VOICE MIMIC SYSTEM USING AN ARTICULATORY CODEBOOK FOR ESTIMATION OF VOCAL TRACT SHAPE

S. Chennoukh, D. Sinder, G. Richard and J.L. Flanagan*
Center for Computer Aids for Industrial Productivity (CAIP), Rutgers University,
Piscataway, NJ 08855-1390, USA
*Matra-Communication, rue J.P. Timbaud, 78392 Bois d'Arcy, France
Tel. +1 908-445-0080, FAX: +1 908 445-4775, E-mail: chenoukh@caip.rutgers.edu

ABSTRACT

Voice mimic systems using articulatory codebooks require an initial estimate of the vocal tract shape in the vicinity of the global optimum. For this purpose, we need to gather a large set of corresponding articulatory and acoustic data in the articulatory codebook. Thus, searching and accessing the codebook becomes a difficult task. In this paper, the design of an articulatory codebook is presented where an acoustic network sub-samples the acoustic space such that vocal tract model shapes are ordered and clustered in the network according to acoustic parameters. Another issue addressed in this paper concerns estimating the trajectory of vocal tract shapes as they change with time. 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 trajectories has been developed. This optimization uses 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 capable of large amount of computation. As a result, Stevens' study towards an articulatory model for speech recognition-synthesis becomes more feasible than it was in the early sixties ([9]). However, an incomplete understanding of speech production and the acoustics of speech prevented us from achieving Stevens' goal. The goal was to mimic input speech signals by recognition-synthesis using a model of the vocal tract area function that can mimic the speech signals without understanding their structure or meaning.

An early attempt at creating a complete computer simulation of articulatory model speech coding using an optimization technique was reported by Flanagan et al. ([4]). The simulation is called "voice mimic". The voice mimic attempts to provide an articulatory description of the vocal tract that corresponds to an arbitrary natural speech input 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, acoustic-to-articulatory mappings are non-unique and, given a cost function, the optimization techniques converge only to a local extremum that may be near the vicinity of the initial parameters. Therefore, one needs to choose accurate startup parameters to initialize the optimization procedure. Schroeter and Sondhi ([8]), who continued along the same lines of Flanagan et al.'s study, used an articulatory codebook proposed earlier by Atal et al. ([1]). Since a codebook is used to obtain the first estimates of the vocal tract shape that may produce a given combination of acoustic parameters, it must be designed such that it spans the natural articulatory space of a speaker. Furthermore, sampling of the space must be fine enough so that an acoustic entry always exists very close to the global optimum. Such codebooks require a large set of matching pairs of vocal tract and acoustic parameters. The complexity of searching a large codebook for 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 the codebook ([6],[7]).

The number and position 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 complicates the inverse problem solution. In the second section of this paper, a new design of the articulatory codebook is presented for which the inversion of the articulatory-to-acoustic mapping is processed during the building of the codebook. This codebook design allows real-time access to the set of acoustically equivalent shapes, regardless the size of the codebook.

Since the inverse mapping from acoustic parameters to model shape does not have a unique solution, several vocal tract shape variations are possible. Schroeter and Sondhi ([7]) proposed the use of dynamic programming to estimate the optimal trajectory of the vocal tract model shape variation path. The dynamic programming requires a delay of several data frames for the speech output ([8]). In the third section, a method is proposed where the articulatory parameters are estimated within one frame. Section

four describes the articulatory speech coder that was built using the new codebook design and the dynamic optimization technique of the vocal tract shape variation. Section five gives test results obtained using this coder. Finally, the conclusion discusses the performance of the articulatory coder and the perspectives of the present study.

