The Third ESCA/COCOSDA Workshop
(ETRW) on Speech Synthesis
Jenolan Caves House, Blue Mountains,
NSW, Australia, November 26-29, 1998

ESTIMATION OF ARTICULATORY PARAMETER TRAJECTORY FROM SPEECH ACOUSTIC DYNAMICS

C. Silva[1] *S. Chennoukh*[2]

[1]Departamento de Electrónica Industrial, Universidade do Minho, Portugal
[2]Center for Computer Aids for Industrial Productivity (CAIP),
Rutgers University, Piscataway, NJ 08854-8088, USA

ABSTRACT

This research aims to perform articulatory analysis as a basis for low bit-rate speech coding. The classical approach consists of gathering a large set of acoustic and articulatory vector pairs in a codebook. Then, based on some criteria, the non-uniqueness of the articulatory trajectories is solved using a dynamic optimization procedure. An articulatory codebook requires a model capable of generating shapes for all possible speech sounds. This paper reports a new approach for incorporating the tongue tip on Ishizaka's vocal tract area function model in order to reproduce more details of certain classes of consonants. Results are reported on the search for an optimized articulatory codebook using different methods of model parameter sampling.

1. INTRODUCTION

Research in speech production collects vast amount of data on the vocal system, its mechanics, the acoustic signal and the information that it encodes. The analysis of this data should establish the characteristics of the speech production device in order to model it. Due to the difficulty of obtaining articulatory data, techniques for estimating the vocal tract area function directly from the speech signal are of interest in studies of the speech production process and as the basis for efficient low bit-rate coding of the speech signal. The problem of estimating the vocal tract shape from the acoustic speech signal is often referred to as the inverse problem. This is a difficult problem because of non-uniqueness of the acoustic-geometry relation.

Our efforts aim to solve the inverse problem by an optimization procedure using an articulatory codebook. A codebook is used to obtain the first estimate 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. However, as the codebook size increases, searching it becomes increasingly complex with increasing number of matching shapes for each entry. This research is devoted to the study of an optimum design of an articulatory codebook. In this paper, we present our results on different codebooks we built using an improved vocal tract model and more appropriate model parameter samplings.

Section 2 discusses the construction of an improved codebook including a new vocal tract area function model. It also gives the different methods we used to sample the model parameters. Results are presented in section 3 on the use of codebooks built with different parameter sampling techniques. Finally, section 4 presents the conclusion of this study.

2. DESIGN OF AN OPTIMIZED ARTICULATORY CODEBOOK

The difficulty with using an articulatory codebook for the voice mimic can be summarized within four 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 seem randomly positioned in the codebook. Thus it would be more convenient to access the codebook if the model parameter vectors were sorted in the acoustic domain. Second, the acoustic-to-articulatory mappings are generally non-unique. Thus, as the codebook grows in size, more shapes are obtained for each speech frame. Third, the mapping is non-linear. Therefore, the interpolation between a pair of articulatory vectors does not match the trajectory obtained from the interpolation between the corresponding pair of acoustic vectors. Fourth, the centroid of a given set of articulatory parameter vectors does not point to the centroid of the corresponding vector in the acoustic domain and also the extent of division of the matching shapes around the centroid is not uniform. Because of these reasons, attempts have been made at reducing the size of the codebook (Schroeter, 1992) and at vocal tract shape clustering (Larar, 1988) 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.

In our previous study, the access and search of a codebook is simplified using an acoustic network that sub-samples the acoustic space (Chennoukh, 1997). The vocal tract model shapes are clustered during their generation in the network that sub-samples the acoustic space. Thus, for each input frame acoustic feature vector, the network is accessed straight forward to obtain all matching model shapes.

2.1. Vocal tract area function model

During speech, the surface of the tongue, which defines almost the whole vocal tract area function, assumes hundreds of subtly different shapes combined with different openings of the lips. Each of these shapes is a particular non-uniform curve, usually one having no simple representation in terms of words or equations. Nonetheless, it becomes clear by simple examining the curves, that they are related to one another in some kind of orderly way. For example, there appear to be systematic patterns of relationships among the tongue shapes used by a given speaker for producing different vowels, as well as patterns relating the shapes used by different speakers to produce the same vowel. Such patterns of tongue shape variation are of great interest to the vocal tract modeling.

