Does knowing speaker sex facilitate vowel recognition at short durations?[a, b)]

David R. R. Smith

Department of Psychology, University of Hull, Cottingham Road, Hull HU6 7RX, United Kingdom.

Running title: Speaker sex and vowel recognition

[a)]Related work was presented in pilot form in "Time to judge sex of speaker: effect of glottal-pulse rate and vocal-tract length," International Congress on Acoustics, Sydney, Australia, 2010.

[b)]Electronic mail: d.r.smith@hull.ac.uk

ABSTRACT

A man, woman or child saying the *same* vowel do so with very *different* voices. The auditory system solves the complex problem of extracting what the man, woman or child has said despite substantial differences in the acoustic properties of their voices. Much of the acoustic variation between the voices of men and woman is due to changes in the underlying anatomical mechanisms for producing speech. If the auditory system knew the sex of the speaker then it could potentially correct for speaker sex related acoustic variation thus facilitating vowel recognition. This study measured the minimum stimulus duration necessary to accurately discriminate whether a brief vowel segment was spoken by a man or woman, and the minimum stimulus duration necessary to accurately recognise what vowel was spoken. Results showed that reliable vowel recognition *precedes* reliable speaker sex discrimination, thus questioning the use of speaker sex information in compensating for speaker sex related acoustic variation in the voice. Furthermore, the pattern of performance across experiments where the fundamental frequency and formant frequency information of speaker's voices were systematically varied, was markedly different depending on whether the task was speaker-sex discrimination or vowel recognition. This argues for there being little relationship between perception of speaker sex (indexical information) and perception of what has been said (linguistic information) at short durations.

Keywords: speaker sex, vowel recognition, duration, indexical information

PsychINFO classification: 2326 Auditory & Speech Perception

1. Introduction

A man, woman or child saying the same vowel do so with very different voices. The auditory system solves the problem of extracting what has been said despite substantial differences in the acoustic properties of the carrying voice. Much of the acoustic variation between the voices of men and women arises from sexual dimorphism in the underlying anatomical mechanisms for producing speech (Fant, 1970; Titze, 1989; Fitch & Giedd, 1999). If the auditory system knew the sex of the speaker then it could potentially compensate for speaker sex related acoustic variation thus facilitating vowel recognition (e.g., Nordström & Lindbolm, 1975; reviewed Johnson, 2005). The purpose of this study was to investigate speaker-sex discrimination and vowel recognition performance using very brief duration vowels. Of particular interest was whether listeners could reliably tell whether a man or woman spoke before they could reliably identify the vowel that was spoken, and how performance was affected in the two tasks when the acoustic properties of the carrying voice was manipulated.

All mammals produce their communication sounds (including the speech sounds of humans) with the same basic physiological mechanism. The action of the diaphragm pushes air against the vocal folds situated in the larynx at the base of the throat. The vocal folds remain closed until air pressure forces them open. With the subsequent release of air pressure the vocal folds close again. This opening-and-closing action produces a glottal pulse and occurs many times per second. The rate of these glottal pulses (GPR) determines the fundamental frequency (*f*0) of the laryngeal source. The perceived pitch of the voice is closely correlated with *f*0. With each open-and-close cycle, a pulse of air enters the space above the larynx called the supralaryngeal vocal tract. The vocal tract acts as an acoustic filter upon the stream of air pulses entering it. Depending on the configuration of the vocal

tract, governed by different placements of the tongue and jaw positions *etc*, the frequency content of the air stream is differentially reinforced by the resonances of the vocal tract. These vocal tract resonances give rise to spectral prominences known as formants, and these formants distinguish the different sounds of speech. For the general principles of speech production see Fant (1970) and Flanagan (1972).

Much of the acoustic difference in the voices of men and women (and children) arises from characteristic differences in GPR (Titze, 1989) and vocal-tract length (Fant, 1970; Fitch & Giedd, 1999). The length and mass of the vocal folds affect the GPR which leads to changes in the *f*0 of the voice. The sexual dimorphism in *f*0 is attributable to increased testosterone at puberty in males which stimulates growth in the laryngeal cartilages (Beckford *et al*., 1985). The relatively longer and more massive vocal folds of adult males cannot physically support as high a GPR as the shorter and lighter vocal folds of adult females and children. The *f*0 of men's voices is about 0.75 of an octave lower than women's voices primarily because the vocal folds of men are about 60% longer than those of women (Titze, 1989). The *f*0 of men's and women's voices is a highly-salient cue to speaker sex, with men typically having a mean *f*0 of around 130 Hz and women typically having a mean *f*0 of around 220 Hz (Peterson & Barney, 1952; Hillenbrand *et al*., 1995). Listeners are highly sensitive to differences in the *f*0 of individual vowels, with the just noticeable difference being around 2% (Smith *et al*., 2005).

The length of the supralaryngeal vocal-tract is highly correlated with speaker height, increasing with age in both sexes (Fitch & Giedd, 1999). As vocal-tract length (VTL) increases the formants in speech shift toward lower frequencies (Fant, 1970). There is an additional spurt in VTL at puberty for males (Fitch & Giedd, 1999) which, added to the generally greater height of adult males compared to adult females, means that the formant

frequencies of adult males decrease by about 30% from their values at age four while the formant frequencies of adult females decrease by about 20% (Huber *et al*., 1999). Consequently, the formant frequencies of adult males are about 15% less than those of adult females (Peterson & Barney, 1952; Hillenbrand *et al*., 1995).

Pattern classification studies have consistently shown that *f*0 and formants capture most of the difference between the speech sounds of adult males and females (Childers & Wu, 1991; Bachorowski & Owren, 1999), with *f*0 and formant information being highly correlated (Childers & Wu, 1991; Wu & Childers, 1991). Bachorowski and Owren (1999) showed that speaker-sex classification is highly accurate using only *f*0 or only formant frequency information, but best using both cues. Perceptual categorization listening experiments have shown that listeners can identify speaker sex from voiceless fricatives (Schwartz, 1968; Ingemann, 1968) and whispered vowels (Schwartz & Rine, 1968) which only have formant-related information. Other studies have reported that *f*0 is a stronger cue to speaker sex than formants (e.g., Lass *et al*., 1976; Whiteside, 1998) while other studies suggest that formant information can be important in discriminating speaker sex (Coleman, 1976). More recent studies, have manipulated the *f*0 and formants of isolated vowels or sentences, to investigate their relative importance in affecting judgements of speaker sex (Smith & Patterson, 2005; Assmann *et al*., 2006; Smith *et al*., 2007; Hillenbrand & Clark, 2009). The consensus in these more recent studies is that *f*0 and formants contribute about equally to the perception of speaker sex.

Speech sounds such as vowels are characterised by different prominent frequencies (formants) which define the vowels within a multidimensional frequency-domain formant space (Peterson & Barney, 1952). The classic study of Peterson and Barney (1952) measured the frequencies of the formants of the vowels of 76 men, women and children. Plotting the

lowest formant frequency (*F*1) against the second lowest formant frequency (*F*2) showed that individual vowels clustered into specific regions within the *F*1-*F*2 space. However, there was both overlap between vowel clusters, and wide variation between different speakers and different speaker groups (men, women and children). The distinguishing characteristics of the voices of men, women and children – the *f*0 and formants of the voice – affect where in the frequency domain acoustic information denoting an individual speech sound's identity is more likely to be found (Fant, 1970; Fitch & Giedd, 1999; Huber *et al*., 1999). For instance, the first three formants *F*1—*F*3 of the vowel /i/ are on average about 270, 2300 and 3000 Hz for men but 300, 2800 and 3300 Hz for women (Howard & Angus, 2001). Given this variability between voices it might be thought advantageous to know the sex of speaker. The idea that information about speaker sex can help facilitate vowel recognition has been widely advanced (e.g., Potter & Steinberg, 1950; Fujisaki & Kawashima, 1968; Wakita, 1977; Traunmüller, 1981; reviewed by Johnson, 2005)

The Peterson and Barney (1952) vowel data set demonstrated both substantial within-vowel dispersion and significant overlap between vowels. Yet listeners in the study were rarely mistaken in their vowel judgements. Bladon *et al* (1984) attempted to normalize for sex of speaker by shifting the auditory spectra of vowels for women down in frequency, thus partially compensating for the higher formants of women compared to men. Other researchers have included information about *f*0 as well as formants (Potter & Steinberg, 1950), added higher formants (Fujisaki & Kawashima, 1968) or scaled formants by some factor related to *f*0 (Miller, 1989). The general idea is that if some measure of speaker sex can be extracted then it can be used to remove some of the difference in the acoustic properties of the voice arising from sexual dimorphism in the underlying anatomical mechanisms for producing speech, thus facilitating vowel recognition.

