Improved Techniques for Voice Mimic Systems Using Articulatory Codebooks

S. Chennoukh, D. Sinder, G. Richard and J.L. Flanagan
Center for Computer Aids for Industrial Productivity (CAIP), CoRe Building, Rutgers University,
Frelinghuysen RD, Piscataway, NJ 08855, USA.

Abstract Voice mimic systems using articulatory codebook require an initial estimate of the vocal tract shape in the vicinity of the global extremum. For this purpose, we need to gather a large set of simultaneous articulatory and acoustic data in the articulatory codebook. Then, the access to this latter becomes a very difficult task. In this paper, a suitable design of an articulatory codebook is presented where an acoustic network sub-samples the acoustic space such that the vocal tract model shapes are ordered and clustered in the network according to the acoustic parameters. Another issue addressed in this paper concerns estimating the trajectory of vocal tract shape time evolution. Since the inverse mapping from acoustic parameters to model shape does not have a unique solution, several vocal tract shape variations are possible. Therefore, a dynamic optimization of the trajectory has been developed. This optimization uses the dynamic properties of each articulatory parameter to estimate the next position.

1 Introduction

The study of speech perception and speech production has been enhanced in the last two decades by the development of computers with high potential of computation. Therefore, Stevens's study toward an articulatory model based speech recognition-synthesis becomes more likely feasible than it was in the early sixties [11]. However, our lack of fully understanding of speech production and the acoustic of speech prevented us from achieving Stevens's goal. The goal was to mimic input speech signals by recognition-synthesis using a model of the vocal tract area function that is like that of a child who can mimic the speech signals he hears without understanding their structure or meaning.

The first attempt at creating a complete computer simulation of an articulatory model speech coding based on an optimization technique was reported by Flanagan and al. [4]. The simulation is called "voice mimic". The voice mimic attempts to provide an articulatory description of the vocal tract that is able to simulate an arbitrary input natural speech and to generate a synthetic signal that, within perceptual accuracy, duplicates the natural one. Central to the effort is the inverse mapping from an acoustic signal to an articulatory description. However, the acoustic-to-articulatory mappings are non-unique and since, given a cost function, the optimization techniques converge only to a local extremum that is at the vicinity of the initial parameters. Therefore, one needs to choose good startup parameters to initialize the optimization procedure. Schroeter and Sondhi [9], who continued along the same lines of Flanagan & al.'s study, used an articulatory codebook proposed earlier by Atal and al. [1]. Since a codebook is used to obtain the first estimates of the vocal tract shape that may produced a given combination of acoustic parameters, one needs to design it such that it spans the total articulatory space of a speaker finely enough so that an acoustic entry always comes very close to the global optimum. Such a codebook requires a large set of matching pairs of vocal tract and acoustic parameters hopefully to cover all the natural articulatory space. Then, the access task to the codebook in order to obtain all possible vocal tract model shapes becomes an issue. For this reason, the voice mimic system needs, in addition to a good articulatory codebook, an efficient procedure for accessing this codebook [6] [7].

The access procedure to the codebook depends on the design of this latter. The usual way to design a codebook is to order the vocal tract model shapes and their corresponding acoustic parameters either according to their iterative order of generation of the model parameter values or according to random generation of these parameter values [9]. Both articulatory codebook designs are access time consuming and this time is randomly distributed.

The number and the positions of the codebook vectors affect the performance of the voice mimic system according to two compromising problems. On one hand increasing the size of the codebook increases the difficulty of the access task and on the other hand reduction of this size can worsen the inverse problem solution [10]. In the second section of this paper, we present a new design of the articulatory codebook where the inversion of the articulatory-to-acoustic mapping is processed during the building of the codebook.

Using the set of articulatory parameter vectors obtained from each input speech frame, Schroeter and Sondhi [7] proposed the dynamic programming to estimate the optimal vocal tract model shape variation path. This dynamic programming gives a good estimate of the vocal tract model shape variation [6] in term of smoothness but it causes in the same time a delay of a certain number of frame time on the speech output [9]. In the third section, we propose another method where the articulatory parameter are estimated within one input speech frame delay. This

method relies on the dynamic properties of the articulators.