One method for describing tongue shapes represented the vocal tract area function by three parameters, the width of the vocal tract at its narrowest point that is specified by the distance of the point from the glottis, the vocal tract area at this point and the lip opening (Dunn, 1950; Stevens, 1955; Fant, 1960; Flanagan, 1980). In our previous studies (Chennoukh, 1997), we used Ishizaka's area function model to describe the vocal tract shape (Flanagan, 1980). We extended this model with two additional parameters, the position of and area at the tongue tip, figure 1. The introduction of these two parameters allows to describe constrictions and occlusions at the alveolar region with more details on the vocal tract area function. This region plays a significant role in the production of stop and fricative sounds.

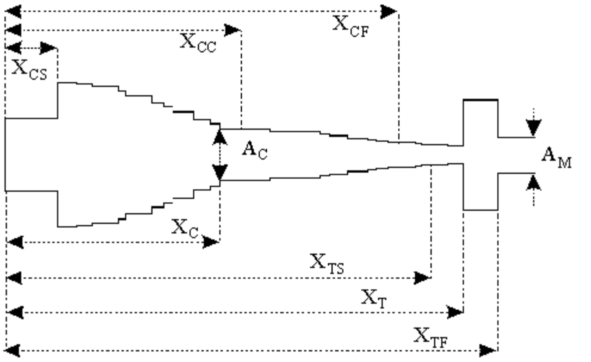

Figure 1: New vocal tract area function model.

The tongue tip is described by its area and its position. The position is obtained using the position of the tongue body by mean of the constant volume of the tongue. When the constriction is around the velar position, as in the production of /o/ and as the volume of the tongue is constant, the tongue tip is pulled to the back and represents one extremity of the tongue tip possible theoritical displacement x_{ts} (Chennoukh, 1995). When the constriction is in the back, as in /a/, or in the front, as in /i/, the tongue tip rests forward to the teeth whose position is represented by the variable x_{tf}. This behavior is described using the following equation:

$$x_t(x_c) = x_{tf} - d_x(x_c),$$

The position of the tongue tip relative to the teeth is represented by x_{tf}, which is measured from the glottis. The displacement, d_x, is given by:

$$d_x(x_c) = \begin{cases} \sqrt{\frac{A}{\pi}} \times \frac{x_c - x_{cs}}{x_{cc} - x_{cs}} & , \ x_c < x_{cc} \\ \sqrt{\frac{A}{\pi}} \times \frac{x_{cf} - x_c}{x_{cf} - x_{cc}} & , \ x_c \geq x_{cc} \end{cases} ,$$

where x_{cc} is the velar position, x_{cs} and x_{cf} are respectively the maximum and minimum position of the constriction with the body of the tongue, x_c. The function A is obtained by:

$$A = \begin{cases} A_n - A_v & , \ A_v < A_n \\ 0 & , \ A_v \geq A_n \end{cases} ,$$

where A_n is the neutral area, and A_v is the area at velar position. These equations were empirically obtained. They model the position of the tongue tip as a function of the volume and the position of the body of the tongue.

The introduction of the tongue tip in the vocal tract model creates an alveolar cavity, A_{alv}. The area of this cavity is fixed at 2.0 cm^2 larger than the area at the lips, A_m. The region between the position of constriction, x_c, and the tongue tip position, x_t, is represented by a superposition of a line and an arc of sine between the coordinates (x_c, A_c) and (x_t, A_t) as follows:

$$Area(x) = k_1 \sin\left(\pi \frac{x - x_c}{x_t - x_c}\right) + \frac{A_t - A_c}{x_t - x_c} \times (x - x_c) + A_c.$$

The factor k_1 gives the amplitude of the arc of sine. It varies according to the distance between the two constrictions:

$$k_1(x) = 0.41 \times (x - 1.5),$$

where x is the Euclidean distance between the two coordinates:

$$x = \sqrt{(x_t - x_c)^2 + (A_t - A_c)^2}.$$

The complete equation of the vocal tract area function model is given in appendix.

Therefore, when A_t is larger than $A_{t_{max}}$, this model behaves exactly as Ishizaka's model, as over this area the tongue tip is not acoustically relevant.

2.2. Methods of sampling the parameters of the model.

Articulatory codebook performance depends on its size, the difference of the vocal tract model shapes and the coverage of the acoustic space. The larger is the number of shapes that are different the more coverage of articulatory space we obtain. The coverage of the acoustic space is the main target in building the codebook. Namely, we need a codebook capable to provide a cluster of shapes for any arbitrary acoustic features. For this purpose, the sampling techniques of model parameters play an important role in spreading the model shapes on the acoustic space. Different techniques have been considered: linear, logarithmic, two empirical techniques and combinations of all these techniques.

The linear sampling method consists in the division of the range in a constant step, given a number of samples.

