6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

PERCEPTION OF OVERLAPPING SYLLABLES

William A. Ainsworth
MacKay Institute of Communication and Neuroscience
School of Life Sciences, Keele University
Keele, Staffordshire ST5 5BG, United Kingdom
w.a.ainsworth@cns.keele.ac.uk

ABSTRACT

Some experiments have been performed in which the perception of two consonant-vowel syllables, the starting time of one was delayed relative to the other, have been studied. The consonant was /w/ or /j/, the vowel /i/ or /a/, and the fundamental frequency of the syllable 100 or 150 Hz. The delay was varied from 0 ms (simultaneous presentation) to 200 ms (consecutive presentation). As expected, the frequency of hearing both consonants correctly increased as the delay in starting times increased and the amount of overlap decreased. Unexpectantly, however, differences in fundamental frequency appeared to have no effect. With certain degrees of overlap, consonants were perceived which were not present in the physical stimulus.

10.21437/Eurospeech.1999-412

1. INTRODUCTION

The speech of one speaker can normally be understood in the presence of other, equally loud, conversations (Cocktail party problem [1]) yet automatic speech recognisers make many errors in this situation. In order to investigate how the human auditory system copes with this problem a number of investigators have studied the perception of sentences masked by speech [2, 3], and simultaneous vowel sounds [4-9] and syllables [10, 11].

In a real situation, however, it is unlikely that two speech sounds would be produced exactly simultaneously. It is more likely that they would merely overlap in time. It is not fruitful to study overlapping vowels as brief vowels can easily be recognised. It is necessary to employ syllables in which the evolution of the formant tracks of the consonants over a longer period is required for their recognition.

2. STIMULI

The syllables /wa/, /ja/, /wi/ and /ji/ were synthesised by a parallel-formant speech synthesiser of the type described by Klatt [12]. The stimuli consisted of 3-formant sounds with a 100ms segment in which the frequency and amplitude of the formants changed followed by a 100ms segment during which the frequency and amplitude remained constant. In order to produce a /w/-like sound F1 began at 250Hz, F2 at 750Hz and F3 at 1500Hz and changed linearly to the steady-state values of the following vowel. The amplitudes of all formants began at 0 and changed linearly to those of the following vowel. To produce a /j/-like sound F1 again began at 250Hz but F2 began at 2500Hz and F3 at 3500Hz. An /i/-syllable was formed with F1 of the steady segment at 250Hz, F2 at 2500Hz and F3 at 3000Hz, whereas an /a/-syllable was formed with F1 at 900Hz, F2 at 1100Hz and F3 at 2500Hz. The four syllables were easily recognisable in isolation. The sounds were generated by software on a personal computer at a 16kHz sampling rate and output via a 16-bit sound card.

3. EXPERIMENTS

In real speech it is very unlikely that two syllables would start and stop exactly simultaneously. It is much more likely that they would merely overlap in time. It is therefore of interest to determine how readily syllables are identified when one leads or lags the other.

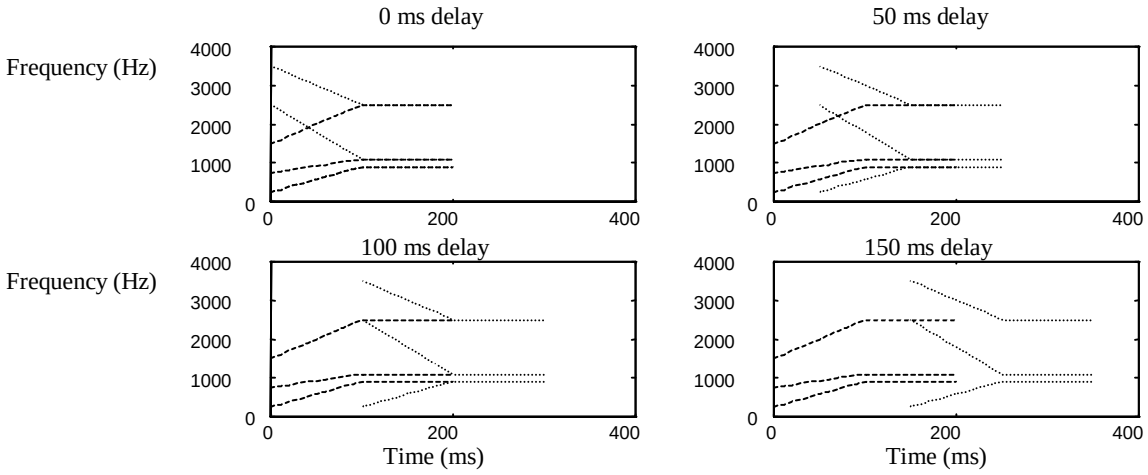

Figure 1. Formant tracks for stimuli consisting of /wa/ (dashed) follwed by /ja/ (dotted) with a delay of 0, 50, 100 and 150ms.