Previous research into the discrimination of speaker sex as a function of vowel duration has shown that the percept is available at very short durations (Whiteside, 1998; Owren *et al*., 2007; Harding & Cooke, 2008). Whiteside found nearly ceiling performance with stimuli as short as 50 or 100 ms. Owren and colleagues tested the ability to judge speaker sex with vowel segments as short as one glottal cycle which equates to a duration of around 5 to 8 ms depending on the sex of the speaker. Owren *et al* found that listeners could discriminate speaker sex at around 1.7 glottal cycles (equivalent to around 8 and 14 ms for women and men respectively). Previous research into vowel recognition as a function of vowel duration has also shown that the percept is available at very short durations (Suen & Beddoes, 1972; Robinson & Patterson, 1995; Harding & Cooke, 2008). For instance, Suen and Beddoes (1972) found that vowels could be identified at durations as short as 10 ms.

When someone speaks information is present in the sound wave in a number of forms. The most obvious form of information is the linguistic message – what the person has just said. However, indexical information relating to sociocultural status, emotional state and physical attributes are also embedded in the sound wave and influence judgements about the speaker (Ladefoged & Broadbent, 1957; Sachs et al., 1972; Giles & Powsland, 1975; Murray & Arnott, 1993; Krause et al., 2002). Whether someone speaking is a man or woman is one of the most important and salient pieces of indexical information available to the listener. The experiments in this paper investigate both speaker-sex discrimination and vowel recognition as a function of vowel duration. Given that speaker-sex categorization is heavily influenced by the *f*0 and formant properties of the cueing voice (e.g., Lass *et al*., 1976; Coleman, 1976; Whiteside, 1998; Smith & Patterson, 2005; Assmann *et al*., 2006; Smith *et al*., 2007; Hillenbrand & Clark, 2009), what happens to speaker sex and vowel recognition performance as *f*0 and formant information become available as vowel duration is increased? Furthermore,

this study systematically manipulated the *f*0 and formant properties of the carrying voice to investigate the relative importance of these two cues to speaker-sex discrimination and vowel recognition.

Specifically, the experiments in this paper measured the minimum stimulus duration necessary for a listener to accurately discriminate whether a brief vowel segment was spoken by a man or woman (min_{sex}), and the minimum stimulus duration necessary to accurately recognize what vowel was spoken (min_{vow}). The hypothesis is that if speaker sex is used to compensate for speaker-sex differences in the acoustic properties of vowel sounds, then the stimulus duration required to make accurate judgements of speaker sex should be less than the stimulus duration required to recognize what vowel was spoken ($min_{sex} < min_{vow}$). By manipulating the *f*0 and formant frequencies of the original speaker voices, it should be possible to slow and/or bias the ability to tell speaker sex. These manipulations consisted of creating vowels with an *f*0 intermediate between those of a man's and woman's vowels, or creating vowels with formant frequencies intermediate between those of a man's and woman's vowels, or creating vowels with both *f*0 and formant frequencies intermediate between those of a man's and woman's vowels. These *f*0 and formant frequency manipulations should allow the relative importance of *f*0 and formants to judgements of speaker sex and vowel recognition to be measured, and to see how the perception of speaker sex (indexical information) might or might not affect the perception of what has been said (linguistic information) at short durations.

The experiments in this paper are arranged in three groups. The first group of experiments (Experiments 1—4) measured min_{sex} and min_{vow} across different *f*0 and formant frequency manipulation conditions. The second group of experiments (Experiments 5—8) investigated the effect of introducing an offset noise mask immediately following the brief vowel

segments. The noise mask was used to deter auditory processing (re-sampling) of the echoic memory of the vowels. The third group of experiments (Experiments 9—12) explored the effect of increasing variability in the stimulus set by allowing four times as many different pairings of the men and women speakers' vowels.

2. Method

2.1 Experiments 1—4

2.1.1 Overview

Listeners were presented isolated vowels recorded from eight different speakers (four adult men and four adult women). The vowels were either not manipulated (Experiment 1), or had their glottal-pulse rate (GPR) modified to the same intermediate value (Experiment 2), or had their simulated vocal-tract length (VTL) modified to the same intermediate value (Experiment 3), or had both their GPR and VTL modified to the same two intermediate values (Experiment 4). The intermediate values chosen for the GPR and VTL modifications in Experiments 2—4 were the geometric mean of the men's and women's vowels for these parameters. Perceptually, GPR is heard as the fundamental frequency (*f*0) of the voice (Titze, 1989). Perceptually, VTL affects the frequencies of the formants with longer VTLs leading to lower frequency formants (Fant, 1970). The vowels were presented at six very brief durations (5, 8, 12, 18, 27 and 40 ms). The ability of listeners to correctly judge the sex of the original speaker and what vowel the speaker had said was measured.

2.1.2 Participants

Twelve native-English speaking listeners participated in Experiments 1—4, six male and six female (age range 18—37 yr, mean=22.8 yr, *SD*=5.1 yr). All listeners had normal audiometric absolute thresholds at both ears at 0.5, 1, 2 and 4 kHz, demonstrating normal hearing. Listeners were naive to the purpose of the experiment and were paid volunteers. Informed consent was given by the participants after the experiments were introduced to them. The experimental procedure was approved by the Hull Psychology Research Ethics Committee.

2.1.3. Stimuli

Examples of the five English vowels /a/, /e/, /i/, /o/, /u/ (/a—u/), corresponding to the vowel sounds in “fa”, “bay”, “bee”, “toe” and “zoo”, of four adult men and four adult women were recorded in a quiet room using a high-quality microphone (Shure SM58-LCE), with a sampling rate of 48 kHz and an amplitude resolution of 16-bits. The speakers were native-English speaking students at the University of Hull. The microphone was connected to a preamp (Xenyx Behringer 502) to boost the signal before recording through the PC sound card. Speakers were required to utter the vowels at a regular relaxed rate at a comfortable effort level. For each speaker, one example vowel that was free of unwanted noise from jaw articulation, lip-smacking and breathing, was selected for further processing. Details of the physical and acoustic characteristics of the speakers are shown in Table I.

TABLE I HERE

The GPR and simulated VTL of the vowels were manipulated using STRAIGHT (Kawahara *et al*., 1999; Kawahara and Irino, 2004). STRAIGHT is a sophisticated vocoder that uses the classical source-filter theory of speech (Fant, 1970) to segregate GPR information from the spectral-envelope information associated with the shape and length of the vocal tract. See Smith *et al* (2005) for a description of how STRAIGHT is used to manipulate vowels to simulate different speakers, and Kawahara and Irino (2004) for the underlying principles. Liu and Kewley-Port (2004) have reviewed STRAIGHT and commented favourably on its ability to manipulate formant-related information.

In Experiment 1, the GPR and simulated VTL of the four men's and four women's vowels were not manipulated. In Experiment 2, the GPR of the men's and women's vowels was modified to the same intermediate value, equal to the geometric mean of the men's and women's GPR. Thus, the vowels of man 1 and woman 1 were set to have a GPR of 119 Hz (=√(95•150), see Table I). Similar manipulations were performed for the GPR of the other three men and women pairs. The simulated VTL was not manipulated in Experiment 2. The geometric mean was chosen as the intermediate point as it was more nearly half-way between the distributions of the men's and women's vowels' GPR in this study than the arithmetic mean. Pilot listening (and later analysis of the results) showed that this intermediate value was still in the ambiguous men—women range. In Experiment 3, the simulated VTL of the men's and women's vowels was modified to the same intermediate value, equal to the geometric mean of the men's and women's estimated VTL. Thus, the vowels of man 1 and woman 1 were set to have a simulated VTL of 15.42 cm (=√(16.32•14.57), see Table I). Similar manipulations were performed for the simulated VTL of the other three men and women pairs. The GPR was not manipulated in Experiment 3. In Experiment 4, both the GPR and the simulated VTL of the men's and women's vowels were modified to the same two

intermediate values, equal to the geometric means of the men's and women's GPR and estimated VTL. Thus, the vowels of man 1 and woman 1 were set to have a GPR of 119 Hz (=√(95•150)) and a simulated VTL of 15.42 cm (=√(16.32•14.57)). Similar manipulations were performed for the GPR and simulated VTL of the other three men and women pairs. Figure 1 shows a schematic of these four types of manipulation.

