Simulation and Visualization of Articulatory Trajectories Estimated from Speech Signals

G. Richard, M. Goirand, D. Sinder, J. Flanagan

CAIP Center, Rutgers Univ., Piscataway, NJ 08855, USA.

Abstract

This paper presents recent research on the estimation of articulatory speech parameters from an acoustic speech signal. The approach is based on a model of speech generation whose parameters are estimated to duplicate an arbitrary speech signal input. It is shown how perceptual articulatory codebooks and limited dynamic programming can improve the efficiency of the initial estimation. An extension of this voice mimic system to unvoiced sounds is also described and demonstrated by synthesizing short utterances of speech such as "she saw a fire". This research aims to advance fundamental understanding of human speech generation and coalesces the problem of speech synthesis, speech recognition, and low bit-rate speech coding into a compact parametric framework.

1 Introduction

The focus of this research is the development of an adaptive voice mimic system which automatically adjusts parameters of articulatory speech synthesis to match arbitrary human speech. This research is expected to result in a new parameterization of speech information that will directly support new methods for speech synthesis, low bit-rate speech coding, and robust speech and speaker recognition. It should also have a direct impact on speech therapy and/or education since the dynamic display of the articulatory trajectories allows visualization of the estimated movement of speech articulators. Significant efforts have been devoted in the past to estimate articulatory parameters from an acoustic speech signal (see [6],[7] for example or [8] for an extensive review of systems). The most sucessful methods for deriving vocal tract shapes from an acoustic speech signal are based on the analysis-by-synthesis approach. In such methods, articulatory speech synthesizer parameters are optimized by comparing the spectra of the synthesized speech with given spectra of natural speech. One of the major drawbacks of such approaches is that almost all optimization techniques will only find a local optimum near the starting point (the initial articulatory parameters given). In other words, the result of the optimization strongly depends on the initial solution given to the optimizer. Also, these studies were limited to voiced sounds excluding plosive bursts and fricatives. There is thus a strong need for efficient means to find accurate initial estimates of vocal tract shapes, referred to here as the open loop steering analysis, for all classes of phonems.

This paper presents an efficient open loop steering analysis and describes an extension of previous work to fricatives (i.e phonems such as /s/, /f/,…). The speech synthesizer currently in use is primarily based on linear acoustics of the vocal system. However, improved techniques for speech synthesis based on numerical solution of the Navier-Stokes equations are also being developed ([15, 10]). The paper is organized as follows: section 2 presents our approach for the open loop estimation of articulatory parameters, starting with a description of the speech generation models used. Section 3 proposes two applications developed using our Voice Mimic system. Finally, section 4 suggests some conclusions.

2 Open Loop Estimation of Articulatory Parameters

The estimation of articulatory parameters is performed in two major steps: an open-loop steering analysis which provides an initial estimation, and a closed-loop (based on optimization algorithms such as the steepest descent algorithm) which refines the initial solution.

2.1 Articulatory Synthesizer

To estimate vocal tract shapes from an acoustic speech signal, a human speech production model is used. This model is referred to as an articulatory synthesizer since it synthesizes speech from slowly varying physiological parameters. The main components of a human speech production system are the mechanism of voiced sound generation (i.e. the vocal cords) and the mechanism for modulating sound timbre (i.e. the vocal tract). Oscillation of the vocal cord results in a glottal volume velocity, U_g, that can be modelled either by a self oscillating vocal cords model (such as the now traditional two-mass model [5]) or by a parametric model of the glottal flow (or similarly a model of the derivative of the glottal flow such as the LF model ([2])). Both the two-mass model and the LF model have been used for this research, however only results with the latter will be presented in this paper.

2.1.1 Vocal Tract Model

The vocal tract shape is defined by means of an articulatory model of the vocal tract. Two different models have been used: Tracttalk (a detailed description may be found in [3]) and the Ishizaka-Flanagan model ([6]). Both models provide stylized vocal tract shapes defined by a small set of parameters. For the research described below, at most 4 parameters have been used: location and size of the main constriction in the vocal tract, mouth aperture and front cavity area section (see Figure 1). It must be noted that the nasal tract has not been considered in this study.

