ma-Ps* values), visually very similar, and received a mixed preference (3/4 in favour of VOCAL).

Like JALI, VOCAL generates a sparse set of keyframes, and user friendly parameters, designed for animators to expressively manipulate (Video 9:25-10:40).

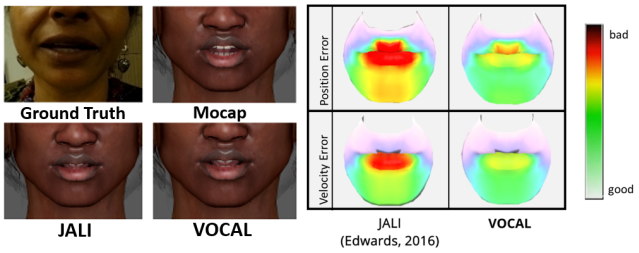

Figure 7: Faceware captured ground truth compared to JALI a nd VOCAL.

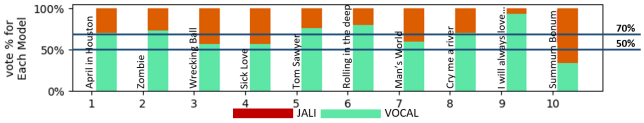

Figure 8: Recorded user preference of 10 clips (Video 13:41-16:43)

6 LIMITATIONS, FUTURE WORK AND CONCLUSION

Phonetic alignment between transcript and audio, is significantly more challenging for singing than for speech, and even our singing-trained aligner [Schulze-Forster et al. 2021], can show errors over long or musically distorted clips (Video 17:01-19:51). While alignment in VOCAL can be manually fixed if needed, it highlights the need for better phonetic alignment models for singing.

Vowel modifications in VOCAL are determined by a threshold value on predicted vowel probability (empirically set to 0.6), which in some cases could produce incorrect vowel predictions (Video 19:51-20:50).

Our biggest limitation relates to the visual expression of effort and breath in singing (Figure 6). Conspicuously missing for example on (james) are sinews and skin tension that evidence muscle effort.

Singing clearly involves more than the motion of the lower face. A complete singing face is rich in emotional and rhythmic paralingual motion of the upper face (eye and brows), head, and neck. While we have demonstrated an audio-driven model for animating the lower face (relying on performance capture/keyframing for the rest), an exciting avenue for future work is the prediction of eyebrows, gaze, blink, and head and neck movements to emotionally and rhythmically accompany the lower face in song.

In summary, VOCAL is a novel visual-singing animation approach that models different singing styles by modifying the contribution of vowel and consonants with a *Ma-Ps* field. To ensure physiological plausibility of sung performance, we also present the use pitch-dependent vowel profiles and vowel modification. Our model captures singing style and produces animator-editable output that is bio-acoustically plausible. We hope our insights on singing will positively impact other modalities of vocal communication, and inspire new directions in expressive facial animation.

REFERENCES

Robert Anderson, Bjorn Stenger, Vincent Wan, and Roberto Cipolla. 2013. *Expressive Visual Text-To-Speech Using Active Appearance Models.* https://doi.org/10.1109/CVPR.2013.434 Journal Abbreviation: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Publication Title: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition.

Stephen F. Austin. 2007. Jaw Opening in Novice and Experienced Classically Trained Singers. *Journal of Voice* 21, 1 (Jan. 2007), 72–79. https://doi.org/10.1016/j.jvoice.2005.08.013

Gérard Bailly. 1997. Learning to speak. Sensori-motor control of speech movements. *Speech Communication* 22, 2 (Aug. 1997), 251–267. https://doi.org/10.1016/S0167-6393(97)00025-3

Elisabetta Bevacqua and Catherine Pelachaud. 2004. Expressive audio-visual speech. *Journal of Visualization and Computer Animation* 15 (July 2004), 297–304. https://doi.org/10.1002/cav.32

V. Blanz, C. Basso, T. Poggio, and T. Vetter. 2003. Reanimating Faces in Images and Video. *Computer Graphics Forum* 22, 3 (2003), 641–650. https://doi.org/10.1111/1467-8659.t01-1-00712 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1467-8659.t01-1-00712.

Paul Boersma and David Weenink. 2001. Praat: doing Phonetics by Computer. (2001). https://www.fon.hum.uva.nl/praat/

Kenneth Bozeman. 2017. *Kinesthetic Voice Pedagogy 2: Motivating Acoustic Efficiency.* Inside View Press. Google-Books-ID: rzopzgEACAAJ.

Kenneth W Bozeman. 2013. Practical Vocal Acoustics. (2013), 162.

