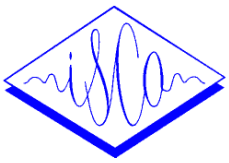

2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

DYNAMIC VOICE SOURCE CHANGES IN NATURAL AND SYNTHETIC SPEECH

Sarah K Palmer and Jill House

Department of Phonetics and Linguistics, University College London,
4, Stephenson Way, London NW1 2HE, UK.

ABSTRACT

This paper describes a series of perceptual tests which demonstrate that listeners are capable of perceiving the systematic cycle-by-cycle changes which occur in the voice source waveform of natural speech as a result of laryngeal coarticulation. Having established that these dynamic changes are perceptible in natural speech we attempt to replicate the effect in synthesis using the KLSYN88 software speech synthesiser. This confirms that our synthesis strategy is appropriate and highlights the most important features of the glottal flow waveform needed to model the changes found in natural excitation. We then turn to detailed analysis of the anticipatory and perseverative coarticulation effects on vowels in the context of British English alveolar obstruents for male and female speakers by means of inverse filtering. The findings demonstrate predictable trends in the voice source parameters associated with different phonetic and allophonic variations and are allowing us to develop rules for the voiced excitation in these phonetic environments for high quality text-to-speech synthesis.

10.21437/ICSLP.1992-38

1. INTRODUCTION

Recent years have witnessed increased research into more detailed modelling of the voice source in an attempt to improve the naturalness of formant speech synthesis. Varying the voice source waveform appropriately has been shown to enhance the naturalness of synthesised female speech [1] and has allowed different voice qualities to be successfully imitated in synthesis [2]. Research in Sweden has further shown that, regardless of an individual's voice quality characteristics, there are consistent changes in the time and frequency domains of the voice source associated with the phonetic environment [3]. Our own work has demonstrated similar predictable trends in the glottal flow waveform for laryngeal coarticulation in British English [4]. Whilst the inclusion of generalised rules for these voice source changes has been reported to improve the naturalness of speech produced by synthesis-by-rule systems, our own work on incorporating detailed cycle-by-cycle changes has yielded limited results [5]. This has been attributed in part to the general quality of synthesis. We believe that it is only as the overall level of naturalness of synthetic speech improves that these dynamic voice source changes will become increasingly important for naturalness. It is possible that naturalness testing, in which synthetic tokens are compared directly with the original digitised speech, is an inappropriate method of studying the importance of dynamic excitation changes at present. Therefore, we have addressed the question of whether listeners are able to perceive these subtle changes in natural quality speech. If we can establish that these changes are perceptible in natural tokens, further detailed analysis of the voice source waveform will be justified in order to develop generalised rules for text-to-speech synthesis, even though demonstration of the dynamic voice source rules' contribution to the naturalness of synthetic speech may remain limited.

2. PERCEPTUAL TESTING OF NATURAL SPEECH

2.1 Introduction

A series of perceptual tests was developed to address the following questions:

1. Are listeners able to predict the following consonant on the basis of the preceding vowel alone?
2. If so, which characteristics of the vowel are the listeners using to make this discrimination?
3. If they are using the voice source as a cue, is it the cycle-by-cycle changes in the source or the average waveshape for the vowel as a whole that is important?
4. Is it possible to incorporate the analysed excitation changes into synthesis and replicate the discrimination effects found in natural speech?

2.2 Test 1

2.2.1 Method: The material used for the perceptual tests was based on the tokens recorded from our earlier work [4] which consisted of the intervocalic glottal fricative [h] and glottal stop [ʔ] in the context of the [ɑ] vowel. These stimuli were chosen due to the clear and contradictory trends in the glottal flow parameters. The sequences were spoken by two males and two females each with three different intonation contours (falling, rising and level) and with pre-consonantal or post-consonantal stress (eg: [ˈɑhɑ] and [ɑˈhɑ]). Analysis revealed a stronger anticipatory than perseverative coarticulation (a finding also noted by [3]). Therefore the initial vowel was studied from the two contexts [ɑhɑ] and [ɑʔɑ] for all the utterances recorded by the four speakers. The simultaneous recordings of the speech pressure waveforms and laryngographic waveforms (Lx) [6] were time aligned in order to correct for the delay for sound to travel from the glottis to the microphone. The positive peaks in the differential of Lx were used to determine the points of excitation. The speech pressure waveforms were then cut at the final excitation marker before the offset of voicing prior to the glottal fricative or glottal stop. The edited vowels were randomised, digitally recorded and presented binaurally via headphones to 15 listeners in a sound-proof environment. The listeners were asked to indicate which of the two consonants the vowel was most likely to precede. After a number of familiarisation examples there was a ten second pause and the test stimuli were presented, separated by three second intervals.