FIGURE 1 HERE

The duration of all vowels was adjusted to have six different durations (5, 8, 12, 18, 27, and 40 ms) by taking different duration length segments from the central portion of each vowel. Each segment was cosine-square gated to ensure that the sounds came on and went off smoothly over the first and last 1 ms respectively. Finally, all the vowel sounds of all durations were normalised to the same root-mean-squared (rms) level of 0.0250 (relative to maximum of ±1). The stimuli were played by a 24-bit sound card (X-fi Xtreme Audio, Sound Blaster, Creative) and presented to the listener diotically over Sennheiser HD600 headphones. Listeners were seated in a single-walled, IAC, sound-attenuating booth. The sound level of the vowels at the headphones was 77 dB SPL.

2.1.4 Procedure

The experiments were performed using a single-interval, two-response paradigm. The listener heard a vowel of a given duration and had to indicate first whether a man or women had spoken the vowel and then second what vowel had been said. There was a 50% chance that either a man or woman had spoken the original vowel. There was a 20% chance that the vowel was a particular vowel from the set of five (/a—u/). The judgement of the sex of the

speaker and the vowel uttered was made by selecting the appropriate buttons on a visual display. The order of the ‘man’ and the ‘woman’ buttons was pseudo-randomly switched at the beginning of each run.

Listeners were first given a practice run of 50 trials with a single vowel duration of 100 ms, where both GPR and VTL information was available. The purpose of the practice was partly to familiarise listeners with the experimental procedure but mainly to ensure that listeners could correctly associate each heard vowel to its orthographic representation (/a/ *etc*) on the response display. The five vowels were each presented in a pseudo-random order 10 times, with half spoken by men and half spoken by women. Listeners invariably found it an easy task to judge the sex of the speaker at this duration (99% correct on average) but some listeners found it relatively hard to correctly identify what vowels were uttered (88% correct on average). Four listeners were given another practice run of 50 trials to reach a criterion performance level of better than 90% on sex discrimination and vowel recognition, and two listeners required a further practice run of 50 trials to reach criterion performance.

Listeners then proceeded on to the main experiments. The listener was given a run of 300 trials, consisting of six durations (5, 8, 12, 18, 27, 40 ms), each repeated 50 times. Half the trials were vowels spoken by men and half the trials were vowels spoken by women (balanced across durations and vowels). The duration, sex and vowel were presented in a pseudo-random order generated by the computer. Which of the four men’s or four women’s vowels was used in any one trial was also pseudo-randomly determined by the computer. There was no feedback. Each experimental run of 300 trials took approximately 15-20 min to complete.

The design was a within-subjects design. Thus all listeners did Experiments 1—4 but the order was counterbalanced to control for the effects of experience and/or fatigue. Each

listener did their experiments in two sessions each lasting approximately one hour. At the start of each session the listener performed a practice run of 50 trials to ensure they were still performing at better than 90% on the sex discrimination and vowel recognition tasks with vowels at a duration of 100 ms.

2.2 Experiments 5—8

In Experiments 5—8, the first four experiments were repeated but with the addition of a noise mask immediately following the offset of the short duration vowel. The Gaussian noise mask was 500 ms in duration, with an onset and offset that was smoothed by a cosine-gating function of 10 ms. The sound level of the Gaussian noise at the headphones was 69 dB SPL.

All other procedural details were the same as for Experiments 1—4 except that the number of repetitions per duration was reduced from 50 to 30. This was to reduce the time spent collecting data for each listener in a situation where participation was for course credit.

A different set of ten native-English speaking listeners participated in Experiments 5—8, three male and seven female (age range 19—38 yr, mean=22.3 yr, *SD*=6.7 yr). Audiometric thresholds were measured at both ears at 0.5, 1, 2 and 4 kHz, and demonstrated normal hearing. Listeners were naive to the purpose of the experiments and participated to earn course credit. Listeners provided informed consent after the experiments were introduced to them. The experimental procedure was approved by the Hull Psychology Research Ethics Committee.

2.3 Experiments 9—12

In Experiments 10—12, the first three experiments were repeated but with the pairing between men's and women's vowels systematically varied. In Experiments 2—4, there were only four different men—women pairings when calculating the intermediate values for GPR and VTL (man1—woman1, man2—woman2, man3—woman3 and man4—woman4). In order to increase the variability in men—women pairings, the possible pairings were systematically varied (man1—woman1, man1—woman2, man1—woman3, man1—woman4, man2—woman1 *etc*). This increased the possible men—women pairings from four to sixteen different pairings. Experiment 9 did not involve any GPR and VTL manipulations, and thus is a replication of Experiment 1 but with a different set of listeners.

All other procedural details were the same as for Experiments 1—4 except that the number of repetitions per duration was reduced from 50 to 30. This was done to reduce data collection time.

A different set of nine native-English speaking listeners participated in Experiments 9—12, two male and seven female (age range 19—31 yr, mean=21.2 yr, *SD*=3.7 yr). Audiometric thresholds were measured at both ears at 0.5, 1, 2 and 4 kHz, and demonstrated normal hearing. Listeners were naive to the purpose of the experiments and participated to earn course credit. Listeners provided informed consent after the experiments were introduced to them. The experimental procedure was approved by the Hull Psychology Research Ethics Committee.

3. Results and Discussion

Fig. 2 shows percentage correct judgement of original speaker sex and percentage correct recognition of vowel, as a function of duration of the vowel, for Experiments 1—4. Fig. 1

represents schematically the experimental manipulations of GPR and VTL for Experiments 1—4. Results are based on the mean data from all twelve listeners. The results presented in Fig. 2 are pooled across both men and women speaker judgements, and across all five vowels. Chance performance, the point when the listener cannot tell whether a man or a woman spoke the vowel, is 50% for the speaker-sex judgement [*d'*=0 in a two-alternative forced-choice (2AFC) task]. The vowel duration at which listeners can reliably tell whether a man or woman spoke (min_{sex}) is taken to be the 75% point [*d'*=1 in a 2AFC task, Macmillan & Creelman (1991)]. Chance performance for the vowel recognition task is 20% [*d'*=0 in a five-alternative forced-choice (5AFC) task]. The vowel duration at which listeners can reliably tell which vowel was spoken (min_{vow}) is taken to be the 50% point [*d'*=1 in a 5AFC task].

FIGURE 2 HERE

Percentage correct scores for the speaker-sex task are the same or only marginally higher than for the vowel-recognition task for all durations in Experiments 1—3. For Experiment 4, the percentage correct scores for the speaker-sex task are markedly lower than for the vowel-recognition task. If information about speaker sex was used to compensate for speaker-sex related acoustic variation to facilitate vowel recognition, then the percept of speaker sex should be available before the ability to recognise vowels. It is clear that the point at which listeners can reliably tell whether a man or woman spoke (min_{sex}) is *not* reliably available before the point at which listeners can reliably tell which vowel was spoken (min_{vow}). This undermines the idea that speaker sex is used as a prior label to allow vowel recognition.

It might be argued that it is not necessary for speaker sex to be known reliably before that information is used to facilitate vowel recognition. If we could discover a similar pattern of change in the two tasks of speaker-sex discrimination and vowel recognition, as we change the GPR and simulated VTL across Experiments 1—4, then this might point to a facilitative relationship between the two tasks. How speaker-sex discrimination performance and vowel-recognition performance change as a function of vowel duration across Experiments 1—4 are treated separately below.

3.1 Building knowledge about sex of speaker over time

Fig. 3 shows percentage correct judgement of original speaker sex, as a function of vowel duration, for Experiments 1 to 4. At short durations (5 and 8 ms), there is little difference between performance levels in Experiment 1 and Experiment 2. However, at durations of 12 ms and longer there is a reduction in performance when having GPR removed as an effective cue to speaker sex (Experiment 2), compared to when having two speaker cues of GPR and VTL (Experiment 1). At 5 ms there is little difference between the performance levels in Experiments 1 and 3, but at durations of 8 and 12 ms, there is a reduction in performance when having VTL removed as an effective cue to speaker sex (Experiment 3), compared to when having two speaker cues of GPR and VTL (Experiment 1). However, at longer durations of 18 to 40 ms there is little difference between performance levels in Experiments 1 and 3. There is a large reduction in speaker-sex discrimination performance when having both GPR and VTL removed as effective cues to speaker sex (Experiment 4), compared to when having both GPR and VTL as cues to speaker sex (Experiment 1), or just VTL as a cue to speaker sex (Experiment 2), or just GPR as a cue to speaker sex (Experiment 3), for all durations tested (5—40 ms).

