

SPEECH SYNTHESIS FROM HEAD AND FACE MOTION

Adriano V. Barbosa Hani C. Yehia
vilela@cpdee.ufmg.br hani@cpdee.ufmg.br
<http://www.cpdee.ufmg.br/~vilela> <http://www.cpdee.ufmg.br/~hani>
CEFALA – Center for Research on Speech, Acoustics, Language and Music
PPGEE – Graduate Program on Electrical Engineering
UFMG – Universidade Federal de Minas Gerais
Av. Antônio Carlos, 6627, Campus - Pampulha
CEP 31270-901 Belo Horizonte, MG – Brazil

Abstract – This paper describes a method to synthesize speech acoustics from head and face motion. Experiments for simultaneous measurement of head motion, face motion and speech acoustics were carried out for a native American English speaker. Facial motion is characterized by the 3D position of markers placed on the speaker's face and tracked at 60 Hz. Rigid body head motion is represented by the rotations and translations about each coordinate axis. The speech acoustics, in turn, is represented by LSP (Line Spectrum Pairs) coefficients, the fundamental frequency (F0) and the RMS amplitude of the speech signal. These parameters are used to train a linear model, which can then be used to evaluate to which extent speech acoustics is determined from face and head motion. The correlation coefficients between measured and estimated trajectories of the speech acoustics parameters are as high as 0.92 for the RMS amplitude, 0.91 for the fundamental frequency and 0.93 for the LSP coefficients.

I. INTRODUCTION

During speech production, the vocal tract motion shapes not only the speech acoustics but also most of facial motion, through the positioning of the jaw, shaping of the lips and motion of the cheeks [1]. This fact results in the existence of an interrelation among these three quantities (vocal tract motion, facial motion and speech acoustics) and suggests that it is possible, for example, to estimate face motion from the speech acoustics [2], [3], and vice-versa [4]. Another interesting acoustic-motion relation during speech production has been observed between head motion and fundamental frequency (F0) [5], [6].

A system capable of mapping speech acoustics onto face and head motion is important, for instance, in parametric facial animation [7], where the parameters used to control a synthetic face can be obtained directly from the acoustic signal. Such a system can be used in videoconferencing, resulting in very low bit-rates, since only the audio signal needs to be transmitted [8].

This work concentrates on estimating speech acoustics from head and face motion. The results are based on simultaneous measurements of 3D facial deformation, 6D head motion and speech acoustics for a native American English speaker during the production of recited sentences. Face deformation and head motion were measured with a high precision marker tracking system and subsequent rigid body analysis [9]. The speech acoustics is characterized by the fundamental

frequency (F0), the RMS amplitude and the LSP (Line Spectrum Pairs) coefficients, which are largely determined by the vocal tract shape.

The procedure for estimating speech acoustics from head and face motion consists of training linear estimators whose inputs are the 6D head motion and the 3D position of the facial markers, and whose outputs are the LSP coefficients, the fundamental frequency (F0) and the RMS amplitude of the acoustic signal. These estimators are then applied to test data. Finally, the estimator outputs are used to synthesize the acoustic signal of speech.

The results obtained show correlation coefficients between measured and estimated data as high as 0.92 for the RMS amplitude, 0.91 for the fundamental frequency and 0.93 for the LSP coefficients. The results can also be evaluated subjectively by listening to the synthesized speech signal.

Only linear estimators were used in this work. Of course, nonlinear estimators could also be used. For example, they can be implemented by means of neural networks. However, linear estimators have the advantage of being simple to implement and, furthermore, they serve as a reference in comparisons with more elaborate mappings.

II. EXPERIMENTATION

The experiments for data acquisition were carried out for a male, native speaker of English. Face motion was measured through the 3D positions of 18 infra-red markers placed on the cheeks, chin and around the lips of the speaker (Fig. 1). The markers were tracked with an OPTOTRAK at 60 Hz. Additional markers were used to measure the 3D head motion (see [9] for details), which were subsequently represented by 6 degrees of freedom of a rigid body (i.e. translation and rotation around each coordinate axis).

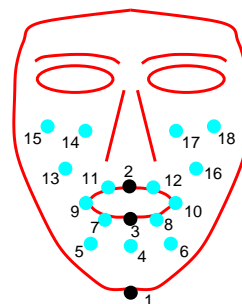


Figure 1. Positions of markers used for face motion measurements.

