ARTICULATORY SPEECH SYNTHESIS USING DIPHONE UNITS

Andrew R. Greenwood

School of Electrical Engineering, Electronics and Physics
Liverpool John Moores University, Byrom Street
Liverpool, United Kingdom, L3 3AF
E-mail: EEEAGREE@LIVJM.AC.UK

ABSTRACT

Two different parametric models of the vocal tract have been developed. These have been used to obtain area functions for use in an articulatory synthesiser based on the Kelly-Lochbaum model. Random sampling of the geometric space spanned by the model has been performed to obtain a codebook for use in spectral copy synthesis. A dynamic programming search of this codebook produces intelligible synthetic speech, but the overall quality is limited by the density of codebook entries in articulatory space. To increase the coverage without significantly increasing the codebook size, a method of generating several small codebooks, each of which covers a small amount of acoustic space has been developed. By using codebooks which map the regions of acoustic space defined by voiced diphones, it has been possible to significantly improve the quality of the synthetic speech.

1. INTRODUCTION

There is great demand for high quality synthetic speech for a variety of applications. While Formant synthesisers have had significant success, they are based on a rather artificial model of the vocal tract. Articulatory synthesisers which attempt to model the actual speech production mechanism promise greater naturalness, but at the expense of a higher computational load.

Articulatory synthesisers model the vocal tract as the concatenation of several uniform cylinders as shown in figure 1. Solution of the wave equations in each tube leads to a time domain filter structure known as the Kelly-Lochbaum structure [1]. The structure for tube k, and its junction with tube k-1 is shown in figure 2. The refection coefficient r_k is given by:

$$r_k = \frac{A_{k-1} - A_k}{A_{k-1} + A_k} \tag{1}$$

The frequency dependent losses are approximated by introducing a fixed attenuation coefficient in the forward and reverse paths with value given by:

$$\alpha_k = 1 - \frac{0.004\Delta x}{\sqrt{A_k}} \tag{2}$$

The lip termination is modelled as an inductor in parallel with a resistor with values:

$$\begin{aligned} R_L &= 1.44 Z_1 \\ L_L &= 24 ac Z_1 \end{aligned} \tag{3}$$

where a is the lip radius, c is the velocity of sound in moist air, and Z_1 is the impedance of the first section ($39.9/A_1$).

The glottal shunt impedance is modelled as a resistor of 91Ω in series with an inductor of 6.8mH [2]. These components are converted to digital equivalents using the bilinear transform.

The sampling rate is determined by the section length. For a 17.5cm tract divided into 20 sections, this is 20kHz. This is then decimated by a factor of two to give a final sampling rate of 10kHz. The high sampling rate reduces the effect of frequency warping introduced by the bilinear transform.

2. ARTICULATORY MODELS

In order to generate synthetic speech it is necessary present sequences of 20 areas (known as area functions) to the synthesiser. While it is possible to specify all 20 areas independently, this is not desirable because of the large amount of parameters that must be specified and because it is easy to generate area profiles that are anatomically impossible.

A better approach is to use some form of parametric model that allows the entire area function to specified by a few variables. Two different parametric models have been developed from measurements made using Magnetic Resonance Imaging [3]. One is based on the Mermelstein model [4] and uses nine parameters to specify the location of key articulators such as the jaw and the tongue. These are used to construct a mid-sagittal vocal tract outline which is then sampled at a uniform spacing to generate an area function. The second model is a quasi articulatory model [5] where nine parameters are used to specify the vocal tract directly. The first model is slightly more flexible while the second is easier to manipulate and requires less computation to map model parameters to area samples.

3. CODEBOOK GENERATION

In order to perform articulatory spectral copy synthesis it is necessary to be able to obtain model parameters directly from the natural speech waveform. This is difficult as this inverse operation can not be specified analytically and is non-unique.

A popular approach is to use a codebook [6]. This is produced by generating thousands of random sets of synthesiser parameters and storing them along with a spectral representation of the corresponding synthesiser output. Synthetic speech may then be generated by dividing the natural speech waveform into a series of frames, and for each frame in turn searching the entire codebook to find the best match using a suitable spectral distance measure. This is shown in figure 3.

The problem of the non-uniqueness of the mapping can be reduced by using a two component cost function when searching the codebook which also penalises changes in area function between successive frames. This can be improved further by applying a dynamic programming search algorithm [7].

4. DIPHONE SYNTHESIS

Using a single large codebook suffers from several disadvantages: some regions of acoustic space that are used by a given speaker may be inadequately represented by the codebook while other regions that are never used may contain many entries. Thus in order to ensure a good coverage of the entire acoustic space an extremely large codebook is required which requires a large amount of storage and takes a prohibitively long time to search.

The approach taken here is to identify regions of space used by a given speaker and to densely populate these areas. t was decided to divide the acoustic space into several zones each defined by a voiced diphone, and to generate a separate codebook for each region.

