Signal Processing for Multimedia
J.S. Byrnes (Ed.)
IOS Press, 1999

Modification of Audible and Visual Speech

Michele Covell and Malcolm Slaney and Christoph Bregler [1] and Margaret Withgott [2]

Interval Research Corporation

ABSTRACT. Speech is one of the most common and richest methods that people use to communicate with one another. Our facility with this communication form makes speech a good interface for communicating with or via computers. At the same time, our familiarity with speech makes it difficult to generate synthetic but natural-sounding speech and synthetic but natural-looking lip-synced faces. One way to reduce the apparent unnaturalness of synthetic audible and visual speech is to modify natural (human-produced) speech. This approach relies on examples of natural speech and on simple models of how to take those examples apart and to put them back together to create new utterances.

We discuss two such techniques in depth. The first technique, *Mach1*, changes the overall timing of an utterance, with little loss in comprehensibility and with no change in the wording of or emphasis within what was said or in the identity of the voice. This ability to speed up (or slow down) speech will make speech a more malleable channel of communication. It gives the listener control over the amount of time that she spends listening to a given oration, even if the presentation of that material is prerecorded. The second technique, *Video Rewrite*, synthesizes images of faces, lip synced to a given utterance. This tool could be useful for reducing the data rate for video conferencing [31], as well as for providing photorealistic avatars.

1. Overview

Speech is one of the most common and richest methods that people use to communicate with one another. We learn to speak earlier than we learn to read or write, and we use speech to interact with other people throughout our lifetimes. In addition to literal meaning of the words, our utterances carry information in emphasis and emotion, as well as indications of the gender, identity, and health of the speaker. When we can see the speaker as well as listen to her, we use both her facial gestures and the audio signal to understand her utterance's basic meaning and the nuances of identity and emotion carried therein.

Our facility with this communication form makes speech a good interface for communicating with or via computers. Certainly the use of speech as an interface medium is growing, most notably in automated dictation-transcription machines and in information services for constrained applications, such as directory assistance. Yet it is difficult to generate synthetic but natural-sounding speech and synthetic but natural-looking *visual speech*—that is, synthetic talking faces. Part of this difficulty arises because of our facility and familiarity with speech: We have listened to and watched other people talk for our whole lives, so we quickly hear and see artifacts in synthetic speech.

One way to reduce the apparent unnaturalness of synthetic audible and visual speech is to modify natural (human-produced) speech. This approach allows us to avoid using a detailed hand-coded or analytic model of how speech is produced and how the facial structures move

The description of Mach1 has been previously published [11]. A more complete description of the listeners' test can be found in that reference. Similarly, the description of Video Rewrite has been previously published [5]. Again, a more complete description of the results of our syntheses can be found in that reference.

[1] C. Bregler is currently affiliated with New York University.

[2] M. Withgott is currently affiliated with Electric Planet, Inc.

TABLE 1. Categories of audible and visual speech modification

changed properties	unchanged properties	Audible Speech	Visual Speech (face, lips)
overall timing	identity, wording	time-scale modification (Mach1)	
wording	identity	concatenative speech synthesis [26]	data-driven synthetic lip sync (Video Rewrite)

to produce it. Rather, it relies on simpler, less comprehensive models; it makes up for this low level of modeling by using the information implicit in the natural-speech samples that have been collected. It replaces the assumption that we can create comprehensive models of vocal tracts, faces, emphasis, and timing by using examples of natural speech and simple models of how to take those examples apart and to put them back together in new sequences.

There are many different types of information provided by speech—wording, timing, emphasis, emotion, identity—so there are many different ways in which speech can be modified (Table 1). Here, we consider methods for changing the timing and the wording of visual speech. Good discussions of identity changes are available elsewhere [2, 27, 35, 39]. Current understanding of how emphasis and emotion are conveyed and how they should be modified is incomplete.

We discuss only two techniques in depth. The first technique, *Mach1*, changes the overall timing of an utterance, with little loss in comprehensibility and with no change in the wording of or emphasis within what was said or in the identity of the voice. This ability to speed up (or slow down) speech will make speech a more malleable channel of communication. It gives the listener control over the amount of time that she spends listening to a given oration, even if the presentation of that material is prerecorded. Other features that would increase the listener's control over a presentation are the ability to jump to new sentences [1], the ability to find and jump to topic changes [38], and the ability to listen to a summary of what was said [8].

The second technique, *Video Rewrite*, provides facial animations, lip synced to a given utterance. This tool could be useful for reducing the data rate for video conferencing [31], as well as for providing photorealistic avatars. With Video Rewrite, a low-data-rate video conferencing system could transmit video models of the participants beforehand and only the audio would be transmitted during the actual conferencing session. Another tool would detect emotion in the audio track [36] and adjust the facial expression accordingly. With synthetic talking faces (such as from Video Rewrite) and text-to-speech capabilities (such as from concatenative speech synthesis [26]), we can build avatars and computer agents who talk to us in speech that is natural and comprehensible (instead of simply putting text on a screen), and who are represented by lip-synced facial displays.

2. Modification of Overall Timing

At times, we may wish to listen to speech at an overall rate faster than the one at which it was originally spoken. For example, voice mail makes it easy and attractive for callers to leave impromptu messages. In contrast, listening to voice-mail messages at their recorded speed is often time-consuming and tedious compared to glancing at a written note. Voice-mail messages are more manageable when the listener can control the playback speed. Similarly, being able to listen to the sped-up audio track of a video in fast forward would be useful when cueing video. However, speeding up the signal would be useful only if we could still understand the content.

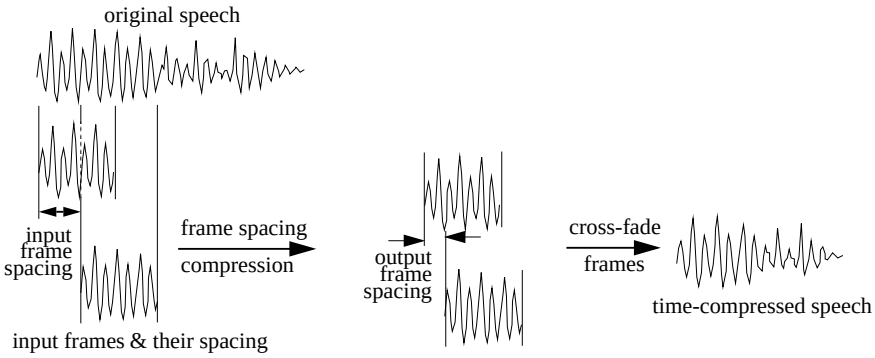

FIGURE 1. Time compression using SOLA.
SOLA starts by "chunking" the input into frames; then, based on the target compression rate, it computes a rough estimate of the output-frame spacing. This final spacing is set to the position of the maximum cross correlation near the desired spacing. SOLA creates the output audio by cross fading frames, using this new frame offset.

At first glance, this modification would seem simple: just play the sound back fast. The most basic version of this approach results in pitch- and formant-frequency shifts: We hear sounds like Mickey Mouse's high-pitched voice. We can use an algorithm called synchronous overlap add (*SOLA*) [32] with digital recordings to avoid these frequency shifts (Figure 1), as long as the playback rate is less than two times faster than the original recording rate. At faster rates, the time-compressed speech quickly takes on an unnatural cadence and becomes incomprehensible to most untrained listeners.

