PARAMETER ESTIMATION FOR SPECTRAL MATCHING IN ARTICULATORY SYNTHESIS

M.G. Rahim and C.C. Goodyear

Abstract

A major difficulty in generating natural sounding speech using articulatory synthesis is the lack of sufficient data on the parameters of the synthesiser and the motions of the articulators. This paper describes a method for estimating the parameters of the nasal section of a vocal tract model, using input speech spectra estimated using a conventional technique. The use of a neural net in estimating the coupling into the nasal tract is also described.

Introduction

The production of synthetic speech from a code sequence which has been derived from natural speech is usually described as synthesis-by-analysis. It involves two basic tasks: (a) that of obtaining a spectral estimate for each frame of natural speech samples, typically lasting 10 to 20ms, (b) the use of this spectral information to adjust the parameters of the synthesiser so that its output will have a spectrum which matches the original as closely as possible. Speech synthesis-by-analysis may thus be regarded as a spectral estimation and matching problem. In this paper the task of parameter estimation for a synthesiser based on an articulatory model is discussed.

The filter model

Most successful synthesisers use a source-filter model of speech production in which, for voiced sounds, the filter modifies the spectrum of a pitch-pulse sequence produced by the source. The filter may be implemented as a serial structure, as in the Klatt synthesiser, or as a parallel combination of band-pass filters, as in the Holmes formant synthesiser. Alternatively, the filter may be designed to model the acoustic behaviour of the vocal tract more directly. In the work reported here we have adopted the last approach, since we believe that transitions between phonemes may be more easily described in terms of the parameters of an articulatory model of this type, than of those of a filter whose structure is more remote from the structure of the vocal system.

The chosen filter models the acoustic equations of a lossless branched tube, constructed from sections of the same fixed length but of adjustable areas, beginning at the glottis and later branching into nasal and oral tracts. The corresponding digital filter structure, based on that of Kelly and Lochbaum (1), is shown in figure 1. The five adjustable areas between the glottis and the velum, five from velum to lips and seven from velum to nostrils provide a total of 17 areas which determine the multipliers shown. The outputs are summed and the net output may be given a 6 dB per octave lift in order to model radiation effects. A simpler alternative, used here, is to de-emphasise the output and to drive the filter using impulses instead of naturally shaped glottal pulses, whose spectra have approximately a 12 dB per octave droop.

Department of Electrical Engineering & Electronics, University of Liverpool.

Parameter estimation

For non-nasal sounds the velum closes the coupling area into the nasal branch so that the coefficient S2 in figure 1 is zero. The area parameters of the resulting unbranched tube may be estimated from the speech waveform using pitch-synchronous covariance analysis on the pre-emphasised signal (2,3). These estimates may then be used as a starting point for optimisation.

During a nasal consonant the nasal coupling area is open and the oral tract is stopped at some point between the velum and the lips. The resulting transfer function contains zeros as well as poles, so that covariance analysis is no longer a valid way to estimate areas. Appropriate area sets for each of the sounds /n/, /m/ and /ŋ/ were therefore estimated using a different technique, as follows. Examples of these phonemes, produced by the same speaker, were recorded. Hamming-windowed 20 ms segments from these were analysed by the autocorrelation method at 14th order, to provide corresponding target spectra. Starting with initial guesses for nasal and vocal tract areas, the synthesiser was driven with fixed energy and pitch and its output spectra were similarly estimated. Distance measures between these spectra and their targets were then minimised using a steepest descent method. The descent vectors included either vocal or nasal tract areas alternatively and the algorithm iterated among the three pairs of spectra, allowing the vocal tract areas to differ but keeping the same nasal tract areas in each case. A zero area was imposed at an appropriate point in the oral tract for each phoneme. In this way it was found possible to achieve acceptable spectral matches, as shown by the example for the consonant /n/ in figure 2.

During synthesis of speech from this speaker, only 11 of the original 17 area parameters now need to be adjusted. For non-nasal vowels these reduce to 10, while for nasal consonants the coupling area into the optimised nasal tract must be included. If the acoustic model is sufficiently accurate, it should also be possible to adjust these parameters to provide matching to the nasalised vowels, such as occur during transitions into and out of nasal consonants.

Recognition of nasal consonants

The presence or absence of coupling into the nasal tract is clearly an important parameter in our synthesis filter and a reliable indicator of nasalisation is needed. Experiments were therefore performed to explore the use of a multi layer perceptron to provide such an indicator. Exemplars for training were obtained from sets of spectra like the one shown in figure 3. This shows natural speech spectra, again from 14th order autocorrelation, during the utterance "mum". The strongest feature of nasalisation is seen at low frequencies and particularly by the appearance of the 'nasal formant' at around 200 Hz. The high frequency information was therefore discarded and ten sample values of the log spectrum between 100 and 1000 Hz were taken as inputs to a two-layer perceptron with five hidden units and a single output unit. The target output was taken to be 0.1 when the spectrum was clearly that of a non-nasal vowel and 0.9 when it was clearly a nasal consonant. Some 20 examples of monosyllables containing nasal consonants were analysed in this way and provided 250 exemplar spectra and targets. Target values intermediate between 0.1 and 0.9 were selected, somewhat arbitrarily, for spectra from transition regions.

After training, the test sentence 'My nanny knew a mean woman', was recorded and spectra taken at 20 ms intervals were computed and applied as inputs to the MLP. The output is shown in figure 4 and is clearly successful in indicating the occurrence of the nasal consonants.