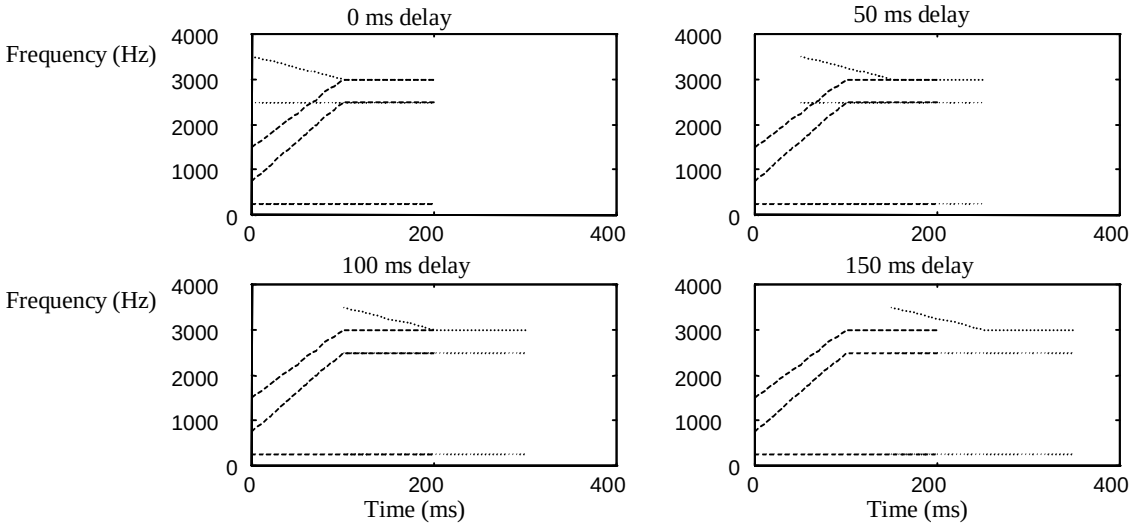

Figure 2. Formant tracks for stimuli consisting of /wi/ (dashed) follwed by /ji/ (dotted) with a delay of 0, 50, 100 and 150ms.

In the experiments syllables having the same vowel (/i/ or /a/) and either the same or different fundamentals (100 or 150Hz) were combined with equal amplitudes and with lags of 0, 50, 100, 150 or 200ms. The stimuli thus consisted of /wV/ followed by /wV/, /jV/ followed by /jV/, /wV/ followed by /jV/ or /jV/ followed by /wV/ where /V/ is /a/ or /i/. The whole of the second syllable was delayed, not truncated at the end of the first syllable. The formant tracks for the stimuli consisting of /wa/ followed by /ja/ with a delay of 0, 50, 100 and 150ms are shown in Figure 1 and those for /wi/ followed by /ji/ in Figure 2.

The listeners were asked whether the consonants they heard were /w/, /j/, /ww/, /jj/, /wj/ or /jw/, and to press an appropriately labelled key. Five listeners took part in these experiments.

4. RESULTS

The listeners mainly heard one syllable when the syllables were coincident. They also only heard one syllable with delays of 50ms. This might have been expected as the auditory system is insensitive to short echoes. (50ms corresponds to a sound path of about 15m.) When the delay was 200ms the syllables were sequential so two syllables were heard. This was also mostly the case with delays of 150ms. At 100ms where the transitions of the lagging syllable were coincident with the vowel of the leading syllable the percept heard depended upon the acoustic structure of stimulus.