Derek Bradley, Wolfgang Heidrich, Tiberiu Popa, and Alla Sheffer. 2010. High Resolution Passive Facial Performance Capture. (2010), 10.

Christoph Bregler, Michele Covell, and Malcolm Slaney. 1997. Video Rewrite: driving visual speech with audio. In *Proceedings of the 24th annual conference on Computer graphics and interactive techniques - SIGGRAPH '97.* ACM Press, Not Known, 353–360. https://doi.org/10.1145/258734.258880

Yong Cao, Wen C. Tien, Petros Faloutsos, and Frédéric Pighin. 2005. Expressive speech-driven facial animation. *ACM Transactions on Graphics* 24, 4 (Oct. 2005), 1283–1302. https://doi.org/10.1145/1095878.1095881

E. Cosatto and H.P. Graf. 2000. Photo-realistic talking-heads from image samples. *IEEE Transactions on Multimedia* 2, 3 (Sept. 2000), 152–163. https://doi.org/10.1109/6046.865480 Conference Name: IEEE Transactions on Multimedia.

P. Cosi, E.M. Caldognetto, G. Perin, and C. Zmarich. 2002. Labial coarticulation modeling for realistic facial animation. In *Proceedings. Fourth IEEE International Conference on Multimodal Interfaces.* 505–510. https://doi.org/10.1109/ICMI.2002.1167047

Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael J. Black. 2019. Capture, Learning, and Synthesis of 3D Speaking Styles. *arXiv:1905.03079 [cs]* (May 2019). http://arxiv.org/abs/1905.03079 arXiv: 1905.03079.

Paul C DiLorenzo, Victor B Zordan, and Benjamin L Sanders. 2008. Laughing out loud: Control for modeling anatomically inspired laughter using audio. In *ACM SIGGRAPH Asia 2008 papers.* 1–8.

Pif Edwards, Chris Landreth, Eugene Fiume, and Karan Singh. 2016. JALI: an animator-centric viseme model for expressive lip synchronization. *ACM Transactions on Graphics* 35, 4 (July 2016), 1–11. https://doi.org/10.1145/2897824.2925984

Pif Edwards, Chris Landreth, Mateusz Popławski, Robert Malinowski, Sarah Watling, Eugene Fiume, and Karan Singh. 2020. JALI-Driven Expressive Facial Animation and Multilingual Speech in Cyberpunk 2077. In *ACM SIGGRAPH 2020 Talks (SIGGRAPH '20).* Association for Computing Machinery, New York, NY, USA, Article 60, 2 pages. https://doi.org/10.1145/3388767.3407339

Paul Ekman and Erika L. Rosenberg. 1997. *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS).* Oxford University Press. Google-Books-ID: KVmZKGZfmfEC.

Faceware. 2017. Analyzer. http://facewaretech.com/products/software/analyzer. (2017).

Yingruo Fan, Zhaojiang Lin, Jun Saito, Wenping Wang, and Taku Komura. 2022. *FaceFormer: Speech-Driven 3D Facial Animation with Transformers.* Technical Report arXiv:2112.05329. arXiv. http://arxiv.org/abs/2112.05329 arXiv:2112.05329 [cs] type: article.

Cletus G. Fisher. 1968. Confusions Among Visually Perceived Consonants. *Journal of Speech and Hearing Research* 11, 4 (Dec. 1968), 796–804. https://doi.org/10.1044/jshr.1104.796 Publisher: American Speech-Language-Hearing Association.

Bryan Gick, Ian Wilson, and Donald Derrick. 2012. *Articulatory Phonetics.* John Wiley & Sons. Google-Books-ID: rrfoJJKmIq4C.

Brian Guenter, Cindy Grimm, Daniel Wood, Henrique Malvar, and Fredric Pighin. 1998. Making faces. In *Proceedings of the 25th annual conference on Computer graphics and interactive techniques (SIGGRAPH '98).* Association for Computing Machinery, New York, NY, USA, 55–66. https://doi.org/10.1145/280814.280822

Liwen Hu. 2017. Avatar digitization from a single image for real-time rendering | ACM Transactions on Graphics. (2017). https://dl.acm.org/doi/10.1145/3130800.31310887

Takayuki Ito, Emi Murano, and Hiroaki Gomi. 2004. Fast force generation dynamics of human articulatory muscles. *Journal of applied physiology (Bethesda, Md. : 1985)* 96

(July 2004), 2318–24; discussion 2317. https://doi.org/10.1152/japplphysiol.01048.2003

Shohei Iwase, Takuya Kato, Shugo Yamaguchi, Tsuchiya Yukitaka, and Shigeo Morishima. 2020. Song2Face: Synthesizing Singing Facial Animation from Audio. In *SIGGRAPH Asia 2020 Technical Communications (SA '20)*. Association for Computing Machinery, New York, NY, USA, 1–4. https://doi.org/10.1145/3410700.3425435