This limited range of playback rates restricts the applications in which time compression of speech can be used. Most people do video searches with consumer VCRs at rates much faster than twice real time. Although people would like to listen to the audio track, as well as to watch the video track, while cueing a tape to the desired material, it is not at all clear that they would accept slower fast-forward rates in exchange for this option.

2.1. *Linear Time Compression*

Time-compression techniques change the playback rate of speech without introducing pitch artifacts. As we stated, human comprehension of linearly time-compressed speech typically degrades at compression rates above two times real time [17]. These degradations are due neither to the speech rate per se nor to the number of words per minute (wpm) [14]. Most people cannot comprehend more than 270 wpm in compressed speech, even though they can understand quickly spoken passages of natural speech at up to 500 wpm [12].

The incomprehensibility of time-compressed speech is due to unnatural timing. Mach1, described in Section 2.2, provides an alternative to linear time compression. Mach1 compresses the components of an utterance in a way that resembles closely the natural timing of fast speech. Section 2.3 describes our test of comprehension and preference levels for Mach1-compressed and linearly compressed speech.

2.2. *Mach1 Time Compression*

Mach1 mimics the compression strategies that people use when they talk fast in natural settings. We used linguistic studies of natural speech [42, 46] to derive these goals:

- Compress pauses and silences the most
- Compress stressed vowels the least
- Compress schwas and other unstressed vowels by an intermediate amount
- Compress consonants based on the stress level of the neighboring vowels
- On average, compress consonants more than vowels

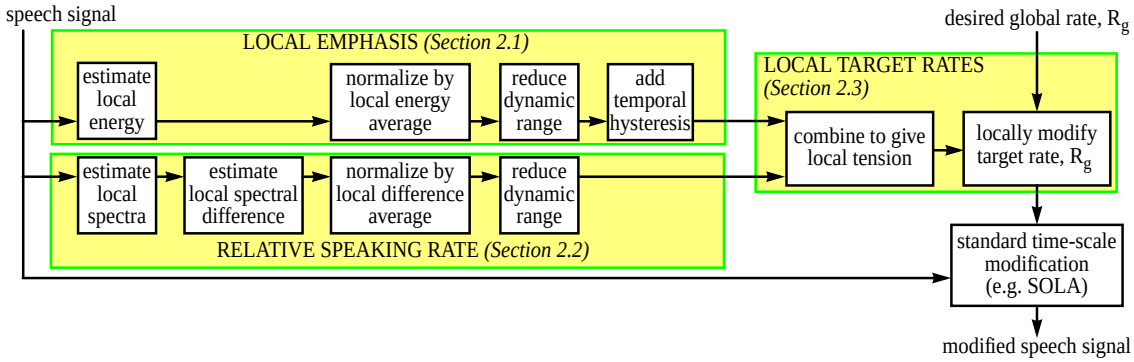

FIGURE 2. Overview of Mach1.
Mach1 first estimates the local emphasis and relative speaking rate. It then locally modifies the global target compression rate using a combination of those measures. The resulting locally varying target rate drives any standard time-scale-modification technique.

Also, to avoid obliterating very short segments, we want to avoid overcompressing already rapid sections of speech.

Unlike previous techniques [1, 21], Mach1 deliberately avoids categorical recognition (such as silence detection and *phoneme*[1] recognition). Instead, as illustrated in Figure 2, it estimates continuous-valued measures of local emphasis and relative speaking rate. Together, the sequences of the values of these two measures estimate what we call *audio tension*: the degree to which a given local speech segment is a poor candidate for faster playback rates. High-tension segments are less compressible than are low-tension segments. Based on the audio tension, we modify the general, preset target compression rate to a local target compression rate on the local speech segment. We use the local target rates to drive a standard, time-scale modification technique (e.g., synchronized overlap add [32]).

In Sections 2.2.1 through 2.2.3, we discuss the local-emphasis measure, the relative-speaking-rate measure, and the technique that we use to combine them. The Mach1 algorithm is explained in greater detail elsewhere [11].

2.2.1. *Measure of Local Emphasis.* We use the *local-emphasis measure* to distinguish among silence, unstressed syllables, and stressed syllables. Emphasis in speech correlates with relative loudness, pitch variations, and duration [8]. Of these, relative loudness is the easiest to estimate. Reliable pitch estimation is notoriously difficult. Reliable duration estimation requires phoneme recognition, because natural durations are highly phoneme dependent. Therefore, we rely on relative loudness to estimate emphasis.

To estimate local emphasis, we first calculate the local energy. Since emphasis is indicated more by relative loudness than by absolute loudness, we normalize our local energy by the local average energy.

These variations of the local relative energy are not linearly related to our goal: controlling the segment-duration variations to mimic those seen in natural speech. Instead, the local relative energy displays much larger upward variations than are observed in emphasized-segment durations, and much smaller downward variations than are seen in pause and unemphasized segment durations [11, 38, 41]. Therefore, we estimate the *frame emphasis* by applying a compressive function to the relative energy. The compressive function reduces the dynamic range of the high-relative-energy segments (the emphasized vowels) and expands the dynamic range of the low-relative-energy segments (the unemphasized vowels and the pauses).

[1] *Phonemes* are the distinct sounds within a language, such as the /IY/ and /P/ in “teapot.”

Human speech perception and production is in part determined by temporal grouping effects, especially within syllables [46] and within silences that immediately proceed or follow sounds [1,38]. To account for these temporal-grouping effects, we apply a 200-msec, tapered, temporal hysteresis to the frame emphasis to give our final local-emphasis estimates.

2.2.2. *Measure of Relative Speaking Rate.* We estimate the speaking rate to avoid overcompressing, and thereby obliterating, already-rapid speech segments. True speaking rate is difficult to measure. We can compute easily, however, measures of acoustic-variation rates, which covary with speaking rates. Conceptually, we are using the phoneme-transition rate to estimate speaking rate: The higher the transition rate, the faster the speaking rate. By lowering our compression during transitions, we effectively lower the compression of rapid speech. This approach also has the advantage of preserving phoneme transitions, which are particularly important for human comprehension [13, 37]. In practice, we use relative acoustic variability, instead of transition labels, to modulate the compression rate, thereby avoiding categorical errors and simplifying the overall estimation process.

Our estimate of relative acoustic variability starts with a local spectral estimate. To avoid unreliable estimates in low-energy regions, we set to the previous frame's values each frame whose energy level is below a dynamic threshold. We then sum the absolute log ratios between the current and the previous frames' values to estimate the local spectral difference. We use a log amplitude scale, instead of a linear one, because intensity-discrimination studies suggest that human perception of acoustic change is more closely approximated by the log scale [24]. Again, to avoid overestimating the spectral difference, we normalize each frame's values by that frame's total energy level, then sum over only the most energetic bins.

Different speaking styles and different recording environments introduce wide deviations in our absolute spectral-difference measure. To avoid heavy influence from these variables, we normalize our spectral difference by the local average difference. These variations of the relative spectral difference overestimate the upward variations in relative speaking rate [11, 41, 42]. Therefore, we estimate the *relative speaking rate* by applying a compressive function to the relative spectral difference.

2.2.3. *Local Target Compression Rates.* The local-emphasis and relative-speaking-rate measures depend purely on the audio signal that we plan to modify. They can be computed as the signal is being recorded. What remains is to combine these two measures to get the *audio tension*, which is a single measure of the compressibility of the underlying speech, and to combine the audio tension with the listener's target compression (or expansion) rate.

