Proceedings of the Institute of Acoustics

ANALYSIS, SYNTHESIS AND PERCEPTION OF LARYNGEAL CO-ARTICULATION

S K Palmer (1), B Allen (1), D M Howard (2), G Lindsey (1) & J House (1)

(1) Department of Phonetics and Linguistics, University College London
(2) Department of Electronics, University of York

0. ABSTRACT

This paper describes experiments on the analysis, synthesis and perception of laryngeal co-articulations in [ɑ:hɑ:] and [ɑ:ʔɑ:], with a view to developing automatic rules for dynamic excitation. Analyses show that there is a significant decrease in larynx closed quotient (CQ) before the glottal fricative and an increase before the glottal stop. Syntheses of the utterances spoken with falling intonation were carried out for one male speaker. Perceptual testing showed no significant preference for utterances synthesised with dynamically varying CQ over utterances synthesised with the CQ set to a fixed value by the experimenter, either in terms of naturalness or similarity to the original utterance. An additional series of perceptual tests was carried out to determine the ability of the listener to discriminate between stimuli with different fixed overall CQ values. The findings suggest that, given the quality of the current synthesis, fine variations over time in the CQ do not need to be modelled. However changes in the fixed overall CQ do have a perceptible affect on the quality of the synthetic speech.

1. INTRODUCTION

Co-articulation effects are often realised at the laryngeal level in different segmental and prosodic environments. One method of measuring these effects is by studying the output waveform from the electrolaryngograph (Lx) [1] and observing the cycle-by-cycle changes in the closed quotient (the percentage of each larynx period that is taken up by the closed phase). A study using this method by Lindsey et al [2] has shown that laryngeal co-articulation can occur in English VCV sequences, specifically in the form of a breathy vowel offset before [s]. Lindsey et al suggest that this results in a reduction in the higher frequency energy associated with increased sub-glottal damping.

The first aim of the present work was a more systematic study of this effect, concentrating on the laryngeal level by choosing intervocalic [h] and [ʔ] as data, in order to test the hypothesis that CQ decreases in anticipation of a following voiceless fricative (during which the vocal folds will be abducted) and that CQ increases before [ʔ] (or a glottally reinforced stop), in preparation for the complete closure at the glottis. (Klatt and Klatt [3], also analysing [h] and [ʔ], but in reiterant sentences, found no anticipatory co-articulation differences between the fricative and the stop. However, analyses were carried out at the midpoint of the vowels and it was hypothesised that any co-articulation effects present did not stretch leftwards from the consonant as far as the midpoints of the vowels studied.)

The second aim of the present work was to replicate the analysed CQ variation in the JSRU parallel formant synthesiser. Holmes [4] has shown that this synthesiser is capable of producing speech which is almost indistinguishable from the natural. However his utterances

were obtained by considerable hand-editing of the formant amplitudes and bandwidths, modelling both changes in vocal tract articulations and dynamic changes in the excitation spectrum. These amplitudes and bandwidths are therefore not strictly appropriate in terms of the source-filter theory of speech production [5] basic to the synthesis system.

The current excitation for the JSRU synthesiser is based on the second time differential of a typical glottal flow pulse, and consists of a stored low pass filtered excitation cycle with a predetermined spectral slope, which is repeated at the desired fundamental frequency, giving the synthesiser excitation waveform EX. The glottal flow model waveform (referred to as GA -- "the glottal area") is available in the synthesis to modify the formant amplitudes and bandwidths on a cycle-by-cycle basis. One of the advantages of the JSRU synthesiser is that it allows experimenters to generate their own excitation waveforms externally to the synthesiser and readily study the effects. An algorithm developed by Howard et al [6] generates the excitation waveforms (EX and GA) cycle-by-cycle for the synthesiser, based on the Fant three-parameter model [7], using parameters measured from the Lx waveform. Pilot studies have shown that this dynamic excitation route added some voice qualities of the original speaker which the stored EX and GA were unable to add.

The third aim of the present work was to assess the naturalness of the syntheses by means of perceptual testing.

