GEOMETRICALLY AND ACOUSTICALLY OPTIMIZED CODEBOOK FOR UNIQUE MAPPING FROM FORMANTS TO VOCAL-TRACT SHAPE

Z.L.Yu and P.C.Ching

Department of Electronic Engineering , The Chinese University of Hong Kong

Shatin, N.T., Hong Kong

E-mail: zlyu@ee.cuhk.edu.hk, pcching@ee.cuhk.edu.hk

ABSTRACT

A method to generate a codebook with distributed formant targets and unique geometric-acoustic mapping from formants to vocal-tract shape by direct acoustic calculation is proposed. Geometric and acoustic constraints are applied to both vocal-tract model parameters and calculated acoustic features to eliminate unacceptable values from the initial codebook which usually has an extremely large codebook size. The vocal-tract length is used as an additional parameter to model the vocal-tract. Restriction on the vocal-tract length based on some measured data is employed. A geometric and acoustic optimization scheme is devised to cluster the constrained codebook into an uniquely mapped codebook with reduced size. The codebook generated by this method is precise and robust and provides a satisfactory solution to the inverse speech production problem.

1. INTRODUCTION

In spite of the many approaches in determining the vocal-tract shape from formant frequencies [1], the non-uniqueness of the acoustic-geometrical mapping remains a problem to be solved. Perturbation based methods have been proposed in [2][3] where the VT shape is modeled by Band Limited Fourier Expansion with odd and even cosine terms (OEBLFE). The acoustic feature associated with the symmetrical component of the VT shape is described by the resonant frequencies, $F_z(k)$, of the lip closed vocal tube. The target acoustics are the formant frequencies of the vowels or their equivalence, the resonant frequencies of the vocal tube $F_p(k)$. End-point initialization for both acoustical and geometrical parameters are first applied. Subsequently, we interpolate the vocal-tract length (*VTL*) and zero frequencies which is followed by a dynamic perturbation procedure to derive the time varying VT shapes. The end-point initialization is based on the root-cell codebook [3]. This method performs well on condition that the end-point formant targets of the vowel-to-vowel (*V-V*) transition are close to the isolated reference vowels. However, this is not always true in practice. If the root-cell codebook is replaced by an enhanced codebook with distributed formant targets in the resonant subspaces, robust end-point initialization can be obtained. Furthermore, automatic inversion for isolated vowel can also be realized by using such codebook.

One way to generate a codebook with distributed formant targets is to establish a geometric-acoustic mapping by direct acoustic calculation of all possible OEBLFE parameters. However, two primary problems need to be considered. The first one is the codebook size. If we use the first six OEBLFE coefficients and *VTL* as the geometrical parameters and apply suitable quantization for each parameter, we might end up with a codebook size as large as $4.48E+8$, which is far too big for practical end-point initialization. The second obstacle is the non-uniqueness problem. Different geometrical vectors might lead to acoustic parameters that have identically or fairly closely locations in the formant space. How to resolve this problem is not trivial.

In this paper, we develop a method to generate a codebook in which the non-uniqueness and the size of the codebook can be efficiently dealt with.

2. CODEBOOK GENERATION SCHEME

Each vector of the codebook contains essentially 7 geometrical components, namely L (*VTL*), $P_o(k)$ (the coefficients of odd BLFE) and $P_e(k)$ (the coefficients of even BLFE) of the first three formants. The derived VT area function is given by

$$A(i) = A_0 - A_0\{\sum_{k=1}^{3} p_o(k)\cos[(2k-1)\pi\frac{i\cdot x_0}{L}] + \sum_{k=1}^{3} P_e(k)\cos[2k\pi\frac{i\cdot x_0}{L}]\} \quad (1)$$

where i is the index of concatenate short tubes from glottal to lips, and x_0 is the unique length of each short tube [2]. From (1), six acoustic components, $F_p(k)$ and $F_z(k), 1 \le k \le 3$, can be computed from $A(i)$ [4][5]. Since some geometrical vectors are inappropriate

because of physical limitations, we can discard them before computing the acoustic vectors. In addition, constraints on acoustic vectors can also be exploited in a similar manner. The final stage of the codebook generation process is to eliminate the redundancies by applying an optimization procedure to the remaining vectors that surpass both geometrical and acoustical constraints. Fig.1 gives the flowchart of the entire procedure of the codebook generation.