2 Design of an articulatory codebook

The difficulty in using an articulatory codebook for the voice mimic can be summarized in three different issues. First, the vocal tract shapes are more likely ordered in the articulatory domain while the access to these shapes is done from the acoustic parameters according to which the shapes are randomly positioned. Second, the acoustic-to-articulatory mappings are generally non-unique. So the larger is the codebook the more possible shapes you may obtain for each speech signal frame. Third, the centroid of a given set of model shape parameter vectors in the articulatory domain does not point to the centroid of the corresponding vectors in the acoustic domain. Because of all these reasons, one talks about reducing the size of the codebook [8][10] and the other talks about vocal tract shape clustering [5] [9] in order to reduce the access time in the codebook.

Since there is not a complete knowledge about the articulatory space as we do not know what are its orthogonal articulatory parameters, we should free a codebook design from the limitation on the size and allow the population of the codebook by all possible model shapes physiologically realistic. However we should also find a suitable technique to access, in a reasonable time, to the model shapes that we are looking for. Then, the goal becomes a database system management issue that looks for the best technique for clustering and searching the articulatory codebook.

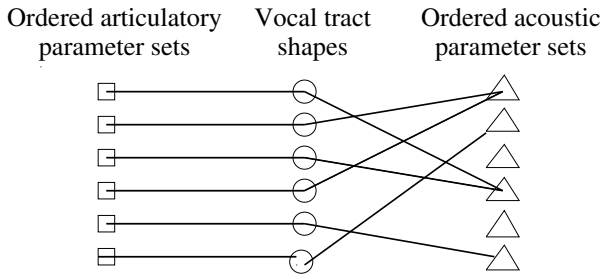

Figure 1: Vocal tract shapes ordered according to both articulatory and acoustic parameters in the articulatory codebook.

The simplest technique, we can start with, is to subsample the acoustic space on a set of ordered clusters and process the inversion of the articulatory-to-acoustic mapping during the building of the articulatory codebook by defining the corresponding acoustic network (figure 1). Each node of the network should point on all the model shapes in the data-base that have acoustic parameters closer to the acoustic centroid representing the node.

Once the codebook is built up, the access task is question of estimating the three first formant frequencies for each frame of the speech signal, determining the coordinates of the corresponding cluster in the network using the sub-sampling period of each formant, and finally pointing to the corresponding node in the acoustic network to get all the possible vocal tract model shapes contained in the articulatory codebook. Once the acoustic parameters are obtained, the matching set of shapes are obtained in few milliseconds. Furthermore, this access time to the searched cluster of shapes does not vary significantly versus of the size of the articulatory codebook.

3 Forward dynamics of the articulatory parameters

The non-uniqueness of the acoustic-to-articulatory mappings spans a non-uniqueness in the vocal tract shape variation trajectory. One needs to address this issue to select the optimal trajectory for the vocal tract shape variation. Based on the slow evolution of the articulation between two successive signal frames, Schroeter and Sondhi [7] proposed a dynamic programming for vocal tract path optimization that relies on the closest vocal tract model shape. The dynamic programming was used to match between a reference and a test speech frame sequence for word recognition. However, this technique includes a delay on the speech voice mimic output and does not take into account directly the physical dynamic features of the articulators that are actors of the speech production.

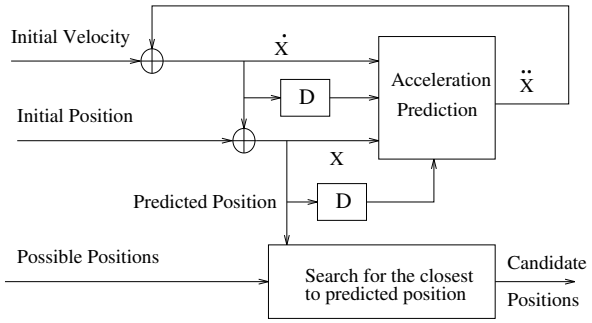

Figure 2: Path optimization network of a single articulatory parameter.

By studying the articulator motion from muscle activity, Bateson and al. [2] described a recurrent algorithm to estimate the position of each articulator from continuous EMG signal. In this study, we propose the same network to our study. The network takes into account the dynamic properties of the articulators and perform the forward dynamics of the articulatory parameters according to the slow variation of their respective acceleration during speech

production (figure 2). The following articulatory parameter position is then estimated from the previous position, velocity and acceleration of the articulatory parameter. The estimate is compared to the different parameter positions of the vocal tract shapes proposed by the articulatory codebook. Then the shape that has all its articulatory model parameters in the candidate positions should be the next vocal tract model shape. This technique led to a recurrent algorithm for path optimization of the vocal tract model shape time evolution.