A diphone is defined to be a segment of speech corresponding to the stationary part of one phoneme to the stationary part of the next phoneme. The sentence "we were away a year" (/wi wər əweɪ ə jɜ:/) can be divided into diphones: wi-iw-wə-ər-rə-əw-weɪ-ɪə-əj-jɜ. This can be synthesised from a total of six diphones: wi,əw,ər,weɪ,ɪə, and jɜ.

A large codebook containing 23686 entries was generated using the random sampling method. The region of acoustic space corresponding to each diphone was identified and the codebook entries lying within that region were selected. Figure 4 shows the entries for codebook /wi/ along with an actual formant track from a real utterance. This codebook contains 686 entries. These codebooks were then extended by a factor of 10 by adding a small amount of random jitter to each entry, this produces a more densely populated codebook. Figure 5 shows the locations of the entries in the extended version of codebook /wi/.

It has been found the quality of speech produced by these codebooks is better than that synthesised from a single codebook. Also there was little difference between the two articulatory models, although the second model was significantly easier to use.

5. RESULTS

A spectral distortion measure was defined as:

$$d = \frac{1}{N_f} \sum_k \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left(P_i^{(k)} - \hat{P}_i^{(k)} \right)^2} \qquad (4)$$

where $P_i^{(k)}$ and $\hat{P}_i^{(k)}$ are the i'th frequency samples for frame k and N_f is the number of voiced frames.

Tests were performed using the utterance "We were away a year". For the single large codebook, the distortion produced by a simple search using a filtered-liftered-cepstral distance measure was 4.32dB, adding a simple geometric penalty reduced this to 4.10dB, and the use of dynamic programming reduced this to 4.02dB. Using the diphone codebooks and dynamic programming produced a spectral distortion of 4.02dB and using the expanded codebooks produced a distortion of 3.86dB. The perceptual quality of the speech was significantly improved

6. CONCLUSION

Two different articulatory models have been developed and used to generate codebooks, for use by an articulatory synthesiser based on the Kelly-Lochbaum structure. A method has been developed for generating small, densely populated codebooks each designed to generate a voiced diphone. The quality of the synthetic speech has been significantly improved.

7. ACKNOWLEDGEMENTS

The author is grateful to Dr. C.C. Goodyear from the University of Liverpool for his work in helping develop the models and for is continual advice and support during the duration of the project.

8. REFERENCES

[1] Kelly J.L, Lochbaum C.C. "Speech Synthesis," 4th Int. Cong. Acoustics, Copenhagen, Denmark, pp. 1-4, 1962.

[2] Flanagan J.L. "Speech Analysis, Synthesis and Perception," 2nd Ed., Springer-Verlang, 1972.

[3] Greenwood A.R., Goodyear C.C., Martin P.A., "Measurements of Vocal Tract Shapes using Magnetic -Resonance Imaging," Proc. IEE Part I,1992, Vol. **139**, pp. 553-560.

[4] Mermelstein P., "Articulatory Model for the Study of Speech Production," J. Acoust. Soc. Am., 1973, Vol. **53**, pp. 1070-1982.

[5] Greenwood A.R., Goodyear C.C., "Articulatory Speech Synthesis using a Parametric Model and a Polynomial Mapping Technique," ISSIPN'94, Hong Kong, pp. 595-598.

[6] Schroeter J, Mayers P, Parthasarthy S., "Evaluation of Improved Codebooks and Codebook Access Distance Measures," ICASSP'92, Albuquerque, USA, pp. 393-396.

[7] Schroeter J. Sondhi M.M., "Dynamic Programming Search of Articulatory Codebooks," ICASSP'89, Glasgow, Scotland.

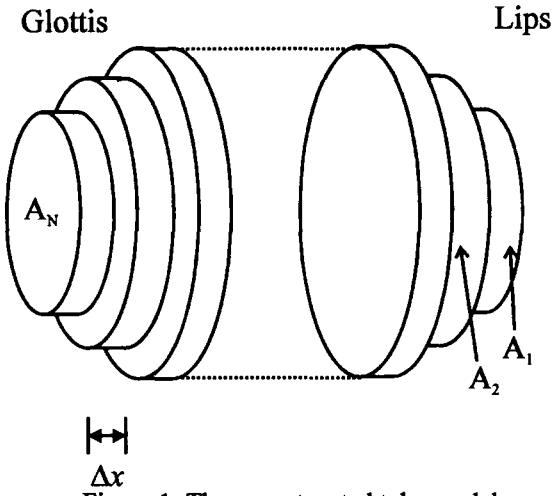

Figure 1: The concatenated tube model.