In Experiments 2—4, the GPR and VTL values of the men's and women's vowels were manipulated to an intermediate value, equal to the geometric mean. An analysis of the speaker sex responses was conducted to check whether listeners showed any bias in responding "man" or "woman" as a consequence of the manipulations. In Experiment 2—4, the proportions of responses across all listeners, regardless of correctness, were 0.48 "man" and 0.52 "woman" (Experiment 2), 0.47 "man" and 0.53 "woman" (Experiment 3), and 0.48 "man" and 0.52 "woman" (Experiment 4). In Experiment 1 where there were no manipulations in GPR or VTL, the proportions were 0.46 "man" and 0.54 "woman". There does not appear to be a bias in listeners "man" and "woman" response rates.

FIGURE 3 HERE

Best-fitting Weibull psychometric functions were fitted to the percentage correct speaker sex data scores in Experiments 1 to 4 using a maximum-likelihood method (Wichmann & Hill, 2001). A series of Monte Carlo tests[1] were performed to determine whether any two psychometric functions could have come from the same underlying distribution of psychometric functions. The tests showed that the psychometric function for Experiment 1 (both GPR and VTL cues present) was significantly different from the psychometric function for Experiment 2 (GPR removed as an effective cue), and the psychometric function for Experiment 3 (VTL removed as an effective cue), and the psychometric function for Experiment 4 (both GPR and VTL removed as effective cue), all at $p<0.001$. The psychometric functions for Experiment 2 and Experiment 3 were also significantly different

[1] Monte Carlo simulation provided by pfcmp (http://bootstrap-software.org/psignifit/faq.php#pfcmp), written by Jeremy Hill [Last checked December 2013]

from each other (p<0.001). The critical alpha was taken to be 0.00625 (Bonferroni corrected =0.05/8).

The point at which listeners can reliably tell whether a man or woman spoke – the duration threshold (min_{sex}) for reliable discrimination – was taken to be the 75% point on the fitted curve (d'=1 for 2AFC). When listeners have access to unmodified voices as in Experiment 1, the vowel duration needs to be 8.8 ms, before the listeners can reliably tell whether a man or woman spoke the original vowel. This value is similar to Harding and Cooke (2008) and Owren *et al*. (2007), who estimate the point of reliable speaker-sex discrimination at between about 10 and 15 ms for experiments similar to Experiment 1. It is clear that the acoustic information in speech relating to speaker sex can be extracted from very short duration stimuli. The early availability of speaker sex information agrees with the idea that many characteristics of the auditory scene are extracted very rapidly (Harding *et al*., 2008). At stimulus durations of about 5 ms speaker-sex discrimination performance is at chance levels but by durations of 25 ms speaker-sex discrimination performance approaches 100%.

When listeners have only one cue (either VTL or GPR), as compared to two cues (GPR and VTL), the vowel duration needs to be longer before the listener can reliably tell whether a man or woman spoke the original vowel. When the GPR of the original speakers was modified to be the same but still leaving VTL as a potential cue (Experiment 2, *cf*. Fig. 1 and Fig. 3), the listener needs a vowel duration of 10.3 ms to reliably tell whether a man or woman spoke the original vowel. When the simulated VTL of the original speakers was modified to be the same but still leaving GPR as a potential cue to speaker sex (Experiment 3, *cf*. Fig. 1 and Fig. 3), the listener needs a vowel duration of 11.4 ms to reliably tell whether a man or woman spoke the original vowel.

Though judgement of original speaker-sex is impaired with the loss of either the GPR or VTL cue, the pattern of impairment across duration is different. With the equalization of GPR (Experiment 2), performance is impaired at durations of 12 ms and longer. For the 5 and 8 ms duration there is no drop in performance compared to having both GPR and VTL cues available (Experiment 1). This is presumably because at the shortest durations there is no pitch cue available – speaker-sex performance is determined by the available cue of VTL. Similarly, work in music perception has shown that note timbre can be identified at durations too short to support pitch-chroma judgements (Robinson & Patterson, 1995). The impaired performance in Experiment 2 at stimulus durations of 12 ms and longer highlights the importance of GPR as a cue to speaker sex.

With the equalization of simulated VTL (Experiment 3), performance is impaired for durations up to and short of about 18 ms but not for durations longer than this, This would suggest that performance in Experiment 3 is impaired at very short durations because of the loss of VTL as a reliable cue to speaker sex and the relative weakness of the available GPR cue. However, at durations around 18 ms and longer, pitch arises as a strong perceptual cue to speaker sex. GPR can be used to support speaker-sex discrimination performance levels at the same errorless levels as having both GPR and VTL cues (Experiment 1).

Finally, when both the GPR and VTL of the original speakers are modified to be the same across men and woman speakers, thus leaving only residual cues to speaker sex other than GPR and VTL (Experiment 4, *cf*. Fig. 1 and Fig. 3), the listener never reaches the criterion performance level of 75% correct for reliable discrimination. The speaker-sex discrimination performance level asymptotes at 69% with vowel durations of 27 ms or longer. However, it is noticeable that the performance at vowel durations of 12 ms and longer is still greater than chance. Greater than chance performance indicates the presence and saliency of other cues to

speaker sex beyond GPR and VTL (e.g., Assmann *et al*., 2006; Smith *et al*., 2007). Other cues to speaker sex beyond GPR and VTL, that would still be present in the very short duration vowels used in this study, could be differences in the pattern of formants in the vowels of men and women consequent upon underlying anatomical differences in the proportions of the vocal tract between men and women (Fant, 1966, 1975). For instance, the pharynx is proportionally longer in adult males than adult females (Fitch & Gield, 1999).

3.2 Building knowledge about vowel identity over time

Fig. 4 shows percentage correct recognition of vowel spoken, as a function of duration of the vowel, for Experiments 1 to 4. We can see that there is little change in performance across all four experimental conditions (Experiments 1—4) for all tested durations.

FIGURE 4 HERE

Best-fitting Weibull psychometric functions were fitted to the percentage correct recognition of vowel spoken data scores in Experiments 1 to 4 using a maximum-likelihood method (Wichmann & Hill, 2001). Monte Carlo tests showed that the psychometric function for Experiment 1 (both GPR and VTL cues present) was just significantly different from the psychometric function for Experiment 2 (GPR removed as an effective cue) p=0.006, but not significantly different from the psychometric function for Experiment 3 (VTL removed as an effective cue) p=0.159, or the psychometric function for Experiment 4 (both GPR and VTL removed as effective cue), p=0.037. The psychometric functions for Experiment 2 and Experiment 3 were not significantly different from each other (p=0.423).

The point at which listeners can reliably recognise what vowel was spoken – the duration threshold (min_{vow}) for accurate vowel recognition – was taken to be the 50% point on the fitted curve (*d'*=1 for 5AFC). The vowel duration needed to accurately recognise the vowel spoken was extrapolated to be 3.7, 4.0, 3.9 and 4.2 ms for Experiments 1—4 respectively. The minimal impairment in vowel recognition across the four experimental conditions, when viewed in light of the marked impairments in speaker sex performance across the four experimental conditions (*cf*. Fig. 3 and Fig. 4), suggests that knowledge of speaker sex (extralinguistic information) has little impact upon vowel recognition (linguistic information) at short durations. For a review of speaker normalisation in speech perception at durations typical of everyday speech see Johnson (2005).

3.2 Auditory sensory memory

A criticism of Experiments 1—4 is that auditory sensory memory (echoic memory) could allow the listener to re-sample the short duration stimuli. Experiments 5—8 repeated Experiments 1—4 but included an offset noise mask immediately following the short duration vowels to deter auditory processing of the echoic memory.

