PERCEPTION OF CONCURRENT APPROXIMANT-VOWEL SYLLABLES

William A. Ainsworth,
Centre for Human and Machine Perception Research,
Department of Communication and Neuroscience,
Keele University, Keele, Staffordshire ST5 5BG, United Kingdom
w.a.ainsworth@cns.keele.ac.uk

ABSTRACT

Some experiments are described which explore the perception of the glides /w/ and /j/ spoken simultaneously. These cannot be spoken in isolation, like vowel sounds, but must be combined with vowels to form syllables. In previous experiments /w/ and /j/ were combined with the vowels /i/ and /a/ to form the four syllables /wi/, /wa, /ji/ and /ja/. It was found that if both the vowels and their pitches differed the consonants could be identified by some of the listeners part of the time. The effect of fundamental frequency on perception has now been explored. Each pair of syllables had different consonants and different vowels but one syllable had a pitch of 100 Hz whilst the other had a pitch of between 100 and 200 Hz. It was found that some syllables were perceived like vowels. The effects one syllable of the pair leading the other have also been systematically explored.

1. INTRODUCTION

It has long been known that it is possible to follow one conversation in the presence of other, equally loud, conversations. Cherry [1] called this the 'cocktail party problem' and demonstrated that two recordings by the same speaker played simultaneously to both ears of a listener are difficult but not impossible to separate. More recently Sheffers [2]) investigated the perception of simultaneous vowel sounds played to both ears and developed a computational auditory model which sought to explain the results. Assman and Summerfield [3], Culling and Darwin [4], de Cheveigné et al. [5] and others have also studied the perception of simultaneous vowels. They found that if the fundamentals of the vowels differed by more than two semitones and they were 200 ms in duration both vowels could be identified.

Bregman [6] has suggested that complex sounds are first segregated into auditory streams which have common features such as pitch. Several models of this process have been proposed based on periodicities in the autocorrelation function of the combined vowel signal. An alternative model has been proposed by Berthommier and Meyer [7]. They suggested that incoming sounds are analysed in a number of frequency-dependent channels by the cochlea and transmitted to higher regions of the auditory system where each channel is further analysed by neural units sensitive to amplitude modulations.

Speech does not consist of a sequence of steady-state vowels. The vowels are interspersed with consonants and it is these consonants which carry the bulk of the spoken message. It is therefore of interest to investigate the perception of simultaneous consonants. Can they be separated on the basis of fundamental frequency alone or are mechanisms employing other features involved? To begin to tackle this question Ainsworth and Meyer [8] have performed some preliminary experiments with the syllables /wi/, /wa/, /ji/ and /ja/. Unlike vowels, most consonants cannot be spoken in isolation. They need to be combined with vowels to produce audible syllables.

In a natural environment it is unlikely that two syllables would occur precisely simultaneously. It is more likely there would simply be some overlap in time. Bregman [6] has suggested that parts of complex sounds which begin or end at the same instant in time would be likely to be grouped together. Further experiments have been carried out to explore this situation.

2. STIMULI

In order to generate truly simultaneous sounds the above syllables were synthesised by a parallel-formant speech synthesiser of the type described by Klatt [9]. Four stimuli corresponding to /wa/, /ja/, /wi/ and /ji/ were synthesised consisting of 3-formant sounds with a 100 ms segment in which the frequency and amplitude of the formants changed followed by a 100 ms segment during which the frequency and amplitude remained constant. In order to produce a /w/-like sound F1 began at 250 Hz, F2 at 750 Hz and F3 at 1500 Hz. To produce a /j/-like sound F1 again began at 250 Hz but F2 began at 2500 Hz and F3 at 3500 Hz. An /i/-syllable was formed with F1 of the steady segment at 250 Hz, F2 at 2500 Hz and F3 at 3000 Hz, whereas an /a/-syllable was formed with F1 at 900 Hz, F2 at 1100 Hz and F3 at 2500 Hz. The formant tracks for all combinations of these syllables are shown in Figure 1. The four syllables were easily recognisable in isolation. The sounds were generated by software on a personal computer and output via a 16-bit sound card.

