Automated Lip-Sync: Direct Translation of Speech-Sound to Mouth-Shape

Barrett E. Koster, Robert D. Rodman and Donald Bitzer

Computer Science Department
North Carolina State University

Abstract

The goal of automatic lip-sync (ALS) is to translate speech sounds into mouth shapes. Although this seems related to speech recognition (SR), the direct map from sound to shape avoids many language problems associated with SR and provides a unique domain for error correction. Among other things, ALS animation may be used for animating cartoons realistically and as an aid to the hearing disabled. Currently, a program named Owie performs speaker dependent ALS for vowels.

Introduction.

In speaking, a person makes a variety of sounds. To change the sound, the vocal tract must assume different shapes. While variation in this process is possible -- for example, ventriloquists produce all their sounds with a single mouth shape and wild convolutions of the tongue -- there is a normal or standard shape for each speech sound. The goal of ALS is to produce continuously the canonical shape that matches the speech sound.

To perform ALS, a program named Owie is being developed. Owie is currently capable of creating speaker dependent lip animation off-line for quasi-static open-mouth vocalizations (basically, vowels).

While Owie's function is related to Speech Recognition (SR) and may ultimately help perform that task, it should be noted that the two are not the same. Owie does not attempt to segment the time signal into phonemes, words or sentences. Owie is not concerned with accent, because it does not analyze sounds phonemically (e.g., distinguishing "pin" and "pen" in Alabama is not an issue).

Owie is language independent. Owie is unaware of what language is being spoken or whether it is language at all. Owie does not even need to identify any particular sounds during operation.

Instead, Owie uses a continuous function, from sound to vocal-tract-shape (informally, "mouth shape"). To parameterize mouth shape, we have defined variables called "articulators". During operation, the articulator values are computed directly from the signal by articulator functions. These functions are trained using known sound-shape pairs. The training sounds may be described phonetically (and it is easy to discuss the program's coverage in terms of these phones), but the functions are in fact continuous and cover all intermediate sounds.

Articulators.

The output of the program is a graphical display of an external view of a face with moving lips. The face is generic, since there is no way of knowing what the speaker actually looks like. The motion is also generic, in the sense that only the canonical way of producing a given sound is displayed.

The animation is driven by the articulators. The ones for the lips and jaw can be measured with a ruler. The tongue parameters must be estimated from linguistic common knowledge. Here are the definitions.

1. **T1x.** Let T1 be the point of greatest constriction between the back of the tongue and the roof of the mouth. T1x measures the distance to T1 from the teeth, with T1x=0 at the teeth and T1x=1 at the back of the throat.
2. **T1m** measures constriction diameter at T1, from stopped (T1m=0) to maximally open (T1m=1).
3. **T2x.** Let T2 be the point of greatest constriction between the tip of the tongue and the roof of the mouth. T2x measures the distance to T2 from the teeth, similar to T1x for T1 and with the same scale. T2x<T1x.
4. **T2m** measures the constriction diameter at T2, similar to T1m for T1, and with the same scale. T2 doesn't move much for vowels.

1058-6393/95 $4.00 © 1995 IEEE

5. **Jaw** measures how far the jaw is open, with 'closed' (Jaw=0) such that teeth tips are even, and 'open' (Jaw=1) for "ah". Note: this is just the range used for speech, not the total range of someone's jaw motion. It is easily measured between fixed points on the nose and chin.
6. **Corners** measures mouth width. The corners of the mouth spread for "see" (Corners=0.5) and converge for "sue", with (Corners=-1) being a complete pucker or kiss shape. Corner separation is measured straight across the widest part of the mouth opening.
7. **LipFlare** measures lip separation in addition to any caused by Jaw or Corners movement. LipFlare=1 is for "sh". LipFlare = -1 is for "m". LipFlare must be measured after the effect of Jaw and Corners on lip separation have been determined.
8. **LowerLipUp** measures lower lip raising asymmetrical to upper lip, as for 'f' and 'v'. LowerLipUp=0 if for no asymmetrical movement. LowerLipUp=1 means the lip is touching the teeth.

The face.

