# Articulatory Analysis using a Codebook for Articulatory based Low Bit-Rate Speech Coding

*C. Silva* and S. Chennoukh***

* Departamento de Electrónica Industrial, Universidade do Minho, Portugal

** Center for Computer Aids for Industrial Productivity (CAIP),
Rutgers University, Piscataway, NJ 08854-8088, USA

## ABSTRACT

Fundamental to the success of the articulatory based speech coding is the mapping from acoustics to articulatory description. As the mapping is not unique and based on articulatory continuity criteria, the non-uniqueness of the articulatory trajectories is solved using a forward dynamic network. In this paper, we present new results on forward dynamic network used to estimate articulatory trajectories when using an improved articulatory codebook for acoustic-to-articulatory mapping. The improvement on the codebook design is based on a new model that provides more details on the vocal tract area function and on more appropriate articulatory parameter samplings according to the articulatory-acoustics relation.

## 1. INTRODUCTION

Speech analysis is concerned with the estimation, from the speech signal, of the parameters of a model for speech production consisting of a slowly time-varying linear system excited by either quasi-periodic (glottal) pulses or random noise. If the model is sufficiently accurate and the parameters are accurately determined, the resulting output of the model is to some extend indistinguishable from the natural speech. When the linear system is represented by a physiological model of the vocal tract, such an analysis-synthesis system is called a voice mimic system (Flanagan, 1980). The voice mimic aims to replicate the human ability to mimic speech signals that he hears without understanding their structure or their meaning. For this purpose, several studies has been conducted to provide a solution to the acoustic-to-articulatory mapping (Chennoukh, 1998). This research is in the same line than these studies. The first objective of this work is to provide a high quality articulatory analysis of speech signal in terms of articulatory features matching the input speech.

The voice mimic of interest in this research is based on an optimization technique (Flanagan, 1980). It is performed in two steps. First, an estimation of the articulatory trajectory, which describes an arbitrary speech input, is processed in an open loop steering. Second, an optimization of the articulatory parameters in a closed loop process improves the acoustic accuracy of the estimated articulatory trajectory compared to the measured acoustic features of the input speech. The voice mimic quality depends strongly on the open loop results. If the open loop provides perceptually correct sequence of vocal tract shapes, the synthetic speech should be comparable to the input signal. Central to the effort is the mapping from an acoustic signal to an articulatory description. However, acoustic-to-articulatory mappings are non-unique and, given a cost function, the optimization techniques converge only to a local extremum that may be in the vicinity of the initial parameters. Therefore, one needs to choose accurate startup parameters to initialize the optimization procedure. For this purpose, we used an articulatory codebook which is a long table of pairs of parameter vectors, on one side are the vocal tract model parameters and on the other side are the corresponding model shape acoustic features. The articulatory codebook should cover as much as possible the entire articulatory space to reach accurate shape estimates for any input acoustic vector.

Second section discusses the construction of an articulatory codebook. A new model of the vocal tract area function is used with appropriate model parameter samplings. The third section describes the technique used to deal with the non-uniqueness problem of model parameter trajectories. Example results of analysis are given in the fourth section and finally the conclusion. Papers are limited to four pages, but we will allow the inclusion of additional sound and image files on the CD-ROM. More information on this is provided below.

## 2. DESIGN OF THE ARTICULATORY CODEBOOK

The access procedure to the codebook depends on the codebook design. The typical method for designing a codebook is to order the vocal tract model shapes and their corresponding acoustic parameters either according to the iterative generation order of the model parameter values or according to random generation of these parameter values (Schroeter, 1992). Search times for both designs are long and are not convenient (Chennoukh, 1998). An off-line inverse mapping have been developed to simplify the on-line access and search of the codebook for the articulatory analysis (Chennoukh, 1997). The new technique for building a codebook clusters the vocal tract model shapes during their generation in an acoustic network that sub-samples the acoustic space figure 1. Such a network is referred straight forward for each input speech frame acoustic feature vector to obtain all matching model shapes (Chennoukh, 1997).

The search of matching shapes in the acoustic network does not always reach a node which clustered model shapes. In this case, a search for an alternative node in the neighborhood is processed in order to be used as a close match cluster of shape estimates for the input speech frame. The search looks for the closest match in terms of perception. As our perception is more sensitive to low frequency spectrum variations, we permit an error on the estimate acoustic features at the higher frequency

spectrum components. Thus, we perturb the node acoustic coordinates that affect high frequency spectrum components with one sampling period of the acoustic space. The perturbation is processed repeatedly for all possible combinations between parameter perturbations until a cluster of shapes is found. If no cluster is found at one sampling period, the search is launched again for two sampling period of the acoustic space away from the right node.