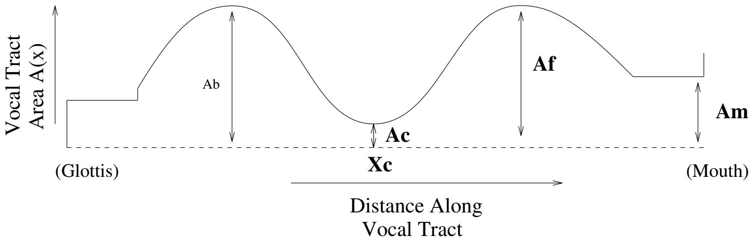


Figure 1: *Vocal tract model and the 4 main articulatory parameters: Main constriction size (A_c in cm^2) and location (X_c in cm), mouth opening (A_m in cm^2) and front cavity area section (A_f in cm^2) (after [6])*

2.1.2 Source Models

The LF model is defined by five main parameters: the fundamental frequency, F_0, and four wave shape parameters (see Figure 2). As it is described below, only two of them are estimated in the Voice Mimic system (the Fundamental frequency, F_0, and the energy, E_e). The other three mainly characterize the voice quality and further work will be needed to achieve an accurate estimation of these parameters.

For fricative generation, a model for the generation of turbulence in the vocal tract has been used ([4]). In this model, random pressure sources are placed at locations of the vocal tract where the Reynolds number[1] exceeds a critical threshold value. The intensity of the sources is proportional to the square of the Reynolds number in excess of the critical value Re_c:

[1] The reynolds number Re is a dimensionless quantity largely used in fluid mechanics to describe the turbulence intensity of a fluid (laminar versus turbulent). $Re = \frac{\rho w u}{\mu}$ where ρ is air density, u is particle velocity, μ is the viscosity coefficient and w is the characteristic width (in our case the width of the constricting passage).

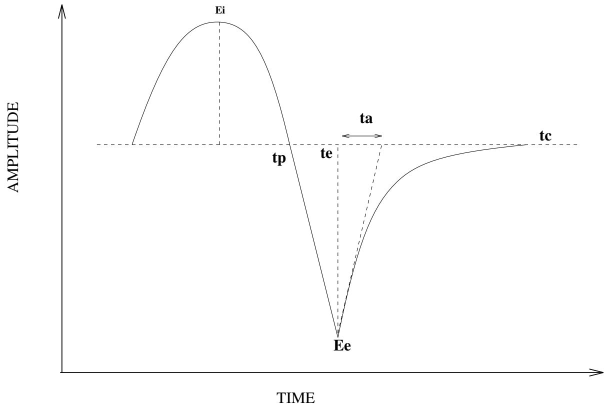


Figure 2: *The LF model of differentiated glottal flow (after [2])*

$$\begin{aligned} I = & \quad \alpha |Re^2 - Re_c^2| & \text{if } Re > Re_c \\ = & \quad 0 & \text{if } Re < Re_c \end{aligned} \tag{1}$$

2.2 Articulatory Codebooks

Best results for the so called open-loop estimation are based on codebooks that provide an acoustic-to-articulatory mapping. Such a codebook, often called an articulatory codebook, associates a discretized set of all possible vocal tract shapes (defined by the model) to corresponding spectral representations of the sounds they produce. Thus, given an acoustic input, its spectra is compared with all spectra in the codebook, and the "closest" spectrum is selected. Then the articulatory shape estimate is taken as the shape which produced that spectrum (see Figure 3).

