Proceedings of the Institute of Acoustics

THE DEVELOPMENT OF DYNAMIC VOICE SOURCE RULES FOR SYNTHESIS

S. K. Palmer & J. House

Dept. of Phonetics and Linguistics, University College London, Wolfson House, 4, Stephenson Way, London NW1 2HE.

1. INTRODUCTION

In recent years researchers have focussed on inadequate modelling of the voice source as one of the factors contributing to the lack of naturalness of formant speech synthesis. It has since been shown that varying the voice source waveform appropriately can enhance the naturalness of synthesised female speech [1] in addition to allowing different voice qualities to be successfully imitated in synthesis [2]. Further work on Swedish speech has shown that, regardless of an individual's voice quality characteristics, there are consistent trends in 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, results remain limited [5]. We attribute this, in part, to the general quality of synthesis, and believe that it is only as the overall level of naturalness of synthetic speech improves that these dynamic voice source changes will become perceptually more important. Our previous work establishes that listeners are capable of perceiving the dynamic voice source changes that occur in natural speech as a result of laryngeal coarticulation [6]. More detailed studies of the voice source changes occurring as a result of coarticulation for British English speakers therefore seems justified in order to incorporate these changes into synthesis.

This paper describes detailed inverse filtering analysis of the anticipatory and perseverative coarticulation effects on vowels in the context of British English alveolar obstruents for male and female speakers. The findings demonstrate predictable trends in the voice source parameters associated with the different phonetic and allophonic variations. Incorporation of these voice source changes into the KLSYN88 software speech synthesiser [7] by means of copy synthesis leads to improved naturalness of the output speech for both male and female speakers. The work has implications for the development of dynamic voice source rules for speaker-dependent synthesis-by-rule of British English.

2. ANALYSIS OF BRITISH ENGLISH

2.1 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 adult male and two adult female British English (RP) speakers recorded five repetitions of each of the nonsense words represented in table 1 below. The tokens are taken from the British English diphone list developed by CSTR [8]. Five repetitions of the steady vowel [ɑ] were also acquired for each speaker.

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

Table 1: The recording stimuli acquired for each speaker. [ɑ] represents the vowel in 'card' and [ə], schwa. A quotation mark denotes stress and a full stop marks the syllable boundaries.

The simultaneous speech pressure and Laryngographic (Lx) [9] 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 [10].

The two parameters studied in detail were the open quotient, OQ ,and the spectral tilt, TL, which are used to model the voice source in the KLSYN88 synthesiser. OQ is the percentage of each fundamental period during which the glottis is open, with the fundamental period being measured from consecutive points of excitation [11]. TL is related to the corner rounding of the glottal flow waveform after the point of excitation until complete closure and is described as the additional attenuation to the source spectrum at 3kHz. It is expressed as a value between 0 and 41 such that when TL = 0 there is no corner rounding and no additional spectral tilt, but when TL = 20 there will be an additional 20dB attenuation of the frequency components at 3kHz in the source spectrum. These two parameters can be derived automatically from the inverse filtered waveform, which is modelled using the L-F model of the first time derivative of the glottal flow waveform [12].

2.2 Results

The analysis reveals consistent variations in OQ and TL in the time domain according to the various CV and VC environments. Table 2 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.

speaker	parameter	mean	aspirated ['t]		unaspirated [t]		glottalised [t]	cluster [st]		[s]		[z]		syll initial [d]		syll final [d]	
			off	on	off	on	off	off	on	off	on	off	on	off	on	off	on
M1	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
F1	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
F2	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

Table 2: 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 [ɑ] vowel spoken in isolation.

Figure 1 shows the speech pressure waveform, OQ and TL variations over time for the nonsense utterance [ə'tɑ.tət] for the male speaker M1 and the female speaker F1. Figure 2 shows the equivalent plots for the utterance [ət'ɑ.tət].