The spectral acoustics is represented by LSP (Line Spectrum Pairs) coefficients [10]. In the experiments, the speech waveform was acquired at 10 kHz, downsampled to 8040 samples per second, and analyzed using a frame length of 16.67

ms. This yields a rate of 60 frames/s, matching the frame rate in which head and face motion was acquired. LPC (Linear Predictive Coding) analysis of order $P = 10$ was applied to each frame. The LPC coefficients were then converted into LSP coefficients. The use of LSP coefficients is justified by the fact that they are strongly related to the speech formants, which are basically determined by the vocal tract shape. The vocal tract motion, in turn, is the main responsible for the facial motion during speech. Figure 3 shows the LSP coefficients for a particular speech frame.

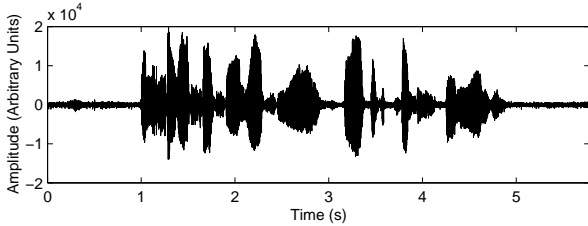


Figure 2. Speech signal.

Besides the LSP coefficients, two other quantities are used to characterize speech acoustics, namely the fundamental frequency (F0) and the RMS amplitude of the speech signal. This is because the LSP coefficients characterize only the spectral envelope of the speech signal; in order to synthesize speech acoustics the fine structure (F0) as well as the energy (RMS amplitude) of the speech signal need to be known.

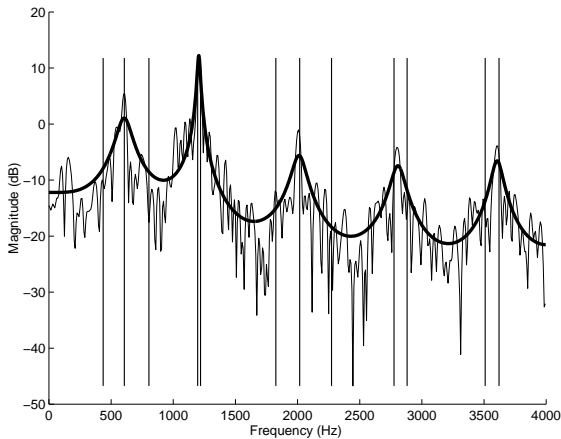


Figure 3. Line Spectrum Pairs (vertical lines) for a speech frame. The envelope is the LPC spectrum.

III. MATHEMATICAL ANALYSIS

In order to perform a mathematical analysis, the data acquired in the experiments are first organized in matricial form as follows: each frame m of face and head motion is represented as a $(3N + 6)$ -dimensional vector in the following way

$$\mathbf{x}_m = [x_{1m} \ x_{2m} \ \dots \ x_{3Nm} \ | \ x_{(3N+1)m} \ \dots \ x_{(3N+6)m}]^t, \quad (1)$$

where $\{x_{1m} \ \dots \ x_{3Nm}\}$ are the 3D positions of the $N = 18$ facial markers and $\{x_{(3N+1)m} \ \dots \ x_{(3N+6)m}\}$ are the head translations and rotations around the coordinate axes. The motion vectors are then grouped in the following matrix

$$\mathbf{X} = [\mathbf{x}_1 \ \mathbf{x}_2 \ \dots \ \mathbf{x}_M], \quad \mathbf{X} \in \mathbb{R}^{(3N+6) \times M} \quad (2)$$

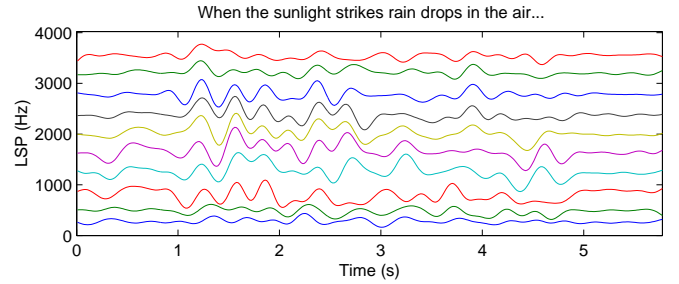


Figure 4. LSP trajectories.

where M is the number of frames. Figure 7 shows the face motion, the head motion, and the combination of face and head motions.