4 Articulatory speech coder

An articulatory codebook that maps the Ishizaka's vocal tract model parameters (figure 3) to the first three formant frequencies has been built. Table 1 gives the formant frequency limits considered as limits of the acoustic network and also gives the subsampling period (step) of each formant frequencies used to determine the dimension of the acoustic network and to define the cluster coordinates. The subsample period represents also the resolution of the acoustic-to-articulatory mapping. Table 2 gives the limits of the Ishizaka's model parameters used to generate 46,080 shapes in the codebook.

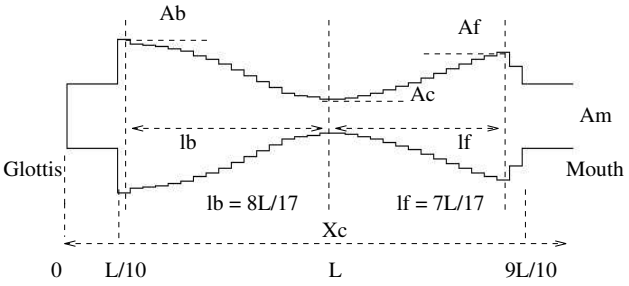

Figure 3: Ishizaka's vocal tract area function model.

Form. Freq.	From (Hz)	To (Hz)	Step (Hz)
F1	150	800	50
F2	300	2900	50
F3	1500	4500	50

Table 1: Acoustic space network of the first three formant frequencies

Model Param.	From	To	Step
Xc	4 cm	13.5 cm	0.5 cm
Ac	0.1 cm^2	3.5 cm^2	0.2 cm^2
Am	0.5 cm^2	8.0 cm^2	0.2 cm^2
Af	10.0 cm^2	10.0 cm^2	0 cm^2
Ab	8.0 cm^2	8.0 cm^2	0 cm^2
L	17.5 cm	17.5 cm	0 cm

Table 2: Articulatory space network of the Ishizaka's area function model

The articulatory codebook is supposed to span the articulatory space. In the case where this hypothesis is verified, the articulatory model voice mimic system will never fall on an empty node (i.e., a node that contain no vocal tract shape). However, this situation of empty node is still possible as long as we do not have an articulatory model that help cover the articulatory space. Thus, this situation should be prevented in the access management. In this case, one obviously searches for the vocal tract model shapes that are in the acoustic clusters of the neighborhood. The search should rely on perceptual effects of the acoustic parameters.

In the case of our study, we used the formant just-noticeable-difference (JND) box measure [3]. So the search procedure should look for the cluster whose the formant frequencies are not farther away from the original than one JND.

In the forward dynamic, the estimation of the closest model shape to the predicted one, among the shapes proposed by the codebook, is obtained by a dynamic threshold for all the model parameters. The threshold increases or decreases according, respectively, if we have a null number of shapes or more than one shape as candidate. This threshold is adjusted until, we obtain one model shape as candidate for the next vocal tract shape in the time sequence. The resulted threshold is used for the next frame articulatory analysis.

5 Results

A sentence "Why were you away?" was spoken by a male speaker. The signal is sampled at 16 kHz and is windowed by a 20 ms Hamming window with 10 ms overlap. Levinson algorithm is used to compute the 23rd order linear prediction model coefficients. Newton-Raphson method estimates the poles of the model. The obtained formant frequencies are then given to the articulatory codebook access procedure to determine the acoustic cluster node which content the model shapes. The set of model shapes are finally filtered by the forward dynamic network to come out with the optimal shape for the present frame of the vocal tract time evolution. Figure 4 shows the results obtained for the time evolution of the vocal tract shape obtained for the spoken sentence.