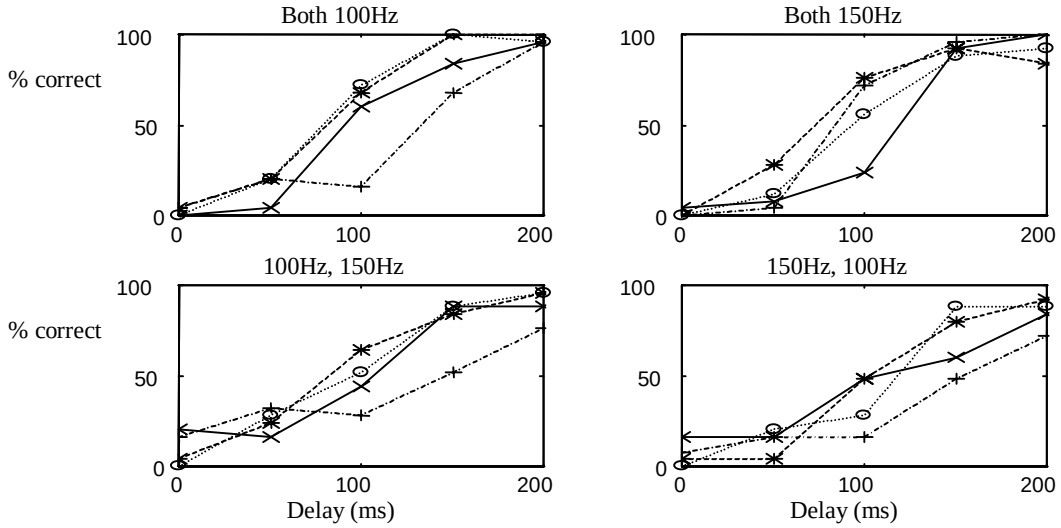

Figure 3. Percentage of /a/ syllable pairs heard correctly as a function of onset delay of the second syllable for both syllables at 100Hz, both at 150Hz, the first at 100Hz and the second at 150Hz and the first at 150Hz and the second at 100Hz. A solid line is used for /wawa/, a dotted line for /waja/, a dashed line for /jawa/ and a dash-dot line for /jaja/.

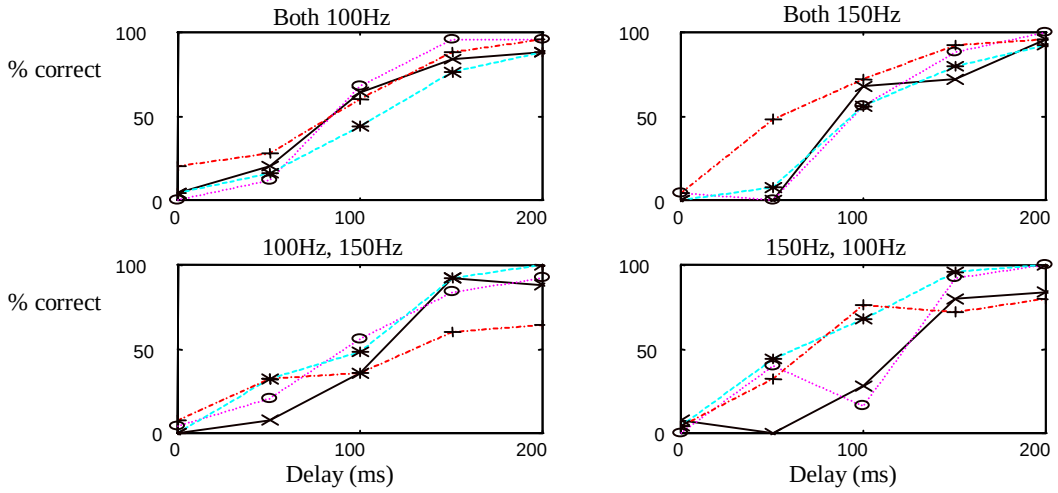

Figure 4. Percentage of /i/ syllable pairs heard correctly as a function of onset delay of the second syllable. A solid line is used for /wiwi/, a dotted line for /wiji/, a dashed line for /jiwi/ and a dash-dot line for /jiji/.

The work of Bregman [13] suggests that short delays in onset might result in both syllables being identified correctly more easily when the fundamentals of two syllables differ than when the fundamentals are the same. However this appears not to be the case. Figure 3 shows the percentage correct identification of both consonants for /a/ syllables (i.e. the key labelled /wj/ was pressed in response to a stimulus consisting of a /wa/ plus a delayed /ja/, etc.) and Figure 4 shows the same data for the /i/ syllables. ANOVA tests showed although there was a significant effect of delay the effect of fundamental frequency was not significant .

