5

Zebra finches and Dutch adults exhibit the same cue weighting bias in vowel perception

Verena R. Ohms, Paola Escudero, Karin Lammers, & Carel ten Cate

Vowels in human speech differ from each other in several acoustic features. A major question in speech perception concerns which of these features are critical to distinguish different vowels, i.e. whether features are weighted differently ('acoustic cue weighting'). Human infants for instance, are more sensitive to low frequency components when discriminating vowels, but it is unclear whether adults are too. Also, while animals are known to perceive speech sound contrasts, it is unknown if they exhibit a cue weighting bias, or if this is a uniquely human trait, linked to using speech. We provided zebra finches (*Taeniopygia guttata*) and human adults with the same task of discriminating words that had incorporated vowels which differed and overlapped in several frequency components. We show that they both exhibit a highly similar acoustic cue weighting bias. In contrast to human infants, however, both zebra finches and human adults pay more attention to high frequency components. Our results demonstrate that cue weighting in speech perception is not a uniquely human characteristic and thus need not be closely linked to experience with speech in general or with vowels in particular. We suggest that both humans and zebra finches are born with specific perceptual biases, which at least for humans might shift developmentally, perhaps as a result of their acoustic environment.

Manuscript

Introduction

The evolution of speech and language is still a fiercely debated topic among scientists from various disciplines (Hauser *et al.* 2002; Pinker & Jackendoff 2005; Anderson 2008; Fitch 2010). The original assumption that 'speech is special' (Liberman 1982) and that the mechanisms underlying speech perception are uniquely human (Lieberman 1975) has been challenged over the years by numerous studies indicating that the general ability of speech perception is widely shared with other species including both mammals (Kuhl & Miller 1975; Hienz *et al.* 1996; Eriksson & Villa 2006) and birds (Kluender et al. 1987; Dooling & Brown 1990; Ohms *et al.* 2010a). The categorical perception of speech sounds previously thought to be uniquely human is just as present in other animals (Kuhl & Miller 1975; Kluender *et al.* 1987) as is the capacity for vocal tract normalization (Ohms *et al.* 2010a). However, most research treated speech sounds as unimodal entities and has not considered the fact that multiple acoustic features are involved in their production and perception.

A major unsolved question in speech perception for humans as well as other species regards the relative contribution that those different acoustic features have in the perception of speech sound contrasts. Vowels are characterized by at least two types of formant frequencies. Formants are vocal tract resonances that are not present in most consonants (Titze 2000). The frequency values of formants vary between vowels and are dependent on the position of the tongue in the mouth cavity. For instance, the vowels in the syllables /dIt/ and /dut/ have different first (F1), second (F2) and third formant (F3) frequency values: /u/ has lower F1, F2 and F3 values than /I/ (Fig. 5.1).

Studies with human infants (Lacerda 1993, 1994; Curtin *et al.* 2009) have shown that both Swedish and Canadian-English babies perceive low formant frequencies, i.e. F1 differences, more readily than high formant frequencies, i.e. differences in F2 and F3, when distinguishing syllables that differ only in their vowel sounds. The authors of the last study explain this as a result of Canadian-English having more vowels which differ more in F1 than in F2 values. Thus, 15-month-old infants seem to exhibit a cue weighting bias towards the acoustic feature that is most important in their native language, a finding that is compatible with the fact that infants start to discriminate only the vowels of their native language, and not those of other languages, by their sixth month of age (Polka & Werker 1994). Interestingly, however, infants aged 3 to 12 months whose native language has more vowels that differ in high formant frequencies than English, namely Swedish, are also better at discriminating F1 than F2 differences

(Lacerda 1993, 1994), which suggests a universal human bias towards lower frequencies in vowel perception. However, to date it remains unclear if these biases are strictly linked to speech sound perception and hence form a uniquely human property that might change developmentally as a result of phonetic experience or not. So far, acoustic cue weighting has not been attested in any other species.