2.2.2 Results: A full logistic regression model was fitted to the data and reveals significant discrimination between the vowels from the two different consonant contexts for all four speakers ($\alpha = 0.05$). Table 1 shows the Chi-Squared values for the effects of consonant

environment (df=1), intonation contour (df=2) and stress (df=1) on the vowel discrimination for the four speakers.

Table 1: The χ^2 values and the associated probability (p) for the influence of consonant environment, intonation and stress respectively on the number of [h] responses.

Speaker	Environment	Intonation	Stress
F1	14.49 (p<0.001)	1.274 (p>0.25)	11.04 (p<0.001)
F2	14.81 (p<0.001)	7.36 (p<0.05)	1.090 (p>0.25)
M1	7.876 (p<0.01)	8.148 (p<0.025)	7.161 (p<0.01)
M2	25.85 (p<0.001)	0.636 (p>0.25)	0.388 (p>0.25)

There is a significant influence of stress on the discrimination of vowels spoken by speaker F1. Examination of the data suggests that for this speaker stressed vowels are more likely to lead to the prediction of a glottal fricative as the phonetic context. Speaker M1 shows a similar but weaker effect for stress. In addition there is a small but significant intonation contour effect in which it appears that vowels spoken with a level intonation contour are more likely to be perceived as coming from the glottal stop environment. Results from speaker F2 also reflect this contour effect.

2.2.3 Discussion: The test demonstrates that listeners are capable of discriminating between natural vowels differing in the phonetic environment from which they were cut. A number of factors could be contributing to this discrimination. Firstly, it is likely that aspiration noise in the vowel preceding the glottal fricative is cueing the discrimination. Secondly, the dynamic changes in the voice source waveform, and as a consequence the interaction phenomena (eg: increased formant damping associated with increased vocal fold abduction) may be serving to aid the discrimination. In addition the vowel length may have an influence. In order to investigate the criteria listeners are using in more detail further tests were developed.

2.3 Test 2

2.3.1 Method: In the second test the initial vowels from the two consonant environments spoken by the female speaker F2 with rising intonation and final stress were used (eg: [aʰha] and [aˀa]. A series of test stimuli were prepared by editing the initial vowel on a cycle-by-cycle basis from the final point of excitation by between 0-12 cycles back to the steady state portion of the vowel. The test stimuli were then randomly recorded three times onto digital tape. Ten listeners, all of whom had completed the first perceptual test, carried out the test in which they were asked to follow the same instructions as test 1.

2.3.2 Results: Statistical analysis shows that there is a significant effect of both the consonant context (χ^2=57.42, df=2, p<0.001) and the vowel length (χ^2=20.75, df=2, p<0.05). There is a significant decrease in the number of [h] responses for fricative stimuli with increasing number of cycles cut back. A much smaller, but still significant, length effect is found for the glottal stop stimuli. The number of cycles cut back at which discrimination between the vowels becomes insignificant is 8.

2.3.3. Discussion: The results demonstrate that, whilst vowel length has a significant effect on discrimination between the two vowels, it is the consonantal environment from which the vowels are taken that remains the most influential. It seems that, as the steady state portion of the vowel is reached, and therefore aspiration noise and dynamic variations in the shape of the voice source removed, discrimination is reduced. Studying the source parameters reveals that it is at the 8th cycle prior to voicing offset that the open quotients (OQ) become equal, (the time from the point of opening of the glottal flow waveform to the following point of excitation, expressed as a percentage of the fundamental period). This leads to the hypothesis that it is the dynamic changes in the excitation that may be cueing the discrimination.

2.4 Test 3

2.4.1 Method: In the third test four vowels were prepared: two from each consonant environment as used in test 2, but with both pre-consonantal and post-consonantal stress). The four vowels were synthesised using LPC resynthesis. In each synthesis the formant frequencies and bandwidths were identical, but the excitation varied in one of three ways. In stimulus (A) the excitation was the raw inverse filtered waveform incorporating dynamic voice source changes and aspiration noise. Stimulus (B) used the dynamically modelled inverse filtered waveform. This was obtained by fitting the L-F four parameter model of the first derivative of the glottal flow waveform [7] to the raw inverse filtered output. This maintains the cycle-by-cycle variations in the voiced excitation but excludes the influence of aspiration noise. Stimulus (C) consisted of a non-dynamic excitation waveshape based on mean source parameters taken from the vowel midpoint, but still maintaining the fundamental frequency and amplitude information of the other excitations. The stimuli were randomly recorded four times on to digital tape and presented to eight of the listeners who took part in the previous tests. The instructions remained the same as before.