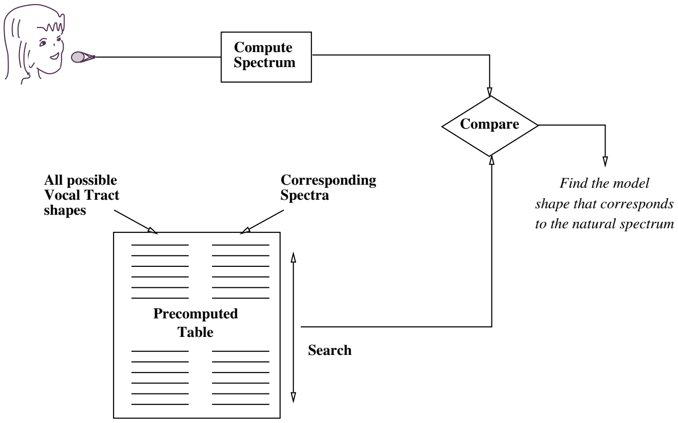

Compute Spectrum
Compare
Find the model shape that corresponds to the natural spectrum
All possible Vocal Tract shapes
Corresponding Spectra
Precomputed Table
Search

Figure 3: *Use of an articulatory codebook*

2.2.1 Codebook design

The design of a codebook is especially important in order to get vectors that will span both the acoustic space and the articulatory domain.

In the work of Schroeter and Sondhi ([8]), the articulatory codebook was built by using a Vector Quantization technique to obtain geometric and acoustic training vectors. Their approach, called the root shape interpolation method, starts from predefined (in an articulatory sense) root shapes and finds the other shapes by interpolation and clustering. Other studies preferred to randomly sample the articulatory space ([9]). This last method has the advantage of spanning all the articulatory space but has the drawback of creating many unreasonable tract shapes and thus produces a larger codebook than necessary. We present below an improvement of the random-sampling method by using a perception criterium.

The idea is to depart from random sampling codebook by keeping only vocal tract shapes that produce perceptually different synthetic sounds. This approach generates a more dense distribution of vectors in regions where a small change in vocal tract shapes has a strong acoustic effect and a more sparse sampling in regions where large changes in vocal tract shapes are necessary to be acoustically effective. The perceptual criterium used ([11]), states that two synthetic signals are perceived differently if the relative difference between either their first or second formant center frequencies exceed 3%.

The generation of the perceptual articulatory codebook is then done as follows:

- initialization: the signal produced by the articulatory synthesizer for the minimum value of each parameter is synthesized and the center frequencies of the first two formants are estimated.
- next, one of the parameters is increased until a difference of 3% or more is seen between formant center frequencies of the current synthesized sound and the previous one stored in the codebook.
- previous step is then iterated for the range of all parameters.

An example of the sampling of the articulatory space for a perceptual codebook with three articulatory parameters is shown Figure 4. One may observe that higher density of elements are seen for small constriction sizes (small Ac) and forward constriction locations (constriction near the lips).

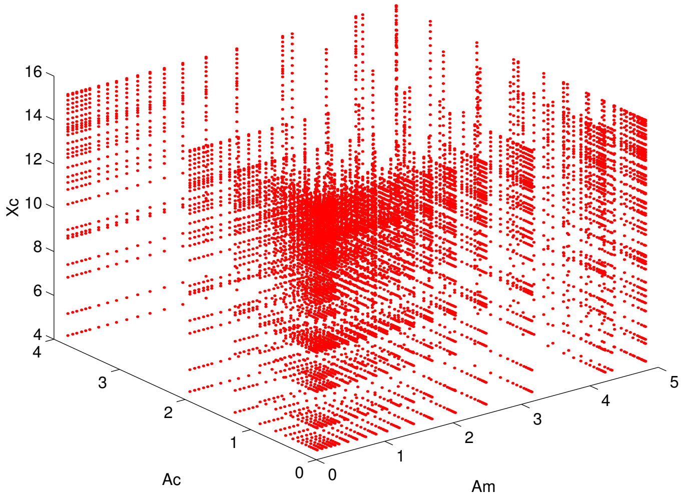


Figure 4: *Perceptual codebook*

The formant extractor used is based on LPC poles obtained by linear prediction. Similarly to [9], only poles with significant amplitudes and bandwidths less than 300 Hz are considered as formants. As any formant estimator, ours may introduce some errors. However, they have a minor consequence here. As a matter of fact, an error will only result in adding elements in the codebook.