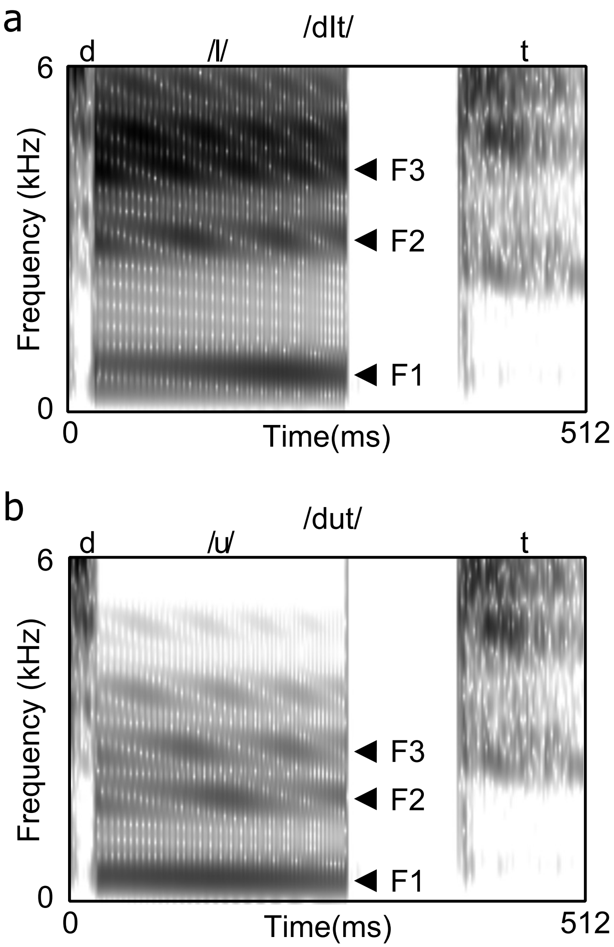

Figure 5.1. Spectrograms of two syllables differing only in their vowels.

This figure shows spectrograms of two synthetic syllables: /dIt/ and /dut/. It is clearly visible that formant frequencies differ between the vowels with lower formant frequencies in /u/ compared to /I/. F1, first formant; F2, second formant; F3, third formant; kHz, kilohertz; ms, milliseconds.

In the current study we used a Go/NoGo operant conditioning paradigm to test acoustic cue weighting in a species assumed to perceive vowel formants in similar ways as humans, namely the zebra finch (*Taeniopygia guttata*) (Ohms *et al.* 2010a; Dooling *et al.* 1995). In their own vocalizations zebra finches show a variety of note types, covering a wide frequency range, which are produced using various articulators (Ohms *et al.* 2010b). Despite similarities in vowel perception, the auditory system of a zebra finch has not been fine-tuned to the perception of human speech and it lacks experience with a particular language. Thus, one can predict that zebra finches utilize high formant frequencies more easily due to an increased sensitivity between approximately 1 and 4 kilohertz (Dooling 2004). Alternatively, existing evidence for a universal formant perception bias towards lower frequencies might transfer to zebra finches because of their human-like perception of vowels. On the other hand it still has to be explored if and how a cue weighting bias in human adults will manifest itself. Although it has been shown that Swedish babies younger than 12 months have the same cue-weighting bias as Canadian-English babies, it has yet to be shown whether extensive experience with Swedish or another language with more F2 and F3 vowel contrasts either makes both cues equally relevant, changes the bias towards higher frequencies, or does not alter the bias in human listeners. Therefore we also tested acoustic cue weighting in vowel perception in Dutch speaking adults using the same stimuli and a highly similar testing procedure as we used for the birds to make the results greatly comparable.

We used four synthetic tokens of each of the vowels /i/, /I/, /u/ and /U/ (Fig. 5.2 and Table A 5.1), which had similar F1 and F2 values to those reported earlier (Curtin *et al.* 2009). Both zebra finches and humans were trained to discriminate two of the four syllables which differed in all formant frequencies following a Go/NoGo paradigm. One syllable was associated to positive feedback, the other to negative feedback (Table 5.1). After subjects had learned to reliably discriminate between the two syllables the remaining two syllables were introduced as probe sounds. Probe sounds were never reinforced and either had the same F1 frequency as the positive stimulus and the same F2 and F3 frequencies as the negative stimulus or the other way around (Fig. 5.2 and Table A 5.1). The responses of birds and humans to the probe sounds allowed us to draw conclusions about how these sounds were perceived by the subjects.