Fig. 5 shows percentage correct judgement of original speaker sex, as a function of vowel duration, for Experiments 5—8. Results are based on the mean data from all ten listeners. All other details in the figure are the same as in Fig. 3, to which Fig. 5 should be compared. The pattern of results for speaker-sex discrimination across Experiments 5—8 are similar to those for Experiments 1—4. Speaker-sex discrimination performance is best for Experiment 5 (when both GPR and VTL cues to speaker sex are available), is worse for Experiments 6 (where the GPR cue to speaker sex has been effectively removed), is slightly worse for Experiment 7 (where the VTL cue to speaker sex has been effectively removed), and worst

for Experiment 8 (where both GPR and VTL cues to speaker sex have been effectively removed).

FIGURE 5 HERE

Best-fitting Weibull psychometric functions were fitted to the percentage correct speaker sex data scores in Experiments 5—8 (Wichmann & Hill, 2001). The psychometric function for Experiment 5 (both GPR and VTL cues present) was significantly different from the psychometric function for Experiment 6 (GPR removed as an effective cue), and the psychometric function for Experiment 8 (both GPR and VTL removed as effective cues), all at $p<0.001$. However, the psychometric function for Experiment 5 was not significantly different ($p=0.099$) from the psychometric function for Experiment 7 (VTL removed as an effective cue). The psychometric functions for Experiment 6 and Experiment 7 were significantly different from each other ($p<0.001$).

The duration threshold for reliable speaker sex discrimination (min_{sex}) was calculated as the 75% point on the best-fitting Weibull psychometric curve. These values were 12.4, 24.5 and 13.5 ms for Experiments 5—7 respectively. In Experiment 8, performance is just under 65% at the longest duration, so the speaker-sex discrimination threshold point was never reached. Compared to the threshold values for Experiments 1—4 (8.8, 10.3, 11.4 ms and asymptote of 69% for Experiment 4 respectively), the threshold values in Experiments 5—8 show that the task is consistently harder when a noise mask immediately follows the short duration vowel. Generally, performance is reduced at each duration for Experiments 5, 6 and 8 by around ten percentage points, accept for where performance is near to chance (50% in 2AFC) or at longer durations (27 ms or more) where there is less opportunity to benefit from

re-sampling the echoic memory. Thus listeners were re-sampling to some extent the echoic memory of the short duration vowels in Experiments 1—4. Comparing Experiment 6 to Experiment 2 (min_{sex} of 24.5 ms *vs* 10.3 ms), shows that preventing re-sampling of the echoic memory which contains VTL information leads to substantially worse performance. However, comparing Experiment 7 to Experiment 3 (min_{sex} of 13.5 ms *vs* 11.4 ms), shows little benefit of re-sampling the echoic memory of the short duration vowels. This is because the echoic memory in Experiment 3 contains GPR information which only becomes useful in discriminating speaker sex at durations of at least 15 ms.

Fig. 6 shows percentage correct recognition of vowel spoken, as a function of duration of the vowel, for Experiments 5—8. There is little change in performance across all four experimental conditions (Experiments 5—8) for all tested durations.

FIGURE 6 HERE

Best-fitting Weibull psychometric functions were fitted to the percentage correct vowel recognition data scores in Experiments 5—8 (Wichmann & Hill, 2001). The psychometric functions for Experiment 5—8 were not significantly different from each other (Experiment 5 to Experiment 6, p=0.979; Experiment 5 to Experiment 7, p=0.337; Experiment 5 to Experiment 8, p=0.991; Experiment 6 to Experiment 7, p=0.219). The vowel duration threshold for reliable vowel recognition (min_{vow}) was calculated as the 50% point on the best-fitting Weibull psychometric curve. These values were 7.3, 7.7, 6.4 and 7.5 ms for Experiments 5—8 respectively. The noise mask immediately following the short duration vowels reduced performance in Experiments 5—8 compared to Experiments 1—4 (3.7, 4.0, 3.9 and 4.2 ms respectively), again suggesting some contribution from echoic memory.

However, the lack of any effect upon vowel recognition when manipulating the GPR and VTL of the spoken vowels across experimental condition in Experiments 5—8, is the same as for Experiments 1—4 (*cf*. Fig. 6 to Fig. 4).

Notably, the same general pattern of minimal impairment in vowel recognition across experimental condition *and* marked impairment in speaker sex discrimination performance across experimental condition as GPR and VTL are systematically manipulated, is the same whether (Experiment 5—8) or not (Experiment 1—4) a noise mask is presented after the short duration vowel. This suggests that knowledge of speaker sex has little influence upon vowel recognition at short durations even when echoic memory is masked.

3.3 Increasing variability of men—women pairings

One consideration of Experiments 2—4 was that the male—female pairings was limited to just four different combinations. Experiments 10—12 investigated this limitation by increasing the variability of men—women pairings by using all possible men—women pairings of the stimulus set. This represented a four-fold increase from four to sixteen men—women pairings. Experiment 9 was a straight forward replication of Experiment 1 with the different set of listeners who participated in Experiments 10—12.

Fig. 7 shows percentage correct judgement of original speaker sex, as a function of vowel duration, for Experiments 9—12. Results are based on the mean data from all nine listeners. All other details are the same as for Fig. 3, to which Fig. 7 should be compared. The pattern of results for speaker-sex discrimination across Experiments 9—12 are similar to those for Experiments 1—4. Performance is best for Experiment 9 (when both GPR and VTL cues to speaker sex are available), is worse for Experiment 10 (where the GPR cue to speaker sex has been effectively removed) and Experiment 11 (where the VTL cue to speaker sex has been

effectively removed), and worst for Experiment 12 (where both GPR and VTL cues to speaker sex have been effectively removed).

FIGURE 7 HERE

Best-fitting Weibull psychometric functions were fitted to the percentage correct speaker sex data scores in Experiments 9—12 (Wichmann & Hill, 2001). The psychometric function for Experiment 9 was significantly different from the psychometric function for Experiment 10 (p=0.001), significantly different from Experiment 11 (p=0.001) and significantly different from Experiment 12 (p<0.001). The psychometric function for Experiment 10 was not significantly different (p=0.064) from the psychometric function for Experiment 11.

The vowel duration threshold for reliable speaker-sex discrimination (min_{sex}) was calculated as the 75% point on the best-fitting Weibull psychometric curve. These values were 8.7, 12.3 and 12.6 ms for Experiments 9—11 respectively. In Experiment 12, performance asymptotes at around 65% for the longer durations (27 and 40 ms), so the speaker-sex discrimination threshold point was never reached. Both the absolute threshold values and the general pattern is essentially the same between Experiments 9—12 (8.7, 12.3, 12.6 ms and undefined respectively) and Experiments 1—4 (8.8, 10.3, 11.4 ms and undefined respectively). Increasing the variability of the pairings by a factor of four has had little effect which suggests that lack of variability in pairings is not a serious issue in these experiments.

Fig. 8 shows percentage correct recognition of vowel spoken, as a function of vowel duration, for Experiments 9—12. There is little change in performance across all four experimental conditions (Experiments 9—12) for all tested durations.

FIGURE 8 HERE

Best-fitting Weibull psychometric functions were fitted to the percentage correct vowel recognition data scores in Experiments 9—12 (Wichmann & Hill, 2001). The psychometric functions for Experiment 9 and Experiment 10 were not significantly different from each other (p=0.012). However, the psychometric function for Experiment 9 was statistically different from that of Experiment 11 (p<0.001) and that of Experiment 12 (p<0.001). The psychometric functions of Experiment 10 and Experiment 11 were not statistically different from each other (p=0.118).

The vowel duration threshold for reliable vowel recognition (min_{vow}) was calculated as the 50% point on the best-fitting Weibull psychometric curve. These values were 6.1, 5.4, 5.2 and 4.6 ms for Experiments 9—12 respectively. These threshold values are similar to those of Experiments 1—4 (3.7, 4.0, 3.9 and 4.2 ms respectively), albeit showing some reduction in performance, but essentially showing little effect upon vowel recognition when manipulating the GPR and VTL of the spoken vowels across experimental condition. This is similar to the lack of effect shown by Experiments 1—4 (*cf.* Fig. 8 to Fig. 4).

The same general pattern of minimal impairment in vowel recognition across experimental condition *and* marked impairment in speaker sex discrimination performance across experimental condition as GPR and VTL are systematically manipulated, is the same whether there are sixteen men—women pairings (Experiment 9—12) or only four men—women pairings (Experiment 1—4). This suggests that knowledge of speaker sex has little influence upon vowel recognition at short durations even with increased variability in the men—women pairings.

