PARAMETERIZED VT AREA FUNCTION INVERSION

Mats Båvegård and Gunnar Fant

Dept. of Speech, Music and Hearing, KTH,
Box 70014, 10044 Stockholm, Sweden
email: Mats.Bavegard@speech.kth.se

ABSTRACT

The purpose of our study is to contribute tools for inversion of articulatory to acoustics relations, in specific to perform an estimate of vocal tract area-function parameters from formant frequencies.

The inversion is performed in two steps. A first approximation is attained from either a codebook or a neural net and a final optimization is performed by an iterative interpolation for finding a perfect or acceptable match.

The study is based on a three-parameter vocal tract model. The codebook relates each of the possible combinations of constriction location, X_c, constriction area, A_c, and the lip parameter, l_0/A_0 to a corresponding F_1, F_2, F_3 pattern. The neural network output provides the same choice of possible VT states as the codebook. The input to the neural network is normally programmed in terms of formant frequencies but other acoustic attributes can be selected or added.

Present experience is limited to vocalic area functions. Our present system provides a rapid conversion of formant frequency data to VT parameters and has provided promising results for short sentences.

1. INTRODUCTION

The basic principle is to decompose any vocal tract area function into two parts, an overall vocalic part with parameters settings appropriate for the particular coarticulation and a parametrically specified consonantal part which substitutes or modifies parts of the vocalic area function.

1.1 The revised vocal tract model

The modeling of VT area functions employs the three traditional independent parameters, X_c and A_c for constriction location and area and a lip parameter l_0/A_0 [1][2]. The detailed area function was constructed by a concatenation of 6 successive segments and a piriformis cavity shunting the outlet of the larynx tube. The constants describing the shapes and sizes of these segment were given default values uniquely determined by the three main control parameters. Some of these constants can be released and given the status of shape parameters for adjustment of the model to a specific speaker. Examples of normal covariation of control parameters and shape constants are the positive correlation of inter-incisor distance with A_c in front vowels and with X_c in back vowel which reflects increasing jaw opening. In back vowels the degree of asymmetry of the pharyngeal constriction varies in accordance with the location of X_c. As expected, there is also a positive covariation between the overall vocal tract length and the degree of lip-rounding l_0/A_0.

In order to conform with natural constraints the model was divided into three parts with respect to the location of the X_c coordinate which has been defined as the distance from the incisors. Thus we defined a range of $X_c<4$ cm were all the front vowels were found at $4>X_c>2.5$. The second region, $4<X_c<7$, was referred to as mid-vowels, were we located [u] at about $X_c=6.5$. The third region of $7<X_c<14$ housed the back vowels in the order of [o], [ɑ] [ae]. These preferred locations are by now quite well documented in several studies [3][4].

The consonantal part of a VT area function is specified by four parameters. Two of these, analogous to X_c and A_c of the vocalic model, have the primary function of specifying the location and degree of consonantal constrictions. The two additional parameters, pertaining to the effective length and shape of the consonantal constriction are dependent on the primary consonantal and vocalic parameters [5][6].

Temporal organization will involve the control of covarying and in part dependent vocalic and consonantal area function parameters subject to articulatory constraints and representative time constants.

2. INVERSION

The inversion from formant frequencies to area function parameters are performed in two stages. The first approximation is performed by a codebook lookup or a neural network and the second is an optimization procedure.

In practice the VT-model may be unable to adapt to a speakers set of formant frequencies and no solutions are attained unless a certain amount of tolerance is allowed. Another obstacle is that two quite different settings of the model may generate almost identical formant patterns. This is less of a problem with a larger than with a smaller number of formants.

A major problem is the initialization of the optimization procedure. Since most algorithms will only find the local minimum of a given cost function close to the initial parameters, it is of great importance to choose suitable startup parameters. A requirement is thus a sufficiently detailed acoustic-to-articulatory mapping. This could be realized with a codebook. Since the articulatory and the acoustic domain needs to be densely sampled, codebooks can become very large and the search very time consuming [7].

As an alternative to large codebook searching, one can use the codebook material to train an artificial neural network (ANN) to perform the inversion. The advantage is that the output from the ANN is computed fast and that the time consuming training of the network is done in advance. There is also the possibility to design networks to handle temporal sequences of data-frames as one input vector [8].

3. METHOD

The basic outline of the revised model with its three prototypes for front, mid and back vowels, see Figure 1, is the same as in the earlier version [2], but the defining equations and constants have been elaborated in several ways [6].

One object has been to allow for a more realistic modeling of the vowel [u], i.e. to secure a sufficiently low F_3. Another is to produce more realistic prototypes of [ae] modeled alternatively as a front or a back vowel.

