ON IMPROVING THE DECISION ALGORITHM FOR ARTICULATORY CODEBOOK SEARCH

C. Silva[1], S. Chennoukh[2] and I. Trancoso[3]

[1]Departamento de Electrónica Industrial, Universidade do Minho,
4709 Braga Codex, Portugal – cas@eng.uminho.pt
[2]Philips Research Labs., Netherlands – chenoukh@natlab.research.philips.com
[3]INESC/IST – Isabel.Trancoso@inesc.pt

ABSTRACT

This paper describes our progress on articulatory voice mimic. The objective is to achieve an articulatory voice mimic system as a basis for low bit-rate speech coding using articulatory codebooks. The articulatory codebook uses a suitable vocal tract model for generating shapes for all possible speech sounds. When building a codebook, unrealistic vocal tract shapes may be generated. In this paper, we describe the used filter based on physiological assumptions to avoid populating the codebook with such shapes. We also propose a new method for the codebook search. This method weights the prediction error for each area function parameter as a function of its contribution to match the input acoustic signal parameters. An interaction between the variations of the area function parameter is created to constraint the articulatory trajectory, which improves the search and increases the output speech quality of the voice mimic system.

1. INTRODUCTION

The problem of estimating the time evolution of the vocal tract shapes from the speech signal, more known as the inverse problem, has attracted the attention of several researchers from distinct areas for decades. The solution would coalesce the speech production, coding and recognition research into a reduced parametric framework.

The acoustic-to-articulatory mapping is not unique. To solve this problem, a two-step approach is often used. First, a table look-up is performed to provide initial estimates of the vocal tract shape. Second, an optimization is applied to improve the accuracy of the articulatory parameters [1]. The look-up table contains vocal tract shapes with their corresponding acoustic vectors. If the look-up table has a fine sampling of the articulatory space, then the second step may be skipped without hopefully major loss in articulatory analysis accuracy. When analyzing continuous speech some mechanism must be used to deal with the non-uniqueness of the acoustic-to-articulatory mapping and ensure a smooth path for the articulatory trajectories through the table look-up.

In our previous work [2], we extended Ishizaka's vocal tract area function model to allow constrictions at the tongue tip [1]. Using this model, several methods have been studied to sample the articulatory parameters. The goal was to obtain maximum coverage of the acoustic space, while minimising the size of the codebook. The design of this codebook turned out to be a very challenging problem. Thus, a more elaborated search mechanism has being studied to minimise the distortions in the articulatory trajectory estimation.

The following section lies down the construction strategy of an articulatory codebook, which prevents the physiologically unrealistic shapes from being inserted in the codebook. The third section describes the decision algorithm of the best candidate. Analysis results are shown in the fourth section. Finally, the conclusions and the perspectives are discussed in the last section.

2. CODEBOOK CONSTRUCTION

In the construction of the codebook, we perform an off-line inversion from the acoustic domain to the articulatory domain [3, 4]. This approach clusters the articulatory vectors in an acoustic network that sub-samples the acoustic space in constant acoustics steps. This network is accessed in a straight manner using the acoustic vector as a search key.

In the searching of the acoustic network, when articulatory shapes in the target cluster are not found, a more thorough search is performed in its vicinity in order to obtain a solution. This search simultaneously considers all shapes at a distance of one acoustic step from the target cluster. If no shape satisfying the criteria for acoustic and articulatory continuity is found, then the search is repeated after increasing the acoustic step. In our previous work, the search was performed only in certain directions, finishing when shapes were found in a cluster. We verified that sometimes it missed the acoustic trajectory, because although it tried to keep the continuity of the acoustic path, it allowed a large mismatch in the articulatory trajectory, causing a wrong prediction of the next state of the forward dynamic network.

The modified version of the vocal tract area function model we have used allows the description of the area function of vocalic and consonant sounds [2]. This is accomplished by the introduction of the tongue tip, which was described by its position and the corresponding area. By working with the area function

instead of a 2D articulatory model we decrease the complexity of the analysis, but we lose the capacity to perform a physiological validation of the shapes inserted in the codebook in a straightforward manner.

