19th INTERNATIONAL CONGRESS ON ACOUSTICS
MADRID, 2-7 SEPTEMBER 2007

SURFACE MODELING OF VOCAL TRACT SHAPES IN TRANSITION SEGMENTS OF VOWEL-CONSONANT-VOWEL SYLLABLES FOR ESTIMATION OF PLACE OF CLOSURE

PACS: 43.72.Ct

Shah, Milind S.; Pandey, Prem C.
EE Dept, Indian Institute of Technology Bombay; Powai Mumbai 400 076, India;
{milind, pcpandey}@ee.iitb.ac.in

ABSTRACT
Production of vowel-stop consonant-vowel syllables involves movement of articulators from the articulatory position of the vowel towards that of the stop closure to that of the vowel. Movement of articulators before and after a stop closure is characterized by formant transitions. The closure portion of a stop have zero or low signal energy and relevant spectral information is not available. Hence, LPC based estimation of vocal tract shape fails during stop closure. A technique has been investigated for estimation of place of stop closure by surface modeling of estimated area values during vowel-consonant and consonant-vowel transition segments, on the assumption that articulatory movements have low order dynamics. Modeling of area values was based on least-squares conic and cubic bivariate polynomials and Delaunay triangulation based surfaces. The technique was applied for estimation of place of closure for bilabial, alveolar, and velar stops, both unvoiced and voiced, in syllables of the type /aCa/ spoken by five speakers. Results were more consistent with conic surface based interpolation, than that of cubic surface and Delaunay triangulation surface based interpolation. The proposed technique can be used for improving effectiveness of speech-training systems for production of stop consonants by providing visual feedback of place of closure.

INTRODUCTION

A good knowledge of the articulatory-acoustic relationship is useful for better understanding of articulation, speech synthesis and recognition, diagnosis of speech disorders, and speech-training for the hearing impaired persons. Shape of the vocal tract can be specified by its cross-sectional area as a function of position along the tract length [1]. Estimation of the vocal tract area function from speech signal, an inverse problem, can be carried out by one of the several techniques: LPC analysis [2], use of formants and factor analysis [3], use of formants and perturbation theory [4], mapping via articulatory codebook [1], etc. Most of these techniques are reported to work satisfactorily for vowels. However, shape estimation fails if spectral information is not available, for example during stop closure duration [1], [2], [5]. Speech-training systems providing visual feedback of articulatory efforts, not visible from outside, are reported to be useful for improving vowel articulation [6] - [10]. Most of these systems are based on visual feedback of vocal tract shape estimated by LPC analysis of speech. However, similar systems for speech training are not available for consonant articulation. Hence, for improving the effectiveness of speech-training systems, it is important to investigate technique for vocal tract shape estimation for consonant articulation.

This paper presents a technique for estimation of place of constriction during stop closures of vowel-stop consonant-vowel (VCV) syllables. Production of VCV syllables involves movement of articulators from the articulatory position of the vowel towards that of the stop closure to that of the vowel. Movement of articulators before and after a stop closure is characterized by transitions in vocal tract shapes as well as formants. The proposed technique is based on surface modeling of estimated area values during vowel-consonant (VC) and consonant-vowel (CV) transition segments, and its two-dimensional (2D) interpolation during stop closure for estimation of place of constriction.

ESTIMATION OF LPC BASED VOCAL TRACT AREA VALUES

Investigations for vocal tract shape estimation were based on Wakita's speech analysis model [2] with reflection coefficients obtained from LPC analysis of speech. LPC based vocal tract shape estimation does not involve automated tracking of formants and is suitable for real-time processing. Hence, despite its several limitations, it is used for developing speech training aids.

For checking the consistency of vocal tract shape estimation, vowels /a/, /i/, and /u/ were synthesized as well as recorded, and analyzed with different combinations of parameters involved in the algorithm: analysis window size and shift, speech sampling frequency, and LPC order. It was observed that vocal tract shapes obtained with analysis window size equal to twice the average pitch period, window shift equal to integer multiple of the pitch period, and LPC order 12 for sampling rate of 10–12 kHz were consistent and realistic. In order to study the consistency of the shape estimation with amplitude and pitch variation in vowels, and to study dynamics of shape estimation during transitions at VC boundaries, we have used *areagram*, a spectrogram-like 2D display of square-root of cubic-spline interpolated vocal tract area values plotted as grey levels as a function of time along x–axis and glottis-to-lips (G-L) distance along y–axis. In this display, each new vertical frame corresponds to shifting the analysis window by 5 ms (~ 55 samples, for F_s = 11.025 kHz).