Further work

Multi-layer perceptrons are becoming useful as a means of recognising speech features and spectral estimates are commonly used as their input data. A main objective of the work described here is to investigate the use of an MLP to provide estimates of the area parameters of the synthesis filter described above. Results will be reported elsewhere which demonstrate the ability of a three-layer perceptron with 26 input units and 11 output units to provide a mapping between spectra and the corresponding areas of our model. These results are encouraging and, together with the work reported here, point to a possible rapid and reliable way of translating between natural speech and parameters for synthesis.

Acknowledgement

The authors are grateful to British Telecom Research Laboratories for their interest in and support of this work.

References

1. Kelly, J. and Lochbaum, C.: 'Speech synthesis', 4th Int. Cong. Acoust., paper G42, 1962.

2. Markel, J.D. and Gray, A.H.: 'Linear prediction of speech', Springer-Verlag, Berlin, 1976.

3. Kuc, R. and Han, H.: 'Errors in determining vocal tract shape from the acoustic signal', ICASSP87, pp.629-632, 1987.

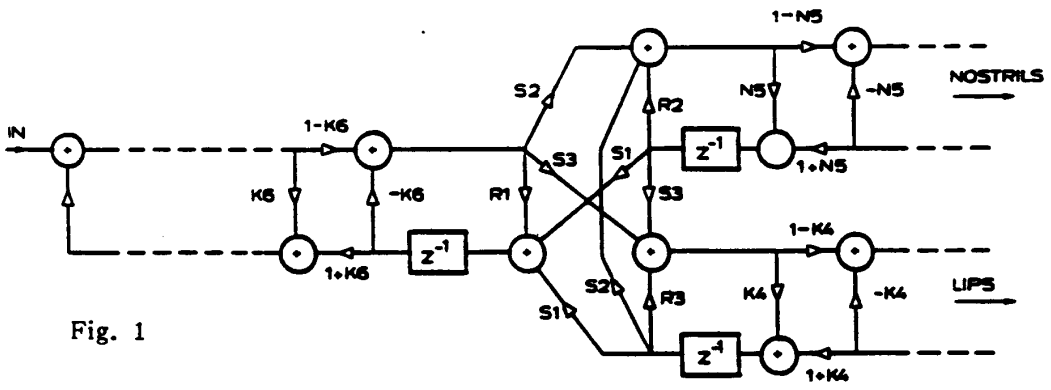

Fig. 1

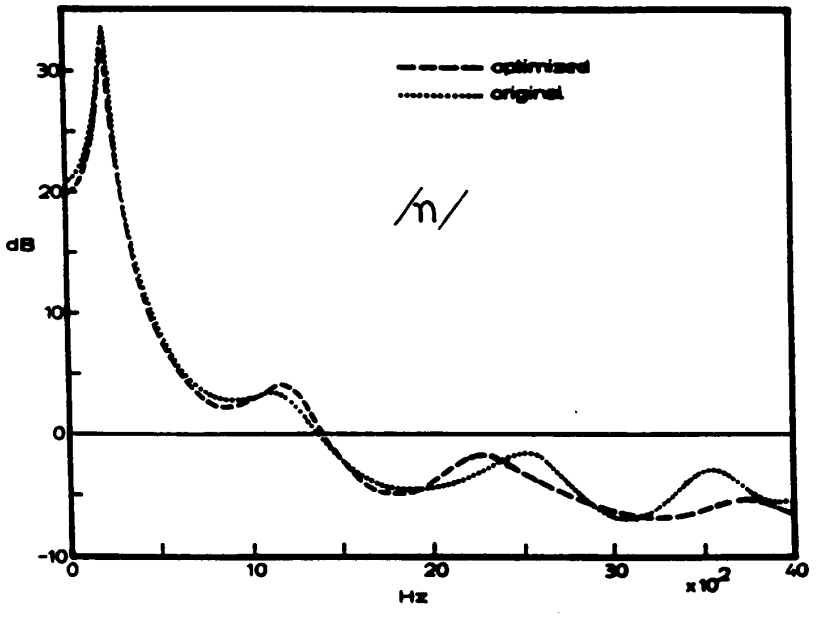

Fig. 2

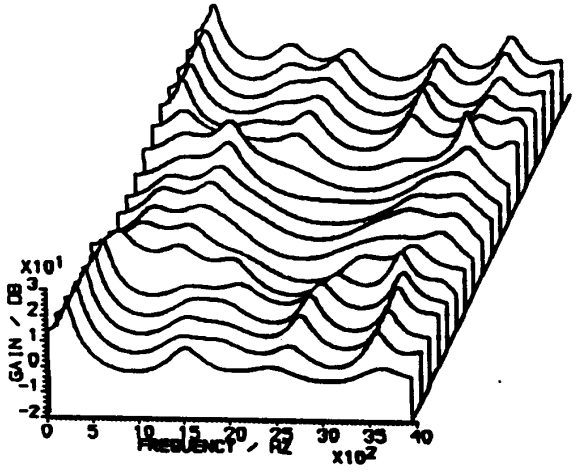

Fig. 3

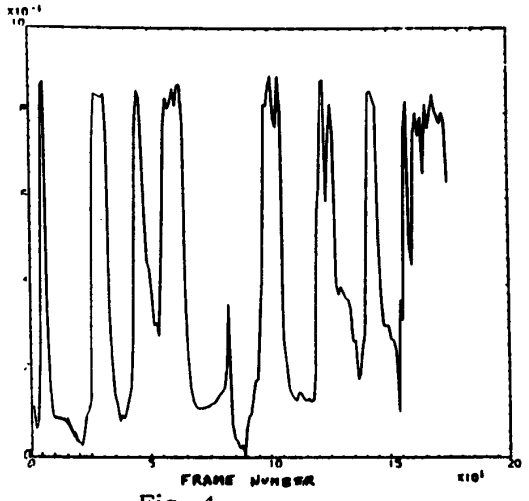

Fig. 4