SPEECH INVERSE FILTERING BY SIMULATED ANNEALING ALGORITHM

Chang-Shiann Wu
Department of Information Management
Shih Chien University
Kaohsiung, Taiwan, R.O.C.
Email: chwu4142@ms5.hinet.net

Yu-Fu Hsieh
Chung Shan Institute of Science and Technology
Lung Tan
Tao Yuan, Taiwan, R.O.C.
Email: aliceufo@ms1.hinet.net

ABSTRACT

The purpose of this study is to develop one solution to the speech inverse filtering problem. A new efficient articulatory speech analysis scheme, identifying the articulatory parameters from the acoustic speech waveforms, was induced. The algorithm is known as simulated annealing, which is constrained to avoid non-unique solutions and local minima problems. The constraints are determined by the articulatory-to-acoustic transformation function and the boundary conditions for the articulatory parameters. The cost function is defined as a percentage of the weighted least-absolute-value error distance between the first four formant frequencies of the articulatory model and the first four formant frequencies determined from speech analysis. It is used to optimize the vocal tract parameters to match a specified set of formant characteristics. A 1% error criterion was found to be both practical and achievable.

1. INTRODUCTION

Articulatory synthesis is the production of speech sounds using a model of the vocal tract, which directly or indirectly simulates the movements of the speech articulators. It provides a means for gaining an understanding of speech production and for studying phonetics. Articulatory synthesis usually consists of two separate components. In the articulatory model, the vocal tract is divided into many small sections and the corresponding cross-sectional areas are used as parameters to represent the vocal tract characteristics. In the acoustic model, each cross-sectional area is approximated by an electrical analog transmission line to simulate the speech sound propagation through the vocal system as well as the physics of the physiological-to-acoustic transformation.

To simulate the movement of the vocal tract, the area functions must change with time. Each sound is designated in terms of a target configuration and the movement of the vocal tract is specified by a separate fast or slow motion of the articulators. The recovery of articulatory movements from the speech signal, known as the speech inverse filtering problem, is difficult due to the non-uniqueness of the solution. This problem has been the subject of research for several applications, including articulatory synthesis, speech recognition, low-bit-rate speech coding, and text-to-speech synthesis. Here we attempt a new solution using the simulated annealing algorithm, which is a "constrained multidimensional nonlinear optimization problem." The coordinates of the jaw, tongue body, tongue tip, lips, velum, and hyoid compose the multidimensional articulatory vector. A comparison between the model-derived and the target-frame first four formant frequencies forms the cost function. There are two constraints: (1) the articulatory-to-acoustic transformation function, and (2) the boundary conditions for the articulatory parameters. The optimum articulatory vector is obtained by finding the minimum cost function. Once the optimum articulatory vector is determined, the articulatory model determines the vocal tract cross-sectional area function which in turn is used by the articulatory speech synthesizer.

2. IMPLEMENTATION OF THE ARTICULATORY MODEL

Geometric data concerning the vocal tract is essential to our understanding of articulation, and is a key factor in speech production. According to the acoustic theory of speech production [6], the human vocal tract can be modeled as an acoustic tube with nonuniform and time-varying cross-sections. It modulates the excitation source to produce various linguistic sounds. The success of articulatory modeling depends to a large extent on the accuracy with which the vocal tract cross-sectional area function can be specified for a particular utterance. Measurement of the vocal tract geometry is difficult. Basically, there are two methods for obtaining the vocal tract cross-sectional area function. Direct measurements of the vocal tract have been made from lateral X-ray images. Unfortunately, these direct measurements and their evaluations are laborious. Magnetic resonance imaging (MRI), which is free from the disadvantages associated with X-ray methods, might appear to be the best available method to directly collect the necessary data.

On the other hand, several researchers have proposed analytical methods to derive the vocal tract cross-sectional area function from acoustic data. Articulatory models can be classified into two major types: parametric area model and midsagittal distance model. The parametric area model describes the area function as a function of distance along the tract, subject to some constraints. The area of the vocal tract is usually represented by a continuous function such as a hyperbola, a parabola, or a sinusoid [9].