We are also using a larger default value of the inter-incisor distance D_t at the teeth. With lip-rounding included the most notable effects of the increased D_t is an increase of about 100 Hz in F_3 of rounded front vowels such as [ʉ] and [ɸ]. A number of simulations and descriptions are found in [6].

3.1 The articulatory codebook

Our articulatory reference material consists of area functions generated with the three-parameter model based on Swedish vowels [6]. The transfer functions of the vocal-tract area functions are computed with FLEA [9].

The codebook provides an initial representation of the three VT-parameters from three given formant frequencies. It is possible to use a fairly small codebook as long as the initial representation is good enough. Our codebook was generated from a set of different VT-parameter combinations, slightly larger than in [12].

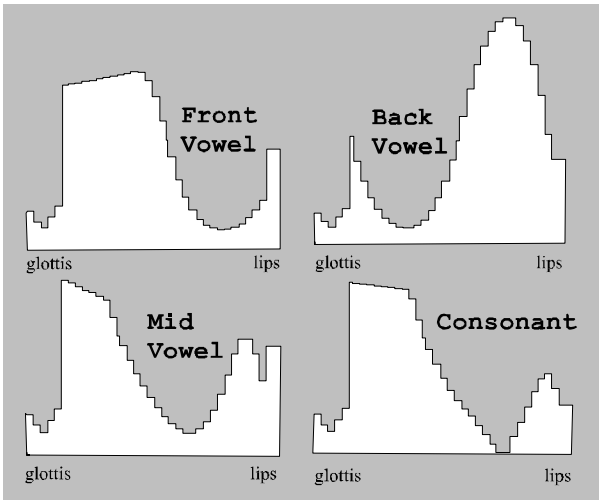

Figure 1. The vocal tract area function model, exemplified by the three different regions defined for the vocalic part and one palatal consonant /g/.

3.2 The distance measure

The difference between a target set of formant frequencies F_{1t}, F_{2t}, F_{3t} and a set of model or codebook generated formant frequencies F_{1m}, F_{2m}, F_{3m} is expressed by a distance measure:

$$D = [D_1^2 + D_2^2 + D_3^2]^{0.5}$$

where:

$$D_n = \text{Bark}(F_{nt}) - \text{Bark}(F_{nm}) \qquad n = 1, 2, 3$$

The perceptual significance of the distance measure D can be argued, but observed values can be quantitatively related to minimal distances between adjacent phonemes in the Swedish vowel system which are of the order of D = 2 BARK.

3.3 The optimization procedure

The estimate of VT-parameters is further enhanced in a optimization procedure. The algorithm we have adopted is based on a perturbation analysis of the differential contribution of each VT-model parameter to each formant F_1, F_2 and F_3. This method was developed by [10], based on the theory of differential contribution [11].

The optimization procedure and the inferred improvements are described in detail in [12].

One important issue here is the accepted error threshold of each formant ΔF_n which may be selected from a combination of acoustic and perceptual criteria. We have chosen to define our convergence criteria so as to allow a maximum error of $\Delta F_1 = 10$ Hz and maximum error of $\Delta F_2 = 2\%$ and $\Delta F_3 = 2\%$.

3.4 The test material

The test material is chosen to evaluate the capacity of the three-parameter model both in terms of accepting a wide range of formant patterns and the possibility of finding reasonable VT-area configurations.

The test material is both the sentence "Ja-Adjö", [ˈjɑːˑaˈjɸː] spoken by one male speaker and the tabulated formant values of Swedish standard vowels [13].

4. RESULTS

The processing of the short sentence, "Ja-Adjö", the first stage involves a codebook matching to formant patterns within each of the 10 ms frames of the utterance.

The three control parameters show dynamic variations which are reasonable and display a fair degree of continuity, see Figure 4.

For the Swedish vowels we obtained similar results, see Figure 2. The average distance measure with codebook lookup only, D^2

= 0.173, was reduced to $D^2 = 0.051$ after the optimization, which gave improvements in 9 of 11 vowels [12].

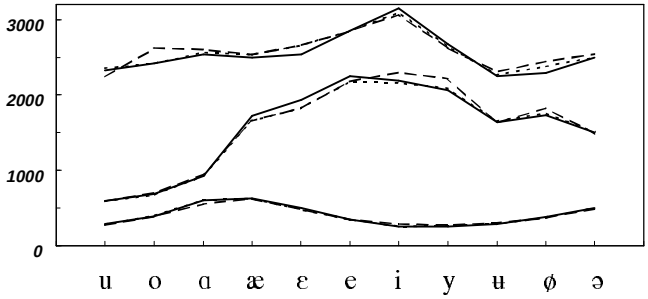

Figure 2. Swedish standard vowels [13], measured formant frequencies (solid), codebook table lookup (dashed) and optimized (dotted).