Material and Methods

Stimuli

We used the software Praat (Boersma 2001) version 4.6.09 freely available at www.praat.org to generate four synthetic tokens of the syllables used before (Curtin *et al.* 2009) namely "deet" (/dit/), "dit" (/dIt/), and "doot" (/dut/). We also synthesized the syllable "dut" (/dUt/) to complete the set of Canadian-English high vowels. F1 and F2 values of these tokens are shown in Table A 5.1. In order to compare the use of F1 and F2 differences in vowel perception, the tokens for the contrasts /dit/-/dIt/ and /dut/-/dUt/ differed in their F1 values, while the tokens for the contrasts /dit/-/dut/ and /dIt/-/dUt/ differed in their F2 values. In terms of F1, the fourth token of each syllable, namely dit4, dIt4, and dut4, had values that fell within one standard deviation of those reported earlier (Curtin *et al.* 2009), while those for dUt4 where identical to dIt4 for F1 and to dut4 for F2. Tokens 1-4 where generated in order to examine whether variation in F1 and F2 values would lead to a different pattern in the use of these dimensions. Listeners heard only one set of tokens, e.g. /dit/1, /dIt/1, /dut/1, and /dUt/1. All synthesized vowel tokens were spliced in the middle of the same natural d_t frame, which was taken from one of the naturally produced /dut/ tokens of the study mentioned earlier (Curtin *et al.* 2009). The vowels had the same fundamental frequency (F0) and duration. They had a falling F0 contour which started at 350 Hz at the vowel onset and fell down to 250 Hz at the vowel offset, with both values being similar to those of a natural female voice. The vowels had the same duration, namely 250 ms, in order for listeners to only use vowel formant differences when discriminating the vowels in the stimuli. The vowels also differed in their F3 values because, in English, vowels with low F2 values, namely back vowels, are always produced with a low F3 value, which gives them their characteristic "rounding" feature. Thus, the script that was used to synthesize the vowels computed F3 values following the formula: F3 = F2 + 1000 Hertz for /i/ and /I/ and the formula F3 = F2 + 400 Hertz for /u/ and /U/.

Zebra finch testing

An extensive description of the testing procedure can be found elsewhere (Ohms *et al.* 2010a). Briefly, eight zebra finches were trained in a Go/NoGo operant conditioning chamber to discriminate between two syllables that differed in all formant frequencies from each other whereby every bird got a different set of stimuli (Table 5.1). One of the syllables was associated to positive feedback, the other to negative feedback (Table 5.1).

Each trial was initiated by the birds pecking a report key which resulted in playback of either the positive or the negative stimulus. The birds had to peck a response key after hearing the positive stimulus (S+), e.g. /dit/, in order to get a food reward while ignoring the negative stimulus (S-), e.g. /dUt/. Responding to the negative stimulus caused a 15 seconds time out in which the light in the experimental chamber went out. Playback of the positive and negative stimulus was randomized with no more than three consecutive positive or negative stimulus presentations. After each bird had reliably learned to discriminate between the two syllables the remaining two syllables were introduced as probe sounds in 20% of the trials. Probe sounds were never reinforced and either had the same F1 frequency as the positive stimulus and the same F2 and F3 frequencies as the negative stimulus or the other way around (Fig. 5.2 and Table A 5.1). The responses of the birds to the probe sounds allowed us to draw conclusions about how these sounds were perceived by the birds. All animal procedures were approved by the animal experimentation committee of Leiden University (DEC number 09058).

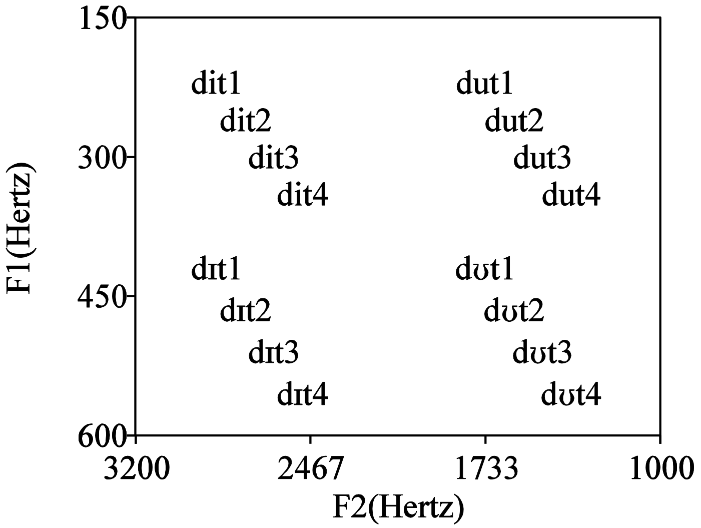

Figure 5.2. Stimuli.