The midsagittal distance model describes the speech organ movements in a midsagittal plane and specifies the position of articulatory parameters to represent the vocal tract shape. Coker and Fujimura (1966) introduced an articulatory model with parameters assigned to the tongue body, tongue tip, and velum. Later this model was modified to control the movements of the articulators by rules [4]. Another articulatory model was designed by Mermelstein [11]. His model can be adjusted to match the midsagittal X-ray tracings accurately. Our articulatory model is a modified version of Mermelstein's model. A set of variables is used to specify the inferior outline of the vocal tract in the midsagittal plane (Figure 1). These variables, called articulatory parameters, are the tongue body center, the tongue tip, jaw, lips, hyoid, and velum. A modification of the lower part of the pharynx and tongue-tip-to-jaw region is also provided and included in our model.

Once the articulatory positions have been specified, the cross-sectional areas are calculated by superimposing a grid structure on the vocal tract outline. These grid lines vary with the positions of the articulators (they are fixed in Mermelstein's model). A total of 60 sections, 59 sections for the vocal tract plus one section (fixed length and area) for the outlet of the glottis, are used in our model. The sagittal distance, g_j of section j, is defined as the grid line segment length between posterior-superior and anterior-inferior outlines. The center line of the vocal tract is formed by connecting the center points of the adjacent grid lines. The length of the center line is considered equivalent to the length of the vocal tract. The sagittal distances are eventually converted to cross-sectional areas by empiric formulas [11].

The calculation of formant frequencies from a given vocal tract cross-sectional area function has been well established in the acoustic theory of speech production [6][1][14][2][7][9][10]. By computing the acoustic transfer function of a given vocal tract configuration, we can decompose the formant frequencies from the denominator of the acoustic transfer function. One of the functions of the articulatory model is to compute the articulatory information (in particular, the vocal tract cross-sectional area) from the acoustic information (the first four formant frequencies in our study) that are obtained from the speech signal. In general, an optimization scheme is used to solve this speech inverse problem. The optimization scheme varies the articulatory parameters iteratively to achieve a match between the model-generated and the desired first four formants.

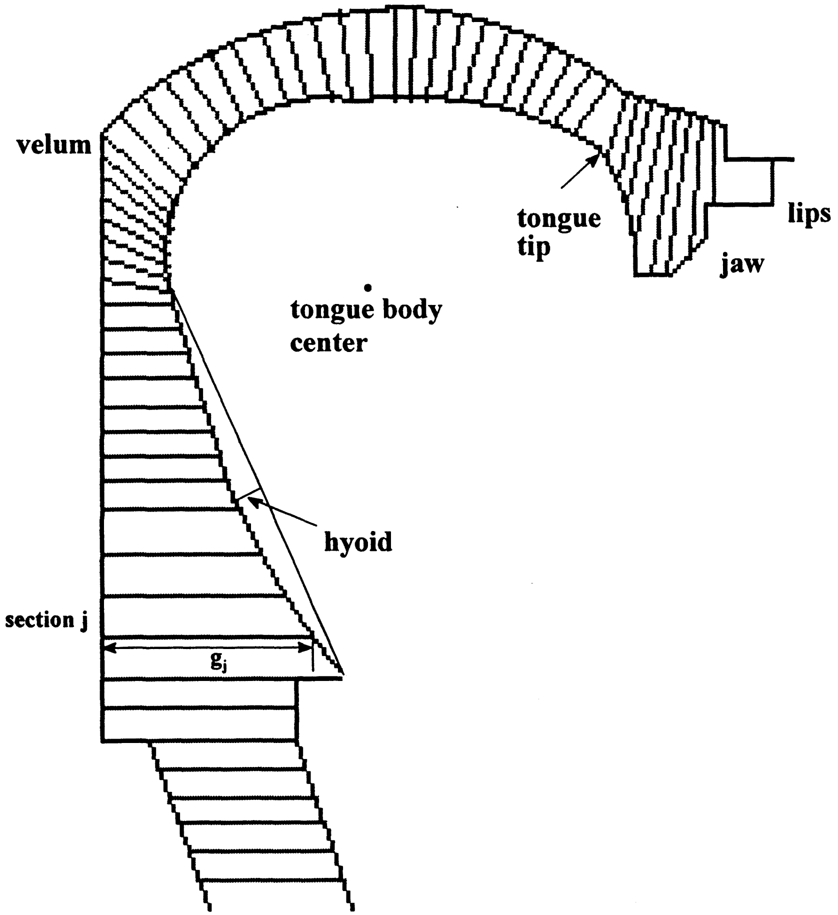


Figure 1: Articulatory model's parameters and midsagittal grids.