2.3 Results

2.3.3 Test 3: Table 2 shows the percentage of [h] responses given by the eight listeners for test 3. Statistical analysis reveals significant discrimination between stimuli synthesised with excitation A and B (χ^2= 41.733 and 30.508 respectively, df=1, p<0.001) but insignificant discrimination for those vowels resynthesised with excitation C (χ^2=3.502, df=1, p> 0.05).

Table 2: Percentage of [h] responses given for vowel stimuli resynthesised with various voice sources.

excitation A [h]	87.5	excitation A [ˀ]	29.7
excitation B [h]	68.8	excitation B [ˀ]	18.8
excitation C [h]	42.2	excitation C [ˀ]	25.0

2.4.3 Discussion: The results suggest that it is the dynamic changes in the voice source that are cueing the discrimination. The addition of aspiration noise increases the significance of this discrimination, but non-dynamic modelling of the excitation waveshape results in a loss of discrimination. Since the formant frequencies and bandwidths were identical for all three types of resynthesis, listeners must be able to perceive the dynamic changes occurring in the glottal flow waveform over time. These variations are therefore likely to be of importance to both the intelligibility and the naturalness of the speech signal, and it seems appropriate to attempt to model them in synthesis. The use of synthetic speech as opposed to natural speech allows for a more detailed examination of the most important features of the waveshape.

3. PERCEPTUAL TESTING OF SYNTHETIC SPEECH

3.1 Method

In order to focus on the excitation, the formant frequency and bandwidth parameters of the KLSYN88 software speech synthesiser [8] were set to appropriate constant values for an unstressed [ɑ] vowel spoken by speaker F2 with rising intonation. The use of fixed formant parameters reduces the overall naturalness of the synthesis but allows the results to be attributed purely to excitation differences. Automatic modelling of the raw inverse filtered waveforms [9] led to the specification of appropriate fundamental frequency (F0) and amplitude of voicing (AV) parameters. In this work, F0 relates to the inverse of the fundamental period measured from the points of excitation in the source waveform [10]. The length of the vowel was fixed at 215ms, based on the mean initial vowel length for all the recordings for speaker F2. The effects of the three excitation parameters used to specify the modified LF source (ss=3) in the synthesiser were examined. These are the open quotient, OQ, (see 2.3), the spectral tilt parameter, TL, which is defined as the additional attenuation to the source spectrum at 3kHz on a scale of 0-41, (TL=0 gives no additional tilt to the spectrum), and the speed quotient, SQ, which is the ratio between the open and closed phase expressed as a percentage. Mean values of these parameters for speaker F2 were OQ=50%, TL=14, and SQ=320%. A reference stimulus was synthesised using constant source parameters set to these values. Test stimuli were then created by synthesising the vowel varying the parameter values, either one at a time or in combination, between the reference values and the values shown in table 3.

Table 3: Values of the voice source parameters specified utterance finally either singly or in combination.

	1	2	3	4	5	6
OQ (%)	80	70	60	40	30	20
TL	26	22	20	12	8	4
SQ (%)	170	220	270	370	420	470

These offset values were based on the analysis findings and were designed to cover the range of variation in each parameter measured for the speaker for the initial vowels from the two consonant environments. Therefore, column 1 represents the maximum offset values found in the vowels preceding a glottal fricative. Column 6 represents the equivalent offset values for the vowels preceding a glottal stop. In order to study the importance of the duration of these dynamic changes in the voice source the parameters varied gradually from the reference to the offset value, over the final 70, 50, 30 or 10ms of the vowel, with linear interpolation between the specified values. The test stimuli and the reference stimulus were randomised and recorded five times onto digital tape. Seven listeners who had participated in the previous tests were presented with the stimuli in a sound-proof room via headphones. The instructions remained the same as for the previous tests.

3.2 Results

A logistic regression analysis was performed on the data. Dynamic variations in the parameters TL and OQ lead to significant discrimination between the stimuli (χ^2= 219.9 and 94.9 respectively, (df=5), $p<0.001$), with TL being the most powerful parameter. Differences in the duration of the voice source changes for these two parameters are insignificant, as are changes to the parameter SQ. When the parameters are varied in combination there is a weak but significant duration effect. Studying the results it appears that when the parameters are varied only over the last 10ms there is no discrimination between stimuli. 41% of the reference stimuli were labelled as coming from the fricative environment.