This being the case the data for all pitch conditions were averaged and the consonants heard plotted against delay for the /a/ syllables (Figure 5) and the /i/ syllables (Figure 6). For /wawa/ the single syllable /wa/ was heard for delays less than about 80ms, after which a double /wa/ was heard. For /jaja/ a single /ja/ was heard for delays of less than 80ms and a double /ja/ was heard for delays of greater than 120ms. Between these two values /jawa/ was heard about half of the time although no /wa/ was present in the stimulus. For /waja/ a /ja/ was heard for delays of less than 50ms and /waja/ thereafter. It had been found earlier that /ja/ dominated /wa/ no matter whether the fundamentals were the same or different [10]. For /jawa/ a single /ja/ was heard for delays of less than 80ms and /jawa/ was heard for longer delays.

For the /i/ syllables a similar but complementary pattern emerged (Figure 6). For /wiwi/ a single /wi/ was heard for delays of less than 80ms and a double /wi/ for longer

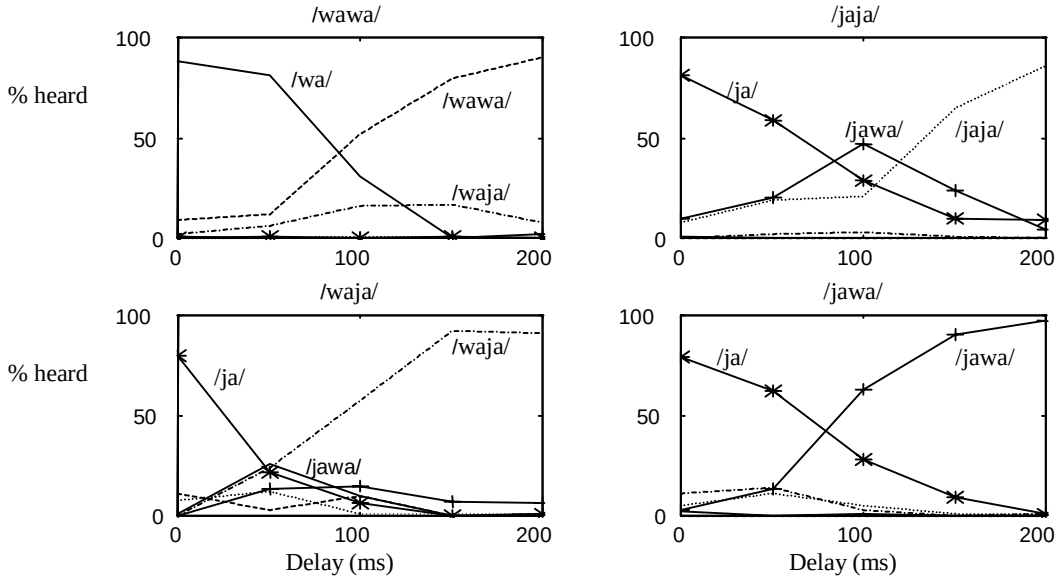

Figure 5. Percentage of syllables heard as a function of onset delay of the second syllable for /wawa/, /jaja/, /waja/ and /jawa/ averaged over all fundamental frequency conditions for /wa/ (_), /wawa/ (--), /waja/ (-.), /ja/ (-*-), /jaja/ (..) and /jawa/ (-+-).