3. SIMULATED ANNEALING ALGORITHM

The articulatory parameters are adjusted and optimized until the synthetic speech features differ minimally from the actual speech features. Advances in computer technology have allowed the solution of optimization problems that require large numbers of complicated function evaluations to be computed on relatively inexpensive machines in a reasonable time. Thus, stochastic methods can be applied to the speech inverse filtering problem. In general, finding the global minimum value of a cost function with many degrees of freedom is difficult, since the cost function tends to have many local minima. A procedure for solving such optimization problems should sample values of the cost function in such a way as to have a high probability of finding a near-optimal solution and should also lend itself to efficient implementation. Over the past few years, simulated annealing has emerged as a viable technique that meets these criteria.

Simulated annealing was first derived from statistical mechanics, where the thermodynamic properties of a large system in thermal equilibrium at a given temperature were studied [12]. A description of the physical annealing process inspired this algorithm. In this situation a solid metal is to be melted at a high temperature. After slow cooling (annealing), the molten metal arrives at a low energy state, since careful cooling brings the material to a highly ordered, crystalline state. Inherent random fluctuations in energy allow the annealing system to escape local energy minima to achieve the global minimum. However, if the material is cooled very quickly (or 'quenched'), it might not escape local energy minima and when fully cooled it may contain more energy than annealed metal. Simulated annealing attempts to minimize an analogue of energy in an annealing process to find the global minimum. Kirkpatrick et al. [8] were the first to propose and demonstrate the application of simulated annealing techniques to problems of combinatorial optimization, specifically to the problems of wire routing and component placement in VLSI design. Both Vanderbilt and Louie [13] and Bohachevsky et al. [3] have modified simulated annealing for continuous variable problems.

However, the Corana et al. [5] implementation of simulated annealing for continuous variable problems appears to offer the best combination of ease of use and robustness, so it is used for our optimization process. In summary, the simulated annealing algorithm starts at some high temperature specified by the user. A sequence of points is then generated until an equilibrium is approached. During this random walk process the step length vector is periodically adjusted to better follow the cost function behavior. After thermal equilibrium, the temperature is reduced and a new sequence of moves is made starting from the current optimum point, until thermal equilibrium is reached again, and so forth. The process is terminated at a low temperature such that no more useful moves can be made, according to the stopping criterion.

4. SPEECH INVERSE FILTERING STRATEGY AND PROCEDURE

In general, the relationship between the shape of the vocal tract and its acoustic output can be represented by a multidimensional function of a multidimensional argument

$$y = f(x) \tag{1}$$

where x is a vector formed by the coordinates of the articulators, y is a vector formed by the corresponding acoustic features, and f is the function relating these vectors. Given an acoustic measurement y_d, the problem is to find an articulatory state x_o such that $f(x_o)$ is the best match to y_d.

4.1 Strategy

Speech inverse filtering is a "constrained multidimensional nonlinear optimization problem." The coordinates of the tongue body (tbodyx, tbodyy), tongue tip (tipx, tipy), lips (lipp, lipo), jaw

(jaw), and hyoid (hyoid) compose the multidimensional articulatory vector , i.e.,

$$x = [tbodyx, tbodyy, tipx, tipy, lipp, lipo, jaw, hyoid] \qquad (2)$$

Note that x is an 8-dimensional vector. Usually, the velum is set at different default positions for nasal, non-nasal, or nasalized phonemes, but it can be optimized for some phonemes. The dimensions of the lower pharynx are also allowed to be optimized whenever this is necessary.

We designate the articulatory vector as

$$x = [x_1, x_2, \ldots , x_M] \qquad (3)$$

where the value of M represents the number of dimensions of the articulatory domain to be optimized. As mentioned in the previous paragraph, M has a value of eight. For nasal and nasalized sounds, we may include the velum as an additional articulatory parameter, i.e., M is set to 9. For middle vowels, some back vowels, and semivowels, three more parameters, related to the height between pharynx and larynx, and their anterior-posterior movements, are included, i.e., M is set to 11. To the extremity, one more parameter, velum, is included, i.e., M=12.