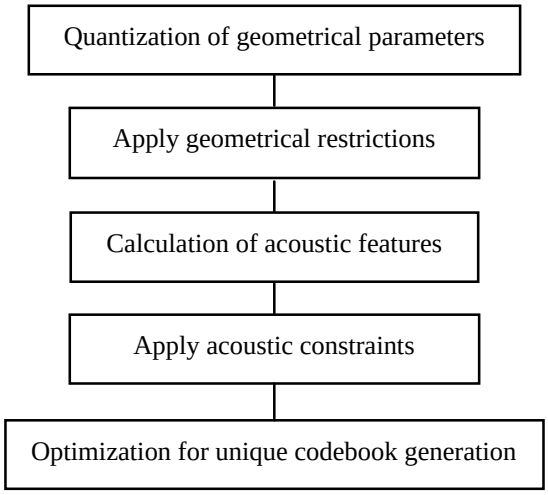

Fig.1 Diagram of the codebook generation

3. GEOMETRIC AND ACOUSTIC CONSTRAINTS

The physical limitations incorporated in the generation of geometrical vectors to avoid inappropriate VT shapes include:

(1) Positive area: the area function derived from the geometrical parameters as generated by equation (1) should be positive. To avoid computational overflow, we apply the following constraint

$$A(i) \geq A_{min} = 0.01cm^2 \tag{2}$$

(2) Maximum area: the area function derived from the geometrical parameters is not allowed to exceed a maximum value, viz.

$$A(i) \leq A_{max} = 15.0cm^2 \tag{3}$$

(3) Glottal constraint: in glottal part, the area function is made less than a pre-defined constant, namely

$$\begin{cases} A(i) \leq A_{g\,max} = 15.0cm^2, & i \leq Ig \\ \sum_{i \leq Ig} A(i) \leq S_{g\,max} = Ig \cdot 4.0cm^2 \end{cases} \tag{4}$$

where Ig is the section number of the glottal part and it is typically determined by the integer value of $2.0/x_0$.

(4) Maximum volume of the total VT: considering the anatomy restriction, the volume of the total VT area function is limited by a maximum threshold

$$\sum_{i=1}^{I} A(i) \leq S_{t\,max} = 17.5 \times 6.0cm^2 \tag{5}$$

The variables of the above constraints are chosen heuristically from the simulation tests, and they have shown to produce fairly good results in our simulations.

The acoustic vectors that are calculated from the geometric vectors that surpass the above geometric constraints are then clustered by using formant pattern knowledge. In each of the $F_1 - F_2$, $F_1 - F_3$ *and* $F_2 - F_3$ subspace, vowels of normal speech usually reside in a confined boundary. Fig.2 shows a typical example of the boundaries of six Russian vowels. As a result, a condition being imposed in the formant domain is to discard a particular acoustic vector if the responding $F_p(k)$ is located outside any one of these formant boundaries.

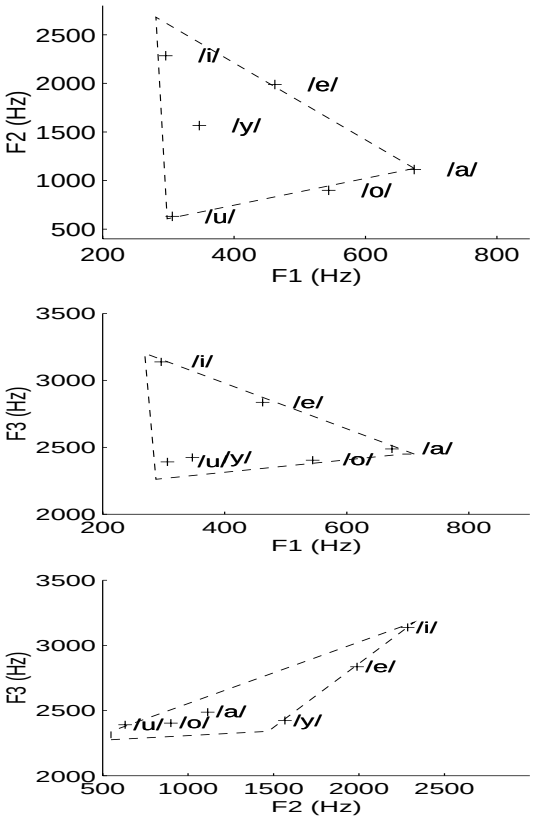

Fig.2 Boundary restriction in formant subspaces

In addition, we use a well defined distributed *VTL* in $F_1 - F_2$ subspace as a combined geometric and acoustic criterion to ensure the uniqueness for a given set of

formants and *VTL*. We restrict the *VTL* on a curve surface of F_1 and F_2, by the equation

$$Z(x,y) = z_0 + aF_1 + bF_2 + cF_1F_2 + dF_1^2 + eF_2^2 \tag{6}$$

With the available reference data, the coefficients, z_0, a, b, c, d and e can be determined by solving a set of linear equations. Fig.3 depicts the constrained *VTL* against the $F_1 - F_2$ plan. If the variables L, $F_p(1)$ and $F_p(2)$ of an acoustic vector do not satisfy (6) within a tolerance, it will be removed. The value of the tolerance σ depends on the quantization of the vocal tract length, and it is made to satisfy the following condition

$$\left| Z(F_p(1), F_p(2)) - L \right| \le \sigma = 0.5 \cdot \Delta L_0 \tag{7}$$

where ΔL_0 is the quantization level of the *VTL*.

