ASYNCHRONY MEASURE OF LIP-TONGUE-JAW MOVEMENTS

H-H. Bothe [1], C. Mooshammer [2], S. Kuhrt [1], and B. Pompino-Marschall [2]
[1] Technical University of Berlin, Electronics Institute, Berlin, Germany
[2] Forschungsschwerpunkt Allgemeine Sprachwissenschaft, Berlin, Germany

ABSTRACT

This paper describes a method of analyzing timing correlations between characteristic *independent movements* of the lips and the tip of the tongue after model based elimination of jaw movements. The Dynamic Time Warping algorithm was applied to time series of articulographic measurements. The investigations lead to a complex similarity measure for the articulatory processes as a degree of coordination as well as to time series of coproduction data. Future goal is to add realistic tongue movements to an existing facial animation computer program.

DATA ACQUISITION

Text Corpus

For data acquisition, one German speaker produced five repetitions of a nonsense-word corpus of the form /ge-CVC-e/ with C=/p, t, k/ and V=all full vowels of German in the carrier-sentence 'Ich habe ... gesagt.' ('I said ...') in randomized order. In this study, we analysed the subset of sequences with lax /a, œ/ and the consonants /p/ and /t/.

Sensor Mounting

To monitor articulatory movements Electromagnetic Articulography (AG100, Carstens Medizinelektronik; see [1]) was used. One sensor $<c_0>$ was mounted on the lower incisors (jaw), one $<c_1>$ on the lower lip, and one on the midline of the tongue, 1 cm from the tip of the tongue $<c_2>$ (see Figure 1). To compensate for head movements two reference coils were attached to the upper incisors and the bridge of the nose. After measuring the occlusal plane the data were translated and rotated with respect to this new (x,y)-coordinate system. Whereas the x-axes is given by inter-section of occlusal and sagittal plane, the y-axes is scaled othogonal to it. The x-offset of which is defined by the coil position of the upper incisors.

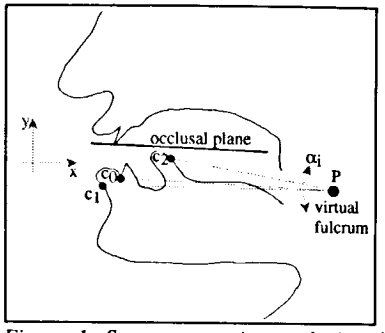

Figure 1. Sensor mounting and virtual fulcrum.

Segmentation of Time Signals

Elimination ofjaw movements and DTW was applied to articulatory time signals, beginning at the onset of the first consonant and ending at the offset of the second consonant (see Figure 2).

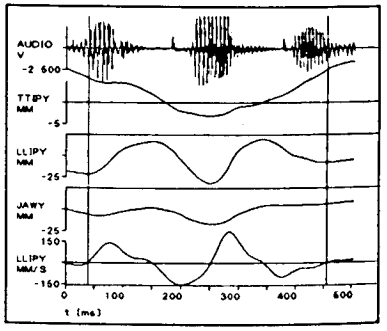

Figure 2. Segmentation of the analysed articulatory time signals for /pap/.

On- and offsets were determined by means of zero-crossings of the vertical velocity signals of the corresponding articulator, i.e. tongue tip coil for apical stop and lower lip for the bilabial stop.

ELIMINATION OF JAW MOVEMENTS

The directly measured cartesian $<c_1>$- and $<c_2>$-coordinates x' and y' were corrected by the corresponding $<c_0>$-rotation around a virtual fulcrum (see Figure 1). Reference was the rest position of the lower incisors $<c_0>_{rest}$.

Goal of the correction is to separate *independent movements* of lips and tip of the tongue from those induced by the jaw motion.

The calculation model of the virtual fulcrum (vf) extends the idea of pure vertical shift and proposes that $<c_0>$ moves approximately on a circle as shown in Figure 3.

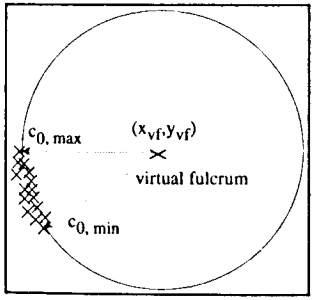

Figure 3. Calculation model of the virtual jaw fulcrum by circle approximation.