In turn, each frame m of digitized speech (acquired simultaneously with face and head motion) is characterized by $P = 10$ LSP coefficients, the speech fundamental frequency (F0) and the RMS amplitude. Thus, a speech frame can be represented as a $(P + 2)$ -dimensional vector in the following way

$$\mathbf{y}_m = [y_{1m} \ y_{2m} \ \dots \ y_{Pm} \ | \ y_{(P+1)m} \ y_{(P+2)m}]^t, \quad (3)$$

where $\{y_{1m} \ \dots \ y_{Pm}\}$ are the LSP coefficients, $y_{(P+1)m}$ is the fundamental frequency and $y_{(P+2)m}$ is the RMS amplitude of the speech frame. These vectors are then grouped in the following matrix

$$\mathbf{Y} = [\mathbf{y}_1 \ \mathbf{y}_2 \ \dots \ \mathbf{y}_M], \quad \mathbf{Y} \in \mathbb{R}^{(P+2) \times M} \quad (4)$$

Figures 4, 5 and 6 show the LSP trajectories, the fundamental frequency (F0) and the RMS amplitude of the speech signal, respectively.

Matrices \mathbf{X} and \mathbf{Y} are then used to determine a *minimum mean squared error* (MMSE) estimator

$$\mathbf{Y} \approx \mathbf{T} \mathbf{X}, \quad (5)$$

$$\mathbf{T} = \mathbf{Y} \mathbf{X}^t (\mathbf{X} \mathbf{X}^t)^{-1}. \quad (6)$$

The linear estimator \mathbf{T} can be used to estimate a matrix \mathbf{Y} of speech coefficients from a matrix \mathbf{X} of head and face motion coefficients according to Equation 5. The speech coefficients in matrix \mathbf{Y} are then used to synthesize the speech acoustics.

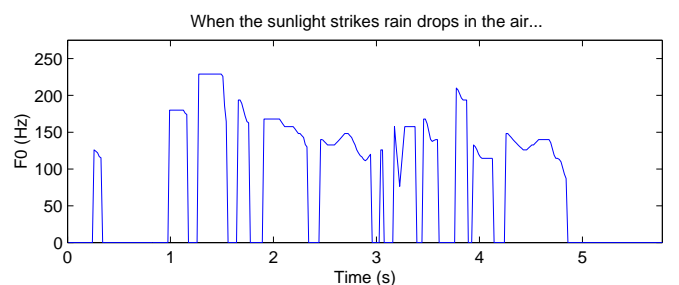


Figure 5. Fundamental frequency (F0) trajectory.