We compute audio tension from local emphasis and relative speaking rate using a simple linear formula. The audio tension increases as the local emphasis increases, from low tension (comparatively large compressions) in regions of silence to high tension (comparatively small compressions) in stressed segments. The audio tension also increases as the relative speaking rate increases, from low tension (large compressions) in regions of slow speech to high tension (small compressions) in regions of fast speech.

From audio tension, $\gamma(t)$, and from a desired global compression rate, R_g, we compute local target rates, $r(t)$, as[2]

$$r(t) = \max\{1, R_g + (1 - R_g)\gamma(t)\}.$$

We use these target local compression rates as an input to standard time-scale-modification techniques. With SOLA (Figure 1), for example, we use the local target rates to set, frame by frame, the target offset between the current and previous frames in the output audio signal.

[2] In this equation for $r(t)$, we assume that compression rates are expressed as numbers greater than 1. Using this convention, the offset between time frames of the output is set to $1/r(t)$ times the input frame offset for compression.

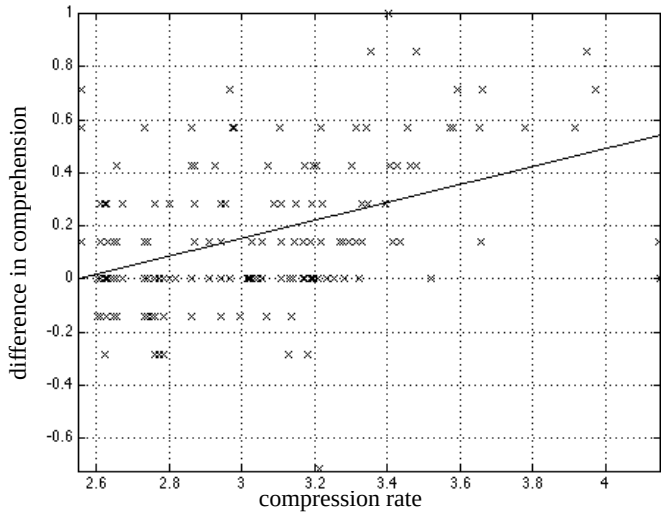

FIGURE 3. Plot of difference in comprehension between Mach1 and linear compression as a function of compression rate. The results of linear regression are shown as a solid line.

The sequence of local compression (or expansion) rates typically gives the overall compression (expansion) rates near the requested global rate, R_g. However, there is no guarantee that this global rate will be achieved. In cases where the global compression rate is important, we add a slow-response feedback loop around the previously described system. This feedback loop acts to correct long-term errors in the overall compression (expansion) rate by adjusting the nominal value of R_g appropriately.

2.3. *Comparison of Mach1-Compressed and Linearly Compressed Speech*

We conducted a listener test comparing Mach1-compressed speech to linearly compressed speech. The details of our method and of the statistical analysis we did on the results are given elsewhere [11]. A portion of the listener test can be found on our web page

$$\texttt{http://www.interval.com/papers/1997-061/}$$

2.3.1. *Method.* We used 108 audio clips, taken from Kaplan's TOEFL study program [33]. These audio clips cover three different discourse styles: short dialogs (1 sentence/turn, 1 turn/speaker); long dialogs (2 to 5 sentences/turn; 3 or 4 turns/speaker); and monologs (9 to 15 sentences). We used these audio clips for comparative studies of comprehension and preference between Mach1 compression and linear compression.

For these tests, we used both Mach1 compression and linear compression on each audio clip. We ran Mach1 compression first, with $R_g = 3$, and without correction of the overall compression rate. We computed the true compression rate achieved by Mach1 on each audio sample. We then recompressed each original, uncompressed clip linearly to the same global rate that the Mach1 compression achieved. This process ensured that the two versions had the same overall compression rate.

This group of 108 time-compressed audio clips were split into two balanced pools and were tested with 14 adult subjects who are fluent in English. In the comprehension sections of the test, one-half of the subjects heard the first audio pool compressed with Mach1 and the second audio pool compressed linearly. The other half of the subjects heard the complementary set of compressed clips: the first audio pool compressed linearly and the second audio pool compressed with Mach1. In the preference section, both Mach1-compressed and linearly compressed versions of the selected audio samples were played, so the subjects could make direct comparisons between the compression techniques.

2.3.2. *Results.* Mach1 compression offers significant improvements in comprehension over linear compression, especially at high compression rates. Mach1 improved comprehension by 25 percent over linear compression, at the same global rates. The difference in comprehension

rates increased with the compression rate (Figure 3). Listeners preferred Mach1-compressed speech over linearly compressed speech in 95 percent of cases; this preference for Mach1 increased with the compression rate.

Mach1 provided the greatest improvement (38 percent on average) in comprehension when used on short dialogs, where linear compression limited comprehension substantially. Improvements were less marked with the longer clips (13 percent with monologs, and 7 percent—not statistically significant at $p < 0.05$—with long dialogs). The short dialogs (average 23 words) are significantly shorter than the other clips (average 144 and 187 words for the long dialogs and monologs). One possible explanation for the lower comprehension of the linearly compressed short dialogs is that the most information is lost at the beginning of the clips, while the subjects adjust to the unnatural speaking style. The absence of a similar decrease in comprehension of Mach1-compressed short dialogs suggests that the listener-adjustment period is much shorter when Mach1 is used.

Note that, with Mach1 compression, there was *no* statistically significant loss in comprehension as a function of compression rate. One hypothesis explaining the uniform comprehension results across achieved compression rates is that the Mach1 audio-tension measure captures the relative compressibility of each audio clip. Mach1 itself determined the distribution of compression rates. It was given a nominal compression target of three times real time, but was allowed to deviate from that target. Mach1 may be providing a predictable overall comprehensibility, rather than a predictable overall compression rate.

Variable-rate compression of speech is a promising notion in time-scale modification. It should allow us to improve our comprehension rates by using approaches suggested by linguistic and text-to-speech studies. Still unanswered, however, is the question of how best to measure paralinguistic qualities, such as emphasis and relative speaking rate. The Mach1 approach avoids categorical labels and relies on easily measurable acoustic correlates. It confers significant improvements in comprehension over linear compression.

3. Modification of Visual Speech

Humans are extremely sensitive to the synchronization between speech and lip motions. Low-data-rate video conferencing and photorealistic avatars raise our expectations for realistic lip motions, making incorrect lip sync especially jarring. Similarly, interfaces to computers that use realistic human faces require high-quality lip sync to maintain the illusion of face-to-face interaction. In this section, we review facial animation systems for lip sync. We then discuss in more detail *Video Rewrite*, a video-based facial animation system for photorealistic synthetic lip sync.

Facial-animation systems build a model of how a person speaking sounds and looks. They use this model to generate a new output sequence, which matches the (new) target utterance. On the model-building side (analysis), there are typically three distinguishing choices: how the facial appearance is learned or described, how the facial appearance is controlled or labeled, and how the *viseme* labels are learned or described. Visemes are the visual counterpart to *phonemes*: Visemes are visually distinct mouth, teeth, and tongue articulations for a language. For example, the phonemes /B/ and /P/ are visually indistinguishable and are grouped into a single viseme.

For output-sequence generation (synthesis), the distinguishing choice among facial-animation systems is how the target utterance is characterized. We review a representative sample of past research in these areas.

Many facial-animation systems use a generic 3D mesh model as the source for facial appearance [15, 22, 29], sometimes adding texture mapping to improve realism [9, 25, 44]. Another synthetic source of face data is hand-drawn images [23]. Other systems use images

of real faces for their source examples, including 3D scans [45] and still images [34]. We use video footage to train Video Rewrite's models.