This figure shows a scatter plot of the first (F1) and second (F2) formant frequencies in Hertz of all 4 tokens used per word. /dit1/ and /dut1/ for example have the same F1 but differ in F2, whereas /dit1/ and /dIt1/ have the same F2 but differ in F1. /dit1/ and /dUt1/ neither overlap in F1 nor in F2.

Table 5.1. Testing scheme.

Bird / Group	S+	S-	Probes
729 / 1	/dit/1	/dUt/1	/dIt/1 and /dut/1
728 / 2	/dUt/2	/dit/2	/dIt/2 and /dut/2
750 / 3	/dit/3	/dUt/3	/dIt/3 and /dut/3
763 / 4	/dUt/4	/dit/4	/dIt/4 and /dut/4
734 / 5	/dIt/1	/dut/1	/dit/1 and /dUt/1
731 / 6	/dut/2	/dIt/2	/dit/2 and /dUt/2
758 / 7	/dIt/3	/dut/3	/dit/3 and /dUt/3
741 / 8	/dut/4	/dIt/4	/dit/4 and /dUt/4

This table shows which tokens of which stimuli were presented as either positive or negative stimulus and probes to individual birds and groups of human participants. S+, positive stimulus; S-, negative stimulus.

Human testing

Testing took place in a quiet room using a PC and a custom-written script in the software E-Prime version 2.0. Stimuli were presented via headphones (Sennheiser HD595). Participants learned to discriminate between two of the syllables, following the same Go/NoGo procedure applied to the birds (Table 5.1). Subjects were randomly allocated to the different test groups (1 to 8) with five persons per group and instructed to follow the instructions displayed in Dutch on the computer screen until a note appeared which announced the end of the experiment. Furthermore it was pointed out that during the experiment something might change, but that they were expected to just continue with the procedure. The Go/NoGo paradigm was not explained beforehand so that the human subjects, just like the birds, had to figure out the correct procedure completely by themselves. The experiment started with the screen displaying the instruction: "Press 'Q' to start the trial". After a subject pressed the button 'Q' either the positive or negative stimulus was played back, followed by the instruction: "Press 'P' after the positive stimulus". A two second interval followed in which the subjects had time to press 'P'. Pressing 'P' after the positive stimulus resulted in the presentation of a happy smiley accompanied by a rewarding 'ding' sound. Not pressing 'P' during these two seconds resulted in the presentation of a sad smiley accompanied by a punishing 'attack' sound. After playback of the negative stimulus pressing 'P' resulted in the presentation of the sad smiley accompanied by the 'attack' sound whereas not pressing 'P' resulted in the presentation of the happy smiley and the 'ding' sound. After this cycle had been completed a new cycle started, again with the instruction: "Press 'Q' to the start the trial" until a total of 10 positive and 10 negative stimulus presentations had taken place. The order of stimulus presentations was random with no more than three positive or negative stimulus playbacks in a row. If a subject had at least 14 correct responses within the first 20 trials (70%) he or she automatically continued to the actual testing phase which was announced by the note: "You are entering the actual testing procedure now". If a subject did not reach the 70% correct responses criterion he or she automatically underwent another training round which was indicated by the sentence: "Your correct score is too low. You will enter another training round.". If a subject still did not achieve 70% correct responses in this second training he or she did not continue to the testing phase and the computer program was terminated with the note: "This is the end of the test. Thank you very much for your participation.". During the testing phase two probe sounds were presented next to the positive and negative stimulus. Each stimulus was presented 16 times in a random order with no more than three consecutive presentations of the same stimulus, resulting in a total of 64 trials. Contrary to the training phase no feedback at all was provided in the testing phase. After the 64 trials a note appeared announcing the end of the experiment and thanking the participants for their participation. The responses to all sounds were automatically saved in E-Prime. The results of one participant of group 7 were not included in the analysis since this person reported to have forgotten which the original positive and negative stimulus was during the testing phase resulting in an 'inverse response', i.e. during testing this person responded to the negative but not to the positive stimulus. Informed consent was obtained from the human participants after the nature of the experiment had been explained.

Results

In the initial training procedure the birds learned to discriminate between the two syllables that differed in all of the formant frequencies (Fig. 5.2 and Table 5.1) after 2143 trials on average (2143.13 ± 257.88 s.e.m., n = 8) following the criterion described earlier (Ohms *et al.* 2010a).