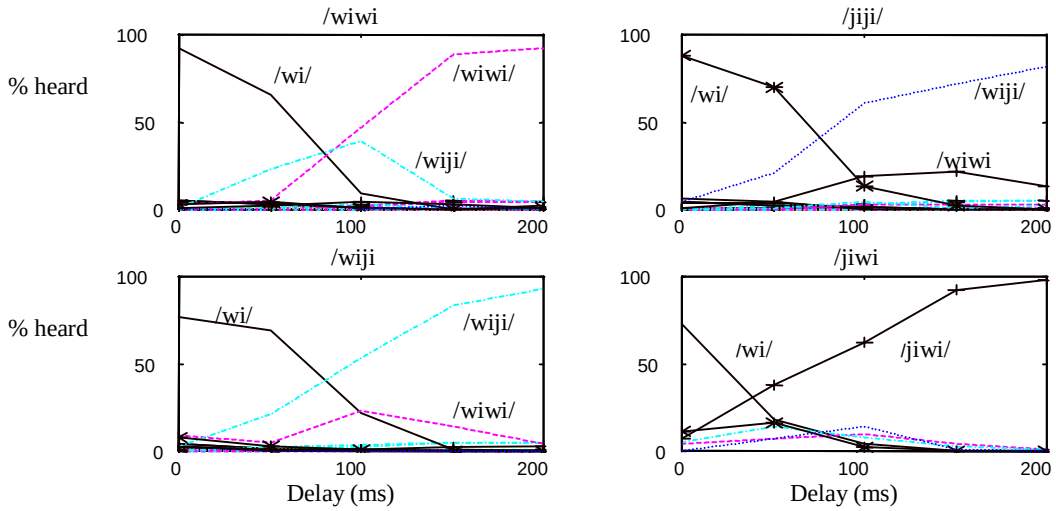

Figure 6. Percentage of syllables heard as a function of onset delay of the second syllable for /wiwi/, /jiji/, /wiji/ and /jiwi/ averaged over all fundamental frequency conditions for /wi/ (_), /wiwi/ (--), /wiji/ (-.), /ji/ (-*-), /jiji/ (..) and /jiwi/ (-+-).

delays. There is a suggestion that /wiji/ was sometimes heard between 80 and 100ms. For /jiji/ a /ji/ was heard up to 75ms and /jiji/ thereafter. Similarly for /wiji/ /wi/ was heard up to 80ms and /wiji/ for longer delays. For /jiwi/ /wi/ was heard for short delays (less than about 40ms) and then /jiwi/. Again this is consistent with the previous finding that /wi/ dominates /ji/ [10].

5. DISCUSSION

As the delay between the start of one syllable and that of the next is increased the identification of the consonants in both syllables improves. The first consonant could then be subtracted from the signal making the second consonant more identifiable. Assmann [14] found that formant transitions in one syllable enabled a concurrent vowel to be more easily recognised.

Unexpectantly a consonant not present in the signal was sometimes heard. Upward spread of masking might obscure the F2 and F3 transitions of the second consonant whilst the F1 transition signalled the presence of a second consonant.

6. CONCLUSIONS

Delaying a second syllable made the initial consonants in both syllables easier to hear, but F0 differences appeared to have no effect on perception. When the vowel of the first syllable overlapped the consonant of the second syllable an extra consonant, not present in the signal, was sometimes heard.

7. REFERENCES

[1] E.C.Cherry 'Some experiments on the recognition of speech, with one and two ears', J. Acoust. Soc. Am. 25, 975-979, 1953.

[2] J.P.L.Brokx and S.G.Nooteboom, ‘Intonation and the perceptual separation of simultaneous voices’, J. Phon. 10, 23-36, 1982.

[3] J.Bird and C.J.Darwin, ‘Effects of a difference in fundamental frequency in separating two sentences’, Psychophysical and Physiological Advances in Hearing (A.R.Palmer, A.Rees, A.Q.Summerfield and R.Meddis, eds.), Whurr Publishers Ltd., London, 263-269, 1998.

[4] M.T.M.Sheffers, 'Sifting Vowels', Doctoral Dissertation, Groningen University, Netherlands, 1983.

[5] P.F.Assmann and Q.Summerfield, 'Modelling the perception of concurrent vowels: vowels with different fundamental frequencies', J. Acoust. Soc. Am. 88, 680-697, 1990.

[6] J.F.Culling and C.J.Darwin, 'Perceptual separation of simultaneous vowels: Within and across-formant grouping by F0', J. Acoust. Soc. Am. 93, 3454-3467, 1993.

[7] A. de Cheveigné, S.McAdams, J.Laroche and M.Rosenberg, 'Identification of concurrent harmonic and inharmonic vowels: A test of the theory of harmonic cancellation and enhancement', J. Acoust. Soc. Am. 97, 3736-3748, 1995.

[8] F.Berthommier and G.F.Meyer, 'Source separation by a functional model of amplitude demodulation', Proc.Eurospeech'95, 135-138, 1995.

[9] J.D.McKeown, ‘Perception of concurrent vowels: The effect of varying their relative level’, Speech Communication 11, 1-13, 1992.

[10] W.A.Ainsworth and G.F.Meyer, ‘Preliminary experiments on the perception of double semivowels’, Proc. Eurospeech’97, 4, 2115-2118, 1997.

[11] W.A.Ainsworth (1998) ‘Perception of concurrent approximant-vowel syllables’, ICSLP, Sydney, vol. 6, 2287-2290, 1998.

[12] D.H.Klatt, ‘Software for a cascade/parallel formant synthesiser’, J. Acoust. Soc. Am., 67(3), 971-995, 1980.

[13] A.L.Bregman, ‘Auditory Scene Analysis’, MIT Press, Cambridge MA, 1990.

[14] P.F.Assmann, ‘The role of formant transitions in the perception of concurrent vowels’, J. Acoust. Soc. Am. 97, 575-584, 1995.