Once a facial model is captured or created, the control parameters that exercise that model must be defined. In systems that rely on a 3D mesh model for appearance, the control parameters are the allowed 3D mesh deformations. Most image-based systems label the positions of specific facial locations (for example, the bottom of the chin) as their control parameters. Most such systems rely on manual labeling of each example image [23, 34]. Video Rewrite creates its video model by automatically labeling specific facial locations (Section 3.2.1).

The final step in the analysis stage is acquiring the viseme labels. Many facial-animation systems label different visual configurations with an associated *phoneme*. These systems then match these phoneme labels with their corresponding labels in the target utterance. With synthetic images, the phoneme labels are artificial or are learned by analogy [25]. When the system uses natural images, taken from a video of a person speaking, the phonemic labels can be generated manually [34] or automatically. Video Rewrite determines the phoneme labels automatically (Section 3.2.2).

As we mentioned, for output-sequence generation (synthesis), the distinguishing choice is how the target utterance is characterized. The goal of facial animation is to generate an image sequence that matches this target utterance. When phoneme labels are used, those for the target utterance can be entered manually [34] or computed automatically [22,25]. Another option for phoneme labeling is to create the new utterance with synthetic speech [9, 29, 44]. Approaches that do not use phoneme labels include motion capture of facial locations that are artificially highlighted [15, 45] and manual control by an animator [23]. Video Rewrite uses a combination of phoneme labels (from the target utterance) and facial-location labels (from the video-model segments). Video Rewrite derives all these labels automatically.

3.1. *Overview of Video Rewrite*

Video Rewrite provides video-based synthetic lip sync, using an approach similar to concatenative speech synthesis. *Concatenative speech synthesis* [26] is a recent and successful approach to changing wording, providing the audio back-end to high-quality text-to-speech translation systems. Instead of modeling the vocal tract, concatenative speech synthesis analyzes a corpus of speech, selects examples of phonemes, and normalizes those examples. Concatenative speech then synthesizes new words by concatenating proper sequences of phonemes, then warping pitch and duration to create speech that sounds natural. This data-driven approach to synthesis is more effective at capturing the nuances of human speech than are approaches that rely primarily on hand-coded rules about speech.

Similarly, Video Rewrite creates new videos in two steps: analysis of a training database and synthesis of new footage. In the **analysis** stage, Video Rewrite automatically segments into phonemes the audio track of the training database. We use these labels to segment the video track as well. We automatically track head pose and facial features in this segmented footage. To track the facial features, Video Rewrite requires a small number (about 25) of hand-labeled images that indicate the locations of specific points on the speaker's face. The hand labeling of these images is the only human annotation or intervention that is required.[3] The steps used in the analysis stage are shown in Figure 4.

In the *synthesis* stage, our system uses this video database with a new utterance. It retrieves the appropriate viseme sequences automatically. Video Rewrite blends the viseme sequences together and into a background scene using morphing techniques. The result is a new video of a person whose lip and jaw movements synchronize to the new audio. The steps used in the synthesis stage are shown in Figure 5.

[3]Even this level of human interaction is not a fundamental requirement: We could use face-independent models instead [10, 18].

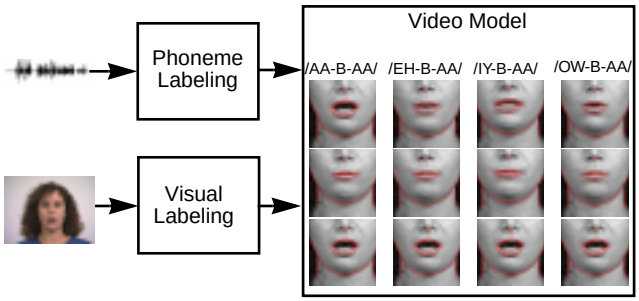

FIGURE 4. Overview of analysis stage.
Video Rewrite uses the audio track to segment the video into triphones. Vision techniques find the orientation of the head, and the shape and position of the mouth and chin, in each image. In the synthesis stage, Video Rewrite selects from this video model to synchronize new lip videos to any given audio.

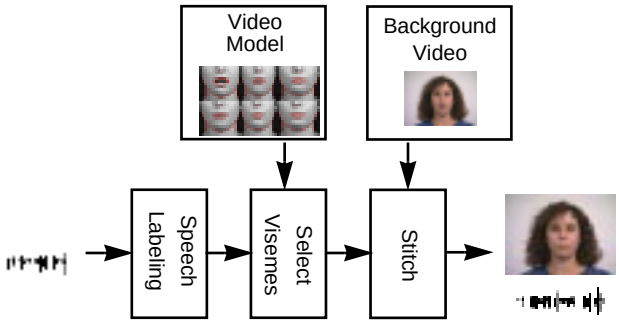

FIGURE 5. Overview of synthesis stage.
Video Rewrite segments new audio and uses it to select triphones from the video model. Based on labels from the analysis stage, the new mouth images are morphed into a new background face.

In the remainder of this section, we describe the analysis (Section 3.2) and synthesis (Section 3.3) stages of Video Rewrite. We then conclude by describing our results (Section 3.4).

3.2. *Analysis Stage of Video Rewrite*

As shown in Figure 4, the analysis stage of Video Rewrite creates an annotated database of example video clips, derived from unconstrained footage. We refer to this collection of annotated examples as a *video model*. This model captures how the subject's mouth and jaw move during speech. These *training videos* are labeled automatically with the phoneme sequence uttered during the video, and with the locations of fiduciary points that outline the lips, teeth, and jaw.

As we shall describe, the phonemic labels are from a time-aligned transcript of the speech, generated by a hidden Markov model (HMM). Video Rewrite uses the phonemic labels from the HMM to segment the input footage into short video clips, each showing three phonemes or a *triphone*. These *triphone videos*, with the fiduciary-point locations and the phoneme labels, are stored in the video model.

In Sections 3.2.1 and 3.2.2, we describe the visual and acoustic analyses of the video footage. In Section 3.3, we explain how Video Rewrite uses this model to synthesize new video.

3.2.1. *Annotation Using Image Analysis.* Video Rewrite can use any footage of the subject speaking; it is not constrained to use footage of actors with artificially highlighted lips. As her face moves within the frame, we need to know the mouth position and the lip shapes at all times; this information is provided by the image annotations of the analysis stage. In the synthesis stage, we use this information to warp overlapping videos such that they have the same lip shapes, and to align the lips with the background face.

Manual labeling of the fiduciary points around the mouth and jaw is error prone and tedious. Instead, we use computer-vision techniques to label the face and to identify the mouth and its shape. A major hurdle to automatic annotation is the low resolution of video images. In a typical scene, the lip region has a width of only 40 pixels. Conventional contour-tracking

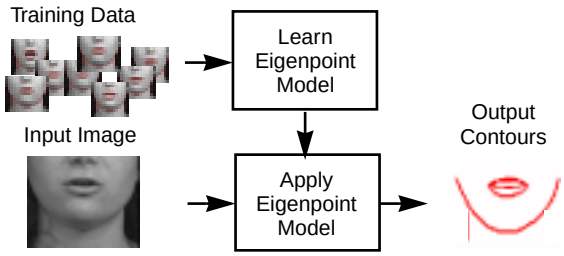

FIGURE 6. Overview of the eigenpoints algorithm.
Eigenpoints uses a small set of hand-labeled facial images to train subspace models. Given a new image, the eigenpoint models tell us the positions of points on the lips and jaw.

