ANALYSIS-RESYNTHESIS: MODELLING SELECTED PHONETIC SEGMENTS OF A WOMAN SPEAKER WITH A GENERAL NORTHERN ACCENT

S. P. Whiteside

Department of Psychology (SSU), University of Sheffield

1. INTRODUCTION

Although there has been increased activity in the synthesis of women's speech [2, 5 &9] much more research is needed. This paper represents a contribution towards this. The purpose of this experiment was to analyse, model and synthesize selected non-nasal, all-voiced phonetic segments of a woman speaker with a General Northern accent, using phoneme-specific segments. A cascade formant synthesizer [4] was used in the synthesis because its unbranched all-pole model mirrors the natural processes of speech production for the phonetic segments in question. In addition, for the purposes of analysis-resynthesis the linear predictive methods used in analysis were more compatible with a cascade configured synthesizer, which would therefore copy the acoustic characteristics of non-nasal, all-voiced speech material more successfully.

2. THE SPEECH DATA: A SET OF PHONEME-SPECIFIC SENTENCES

A group of ten phoneme-specific sentences was devised to contain non-nasal, all-voiced speech sounds. This included vowels, dipthongs, liquids and glides. The main aim of constructing the sentences was to focus of the labial-velar glide /w/ and place it within different vowel contexts. All sentences were well formed syntactically but semantically anomolous. The sentences are listed below.

1. Were you early you owl?
2. We were early you owl.
3. Why will you yell?
4. Where will you yell?
5. Wail your yell away.
6. Weigh your yellow ruler.
7.Wheel your wallaroo away.
8. Rule your row warily.
9. Reel your wheel away.
10. You will reel your wool.

It has been shown that the formant patterns of /l r w j/ vary with vowel context [7 &8]. It was therefore decided to analyse, characterise and model the formant patterns of /w/ for a woman speaker, within the different vowel contexts represented in the phoneme-specific sentences.

3. COLLECTION OF DATA AND A PRELIMINARY ACOUSTIC ANALYSIS

The phoneme-specific sentences were read by a woman speaker (cf) with a General Northern accent in a recording studio. Both speech pressure and laryngograph waveforms were recorded using the protocol described in Lindsey et al [6] for the APLAWD database.

The sentences were loaded onto a MASSCOMP 5500 and converted into processable speech files using a SPARbase format [3]. A preliminary acoustic analysis was carried out to obtain fundamental frequency and formant frequency data which could be displayed. The visual displays were used in the acoustic characterisations and phonetic modelling.

To facilitate the recognition and association of formant patterns with their corresponding phonetic 'segments' a reference framework was necessary. This was provided by the Machine Readable Phonetic Alphabet (MRPA) within the SPAR speech filing system These were assigned to the phoneme-specific sentences manually. Auditory and visual judgements of formant patterns and the speech pressure waveforms were used in assigning the labels [11]. Figure 1 illustrates the results of the preliminary acoustic analysis and the annotation procedure for the sentence 'Were you early you owl?'.

4. A CHARACTERISATION OF FORMANT PATTERNS FROM ACOUSTIC ANALYSIS

4.1 /w-vocoid/ syllables from the phoneme-specific sentences: an acoustic analysis

Using the formant track estimates and annotations of all the phoneme-specific sentences each /w-vocoid/ syllable was located and identified. Additionally formant target values for /w/, vowels and diphthongs were estimated together with formant transution rates and their durations. The focus here was on F2. The durations of the /w-vocoid/ syllables were marked from the start of the steady state onset for /w/ until the end of the steady state F2 pattern of the vocoid.

Analysis results

A sample of the acoustic analyses for the /w-vocoid/ syllables in the phoneme-specific sentences are presented below. Phonemic, narrow phonetic transcriptions (IPA) and MRPA representations are provided in this order for each syllable.

i) Were you early you owl? - Stressed, initial position - /wɜː/ - ['wɜ˔]
- [w3:] (MRPA)

Steady state onset		duration	F2 transition dur.	F2transition rate	steady state	+dur
[w](Hz)		ms	ms	Hz/ms	[ɜ˔](Hz)	ms
F1-	400	60	68	11	F1-650	80
F2-	900				F2-1650	
F3-	2400				F3-2900	
F4-	3000				F4-4100	

Total duration of syllable - 208ms

ii) We were early you owl.- stressed initial position. - /wiː/ - ['wiː] - [wi:].