LPC based vocal tract shape estimation for vowels were found to be independent of pitch variation and amplitude variation (over an attenuation range of 0–40 dB) [5]. For verifying shape tracking ability of LPC based estimation, vowel-semivowel-vowel syllables were analyzed. Areagram results showed proper transition in vocal tract shapes and correct estimation of place of articulation for semivowels. Next, VCV syllables involving stop consonants were analyzed. Estimated place of articulation was proper for vowel segments, but area estimates were random and unrelated to place of constriction during stop closure due to zero or very low signal energy and unavailability of relevant spectral information. However, for various VCV syllables, area values and spectrogram display during VC and CV transition segments were distinctly different. It indicated that transition in estimated area values during VC and CV segments may contain information related to place of articulation for stop consonant. Hence, investigations for estimation of place of closure were carried out based on surface modeling of area values during transition segments.

SURFACE MODELING FOR ESTIMATION OF PLACE OF CLOSURE

Production of VCV syllable involves dynamic variation in vocal tract shape before and after stop closure. A technique is proposed and investigated for estimation of place of stop closure by surface modeling of estimated area values during VC and CV transition segments, on the assumption that articulatory movements have low order dynamics. Modeling of area values were based on least-squares conic and cubic bivariate polynomials [11], [12] and Delaunay triangulation [13], [14] based surfaces. The place of constriction was estimated by carrying out 2D interpolation of the surfaces during closure duration.

Least-squares bivariate polynomial surface modeling and interpolation

The aim of least-squares bivariate polynomial modeling was to obtain a surface that models estimated area values during transition segments in the least-squares sense. Investigations were restricted to second order (conic) and third order (cubic) polynomials. The conic bivariate polynomial modeling is given by

$$f(x,y) = c_0 + c_1 x + c_2 y + c_3 xy + c_4 x^2 + c_5 y^2 \qquad \text{(Eq. 1)}$$

where $f(x, y)$ is the estimated area at analysis frame 'x' (along time axis) and lip-glottis distance 'y', and c_0-c_5 are the conic polynomial coefficients. The cubic bivariate polynomial modeling is given by

$$f(x,y) = d_0 + d_1 x + d_2 x^2 + d_3 x^3 + d_4 y + d_5 y^2 + d_6 y^3 + d_7 xy + d_8 x^2 y + d_9 xy^2 \qquad \text{(Eq. 2)}$$

where d_0-d_9 are the cubic polynomial coefficients. These equations for a set of q points, with q > 6 for conic and q > 10 for cubic polynomial approximation, result in over-determined system of simultaneous linear equations expressed in matrix notation as

$$\mathbf{Az} = \mathbf{B} \qquad \text{(Eq. 3)}$$

where

$$\mathbf{B}^T = \left[f(x_0, y_0) \quad f(x_1, y_1) \quad \ldots \quad f(x_{q-1}, y_{q-1}) \right]. \qquad \text{(Eq. 4)}$$

For conic polynomial approximation, $\mathbf{A}$ and $\mathbf{z}^T$ are given by

$$\mathbf{A} = \begin{bmatrix} 1 & x_0 & y_0 & x_0 y_0 & x_0^2 & y_0^2 \\ 1 & x_1 & y_1 & x_1 y_1 & x_1^2 & y_1^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{q-1} & y_{q-1} & x_{q-1} y_{q-1} & x_{q-1}^2 & y_{q-1}^2 \end{bmatrix} \qquad \text{(Eq. 5)}$$

$$\mathbf{z}^T = \begin{bmatrix} c_0 & c_1 & c_2 & c_3 & c_4 & c_5 \end{bmatrix}. \qquad \text{(Eq. 6)}$$

For cubic polynomial approximation, $\mathbf{A}$ and $\mathbf{z}^T$ are given by

$$\mathbf{A} = \begin{bmatrix} 1 & x_0 & x_0^2 & x_0^3 & y_0 & y_0^2 & y_0^3 & x_0 y_0 & x_0^2 y_0 & x_0 y_0^2 \\ 1 & x_1 & x_1^2 & x_1^3 & y_1 & y_1^2 & y_1^3 & x_1 y_1 & x_1^2 y_1 & x_1 y_1^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{q-1} & x_{q-1}^2 & x_{q-1}^3 & y_{q-1} & y_{q-1}^2 & y_{q-1}^3 & x_{q-1} y_{q-1} & x_{q-1}^2 y_{q-1} & x_{q-1} y_{q-1}^2 \end{bmatrix} \qquad \text{(Eq. 7)}$$

$$\mathbf{z}^T = \begin{bmatrix} d_0 & d_1 & d_2 & d_3 & d_4 & d_5 & d_6 & d_7 & d_8 & d_9 \end{bmatrix}. \qquad \text{(Eq. 8)}$$