FIGURE 7. Mask that Video Rewrite uses to estimate the global warp.
Each image is warped to account for changes in the head's position, size, and rotation. The transform minimizes the difference between the transformed images and the face template. The mask (left) forces the minimization to consider only the upper face (right).

algorithms [16, 47] work well on high-contrast outer-lip boundaries with some user interaction, but fail on inner-lip boundaries at this resolution, due to the low signal-to-noise ratios. Grayscale-based algorithms, such as eigenimages [18, 40], work well at low resolutions, but estimate only the location of the lips or jaw, rather than estimating the desired fiduciary points. The *eigenpoints* algorithm [10], and other extensions of eigenimages [20], estimate control points reliably and automatically, even in such low-resolution images. As shown in Figure 6, eigenpoints learns how fiduciary points move as a function of the image appearance, and then uses this model to label new footage.

Video Rewrite labels each image in the training video using a total of 54 eigenpoints: 34 on the mouth (20 on the outer boundary, 12 on the inner boundary, 1 at the bottom of the upper teeth, and 1 at the top of the lower teeth) and 20 on the chin and jaw line. There are two separate eigenpoint analyses. The first *eigenspace* controls the placement of the 34 fiduciary points on the mouth, using pixels around the *nominal mouth location*—a region that covers the mouth completely. The second eigenspace controls the placement of the 20 fiduciary points on the chin and jaw line, using pixels around the *nominal chin location*—a region that covers the upper neck and the lower part of the face.

We create the two eigenpoint models for locating the fiduciary points from a small number of images: usually 20-30, depending of the variation in head pose. We extend the hand-annotated dataset by left-right flipping images and by morphing pairs of annotated images to form intermediate images, expanding the original set of n hand-annotated images to $2n^2 + n$ annotated images without any additional manual work. We then derive eigenpoints models using this extended data set.

We use eigenpoints to find the mouth and jaw and to label their contours. The derived eigenpoint models locate the facial features using six basis vectors for the mouth and six different vectors for the jaw. Eigenpoints then places the fiduciary points around the feature locations: 32 basis vectors place points around the lips, and 64 basis vectors place points around the jaw.

Eigenpoints assumes that the features (the mouth or the jaw) are undergoing pure translational motion. It does a comparatively poor job at modeling rotations and scale changes. Yet, Video Rewrite is designed to use unconstrained footage. We expect rotations and scale changes. Subjects may lean toward the camera or turn away from it, tilt their heads to the side, or look up from under their eyelashes.

To allow for a variety of motions, we warp each face image into a standard reference orientation, prior to eigenpoints labeling. We find the global transform that minimizes the

mean-squared error between a large portion of the face image and a facial template. We currently use an ellipsoidal transform, followed by an affine transform [4]. The ellipsoid allows us to describe the curvature of the face and to compensate for changes in pose. Subsequent processing using an affine transform provides more accurate and reliable estimates of the head's translation and rotation. The mask shown in Figure 7 defines the support of these minimization integrals. Once the best global mapping is found, it is inverted and applied to the image, putting that face into the standard coordinate frame. We then perform eigenpoints analysis on this pre-warped image to find the fiduciary points. Finally, we back-project the fiduciary points through the global warp to place them on the original face image.

3.2.2. *Annotation Using Audio Analysis.* All the speech data in Video Rewrite (and their associated video clips) are segmented into sequences of phonemes. Although single phonemes constitute a convenient representation for linguistic analysis, they are not appropriate for Video Rewrite. We want to capture the visual dynamics of speech. To do so correctly, we must consider *coarticulation*, which causes the lip shapes for many phonemes to be modified based on the phoneme's context. For example, the /T/ in "beet" looks different from the /T/ in "boot."

Therefore, Video Rewrite segments speech and video into triphones: collections of three sequential phonemes. The word "teapot" is split into the sequence of triphones /SIL–T–IY/,[4] /T–IY–P/, /IY–P–AA/, /P–AA–T/, and /AA–T–SIL/. When we synthesize a video clip, we emphasize the middle of each triphone, and cross-fade the overlapping regions of neighboring triphones. We thus ensure that the precise transition points are not critical, and that we can capture effectively many of the dynamics of both forward and backward coarticulation.

Video Rewrite uses HMMs [30] to label the training footage with phonemes. We trained the HMMs using the TIMIT speech database [19], a collection of 4200 utterances with phonemic transcriptions that gives the uttered phonemes and their timing. Each of the 61 phoneme categories in TIMIT is modeled with a separate three-state HMM. The emission probabilities of each state are modeled with mixtures of eight Gaussians with diagonal covariances. For robustness, we split the available data by gender and train two speaker-independent, gender-specific systems, one based on 1300 female utterances, and one based on 2900 male utterances.

We use these gender-specific HMMs to create a fine-grained phonemic transcription of our input footage, using *forced Viterbi search* [43]. Forced Viterbi uses unaligned sentence-level transcriptions and a phoneme-level pronunciation dictionary to create a time-aligned phoneme-level transcript of the speech. From this transcript, Video Rewrite segments the video automatically into triphone videos, labels them, and includes them in the video model.

3.3. *Synthesis Stage of Video Rewrite*

As shown in Figure 5, Video Rewrite synthesizes the final lip-synced video by labeling the new speech track, selecting the sequence of triphone videos that most accurately matches the new speech utterance, and stitching these images into a background video.

The background video sets the scene and provides the desired head position and movement. The background sequence in Video Rewrite includes most of the subject's face, as well as the scene behind the subject. The frames of the background video are taken from the source footage in the same order as they were shot. The head tilts and the eyes blink, based on the background frames.

In contrast, the different triphone videos are used in whatever order is needed. They simply show the motions associated with articulation. The triphone images include the mouth,

[4] /SIL/ indicates silence. Two /SIL/ in a row are used at the beginnings and ends of utterances to allow all segments-including the beginning and end-to be treated as triphones.

chin, and part of the cheeks, so that the chin and jaw move and the cheeks dimple appropriately as the mouth articulates. We use illumination-matching techniques [7] to avoid visible seams between the triphone and background images.

The first step in synthesis (Figure 5) is labeling the new soundtrack. We label the new utterance with the same HMM that we used to create the video-model phoneme labels. In Sections 3.3.1 and 3.3.2, we describe the remaining steps: selecting triphone videos and stitching them into the background.

3.3.1. *Selection of Triphone Videos.* The new speech utterance determines the target sequence of speech sounds, marked with phoneme labels. We would like to find a sequence of triphone videos from our database that matches this new speech utterance. For each triphone in the new utterance, our goal is to find a video example with exactly the transition we need, and with lip shapes that match the lip shapes in neighboring triphone videos. Since this goal often is not attainable, we compromise by a choosing a sequence of clips that approximates the desired transitions and shape continuity.

Given a triphone in the new speech utterance, we compute a matching distance to each triphone in the video database. The matching metric has two terms: the *phoneme-context distance,* D_p, and the *distance between lip shapes* in overlapping visual triphones, D_s**.** The total error is

$$\text{error} = \alpha D_p + (1 - \alpha) D_s$$

where the weight, α, is a constant that trades off the two factors.

The phoneme-context distance, D_p, is based on categorical distances between phoneme categories and between viseme classes. Since Video Rewrite does not need to create a new soundtrack (it needs only a new video track), we can cluster phonemes into viseme classes, based on their visual appearance.