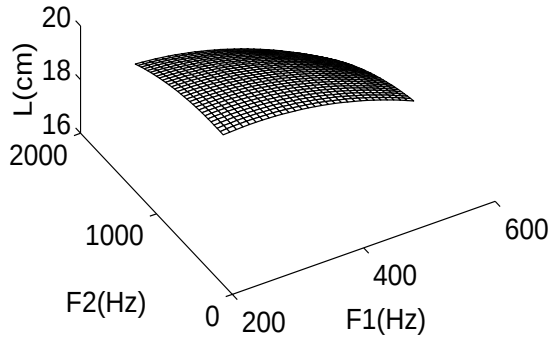

Fig.3 Vocal-tract length distribution versus F_1-F_2 plan

4. OPTIMIZATION FOR UNIQUE MAPPING VECTORS

To eliminate the non-uniqueness of the codebook completely, an optimization procedure is necessary after applying the above constraints. The formant space is quantized by $M_1 \cdot M_2 \cdot M_3$ cubic cells. For each cell, an optimal vector is chosen from the candidates residing within its area. The geometric cost function for optimization is

$$\begin{aligned} C_G(j) = & \sum_{n=1}^{M_r} \sum_{i=1}^{I} w_n^r(i) \cdot [A_j(i) - A_n^r(i)]^2 \\ & + \sum_{n=1}^{M_r} \sum_{i=1}^{I} w_n^r(i) \cdot \{[A_j(i) - A_j(i-1)] \\ & - [A_n^r(i) - A_n^r(i-1)]\}^2 \end{aligned} \tag{8}$$

where $A_j(i)$ represents the VT area, $A_n^r(i)$ signifies the referenced area of the measured isolated vowel n, and $w_n^r(i)$ is a weighting function. The weighting function is designed to emphasize the importance of those special parts of the VT, which include the glottal, lips and the restriction parts, and the peak point of the front/back cavity, if any, of the reference VT shape. The weighting profiles are shown in Fig.4.

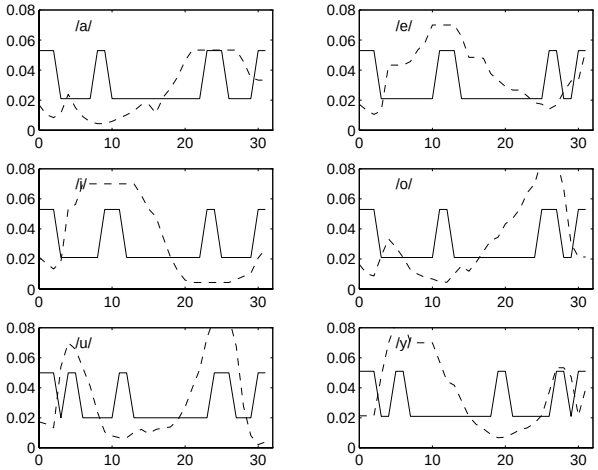

Fig.4 Weighting function profiles for the six reference VT shapes (—— weighting function, - - - - area shapes)

The acoustic optimization cost function is given by,

$$C_A(j) = \sum_{n=1}^{N_r} \sum_{k=1}^{3} \left| F_{z,j}(k) - F_{z,n}^r(k) \right|^2 \tag{9}$$

where $F_{z,j}(k)$ is the zero frequency of the $j'th$ candidate and $F_{z,n}^r(k)$ is that of the $n'th$ reference vowel.

The overall cost function is computed from

$$C_T(j) = \alpha_1 \cdot C_G(j) + \alpha_2 \cdot C_A(j) \tag{10}$$

where α_1 and α_2 are weighting factors. The final codebook size depends on the quantization level of the formant space. We have conducted various tests and the results revealed that a codebook consisting of *7819* vectors could be obtained which is quite adequate for most applications.

5. EXPERIMENTAL RESULT

We take the Russian vowels as reference data [4]. Two validations are carried out. One is for isolated vowels and another is for dynamic *V-V* transition. For isolated vowel, $F_z(k)$ and *VTL* are taken from the matched

vectors of the optimal codebook for the specific formant targets $F_p(k)$. Whereas for the dynamic case, the end-point $F_z(k)$ and *VTL* are obtained from the matched vectors of the optimal codebook for the desired end-point formant targets. The dynamic $F_z(k)$ and *VTL* trajectories are interpolated. The interpolated $F_z(k)$ and *VTL* are then merged with the original $F_p(k)$ to form a joint target. For this joint target, the inverse solution is derived from the procedure described in [3]. Fig.5 and Fig.6 show the experimental results of two typical case of the inversion of isolated vowels and dynamic vowel-to-vowel transitions, respectively. The results illustrate good end-point matching and smooth dynamic trajectory.