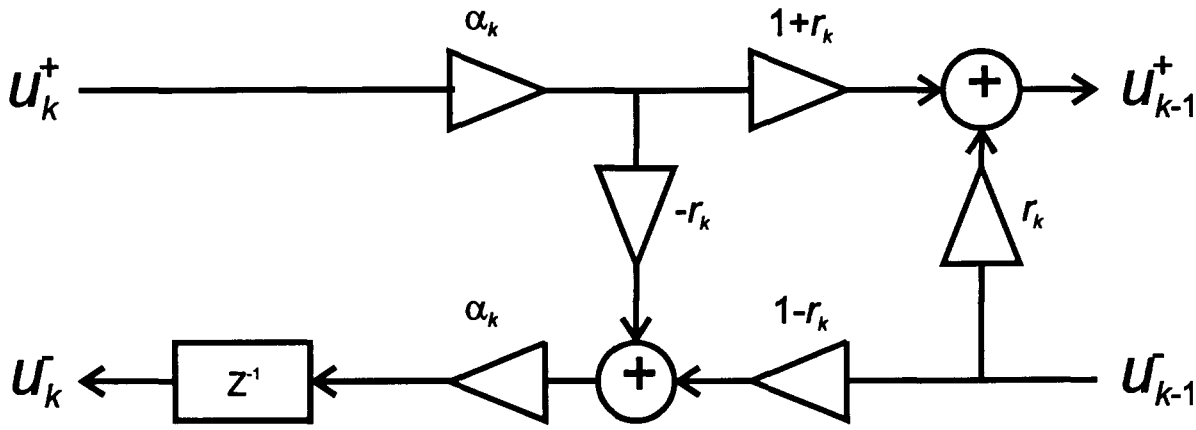

Figure 2: The Kelly-Lochbaum model.

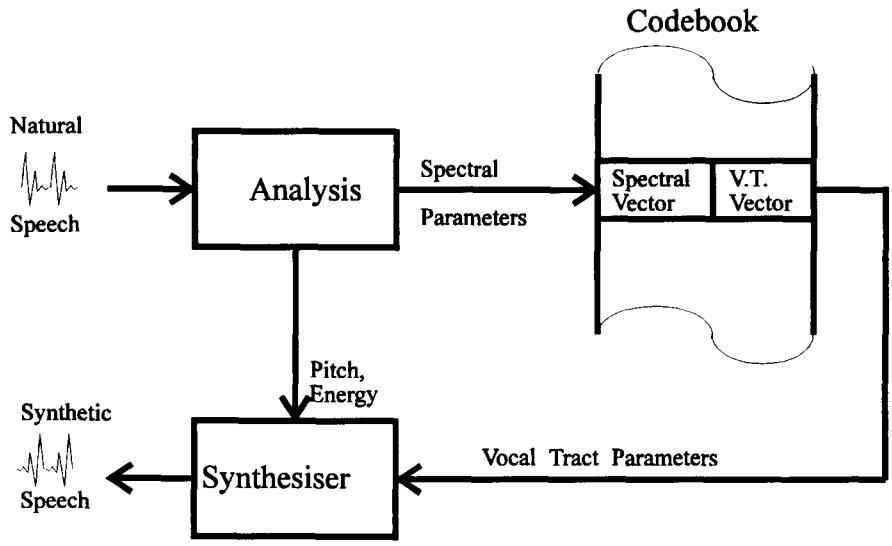

Figure 3: Spectral copy synthesis using a codebook

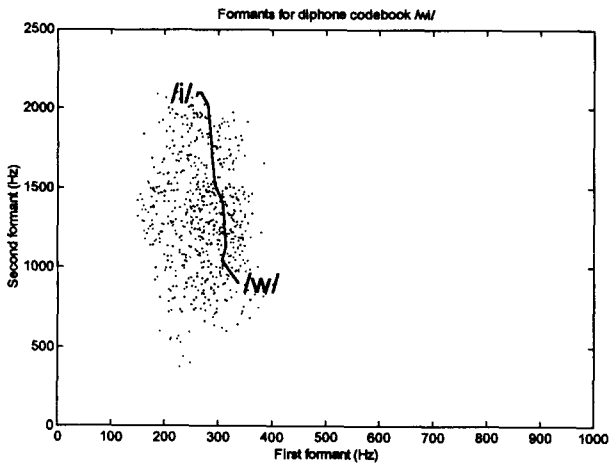

Figure 4: F1/F2 Space of codebook /wi/ along with formant track.

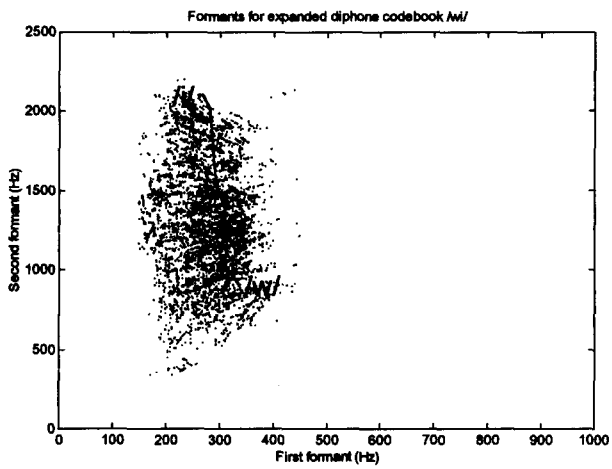

Figure 5: F1/F2 Space of expanded version of with formant tract.