5. DISCUSSION

5.1 VT area function model

Our revised three-parameter model has gained in overall performance and flexibility to adapt to a wide range of vowel articulations, in the first place for male speakers. This has been achieved by a greater attention to system constants some of which may be given the role of *shape parameters*. Shaping the VT-model to an individual speaker. A model of this type is not free from inherent constraints. We encounter some difficulties in realizing a complete range of human variations and conversely the model may generate patterns beyond human capability.

The latter is less of a problem and can be avoided with proper constraints. An insufficient coverage can also be overcome by a proper choice of system constants and shape parameters. We do not claim that this is a final version.

The extension to include consonantal articulations is based on the general philosophy that any consonant can be regarded as superimposed on a vowel configuration which is set by the particular coarticulatory pattern [14]. Conversely, the consonantal part of the articulation must be adapted to the specific vowel configuration.

A second phase will be to adapt the model to match dynamic patterns of human speech. In practice, the coarticulatory pattern will be influenced not only by the specific sequence of phonemic entities but also by the superimposed prosodic pattern, e.g. by the vowel-consonant contrast implied by the relative emphasis. We have not yet been in a position to model such variations but we believe that they could be handled with the present inventory of vocal tract area-function parameters.

5.2 Inversion

The inversion is performed in two steps, a primary selection by codebook lookup or a neural net processing, followed by an optimization through complex interpolation in the codebook domain.

The codebook contains sets of the three basic control parameters X_c, A_c, l_0/A_0 and corresponding formant patterns F_1, F_2 and F_3 constrained to a domain of articulatory-acoustic relevance with bounds set by the rich Swedish vowel system allowing for extreme values.

It is well known that the non-uniqueness of the inverse transform can be partially counteracted by such constraints [4]. In practice we need to accept approximate solutions, the accuracy of which we have quantified in terms of a vectorial distance measure between the target formant pattern and the closest model generated pattern expressed as a vector in the BARK-domain and its components in the F_1, F_2 and F_3 dimensions.

The typical outcome of our inversion experiments is a vectorial distance measure D of the order 0.3 BARK, to be compared with the average minimal distance between adjacent Swedish vowel prototypes 2.0 BARK.

The inversion performs well for formant patterns generated by our standard male model. The major problem for inversion of human speech is the difference between a model and the human vocal tract. Representative solutions may not be found unless the shape parameters of the model can be adjusted closely to the speaker. Future work should accordingly be directed to studies of how the shape parameters can be adjusted for a specific subject.

The codebook look up guarantees a fair match of formant patterns which is further improved by the optimization, but what about the underlying VT parameters and the demand for articulatory realism and continuity? Our results from the voiced sentence are on the whole promising and provide some indirect insight in the gestural sequence, including a representative final schwa /ə/. However the optimization procedure is solely based on distance measure minimization. There are no continuity or parameter smoothing constraints built in, which gives us some discontinuities. This is shown in the transition from /j/ to /a/ and from /ɑ/ to /j/ i.e passing through a neutral shape, see figure 4. If we analyze this further, we have found that there is a fairly large distance measure D^2 at these points, mainly dependent of F_3. The solutions found implies that the VT model needs further adaptation to the specific speaker. In our future work we will further exploit the continuity constraints and VT model adaptation.

The neural network is an interesting alternative to the codebook look up. It is well known that three- or four- layer neural networks have the ability to approximate any arbitrary nonlinear function. among others, [15] have revealed the possibilities of estimating articulatory parameters by neural networks. The performance is often less accurate, thus the method is robust and fast. In our tests the neural network inversion gave larger distance measure, D^2 than the codebook lookup. There are two possible explanations. There is no minimization involved in the neural network inversion, which means that in the ideal case of perfect mapping with the neural network in this application, produces results similar to the codebook table lookup. Also there are no constraints to the output of the neural network, which

means that it might generate VT-parameter combination that where sorted out, as forbidden, in the codebook and training material. However the strength of a neural network application is its ability to make generalizations within the training data material while our intention is to select one VT-parameter combination out of a fixed number of possible combinations.

6. ACKNOWLEDGEMENTS

This work has been funded by ESPRIT/BR (6975), SPEECHMAPS, in part financed by NUTEK and by grants from TFR, the Swedish Research Council for Engineering Sciences.