2. METHOD

2.1 Analysis

Recordings of twelve VCV sequences spoken by four phonetically-trained male speakers were made under anechoic conditions. The consonants used in this study were the glottal fricative [h] and the glottal stop [ʔ], chosen in order to minimise the filtering effects of the vocal tract. The sequences [ɑːhɑː] and [ɑːʔɑː] were spoken with three different intonation contours (falling, rising or level) and with two stress patterns (initial stress or final stress). The speech pressure waveform and the Lx waveform were recorded simultaneously onto two channels of a digital Sony PCM-F1 video recording setup.

The data was acquired onto a Masscomp 5600 at a 20kHz sampling rate, then analysed using an algorithm developed by Davies et al [8] under the "Speech Filing System" (SFS) protocols developed by Huckvale et al [9]. This outputs the following parameters for each cycle of the Lx waveform: the Lx period (Tx), the closed phase and open phase durations calculated by three different methods, and the corresponding CQ.

In all the methods the point of vocal fold closure is taken as the positive peak in the differential of the Lx waveform (Lx is first polarised with closing slope positive going). The point of opening is derived by three alternative methods: (i) as the negative peak in the Lx differential, (ii) as the crossing-point of a fixed ratio of 70:30 of the current cycle's peak to peak amplitude, or (iii) as the point at which the opening slope has the same amplitude value as the point of closure for that cycle. A comparison of the measurements of glottal closed phase derived on the one hand from inverse filtering, and on the other hand from the Lx waveform using the

second above-mentioned method on the other, has shown a high degree of correlation between measurements made from the two methods [10].

2.2 Synthesis

A pair of VCV sequences [ɑ:hɑ:] and [ɑ:ʔɑ:] for one male speaker was chosen for a study of the effects of dynamic excitation on the naturalness of the synthetic output. These utterances were spoken with a falling intonation contour and the stress on the second syllable, and showed clearly the effects of laryngeal co-articulation. Closed-phase LPC analysis was performed to derive the formant frequencies, and the amplitudes were mapped from a FFT analysis. In order to obtain an optimal utterance as a basis for the study of dynamic excitation the formant frequencies were hand-edited where necessary. EX and GA waveforms were obtained using the algorithm developed by Howard [7], which uses the Davies et al algorithm [8] (method two) to obtain its parameters. Two perceptual studies were then performed.

2.3 Perception

2.3.1 Test 1. The first set of perceptual tests required 12 subjects to perform a ranking task for sets of four stimuli, both in terms of their perceived naturalness and in terms of their similarity to the original utterance. There were two sets of tests for each consonant environment, one in which the natural utterance was played as a reference and one in which it was absent. The four test stimuli were synthesised versions of the natural which differed only in terms of their CQ values. The CQ was either kept constant throughout the utterance at 30%, 50% or 70%, or it was varied on a cycle-to-cycle basis according to the values obtained from the analysis of the Lx waveform. Each set of stimuli was randomised and plotted on the Masscomp terminal in a mouse-driven environment, enabling the subject to play each stimulus via headphones as often as required before ranking them. In each case the lowest rank was assigned to the most natural, or to the most similar to the original.

2.3.2 Test 2. A same-different discrimination test was carried out to determine the extent of CQ change necessary for subjects to detect a difference in the quality of the synthetic speech. Two sets of material were prepared, one for the glottal fricative and one for the glottal stop. The CQ was fixed in each utterance to a value between 30% and 70% inclusive, in intervals of 5%. This range of values represented the approximate range of CQ values found in the analysis of the male data. Eight subjects were presented with 128 stimulus pairs for each set of material. Each set consisted of all possible combinations of CQ intervals from 5% to 40%, e.g. 30% with 35%; 30% with 40%; 30% with 45% etc. In total there were eight pairs of each possible CQ interval making 64 pairs, and an equivalent number of pairs in which the stimuli were the same. The sets were then randomised and divided into six lists lasting approximately eight minutes each. Each stimulus pair was repeated three times. Tests were run in a sound-proof room and presented binaurally through headphones via audio-cassette. No familiarisation was considered necessary since all the subjects participating in this part of the study had completed Test 1.