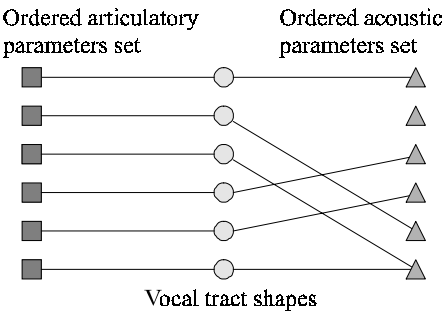


**Figure 1**: The inversion from acoustic to articulatory is performed during the construction of the codebook.

In this study, we used a codebook that maps vocal tract model shapes with the first three formants frequencies. The next paragraph discusses the vocal tract model used to describe the articulatory space of the vocal tract and the following paragraph provides the optimal setting for the model parameter sampling used to build our codebook.

## 2.1. Vocal tract area function model

In our previous study, we used Ishizaka's model to describe vocal tract area functions of vocalic sounds (Chennoukh, 1997). We have extended this model with two additional parameters to be able to model more complex vocal tract shapes such as vowel-consonant co-articulation or vice versa and consonant productions (Silva, 1998). The two parameters are the position of and the vocal tract area at the tongue tip, figure 2. A validation procedure interfaces the model parameter generation and the insertion in the codebook to insure an articulatory codebook of physiologically realistic shapes (Silva, 1998).

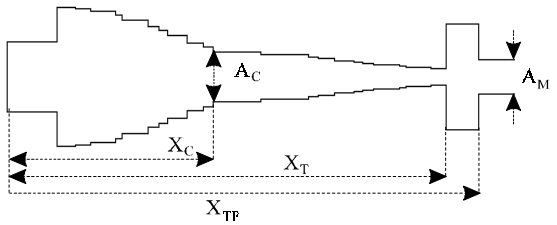


**Figure 2**: Vocal tract area function parameters. $X_{TF}$ is the maximum displacement of the tongue tip.

## 2.2. Articulatory parameter sampling

The network used to access the articulatory codebook samples linearly the acoustic space. Each node of the acoustic network should cluster model shapes when model shapes inserted in the codebook are best matched with this acoustic node. Ideally, one needs to generate model shapes that spans the articulatory space such as to hit all nodes within the limits of the speech acoustic space. This requires an appropriate sampling of the vocal tract model parameters.

| Parameter | From (cm2) | To (cm2) | Nr of pts | Sampling |
|---|---|---|---|---|
| Ac | 0.001 | 4.0 | 12 | Log. |
| Xc | 4.0 | 13.5 | 15 | Log. |
| Am | 0.5 | 8.0 | 12 | Log. |
| Ab | 2.0 | 10.0 | 3 | Linear |
| Af | 4.0 | 12.0 | 3 | Linear |
| At | 0.001 | 4.0 | 12 | Log. |

**Table 1**: Table with the parameter's variation of the model used in the construction of the codebook.

In our recent study, several sampling techniques of model parameters have been tested to improve the access to the codebook for the best match model shape estimates (Silva, 1998). It is showed that a codebook built with appropriate parameter samplings can outperform a much larger codebook with different sampling.

For our codebook we used the following samplings. The forward and backward cavity to the constriction area, *Af* and *Ab* respectively, were sampled in a linear scale, the area of the constriction *Ac*, the area of the tongue tip *At* and the place of constriction *Xc* were sampled in a logarithm scale. *Xc* in our model represents the distance from the teeth such as the sampling in linear scale provides a high sampling frequency at the alveolar region and low sampling frequency at the pharynx region. Table 1 shows the limits of the model parameters used to build the codebook. With these specifications, the codebook stored 199.836 shapes.

# 3. NON-UNIQUENESS OF THE ARTICULATORY TRAJECTORIES

The acoustic-to-articulatory mapping is non-unique. As a consequence, the articulatory trajectory is also non-unique. Schroeter (Schroeter, 1989) proposed a delayed decision method to solve this problem using the dynamic programming. This approach takes a decision on the optimal articulatory trajectory until a segment of about 10 frames of speech is analysed. The choice of the optimal trajectory is made according to a cost function which takes into consideration two components, the acoustic distance between the frame acoustic features and the shape acoustic feature and the geometric distance between the previous and the present model shape. The decision is taken on the articulatory trajectory cumulative distances. This approach is computationally intensive and its further development will highly increase the complexity of analysis.