7. REFERENCES

1. Fant, G (1960). Acoustic theory of speech production, Mouton & Co., The Hague, Second edition, 1970, Walter de Gruyter, Berlin.
2. Fant, G (1992). Vocal tract area functions of Swedish vowels and a new three-parameter model, Proceedings ICLSP 92, 1: 807-810.
3. Wood S (1979). A radiographic analysis of constriction locations for vowels, J of Phonetics, 7: 25-43.
4. Boë L-J, Perrier P & Bailly G (1992). The geometric vocal tract variables con-trolled for vowel production: proposals for constraining acoustic- to- articulatory inversion. J of Phonetics, 20: 27-38.
5. Båvegård, M. (1995), Introducing a consonantal model to the articulatory speech synthesizer, Proceedings Eurospeech‘95, Madrid. 1857-1860.
6. Fant G & Båvegård M (1995). Parametric model of VT area functions: Vowels and Consonants, ESPRIT/BR SPEECHMAPS (6975), Delivery 28, WP2.2, 1-30.
7. Schroeter J, Meyer P & Parthasarathy S (1990). Evaluations of improved articulatory codebooks and codebook access distance measures. IEEE ICASSP, 1: 393-396.
8. Papcun G, Hochberg J, Thomas TR, Laroche F, Zacks J & Levy S (1992), Inferring articulation and recognizing gestures from acoustics with a neural network trained on x-ray microbeam data, J Acoust Soc Am 92, 2: 688-700.
9. Lin, Q (1990). Speech production theory and articulatory speech synthesis, Ph.D thesis, Dept of Speech Communication and Music Acoustics, KTH, Stockholm
10. Lin Q & Fant G (1989). Vocal-tract area-function parameters from formant frequencies. Proceedings Eurospeech '89, Paris, 673-676.
11. Atal BS, Chang JJ, Mathews MV & Tukey JW (1978). Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique. J Acoust Soc Am, 63: 1535-1555.
12. Båvegård M & Fant G (1995). From formant frequencies to VT area function parameters, STL-QPSR 4/1995, 55-66
13. Fant G, Henningsson G. & Stålhammar U (1969). Formant frequencies of Swedish vowels. STL-QPSR 4/1969, 26-31.
14. Öhman, S (1967). Studies of articulatory coordination, STL-QPSR 1/1967, 15-20.
15. Shirai K & Kobayashi T (1991). Estimation of articulatory motion using neural networks, J of Phonetics, 19: 379-385.

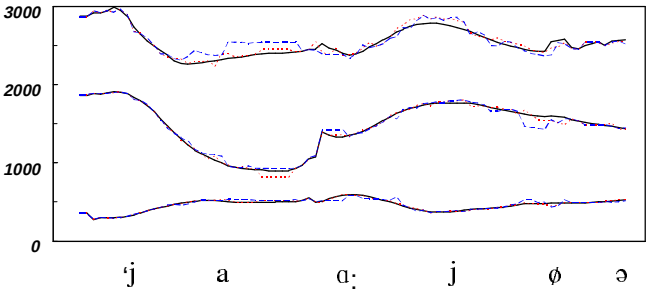

Figure 3. Inversion of the short sentence "Ja-Adjö". Measured formant frequencies (solid line), codebook table lookup (dashed) and after optimization procedure (dotted).

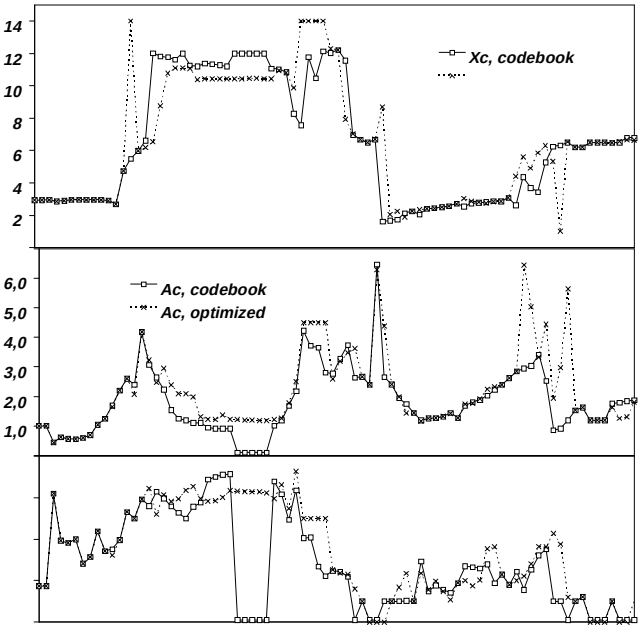

Figure 4, The resulting VT parameters, codebook table lookup (solid) and after optimization (dashed). Top: The location of the constriction X_c, the area of the constriction, A_c in the middle, and bottom, the resulting lip parameter, l_0/A_0.