2. DESIGN OF AN ARTICULATORY CODEBOOK

The difficulty with using an articulatory codebook for the voice mimic can be summarized within 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 in the codebook. Second, the acoustic-to-articulatory mappings are generally non-unique. Thus, as the codebook grows in size the more shapes are obtained for each speech frame. Third, the centroid of a given set of articulatory parameter vectors does not point to the centroid of the corresponding vectors in the acoustic domain. Because of these reasons, attempts have been made at reducing the size of the codebook ([8]) and at vocal tract shape clustering ([5]) in order to reduce codebook access time.

Since orthogonal parameters for the vocal tract shape are not well established, the codebook design should be free of size limitation. Then, the codebook can be populated with all physiologically realistic model shapes. However a suitable technique for accessing the codebook in a reasonable amount of time must be found. Then, the problem is reduced to a database management issue that looks for the best technique for clustering and searching the vocal tract shapes.

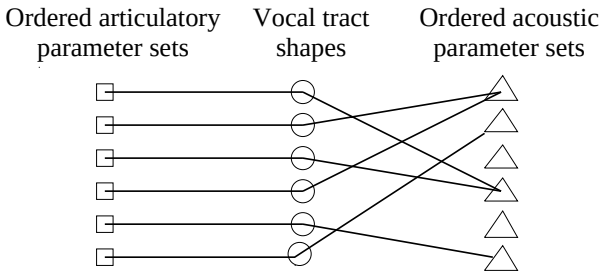

Figure 1: Schematic representation of vocal tract shape clustering into an acoustic network.

The simplest technique is to sub-sample the acoustic space on a set of ordered acoustic clusters that gives rise to the acoustic network (Figure 1). The inversion of the articulatory-to-acoustic mapping is processed during the building of the articulatory codebook as follows. For each generated vocal tract shape, acoustic parameters are determined. Using the sub-sampling period for each acoustic parameter, the closest node in the network is determined and notified about the position of the shape in the codebook. Thus, each node of the network points to all the model shapes in the codebook that have acoustic parameters close to the acoustic centroid represented by the node.

Once the codebook is built, the access task simply requires estimating the acoustic parameters for each frame of the speech signal, determining the coordinates of the corresponding cluster node in the network using the sub-sampling period of each parameter, and retrieving all possible vocal tract model shapes to which the acoustic node points.

Once the acoustic parameters are obtained, the matching shapes are obtained in a few milliseconds on a SPARC 5 workstation. Furthermore, this access time to the cluster of shapes does not vary significantly with the size of the articulatory codebook.

3. FORWARD DYNAMIC OF THE ARTICULATORY PARAMETERS

The non-uniqueness of the acoustic-to-articulatory mappings leads to a non-uniqueness in the vocal tract shape variation trajectory. One needs to address this issue to select the most probable vocal tract shape variation. Based on the slow evolution of the articulation between two successive signal frames, Schroeter and Sondhi ([7]) proposed dynamic programming for vocal tract path optimization that relies on the closest vocal tract model shape. However, this technique imposes a delay on the voice mimic output and does not take into account directly the physical dynamic features of the articulators.

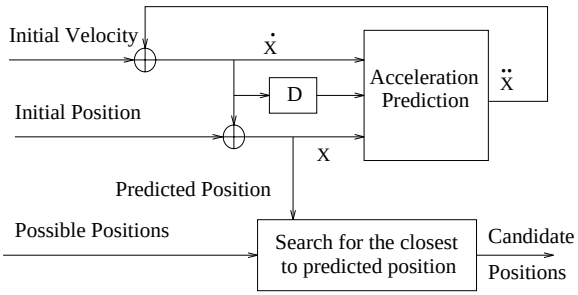

Figure 2: Path optimization network for a single articulatory parameter.

By studying the articulator motion from muscle activity, Bateson et al. ([2]) described a recurrent algorithm to estimate the position of each articulator from continuous EMG signals. In this study, a similar network is proposed. The network takes into account the dynamic properties of the articulators and performs 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, and from the velocity and acceleration of the artic-

ulatory parameter. The estimate is compared to the different parameter positions of the shapes proposed by the articulatory codebook. Then, the shape that has its articulatory model parameters in the candidate positions is chosen as the next vocal tract model shape. This technique leads to a recurrent algorithm for optimization of the vocal tract model shape time evolution.