3.3 Discussion

The findings from our synthesis work suggest that it is the dynamic variations in OQ and TL that are capable of distinguishing two vowels in synthesis. The parameter SQ has no effect on this distinction, despite the consistent variations in natural speech. Other researchers [11] have found this to be a useful parameter for specifying different voice types, but it may not be necessary to model dynamically for a particular speaker. The aspiration noise associated with breathiness has not been studied at present, although it has been shown that this is an important feature to model [8]. The results lead to more detailed analysis of OQ and TL variations in natural speech in order to gain an insight into the perceptually important dynamic changes associated with coarticulation.

4. ANALYSIS OF BRITISH ENGLISH

4.1 Method

Although our aim is to develop a comprehensive set of dynamic voice source rules for British English only the effects of alveolar obstruents on adjacent vowels will be discussed in this paper. One male and two female British English (RP) speakers recorded five repetitions of each of the nonsense words represented in table 4 below. The tokens are taken from the British English diphone list developed by CSTR [12]. Five repetitions of the steady vowel [ɑ] were also acquired.

Table 4: The British English diphones for alveolar obstruents. [ɑ] represents the vowel in 'card' and [ə], schwa. A quotation mark denotes stress and a full stop marks the syllable boundaries.

ə'tɑ.tət	'tɑ.tə	ə'dɑ.təd	'tɑ.də	ə'zɑ.təz	'tɑsə
ət'ɑ.tət	'tat.hə	əd'ɑ.təd	'tad.hə	ə'sɑ.təs	'tɑzə
ə'stɑ.tət					

The simultaneous speech pressure and Lx waveforms were acquired interactively onto a Masscomp 5600 at a 20kHz sampling rate. The phase-corrected speech and Lx were time aligned and fully automatic inverse filtering performed based on a closed phase LPC analysis. Parameters appropriate for the KLSYN88 synthesiser, described in section 3.1, were automatically derived from the modelled inverse filtered waveform.

4.2 Results: The analysis reveals consistent variations in OQ and TL in the time domain according to the various CV and VC environments. Table 5 shows the mean values of OQ and TL at the vowel offsets and onsets for the different phonetic environments for the three speakers. The first column shows the mean OQ and TL values averaged over the middle twenty cycles of the five isolated vowels.

Figure 1 shows typical changes in OQ associated with the different voiceless plosive environments for the male speaker M3. The female speakers show a similar pattern and TL for all three speakers also follows these trends.

Table 5: Mean values of OQ and TL at the vowel offsets and onsets for the three speakers. The first column shows the mean parameter values taken from the mid-point of the [a] vowel spoken in isolation.

speaker	parameter	mean	aspirated [t]		unaspirated [t]		glottalised [t]		cluster [st]		[s]		[z]		syll initial [d]		syll final [d]	
			off	on	off	on	off	on	off	on	off	on	off	on	off	on	off	on
M3	OQ	47	67	71	69	60	30	71	53	73	64	70	56	72	56	48	38
	TL	16	26	22	26	16	17	33	13	30	18	31	23	31	12	21	14
F3	OQ	55	73	74	69	60	32	77	60	71	64	61	52	65	55	58	40
	TL	15	26	13	23	14	16	18	8	21	12	14	15	20	13	20	7
F4	OQ	55	62	74	66	52	32	71	66	64	71	67	52	65	52	60	41
	TL	13	18	12	13	15	11	19	9	20	12	17	19	17	13	18	10

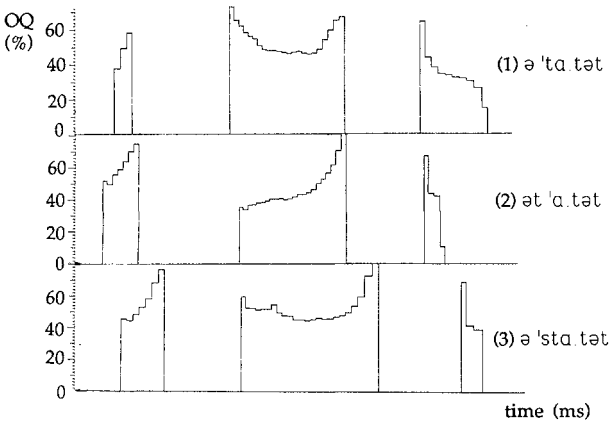

Figure 1. The variations in OQ with time for the nonsense utterances (1) ə'tɑ tə, (2) ət'ɑ tət and (3) ə'stɑ tət for the male speaker.