The acoustic vector is composed of the first four formant frequencies, i.e., $y = [F_1, F_2, F_3, F_4]$. The cost function (error distance) is derived from a comparison of between the first four formant frequencies of the articulatory model and the first four formant frequencies determined from speech analysis. A percentage of the weighted least-absolute-value (l_1-norm) error distance is defined as:

$$\sum (W_i \mid F_{mi}(x) - F_{ti} \mid)/ F_{ti} \quad \% , i = 1, 2, 3, 4. \qquad (4)$$

where F_{mi} is the ith model-derived formant which is function of articulatory vector, F_{ti} is the ith target-frame formant estimated from the analysis of speech signal, and W_i is the assigned weight. The constraints include the articulatory-to-acoustic transformation function f, and the lower and upper boundary conditions of the articulatory parameters.

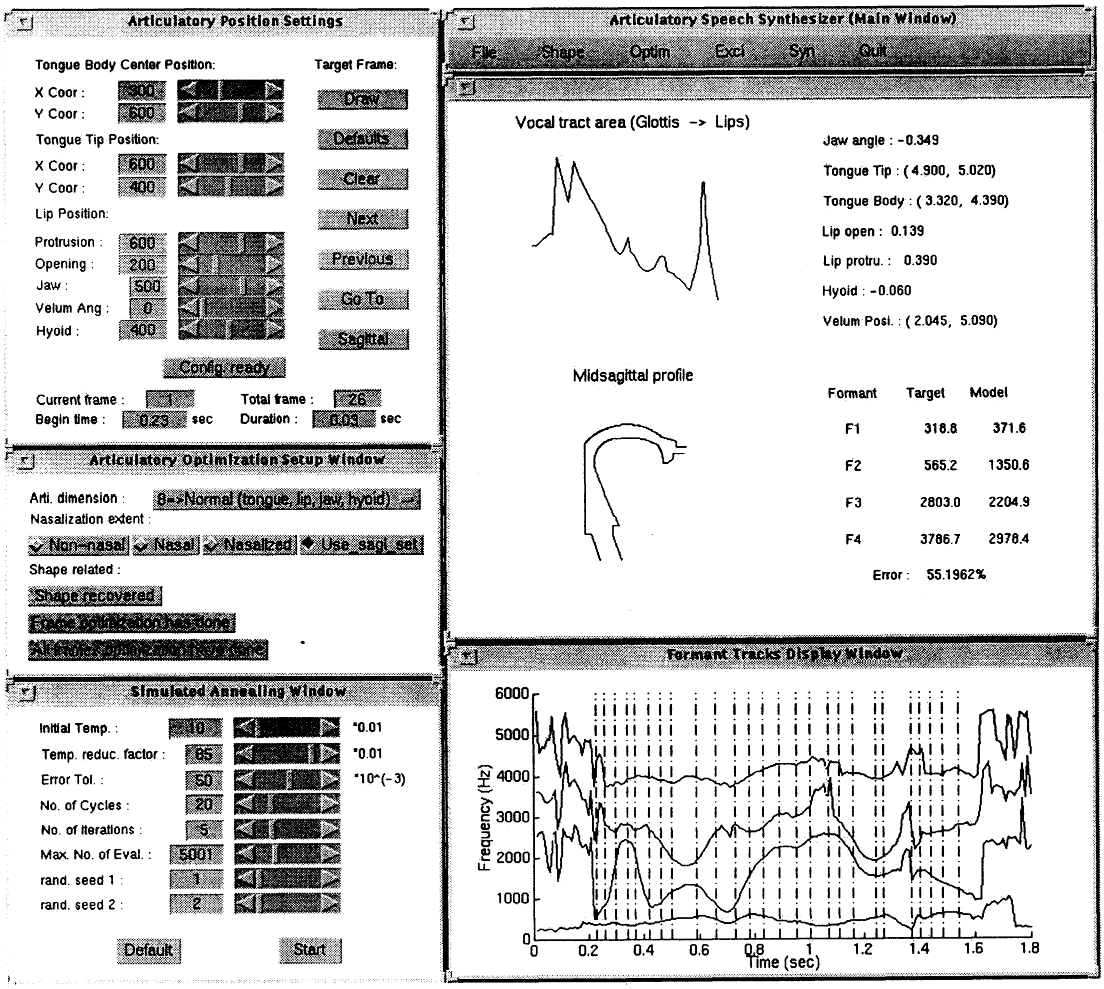
The object of the optimization process is to find the optimal articulatory vector that generates the acoustic vector (model-derived) as close to the desired (target-frame) as possible. The ideal minimum value is 0%, but some approximations used in the articulatory model make this value hard to reach. The first approximation is related to the articulatory model. A non-robust representation of the lower part of the pharynx and the tongue tip-to-jaw region may cause some deviations on the midsagittal vocal tract outline. The second, and more significant deviation, is the uncertainty of the sagittal distance to cross-sectional area transformations. The final one is the area to formant frequency conversion. We have determined that an error criterion requiring the final value of error distance function to be less than 1% appears adequate.