In the logarithmic sampling method, we fix a number of samples and apply a logarithmic scale. If the parameter is an area, the sampling was done in such a way that the sampling is dense in lower values. In the case of the position of constriction, the sampling was done such that most of the sampling period is shorter near the teeth.

In the percentage sampling method, one of the empirical techniques, the next sample is given by:

$x_{next} = (1 + k) \times x_{prev}$,

where $0.0 < k < 1.0$.

In the irregular sampling method, the next sample is obtained as $x_{next} = k \times x_{prev}$. The multiplier factor, k, is updated at each iteration according to the following expression:

$$k_n = \begin{cases} \lceil \frac{k_{n-1}}{2.0} \rceil & , \; k_{n-1} > 3 \\ 1.5 & , \; k_{n-1} \leq 3 \end{cases} .$$

The operator $\lceil . \rceil$ return the smallest integer greater than its argument.

All these methods have as a goal to obtain most of its samples in the range of small areas (areas > 1.0). The irregular sampling methods were not employed for the position of the constriction, since they would obtain too many samples near the teeth.

3. RESULTS

In order to assess the new vocal tract model and parameter sampling methods compared to the Ishizaka's model, we tested several codebooks. The performance of these codebooks is represented as a function of parameter sampling and as a function of size. Table 1 shows the range for each parameter of the vocal tract model used to build the codebooks.

Parameter	*Minimum value*	*Maximum value*	*Unit*
A_c	0.3	4.0	cm^2
x_c	4.0	13.5	cm
A_m	0.5	8.0	cm^2
A_{tip}	0.001	4.0	cm^2

Table 1: Range of the variation of the parameters of the models.

The codebooks were evaluated according to the percentage of first access hit. The percentage of access hit is defined as the percentage of number of frames in the sentence for which a cluster of shapes is found in the node pointed by the formant frequencies of each frame.

Three test sentences were used, "i a i a o", "who are you" and "Where are you". They were spoken by three different male speakers. The signals were sampled at $16kHz$ and windowed by a $32ms$ Hamming window with $15ms$ overlap. Levinson-Durbin's algorithm was used to compute the $18^{\underline{th}}$ order linear prediction coefficients. Newton-Raphson's method was used to estimate the poles of the transfer function of the model. The obtained formants were used to access and search the codebook. Then, the codebook provides the best match cluster of model shape parameter vectors.

Table 2 shows our result collection of codebook testing using Ishizaka's model and our upgraded model with different parameter sampling techniques. In order to evaluate the models and the model parameter sampling techniques, we perform the comparison at comparable codebook sizes. If we consider the linear parameter sampling, the use of the new model shows better performance than Ishizaka's model except for the second sentence "Who are you?". The new model has one more parameter than Ishizaka's model. For comparable sizes, the codebook built using the new model has lower resolution on the parameters A_c, x_c and A_m. Therefore, some degradation in performance of the codebooks built using the new model is observed for small codebook sizes.

We can notice from Table 2 that as the size of the codebook increases the codebook performance using the new model increases. Indeed, the codebook *cdb-64.3* performs 18% better than *cdb-64.1*, which is built using Ishizaka's model. Comparing the codebooks *cdb-10.3* and *cdb-64.1* applied to the third sentence, we show that the introduction of the tongue tip allowed a codebook to outperform other codebooks based on Ishizaka's model that are 8 times larger.

According to the results of using different parameter samplings, it is also clear that the logarithmic sampling method and logarithmic combined with irregular technique were the methods that outcome with the best results. However, aditional tests are needed to conclude on the best technique. We also showed that the difference in parameter sampling methods decreases for large codebook sizes.

4. CONCLUSIONS AND FUTURE WORK

In this article we have reported our results in the construction of an improved codebooks. We verified that the inclusion of a parameter to describe constriction with the tongue tip on the vocal tract area function model increased the performance of the analysis. We, also, verified that the choice of an appropriated parameter sampling method was relevant to the performance of the codebook. However, we need to perform the tuning of some constants in our model with physiological data. It is also necessary to design more tests with the new model and with the sampling methods with specific classes of phonemes in order to assess broadly the efficiency of the model and of the methods. However, to succeed in this task we need to change the access method from formants to another acoustic feature, since it is very difficult, or even impossible, to obtain the formants for certain speech signals.

5. REFERENCES

1. S. Chennoukh, D. Sinder, G. Richard, and J. Flanagan, "Voice mimic system using articulatory codebook for estimation of vocal tract shape," *EuroSpeech'97*, Patras, Greece, Septembre 1997.

2. 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.

3. 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.

