Speech Optimization By Articulatory Analysis

Chang-Shiann Wu
Department of Information Management
National Formosa University
cswu@nfu.edu.tw

Abstract

The purpose of this study is to develop one solution to the speech optimization 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.

Keyword：optimization, speech analysis

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 is difficult due to the non-uniqueness of the solution. 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. Articulatory Model

According to the acoustic theory of speech production, 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. 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.

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 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. 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, as shown in 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. The sagittal distances are eventually converted to cross-sectional areas by empiric formulas. 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. 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 optimization problem. The optimization scheme varies the articulatory parameters iteratively to achieve a match between the model-generated and the desired first four formants.

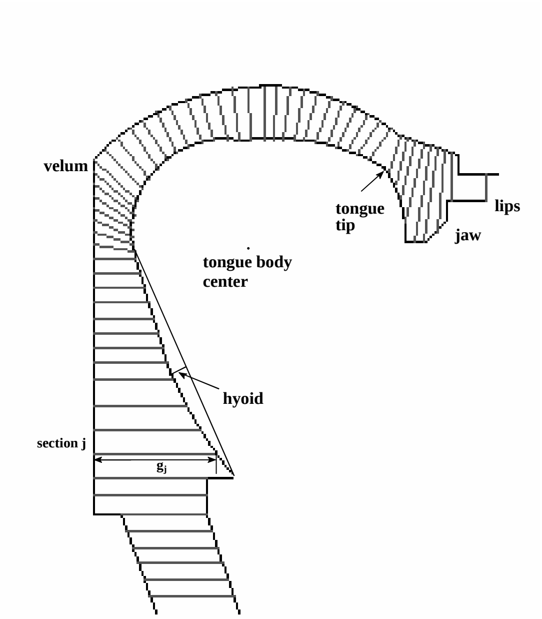

Figure 1: Articulatory model's parameters and midsagittal grids.

The articulatory parameters are adjusted and optimized until the synthetic speech features differ minimally from the actual speech features. The Corana et al. 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.

3. Speech Optimization Strategy

Speech optimization 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] (1)

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

$$x = [x_1, x_2, \dots , x_M] \qquad (2)$$

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:

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

where F_{mi} is the i^{th} model-derived formant which is function of articulatory vector, F_{ti} is the i^{th} 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. We have determined that an error criterion requiring the final value of error distance function to be less than 1% appears adequate.

The speech optimization procedure, as shown in Figure 2, is applied to each target frame to obtain the optimum articulatory parameters. 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. 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 optimization 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.

4. Results and Conclusion

Figure 3 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 optimization 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.

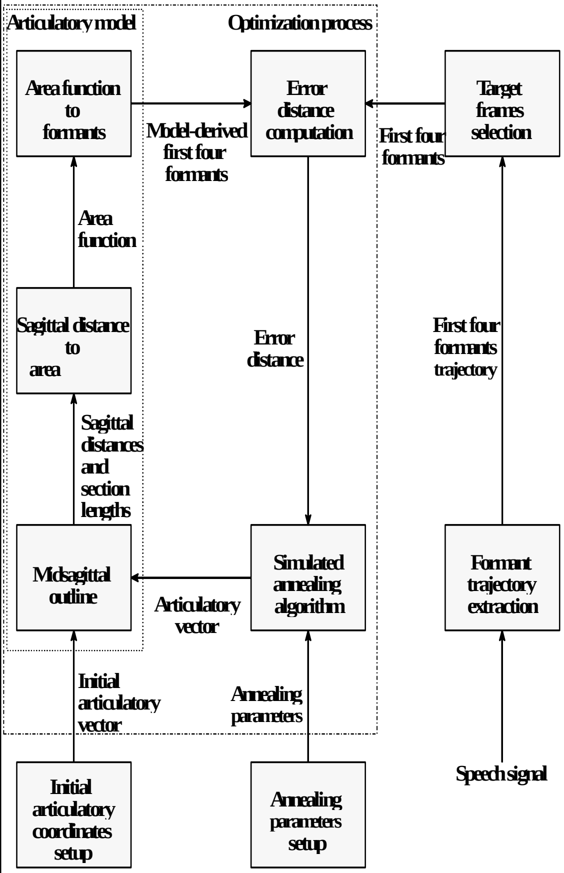
Articulatory model
Optimization process
Area function to formants
Error distance computation
Target frames selection
Model-derived first four formants
First four formants
Area function
Sagittal distance to area
Error distance
First four formants trajectory
Sagittal distances and section lengths
Midsagittal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

Figure 2: The block diagram of the speech optimization procedure.

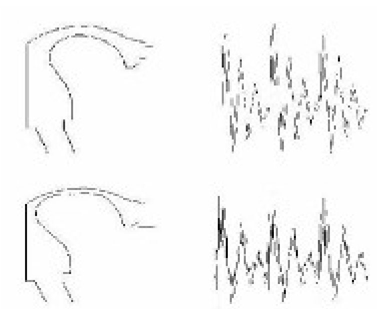

Figure 3:The optimized articulatory characteristics for /I/ and /i/.

5. 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.