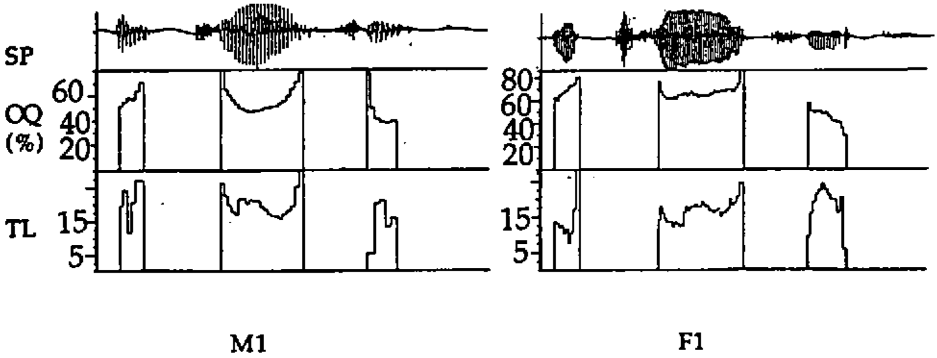


Figure 1. The variations in the speech pressure waveform, OQ and TL parameters with time for the nonsense utterance [ə'tɑ.tət] spoken by the male speaker M1 (left) and the female speaker F1 (right).

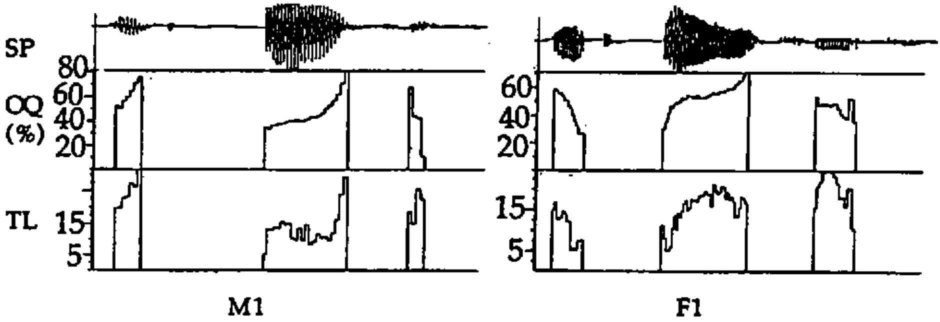


Figure 2. The variations in the speech pressure waveform, OQ and TL parameters with time for the nonsense utterance [ət'ɑ.tət] spoken by the male speaker M1 (left) and the female speaker F1 (right).

2.3 Discussion

The analysis findings confirm those of other researchers [3, 13]. It can be seen from figure 1 that for both speakers 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. For the male speaker the OQ and TL then fall to a more average value for the speaker as the folds adduct before rising again in anticipation of the following syllable initial voiceless plosive. 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. The female speaker F1 shows a similar high OQ onset following the aspirated release, but the OQ remains high in earlier anticipation of the following syllable initial voiceless stop. Therefore the OQ does not recover its average value reflecting a different degree of coarticulation. Both speakers tend to glottalise the utterance final voiceless plosive and this is reflected by the falling values of OQ and TL word finally. Our other female speaker F2 exhibits similar trends to the female speaker F1.

In figure 2 it can be seen that the speakers M1 and F1 show different voice source trends associated with the syllable final voiceless plosive. The male speaker shows a similar high OQ and TL offset as that for the syllable initial voiceless [t] and aspirated [t]. However, unlike the coarticulation associated with the aspirated plosive shown in figure 1 the effects are mainly anticipatory with the onset values of OQ and TL being close to their average values for the speaker.

The female speaker F1 tends to glottalise not only the utterance final but also the syllable final [t] in these tokens, and therefore a falling OQ and TL is seen towards vowel offset and a rising OQ and TL from vowel onset either side of this plosive. Our other female speaker, F2, shows similar voice source trends to the male speaker for this utterance and therefore this syllable final glottalisation is not a sex difference but perhaps a difference in interpretation of syllable boundaries of the recording stimuli. However it gives a clear example of the anticipatory and perseverative coarticulation effects of glottalisation on the voice source parameters.