Steady state onset		duration	F2 transition dur.	F2 transition rate	steady state	+dur
[w] (Hz)		ms	ms	Hz/ms	[iː](Hz)	ms
F1-	400	80	68	24.3	F1-500	80
F2-	1000				F2-2650	
F3-	2600				F3-3250	
F4-	3400				F4-4150	

Total duration of syllable - 228ms.

iii) Why will you yell? - Stressed, initial position - /waɪ˞/-['waɪ˞]-[waI]

Steady state onset		duration	F2 transition dur.	F2 transition rate	steady state+dur	
[w] (Hz)ms		ms	Hz/ms	[aɪɹ](Hz)	ms	
F1-	400	70	60	10	F1:800-700	170
F2-	900				F2:1500-1900	
F3-	2500				F3:2500-2900	
F4-	3400				F4:3800-4000	

Total duration of syllable - 300ms.

4.2 Characterising and modelling speech patterns using descriptive rules

The main aim here was to construct rules for formant patterns representing /w-vocoid/ syllables in different phonetic contexts based on the analyses of the limited set of phoneme-specific sentences spoken by cf. The objective was not to mimic the woman speaker but to characterise her formant patterns. This was done through the construction of twenty English phrases which contained /w/ in different phonetic environments. The rules representing the phrases are therefore to be viewed as idealised speech models of observed acoustic phonetic phenomena.

The constructed phrases: a description

The twenty phrases that were constructed are listed below.

1. We were
2. Were we
3. Why we were
4. Why were we?
5. Where we were
6. Where were we?
7. Where you were
8. Where were you?
9. Why you were
10. Why were you?
11. Why we yell.
12. Why you yell.
13. Where we yell
14. Where you yell.
15. Why we wail.
16. Why you wail.
17. Where we wail.
18. Where you wail.
19. We were well.
20. You were well.

Modelling formant patterns- a sample

The descriptive rules used for the modelling concentrated on the segmental aspects of [w]. This was based on observed patterns and primarily concentrated on F2. Additionally some secondary modelling of the acoustic effects of stress and intonation through modelling duration and varying F0 contours was also considered. The values for the latter were chosen to reflect the trends observed in the speech patterns of cf [11].

The rules used for the construction of the phrase 'We were' are presented here as a sample.
1. We were - /wiː ˈwɜː/ - [ˈwiː ˈwɜː ɹ] - [wiː w3ː]

Modelling formant patterns
Values for the phrase initial syllable [wiː] were based on the values of the stressed initial syllable in **We were early you owl.** and [wɜː ɹ] was based on the values of the stressed initial syllable in **Were you early you owl?.** F2 transition durations for /vocoid-w/ sequences were based on the those of /w-vocoid/ syllables in light of their similar patterns as observed for /wiː w/ in **We were early you**

owl. This was also the approach adopted by Lisker [7].

A) For a duration of 80 ms F1=400Hz, F2=1000Hz, F3 = 2600Hz and F4=3400Hz.This represents [w] in a 'steady state' in an initial position before the close front vowel [i:].

B) Over a duration of 68 ms F1: 400Hz ---> 500Hz, F2:1000Hz ---> 2650Hz,F3: 2600Hz ---> 3250Hz and F4: 3400Hz ---> 4150Hz. This represents the 'transition state' of [w] ---> [i:], where ---> indicates the move from a set of formant frequencies representing a segment to those of another. This will also apply to the remaining descriptive rules.

C) For a duration of 80ms F1= 500Hz, F2=2650Hz, F3=3250Hz and F4=4150Hz. This represents [i:] in a 'steady state' between two occurrences of the glide [w].

D) Over a duration of 68 ms F1: 500Hz ---> 400Hz, F2:2650Hz ---> 900 Hz, F3: 3250Hz ---> 2400Hz and F4:4150Hz ---> 3000Hz. This represents the 'transition state' of [i:] ---> [w] of the syllable [wɜ: ɹ].

E) For a duration of 60 ms F1 = 4000Hz, F2 = 900Hz, F3 = 2400Hz and F4=3000Hz. This represents the 'steady state' of [w] in a medial position between the close front vowel [i:] and the neutral vowel [ɜ: ɹ].

F) Over duration of 68ms F1: 400Hz ---> 650 Hz, F2: 900Hz ---> 1650Hz, F3: 2400Hz ---> 4100Hz and F4: 3000Hz ---> 4100Hz. This represents the 'transition state' of [w] ---> [ɜ: ɹ].

G) For a duration of 80 ms x 2.5, F1 = 650Hz, F2 = 1650Hz, F3 = 2900Hz and F4 = 4100Hz. This represents the 'steady state' of [ɜ: ɹ] in phrase-final position.

General rule
Relative to its duration in an initial stressed syllable, the vowel of the phrase-final syllable was lengthened by a factor of 2.5. This was an arbitrary value.

F0 contour modelling
Due to its declarative nature a falling F0 contour from 240Hz to 220Hz was modelled. These values are representative of those observed for cf in the phoneme-specific sentences.