After this initial discrimination stage the two non-reinforced probe sounds were introduced in 20% of the stimulus presentations. The response pattern of the birds to these probe sounds compared to the training stimuli is given in figure 5.3 a,b. Recall that Canadian-English infants used only F1 differences, but not F2 or F3, to distinguish between the vowels of d-vowel-t syllables (Curtin *et al.* 2009). The results of the present study are reversed for zebra finches: they utilized F2 and F3 differences to a greater extent than F1 differences because they categorized stimuli primarily based on differences in F2 and F3 (Fig. 5.3 a,b) by responding to probe sounds that had the same F2 and F3 frequencies as the positive stimulus while ignoring probe sounds with the same F2 and F3 frequencies as the negative stimulus. In other words birds did not weight F1 differences between sounds as strong as F2 and F3 differences as they responded similarly to stimuli and probe sounds which differed in F1. Thus, if a bird was trained to respond to e.g. /dit/ it also responded to /dIt/, whereas if it was trained to respond to /dUt/ it also responded to /dut/. Therefore zebra finches indeed seem to weight higher frequencies, i. e. those for which their auditory system is more sensitive, stronger.

Surprisingly, the results of the human subjects (n = 39, average 24.28 years, ranging from 19 to 34 years) are highly similar compared to the results of the zebra finches (Fig. 5.3) and therefore opposite to the classification pattern of the babies found in earlier studies (Lacerda 1993, 1994; Curtin *et al.* 2009).

Repeated measures ANOVA revealed that the human subjects responded significantly slower to probe sounds compared to training stimuli (n=39, F=7.519, p< 0.01) indicating that they did perceive a difference between the sounds but nevertheless treated them as tokens of the same category. For the birds on the other hand no significant difference between reaction times was detected (n=8, F=0.764, p=0,383) although it is highly likely that they also perceived a difference between training and test stimuli since they responded significantly less to the probe sound that they otherwise treated like the positive stimulus (Fig 5.3 a,b).

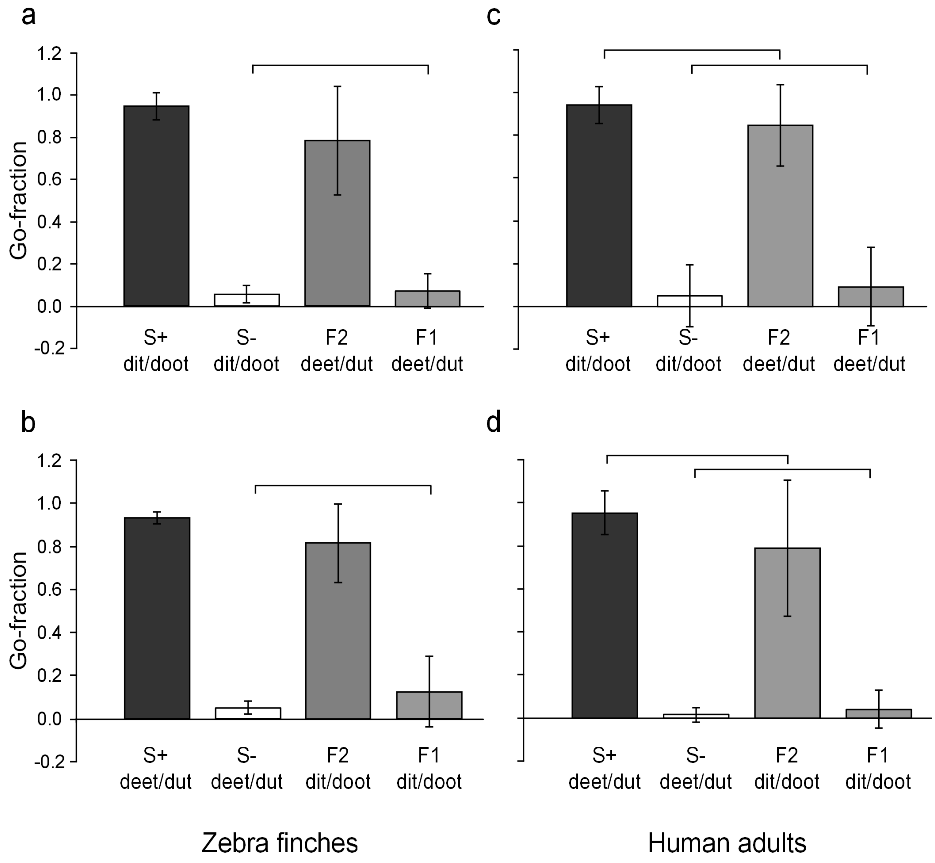

Figure 5.3. Categorization patterns of training stimuli and probe sounds of both zebra finches and Dutch adults.