4.2 Procedure

Figure 2 shows the user interactive windows used during the speech inverse filtering phase. To extract the articulatory trajectories from a speech sentence, the first step is to obtain a smoothed formant trajectory from the speech signal. Then N target frames are selected. The target frame selection is based on the results of the speech analysis, which include the formant trajectory, the location of the word endpoints, and the estimated phoneme boundaries of the speech signal. The speech inverse filtering procedure, as shown in Figure 3, is applied to each target frame to obtain the optimum articulatory parameters. For each target frame, an initial value of the error distance function (cost function) is evaluated from the initial articulatory vector.

The error distance function evaluation includes the computations of the sagittal distances and the section lengths, the calculations of the vocal tract cross-sectional area and the acoustic transfer function, the decomposition of the first four formants from the acoustic transfer function, and the calculation of the error distance.

Then the simulated annealing algorithm controls the movement of the search path. Each movement requires the generation of a next candidate point, the error distance function evaluation for the candidate point, and the decision to move. After a number of steps, the temperature is lowered and a new search begins. The process stops if the near-global minimum is reached or the maximum allowed number of function evaluations is exceeded. The speech inverse filtering procedure terminates when all target frames are optimized. The articulatory parameters and the vocal tract cross-sectional areas of all the optimized N target frames can be saved as disk file for later use or can be directly passed to the articulatory synthesizer for synthesis.


Articulatory Position Settings
Target Frame:
Draw
Defaults
Clear
Next
Previous
Go To
Sagittal
Config. ready
Articulatory Optimization Setup Window
Shape recovered
Frame optimization has done
All frames optimization have done
Simulated Annealing Window
Default
Start
Articulatory Speech Synthesizer (Main Window)
File
Shape
Optim
Excl
Syn
Quit
Vocal tract area (Glottis -> Lips)
Midsagittal profile
Formant Tracks Display Window
6000
5000
4000
3000
2000
1000
0
Frequency (Hz)
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
Time (sec)

Tongue Body Center Position:
X Coor : 300
Y Coor : 600
Tongue Tip Position:
X Coor : 600
Y Coor : 400
Lip Position:
Protrusion : 600
Opening : 200
Jaw : 500
Velum Ang : 0
Hyoid : 400

Current frame : 1 Total frame : 26
Begin time : 0.23 sec Duration : 0.03 sec

Arti. dimension : 8->Normal (tongue, lip, jaw, hyoid)
Nasalization extent :
✓ Non-nasal ✓ Nasal ✓ Nasalized ◆ Use_sagi_set
Shape related :

Initial Temp. : 10 *0.01
Temp. reduc. factor : 85 *0.01
Error Tol. : 50 *10^(-3)
No. of Cycles : 20
No. of Iterations : 5
Max. No. of Eval. : 5001
rand. seed 1 : 1
rand. seed 2 : 2