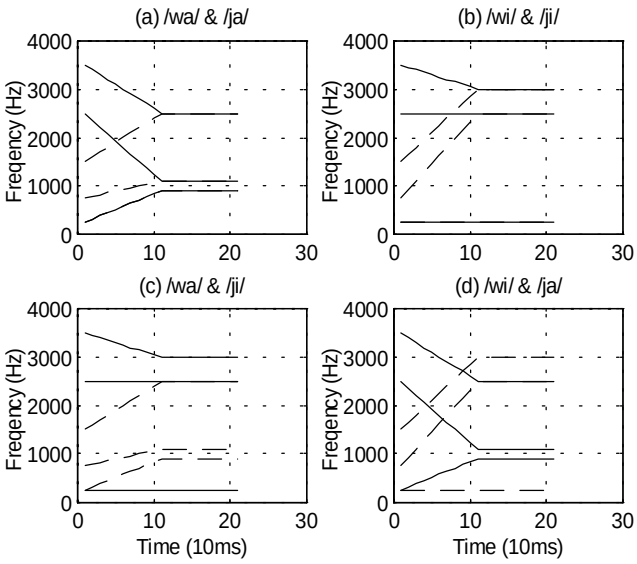

Figure 1. Formant tracks of the stimulus pairs. /wa/ and /wi/ are shown with dashed lines and /ja/ and /ji/ with solid lines.

3. PREVIOUS EXPERIMENTS

Two experiments have been carried out previously [8]. In the first the effects of fundamental frequency and vowel differences were explored and in the second the identification of both the consonants and vowels was tested.

3.1. Effects of vowel and fundamental frequency differences

In each experiment the listeners heard two syllables played simultaneously and were asked whether the consonant they heard at the beginning of the syllable was /w/, /j/ or whether they heard both, and to press W, Y or B appropriately. In each experiment /w/-syllables were added to /j/-syllables such that the ratios of their amplitudes were 1:9, 3:7, 5:5, 7:3 and 9:1 (approximately -20, -7, 0, +7 and +20 dB). Various combinations of vowel (/i/ or /a/) and fundamental frequency (100 or 150 Hz) were employed in the different experiments (Table 1).

Condition	Vowel	F0
1	Same	Same
2	Same	Different
3	Different	Same
4	Different	Different

Table 1: Combinations of vowels and fundamental frequencies.

In the first condition /wa/ was combined with /ja/ and /wi/ with /ji/ with both syllables in the pair at 100 or 150 Hz. As expected, for ratios of 1:9 and 3:7 more syllables beginning with /j/ were heard and for 7:3 and 9:1 more /w/ syllables. With two equally loud syllables /wi/ was heard more often than /ji/ and /ja/ was heard more often than /wa/. Averaged over all the listeners 15.0% of the stimuli were identified as containing both /w/ and /j/. These were fairly independent of the mixture ratio.

In the second condition the same pairs of syllables were employed except one syllable had a fundamental of 100 Hz and the other one of 150 Hz. Again /wi/ and /ja/ were dominant for equal mixtures. Some 16.1% of the stimuli were identified as containing both /w/ and /j/. There was a slight peak in the proportion of dual responses when the two syllables were equally intense.

In the third condition /wi/ was combined with /ja/ and /wa/ with /ji/. Both syllables in the pair had a fundamental frequency of either 100 Hz or 150 Hz. The proportion of stimuli estimated to contain both consonants rose to 23.5%. For /wi/ and /ja/ both at 150 Hz the proportion of stimuli for which both consonants were heard peaked at about 70%.

In the fourth condition the same syllable combinations were employed as in the third condition but one syllable of the pair had a fundamental frequency of 100 Hz and the other one of 150 Hz. This condition increased the proportion of stimuli perceived as containing both consonants to 33.3%. For equal mixtures of syllables more stimuli were identified as consisting of both consonants than a single consonant for all vowel and fundamental frequency combinations.

3.2. Different vowels and fundamental frequencies

In the last two conditions of the previous experiment different vowels were combined as well as different consonants. It is possible that some of the listeners realised this and pressed B when they heard two vowels, although they were asked to do this only when they heard two consonants. It is known from previous work by Assmann and Summerfield [3] that the vowels /i/ and /a/ with fundamentals of 100 Hz and 142 Hz can be readily identified.