3. RESULTS

3.1 Analysis

Analyses show clearly the anticipatory laryngeal co-articulations. Figure 1 shows a display of the speech pressure waveform, Lx waveform, Tx markers, the open and closed phase times and a plot of CQ for the glottal fricative, spoken by a male speaker with falling intonation and the stress on the second syllable. Figure 2 shows the equivalent plot for the glottal stop.

A series of Jonckheere trend tests [11] revealed that for all four subjects recorded in this study there was a significant decrease in the CQ before [h], as measured from the mid-point of the first vowel, and for three of the four a significant rise before the glottal stop. The mean of CQ means for each utterance was 47%.

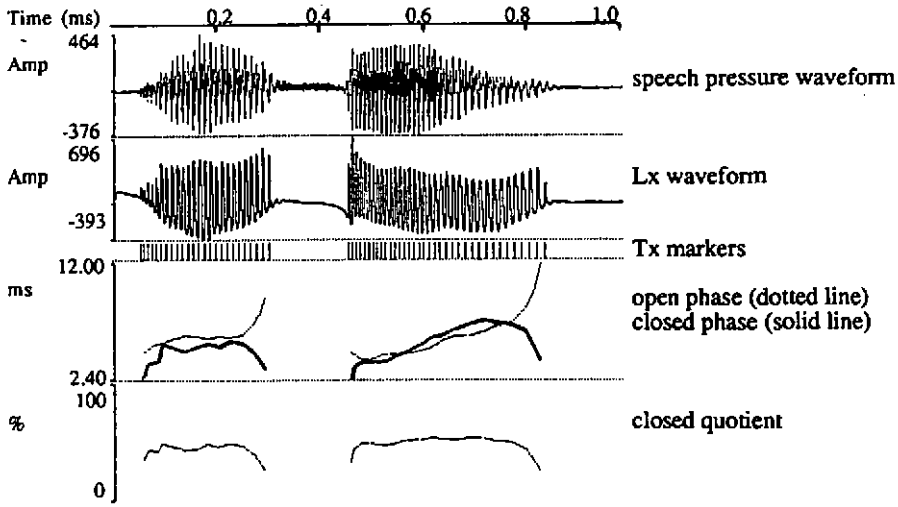


Figure 1 : [ɑ:hɑ:] spoken by a male speaker with falling intonation

ANALYSIS, SYNTHESIS AND PERCEPTION OF LARYNGEAL CO-ARTICULATION

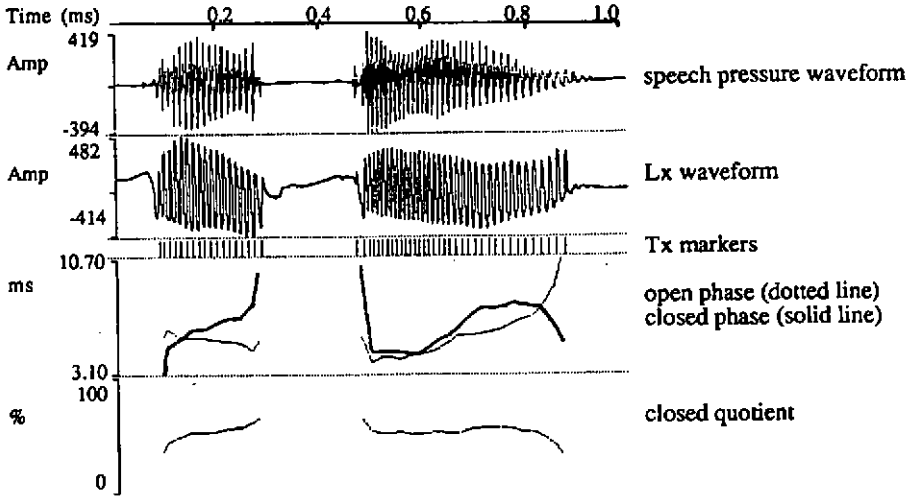


Figure 2 : [ɑ:ʔɑ:] spoken by a male speaker with falling intonation

3.2 Perception

3.2.1 Test 1. Tables 1 and 2 show the rank given to the test stimuli averaged over the 12 subjects. Each table shows the mean overall rank of the stimulus when presented with or without the natural stimulus.