In order to validate the insertion of shapes in the codebook we took into account the boundary positions of the articulators [2]. For instance, the production of allophones of /a/ in the lower pharynx is accomplished by a contraction of the hyoglossus muscle and by the pharyngeal constrictors [5]. This results in a backward movement of the tongue body and a downward movement of the tongue blade. The movement of the tongue blade brings the tongue tip to the front, near the teeth, allowing it to raise and perform a front constriction with the help of the mandible. The same effect over the tongue tip can be seen in a front vowel such as /i/. However, in the production of a constriction near the velum, as in the vowel /u/, the raise of the hump of the tongue is accomplished by contraction of the posterior genioglossus and contraction of the styloglossus [5]. The latter has the effect of bringing the tongue tip to the back and limiting its movement in the upward direction. Taking this physiological process into account, we have performed the pruning of the shapes inserted in the codebook by defining a forbidden zone for the tongue tip:

$$A_{T_{forbidden}} = A_{T_{\min}} - A_{T_{\min}} \frac{(x_T - x_{TS})}{(x_{TF} - x_{TS})}$$

In the above equation, X_T is the tongue tip position, and X_{TS} and X_{TF} are the minimum and the maximum position of the tongue tip, respectively. When the tongue tip position is at its maximum, X_{TF}, the minimum forbidden area is zero, which means that the tongue tip has a complete freedom of movement. The minimum tongue tip area, AT_{min}, is a function of the tongue body constriction area:

$$A_{T_{\min}}(A_C) = A_{C_{\max}} - A_C$$

where Ac_{max} is the maximum area of the constriction of the tongue body, A_C. Therefore, when A_C is minimum, the minimum forbidden area for the tongue tip is maximum. This minimum area is weighted by the position of the constriction of the tongue body, which is implicitly contained in the tongue tip position [2].

3. DECISION ALGORITHM

A difficult problem in articulatory speech analysis is the possibility of having different shapes of the vocal tract for the same acoustic vector. The usual approach to tackle the non-uniqueness consists of imposing continuity on the trajectories of the states of the speech production mechanism over a segment of speech. Due to the computational complexity of this solution, we follow a different strategy to deal with the non-uniqueness problem. We use a predictive technique to estimate the future state of the system based on the previous position, velocity and acceleration of the parameters of the area function model. This estimate is then used to solve the non-uniqueness problem by taking into consideration the dynamics of the system and ensuring the smoothness of its parameters according to a cost function. We have adopted as a cost function the sum of the square Euclidean distance between the input shape parameters and their predicted position [4].

When testing our approach, we verified that the smoothness of the acoustic trajectory varied according to the acoustic coverage of the articulatory codebook, since the inversion was poor when the percentage of empty target nodes increased. By analysing the data of the decision path of the algorithm, we verified that the algorithm sometimes did not make the best decision, when considering such factors as the type of parameter and its range. So, we concluded that the algorithm was not robust enough to choose the best representative shape in the case of a miss in the search of the target cluster.

To tackle this limitation we tried to introduce more physiological insight into the algorithm, taking into account the properties of each parameter of the model. We did it by noting Stevens' discussion on the quantal nature of speech [6], which draws one's attention to the saturation effects during speech production. These saturation effects occur when an articulatory parameter is moved by the muscles along its dynamic range, causing a non-linear pattern on some acoustic parameter. For example, in the production of a labial or lingual closure, the muscles contract decreasing the constriction area until it saturates at zero, which is followed by a saturation of the total acoustic energy at a low level. This king of saturation effects is found in consonant sounds. In the case of the vowels /i/, /a/ and /u/, the saturation effects are attained in a different manner, where the acoustic cavity interaction causes a relative insensitivity of its formants on small variations of the constriction location [7, 8]. Indeed, several studies on the repeatability of the production of these vowels and certain consonants show that muscles interact to maintain steadiness in the area of the constriction, while allowing some variation on its position [9, 10]. However, the same mechanism is not found for vowels with a higher constriction area [10].

This insight prompted a new two-stage decision mechanism. In the first stage, a pruning of all possible shapes in the analysed cluster is performed, where every shape whose parameter is above a certain threshold is ignored. This parameter-specific threshold varies according to the sub-range of the parameter. The effect of the first stage is to leave only the shapes that are close enough to the predicted one to be considered as a plausible solution. In the second stage, we apply a weighted square Euclidean distance as a cost function to the remaining shapes:

$$Error = \sum w_i (p_i - s_i)^2$$

where w_i is the weight, p_i the predicted position and a_i the value of the parameter found in the codebook. The weights vary according to the kind of parameter and its value range. This variation tries to mimic the interplay of the muscles and articulators in a simple fashion.