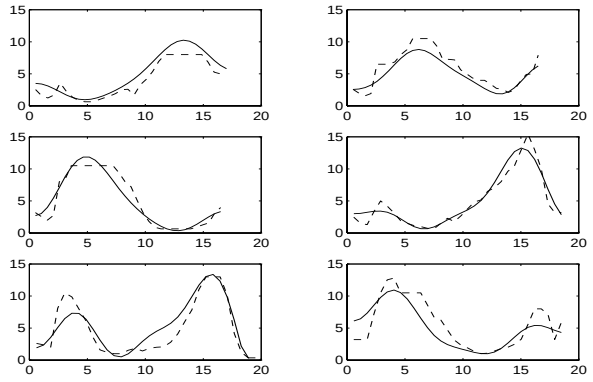

Fig.5 Codebook matched VT for formants of isolated vowels

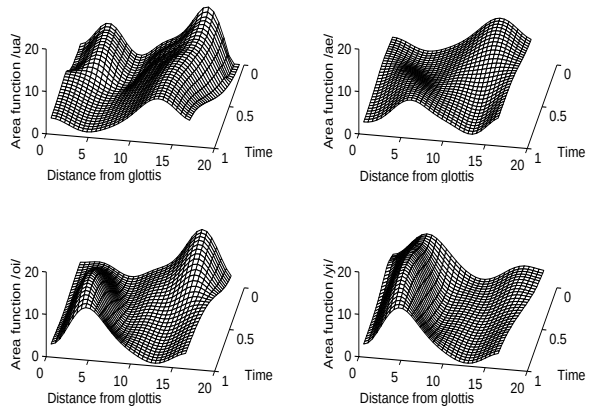

Fig.6 Dynamic inverse solution of V-V transitions based on the codebook matching initialization

6. DISCUSSION AND CONCLUSION

The proposed speech inversion technique has several advantages. First, the optimal codebook provides an unique mapping relation from formants to VT shapes. Second, the codebook generation initially covers a large range of VT shapes. Third, the geometric and acoustic restriction eliminate invalid vectors which in turn saves computation and storage. Simulations show that the optimal codebook not only provides good matching for the end-points of *V-V* transition, but is also useful to solve the static case of the isolated vowel's inversion without a prior specification of $F_z(k)$ and *VTL*. Considering a static case, if there is no such a codebook, the $F_p(k)$ and *VTL* have to be specified artificially before the perturbation procedure can be applied, as it is done in the work of [2] where $F_p(k)$ are set to zero and *VTL* is specified to coincide with the reference data. Even if the root-cell codebook is used for this purpose, though the artificial specification can be avoided, the applicability of the assemble targets can be argued because the root-cell book has only distinctive values of $F_p(k)$, $F_z(k)$ and L while the targets of end-point $F_p(k)$ may varies very differently. Therefore, the robustness of the proposed approach is to a certain extent guaranteed.

The current work is a part of the entire profile of the inverse solution of speech production based on the perturbation theory. Although the determined VT shape is on-line validated by the method itself, the evaluation of the inverse result by synthesis will also be necessary. Our future work will be focused on the reproduction of the original sound with the inversed VT area which is being used to drive the K-L synthesis model [6][7]. This work is currently undergoing.

7. REFERENCES

[1]. J. Schroeter & M. M. Sondhi (1994): "Techniques for estimating vocal-tract shapes from the speech signal", IEEE Trans. Speech & Audio Proc., vol.2, no.1, pp.133-150 .

[2]. Z. L. Yu (1993): "A method to determine the area function of speech based on perturbation theory," STL - QPSR, 4/1993, pp.77-95.

[3]. Z. L. Yu & P. C. Ching (1996): "Determination of vocal-tract shapes from formant frequencies based on perturbation theory and interpolation method," ICASSP96, Vol.1, pp.786-789, Atlanta, USA.

[4]. G. Fant (1960): "Acoustic theory of speech production," the Huge: Mouton (2nd edition, 1970).

[5]. P. Badin & G. Fant (1984): "Notes on vocal tract computation," STL - QPSR, 2-3/1984, pp.53-107.

[6]. J. L. Kelly & C. C. Lockbaum (1962): "Speech synthesis," Proc. 4th Int. Congress on Acoustics, Copenhagen, PaperG-42,pp.1-4

[7]. J. Liljencrants (1985): "Speech synthesis with a reflection-type line analog," Doctoral Thesis, KTH, Stockholm.