For clusters vocal fold adduction has been shown to occur during the stop of the cluster [13] which is reflected in our findings by a closer to average OQ onset than seen for the aspirated plosive (see table 2). It has also been reported that the airflow at vowel offset preceding this cluster 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 than the stop.

3. NATURALNESS TESTING

3.1 Introduction

We predicted that these consistent voice source changes for British English speakers, associated with coarticulation, would improve the naturalness of the output speech if modelled in synthesis. We therefore carried out preliminary perceptual testing to study our hypothesis.

3.2 Method

Using the utterance [ə'tɑ.tət] spoken by the male speaker M1 and the female speaker F1, as shown in figure 1 of this paper, a basic copy synthesis for each speaker was prepared with the fundamental frequency and amplitude of voicing varying according to the natural utterance and the formant frequencies and bandwidths smoothed to give as close as possible a representation of the original utterances. Both of the utterances were then synthesised using one of the five excitation types shown below:

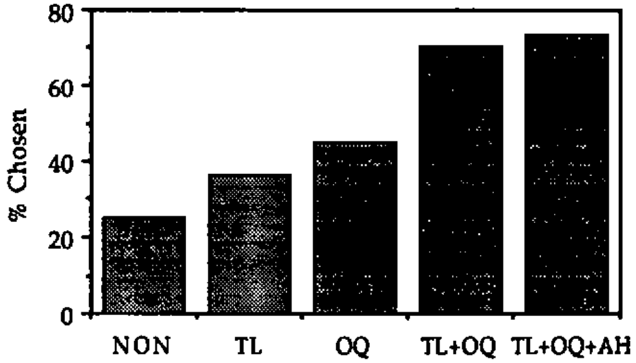
1. Non-dynamic excitation with the OQ and TL fixed to average speaker values throughout the utterance.

2. Dynamically varying OQ with constant TL.
3. Dynamically varying TL with constant OQ.
4. Dynamically varying OQ and TL.
5. Dynamically varying OQ and TL (with added aspiration noise, AH increasing with OQ when OQ > 60%).

The stimuli were grouped into pairs of all possible combinations and these pairs then randomised and recorded onto both channels of a digital audio tape. The stimuli were presented via headphones in a soundproof room to ten listeners who were asked to indicate which stimulus from each pair they perceived as being the more natural.

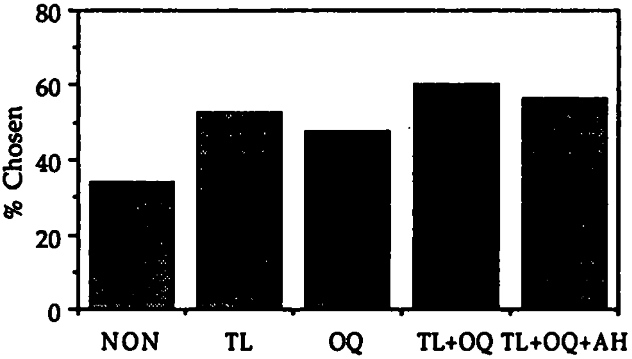
3.3 Results

The results are presented in graphs 1 and 2. Graph 1 shows the results for the male speaker and graph 2 the results for the female speaker. The five excitation types used are shown on the horizontal axis. The vertical axis gives the percentage of the total number of presentations of each stimulus when it was chosen as being the more natural shown on the vertical axis.



Graph 1: Results of naturalness testing for male synthesis

Statistical analysis of the male data reveals that when synthesis is performed using an excitation in which both the OQ and TL are varying dynamically, either with or without additional aspiration noise, the stimuli are perceived as being more natural than the utterance synthesised using an excitation in which either the OQ or the TL are varying dynamically alone. In turn, all the utterances synthesised with either TL or OQ varying dynamically alone or in combination with each other are perceived as being significantly more natural than the stimulus which was synthesised using a non-dynamic excitation. For the female data synthesis using a combination of dynamically varying OQ and TL is perceived as being significantly more natural than those utterances synthesised with either the non-dynamic excitation or excitation in which either just the OQ or the TL is dynamic. Incorporation of dynamic OQ or dynamic TL alone appears to give no significant improvement to the perceived naturalness compared with the non-dynamic synthesis.