4. General Discussion

If knowledge of speaker sex is used to compensate for speaker sex difference in the acoustic properties of vowels in order to facilitate vowel recognition we might expect reliable information about speaker sex to be available *before* vowel identity. This is certainly not the case – we know what vowel was spoken before we know the sex of the speaker – and if we systematically manipulate acoustic cues affecting the perceived sex of the speaker, such as *f*0 and the formant properties of the vowels, we get significant changes in judgement of speaker sex with no effect upon vowel recognition performance.

Overall, the pattern of speaker-sex discrimination performance as a function of duration and across different manipulations of *f*0 and formants, suggests that in very brief duration vowel sounds the listener uses VTL-related perceptual cues (frequencies of the formants) to distinguish men's voices from women's voices. However, at the point at which the percept is available the listener switches to increasingly using GPR-related perceptual cues (voice pitch). Speculatively, when constructing a hypothesis the listener combines what information is available, using fast but less reliable information at the start and updating that hypothesis with slower but more reliable information as time exposed to the stimulus increases. Such an approach maximises performance in a rapidly changing dynamic environment.

One consideration in this study is that equal duration stimuli were used. This introduces a systematic confound in that for any given duration vowel, there will be twice as much GPR information in a woman's vowel compared to a man's vowel. For this reason other researchers (e.g., Robinson & Patterson, 1995; Owren *et al*., 2007) have used men's and women's vowels equated for number of glottal cycles. However, equating for number of glottal cycles produces in turn a systematic confound in that for any given vowel, the man's

vowel will be twice as long as the woman's vowel. A particular point of the present study was to explore how speaker-sex discrimination and vowel recognition performance builds up over time. In everyday interactions the listener has to make discriminations about whether a man or woman is speaking based on voice information that arrives with this confound as well. For these reasons it was decided to equate for duration and allow GPR information availability to differ between men's and women's vowels.

In summary, listeners were presented with very brief duration vowels (5—40 ms) spoken by either men or women. Listeners were required to judge whether a man or a woman had spoken the original vowel and what vowel had been spoken. In Experiment 1, the vowels were untouched. In Experiment 2, the GPR of the vowels of the men and women were modified to the same intermediate value. In Experiment 3, the VTL of the vowels of the men and women were modified to the same intermediate value. In Experiment 4, both the GPR and the VTL of the vowels of the men and women speakers were modified to the same intermediate values. The results show that the stimulus duration required to make accurate judgements of speaker sex (min_{sex}) is *greater* than the stimulus duration required to recognise what vowel (min_{vow}) was spoken. Thus reliable vowel recognition *precedes* reliable speaker-sex discrimination. The auditory system does not seem to need to extract the sex of the speaker to correct for speaker-sex related acoustic variation in order to facilitate vowel recognition. Furthermore, the pattern of performance across Experiments 1 to 4, where GPR and VTL information were systematically varied, is markedly different depending on whether the task is speaker-sex discrimination (Fig. 3) or vowel recognition (Fig. 4). This basic pattern of results was found when all experiments were repeated with the addition of an offset noise mask (Experiments 5—8), and when the number of men—women pairings was increased four-fold (Experiments 9—12). The general pattern of results across the study

argues for there being little relationship between perception of speaker sex (extralinguistic information) and perception of what has been said (linguistic information) at short durations.

Acknowledgements

This research was supported by a grant from the University of Hull.

References

Assmann, P. F., Nearey, T. M., & Dembling, S. (2006). Effects of frequency shifts on perceived naturalness and gender information in speech. In *Proceedings of the 9th International Conference on Spoken Language Processing*, 889-892, Pittsburgh.

Bachorowski, J., & Owren, M. J. (1999). Acoustic correlates of talker sex and individual talker sex identity are present in a short vowel segment produced in running speech. *Journal of the Acoustical Society of America*, *106*, 1054-1063.

Beckford, N. S., Rood, S. R., & Schaid, D. (1985). Androgen stimulation and laryngeal development. *Ann. Otol. Rhinol. Laryngol.*, *94*, 634-640.

Bladon, R. A., Henton, C. G., & Pickering, J. B. (1984). Towards an auditory theory of speaker normalization. *Language Communication*, *4*, 59-69.

Childers, D. G., & Wu, K. (1991). Gender recognition from speech. II Fine analysis. *Journal of the Acoustical Society of America*, *90*, 1841-1856.

Coleman, R. O. (1976). A comparison of the contribution of two voice quality characteristics to the perception of maleness and femaleness in the voice. *Journal of Speech and Hearing Research*, *19*, 168-180.

Fant, G. (1966). A note on vocal tract size factors and non-uniform F-pattern scalings. STL-QPSR, 4, 22-30.

Fant, G. (1970). *Acoustic Theory of Speech Production* (2nd ed.). Mouton, Paris: Mouton.

Fant, G. (1975). Non-uniform vowel normalization. STL-QPSR, 2-3, 1-19.

Fitch, W. T., & Giedd, J. (1999). Morphology and development of the human vocal tract: A study using magnetic resonance imaging. *Journal of the Acoustical Society of America*, *106*, 1511-1522.

Flanagan, J. (1972). *Speech Analysis and Perception* (2nd ed.). New York and Berlin: Springer-Verlag.

Fujisaki, H., & Kawashima, T. (1968). The roles of pitch and higher formants in the perception of vowels. *IEEE Transactions on Audio and Electroacoustics Au-16*, 73-77.

Giles, H., & Powsland, N. F. (1975). *Speech style and evaluation*. New York: Academic Press.

Harding, S., & Cooke, M. P. (2008). Perception of speech properties from extremely brief segments. *Journal of the Acoustical Society of America*, 123, 3724.

Harding, S., Cooke, M., & König, P. (2008). Auditory gist perception: an alternative to attentional selection of auditory streams? In *Attention in Cognitive Systems* Paletta, L. and

Rome, E (eds), Lecture Notes in Artificial Intelligence 4840 Springer, Berlin/Heidelberg, 399-416.

Hillenbrand, J. M., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of Americam English vowels. *Journal of the Acoustical Society of America*, 97, 3099-3111.

Hillenbrand, J. M., & Clark, M. J. (2009). The role of f_0 and formant frequencies in distinguishing the voices of men and women. *Attention, Perception & Psychophysics*, *71*, 1150-1166.

Howard, D. M., & Angus, J. (2001). *Acoustics and psychoacoustics*. Oxford: Focus Press.

Huber, J. E., Stathopoulos, E. T., Curione, G. M., Ash, T., & Johnson, K. (1999). Formants of children, women and men: The effects of vocal intensity variation. *Journal of the Acoustical Society of America*, *106*, 1532-1542.

Ingemann, F. (1968). Identification of the speaker's sex from voiceless fricatives. *Journal of the Acoustical Society of America*, *44*, 1142-1144.

Johnson, K. (2005). Speaker Normalization in Speech Perception. In D. B. Pisoni and R. E. Remez (Eds), *The handbook of speech perception* (pp. 363-389). Molden, MA: Blackwell.

Kawahara, H., Masuda-Kasuse, I., & de Cheveigne, A. (1999). Restructuring speech representations using pitch-adaptive time-frequency smoothing and instantaneous-frequency-

based F0 extraction: Possible role of repetitive structure in sounds. *Speech Communication, 27*, 187-207.

Kawahara, H., and Irino, T. (2004). Underlying principles of a high-quality speech manipulation system STRAIGHT and its application to speech segregation. In P. Divenyi (Ed.), *Speech Separation by Humans and Machines* (pp. 67-180) Massachusetts: Kluwer Academic.

Krause, R. M., Freyberg, R., & Morsella, E. (2002). Inferring speakers' physical attributes from their voices. *Journal of Experimental Social Psychology, 38,* 618-625.

Ladefoged, P., & Broadbent, D. E. (1957). The information conveyed by vowels. *Journal of the Acoustical Society of America, 39,* 98-104.

Lass, N. J., Hughes, K. R., Bowyer, M. D., Waters, L. T., & Bourne, V. T. (1976). Speaker sex identification from voiced, whispered, and filtered isolated vowels. *Journal of the Acoustical Society of America, 59,* 675-678.

Liu, C., & Kewley-Port, D. (2004). STRAIGHT: a new speech synthesizer for vowel formant discrimination. *Acoustics Research Letters Online, 5*, 31-36.