Tero Karras, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. 2017. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Transactions on Graphics* 36, 4 (July 2017), 1–12. https://doi.org/10.1145/3072959.3073658

Namjung Kim and Kyoungju Park. 2020. Singing Lip Sync Animation System Using Audio Spectrum. In *Advances in Computer Science and Ubiquitous Computing*. Springer, Singapore, 135–140. https://doi.org/10.1007/978-981-13-9341-9_23

Scott A. King and Richard E. Parent. 2004. Animating song. *Computer Animation and Virtual Worlds* 15, 1 (2004), 53–61. https://doi.org/10.1002/cav.7 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cav.7.

Scott A. King and Richard E. Parent. 2005. Creating speech-synchronized animation. *IEEE transactions on visualization and computer graphics* 11, 3 (June 2005), 341–352. https://doi.org/10.1109/TVCG.2005.43

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]* (Jan. 2017). http://arxiv.org/abs/1412.6980 arXiv: 1412.6980.

H. Kuwabara. 1996. Acoustic properties of phonemes in continuous speech for different speaking rate. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, Vol. 4. 2435–2438 vol.4. https://doi.org/10.1109/ICSLP.1996.607301

B. E. Lindblom and J. E. Sundberg. 1971. Acoustical consequences of lip, tongue, jaw, and larynx movement. *The Journal of the Acoustical Society of America* 50, 4 (Oct. 1971), 1166–1179. https://doi.org/10.1121/1.1912750

Yilong Liu, Feng Xu, Jinxiang Chai, Xin Tong, Lijuan Wang, and Qiang Huo. 2015. Video-Audio Driven Real-Time Facial Animation. *ACM Trans. Graph.* 34, 6, Article 182 (oct 2015), 10 pages. https://doi.org/10.1145/2816795.2818122

D. W. Massaro, M. M. Cohen, R. Clark, M. Tabain, and Jonas Beskow. 2001. Animated speech : Research progress and applications. Cambridge University Press, 309–345. http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-167652

James McCrae and Karan Singh. 2009. Sketching piecewise clothoid curves. *Computers & Graphics* 33, 4 (Aug. 2009), 452–461. https://doi.org/10.1016/j.cag.2009.05.006

Ulrich Neumann, J.P. Lewis, Tae Kim, Murtaza Bulut, and Shrikanth Narayanan. 2006. Expressive Facial Animation Synthesis by Learning Speech Coarticulation and Expression Spaces. *IEEE Transactions on Visualization and Computer Graphics* 12 (Nov. 2006), 1523–1534. https://doi.org/10.1109/TVCG.2006.90

John Nix. 2015. Speaking vs Singing. (Sept. 2015). http://music.utsa.edu/pdfs/61_SpeakingvsSinging.pdf

Kyle Olszewski, Joseph J. Lim, Shunsuke Saito, and Hao Li. 2016. High-fidelity facial and speech animation for VR HMDs. *ACM Transactions on Graphics* 35, 6 (Nov. 2016), 1–14. https://doi.org/10.1145/2980179.2980252

Guilherme Pecoraro, Daniella Curcio, and Mara Behlau. 2013. Vibrato rate variability in three professional singing styles: Opera, Rock and Brazilian country. *The Journal of the Acoustical Society of America* 133 (May 2013), 3321. https://doi.org/10.1121/1.4805550

Alexander Richard, Michael Zollhoefer, Yandong Wen, Fernando de la Torre, and Yaser Sheikh. 2021. MeshTalk: 3D Face Animation from Speech using Cross-Modality Disentanglement. *arXiv:2104.08223 [cs]* (April 2021). http://arxiv.org/abs/2104.08223 arXiv: 2104.08223.

Kilian Schulze-Forster, Clement S. J. Doire, Gaël Richard, and Roland Badeau. 2021. Phoneme Level Lyrics Alignment and Text-Informed Singing Voice Separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), 2382–2395. https://doi.org/10.1109/TASLP.2021.3091817 Conference Name: IEEE/ACM Transactions on Audio, Speech, and Language Processing.

J. Sundberg. 1970. Formant Structure and Articulation of Spoken and Sung Vowels. *Folia Phoniatrica et Logopaedica* 22, 1 (1970), 28–48. https://doi.org/10.1159/000263365 Publisher: Karger Publishers.