Modelling formant bandwidths
Formant bandwidth estimates [11] were used to model the segments in [wi:] and [wɜ: ɹ]. While the use of these values is an abstraction from real speech where bandwidths will vary temporally within a segment, the values in question displayed some of the general acoustic patterns for [w]. Wider bandwidths were found for example for **b1-b4**. This reflects the weakening of formants during the constrictions formed for the production of [w]. This was illustrated by the diffuse appearance of the formant tracks for [w] in the phoneme-specific sentences.

5. SYNTHESIZING THE MODELLED PHRASES

The twenty modelled phrases were sythesized using an implementation of the Klatt sythesizer [4]. The cascade connection of the synthesizer was used given the acoustic phonetic content of the constructed phrases. The interactive nature of the synthesizer allows the provision of acoustic parameters at discrete time intervals with linear interpolation for synthesis. Using this facility acoustic parameters derived from the acoustic analysis were input to the synthesizer. The descriptive rules as exemplified above served as the acoustic framework for each phrase on a temporal basis. In addition to fundamental frequency, formant frequencies and bandwidths several other parameters were considered during synthesis. These included:-

- modelling gradual onsets and offsets of voicing amplitude
- using the 'natural pulse train' [4] which was more suitable for synthesizing a female voice [9]
- modelling a wider glottal pulse (a larger open quotient) which is obeserved typically for women speakers [1] and which has been used to synthesize the speech of a woman speaker [5]
- using a suitable spectral tilt of voicing (an attenuation of 24db in the freqency components above 3kHz relative to an attenuation of 0db) to model the rounded corner at the time of glottal closure.

All the synthetic phrases were analysed and annotated. Figure 2 illustrates the results of this analysis for the phrase Why were you?

6. A FORMAL ASSESSMENT OF THE SYNTHETIC PHRASES

Two repetitions of each synthetic phrase were transcribed orthographically and phonetically by an experienced phonetician. The transcriptions were used to compute intelligibility scores.

6.1 An analysis of the transcriptions: a few selected observations

i) Intelligibility of /w/
Every single occurrence of /w/ was recognised, thus giving an intelligibility score of 100%.

ii) Phrase intelligibility
Out of a total of forty phrases, seventeen were assigned fully correct orthographic transcriptions. This implied a phrase intelligibility score of 42.5%.

From the seventeen correct responses the following phrases were noted; both repetitions of 1.We were, 2. Were we?, 4. Why were we? 10. Why were you?, 15. Why we wail, 19. We were well and 20. You were well and one repetition of 6. Where were we?, 7. Where you were and 8. Where were you?

The remaining twenty three phrases were partially intelligible. Here, either the extra word 'are' was heard as part of a phrase, or words were misidentified. Some of these errors are examined below.

iii) Word intelligibility
Out of a total of 116 words, 93 were recognised. This gave a word intelligibility score of 80.2%. The table below represents a more detailed view of each word, its frequency in the phrases and how often it was recognised.

Word	Frequency	Correct recognitions
we	22	22
were	24	24
why	16	12
where	16	3
you	18	18
yell	8	4
wail	8	6
well	4	4
Total	**116**	**93**

These results suggest that the acoustic patterns representing 'we', 'were', 'you' and 'well' were accurate models. However, the poorer recognition scores of 'why', 'wail', 'where' and 'yell' suggested the contrary.

Word error analysis - a selection:*where' (13 errors)*

Phrase	Response for word
5. Where we were	a) wire
	b) while
6. Where were we	a) why
7. Where you were	b) why are
8. Where were you	a) why
13. Where we yell	a) why are
	b) why are
14. Where you yell	a) why are
	b) why are
17. Where we wail	a) why are
	b) why are
18. Where you wail	a) why are
	b) why are

From the thirteen errors above, nine were cases where 'why are' was heard instead of 'where'. To explain these observations, the modelling descriptions of each of the relevant phrases were examined. These revealed that 'Where you' was modelled identically for phrases 7., 14. and 18. The same applied 'Where we' in phrases 13. and 17. The common characteristics in all these cases were the representations for 'where' as [we:ɹ]-[we:] and the transition durations of 90ms into [j] and [w}. This therefore explains the recurrence of errors for these examples.

The addition of 'are' corresponded to occurrences of the schwa in the narrow phonetic transcriptions provided. It is suspected that this was related to the transition duration of 90ms mentioned above. This may have been too long and therefore had the perceptual effect of introducing the schwa as an extra segment. This phenomenon has been discussed elsewhere [10].