Chennoukh (Chennoukh, 1997) coped with the non-uniqueness problem using a forward dynamic network (FDN) that is a predictive technique to estimate the next model shape according to the previous velocity and acceleration of the model

parameters, figure 3. The estimate is then used to decide on the best matched vocal tract shape according to a cost function among the clustered shapes. The cost function is the sum of square euclidean distance between the input shape parameters and their corresponding predicted position:

$$error = \sum_{i=0}^{N-1} (a_i - p_i)^2$$

where $N$ is the number of parameters, $a$ is the value of the parameter found in the cluster and $p$ is the value predicted by the FDN. The approach takes into account the dynamic properties of the articulators. This dynamic is updated frame-by-frame by the best matched shape according to the cost function. Thus, the articulatory trajectory is estimated frame-by-frame during the analysis.

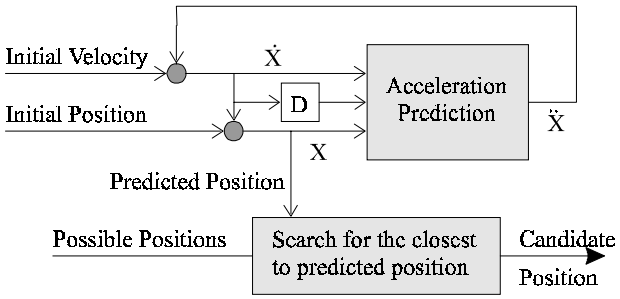


**Figure 3**: Dynamic Forward Network.

The computation complexity is several orders less than dynamic programming, allowing the introduction of more specifications on the dynamics of the articulators and consequently the vocal tract area function to cope with the complex processes of speech production.

## 4. RESULTS

The sentence "*Where are you?*" [SOUND 0899_01.WAV] was spoken by a male speaker. The signal was sampled at 16 KHz and windowed by a 32 ms Hamming window with 15 ms overlap. Levinson-Durbin's algorithm was used to compute the 18th order linear prediction model coefficients. Newton-Raphson's method was used to estimate the poles of the transfer function of the model. The obtained formants are used to access and search the codebook. Then, the codebook provides the best matched cluster of model shape vectors of model parameters which are in turn used as input to the FDN. The FDN outputs the optimal vector of parameters for the current frame. The figures 4a and figure 4b show the spectrograms for the original and the mimic result of the sentence. It can also be found in the CD-ROM the mimic result [SOUND 0899_2.WAV] of our current codebook and the mimic result [SOUND 0899_03.WAV] and the spectrogram [IMAGE 0899.GIF] using our previous codebook.

## 5. CONCLUSION

Numerical simulation of a voice mimic based on physiological model of speech production requires a knowledge of the optimal number of geometrical, acoustical and mechanical parameters in order to account for the complexity of speech production. One main difficulty in the simulation is to account for the complete conditions of various physiological processes involved in human speech production using the chosen parameters. The quality of the voice mimic output signal is, to some extend, a measure of the efficiency of the acoustic-to-articulatory mapping algorithm and the amount of physiological and articulatory conditions included in the conception of the vocal system simulation program. In a number of cases, however, the sound quality also depends on the degree of accuracy of the physical parameters involved in the vocal tract model.

In this paper, we reported our progress toward this objective using the forward dynamic network with an improved codebook design which uses a new area function model and appropriate model parameter samplings. We have increased the intelligibility of the mimic results compared to our previous system (Chennoukh, 1997). Efforts are devoted to improve the forward dynamic network by including articulatory constraints that will modulate the dynamic properties of the parameters of the model. This approach should lead us to uncover the co-articulation during the consonant production and to capture the consonant place of articulation. Another issue concerns the use of formant as acoustic features. Their detection becomes very difficult, if not impossible, at the closure and during the consonant production. Some efforts are also been devoted to the use of additional acoustic features to handle the mimic of sentences with consonant contexts.

## 5. REFERENCES


1. Chennoukh, S., Sinder, D., and Flanagan, J., "Voice Mimic System'", CAIP Technical Report TR-228, New-Jersey, 1998.

2. Chennoukh, S., Sinder, D., Richard, G., and Flanagan, J., "Voice Mimic System Using Articulatory Codebook For Estimation of Vocal Tract Shape," *EuroSpeech'97*, Patras, Greece, September 1997.

3. Flanagan, J., Ishizaka, K., and Shipley, K., "Signal models for low bite-rate coding of speech'", *J.Acoust. Soc. Am. 68*, pp. 780-791, 1980.

4. Schroeter, J., and Sondhi, M.M., "Dynamic Programming Search of Articulatory Codebooks", *ICASSP*, Glasgow, 1989.

5. Schroeter, J., and Sondhi, M.M., "Speech coding based on physiological models of speech production", in: S. Furui and M.M. Sondhi Eds., *Advances in Speech Signal Processing*, Marcel Dekker, New York, pp. 231-268, 1992.