The (x_{vf},y_{vf})-location is calculated by curve fitting of sample jaw data (×) by means of a Genetic Algorithm. This iterative optimization method assumes a family of parameter vectors, eachof which composed of the free circle parameters x_{vf} and y_{vf} which are lined up in a bit string and Gray-coded ([2]; Figure 4).

The parameters are optimized by random change with the help of *mutation* (single bit change) and *crossover* (changes of longer parts of the string within a family of parameter sets). The used algorithm is described in [3].

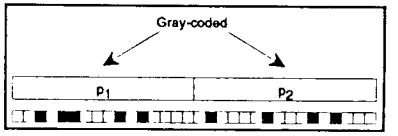

Figure 4. Gray-coded parameter string.

It shows up that the (vf)-position becomes stable at (82.4,-6.1) after a high number of iterations (>15000). The courses of the corrected secondary positions (x_i,y_i) were then compared by applying the DTW algorithm.

DYNAMIC TIME WARPING (DTW) OF 2D TIME DISCRETE SIGNALS

The DTW was applied to two prototypes of the same two-dimensional CVC time series of interest. Two principle courses of discrete (x,y)-positions over time τ with a sample rate of 250 [Hz] are shown in Figure 5.

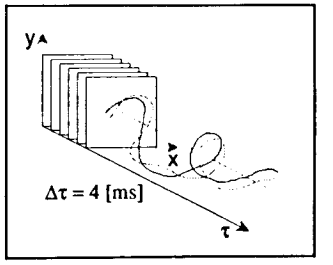

Figure 5. Time series of discrete (x,y)-positions of two sensors over time τ.

Applying DTW to a time discrete reference curve $C_1(\tau)$ and a test curve $C_2(\tau)$ results in a nonlinear projection between them together with a global distance measure of similarity. The nonlinear stretching is necessary with respect to different speed of the articulatory movements and different lengths of the phonemes. The principle of the algorithm is shown in Figure 6 (see also [4]).

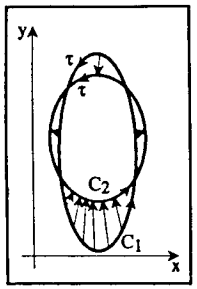

Figure 6. Nonlinear mapping of two sample curves $C_1(\tau)$ on $C_2(\tau)$ in (x,y) head coordinates over time τ.

In order to find the optimum projection at first a matrix **d** of local distances $d(\tau_i,\tau_j)$ between all C_1 (τ_i) and $C_2(\tau_j)$ has to be calculated. Applying the Euclidian distance measure $d_E(\tau_i,\tau_j)$ results in

$$d_E(\tau_i,\tau_j)^2=[x_1(\tau_i)-x_2(\tau_j)]^2+[y_1(\tau_j)-y_2(\tau_j)]^2.$$

The basic idea behind DTW is to find the path from the starting point $\tau_0=0$ to the end point τ_I on which the accumulated local distances become a minimum.

The sum $D(\tau_i,\tau_j)$ of the distances can be calculated by the recursive formula

$$D(\tau_i,\tau_j) = d(\tau_i,\tau_j) + \min [D(\tau_{i-1},\tau_{j-1}), D(\tau_{i-1},\tau_j), D(\tau_i,\tau_{j-1})]$$

as the sum of the local distance $d(\tau_i,\tau_j)$ and the minimum of the accumulated distances

$D(\tau_{i-1},\tau_{j-1})$, $D(\tau_{i-1},\tau_j)$ and $D(\tau_i,\tau_{j-1})$.

Solving the above equation requires a column oriented calculation in the three-dimensional (x,y,τ)-universe.

For the time being we are interested only in the vertical articulatory movements $y_{lips,\ tongue}(\tau)$ in order to position a two-dimensional bit pattern of the tongue in the later computer animation.

The calculation scheme for the o-path is shown in Figure 7.

The value $D(\tau_i,\tau_j)$ is interpreted as a measure of global similarity.

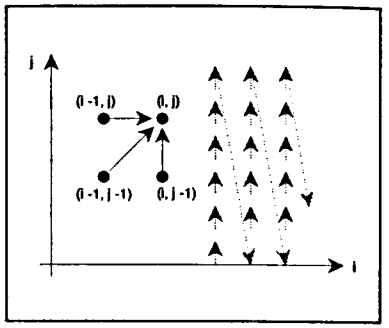

Figure 7: Column oriented calculation of the accumulated distance $D(\tau_i,\tau_j)$.