CQ	mean rank without natural	mean rank with natural
dynamic	1.83	2.33
30 %	1.92	1.33
50 %	2.17	2.33
70 %	4.00	4.00

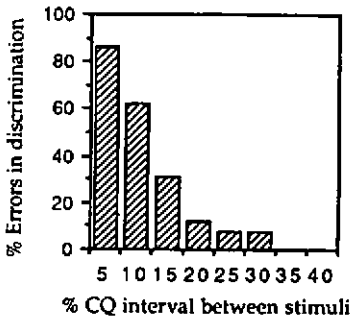
Table 1 : Showing the mean overall naturalness ranking for [ɑ:hɑ:]

CQ	mean rank without natural	mean rank with natural
dynamic	2.42	2.00
30 %	1.92	1.50
50 %	1.83	2.33
70 %	3.83	4.00

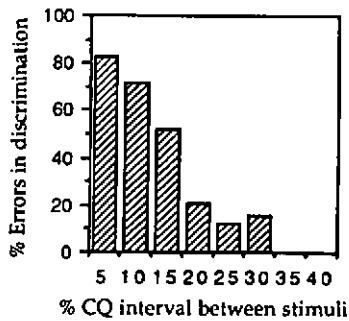
Table 2 : Showing the mean overall naturalness ranking for [ɑ:ʔɑ:]

A number of subjects commented on the differences between the utterance with a fixed CQ at 30% and 70%. The 30% utterance was considered to be more "breathy" and "clearer" whilst the 70% utterance was thought to be "tinny", "muffled" and more "American".

3.2.2 Test 2. Results of the discrimination tests are shown in graphs 1 and 2. These bar charts show the percentage errors on the vertical axis with each CQ interval being represented along the horizontal axis. A statistical analysis of the number of errors when the two stimuli were the same shows that subjects were not making random guesses, the percentage of errors for the [ɑ:hɑ:] stimuli being 7.4 % and 7.6% for the [ɑ:ʔɑ:] stimuli. There was significant discrimination between utterances differing by a CQ of 15% for the glottal fricative as defined by a relative standing z-score at a significance level of 0.05%. The discrimination was not as clear for the glottal stop and was significant for a 20% difference.



Graph 1: The errors in discrimation for different CQ intervals for [ɑ:hɑ:]



Graph 2: The errors in discrimation for different CQ intervals for [ɑ:ʔɑ:]

ANALYSIS, SYNTHESIS AND PERCEPTION OF LARYNGEAL CO-ARTICULATION

4. DISCUSSION

The displays of the Lx parameters shown in figures 1 and 2 are typical of all but one of the speakers, showing an increase in CQ before the glottal stop and a decrease before the fricative.

Perceptual testing has failed to indicate a preference for the dynamic as opposed to the static CQ. Presumably a CQ change needs to be sustained for a certain length of time before it alters the perceived quality of the synthetic speech. Differences in "breathiness" and "clarity" of the voice were reported, suggesting that appropriate modelling of the CQ changes should lead to more acceptable synthesis. CQ interval discrimination for the glottal stop was found to be slightly worse than for the glottal fricative.

The lack of a preference for the dynamic as opposed to the static CQ may be due to a number of factors which have yet to be taken into account:

Firstly, a method is needed by which the formant amplitude values obtained in analysis represent the vocal tract filter function only. LPC analysis assumes a flat excitation spectrum and therefore the formant frequencies and amplitudes obtained may not only correspond to the vocal tract configuration but also to features of the dynamically changing voice source. This is an appropriate analysis route for the JSRU synthesiser since it makes use of an excitation which has a flat spectrum. The effect of lip radiation is added later by means of -6dB/octave filter. However, when using a dynamically varying voice source with the JSRU analysis route, the associated formant amplitudes and bandwidths are effectively applied twice. Thus no direct comparison has been made here between the static excitation usually used in the JSRU synthesiser and our dynamic excitation versions.