Owie draws faces using several methods. In one version, Bezier patches are used to create a shaded 3-D look. Bezier patches are extremely versatile, so only one is needed for each lip, and the surrounding skin is done with only 4 more.

Each patch has 16 control points. Each control point has a position vector, and a velocity vector array that specifies how to move it with each articulator. Multiple articular effects are combined linearly.

The effect is flexible and shapely looking, and it is computed easily enough to run at 60 frames per second on a graphics work station.

There are, however, some non-linear speech motions that may require more control point positioning. Consider the fact that the lower lip comes up when the corners come in *only* when the mouth is open. One solution for this interaction is to be able to specify a point's position on the line between two other points instead of as absolute x, y and z. This is suitable for parameters that change proportionally, which ours do, and is needed besides to specify the colinearity that keeps the patch boundaries strictly flush.

To verify program operation on slow machines, the shaded patches can be replaced with stick drawing.

Sound processing.

The input signal is processed one glottal pulse ("GP") at a time. To find the GPs, a number of techniques are combined. GP length is estimated from the first maximum of the autocorrelation function and by adjusting the period to minimize (f1 +f2 + f3 + f5 + f6 + f7) / f4 over 4 periods [3].

Each GP is then identified as either the phase=0 point of the fundamental or the point of maximum damping, depending on whether the fundamental frequency is actually present in the signal or merely modulates higher frequencies. GP-sync allows for a very short window without the usual problem of fractional harmonics of the fundamental in the FFT.

From each GP, a spectrum is computed (standard FFT). From the spectrum, 'moment' functions are computed, first moments, second moments, and other functions which are not all moments. During training, the set is quite large and effectively arbitrary. As is explained in the next section, some of these moment functions turn out to have strong correlates with facial parameters. Thus the set of functions can be culled, and during program operation, only the useful ones need be computed.

Correlating the moment functions to the articulators.

The relationship between sound and mouth shape is contained in a 'model'. The model is made by having the speaker record a standard set of sounds, typically, a set of vowels. For each sound, there is mouth shape that makes that sound. Instead of measuring each speaker's mouth, we use a standard set of articulator values, so that the resulting motion will be the canonical motion desired. Each model point contains a sound from the speaker paired with the standard articulator values that go with that sound.

To create an articulator function from the model, we compute the set of arbitrary functions on all the sounds in the model, and then correlate the resulting values to the desired articulator values. The ones with the highest correlation are used during operation.

A priori , there is no guarantee that a correlation to each articulator can be found among *any* easily guessed moment functions. However, we have already found ones that do correlate to most of the articulators.

For example, the second moment of the spectrum correlates to Corners. Other functions are listed below.

$$f_0 = \int_0^H a(h)\, dh$$

$$f_1 = \frac{1}{f_0} \int_0^H a(h)\, h\, dh$$

$$f_2 = \frac{1}{f_0} \int_0^H a(h)\, h^2\, dh$$

$$f_3 = f_2 - f_1^2$$

$$f_8 = \frac{1}{f_3} \int_0^H a(h)\, (h - f_1)^3\, dh$$

$$f_{33} = \frac{1}{f_0} \int_{f_1}^H a(h)\, h\, dh$$

$$f_{34} = \frac{1}{f_0} \int_0^{f_1} a(h)\, h\, dh$$

Jaw correlates to f_{34}. T1x correlates to f_{33}. Corners correlates to f_8, as well as f_1, f_2 and f_{33}. This list is by no means complete.

For the moment functions that correlate to an articulator, we compute the slope and intercept of the line of best fit. I.e.,

articulator = slope * moment + intercept

During program operation, the moment is computed on the signal, and then the line for that moment-articulator correlation is used to get the articulator value. The only part of the model data needed, therefore, is a table of slopes and intercepts.

The general framework described here is rich enough to allow multiple approaches to the ALS problem. For example, a single moment-line can be used, or the articulator values from several moment-lines can be averaged. Furthermore, instead of computing correlation lines, the correlations can be computed for planes (two moments at once). Planes have shown non-trivial improvement over single variable correlation for at least one articulator.

Smoothing.

After articulator values are determined for each GP, continuity of facial movements is enforced. Each articulator sequence is 'smoothed' using quadratic fits over some time interval. The resulting shape stream is used to drive the display.