In order to investigate this a new experiment was carried out using the same sounds as in the fourth condition of the previous experiment but this time the listeners were asked to press W or J for /w/ or /j/ alone, but A if they heard /wa/ and /ji/ together and I if they heard /wi/ and /ja/. By examining the results it was possible to determine the proportion of syllable combinations correctly recognised. They heard /wa/ and /ji/ correctly 67% of the time and /wi/ and /ja/ correctly 70% of the time. These figures, however, mask the individual variations. They ranged from 88.5% for one listener to 51.3% for another. As chance level is 50% it is unlikely that this latter listener was able to perform the task.

4. EFFECT OF FUNDAMENTAL FREQUENCY

The results of the second condition of the first experiment suggest that it is difficult to segregate /w/ and /j/ syllables with the same vowel even though the fundamentals of the two syllables are different. In this experiment the syllables had fundamental frequencies of 100Hz and 150Hz. These values were chosen so that the fundamental of the second syllable lay midway between the fundamental of the first syllable and its first harmonic. This has the consequence that the second harmonic of the first syllable and the first harmonic of the second syllable are both 300Hz. Also the fifth harmonic of the first syllable and the third harmonic of the second syllable are both 600Hz, etc. This harmonicity of the two syllables may cause them to be grouped into a single perceptual stream and so make recognition of the individual consonants difficult.

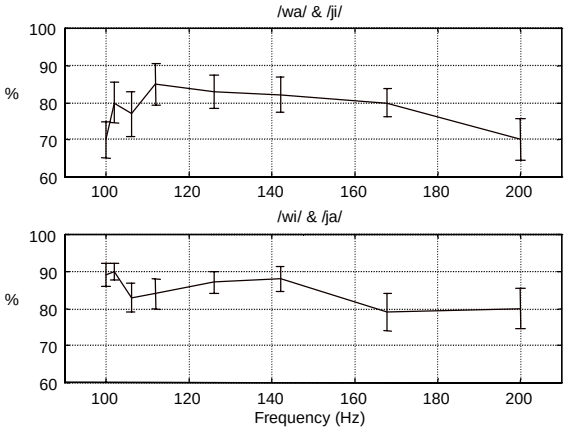

Figure 2. Proportion of pairs of syllables heard correctly as /wa/ & /ji/ (top) and /wi/ & /ja/ (bottom) as a function of fundamental frequency.

In order to investigate this possibility a new experiment was performed in which the pitches of the syllables were not harmonically related. The fundamental frequency of one syllable was always 100Hz and that of the other was 100, 102, 106, 112, 126, 142, 168 or 200Hz. These were the integers nearest to the values used by Assman and Summerfield [3] in their experiments with concurrent vowels. They found a peak in the identification curve at about 142Hz.

There were 32 stimuli consisting of two syllables added together in equal proportions: /ja/ at 100Hz with /wi/ at each of the eight frequencies above, /ji/ at 100Hz with /wa/ at the eight frequencies, /wa/ at 100Hz with /ji/ at the eight frequencies and /wi/ at 100Hz with /ja/ at the eight frequencies. Each listener heard all stimuli twice with a different ordering on each occasion.

Five listeners took part in this experiment. The listeners were asked to identify the syllable(s) they heard and press W if they heard /wi/ or /wa/, to press J if they heard /ji/ or /ja/, to press 1 if they heard /wi/ and /ja/ and to press 0 if they heard /wa/ and /ji/.

Overall the results were similar to those obtained previously by others (e.g. [3]) with pairs of concurrent vowels but the effect of fundamental frequency was less marked. As the difference in fundamental increased the proportion of syllables correctly recognised increased but decreased again as the fundamental of the second vowel approached the first harmonic of the first vowel.