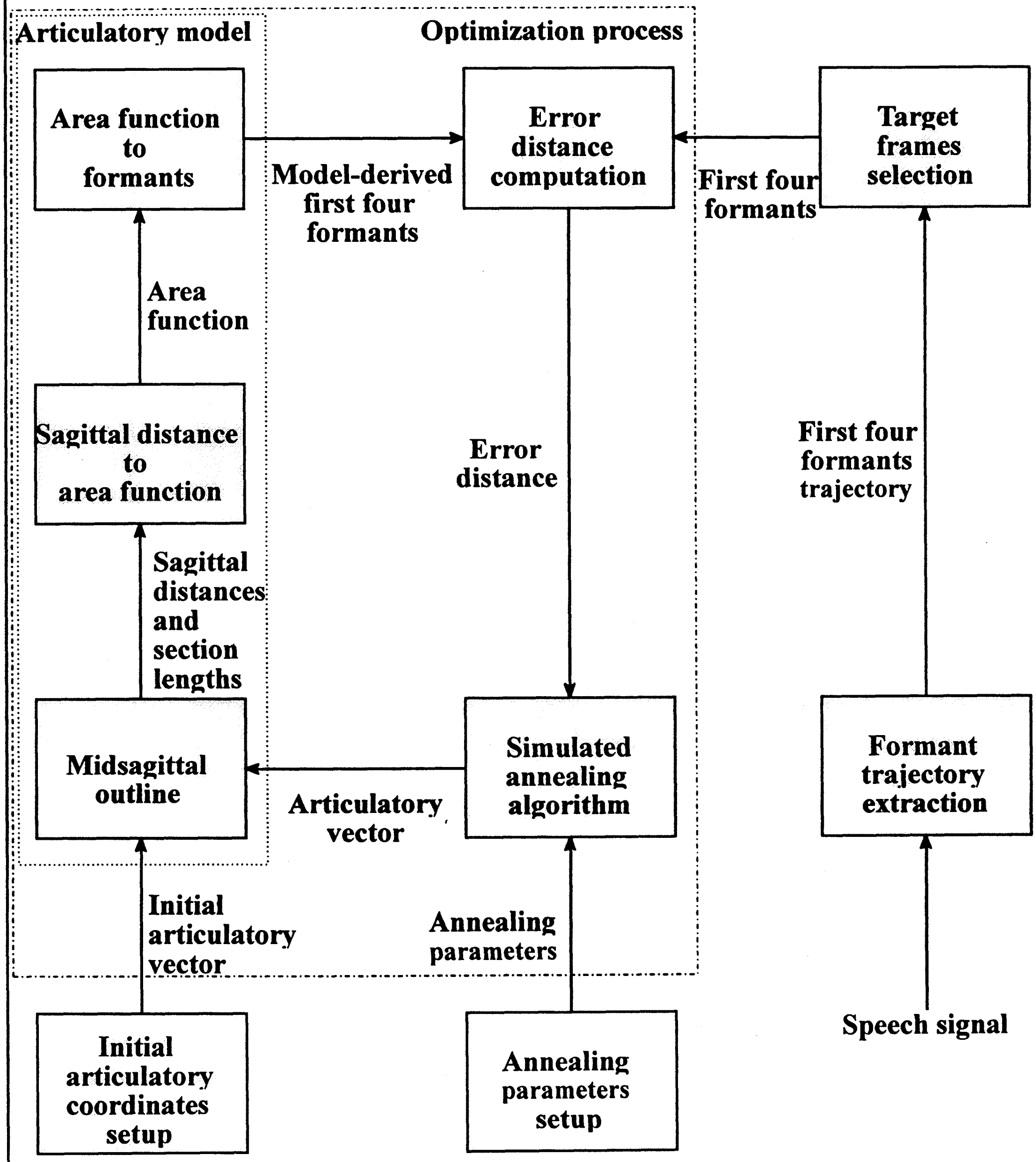
Jaw angle : -0.349
Tongue Tip : (4.900, 5.020)
Tongue Body : (3.320, 4.390)
Lip open : 0.139
Lip protru. : 0.390
Hyoid : -0.060
Velum Posi. : (2.045, 5.090)

Formant	Target	Model
F1	318.8	371.6
F2	565.2	1350.6
F3	2803.0	2204.9
F4	3786.7	2978.4

Error : 55.1962%

Figure 2: Speech inverse filtering related windows.

The analogues between the annealing process and the articulatory problem can be identified as following. First, the percentage of the weighted least-absolute-value (l_1-norm) error distance, equation 4, corresponds to the energy of the material. The articulatory vector, equation 2 or 3, corresponds to the configuration of particles. The change of articulatory parameters corresponds to the rearrangement of particles. Finding a near-optimal articulatory vector corresponds to finding a low-energy configuration. The temperature of the annealing process becomes the control parameter for the speech inverse filtering process. Second, the Metropolis algorithm corresponds to the random fluctuations in energy. Third, the temperature reduction coefficient corresponds to the cooling rate. Fourth, the finite number of moves at each downward control temperature value corresponds to the amount of time spent at each temperature.