The matrix $\mathbf{B}$ consists of selected area values along x (time) and y (lip-to-glottis) axes, in VC and CV transition regions, as shown in Fig. 1. The surface modeling was carried out starting from area values at the top of Fig. 1 (i.e., lip end) and gradually moving down towards glottis end. The number of frames to the left and right of stop closure, used for surface modeling, is represented by L_{col} and R_{col} respectively. In order to evaluate the unknown polynomial coefficients in Eqs. 1 and 2, we need to have $j = n_2 - n_1 \geq 2$ for conic and 3 for cubic polynomials respectively. As the 12–section values are plotted along y–axis, we get $j \leq n_2 \leq 12$. The polynomial coefficient matrix $\mathbf{z}$ is obtained from Eq. 3 by

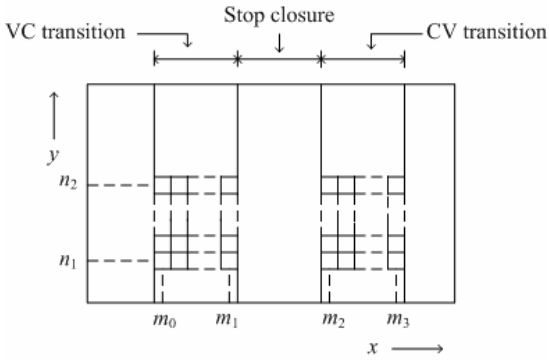

Figure 1.- Selection of area values during transition segments for 2D surface modeling.

$$\mathbf{z} = \left(\mathbf{A}^T \mathbf{A} \right)^{-1} \mathbf{A}^T \mathbf{B} \qquad \text{(Eq. 9)}$$

where matrix $(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T$ is the pseudo-inverse of $\mathbf{A}$ [15]. The computed conic and cubic polynomial coefficients give the bivariate surfaces which approximate area values during VC and CV transition regions. Two-dimensional interpolation of these surfaces during stop closure duration will result in estimation of area values during stop closure. For $m_1 < x < m_2$, $y = n_2$, and $\mathbf{z}$ as obtained from Eq. 9, the 2D interpolation of area values during stop closure frames, along the glottis-to-lip position $y = n_2$ can be performed using

$$\hat{\mathbf{B}} = \hat{\mathbf{A}}\,\mathbf{z}\,. \qquad \text{(Eq. 10)}$$

Interpolation is carried out for $y = n_2$ (and not for $y = n_1$ or any other intermediate value) because this will result in a proper estimation of area values towards the lips region and a slightly improper estimation of area values near the glottis region, which is acceptable. For $n_2 = j$, $\hat{\mathbf{B}}$ in Eq. 10 is evaluated by varying y over the range $1 \leq y \leq j-1$, thus giving the interpolated area values along the first $j-1$ glottis-to-lips positions.

Delaunay triangulation based surface modeling and interpolation

In addition to bivariate polynomial surface approximation, the estimated area values were modeled by the surfaces based on Delaunay triangulation method. Given a set of data points, Delaunay triangulation is a set of lines connecting each point to its natural neighbor. Delaunay triangulation generates triangles having good aspect ratio, with scattered data points as vertices, and is useful for surface interpolation. Use of Delaunay triangulation is particularly suited when one does not want to force any constraints on the set of data points. Also, it

minimizes the number of spikes in the generated surface and satisfies the following two properties [16]:
(a) No data points are contained within a circle circumscribing the triangles;
(b) The smallest angle over all triangulation is maximized thus avoiding triangles with small angles, which is useful for finite element analysis technique.

Similar to the least-squares approximation method, $f(x, y)$ represents the estimated area value at analysis frame 'x' (along time axis) and lip-glottis distance 'y'. Based on Delaunay triangulation method, surface fitting of the estimated area values during VC and CV transition segments was performed. Finally, place of constriction was estimated by 2D interpolation of these surfaces during stop closure of VCV syllables. Delaunay triangulation of given set of area values and surface interpolation was carried out using in-built functions in Matlab®.

