DYNAMIC VOICE SOURCE SYNTHESIS

Sarah K Palmer (1) and David M Howard (2)

(1) Phonetics and Linguistics Department, University College London, Wolfson House, 4 Stephenson Way, London NW1 2HE, UK.
(2) Signal Processing: Voice and Hearing Research Group; Electronics Department, York University, Heslington, York YO1 5DD, UK

ABSTRACT

Analysis of the excitation waveform modelled by a four parameter model of glottal flow has revealed consistent variations in shape due to laryngeal co-articulation. LPC resynthesis using the model demonstrates that changes to the shape of the excitation affect the quality of the speech in a way predicted by the findings of the analysis. Pilot studies suggest that dynamic excitation provides a more natural LPC resynthesis than non-dynamic excitation.

1. INTRODUCTION

Previous work using the JSRU parallel formant synthesiser [8] has failed to establish a preference for utterances synthesised with a dynamically varying excitation based on the three parameter model of glottal flow [2] over those synthesised using a static excitation waveform with pre-determined spectral slope.

A number of reasons were identified which could be contributing to this lack of preference. Firstly there are difficulties in obtaining formant amplitudes for the parallel formant synthesiser which correspond to the supra-glottal tract configuration alone. Secondly, previous work was based on a three parameter model of glottal flow [4] derived from the laryngographic waveform (Lx). More detailed models of glottal flow exist (eg: [3]). In addition the use of Lx (a measure of vocal fold contact) to derive a model of glottal flow has yet to be justified.

Thirdly, the test stimuli used as a basis for naturalness testing [8] were [ɑ:hɑ:] and [ɑ:ʔɑ:]. Whilst these stimuli demonstrate clear differences in the closed quotient trends due to laryngeal co-articulation, they rarely occur in natural British English speech and are not ideal for naturalness testing. The quality of the parallel formant resynthesis of these tokens was rather poor, and changing the voice source parameters gave no reliable perceptual judgements. Whilst this could be a function of our voice source model, we believe that subjects will become sensitive to voice quality differences *only* as the overall intelligibility of the synthesis improves. This view is supported by Pickering [9].

The aim of this work is to re-analyse the natural data using a four parameter model of glottal flow [3], to investigate the changes in the time course of the flow parameters, and to study the effects of altering these parameters based on LPC resynthesis via perceptual tests. The problems of formant amplitude estimation in parallel formant synthesis are therefore avoided.

2. METHOD

The fully automatic inverse filtering programs used in this work were developed by Chan and Brookes [1]. They carry out an LPC closed phase analysis from which the inverse filter is then calculated. This filter is then applied to the speech pressure waveform and a raw estimate of the differentiated glottal flow obtained. A four parameter Fant, Liljencrants and

Lin [3] model of the derivative of glottal flow (L-F) can then be fitted to the raw inverse filtered waveform. The LPC coefficients are then re-estimated using the modelled excitation waveform.

The output of the inverse filter was modelled automatically using the L-F model of glottal flow for the utterances [ɑ:hɑ:] and [ɑ:ʔɑ:] spoken by four male and four female speakers. Figure 1 shows a typical cycle of the L-F model and the time alligned equivalent cycle of glottal flow.

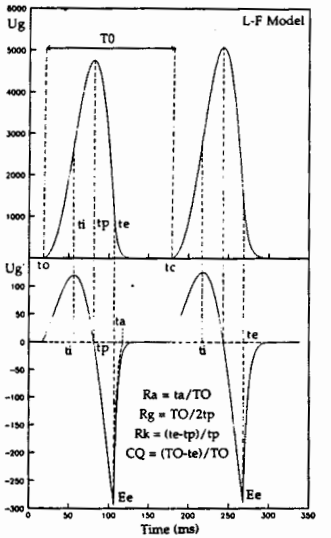

Figure 1: The four parameter L-F model of the derivative of glottal flow (Ug') with the derived glottal flow (Ug).

Analysis of the utterances was carried out in terms of five measurements: Rk (a measure of the asymmetry of the flow), Ra (the return phase ratio), Rg (the glottal frequency), CQ (the closed quotient) and Ee (the strength of the excitation). The way in which these ratios have been calculated is indicated in figure 1 along with the parameters from which they are derived: T0, tp, te and ta. Further details of the analysis ratios can be found in [5] and [7].