Macmillan, N. A., & Creelman, C. D. (1991). *Detection Theory: A User's Guide*. Cambridge: Cambridge University Press.

Miller, J. D. (1989). Auditory-perceptual interpretation of the vowel. *Journal of the Acoustical Society of America, 85,* 2114-2134.

Murray, I. R., & Arnott, J. L. (1993). Toward the simulation of emotion in synthetic speech – A review of the literature on human vocal emotion. *Journal of the Acoustical Society of America, 93,* 1097-1108.

Nordström, P.-E., & Lindbolm, B. (1975). A normalization procedure for vowel formant data. *Proceedings of the 8th International Congress of Phonetic Sciences*, Leeds, UK.

Owren, M. J., Berkowitz, M., & Bachorowski, J.-A. (2007). Listeners judge talker sex more efficiently from male than from female vowels. *Perception and Psychophysics,* 69, 930-941.

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America, 24,* 175-184.

Potter, R. K., & Steinberg, J. C. (1950). Toward the specification of speech. *Journal of the Acoustical Society of America, 22,* 807-820.

Robinson, K., & Patterson, R. D. (1995). The duration required to identify the instrument, the octave, or the pitch chroma of a musical note. *Music Perception,* 13, 1-15.

Sachs, J., Lieberman, P., & Erikson, D. (1972). Anatomical and cultural determinants of male and female speech. In R. W. Shuy & R. W. Fasold (Eds.), *Language attitudes: Current trends and prospects* (pp. 74-84). Washington, DC: Georgetown University Press.

Schwartz, M. F. (1968). Identification of speaker sex from isolated voiceless fricatives. *Journal of the Acoustical Society of America*, *43*, 1178-1179.

Schwartz, M. F., & Rine, H. E. (1968). Identification of speaker sex from isolated, whispered vowels. *Journal of the Acoustical Society of America*, *44*, 1736-1737.

Smith, D. R. R., & Patterson, R. D. (2005). The interaction of glottal-pulse rate and vocal-tract length in judgements of speaker size, sex, and age. *Journal of the Acoustical Society of America*, *118*, 3177-3186.

Smith, D. R. R., Patterson, R. D., Turner, R., Kawahara, H., & Irino, T. (2005). The processing and perception of size information in speech sounds. *Journal of the Acoustical Society of America*, *117*, 305-318.

Smith, D. R. R., Walters, T. C., & Patterson, R. D. (2007). Discrimination of speaker sex and size when glottal-pulse rate and vocal-tract length are controlled. *Journal of the Acoustical Society of America*, *112*, 3628-3639.

Smith, D. R. R. (2010). Time to judge sex of speaker: effect of glottal-pulse rate and vocal-tract length. In *Proceedings of the 20th International Congress on Acoustics ICA2010*, Sydney, Australia, 355.

Suen, C. Y., & Beddoes, M. P. (1972). Discrimination of vowel sounds of very short duration. *Perception and Psychophysics, 11*, 417-419,

Titze, I. R. (1989). Physiologic and acoustic differences between male and female voices. *Journal of the Acoustical Society of America, 85*, 1699-1707.

Traunmüller, H. (1981). Perceptual dimension of openness in vowels. *Journal of the Acoustical Society of America, 69*, 1465-1475.

Turner, R. E., Walters, T. C., Monaghan, J. J. M., & Patterson, R. D. (2009). A statistical, formant-pattern model for segregating vowel type and vocal-tract length in developmental formant data. *Journal of the Acoustical Society of America, 125*, 2374-2386.

Wakita, H. (1977). Normalization of vowels by vocal-tract length and its application to vowel identification. *IEEE Transactions on Acoustics, Speech and Signal Processing, ASSP, 25*, 183-192.

Whiteside, S. P. (1998). Identification of a speaker's sex from synthesized vowels. *Perceptual and Motor Skills, 86*, 595-600.

Wichmann, F. A., & Hill, N. J. (2001). The psychometric function: I. Fitting, sampling, and goodness of fit. *Perception and Psychophysics, 63*, 1293-1313.

Wu, K., & Childers, D. G. (1991). Gender recognition from speech. I Coarse analysis. *Journal of the Acoustical Society of America, 90*, 1828-1840.

Figure Captions

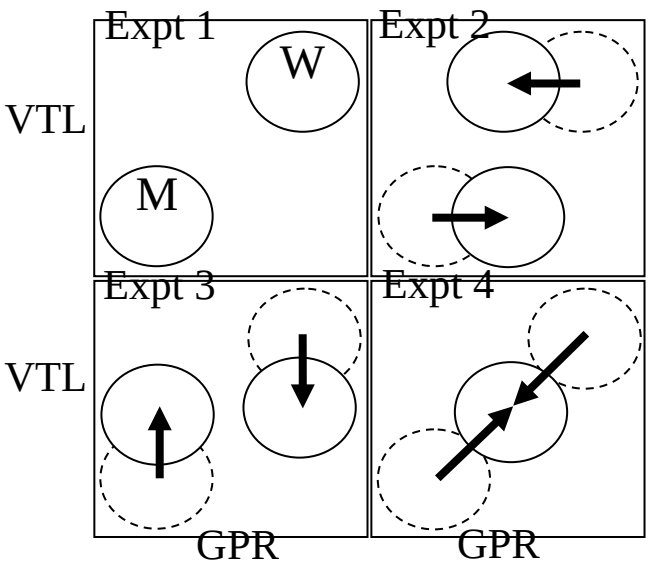
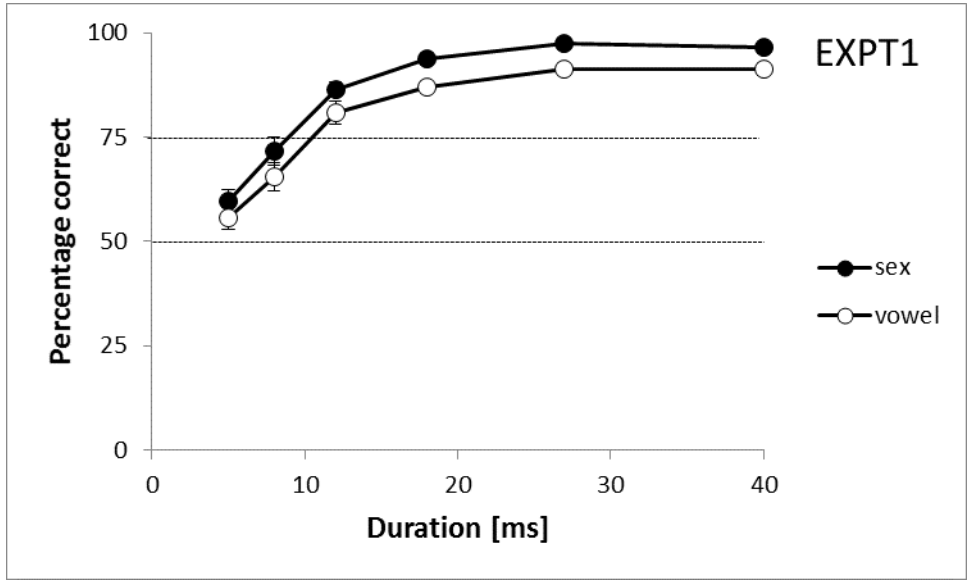
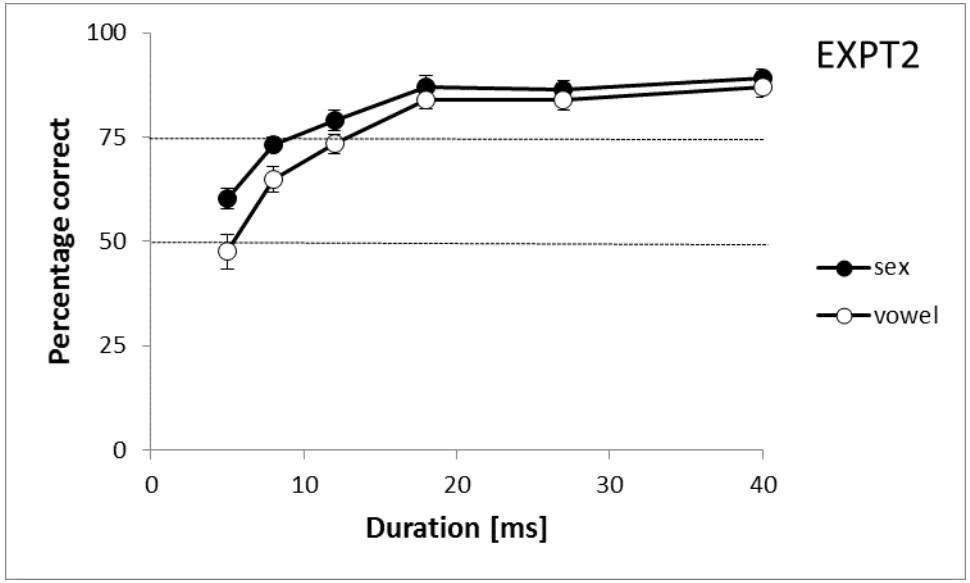
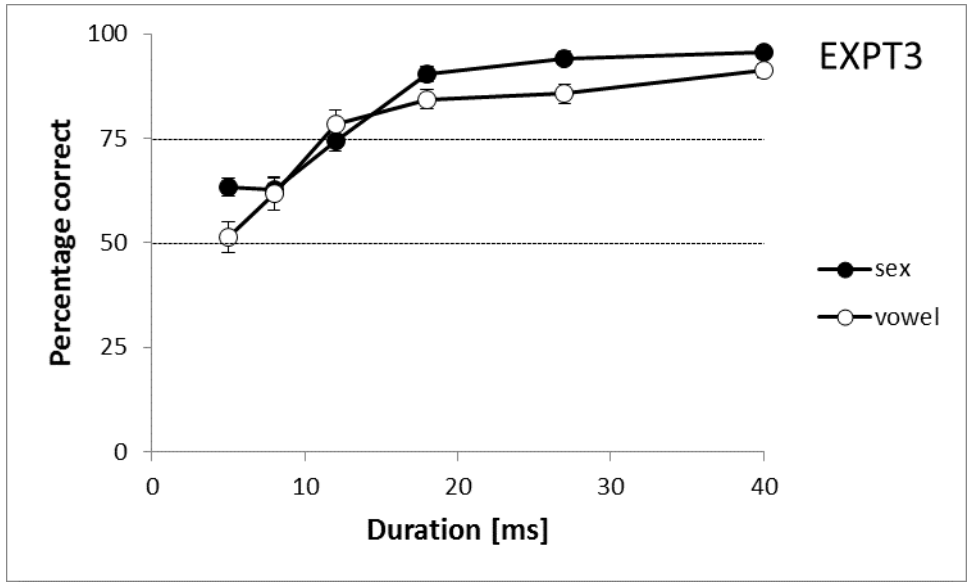