The projection of $C_2(\tau)$ on $C_1(\tau)$ is related to the minimum path calculated from the end point to the starting point of the matrix of accumulated distances **D** (a principal example of a one-dimensional mapping of $y_1(\tau)$ and $y_2(\tau)$ is shown in Figure 8).

In order to reduce the calculation time, a desired area of interest is taken into account around the diagonal of **D**.

The nonlinear stretching coefficients $d_{OP}(\tau_i,\tau_j)$ of the optimum path (o-path) represent the local similarities of the projection of $C_2(x,y,\tau)$ on $C_1(x,y,\tau)$.

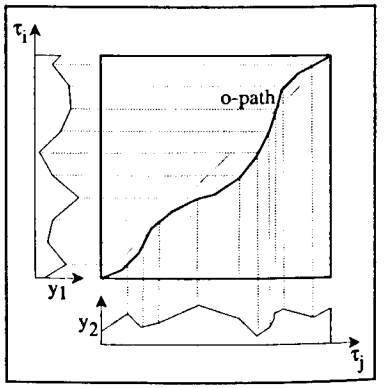

Figure 8. Principal of 1D- DTW mapping.

Restricting these coefficients to relatively small ε-values by $d_{OP}(\tau_i,\tau_j) < \epsilon$ can be interpreted as fixing the zones of similarity.

Some Exemplary Results

The DTW results in a diagonal either for total similarity of both curves or for uncorreleted curves.

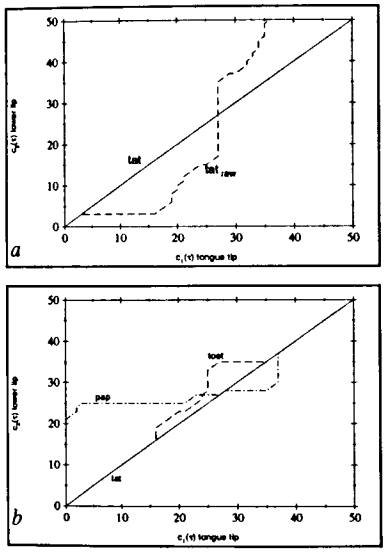

Figure 9. DTW path for a. /tat/raw, /tat/, b. corrected curves /tat/, /pap/, /tøt/.

Figure 9a shows that for uncorrected $/tat/_{raw}$ the time intervall $\Delta\tau=[16, 35]$ in msec of the tongue tip is mapped on [4, 50] of the lower lip, whereas after correction the movements are similar. DTW curve /pap/ in Figure 9b shows that [0, 40] of the tongue tip is mapped on a nearly constant lower lip position at ca. 25, and thus, determines highly independent movement during this time. The course of similarity for /tøt/ can be interpreted as an achievment of lip rounding.

Summary and Conclusion

The above work shows basic investigations for the design of a model based computer animation program that displays visual articulation movements with moving lips, teeth and tongue tip on the computer screen.

Whereas in the existing program the grey-scale film is created by a codebook of key-pictures and a morphing algo-rithm for lips, skin, and teeth [5], the above investigations make possible the implementation of coordinated movements of the tongue.

Acknowledgement

The text corpus has been designed and recorded within a DFG project (DFG = German Research Council) under grant Ti 69/29-2. Many thanks for help especially to Phil Hoole.

References

[1] Perkell, J.S., Cohen, M.H., Svirsky, M.A., Matthies, M.L., Garabieta, I., Jackson, M.T.T. (1992), "Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements", *J. Ac. Soc. of America*, vol. 92, pp. 3078-3096.

[2] Mathias, K.E., Whitley, L.D., *Transforming the search space with Gray-coding. Proc. 1st IEEE Conf. on Evolutionary Computation*, pp. 513-518, Orlando, USA.

[3] Thierens, D., Goldberg, D. (1994), *Elitist Recombination: An Integrated Selection Recombination GA*. Proc. 1st IEEE Conf. on Evolutionary Computation, pp. 508-512, Orlando, USA.

[4] Bothe, H.H., Rieger, F., Tackmann, R. (1993), *Visual coarticulation effects in syllable environment*, Proc. EUROSPEECH '93, pp. 1741-1744, Berlin, Germany.

[5] Bothe, H.H., Rieger, F. (1993), Visual speech and coarticulation effects. Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP) '93, pp. V634-V637, Minneapolis, USA.