Graph 2: Results of naturalness testing for female synthesis

3.3 Discussion

The perceptual test shows that detailed modelling of the naturally occurring changes in OQ and TL leads to increased naturalness of the speech synthesised using the KLSYN88 formant speech synthesiser, The results for the female speaker are not as clear as those for our male speaker. We speculate that the overall quality of female synthesis is still not as high as that for male speech. Therefore the more subtle voice source changes which are perceptible in the higher quality male synthesis, such as those that occur when just one of voice source parameters is modelled dynamically, do not significantly increase the naturalness of the female speech. It is only when a combination of voice source features are modelled dynamically that an effect is seen. It also seems that more accurate modelling of the aspiration noise is needed to enhance the naturalness of both the male and female synthesis. Larger scale testing, using more speakers is planned.

4. DISCUSSION

Analysis has revealed consistent trends in the voice source associated with coarticulation for British English speakers. Incorporation of these dynamic changes in OQ and TL into synthesis on a cycle-by-cycle basis leads to improved naturalness for both the male and female speakers studied. However, it may be that more complex modelling of the voice source is required at present for female speakers due to the lower overall quality of female synthesis in general. Aspiration noise is also likely to be an important factor in the perceived naturalness of synthetic speech and further research into its modelling is planned. This work leads to the development of speaker-dependent rules which attempt to model the magnitude and duration of changes in OQ and TL associated with coarticulation for British English speakers.

5. ACKNOWLEDGEMENTS

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

6. REFERENCES

[1] I. Karlsson, "Voice source dynamics for female speakers,"*ICSLP*, 1, p69 (1990).
[2] C. Gobl, "Preliminary studies of acoustic voice quality correlates," *STL: QPSR* , Stockholm, 4, p9 (1989).
[3] C. Gobl & A. Ní Chasaide, "The effects of adjacent voiced/voiceless consonants on the vowel voice source: a cross language study," *STL: QPSR* , Stockholm, 2-3, p23 (1988).
[4] S. K. Palmer & D. M. Howard, "Dynamic voice source synthesis," *Proc of 12th ICPhS*," 5, p442 (1991).
[5] S. K. Palmer, B. Allen, D. M. Howard, G. Lindsey, & J. House, "Analysis, synthesis and perception of laryngeal coarticulation," *Proc of IOA* , 12, part 10, p17 (1990).
[6] S. K. Palmer & J. House, " Dynamic voice source changes in natural and synthetic speech", *ICSLP* , 1, p129 (1992)
[7] D. H. Klatt & L. C. Klatt, "Analysis, synthesis and perception of voice quality variations among female and male talkers," *JASA* , 87(2), p820 (1990).
[8] S. D. Isard & D. A. Miller, "Diphone Synthesis Techniques," *Int Conf on Speech Input/Output Techniques and Applications* , IEE no. 258, p77 (1986).
[9] A. J. Fourcin & E. M. R. Abberton, "First Applications of a new Laryngograph", *Med and Biol Illust* , 21, p172 (1971).
[10] D. S. F. Chan & D. M. Brookes, "Variability of excitation parameters derived from robust closed phase inverse filtering," *Eurospeech* , Paris, p199 (1989).
[11] S. K. Palmer, "Measurement of the fundamental period in dynamic voice source models for speech synthesis," *SHL work in progress* :UCL, 6, p169 (1992).
[12] G. Fant J. Liljencrants & Q. G. Lin, "A Four parameter model of glottal flow," *STL: QPSR* , Stockholm, 4, p1 (1985).
[13] A. Löfqvist & R. S. McGowan, "Voice source variations in running speech," in *Vocal Fold Physiology* , Ed. J. Gauffin, and B. Hammarberg, Whurr, London, 1992.