4.3 Discussion: The analysis findings confirm those of other researchers [3, 13]. It can be seen from figure 1 that OQ rises in anticipation of the initial stressed voiceless plosive reflecting increased abduction of the vocal folds. It remains high for the vowel onset as a result of the relatively high airflow through the glottis following this aspirated stop, before falling to a more average value for the speaker as the folds adduct (1). One of the difficulties of quantifying these dynamic changes for synthesis lies in specifying reference values from which the source changes start in addition to the timing of the variations. Both female speakers exhibit a similar high OQ onset following the aspirated release but the OQ remains high in anticipation of the following syllable initial voiceless stop. Therefore the OQ does not recover its average value reflecting a different degree of coarticulation. When the stop is unaspirated, as in the syllable final position (2) or in a consonant cluster (3), the coarticulation effects are mainly anticipatory with a high OQ offset but a more level OQ onset. Vocal fold adduction has been shown to occur during the stop of the cluster which is reflected by a more level OQ onset [13]. It has also been reported that the airflow at vowel offset preceding the cluster and voiceless fricative [s] is higher than for the plosives possibly due to the need to maintain high oral flow in an effort to sustain turbulent airflow. This finding seems to be confirmed by our data with the average offset value for all three speakers tending to be higher before the cluster and fricative than the stop (see table 5). Speaker F3 consistently glottalises the syllable final stop, both in the word medial and word final position. Speakers M3 and F4 only tend to glottalise word finally. This is shown by a falling OQ at the vowel offset, and for speaker F3, a rising OQ at the vowel onset following the word medial stop. The final vowels in figure 1 show this decrease in OQ.

Analysis of the voiced alveolar obstruents reveals a smaller coarticulatory effect with a small but consistent rise in OQ and TL before syllable initial [d] and [z]. This effect is mainly anticipatory with vowel onset close to the average values of OQ for that speaker.

The analysis has revealed consistent voice source trends for coarticulation from which we are deriving generalised speaker-dependent rules for more natural sounding synthesis. The effects of these rules are currently being evaluated.

5. ACKNOWLEDGEMENTS

The authors would like to thank both the speakers and listeners who kindly participated in the recordings and the tests. This work was funded by SERC grant no GR/F/30642

6. REFERENCES

[1] I. Karlsson. "Voice source dynamics for female speakers," *Proc of ICSLP*, vol 1, pp. 69-72, 1990.

[2] C. Gobl. "Preliminary studies of acoustic voice quality correlates," *STL: QPSR* , Stockholm, vol 4, pp. 9-22, 1989.

[3] C. Gobl, and A. Ní Chasaide. "The effects of adjacent voiced/voiceless consonants on the vowel voice source: a cross language study," *STL: QPSR* , Stockholm, vol 2-3, pp. 23-59, 1988.

[4] S. K. Palmer and D. M. Howard. "Dynamic voice source synthesis," *Proc of 12th ICPhS*," vol 5, pp. 442-445, 1991.

[5] S. K. Palmer, B. Allen, D. M. Howard, G. Lindsey, and J. House. "Analysis, synthesis and perception of laryngeal coarticulation," *Proc of IOA* , vol 12, part 10, pp. 17-24, 1990.

[6] A. J. Fourcin, and E. M. R. Abberton. "First Applications of a new Laryngograph", *Med and Biol Illust* , vol 21, pp. 172-181, 1971.

[7] G. Fant, J. Liljencrants, and Q. G. Lin. "A Four parameter model of glottal flow," *STL: QPSR* , Stockholm, vol 4, pp. 1-13, 1985.

[8] D. H. Klatt, and L. C. Klatt. "Analysis, synthesis and perception of voice quality variations among female and male talkers," *JASA* , 87(2), pp. 820-857, 1990.

[9] D. S. F. Chan, and D. M. Brookes. "Variability of excitation parameters derived from robust closed phase inverse filtering," *Eurospeech* , Paris, pp.199-202, 1989.

[10] S. K. Palmer. "Measurement of the fundamental period in dynamic voice source models for speech synthesis," *Speech, Hearing and Language: work in progress* , UCL, England, vol 6, pp. 169-177, 1992.

[11] C. K. Lee, and D. G. Childers. "Some acoustical, perceptual, and physiological aspects of vocal quality," in *Vocal Fold Physiology* , Ed. J. Gauffin, and B. Hammarberg, Whurr, London, 1992.

[12] S. D. Isard, and D. A. Miller. "Diphone Synthesis Techniques," *Int Conf on Speech Input/Output Techniques and Applications* , IEE no. 258, pp. 77-82, 1986.

[13] A. Löfqvist, and R. S. McGowan. "Voice source variations in running speech," in *Vocal Fold Physiology* , Ed. J. Gauffin, and B. Hammarberg, Whurr, London, 1992.