In our current work, only the weights for A_T, A_C and X_C are variable, while the weights of the other parameters are considered to be unitary. The weights for A_C and A_T have maximum value for the minimum value of the respective parameter, linearly decreasing down to one for the maximum value. The weight for X_C has the inverse behaviour of the weight of A_C, assuming its minimum value when A_C is minimum.

This two-stage approach is important in our strategy. It ensures that we can detect false solutions during the decision. Algorithmically, we have:

WHILE [best shape not found] **DO**:
  FOR [shapes in the cluster] **DO**:
    IF [shape is bounded by the threshold] **THEN**
      [compute cost function]
  IF [shapes not found] **THEN**
    [enlarge thresholds]

This procedure is applied to every shape in the target cluster and, when the target cluster is empty, to all the shapes in the neighbourhood; only then can we consider to have found the best shape.

4. RESULTS

We used one of our codebooks to assess the performance of the decision algorithms. The parameters' ranges are shown in Table I. The area of the tongue body, the tongue tip, the area at the lips and the place of the tongue body constriction were sampled logarithmically. Two sentences were used in the assessment, "Who are you?" and "Where are you?". These were spoken by different male speakers. The signals were analysed using the Waves+ software package (Entropics). Formant analysis was done using the autocorrelation method, with a Hamming window of 25 ms, updated every 10 ms. The obtained formants were used to access and search the codebook. Then, the codebook provided the matched cluster of model parameter vectors to the decision algorithm.

We verified that with the new decision algorithm the system performs a better tracking of the formants, even when the target cluster was empty. This is illustrated in fig. 1, where we show the spectrograms of the original and synthetic utterances of the sentence “Who are you?”, processed by the two decision algorithms. We also verified that the new decision algorithm provided smoother trajectory for the parameters of vocal tract area function model.

In general, we have verified that when the computed formant values area accurate enough, the decision algorithm performed quite well. However, when the formant analysis failed, the algorithm could not perform a correct tracking.

Parameter	Min	Max	Nr. Pts	sampling
A_C	0.001	4.0	18	Log.
X_C	4.0	14.0	23	Linear
A_M	0.1	8.0	15	Log.
A_T	0.001	2.0	5	Log.
A_B	2.0	8.0	3	Linear
A_F	4.0	10.0	3	Linear

Table 1: Value ranges for the parameters used in the construction of the codebook. The codebook includes 235.845 shapes.

5. CONCLUSION

In this paper, results on the pruning of physiologically unrealistic shapes during codebook construction are presented. The pruning was implemented by taking into account physiological constraints, which led to a reduced size of the articulatory codebook. The relevance of the multidimensional constraints used to prune the codebook has not been checked to make sure that only unrealistic shapes have been rejected. So, further studies are needed for a better assessment of the codebook construction technique. The approach is promising as we improve the vocal tract area function model that we validate according to physiological considerations. A new decision algorithm is also been developed. The robustness of our decision algorithm has been increased by including articulatory constraints. By taking into account the contribution of the constriction to the acoustic transfer functions, we succeeded to cope with the problem of empty nodes. As a future direction, we are considering new techniques to handle the errors in the formant analysis and its effects on the codebook search. The introduction of information about compensation effects between the articulators as additional articulatory constraints is also in investigation for a better decision algorithm.

REFERENCES

1. J. Flanagan, K. Ishizaka, and K. Shipley, “Signal models for low bit-rate coding speech”, J. Acoust. Soc. Am., 68, pp. 780-791, 1980.
2. C. Silva, and S. Chennoukh, “Estimation of articulatory parameter trajectory from speech acoustic dynamics”, The Third ESCA/COSCODA Workshop on Speech Synthesis, Blue Mountains, Australia, 1998.
3. S. Chennoukh, D. Sinder, and J. Flanagan, “Voice Mimic System”, CAIP Technical Report TR-228, New Jersey, 1998.

4. C. Silva, and S. Chennoukh, “Articulatory analysis using a codebook for articulatory based low bit-rate speech coding”, ICSLP’98, Sydney, Australia, 1998.