There are two syllable combinations: /wa/ & /ji/ and /wi/ & /ja/. With the syllables /wi/ & /ja/ the formants cross but with /wa/ & /ji/ they do not (Figure 1). If these two conditions are analysed separately interesting results emerge. The top panel of Figure 2 shows that for /wa/ & /ji/ the results are very similar to those obtained with concurrent vowels [3, 7]. In this condition the formants do not cross, as is also the case with isolated vowels. For /wi/ & /ja/ a different pattern emerges (Figure 2, bottom panel). The highest scores were obtained when the fundamentals of the two syllables were near to each other.

5. EFFECT OF DELAY OF ONSET TIME

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.

In the next experiment syllables having the same vowel (/i/ or /a/) and either the same or different fundamentals (100 or 150Hz) were combined with lags of 0, 50, 100, 150 or 200ms. The stimuli thus consisted of /wV/ followed by /wV/, /jV/ then /jV/, /wV/ then /jV/ or /jV/ then /wV/ where /V/ is /a/ or /i/. The listeners were asked whether the consonants they heard were /w/, /j/, /ww/, /jj/, /wj/ or /jw/. Five listeners took part in these experiments.

The listeners mainly heard one syllable when the syllables were coincident, as expected from the first experiment. They also only heard one syllable with a delay of 50ms. This also 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.

The work of Bregman [6] suggests that, when the fundamentals of two syllables do not differ, short delays in onset might result in both syllables being identified. However it appears that this is independent of fundamental frequency. Figure 3 shows the percentage correct identification of both consonants for /a/ syllables. Similar results were found with the /i/ syllables. ANOVA tests showed there was no significant effect of fundamental frequency.

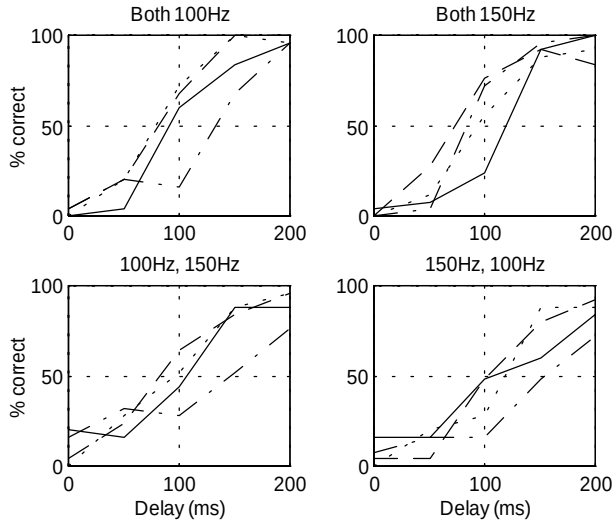

Figure 3. Percentage of /a/ syllable pairs heard correctly as a function of onset delay of the second syllable for both syllables at 100Hz, both at 150Hz, the first at 100Hz and the second at 150Hz and the first at 150Hz and the second at 100Hz. A solid line is used for /wawa/, a dotted line for /waja/,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

This being the case the data for all pitch conditions were averaged and the consonants heard plotted against delay for the /a/ syllables (Figure 4). 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. This is consistent with conditions 1 and 2 of the previous experiment where /ja/ dominated /wa/ no matter whether the fundamentals were the same or different. 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. For /wiwi/ a single /wi/ was heard for delays of less than 80ms and a double /wi/ for longer 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 /wi/ dominating /ji/ as in the previous experiment.

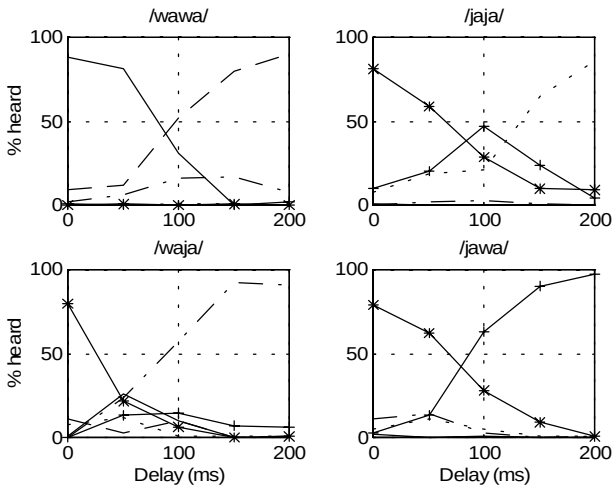

Figure 4. Percentage of syllables heard as a function of onset delay of the second syllable for /wawa/, /jaja/, /waja/ and /jawa/ averaged over all fundamental frequency conditions for /wa/ (_), /wawa/ (--), /waja/ (-.), /ja/ (-*-), /jaja/ (..) and /jawa/ (-+-).