We use 26 viseme classes. Ten are consonant classes: /CH/, /JH/, /SH/, /ZH/; /K/, /G/, /N/, /L/; /T/, /D/, /S/, /Z/; /P/, /B/, /M/; /F/, /V/; /TH/, /DH/; /W/, /R/; /HH/; /Y/; and /NG/. Fifteen are vowel classes: one each for /EH/, /EY/, /ER/, /UH/, /AA/, /AO/, /AW/, /AY/, /UW/, /OW/, /OY/, /IY/, /IH/, /AE/, /AH/. One class is for silence, /SIL/.

The phoneme-context distance, D_p, is the weighted sum of phoneme distances between the target phonemes and the video-model phonemes within the context of the triphone. If the phonemic categories are the same (for example, /P/ and /P/), then this distance is 0. If they are in different viseme classes (/P/ and /IY/), then the distance is 1. If they are in different phonemic categories but are in the same viseme class (/P/ and /B/), then the distance is a value between 0 and 1. The intraclass distances are derived from published confusion matrices [28].

In D_p, the center phoneme of the triphone has the largest weight, and the weights decrease smoothly from that center weight. Although the video model stores only triphone images, we consider the triphone's original context when picking the best-fitting sequence. In current animations, this context covers the triphone itself, plus one phoneme on either side.

The second term, D_s, measures how closely the mouth contours match in overlapping segments of adjacent triphone videos. In synthesizing the mouth shapes for "teapot," we want the contours for the /IY/ and /P/ in the lip sequence used for /T–IY–P/ to match the contours for the /IY/ and /P/ in the sequence used for /IY–P–AA/. We measure this similarity by computing the Euclidean distance, frame by frame, between four-element feature vectors containing the overall lip width, overall lip height, inner lip height, and height of visible teeth.

The lip-shape distance (D_s) between two triphone videos is minimized with the correct time alignment. For example, consider the overlapping contours for the /P/ in /T–IY–P/ and /IY–P–AA/. The /P/ phoneme includes both a silence, when the lips remain pressed together, and an audible release, when the lips move rapidly apart. The durations of the initial silences within the /P/ phoneme may be different. The phoneme labels do not provide us with this level of detailed timing. Yet, if the silence durations are different, the lip-shape distance for two

otherwise-well-matched videos will be large. This problem is exacerbated by imprecision in the HMM phonemic labels.

We want to find the temporal overlap between neighboring triphones that maximizes the similarity between the two lip shapes. We shift the two triphones relative to each other to find the best temporal offset and duration. We then use this optimal overlap both in computing the lip-shape distance, D_s, and in cross-fading the triphone videos during the stitching step. The optimal overlap is the one that minimizes D_s while maintaining a minimum-allowed overlap.

Since the fitness measure for each triphone segment depends on that segment's neighbors in both directions, we select the sequence of triphone segments using dynamic programming over the entire utterance. This procedure ensures the selection of the optimal segments.

3.3.2. *Stitching It Together*. Video Rewrite produces the final video by stitching together the appropriate entries from the video database. At this point, we have already selected the sequence of triphone videos that most closely matches the target audio. We need to align the overlapping lip images temporally. This internally time-aligned sequence of videos is then time aligned to the new speech utterance. Finally, the resulting sequences of lip images are spatially aligned and are stitched into the background face. We describe each step in turn.

We have a sequence of triphone videos that we must combine to form a new mouth movie. In combining the videos, we want to maintain the dynamics of the phonemes and their transitions. We need to time align the triphone videos carefully before blending them, because otherwise the mouth will appear to flutter open and closed inappropriately. We align the triphone videos by choosing a portion of the overlapping triphones where the two lips shapes are as similar as possible. We make this choice when we evaluate D_s to choose the sequence of triphone videos (Section 3.3.1). We use the overlap duration and shift that provide the minimum value of D_s for the given videos.

We now have a self-consistent temporal alignment for the triphone videos. We have the correct articulatory motions, in the correct order to match the target utterance, but these articulations are not yet time aligned with the target utterance. We align the lip motions with the target utterance by comparing the corresponding phoneme transcripts. The starting time of the center phoneme in the triphone sequence is aligned with the corresponding label in the target transcript. The triphone videos are then stretched or compressed such that they fit the time needed between the phoneme boundaries in the target utterance.

The remaining task is to stitch the triphone videos into the background sequence. The correctness of the facial alignment is critical to the success of the recombination. The lips and head are constantly moving in the triphone and background footage. Yet, we need to align them all so that the new mouth is firmly planted on the face. Any error in spatial alignment causes the mouth to jitter relative to the face—an extremely disturbing effect. We again use the mask from Figure 7 to help us find the optimal global transform to register the faces from the triphone videos with the background face. The combined transforms from the mouth and background images to the template face (Section 3.2.1) give our starting estimate in this search. Re-estimating the global transform by directly matching the triphone images to the background improves the accuracy of the mapping.

We use a replacement mask (Figure 3.3.2) to specify which portions of the final video come from the triphone images and which come from the background video. This replacement mask warps to fit the new mouth shape in the triphone image and to fit the jaw shape in the background image.

Local deformations are required to stitch the shape of the mouth and jaw line correctly. These two shapes are handled differently. The mouth's shape is completely determined by the triphone images. The only changes made to these mouth shapes are imposed to align the mouths within the overlapping triphone images: The lip shapes are linearly cross-faded between the shapes in the overlapping segments of the triphone videos.

FIGURE 8. Facial fading mask. This mask determines which portions of the final video frames come from the background frame, and which come from the triphone database. The mask should be large enough to include the mouth and chin. These images show the replacement mask applied to a triphone image, and its inverse applied to a background image. The mask warps according to the mouth and chin motions.

The jaw's shape, on the other hand, is a combination of the background jaw line and the two triphone jaw lines. Near the ears, we want to preserve the background video's jaw line. At the center of the jaw line (the chin), the shape and position are determined completely by what the mouth is doing. The final image of the jaw must join smoothly together the motion of the chin with the motion near the ears. Therefore, we vary smoothly the weighting of the background and triphone shapes as we move along the jawline from the chin toward the ears.

The final stitching process is a three-way tradeoff in shape and texture among the fade-out lip image, the fade-in lip image, and the background image. As we move from phoneme to phoneme, the relative weights of the mouth shapes associated with the overlapping triphone-video images are changed. Within each frame, we vary spatially the relative weighting of the jaw shapes contributed by the background image and of the triphone-video images.

The derived fiduciary positions are used as control points in morphing. All morphs are done with the Beier-Neely algorithm [3]. For each frame of the output image, we need to warp four images: the two triphones, the replacement mask, and the background face. The warping is straightforward, since we generate high-quality control points automatically using the eigenpoints algorithm.

3.4. *Results from Video Rewrite*

We have applied Video Rewrite to several different training databases. We recorded one video dataset specifically for our evaluations. Section 3.4.1 describes the methods that we used to collect these data and to create lip-sync videos, as well as our evaluation of the resulting videos. More details about our method and our evaluation for this experiment are given elsewhere [5].

We also used old footage of John F. Kennedy to evaluate the system's performance on extremely small databases. We summarize the results from these experiments in Section 3.4.2. Details about our method and our evaluation of this specific experiment were reported elsewhere [6].

3.4.1. *Reanimation of High-Quality Footage.* We recorded about 8 minutes of video, containing 109 sentences, of a woman narrating a fairy tale. The subject was also asked to wear a hat during the filming. We use this landmark to provide a quantitative evaluation of our global alignment. The hat is strictly outside all our alignment masks and our eigenpoints models. Thus, having the subject wear the hat does not affect the magnitude or type of errors that we expect to see in the animations—it simply provides us with a reference marker for the position and movement of her head.