4. S. Chennoukh, "Modélisation du conduit Vocal en Régions Distinctives. Synthèse d'ensembles Voyelle-Voyelle et Voyelle-Consonne-Voyele," Doctorate Dissertation, ENST, Paris, 1995.

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

6. G. Fant, "Acoustic Theory of Speech", Mouton eds-Grevenhage, The Netherlands, 1960.

7. H. Dunn, "The calculation of vowel resonances", *J.Acoust. Soc. Am.* 22, pp. 740-753, 1950.

8. K. Stevens and A. House, "Development of a quantitative of vowel articulation", *J.Acoust. Soc. Am.* 27, pp. 484-493, 1955.

$$Area(x) = \begin{cases} 2.0 & , \; x < x_{cs} \\ \frac{A_b + A_c}{2} - \frac{A_b - A_c}{2} \cos\left(\pi \frac{x_c - x}{l_b}\right) & , \; x_{cs} \leq x \leq x_c \\ \frac{A_f + A_c}{2} - \frac{A_f - A_c}{2} \cos\left\{\pi \left[0.4 + 0.6 \frac{x - x_c}{l_f}\right] \frac{x - x_c}{l_f}\right\} & , \; x_c < x \leq x_{tf} \wedge A_t \geq A_{t_{max}} \\ k_1 \sin\left(\pi \frac{x - x_c}{x_t - x_c}\right) + \frac{A_t - A_c}{x_t - x_c} \times (x - x_c) + A_c & , \; x_c < x \leq x_t \wedge A_t < A_{t_{max}} \\ A_m + A_{alv} & , \; x_t < x \leq x_{tf} \wedge A_t < A_{t_{max}} \\ A_m & , \; x_{tf} < x < L \end{cases}$$

Equation 1: Complete equation of the vocal tract area function model.

				Sentence I	**Sentence II**	**Sentence III**
Codebook	*Model*[1]	*Sampling*[2]	*Size*	*Hits (%)*	*Hits (%)*	*Hits (%)*
Cdb-10.1	Ish	L3	8320	43.1	28.8	6.6
Cdb-10.2	Ish	O3	8320	57.6	37.3	13.2
Cdb-10.3	NM	L4	7956	47.7	15.3	14.5
Cdb-10.4	NM	O4	9236	60.3	30.5	21
Cdb-10.5	NM	P4	8.541	55.6	37.3	11.8
Cdb-10.6	NM	P3I	8840	63.6	44.1	11.8
Cdb-10.7	NM	O3I	8931	59	35.6	13.2
Cdb-20.1	Ish	L3	19800	42.4	35.6	9.2
Cdb-20.2	Ish	O3	19800	57	37.3	13.2
Cdb-20.3	NM	L4	19965	67.6	25.4	—
Cdb-20.4	NM	O4	19710	69.5	39	25
Cdb-20.5	NM	P4	16530	55.6	32.2	17.1
Cdb-20.6	NM	P3I	17595	68.2	33.9	23.7
Cdb-20.7	NM	O3I	19140	63.6	33.9	25
Cdb-40.1	Ish	L3	39600	54.3	40.7	11.8
Cdb-40.2	Ish	O3	39600	61.6	42.4	13.2
Cdb-40.3	NM	L4	39120	87.4	32.2	25
Cdb-40.4	NM	O4	38368	72.9	47.5	25
Cdb-40.5	NM	P4	33768	57	42.4	33.4
Cdb-40.6	NM	P3I	36162	73.5	45.8	26.3
Cdb-40.7	NM	O3I	35802	65.6	42.4	23.7
Cdb-64.1	Ish	L3	64000	61.6	40.5	13.2
Cdb-64.2	Ish	O3	64000	62.3	42.4	13.2
Cdb-64.3	NM	L4	65076	85.4	47.8	26.3
Cdb-64.4	NM	O4	63580	76.8	49.2	29
Cdb-64.5	NM	P4	54802	65.6	39	15.8
Cdb-64.6	NM	P3I	61532	80.1	45.8	29
Cdb-64.7	NM	O3I	63206	88.1	49.2	21.4

Table 2: Performance of different codebooks built with different sizes and model parameter samplings. 1) Vocal tract area function model: Ish (Ishizaka's Model), NM (New Model). 2) This column characterize the sampling method used; the notation is: the letter refer to the sampling method and the number refer to the number of parameters that used such sampling method. The methods were: L (linear), O (logarithmic), P (percentage) and I (irregular). The order of the parameters are A_c, x_c, A_m and A_t. For instance, $P3I$ means that A_c, x_c and A_m used percentage sampling method, while A_t used irregular sampling method.