Location of stop closures

For bivariate surface generation and interpolation, the stop closure boundary locations need to be known. These locations were estimated using a two step process; estimation of the beginning and ending points of VCV syllables, followed by the estimation of stop closure boundary locations within a VCV syllable. These estimations were based on computation of short-time average magnitude and use of empirically selected threshold values [17].

Validation of the proposed technique

Vowel-semivowel-vowel syllables represent a dynamically varying vocal tract shape with known transition in place of articulation from vowel towards semivowel to that of the vowel. Recovery of place of articulation during artificially silenced central speech segment of different durations, based on partially available area values during vowel-to-semivowel and semivowel-to-vowel transitions, was used as the first step in validation of the proposed technique. Vowel-semivowel-vowel syllables /aja/ and /awa/ were recorded for three male and two female speakers and analyzed for proper recovery of place of articulation during artificially silenced speech segments.

Figure 2 shows analysis results for /aja/ for a female speaker for one of the cases in which a middle segment of 155 ms was artificially silenced and approximately 30 ms of VC and CV transition segments were available for surface modeling. Parts (a), (b), and (c) of Fig. 2 show speech waveform, spectrogram, and areagram respectively while parts (d), (e), and (f) show areagrams obtained after 2D interpolation of conic, cubic, and Delaunay triangulation based surfaces respectively. Part (g) of the figure shows areagram for original /aja/, i.e., without any artificially introduced silence gap. Estimated end-points and silence interval boundary locations are indicated by downward arrows

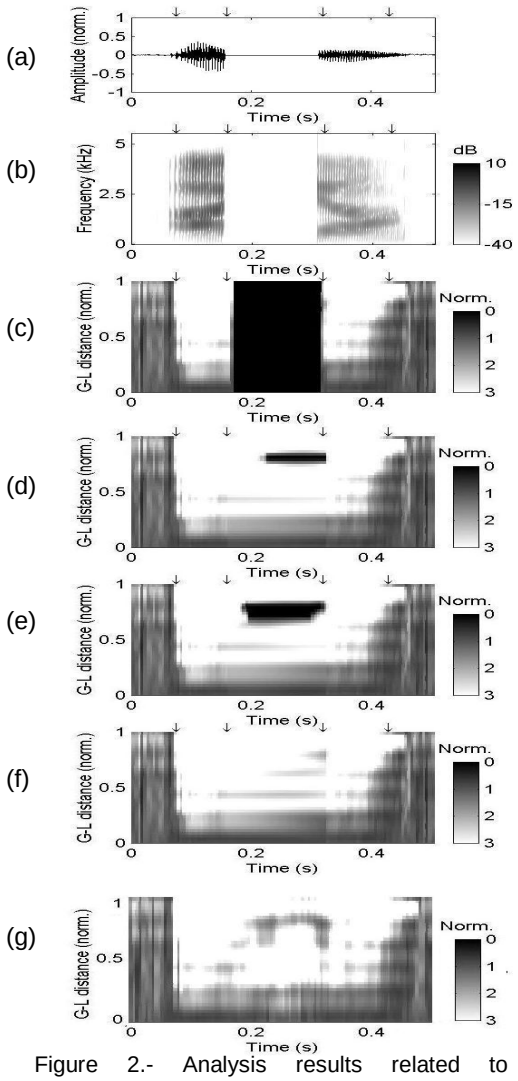

Figure 2.- Analysis results related to 2D interpolation of area values for /aja/ (silence interval = 155 ms): (a) waveform; (b) spectrogram (Δf = 300 Hz); (c) areagram; (d), (e), and (f) areagrams obtained after 2D interpolation of conic, cubic, Delaunay triangulation based surfaces respectively (surface generation parameters j = 3, L_{col} = 6, and R_{col} = 6); (g) areagram for original /aja/ without any silence gap.

along the upper side of each part of the figure. Outer pair of arrows shows end-points, while inner pair of arrows shows silence-interval boundary locations.

After comparing 2D interpolation results with the known place of articulation for palatal /j/ as shown in part (g) of Fig. 2, it is observed that conic and cubic surface modeling of area values, based on minimum available VC and CV transition segments of 30 ms, and its 2D interpolation during silence interval is capable of estimating proper place of articulation. Estimation of place of articulation based on Delaunay triangulation was found to be proper for other cases where minimum required transition segments for surface modeling were of duration more than 30 ms. Analysis of /aja/ and /awa/ for all the five speakers showed that conic surface based modeling of area values was more consistent than cubic and Delaunay triangulation based surfaces in estimating proper place of articulation.