To create a video model, we trained the system on still-head footage. We hand annotated only 26 images (of 14,218 images total; about 0.2 percent). Video Rewrite then constructed

FIGURE 9. Examples of synthesized output frames.
These frames show the high quality of Video Rewrite's output after triphone segments have been stitched into different background video frames.

and annotated the video model automatically with just under 3500 triphone videos, using HMM labeling of triphones and eigenpoint labeling of facial contours.

Video Rewrite was then given the target sentence, and was asked to construct the corresponding image sequence. We evaluated the output footage both qualitatively and quantitatively. Our qualitative evaluation was done informally, by a panel of observers; only the (global) spatial registration was evaluated quantitatively. Because the narrator's hat moved rigidly with her upper head, we were able to measure quantitatively our global-registration error on this footage.

Examples of our output footage can be viewed at

$$\texttt{http://www.interval.com/papers/1997-012/}$$

The top row of Figure 9 shows example frames, extracted from these videos. This section describes our evaluation criteria and results.

- There are visible timing errors in less than 1 percent of the phonemes. These timing errors all occur during plosives and stops. There are no visible artifacts due to synchronization errors between triphone videos.
- No registration errors are visible. In quantitative terms, using the hat registration as a metric, the mean, median, and maximum errors in the still-head videos were 0.6, 0.5, and 1.2 pixels (standard deviation 0.3); those in the moving-head videos were 1.0, 1.0, and 2.0 pixels (standard deviation 0.4). For comparison, the face covers approximately 85 X 120 pixels.
- We did not observe unnatural out-of-plane distortions of the lips. Such distortion would be caused by mistakes in the estimate of out-of-plane facial curvature. However, the out-of-plane motions within our database and our background movies were fairly limited. The JFK footage, discussed in Section 3.4.2, provided a better test for this type of error.
- The illumination matching is accurate. Without illumination correction, we see artifacts in some of the moving-head videos; for example, when the narrator looked down, the lighting on her face changed significantly. These artifacts disappear with adaptive illumination correction [7].

- Artifacts are occasionally visible near the outer edges of the jaw and neck. These artifacts are due to incorrect warping of the background image.
- No unnatural-looking articulation artifacts could be traced to triphone-sequence replacement. Video Rewrite approximated about 31 percent of triphone sequences using other video sequences. Even so, none of the visible artifacts seemed to correlate with these replacements.
- Despite the foregoing occasional artifacts, the overall quality of the final video was judged informally to be excellent.

3.4.2. *Reanimation of Historic Footage.* We also applied Video Rewrite to public-domain footage of John F. Kennedy. For this application, we digitized 2 minutes (1157 triphones) of Kennedy speaking during the Cuban missile crisis. Forty-five seconds of this footage are from a close-up camera, placed about 30 degrees to Kennedy's left. The remaining images are medium-range shots from the same side. The size ratio is approximately 5:3 between the close-up and medium-range shots. During the footage, Kennedy moves his head about 20 degrees vertically, looking down to read his speech from notes on his desk and then looking up to make eye contact with a center camera (film from which we do not have).

We used this video model to synthesize new animations of Kennedy saying, for example, "Read my lips" and "I never met Forrest Gump." These animations combine the footage from both camera shots (close-up and medium-range) and from all head positions. The resulting videos are shown on

$$\texttt{http://www.interval.com/papers/1997-012/}$$

The bottom row of Figure 9 shows example frames, extracted from these videos.

We evaluated our Kennedy results qualitatively along the following dimensions: synchronization between lip videos and between the composite lips and the utterance; spatial registration between the lip videos and between the composite lips and the background head; quality of the illumination matching between the lips and the background head; visibility of the chosen fading-mask extent and of the background warping; naturalness of the composited articulation; and the overall quality of the video.

- There are visible timing errors in about 1 percent of the phonemes. These timing errors all occur during plosives and stops. There are no visible artifacts due to synchronization errors between triphone videos.
- The lips are distorted unnaturally in 8 percent of the output frames. This distortion is caused by mistakes in the estimate of out-of-plane facial curvature. We see no other errors in the alignment between the lips and the background face.
- The illumination matching is accurate. There are no visible artifacts from illumination mismatches.
- The fading mask occasionally includes nonfacial regions (for example, the flag behind Kennedy or the President's shirt collar). This error results in visible artifacts in 4 percent of the output frames, when lips from one head position are warped into another head position.
- Unnatural-looking articulation results occasionally from replacement of a desired (but unavailable) triphone sequence. In our experiments with the Kennedy images, this type of replacement occurs on 94 percent of the triphone videos. Of those replacements, 4 percent were judged informally to be unnatural looking.
- Despite the foregoing occasional artifacts, the overall quality of the final video was judged informally to be very good.

4. Conclusions

We have reviewed two techniques for modifying audible and visual speech. These tools and others that change the wording [26], the speaker's identity [2, 35], the emotional content, and the emphasis of audible and visual speech are particularly useful as we rely more on computer-mediated speech communication between people. They, as well as speech recognition and speaker identification, will be needed for speech-based human-computer interfaces.

Acknowledgments

For work on Mach1: We thank Gerald McRoberts and Dan Levitin for their advice on designing the listener test, Jennifer Orton for running the listener tests, Jennifer Smith for her guidance and work in statistical analysis, and Tom Ngo and Lyn Dupré for their editing. We also thank all our listener-test subjects who gave us an hour of their time and attention to test our approach.

For work on Video Rewrite: Many colleagues helped us. Ellen Tauber and Marc Davis graciously submitted to our experimental manipulation. Trevor Darrell and Subutai Ahmad contributed many good ideas to the algorithm development. Trevor, Subutai, John Lewis, Bud Lassiter, Gaile Gordon, Kris Rahardja, Michael Bajura, Frank Crow, Bill Verplank, and John Woodfill helped us to evaluate our results. Bud Lassiter and Chris Seguine helped us with the video production. Lyn Dupré helped us to evaluate, correct, and edit our description. We offer many thanks to all.

References

[1] B. Arons, 1994. "Interactively Skimming Recorded Speech," Ph.D. dissertation, Massachusetts Institute of Technology, Boston MA.

[2] L. Arslan, D. Talkin, 1997. "Voice Conversion by Codebook Mapping of Line Spectral Frequencies and Excitation Spectrum," *Eurospeech*, vol. 3, pp. 1347–1350, Rhodes, Greece.

[3] T. Beier, S. Neely, 1992. "Feature-Based Image Metamorphosis." *Computer Graphics*, 26(2): 35–42.

[4] J. Bergen, P. Anandan, K. Hanna, R. Hingorani, 1992. "Hierarchical Model-Based Motion Estimation," *Proc. ECCV*, pp. 237–252, Liguria, Italy.

[5] C. Bregler, M. Covell, M. Slaney, 1997. "Video Rewrite: Driving Visual Speech with Audio." *SIGGRAPH'97*, pp. 353–356, Los Angeles CA.

[6] C. Bregler, M. Covell, M. Slaney, 1997. "Video Rewrite: Visual Speech Synthesis from Video." *Workshop on Audio-Visual Speech Processing*, pp. 153–156, Rhodes, Greece.

[7] P. Burt, E. Adelson, 1983. "A Multiresolution Spline with Application to Image Mosaics." *ACM Trans. Graphics*, 2(4): 217–236.

[8] F. Chen, M. Withgott, 1992. "The Use of Emphasis to Automatically Summarize a Spoken Discourse," *Proc. IEEE ICASSP*, vol. 1, pp. 229–232, San Francisco CA.