4. ARTICULATORY SPEECH CODER

An articulatory codebook that maps 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 sub-sampling period (step) of each formant frequency used to determine the dimension of the acoustic network and to define the cluster coordinates. The sub-sample period also represents the resolution of the acoustic-to-articulatory mapping. Table 2 gives the limits of Ishizaka's model parameters used to generate 46,080 shapes in the codebook.

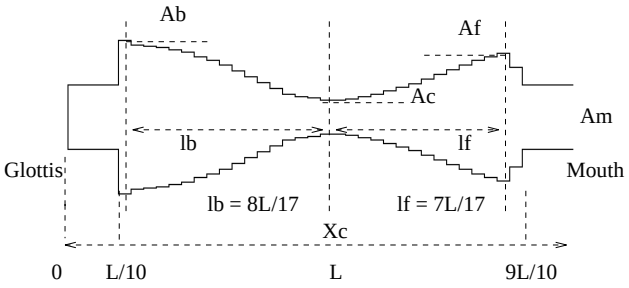

Figure 3: Ishizaka's vocal tract area function model.

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 limits 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 limits for Ishizaka's area function model

Ideally, the codebook should span the articulatory space. If this is true, the voice mimic system will never point to an empty acoustic node (i.e., a node that contain no vocal tract shape). However, in practice an empty node is still possible since the articulatory model does not cover the entire articulatory space of natural speech. Thus, this situation should be prevented in the codebook access management. In this case, one obviously searches for the vocal tract model shapes that are in acoustic clusters which are neighbors of the desired node. In addition, perceptual effects are considered when choosing the cluster. In the case of this study, the formant just-noticeable-difference (JND) measure ([3]) is used. At first, the search procedure looks for the cluster whose formant frequencies are not farther away from the original than one JND. If this search fails the procedure goes to two JND's and so on, until a non-empty node is found.

In the forward dynamic network, the closest model shape to the predicted one must be chosen from among the shapes proposed by the codebook. This is accomplished using a dynamic threshold for all model parameters. The threshold increases if there is a null number of candidate shapes and decreases if there is more than one candidate shape. The threshold is adjusted until one candidate shape is obtained for the next vocal tract shape in the time sequence. The resulting threshold is used to initialize the next frame's articulatory analysis.

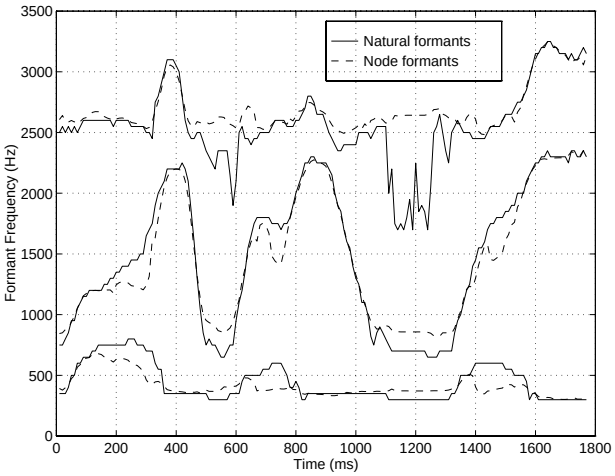

Figure 4: Superposition of the natural formant frequency trajectories (solid) and the acoustic node coordinates (dash) for "Why were you away?" spoken by a non-native english male speaker.