Compared to uniformly sampled codebooks, perceptual articulatory codebooks are much reduced in size and produce equivalent speech coding quality. However, it is reasonable to think that better codebooks should be obtained by using improved perception criteria based on center frequencies and bandwidths of the first 4 formants.

2.2.2 Unvoiced codebook

Evident discrepencies exist in the frequency content between sounds produced by a source at the glottis (vibration of vocal cords) and sounds produced by a noise source at a constriction in the vocal tract (as for fricatives). These discrepencies make necessary the use of multiple codebooks. However, while the voiced codebook may be based on perceptual criteria, it is not clear how to develop a perceptual codebook for unvoiced sounds since the difference between two "noises" is difficult to define. For the sake of simplicity, the same sampling defined for the voiced codebook is used by re-synthesizing all sounds for those shapes without excitation source at glottis.

2.3 Limited Dynamic Programming

Codebook searches ensure that the best acoustic match is obtained for each successive analysis frame; however, the overal solution might be physiologically unrealistic (due to abrupt changes of articulatory parameters) and thus might provide a low quality synthetic speech signal. A technique based on dynamic programming was introduced by Schroeter et al. ([8]) to extract smoothly evolving articulatory trajectory by penalizing fast changes of the articulatory parameters (or vocal tract shapes) under the constraint of matching a given sequence of spectra. With such an approach, it is not necessary to have accurate information on the dynamics of the vocal tract. The other advantage is its simplicity. However, the dynamic programming approach is computationnaly intensive when large codebooks are used. We present below a modification of the original implementation which allows improvement of the efficiency of the original algorithm.

In the process of dynamic programming, it is necessary to minimize the accumulated composite cost:

$$C_t = d(x_0, x_{j(0)}) + \sum_{t=1}^{T} \Big(d(x_j, x_{j(t)}) + D(Y_{j(t-1)}, Y_{j(t)}) \Big) \qquad (2)$$

where $D[Y_{j(t-1)}, Y_{j(t)}]$ is the geometric cost of making a transition from shape $Y_{j(t-1)}$ at time $t-1$ to shape $Y_j(t)$ at time t, and $d(x_j, x_{j(t)})$ is the acoustic distance between the given acoustic vector x_i and the acoustic vector $x_{j(t)}$ related to the candidate tract shape (both at time t). This accumulated composite cost is calculated iteratively for all paths in the codebook. This leads, for a codebook of M entries, to M^2 distance computations to extend the optimal paths from time t to time $t+1$.

In the original implementation, all possible transitions are considered- even those that are unrealistic. More precisely, it means that the acoustic and geometric costs (respectively $d(x_j, x_{j(t)})$ and $D(Y_{j(t-1)}, Y_{j(t)})$) are computed for all shapes $Y_{j(t-1)}$ in the codebook. It is thus possible, by taking advantage of vocal tract dynamics, to consider only shapes $Y_{j(t-1)}$ that are "close" to the considered shape $Y_{j(t)}$ to update the optimal path. The Figure 5 below illustrates this *limited dynamic programming*. Practically, the number of vectors used for the update can be reduced by a factor 8 leading to a total reduction of the number of distance computation by a factor 64 to extend the optimal paths from time t to time $t+1$, and this is done without affecting the final solution.

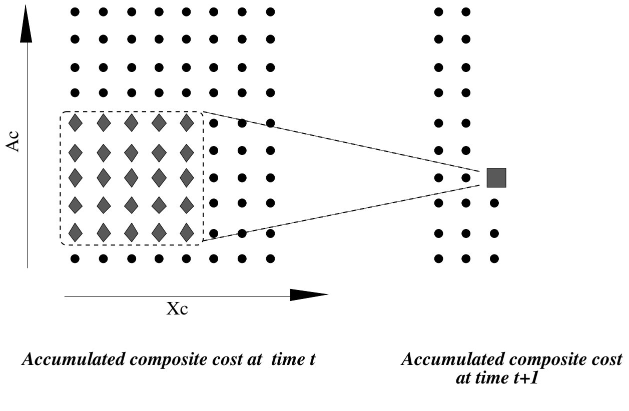

Ac
Xc
Accumulated composite cost at time t
Accumulated composite cost at time t+1