5. J. S. Perkell, “Properties of the tongue help to define vowel categories: hypotheses based on physiologically-oriented modeling”, J. of Phonetics, 24, pp. 3-22, 1996.

6. K. N. Stevens, “On the quantal nature of speech”, J. of Phonetics, 17, pp. 3-45, 1989.

7. J. S. Perkell, M. Matthies, H. Lane, F. Guenther, R. Tricarico, J. Wozniak, and P. Guiod, “Speech motor control: Acoustic goals, saturation effects, auditory feedback and internal models”, Spec. Comm., 22, pp. 227-250, 1997.

8. S. Wood, “A radiographic analysis of constriction locations for vowels”, J. of Phonetics, 7, pp. 25-43, 1979.

9. J. S. Perkell, and W. L. Nelson, “Articulatory targets and speech motor control: A study of vowel production”, in S. Grillner, A. Persson, B. Lindblom, and J. Lubker, (eds.), Speech Motor Control, pp. 187-204, 1982.

10. C-K. Chuang, J. H. Abbs, and R. Netsell, “A study of tongue-jaw coordination: Possible role of tongue-hard palate contact in vowel production”, J. Acoust. Soc. Am., 63, S33 (A), 1978.

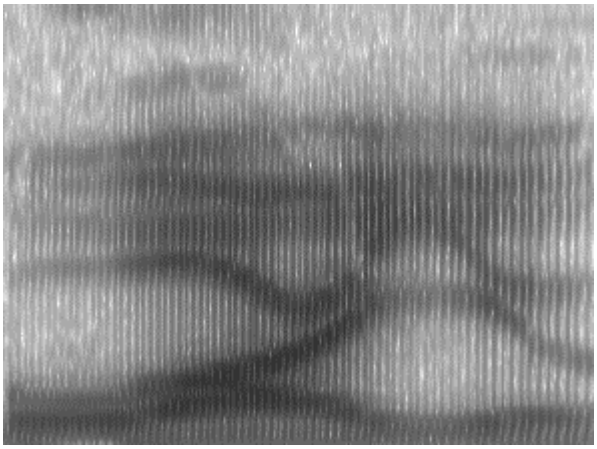

Figure 1a: Spectrogram of the original sentence.

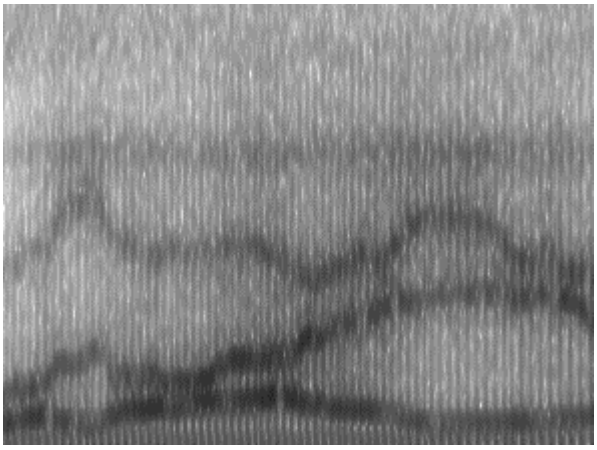

Figure 1b: Spectrogram of the synthetic sentence – old decision algorithm.

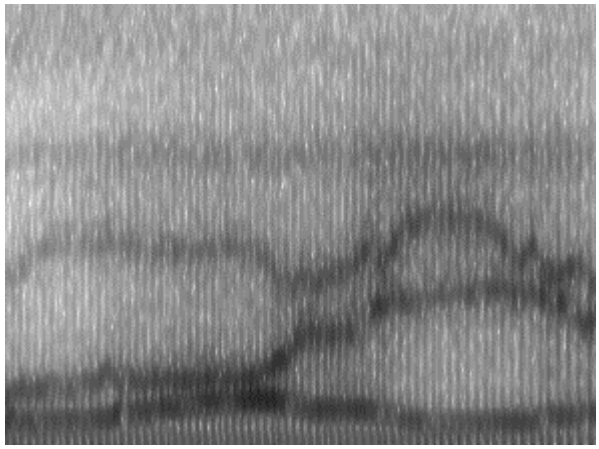

Figure 1c: Spectrogram of the original sentence – new decision algorithm.