6. Schroeter, J., and Sondhi, M.M., "Techniques for Estimating Vocal-Tract Shapes from the Speech Signal",

*IEEE trans. on Speech and Audio Processing* 1, pp. 133-150, 1994.

7. Silva, C., and Chennoukh, S., "Estimation of Articulatory Parameter Trajectory from Speech Acoustic Dynamics," To be presented at the *Third International Workshop on Speech Synthesis*, ESCA, 1998.

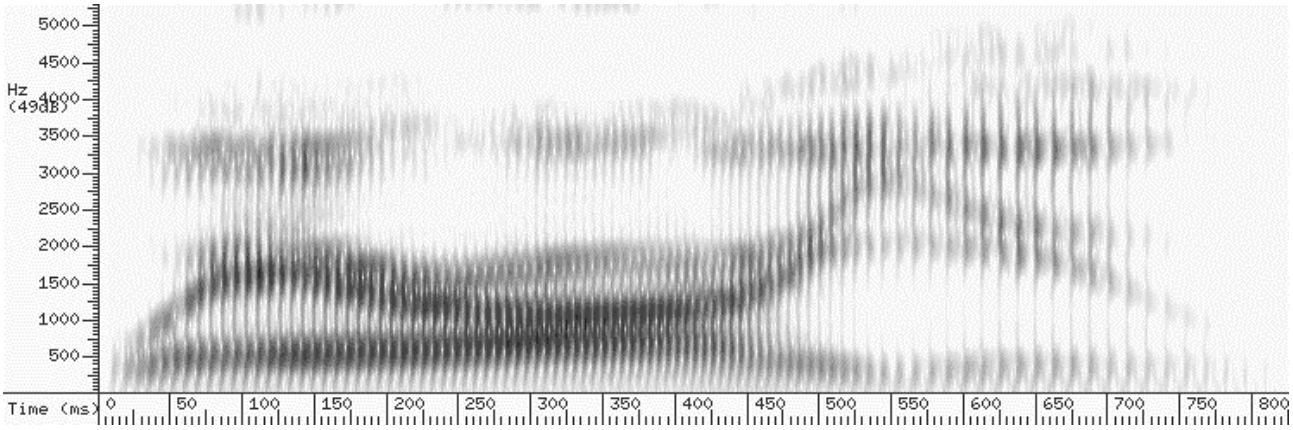


**Figure 4a**: Spectrogram of the natural sentence "*Where are you*".

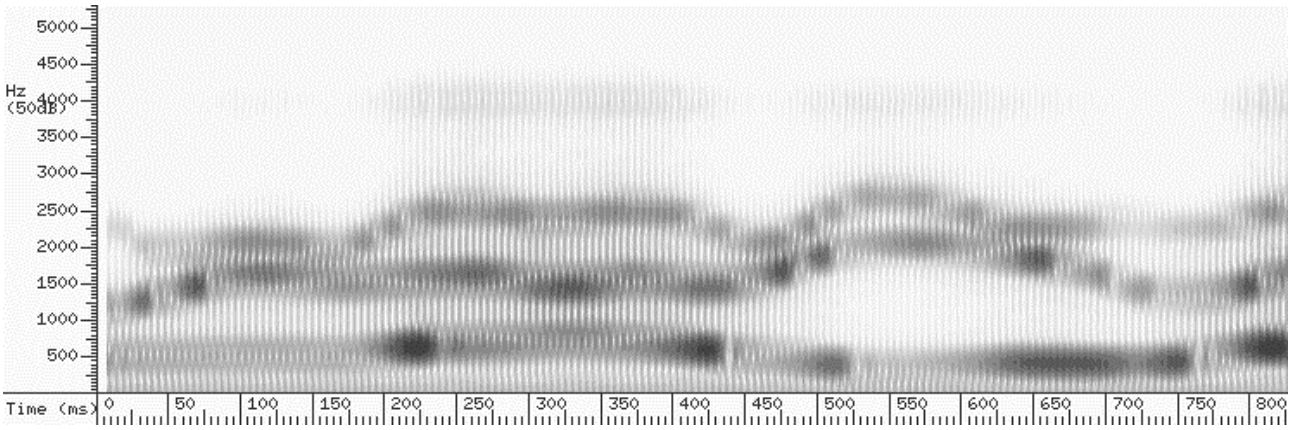


**Figure 4b**: Spectrogram of the mimic sentence "*Where are you*".