Figure 5: *Limited dynamic programming algorithm in the case of 2 articulatory parameters. The accumulated composite cost at time* $t+1$ *(for the node represented by a square) is updated using only paths arriving at the nodes represented by a diamond*

The choice of the acoustic distance $d(x_j, x_{j(t)})$ is especially crucial to obtain accurate estimations. We used the classical cepstral weighting techniques $d(s, n) = \sum_{k=1}^{15} w_k^2 (s_k - n_k)^2$ for its ability of representing all classes of phonemes and for its robustness against glottal variabilities.

2.4 General Scheme: Summary

The first step of the Voice Mimic system is the voiced and unvoiced codebooks generation. It is done by synthesizing 3 frames of 12.8 ms using a constant pitch glottal excitation for the voiced codebook and constant glottal air flow for the unvoiced one. The first 15 Mel Frequency Cepstrum Coefficients (MFCC) are computed on the middle frame and stored with the corresponding articulatory shape.

Once the codebooks are built, the estimation of articulatory parameters can be summarized as follows (see Figure 6):

1. The incoming speech signal is sliced into frames of 12.8 ms
2. The fundamental frequency is estimated using a Voiced/Unvoiced (V/U)) decision followed by a pitch extractor when the signal is declared voiced. The pitch estimator combines a comb filter approach followed by a dynamic programming module that avoids unrealistic F_0 changes and helps to reduce octaviation ([13], [14]).
3. The first 15 MFCC coefficients (excluding c_0) are calculated.
4. The appropriate codebook is selected based on the V/U decision on each frame.
5. Using Dynamic programming an optimal path is found in the articulatory domain.
6. A closed-loop optimization is started to refine the source energy parameter.
7. Ultimately, vocal tract parameters will also be refined using optimization techniques, which is needed for high quality speech coding applications.

It must be specified here that steps 1 to 4 can be done for a fixed number of analysis frames (typically 20) and that the optimal path found by dynamic programming does not need to be computed all at once on the total length of the incoming speech signal.