Articulatory model
Optimization process
Area function to formants
Model-derived first four formants
Error distance computation
First four formants
Target frames selection
Area function
Sagittal distance to area function
Error distance
First four formants trajectory
Sagittal distances and section lengths
Midsagittal outline
Articulatory vector
Simulated annealing algorithm
Formant trajectory extraction
Initial articulatory vector
Annealing parameters
Initial articulatory coordinates setup
Annealing parameters setup
Speech signal

Figure 3: The block diagram of the speech inverse filtering procedure.

Reasonable values, found after some experimental tests, of the parameters (Table 1) are used as defaults for the optimization process. The following guideline provides some rules for adjusting the appropriate annealing parameters:

i. Set the desired nasalization extent and set the number of dimensions of the articulatory vector at the appropriate dimensions, e.g., M=8 for front vowels, M=9 for nasalized front vowels, M=11 for middle, back vowels, and semivowels, and M=12 for nasalized vowels. Start the optimization process with the default initial articulatory vector and the default annealing parameters. If the error distance is less than 1% after the process stops, go to step v. If not, go to step ii.

ii. Check if the current vocal tract shape (or cross-sectional area) is reasonable. If the shape is not reasonable go to step iii. Otherwise, record the error distance ε_p and the current final temperature as T_p. Then set the initial temperature T equal to the floor value of T_p. Start the optimization process again. If the new error distance is less than ε_p, then this step is repeated until the error criterion is met. If not go to step iv.

iii. Use the control button with label Shape Recovered on the Articulatory Optimization Setup Window to recover the vocal tract shape (articulatory vector) to the initial settings. Several adjustments of annealing parameters can be used. The following order of adjustments are recommended: raise the initial temperature, increase the value of the reduction factor, increase the total number of evaluations, and change the other annealing parameters. Then begin the process and apply step ii.

iv. Recover the vocal tract shape as described in step iii. Increase the number of dimensions of the articulatory vector from M=8 to M=11 and start the process. Apply step ii.

v. Check the vocal tract shape with X-ray tracings or schematic vocal tract profiles as published in the literature. If the vocal tract outline is similar to those published, then the optimization process is done. If not, this may mean that the "ventriloquist effect" has occurred. One can adjust the settings of the nine articulatory sliders in the Articulatory Position Settings window so that initial configuration is closer to the true outline. Then go to step i to start the optimization process again.

vi. If the above steps have been tried and the error criterion is still not satisfied, then go back to the target-frame selection phase and reselect the current target frame. Start the optimization process from the step i.

5. RESULTS AND CONCLUSION

Figure 4 presents the articulatory characteristics for /I/ and /i/ vowels. The midsagittal vocal tract outline and the corresponding synthetic speech waveform are obtained from sustained vowel phonations by using the simulated annealing algorithm. We can see that the simulated annealing optimization algorithm works well, since most of the error distances are less than 0.5%.

The simulated annealing algorithm is also applied to perform the speech inverse filtering for two speech signals that were obtained from one speech token spoken by two male subjects. The simulated annealing algorithm performs well. On the average, over 87% of the total frames have an error distance less than 0.1%.

The above results illustrate the usefulness of the simulated annealing algorithm, which has proved to be efficient and very flexible in dealing with the problems that are inherent to the acoustic-to-articulatory transformation. However, the selection of parameters for the annealing schedule is an obstacle for the simulated annealing algorithm, since we know little about the relation between the argument domain (articulatory vector) and the technology (the algorithm). The guideline and the default annealing parameter values in Table 1 are considered a good procedure at this time. The evaluation of the error distance function is the most computationally intensive part of the program.