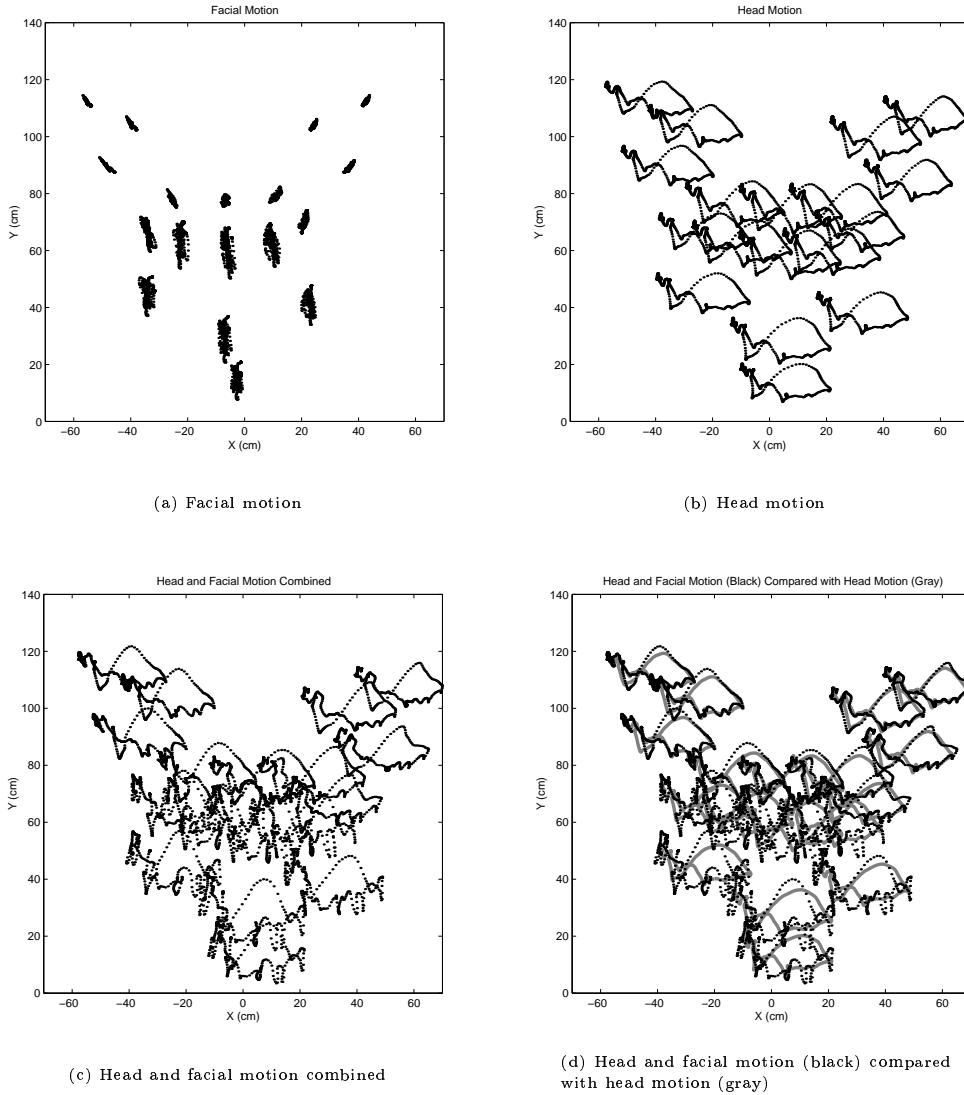


Figure 7. Head and face motion.

IV. RESULTS

The methodology described was applied to a sentence uttered by a native American English speaker. The sentence used in the experiments was “*When the sunlight strikes raindrops in the air, they act like a prism and form a rainbow*”. After motion and speech acoustics parameters have been extracted, they were used to train a linear estimator (Eq. 6).

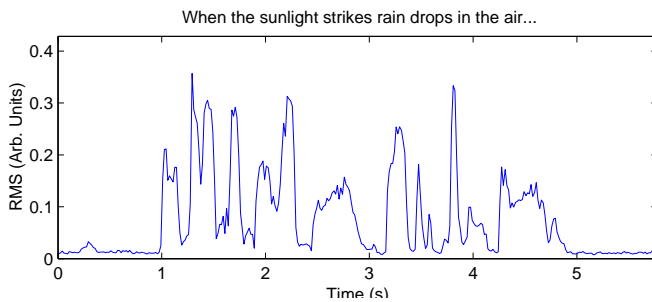


Figure 6. RMS amplitude.

The motion parameters were then fed into the linear estimator, that gives as outputs the LSP coefficients, the fundamental frequency (F0) and the RMS amplitude. Finally, the output parameters provided by the linear estimator were used to synthesize the speech signal.

Table I shows the correlation coefficients between the parameters extracted from the speech acoustics and those provided by the linear estimator. The correlation coefficients for both F0 and RMS are greater than 90%, and those for the LSP coefficients are in the range from 81% up to 93%. Figure 8 shows the time trajectories for the measured and estimated speech acoustics parameters.

We should notice that the data set used in this work is quite small. This does not invalidate the results obtained. However, in order to get a better understanding of the subject, more experiments should be conducted with larger data sets, including speakers of other languages. A larger data set containing more sentences would allow us to study other aspects of the relation between motion and acoustics domains, such as generalization issues.

TABLE I
Correlation coefficients

Amp. RMS	F0	LSP #1	LSP #2	LSP #3	LSP #4	LSP #5	LSP #6	LSP #7	LSP #8	LSP #9	LSP #10
0.9212	0.9125	0.8790	0.8894	0.9030	0.8985	0.9339	0.9142	0.9149	0.9066	0.8144	0.8696