Secondly, while this pilot investigation assumed the three-parameter model of glottal flow, it is clear that a more detailed model is required. The Liljencrants and Fant (L-F) four-parameter model [12] includes an exponentially decaying return phase from the point of maximum closing discontinuity towards maximum closure. This is of particular importance in modelling breathy voice, where it has been suggested that "discontinuity may occur in the middle of the descending branch of the flow followed by a less steep descent and a final trailing off corner effect" [12]. The current three-parameter model is unable to model this effect. Investigations using the L-F model are in hand [7]. It is also acknowledged that a problem remains in justifying the use of Lx (a measure of the vocal contact area) in the development of a glottal flow model.

Finally, it may be that the overall quality of the synthesised utterance affects the perception of CQ changes. The increasing sensitivity of the listener to voice quality as the overall intelligibility improves has already been suggested by Pickering et al [13]. Such a finding would suggest that, as the overall level of naturalness achieved in speech synthesis improves, appropriate modelling of CQ variations of the type discussed here will become increasingly important.

ANALYSIS, SYNTHESIS AND PERCEPTION OF LARYNGEAL CO-ARTICULATION

5. ACKNOWLEDGEMENTS

This work was supported by SERC research grant number GR/F 30642 and SERC award reference number 88303077. The authors would like to thank all the speakers who took part in the recordings and the subjects who gave their time freely to take part in the listening tests.

6. REFERENCES

[1] A J FOURCIN & E M R ABBERTON, 'First Applications of a new Laryngograph', *Medical and Biological Illustration*, 21 p172 (1971)
[2] G LINDSDEY, P DAVIES & A J FOURCIN, 'Laryngeal Coarticulation Effects in English VCV Sequences', *Proceedings of the IEE Conference on Speech Input and Output* , p99 (1986)
[3] D H KLATT & L C KLATT, 'Analysis, Synthesis and Perception of Voice Quality Variations among Female and Male Talkers', *Journal of the Acoustical Society of America*, 87 p820 (1990)
[4] J N HOLMES, 'The Influence of Glottal Waveform on the Naturalness of Speech from a Parallel Formant Synthesiser ', *IEEE Transactions on Audio and Electroacoustics* , AU21 3 p298 (1973)
[5] G FANT, 'Acoustic Theory of Speech Production', Mouton, The Hague, The Netherlands (1960)
[6] D M HOWARD, D M BROOKES & D S F CHAN, 'Dynamic Excitation Control in Parallel Formant Speech Synthesis', *Proceedings of the 7th FASE Symposium* , Edinburgh 3 p1123 (1988)
[7] G FANT, 'Glottal Source and Excitation Analysis', *Speech Transmission Laboratories: Quarterly Progress and Status Report* , 1 Royal Institute of Technology, Stockholm p85 (1979)
[8] P DAVIES, G LINDSEY, H FULLER & A FOURCIN, 'Variation in Glottal Open and Closed Phase for Speakers of English', *Proceedings of the Institute of Acoustics*, 8 p539 (1986)
[9] M A HUCKVALE, D M BROOKES, L T DWORKIN, M E JOHNSON, D J PEARCE & L WHITAKER, 'The SPAR Speech Filing System', *European Conference on Speech Communication and Technology*, Edinburgh p305 (1987)
[10] D M HOWARD, G A LINDSEY & B ALLEN, 'Toward the Quantification of Vocal Efficiency', *Journal of Voice*, 4 p205 (1990)
[11] A R JONCKHEERE & G H BOWER, 'Non-parametric Trend Tests for Learning Data', *The British Journal of Mathematical and Statistical Psychology*, 20 p163 (1967)
[12] G FANT, J LILJENCRANTS & Q G LIN, 'A Four Parameter Model of Glottal Flow', *Speech Transmission Laboratories: Quarterly Progress and Status Report*, 4 Royal Institute of Technology, Stockholm p1 (1985)
[13] J B PICKERING, 'Effects of Voice Type and Quality on the Intelligibility of a Text-to-Speech System', *European Conference on Speech Communication and Technology*, Paris p637 (1989)