This figure shows the average proportions including standard deviation of go-responses of birds and humans to training and test stimuli. Every bird got between 50 and 100 probe sound presentations, whereas every human subject got 16 presentations per probe. Horizontal brackets indicate which go-responses did not differ significantly from each other (p<0.05) analyzed with a simultaneous testing procedure based on G-tests of independence (Sokal & Rolf 1995). **(a)** and **(c)**, Go-responses of zebra finches (n = 4) and humans (n = 19) respectively that were first trained to discriminate /dIt/ and /dut/ and afterwards got /dit/ and /dUt/ as probe sounds. F2 beneath the bars indicates the go-response to the probe sound that had the same F2 and F3 frequencies as the positive stimulus but the same F1 frequency as the negative stimulus, whereas F1 indicates the go-response to the probe sound that had the same F1 frequency as the positive stimulus but the same F2 and F3 frequencies as the negative stimulus. **(b)** and **(d)**, show the same information as panels (a) and (c) but for those birds (n = 4) and humans (n = 20) that were trained to discriminate /dit/ and /dUt/ and got /dIt/ and /dut/ as probe sounds. S+, positively reinforced stimulus; S-, negatively reinforced stimulus.

Discussion

The results of our study are striking as they reveal a hitherto undiscovered parallel in speech perception between humans and birds. Up to now differences in acoustic cue weighting strategies in speech perception have been attributed to developmental differences between ages (Curtin *et al.* 2009; Nittrouer 1996; Mayo *et al.* 2003; Mayo & Turk 2004) and linguistic background (Escudero *et al.* 2009; Ylinen *et al.* 2009). We now added a new perspective on cue weighting differences by including a non-related, but highly vocal, species. The discovery that both zebra finches and adult Dutch listeners exhibit the same cue weighting strategy for vowel perception might be explained by the fact that both humans and birds show increased sensitivity in higher frequency regions between approximately 1 and 4 kilohertz, i.e. it might not be attributed to linguistic background at all, given that zebra finches obviously lack comparable experience with the Dutch language.

Why then do Canadian-English infants at 15 months of age as well as Swedish infants between 3 and 12 months exhibit an opposite cue weighting bias? Maybe the reason for that lies in initial difficulties of the auditory system to process noisy sounds or sound components that are spectrally less prominent (Nittrouer & Lowenstein 2009) such as F2 and F3 whereas F1, the most prominent spectral feature of a vowel, dictates categorization in an early stage of vocal learning. For normally raised adult zebra finches, which lack experience with human speech, the sensitivity matches the region with the most prominent frequency range of their natural songs. Whether this sensitivity arises from their exposure to a rich conspecific acoustic environment consisting, like human speech, of complex broad-band, amplitude- and frequency-modulated sounds (Lachlan *et al.* 2010) or whether their sensitivity is independent of such an acoustic experience remains an open question. Whatever the causes, our findings do demonstrate that acoustic cue weighting underlying vowel perception in humans does not need to be a highly derived feature linked to the evolution of speech.

Acknowledgements

We thank Dirk Vet from the Institute of Phonetic Sciences at the University of Amsterdam for helping us writing the E-Prime script to test human listeners and all subjects who volunteered to participate in this study. Gabriël J.L. Beckers from the Max-Planck-Institute for Ornithology helped extracting latencies from the data files of the birds. Funding was provided by the Netherlands Organization for Scientific Research (NWO).

Appendix

Table A 5.1. Formant values of the synthesized stimuli.

	/dit/		/dIt/		/dut/		/dUt/	
	F1	**F2**	**F1**	**F2**	**F1**	**F2**	**F1**	**F2**
T1	220	2862	420	2862	220	1736	420	1736
T2	260	2742	465	2742	260	1616	465	1616
T3	300	2622	510	2622	300	1496	510	1496
T4	340	2502	555	2502	340	1376	555	1376

Table A 5.1 gives the frequency values in Hertz of the first two formants of all synthesized stimuli used in this study. T, token; F1, first formant; F2, second formant.