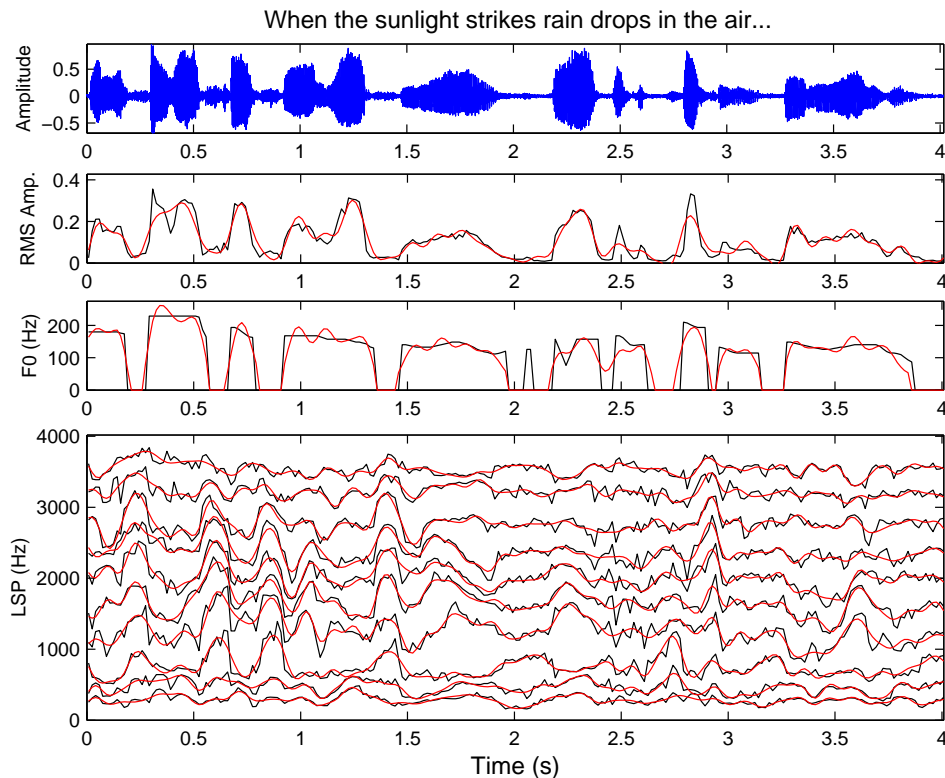


Figure 8. Measured (black lines) and estimated (gray lines) speech acoustics parameters.

V. CONCLUSIONS

In this paper, a method to synthesize speech acoustics from head and face motion was described. Rigid body head motion was represented by the translations and rotations about each coordinate axes; face motion was characterized by the 3D positions of markers placed on the speaker's face and the speech acoustics was represented by LSP coefficients, the fundamental frequency and the RMS amplitude of the speech signal. The data representing the three domains were parameterized and used to train a linear estimator. This estimator was evaluated with the same data used for training, and provided correlation coefficients as high as 0.93. The results can also be subjectively evaluated by listening to the synthesized acoustic signal.

ACKNOWLEDGEMENTS

The authors thank ATR – Information Sciences Division (Kyoto, Japan), in particular, Eric Vatikiotis-Bateson and Takaaki Kuratate.

REFERENCES

- [1] Hani Camille Yehia, P. Rubin, and Eric Vatikiotis-Bateson, "Quantitative association of vocal-tract and facial behavior," *Speech Communication*, vol. 26, pp. 23–43, 1998.
- [2] Hani Camille Yehia, Takaaki Kuratate, and Eric Vatikiotis-Bateson, "Using speech acoustics to drive facial motion," in *14th International Congress of Phonetic Sciences – ICPH'99*, August 1999, vol. 1, pp. 631–634.
- [3] Adriano Vilela Barbosa and Hani Camille Yehia, "Measuring the relation

- between speech acoustics and 2D facial motion," in *26th International Conference on Acoustics, Speech, and Signal Processing – ICASSP'2001*, Salt Lake City, USA, 2001, vol. 1, pp. 181–184.
- [4] Takaaki Kuratate, Hani Camille Yehia, and Eric Vatikiotis-Bateson, "Kinematics-based synthesis of realistic talking faces," in *Proceedings of the International Conference on Auditory-Visual Speech Processing – AVSP'98*, 1998, pp. 185–190.
- [5] Takaaki Kuratate, Kevin G. Munhall, Philip Rubin, Eric Vatikiotis-Bateson, and Hani Camille Yehia, "Audio-visual synthesis of talking faces from speech production correlates," in *6th European Conference on Speech Communication and Technology – EUROSPEECH'99*, September 1999, vol. 3, pp. 1279–1282.
- [6] Hani Camille Yehia, Takaaki Kuratate, and Eric Vatikiotis-Bateson, "Facial animation and head motion driven by speech acoustics," in *IV Speech Production Seminar*, Munich, May 2000.
- [7] Eric Vatikiotis-Bateson, Takaaki Kuratate, M. Kamachi, and Hani Camille Yehia, "Facial deformation parameters for audiovisual synthesis," in *Proceedings of the International Conference on Auditory-Visual Speech Processing – AVSP'99*, August 1999, pp. 118–122.
- [8] Adriano Vilela Barbosa, "Codificação audiovisual integrada da fala," M.S. thesis, Programa de Pós-Graduação em Engenharia Elétrica, UFMG, 2000, in Portuguese.
- [9] Eric Vatikiotis-Bateson and David J. Ostry, "An analysis of the dimensionality of jaw motion in speech," *Journal of Phonetics*, vol. 23, pp. 101–117, 1995.
- [10] N. Sugamura and F. Itakura, "Speech analysis and synthesis methods developed at ECL in NTT - from LPC to LSP," *Speech Communication*, vol. 5, pp. 199–215, 1986.