LPC resynthesis was carried out using the re-estimated filter function and various modelled excitation waveforms based on the five measurements above. A pilot test was carried out to compare the natural utterances of [ɑ:ʔɑ:] with three synthesised versions for three speakers, two male and one female. For each speaker the filter function remained constant whilst the excitation model was changed as follows:

(a) the shape of the excitation was varied dynamically on a cycle-by-cycle basis,
(b) the excitation was fixed according to the average values of Rk, Ra and Rg for the whole utterance,
(c) the excitation was fixed according to the average ratio values of Rk, Ra and Rg measured from the mid-portion of the second vowel in the utterance.

Perceptual testing was carried out to evaluate the naturalness of the stimuli. Four subjects were able to replay each stimulus through headphones as many times as required, and they were asked to mark down the stimulus which they perceived to be most like the natural utterance. For three of the tests subjects listened to the whole utterance whilst in a further three tests they only heard the initial vowel.

A further set of perceptual tests was carried out to study the effect of changes in the excitation on voice quality. Three stimuli were prepared based on the voice source analysis of one female speaker as follows:

(a) the excitation shape was fixed to the mean ratio values of the mid-portion of the second vowel in the utterance,
(b) Ra and Rk were increased and Rg decreased by 30% of their mean values to simulate a more breathy voice quality,
(c) Ra and Rk were reduced and Rg increased by 30% of the mean value to simulate a less breathy voice quality.

Five subjects were asked to rank the stimuli (presented over headphones in a sound-proof room) in terms of 'breathiness'.

3 RESULTS

3.1 Analysis

Results of the analysis confirm the previous finding, using Lx, that closed quotient increases before a glottal stop and decreases before a glottal fricative. This result is to be expected from an inverse filter analysis given a previously demonstrated high correlation between closed phase measurements made from Lx and inverse filtering [6].

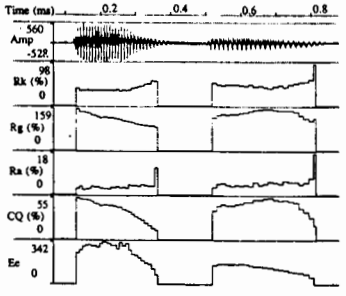

Figure 2. Analysis of [ɑ:hɑ:] showing the speech pressure waveform and changes in Rk, Rg, Ra, CQ and Ee over time.

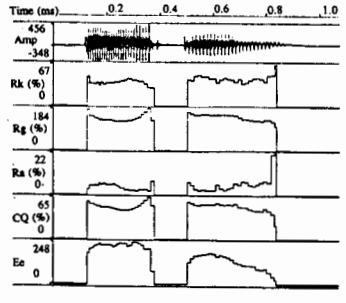

Figure 3: Analysis of [ɑ:ʔɑ:] showing the speech pressure waveform and changes in Rk, Rg, Ra, CQ and Ee over time.

Figures 2 and 3 show typical analysis findings for the utterances [ɑ:hɑ:] and [ɑ:ʔɑ:] spoken with falling intonation and the stress on the initial syllable for one male speaker.

3.2 Perceptual Tests

Results of the comparison between test stimuli and the natural utterance show that two of the subjects had a significant preference for utterances resynthesised using a dynamic excitation over those resynthesised with a fixed excitation shape (α=0.05). However, results from two further subjects were insignificant Overall there was no significant preference for the dynamic excitation. Focussing on the initial part of the utterance, where the main changes in the excitation are taking place, seems to have no effect on the preference results.

In the second test all the subjects chose stimulus (b), in which Ra and Rk had been increased and Rg decreased by 30%, as having the most breathy voice quality and stimulus (c) as having the least breathy quality.