7. CONCLUSIONS

The very high intelligibility score for /w/(100%) suggests that the formant patterns used to represent this segment in different phonetic contexts in the synthetic phrases were acceptable acoustic characterisations of a woman speaker with a General Northern accent. However, the lower word intelligibility score of 80.2% indicates the unsatisfactory nature of the acoustic models for the steady states and formant transitions representing 'why', 'where', 'yell' and 'wail'. These words were often misidentified and therefore had the effect of reducing the phrase intelligibility score to 42.5%. Observations for 'where' showed that if two different /w-vocoid/ syllables with comparable durations have vocoids with similar F1 and F2 frequencies, they will be perceived as having a similar phonetic quality. This is likely to cause misidentifications and was exemplified for 'where', which was often transcribed as 'why'. This suggests the need to analyse more natural speech data containing 'where' in an initial stressed context to improve the intelligibility of modelled synthetic speech.

The observations of the intrusive schwa also had the effect of reducing word and phrase intelligibility. It was suggested above that the transition durations between the syllables representing 'why' and 'you' (100ms); 'where' and 'you' (90ms) and 'where' and 'we' (90ms) were too long. These resulted in slow formant transitions which gave the auditory impression of extra segments. This suggests that further analysis and some remodelling is necessary to speed up the formant transitions and thereby eliminate the auditory perception of 'extra' segments. Therefore indicating again that natural speech data should provide the basis for all acoustic modelling and subsequent speech synthesis.

8. REFERENCES

[1] Holmberg, E. B., Hillman, R.E. & Perkell, J. S. (1987) 'Glottal airflow and pressure measurements for female and male speakers in soft, normal and loud voice', JASA, 84, 511-529.

[2] Holmes, W. J., Holmes, J. N. & Judd, M. W. (1990) 'Extension of the bandwidth of the JSRU synthesizer for high qulaity synthesis of male and female speech', ICASSP, 313-316.

[3] Huckvale, M. A., Brookes, D. M., Dworkin, L. T., Johnson, M. E., Pearce, D. J. & Whitaker, L. (1987) 'The spar speech filing system', European Conference on Speech Technology, Vol.1, 305-308.

[4] Klatt, D. H. (1980) 'Software for a cascade/parallel formant synthesizer', JASA, 67, 971-955.

[5] Klatt, D. H. (1987b) 'Review of text-to-speech conversion for English', JASA, 82(3), 737-793.

[6] Lindsey, G., Breen, A. & Nevard, S. (1987) 'Spar's archivable actual-word databases'. UCL Report on SPAR Project.

[7] Lisker, L. (1957) 'Minimal cues for separating /w, r, l, y/ in intervocalic position', Word,13(2), 256-267.

[8] O'Connor, J. D., Gerstmann, L. J., Liberman, A. M, Delattre, P. C. & Cooper, F. S. (1957) 'Acoustic cues for the perception of initial /w, y, r, l/ in English', Word, 13, 24-43.

[9] Pickering, J. B. (1988) ' Glottal pulse shapes, naturalness, and the synthesis of female speech', 7th FASE Symposium, 1107-1114.

[10] Scully, C. (1973) 'The problem of unstressed vowels and their coarticulation within consonantal clusters in English', JIPA, (3) No.1, 4-9.

[11] Whiteside, S. P. (1991) Towards improved synthesis of women's speech: British General Northern accent, Unpublished PhD. Thesis, University of Leeds.

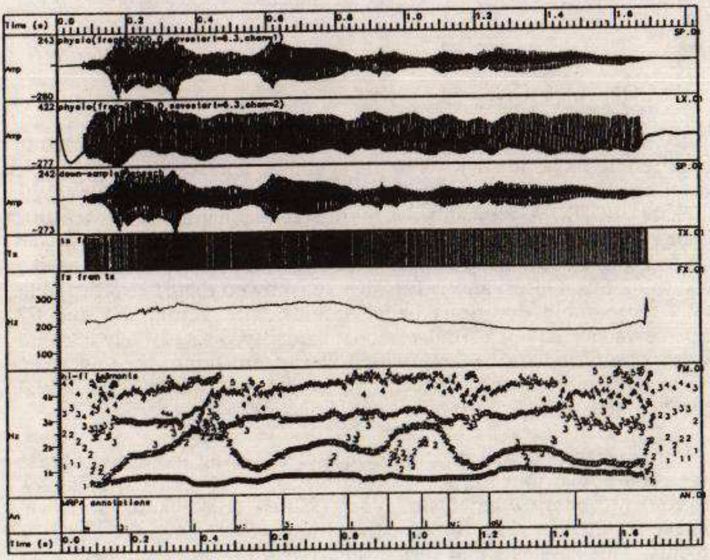

Figure 1 Were you early you owl?

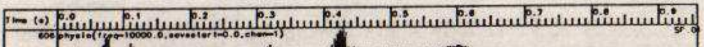

Figure 2. Why were you ? (synthetic phrase)