Experimental results.

Several speakers have been analyzed so far. The first one was Barrett Koster (first author), male. I provided two training sets consisting of pure vowel sounds made separately. High correlations to the functions listed above were observed. In a subsequent animation of the sound "owie", the program produced the correct mouth shapes.

In order to indicate that this is not a single speaker phenomenon, Donna Kelly of CNN was recorded from television. The training set of pure vowels was not available, so samples had to be sliced out manually. Some of the samples were extremely short, but because the program processes sound in single GPs, this was not a problem. Also, the set was incomplete and somewhat inaccurate, as not all of the training sounds could be located in Ms. Kelly's speech.

Nevertheless, the set was analyzed and high correlations were seen in many of the same moment-articulator pairs as for my voice. The indication is that this method is usable on other speakers. Furthermore, minimal or incomplete training is acceptable because the program does not have to identify individual sounds. It only has to verify the trend of the sound-moment correlations over the range of sound.

Extensions.

There are a number of possible extensions to this work. Techniques have been studied to allow speaker-independent operation. Automatic calibration for speaker independence may be possible. If there are functions that correlate to the same articulators for all people (as appears likely in our preliminary data), it may only be a matter of adjusting some ranges when the speaker first starts to talk.

Techniques have also been studied to cover unvoiced sounds and sounds where the mouth shape changes rapidly (for the most part, consonants). We believe consonants should yield to our methodology. Toward this end, several program features are already in place. The GP-synched FFT can get a spectrum from a single pulse, and in particular, from plosive sounds where there *is* only one pulse. The Smoothing is also set up for consonants. Even when the sound is discontinuous, physical position and velocity of mouth parts must be continuous. For

example, "pa" seems discontinuous, with a pop, silence and then a voiced sound, but the jaw opening simply increases from 0. The other articulators will also follow simple trajectories.

Conclusion.

Owie was designed for animation of cartoons and so that the deaf can lip-read telephone and radio. Mouth shape turns out to be a useful domain for representing speech and smoothing out noise. Concentration on acoustics avoids many of the pitfalls of word-based processing and has led so far to animation of vowels.

Bibliography.

[1]. Cannon, Robert H., Jr.. *Dynamics of Physical Systems.* McGraw-Hill, Inc. NY. 1967.

[2]. Exploratorium in San Francisco. Hands-on artificial adjustable see-through voice box.

[3]. Glinski, Stephen Charles. *Diphone Speech Synthesis Based On A Pitch-Adaptive Short-Time Fourier Transform.* PhD Thesis in EE at U of IL at Urbana-Champaign. 1981.

[4]. Greenwald, Audrey B. *Lipreading Made Easy* . Alexander Graham Bell Association for the Deaf. 3417 Volta Place NW, Wash., DC 20007. 1984.

[5]. Halas, John. *Film Animation: a simplified approach.* 1976. United Nations Educational, Scientific and Cultural Organization (UNESCO).

[6]. Kopp, George A., Kopp, Harriet Green, and Angelocci, Angelo. *Visible Speech Manual* . Wayne State U. Press, Detriot MI. 48202. 1967.

[7]. Ladefoged, Peter. *A Course in Phonetics* . Harcourt Brace Jovanovich, Inc. Atlanta. 1975.

[8]. McGrath, Matthew and Summerfield, Quentin. "Intermodal timing relations and audio-visual speech recognition by normal-hearing adults". *JASA: Journal of the Acoustical Society of America.* V77 (2), Feb 1985. pp. 678-685.

[9]. Norton, Michael Peter. *Fundamentals of noise and vibration analysis for engineers.* Cambridge University Press. 1989.

[10]. Rabiner, Lawrence R. and Schafer, Ronald W. *Digital Processing of Speech Signals.* Prentice Hall, Inc. Englewood Cliffs, NJ. 1978 by Bell Labs.

[11]. Slager, Robert P.. "Device could help hearing impaired with telephone talk". *Western News.* Western Michigan University. V18 #18. (30 Jan 1992) p. 1-2.

[12]. Steigerwald, Silvi K. *Development Tools For A Speech Mimicking System.* MS thesis for EE at U of IL at Urbana-Champaign. 1986.