4. DISCUSSION

The changes occuring in Rk, Ra and Rg have previously been linked to variations in the strength of the excitation which correlates strongly to the parameter Ee. It was hypothesised [5] that the sharpness of closure ratio Ra should vary inversely with Ee, (the stronger the excitation the more rapid the closure and therefore the shorter the return phase). When the glottal frequency Rg is fairly constant Rk varies with Ra in such a way that if the excitation strength Ee is large and Ra is small, the asymmetry of the pulse will increase due to a relatively shorter return phase (ie: Rk is small). On the other hand if the excitation strength is weak and Ra is high, the asymmetry will decrease and Rk will be high. Therefore at the onset and offset of voicing one would expect Rk and Ra to be higher than during the mid-point of a vowel in the utterance. This is confirmed by some of our data but the presence of a glottal stop seems to affect the relationship between these parameters.

Perhaps a clearer explanation is offered by studying the changes taking place in the closed quotient. When the closed quotient (CQ) is rising, as it does before a glottal stop (see figure 3), the length of the open phase becomes relatively shorter, resulting in a smaller value of 'tp'. This results in a rising Rg value since it is proportional to the inverse of 'tp', and the asymmetry of the pulse Rk decreases. The opposite effect is shown when CQ decreases in figure 2. CQ tends to be lower utterance initially and utterance finally and can therefore account for the changes taking place in Rk and Rg in these regions. The interpretation of the variation in the parameter Ra in our data is not clear, but it seems to depend mainly upon the the strength of the excitation Ee.

This work has demonstrated that voice source changes found in natural speech can be resynthesised and that modifications to the shape of the excitation waveform in LPC re-synthesis can alter perceived speaker quality appropriately, both for male and female speech. For some listeners dynamic excitation provides a closer perceived match to the natural speech than an excitation with a fixed waveshape. Whilst the number of subjects used in this test was limited and a more widespread study is needed, it is thought that the use of sentence level material instead of isolated utterances will produce a clearer preference for a time varying excitation due to different intrinsic properties of the phonemes within the utterance as well as co-articulation, stress and intonation effects. Our experience suggests that the overall quality of synthesis from the JSRU synthesiser requires improvement before excitation changes will affect the perceived naturalness significantly. Work is in progress to modify the methods used to specify the formant amplitude data used by the JSRU system.

5. ACKNOWLEDGEMENTS

This work was supported by the SERC research grant number GR/F/30642.

6. REFERENCES

[1] CHAN D.S.F.& BROOKES D.M, (1989), "Variability of Excitation Parameters Derived from Robust Closed Phase Glottal Inverse Filtering", *European Conference on Speech Communication and Technology*, Paris 199-202.

[2] FANT G, (1979), "Glottal Source and Excitation Analysis", *Speech Transmission Lab: Quarterly Progress and Status Report, 1*, Royal Inst. of Technology, Stockholm, 85-107 .

[3] FANT G., LILJENCRANTS J. and LIN Q, (1985), "A Four Parameter Model of Glottal Flow", *Speech Transmission Lab: Quarterly Progress and Status Report, 4*, Royal Inst. of Technology, Stockholm, 1-13.

[4] HOWARD D.M. and BREEN A. P, (1989), "Methods for Dynamic Excitation control in Parallel formant speech synthesis", *IEEE Proceedings of the International Conference on Acoustics, Speech and Signal Processing - 89, 1*, 215-218.

[5] GOBL C. and CHASAIDE A, (1990), "Linguistic and Paralinguistic Variation in the Voice Source", *Proceedings of the International Conference on Spoken Language Processing*, Japan, 85-88.

[6] HOWARD D.M, LINDSEY G. and ALLEN B, (1990), "Towards the Quantification of Vocal Efficiency", *Journal of Voice, 4*, 205-212.

[7] KARLSSON I, (1990), 'Voice Source Dynamics for Female Speakers', *Proceedings of the International Conference on Spoken Language Processing*, Japan, 69-72.

[8] PALMER S.K, ALLEN B, HOWARD D.M, LINDSEY G. and HOUSE J, (1990), 'Analysis, Synthesis and Perception of Laryngeal Co-articulation', *Proceedings of the Institute of Acoustics, 10*, 17-24.

[9] PICKERING J. B, (1989), 'Effects of Voice Type and Quality on the Intelligibility of a Text-to-Speech System', *European Conference on Speech Communication and Technology*, Paris, 637- 639.