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. The Levinson algorithm is used to compute the 23rd order linear prediction model coefficients. The Newton-Raphson method is used to estimate the poles of the model. The obtained formant frequencies are then given to the articulatory codebook search procedure to determine the acoustic cluster node which, in turn, points to the model

shapes in the codebook. The set of model shapes are finally filtered by the forward dynamic network which outputs the optimal shape for the present frame of the vocal tract time evolution. Figure 4 shows the formant frequency trajectories obtained from the signal processing and the acoustic node coordinate trajectories obtained from the articulatory speech coder. The gaps between the natural and the articulatory speech coder formant trajectories are due to the lack of model shapes having those natural formant frequency value combinations. Thus, the forward dynamic network chooses different and closer trajectories until the natural speech again appears with formant frequencies for which the articulatory codebook has matching model shapes. Figure 5 shows the time evolution of the vocal tract shape obtained for the spoken sentence.

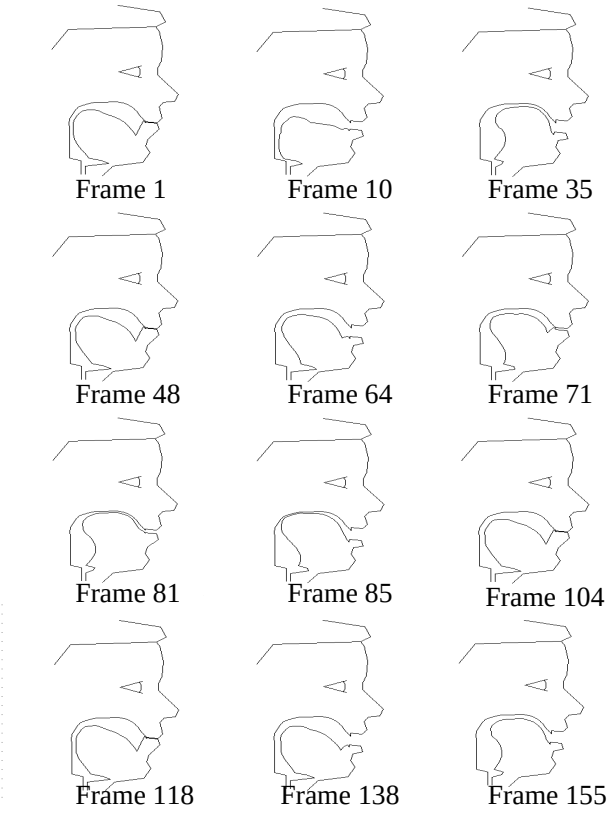

Figure 5: Vocal tract shape time evolution for "Why were you away?" spoken by a male speaker.

6. CONCLUSIONS

This paper describes the design of an articulatory codebook that allows real-time access to vocal tract model shapes that best match natural acoustic parameters. A simple model of the vocal tract area function was used for this study. The resulting articulatory codebook is able to approximate shapes such as vowels or diphtongs. In order to extend this application to more complicated vocalic phoneme shapes, one needs to use an articulatory model with more degrees of freedom capable of generating all possible realistic shapes. This requires distinguishing between realistic and unrealistic shapes. Furthermore, investigation into articulatory models that span the articulatory space will provide material for progress toward a robust voice mimic for speech coding.

A forward dynamic network is utilized to estimate the vocal tract shape time evolution. The prediction is based only on the acceleration of the articulatory parameters, each independent of the others. The system performs well for vowel strings in terms of articulatory features matching the succession of the phonemes contained in the spoken sentence and in term of smooth vocal tract trajectories. However, no performance measure exists regarding coarticulation. The prediction should be improved such that it integrates the dynamic constraints of the vocal tract articulators. The improvements should include relative relationships between articulatory parameters as well as constraints on the articulatory dynamics.

ACKNOWLEDGMENT

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

7. REFERENCES

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

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

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

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

[5] J. Larar, J. Schroeter and M.M. Sondhi, "Vector Quatification of the Articulatory Space, *IEEE trans. on Acoustic, Speech and Signal Processing"*, pp. 1812-1818, 1988.

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

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

[8] J. Schroeter and M.M. Sondhi, "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, 1992.

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