RESULTS AND DISCUSSION

Vowel-consonant-vowel syllables of the type /aCa/ with stop consonants /p/, /b/, /t/, /d/, /k/, and /g/ were recorded for three male and two female speakers and analyzed for estimation of the vocal tract shape during stop closure. Estimated place of constriction for stop consonants was compared against the values reported earlier. From the available data based on MRI [18] and X-ray images [19] values of place of constriction on the normalized scale of 0 to 1 (0 corresponds to glottis position and 1 corresponds to lip position) for bilabial, alveolar, and velar stops are 1.0, 0.75–0.89, and 0.47–0.7 res–pectively. These data were used for validation of the estimated place of stop closures.

Interpolation results for VCV syllable /aga/ based on conic, cubic, and Delaunay triangulation based surface approximation of area values are shown in Fig. 3. Parts (a), (b), and (c) of the figure show speech waveform, wideband spectrogram, and original areagram respectively. Parts (d), (e), and (f) show areagram results obtained after performing 2D interpolation of conic, cubic, and Delaunay triangulation based surfaces respectively. It is observed that 2D interpolation of conic, cubic, and Delaunay surfaces show less area values around the normalized glottis-to-lip distance of 0.55, which corresponds to the place of constriction for velar stops.

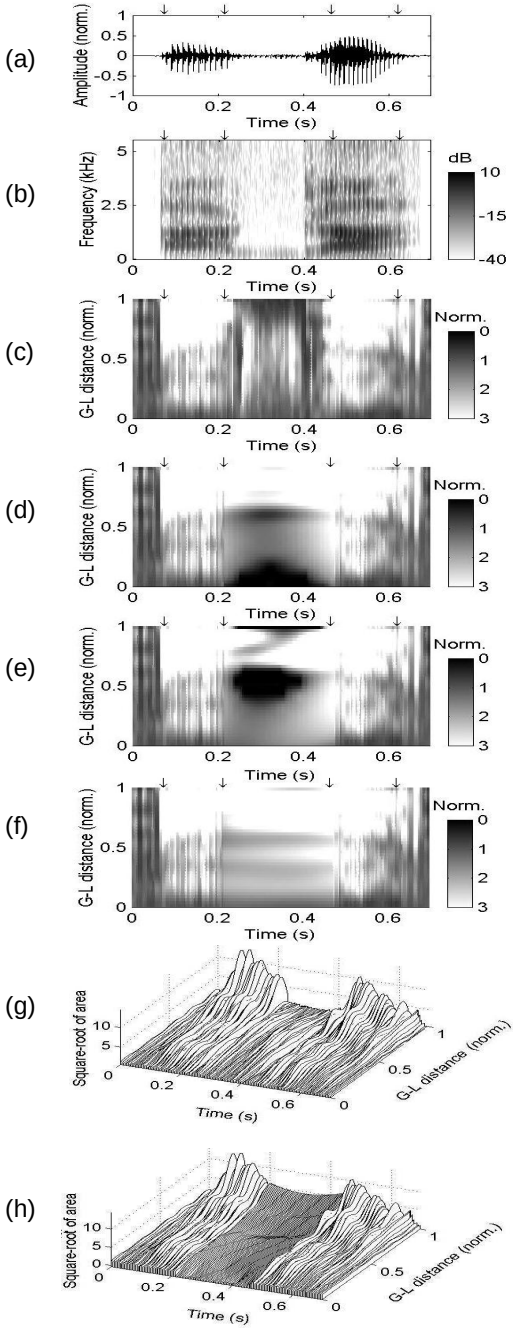

Figure 3.- Analysis results related to 2D interpolation of area values for /aga/: (a) waveform; (b) spectrogram (Δf = 300 Hz); (c) original areagram; (d), (e), and (f) areagrams obtained after 2D interpolation of conic, cubic, and Delaunay triangulation based surfaces respectively (surface generation parameters j = 7, L_{col} = 5, and R_{col} = 3); (g) and (h) waterfall diagrams corresponding to areagrams shown in part (c) and (d) of this figure.