6. DISCUSSION

It appears that when listening to simultaneous consonants, in order to hear both of them, it is not sufficient for the fundamentals of the two syllables to be different as is the case for isolated vowels. It is also necessary for the vowels of the syllables to be different, at least for the restricted cases considered here. It is likely that some sequential analysis of the formant transitions takes place which is facilitated if the formants of the two syllables are moving towards different vowel targets. When the formant tracks converge a single vowel is heard. This makes it unlikely that two syllables have been spoken so only a single consonant is perceived. However when the formants remain apart two vowels may be heard which increases the chance that two syllables have been spoken. In this case two consonants may be heard.

When /wi/ and /ji/ are heard simultaneously /wi/ dominates and when /wa/ and /ja/ are heard simultaneously /ja/ dominates. When the dominant syllables are paired they are easier to recognise when they do not differ in fundamental than the less dominant syllables. This might be because of the greater frequency range of the transitions (Figure 1) or it may have something to do with the crossing formants. With a frame-by-frame analysis crossing formants might be expected to lead to greater confusions but with separate analyses in different frequency regions [10] which take into account the dynamics of the formants the more complex structure may be an advantage. Another possibility is that /wi/ and /ja/ are more dominant because of some top-down process relating to the fact that they are linguistically meaningful. The word “we” is common in English and “ja” is common German. The word “ja”, however, is known to most English speakers and is occasionally used as a substitute for “yes” in some British English dialects. On the other hand the syllable /wa/ is linguistically meaningless and the word “ye” is archaic and seldom used.

7. CONCLUSIONS

Experiments on the perception of concurrent synthesised approximant-vowel syllables suggest that differences in fundamental frequency are effective in segregating the syllables only when the vowels in the syllables are different, although even with different vowels in the two syllables this is not always the case. Perception of syllables with crossing formant transitions appears to be an exception. The present experiments, however, were all conducted with steady fundamental frequencies. Further experimentation is required to examine the effects of changing fundamentals.

8. ACKNOWLEDGEMENTS

The work was supported in part by Grant GR/L05655 from the UK Engineering and Physical Sciences Research Council.

9. REFERENCES

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

2. Sheffers, M.T.M. *Sifting Vowels: Auditory Pitch Analysis and Sound segregation*, Doctoral Dissertation, Groningen University, Netherlands, 1983.

3. Assmann, P.F. and Summerfield, Q. “Modelling the perception of concurrent vowels: vowels with different fundamental frequencies”, *J. Acoust. Soc. Am.* 88, 680-697, 1990.

4. Culling, J.F. and Darwin, C.J.) “Perceptual separation of simultaneous vowels: Within and across-formant grouping by F0”, *J. Acoust. Soc. Am.* 93, 3454-3467, 1993.

5. Cheveigné, A.de, McAdams, S., Laroche, J. and Rosenberg, M. “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.

6. Bregman, A.L. *Auditory Scene Analysis*, MIT Press, Cambridge MA, 1990.

7. Berthommier, F. and Meyer, G.F “Source separation by a functional model of amplitude demodulation”, *Proc.Eurospeech'95*, 135-138, 1995.

8. Ainsworth, W.A. and Meyer, G.F. “Preliminary experiments on the perception of double semivowels”, *Proc. Eurospeech’97*, 4, 2115-2118, 1997.

9. Klatt, D.H. “Software for a cascade/parallel formant synthesiser”, *J. Acoust. Soc. Am.*, 67(3), 971-995, 1980.

10. Greenberg, S. “Understanding speech understanding: towards a unified theory of speech perception”, *Proc. of ESCA Workshop on the Auditory Basis of Speech Perception* (W.A.Ainsworth & S.Greenberg, eds.), Keele University, 1-8, 1996.