6 Conclusions

In this paper, we described a suitable design of an articulatory codebook that allows a fast access to vocal tract model shapes that best match with given acoustic parameters. In this study, we used a simple model of the vocal tract area function for the voice mimic. The resulted articulatory codebook is able to approximate the shapes of vowel like phonemes, namely, voiced and with no more than one constriction along

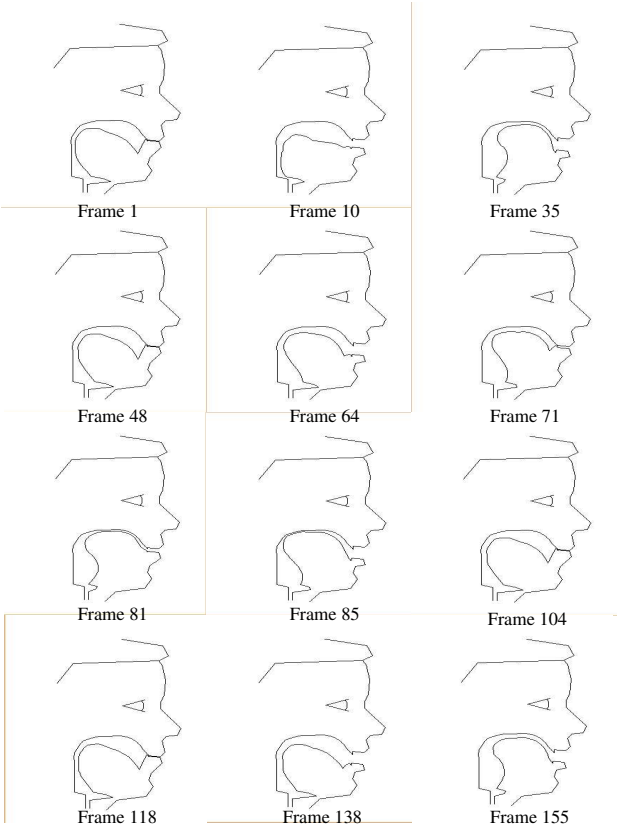

Figure 4: Vocal tract shape time evolution for "Why were you away?" spoken by a male speaker.

the vocal tract. In order to extend this application for more complicated vocalic phoneme shapes, one needs to use an articulatory model with a generation procedure of realistic vocal tract shape. This requires distinguishing between realistic shapes and unrealistic shapes. Furthermore, an investigation of an articulatory model that covers the articulatory space should provide a material for progress toward a robust voice mimic for speech coding.

A forward dynamic network is proposed to estimate the shape variation trajectory. The predictor is based only on the acceleration of the articulatory parameters independently of each others. It already gives good results in term of articulatory features matching the succession of the phoneme contained in the sentence and in term of smoothness of the vocal tract shape time evolution. However, since we do not have any other performance measure of the voice mimic results regarding the coarticulation, we should improve the predictor such that it integrates the dynamic constraints of the vocal tract articulators. This should include on one hand the relationship between the articulatory parameters them-self and on the other hand the articulatory dynamic constraints.

Acknowledgment

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

References

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

[2] E. Bateson, M. Hirayama and Y. Wada, 1993, Generating Articulator Motion from Muscle Activity Using Artificial Neural Networks, *ATR HIP Res. Labs.* 2, pp. 264-274.

[3] J.L. Flanagan, 1955, A difference limen for vowel formant frequency, *J. Acoust. Soc. Am.* 27, pp. 613-617.

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

[5] J. Larar, J. Schroeter and M.M. Sondhi, 1988, Vector Quatification of the Articulatory Space, *IEEE trans. on Acoustic, Speech and Signal Processing*

[6] G. Richard, M. Goirand, D. Sinder, J. Flanagan, 1997, Simulation and Visualization of Articulatory trajectories estimated from speech signals, *International Symposium in Simulation, Visualization and Auralization for Acoustic Research and Education*, Tokyo, Japan.

[7] J. Schroeter and M.M. Sondhi, 1989, Dynamic Programming Search of Articulatory Codebooks, *ICASSP*, Glasgow.

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

[9] J. Schroeter and M.M. Sondhi, 1994, Techniques for Estimating Vocal-Tract Shapes from the Speech Signal, *IEEE trans. on Speech and Audio Processing* 1, pp. 133-150.

[10] V.N. Sorokin and A.V. Trushkin, 1996, Articulatory-to-Acoustic mapping for inverse problem, *Speech Communication* 19, pp. 105-118.

[11] K. Stevens, 1960, Toward a Model for Speech Recognition, *J.Acoust. Soc. Am.* 32, pp. 47-55.