[9] M. Cohen, D. Massaro, 1993. "Modeling Coarticulation in Synthetic Visual Speech." In *Models and Techniques in Computer Animation*, ed. N.M Thalman, D. Thalman, pp. 139–156, Tokyo: Springer-Verlag.

[10] M. Covell, C. Bregler, 1996. "Eigenpoints." *Proc. IEEE ICIP*, vol. 3, pp. 471–474, Lausanne, Switzerland.

[11] M. Covell, M. Withgott, M. Slaney, 1998. "Mach1: Nonuniform Time-Scale Modification of Speech." *Proc. IEEE ICASSP*, vol. 1, pp. 349–352, Seattle WA.

[12] C. Fulford, 1993. "Can Learning Be More Efficient? Using Compressed Audio Tapes to Enhance Systematically Designed Text," *Educational Technology*, 33(2): 51–59.

[13] S. Furui, 1986. "On the Role of Spectral Transition for Speech Perception," *JASA*, 80(4): 1016–1025.

[14] P. Gade, C. Mills, 1989. "Listening Rate and Comprehension as a Function of Preference for and Exposure to Time-Altered Speech," *Perceptual and Motor Skills*, 68(2): 531–538.

[15] T. Guiard-Marigny, A. Adjoudani, C. Benoit, 1994. "A 3-D Model of the Lips for Visual Speech Synthesis." *Proc. ESCA/IEEE Workshop on Speech Synthesis*, pp. 49–52, New Paltz NY.

[16] M. Kass, A. Witkin, D. Terzopoulos, 1987. "Snakes: Active Contour Models." *Int. J. Computer Vision*, 1(4): 321–331.

[17] P. King, R. Behnke, 1989. "The Effect of Time-Compressed Speech on Comprehensive, Interpretive, and Short-Term Listening," *Human Communication Research*, 15(3): 428–443.

[18] M. Kirby, L. Sirovich, 1990. "Application of the Karhunen-Loeve Procedure for the Characterization of Human Faces." *IEEE PAMI*, 12(1): 103–108.

[19] L. Lamel, R. Kessel, S. Seneff, 1986. "Speech Database Development: Design and Analysis of the Acoustic-Phonetic Corpus." *Proc. Speech Recognition Workshop (DARPA)*, Report #SAIC-86/1546, pp. 100–109, McLean VA: Science Applications International Corp.
[20] A. Lanitis, C.J. Taylor, T.F. Cootes, 1995. "A Unified Approach for Coding and Interpreting Face Images." *Proc. IEEE ICCV*, pp. 368–373, Cambridge MA.
[21] S. Lee, H. Kim, et al., 1997. "Variable Time-Scale Modification of Speech Using Transient Information," *Proc. IEEE ICASSP*, vol. 2, pp. 1319–1322, Munich, Germany.
[22] J. Lewis, 1991. "Automated Lip-Sync: Background and Techniques." *J.Visualization and Computer Animation*, 2(4): 118–122.
[23] P. Litwinowicz, L. Williams, 1994. "Animating Images with Drawings." *SIGGRAPH 94*, pp. 409–412, Orlando FL.
[24] B. Moore, 1995. *Hearing*, Academic Press, San Diego CA.
[25] S. Morishima, H. Harashima, 1991. "A Media Conversion from Speech to Facial Image for Intelligent Man-Machine Interface." *IEEE J Selected Areas Communications*, 9 (4): 594–600.
[26] E. Moulines, P. Emerard, et al., 1990. "A Real-Time French Text-to-Speech System Generating High-Quality Synthetic Speech." *Proc. IEEE ICASSP*, pp. 309–312, Albuquerque NM.
[27] E. Moulines, Y. Sagisak (editors), 1995. "Voice Conversion: State of the Art and Perspective," Special issue of *Speech Communications*, 16, pp.125–216.
[28] E. Owens, B. Blazek, 1985. "Visemes Observed by Hearing-Impaired and Normal-Hearing Adult Viewers." *J. Speech and Hearing Research*, 28: 381–393.
[29] F. Parke, 1972. "Computer Generated Animation of Faces." *Proc. ACM National Conf.*, pp. 451–457.
[30] L. Rabiner, 1989. "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition." *Readings in Speech Recognition*, ed. A. Waibel, K. F. Lee, pp. 267–296, San Mateo CA: Morgan Kaufmann Publishers.
[31] R. Rao, T. Chen, R. Mersereau, 1998. "Audio-to-Visual Conversion for Multimedia Communication." *IEEE Trans on Industrial Electronics*, 45(1): 15–22.
[32] S. Roucous, A. Wilgus, 1985. "High Quality Time-Scale Modification for Speech," *Proc. IEEE ICASSP*, vol. 2, pp. 493–496, Tampa FL.
[33] M. Rymniak, G. Kurlandski, et al., 1997. *The Essential Review: TOEFL (Test of English as a Foreign Language)*, Kaplan Educational Centers, New York NY.
[34] K. Scott, D. Kagels, et al., 1994. "Synthesis of Speaker Facial Movement to Match Selected Speech Sequences." *Proc. Australian Conf. Speech Science and Technology*, pp. 620–625, Perth, Australia.
[35] M. Slaney, M. Covell, C. Lassiter, 1996. "Automatic Audio Morphing," *Proc. IEEE ICASSP*, vol. 2, pp. 1001–1004, Atlanta GA.
[36] M. Slaney, G. McRoberts, 1998. "Baby Ears: A Recognition System for Affective Vocalizations." *Proc. IEEE ICASSP*, vol. 2, pp. 985–988, Seattle WA.
[37] K. Stevens, 1980. "Acoustic Correlates of Some Phonetic Categories," *JASA*, 68(3): 836–842.
[38] L. Stifelman, 1997. "The Audio Notebook: Paper and Pen Interaction with Structured Speech," Ph.D. dissertation, Massachusetts Institute of Technology, Boston MA.
[39] E. Tellman, L. Haken, B. Holloway, 1995. "Timbre Morphing of Sounds with Unequal Numbers of Features," *J. of AES*, 43(9), pp. 678–689.
[40] M. Turk, A. Pentland, 1991. "Eigenfaces for Recognition." *J. Cognitive Neuroscience*, 3(1): 71–86.
[41] J. van Santen, 1992. "Contextual Effects on Vowel Duration," *Speech Communication*, 11(6): 513–546.
[42] J. van Santen, 1994. "Assignment of Segmental Duration in Text-to-Speech Synthesis," *Computer Speech and Language*, 8(2): 95–128.
[43] A. Viterbi, 1967. "Error Bounds for Convolutional Codes and an Asymptotically Optimal Decoding Algorithm." *IEEE Trans. Informat. Theory*, IT-13: 260–269.
[44] K. Waters, T. Levergood, 1995. "DECface: A System for Synthetic Face Applications." *J. Multimedia Tools and Applications*, 1(4): 349–366.
[45] L. Williams, 1990. "Performance-Driven Facial Animation." *Computer Graphics* (Proc. SIGGRAPH 90), 24(4): 235–242.
[46] M. Withgott, F. Chen, 1993. *Computational Models of American Speech, CSLI Lecture Notes #32*, Center for the Study of Language and Information, Stanford CA.
[47] A. Yuille, D.S. Cohen, P.W. Hallinan, 1989. "Feature Extraction from Faces using Deformable Templates." *Proc. IEEE CVPR*, pp. 104–109, San Diego CA.