Supasorn Suwajanakorn, Steven M. Seitz, and Ira Kemelmacher-Shlizerman. 2017. Synthesizing Obama: learning lip sync from audio. *ACM Transactions on Graphics* 36, 4 (July 2017), 1–13. https://doi.org/10.1145/3072959.3073640

Ken Tamplin. 2016. *How To Sing Any Song: Voice Lessons, Tamplin Vocal Academy.* $https://www.youtube.com/watch?v=ZATunybJm_4\&t=57s$.

Sarah Taylor, Taehwan Kim, Yisong Yue, Moshe Mahler, James Krahe, Anastasio Garcia Rodriguez, Jessica Hodgins, and Iain Matthews. 2017. A deep learning approach for generalized speech animation. *ACM Transactions on Graphics* 36, 4 (July 2017), 1–11. https://doi.org/10.1145/3072959.3073699

Sarah L. Taylor, Moshe Mahler, Barry-John Theobald, and Iain Matthews. 2012. Dynamic Units of Visual Speech. In *Proc. SCA*.

Justus Thies, Mohamed A. Elgharib, Ayush Tewari, C. Theobalt, and M. Nießner. 2020. Neural Voice Puppetry: Audio-driven Facial Reenactment. In *ECCV*. https://doi.org/10.1007/978-3-030-58517-4_42

Ingo Titze. 2011. Formant Frequency Shifts for Classical and Theater Belt Vowel Modification. (2011), 2.

Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. 2018. End-to-End Speech-Driven Facial Animation with Temporal GANs. *arXiv:1805.09313 [cs, eess]* (July 2018). http://arxiv.org/abs/1805.09313 arXiv: 1805.09313.

Alice Wang, Michael Emmi, and Petros Faloutsos. 2007. Assembling an expressive facial animation system. In *Proceedings of the 2007 ACM SIGGRAPH symposium on Video games - Sandbox '07*. ACM Press, San Diego, California, 21. https://doi.org/10.1145/1274940.1274947

Lijuan Wang, Wei Han, and Frank K. Soong. 2012. High Quality Lip-Sync Animation for 3d Photo-Realistic Talking Head. (2012).

Thibaut Weise, Hao Li, Luc Van Gool, and Mark Pauly. 2009. Face/Off: live facial puppetry. In *Proceedings of the 2009 ACM SIGGRAPH/Eurographics Symposium on Computer Animation - SCA '09*. ACM Press, New Orleans, Louisiana, 7. https://doi.org/10.1145/1599470.1599472

Julia Wilkins, Prem Seetharaman, Alison Wahl, and Bryan Pardo. 2018. Vocalset: A Singing Voice Dataset. (March 2018). https://doi.org/10.5281/ZENODO.1203819 Type: dataset.

Lance Williams. 1990. Performance-driven Facial Animation. In *Proc. SIGGRAPH*.

Yuyu Xu, Andrew W. Feng, Stacy Marsella, and Ari Shapiro. 2013. A Practical and Configurable Lip Sync Method for Games. In *Proceedings of Motion on Games*. ACM, Dublin 2 Ireland, 131–140. https://doi.org/10.1145/2522628.2522904

Jun Yu, Chang Wen Chen, and Zengfu Wang. 2019. 3D Singing Head for Music VR: Learning External and Internal Articulatory Synchronicity from Lyric, Audio and Notes *(MM '19)*. ACM, 945–952. https://doi.org/10.1145/3343031.3350865 Book Title: Proceedings of the 27th ACM International Conference on multimedia.

Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. 2019. Few-Shot Adversarial Learning of Realistic Neural Talking Head Models. 9458–9467. https://doi.org/10.1109/ICCV.2019.00955

Hang Zhou, Yu Liu, Ziwei Liu, Ping Luo, and Xiaogang Wang. 2019. Talking Face Generation by Adversarially Disentangled Audio-Visual Representation. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence (AAAI'19/IAAI'19/EAAI'19)*. AAAI Press, Article 1141, 8 pages. https://doi.org/10.1609/aaai.v33i01.33019299

Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. 2020. MakeItTalk: Speaker-Aware Talking-Head Animation. *ACM Trans. Graph.* 39, 6, Article 221 (nov 2020), 15 pages. https://doi.org/10.1145/3414685.3417774

Yang Zhou, Zhan Xu, Chris Landreth, Evangelos Kalogerakis, Subhransu Maji, and Karan Singh. 2018. Visemenet: audio-driven animator-centric speech animation. *ACM Transactions on Graphics* 37, 4 (Aug. 2018), 1–10. https://doi.org/10.1145/3197517.3201292

Victor Brian Zordan, Bhrigu Celly, Bill Chiu, and Paul C DiLorenzo. 2004. Breathe easy: model and control of simulated respiration for animation. In *Proceedings of the 2004 ACM SIGGRAPH/Eurographics symposium on Computer animation*. 29–37.