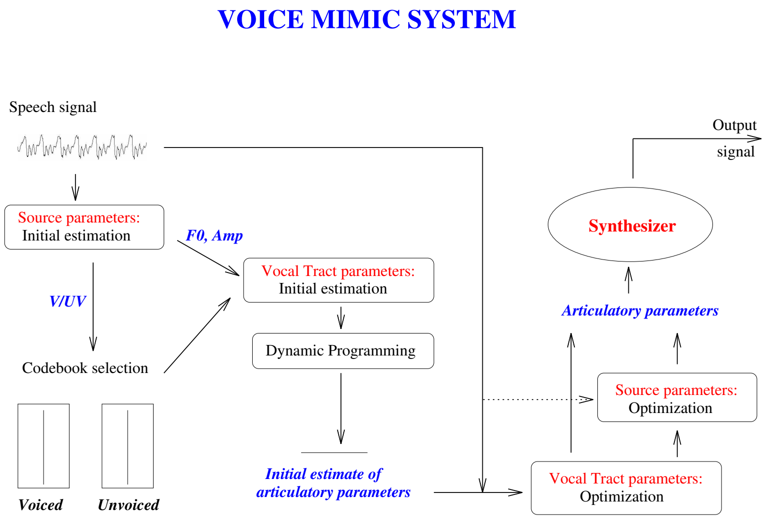


Figure 6: *General scheme of the Voice Mimic system*

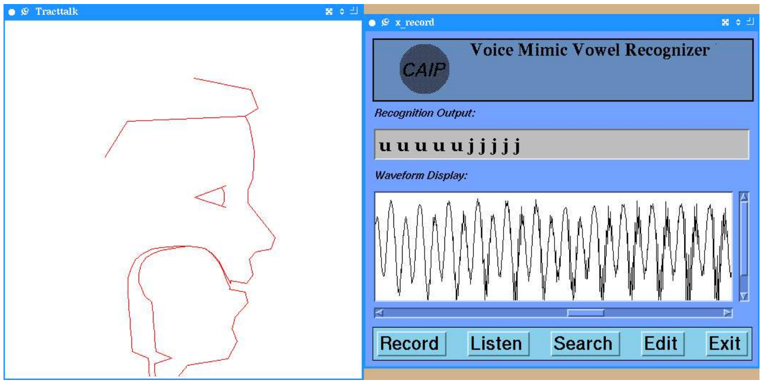
3 Applications

3.1 Vowel recognition

Using a spectral representation based on linear-predictive poles and a reduced number of articulatory parameters, a vowel recognition system based on an articulatory representation of speech signals has been designed. In contrast to the articulatory based approach, traditional speech recognition systems have relied on spectral and/or cepstral features. Despite considerable efforts seeking more accurate, compact, and reliable features for robust speech recognition, the articulatory representation of speech has not been exploited due to the difficulty and computational intensity involved in estimating articulatory parameters from speech waveforms. Adaptive voice mimic with optimized open-loop steering and efficient closed-loop control provides a promising solution to the challenge.

A nearly real-time laboratory prototype of the articulatory based recognition system has been implemented and demonstrated. The system can recognize both isolated vowels and vowel strings (see Figure 7). During the recognition computation, dynamically changing sagittal profiles of the vocal tract (corresponding to the input speech) are displayed.

It must be emphasized that the objective of this experiment was to demonstrate the potential of the articulatory representation of speech for speech recognition but further work is needed to include articulatory features in state of the art speech recognition systems.



Voice Mimic Vowel Recognizer

Recognition Output:

u u u u u j j j j j

Figure 7: *User interface of the vowel recognition prototype system*

3.2 Speech coding

This research also produced an adaptive voice mimic for vowels and fricative consonants (such as the /s/ in ”sea” and /f/ in ”fire”). Unlike the system of Badin et al. ([12]), this system is based on the sole information given by the acoustic speech signal.

This system has produced vowel/consonant/vowel utterances and short sentences (such as “she saw a fire”) of very encouraging quality. The speech quality obtained shows the potential of this approach for high quality, very low bit rate speech coding. Furthermore, it allows visualization of important features of speech production. As an example, Figure 8 provides the successive vocal tract shapes for the speech segment /usu/.

Since the voice mimic system provides a parametric representation of the signal, speech modifications are very simple. For instance, it is possible to change the rate of the speech by simply changing the rate of change of the articulatory parameters, or to change the overall size of vocal tract (and concurrently with the fundamental frequency) to transform a male voice to the voice of a female or child.

4 Conclusion

In this paper, recent results on the estimation of articulatory parameters from an acoustic speech signal are presented. It is shown that, using a perceptual articulatory codebook with limited dynamic programming, the efficiency of the algorithm proposed is greatly improved. Furthermore, this work represents an extension of previous studies to the class of unvoiced sounds which are not usually modelled in such systems. However, a rather simple noise generation model is used and significant improvement should be obtained by using an improved frication model. Future work will be devoted to closed loop optimization in order to refine the estimation of vocal tract parameters, which is needed for high quality speech coding applications.

Acknowledgement

This research is supported by the Advanced Research Project Agency (ARPA) under contract ARPA DAST 69-93-C-0064.

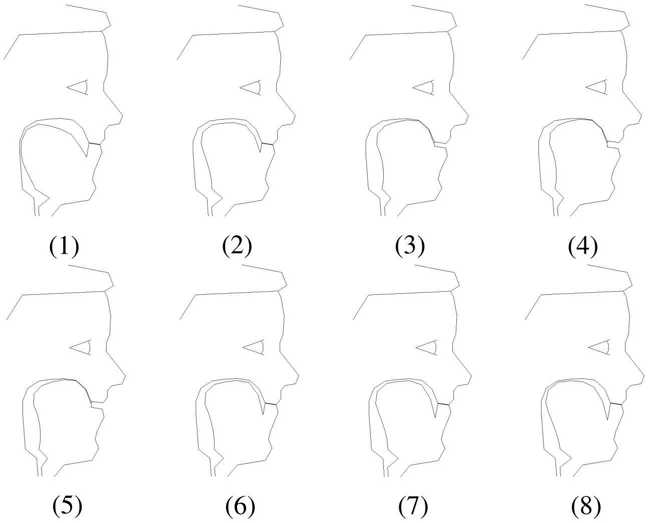


Figure 8: *Mid-sagittal profiles obtained by Voice Mimicking (speech segment /usu/)*