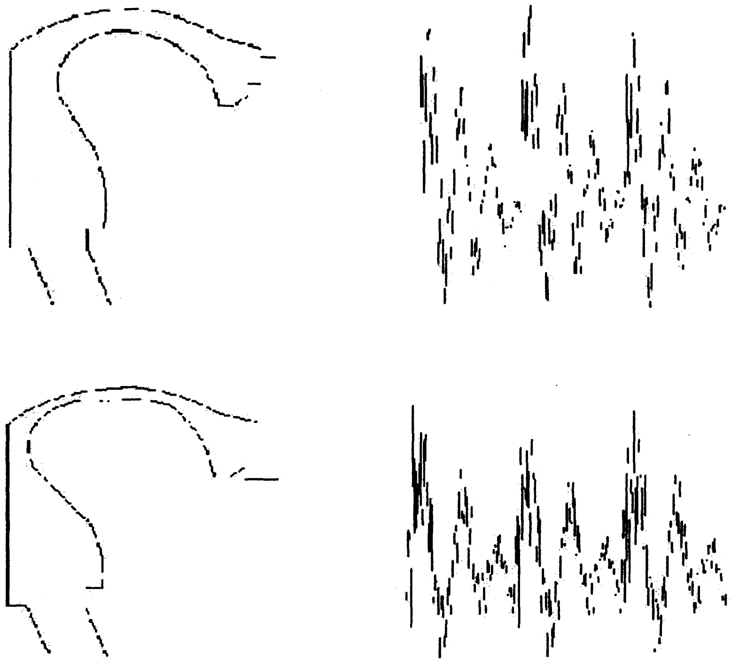


Figure 4: The optimized articulatory characteristics for /I/ and /i/.

Table 1: Default annealing parameter

Annealing parameters	Default values
artificial temperature (as control parameter)	0.1 – 0.2 degrees
temperature reduction coefficient	0.85
number of steps to adjust the step length vector	20
number of adjustments at each temperature	5
number of successive temperatures to test for stopping	4
termination criterion	0.005
total number of function evaluations	5001
step length; where i = 1, 2, ..., M	3.0

6. REFERENCES

[1] Atal, B. S., Chang, J. J., Mathews, M. V., and Tukey, J. W. (1978). "Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique," J. Acoust. Soc. Am., 63(5), 1535-1555.

[2] Badin, P., and Fant, G. (1984). "Notes on vocal tract computation," STL-QPSR, Royal Institute of Technology, Stockholm, Sweden, 2-3, 53-108.

[3] Bohachevsky, I. O., Johnson, M. E., and Stein, M. L. (1986). "Generalized simulated annealing for function optimization," Technometrics, 28(3), 209-217.

[4] Coker, C. H. (1976). "A model of articulatory dynamics and control," Proc. IEEE, 64(4), 452-460.

[5] Corana, A. C., Marchesi, M., Martini, C., and Ridella, S. (1987). "Minimizing multimodal functions of continuous variables with the 'simulated annealing' algorithm," ACM Transactions on Mathematical Software, 13(3), 262-280.

[6] Fant, G. (1960). Acoustic Theory of Speech Production, Mouton and Co., Gravenhage, The Netherlands.

[7] Fant, G. (1985). "The vocal tract in your pocket calculator," STL-QPSR, Royal Institute of Technology, Stockholm, Sweden, 2-3, 1-19.

[8] Kirkpatrick, S., Gelatt, C. D., Jr., and Vecchi, M. P. (1983). "Optimization by simulated annealing," Science, 220(4598), 671-680.

[9] Lin, Q. G. (1990). "Speech production theory and articulatory speech synthesis," Ph.D. dissertation, Royal Institute of Technology, Stockholm, Sweden.

[10] Lin, Q. G. (1992). "Vocal-tract computation: How to make it more robust and faster," STL-QPSR, Royal Institute of Technology, Stockholm, Sweden, 4, 29-42.

[11] Mermelstein, P. (1973). "Articulatory model for the study of speech production," J. Acoust. Soc. Am., 53(4), 1070-1082.

[12] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). "Equation of state calculations by fast computing machines," Journal of Chemical Physics, 21, 1087-1092.

[13] Vanderbilt, D., and Louie, S. G. (1984). "A Monte Carlo simulated annealing approach to optimization over continuous variables," Journal of Computational Physics, 56, 259-271.

[14] Wakita, H., and Fant, G. (1978). "Toward a better vocal tract model," STL-QPSR, Royal Institute of Technology, Stockholm, Sweden, 1, 9-29.