From the analysis results for VCV syllables of the type /aCa/ with the six stop consonants across all the five speakers, it was observed that estimation of place of constriction for bilabial, alveolar, and velar stop consonants, both voiced and unvoiced, was more consistent with conic surface modeling of area values than cubic and Delaunay triangulation based surfaces. This is in conformity with observations during initial validation of the technique with artificially introduced silence gaps in semivowels. Hence, it can be said that conic polynomial is better suited for modeling of articulatory movements. For an alternate visualization, the temporal variations in the vocal tract shapes were plotted as waterfall diagram. These patterns corresponding to the original areagram (Fig. 3 (c)) and the areagram obtained by conic surface based interpolation (Fig. 3 (d)) are shown in parts (g) and (h) of Fig. 3.

CONCLUSIONS

Conic surface based modeling of vocal tract area values during VC and CV transition segments of VCV syllables of the type /aCa/, and its 2D interpolation during stop closure, can be used for estimating the place of closure for bilabial, alveolar, and velar stops, both voiced and unvoiced. The proposed technique may be used for development of speech-training aid for consonant articulation, providing dynamic display of vocal tract shape and place of constriction.

References

[1] J. Schroeter, M. M. Sondhi: Techniques for estimating vocal-tract shapes from the speech signal. IEEE Transactions on Speech and Audio Processing **2, No. 1** (1994) 133–150.

[2] H. Wakita: Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms. IEEE Transactions on Audio Electroacoustics **21, No. 5** (1973) 417–427.

[3] P. Ladefoged, R. Harshman, L. Goldstein, L. Rice: Generating vocal tract shapes from formant frequencies. Journal of the Acoustical Society of America **64, No. 4** (1978) 1027–1035.

[4] Z. Yu, P. C. Ching: Determination of vocal-tract shapes from formant frequencies based on perturbation theory and interpolation method. Proceedings of the IEEE International Conference on Acoustic, Speech, and Signal Processing (1996) 369–372.

[5] M. S. Shah, P. C. Pandey: Estimation of vocal tract shape for VCV syllables for a speech training aid. Proceedings of the IEEE International Conference on Engineering in Medicine and Biology Society (2005) 6642–6645.

[6] R. G. Crichton, F. Fallside: Linear prediction model of speech production with applications to deaf speech training. IEE Control and Sciences **121** (1974) 865–873.

[7] M. Shigenaga, H. Kubo: Speech training system for handicapped children using vocal tract lateral shapes. Proceedings of the IEEE International Conference on Acoustic, Speech, and Signal Processing (1986) 637–640.

[8] N. D. Black: Application of vocal tract shapes to vowel production. Proceedings of the IEEE International Conference on Engineering in Medicine and Biology Society (1988) 1535–1536.

[9] S. H. Park, D. J. Kim, J. H. Lee, T. S. Yoon: Integrated speech training system for hearing impaired. IEEE Transactions on Rehabilitation Engineering **2, No. 4** (1994) 189–196, 1994.

[10] P. M. T. de Oliveira, M. N. Souza: Speech aid for the deaf based on a representation of the vocal tract: the vowel module. Proceedings of the IEEE International Conference on Engineering in Medicine and Biology Society (1997) 1757–1759.

[11] G. M. Philips: Interpolation and Approximation by Polynomials. New York: Springer-Verlag, 2003.

[12] V. Pratt: Direct least-squares fitting of algebraic surfaces. Computer Graphics **21, No. 4** (1987) 145–152.

[13] J. O'Rourke: Computational Geometry in C. Cambridge: Cambridge University Press, 1994.

[14] T. Lambert: An optimal algorithm for realizing a Delaunay triangulation. Information Processing Letters **62** (1997) 245–250.

[15] H. W. Brinkmann, E. A. Klotz: Linear Algebra and Analytic Geometry. Reading, Massachusetts: Addison-Wesley, 1971.

[16] O. R. Musin: Properties of Delaunay triangulation. Proceedings of the 13th Annual ACM Symposium on Computational Geometry (Nice, France) (1997) 424–426.

[17] L. R. Rabiner, M. R. Sambur: An algorithm for determining the endpoints of isolated utterances. The Bell System Technical Journal **54, No. 2** (1975) 297–315.

[18] B. H. Story, I. R. Titze, E. A. Hoffman: Vocal tract area functions from magnetic resonance imaging. Journal of the Acoustical Society of America **100, No. 1** (1996) 537–554.

[19] J. L. Flanagan: Speech Analysis, Synthesis, and Perception, 2nd ed. New York: Springer-Verlag, 1975.