References

[1] G. Richard, Q. Lin, F. Zussa, D. Sinder, C. Che, and J. Flanagan,"Vowel recognition using an articulatory representation," 130th meeting of Acoustical Society of America, St. Louis, MO, November 1995.

[2] G. Fant, J Liljencrants, Q. Lin, "A four Parameter Model of Glottal Flow", Quaterly Progress and Status Report, STL-QPSR 4/1995, pp 1-13, 1985.

[3] Q. Lin, 'Speech Production Theory and Articulatory Speech Synthesis'" Ph.D. thesis, KTH Stockholm, Trita-Töm 90-1, ISSN 0280-9850, 1990.

[4] J. Flanagan and K. Ishizaka, "Automatic Generation of Voiceless Excitation in a Vocal Cord-Vocal Tract Speech Synthesizer", IEEE Trans. on ASSP, 24, No 2, pp 163 - 170, April 1976.

[5] K. Ishizaka and J. Flanagan, "Synthesis of Voiced Sounds from a Two-Mass Model of the Vocal Cords," Bell Syst. Tech. Journal 50, pp 1233-1268, 1972.

[6] J. Flanagan, K. Ishizaka, and K. Shipley, "Signal models for low bit-rate coding of speech", *J. Acoust. Soc. Am.* 68(3), pp. 780-791, 1980.

[7] J. Schroeter, J. Larar, and M. Sondhi. "Speech parameter estimation using a Vocal Tract/Cord model", *Proc. ICASSP-87,* Paper 8.6, vol 1, pp. 308-311, 1987.

[8] J. Schroeter and M. Sondhi, "Techniques for Estimating Vocal-Tract Shapes from the Speech Signal", IEEE trans. on Audio ...

[9] J. Schroeter, P. Meyer and S. Parthasarathy, "Evaluation of improved articulatory codebooks and codebook access distance measures", *Proc. of ICASSP-90,*, Paper S7.6, vol 1, pp 393-397, 1990.

[10] D. Sinder, G. Richard, H. Duncan, J. Flanagan, M. Krane, S. Levinson, S. Slimon, D. Davis, "Flow Visualization in Stylized Vocal Tracts", International Symposium in Simulation, Visualization and Auralization for Acoustic Research and Education, Tokyo, Japan, April 2-4, 1997.

[11] J. Flanagan, "Difference limen for the intensity of a vowel sound", *J. Acoust. Soc. Am.* 27, pp. 1223-1225, 1955.

[12] P. Badin, K. Mawass, G. Bailly, C. Vescovi, D. Beautemps, X. Pelorson, "Articulatory Synthesis of Fricative Consonants: Data and Models", first ESCA Tutorial and Research Workshop on Speech Production Modeling: From control strategies to Acoustic, Autrans, France, pp 221-224, May 21-24, 1996.

[13] P. Martin. " Extraction of the fundamental frequency using a comb filter,", *Proc. IEEE ICASSP82,*, 1982

[14] B. Secrest and G. Doddington, " An integrated Pitch Tracking Algorithm for Speech Systems". Proc. ICASSP83 , vol 1, pp 1352-1355, 1983.

[15] D. Sinder, G. Richard, H. Duncan, Q. Lin, J. Flanagan, S. Levinson, D. Davis, and S. Slimon, "A fluid flow approach to speech generation", first ESCA Tutorial and Research Workshop on Speech Production Modeling: From control strategies to Acoustic, Autrans, France, May 21-24, 1996.