FIG 1. Schematic of glottal-pulse rate (GPR) and vocal-tract length (VTL) manipulations in Experiments 1—4. Men's vowels (M) tend to be located in the bottom-left corner (low GPR and long VTL) and women's vowels (W) tend to be located in the top-right corner (high GPR and short VTL) of the GPR-VTL plane. Dashed circles represent the original location of non-manipulated men's and women's vowels. Arrows indicate manipulations of GPR and VTL. Experiment 1: No manipulation. Experiment 2: The GPR of the men's and women's vowels were modified to the same intermediate value (equal to the geometric mean). Experiment 3: The simulated VTL of the men's and women's vowels were modified to the same intermediate value (equal to the geometric mean). Experiment 4: Both the GPR *and* the simulated VTL of the men's and women's vowels were modified to the same two intermediate values (equal to the geometric means of the men's and women's GPR and estimated VTL).







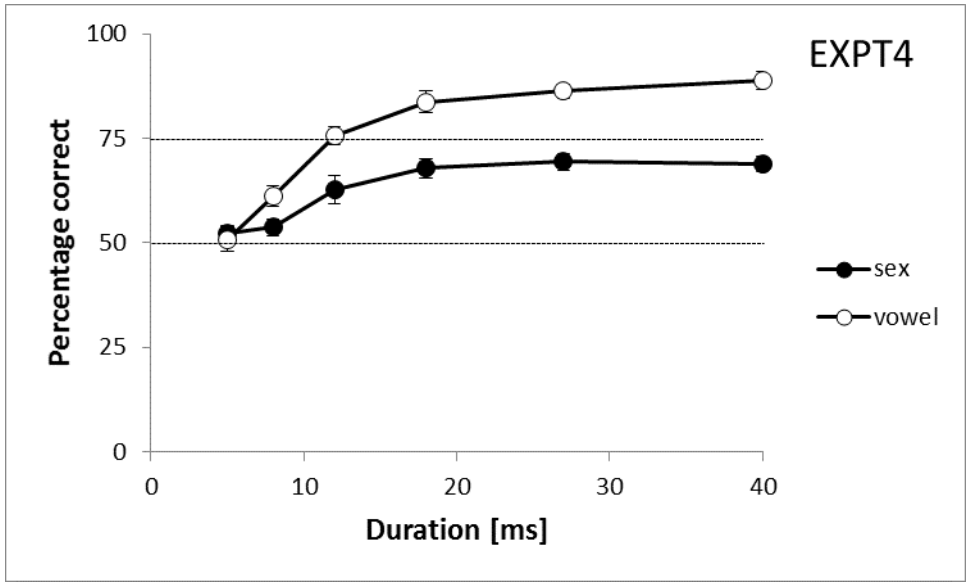


FIG. 2. Percentage correct judgement of original speaker sex and vowel recognition, as a function of vowel duration, for the four experimental manipulations. See Fig. 1 for the schematic representation of the experimental manipulations for each experiment. Data collapsed across correct judgements of both men and women speakers, and across all five vowels. Each point shown for each duration is based on 600 trials [(25 men + 25 women speaker repetitions) X 12 listeners]. Error bars are standard error of the mean across the twelve listeners. The dotted line at 50% shows the threshold point for reliable vowel recognition (d'=1 in a 5AFC task). The dotted line at 75% shows the threshold point for reliable discrimination of speaker sex (d'=1 in a 2AFC task).

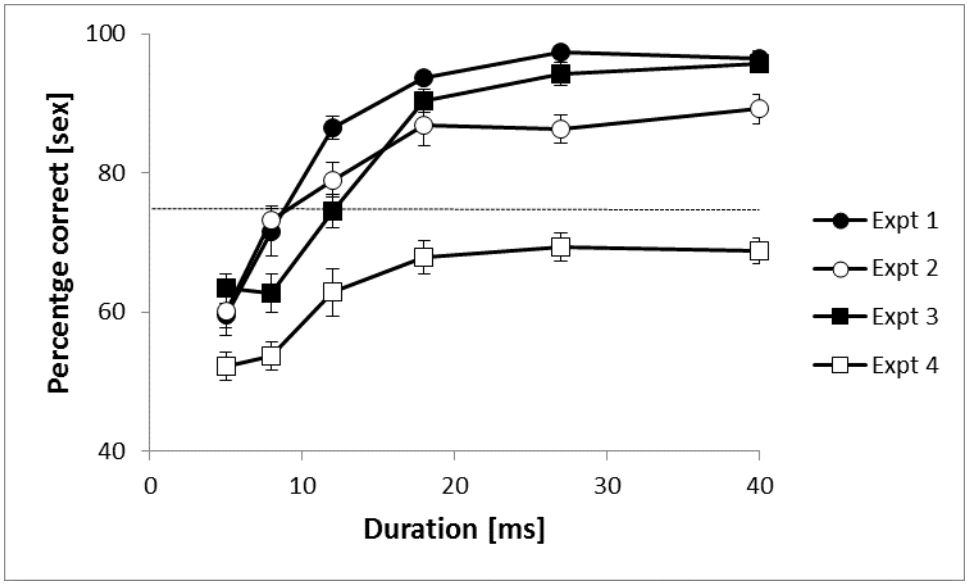


FIG. 3. Percentage correct judgement of original speaker sex, as a function of vowel duration, for Experiments 1—4. See Fig. 1 for the schematic representation of the experimental manipulations for each experiment. Data collapsed across correct judgements of both men and women speakers, and across all five vowels. Each point shown for each duration is based on 600 trials [(25 men + 25 women speaker repetitions) X 12 listeners]. Error bars are standard error of the mean across the twelve listeners. The dotted line at 75% shows the threshold point for reliable discrimination of speaker sex (d'=1 in a 2AFC task)

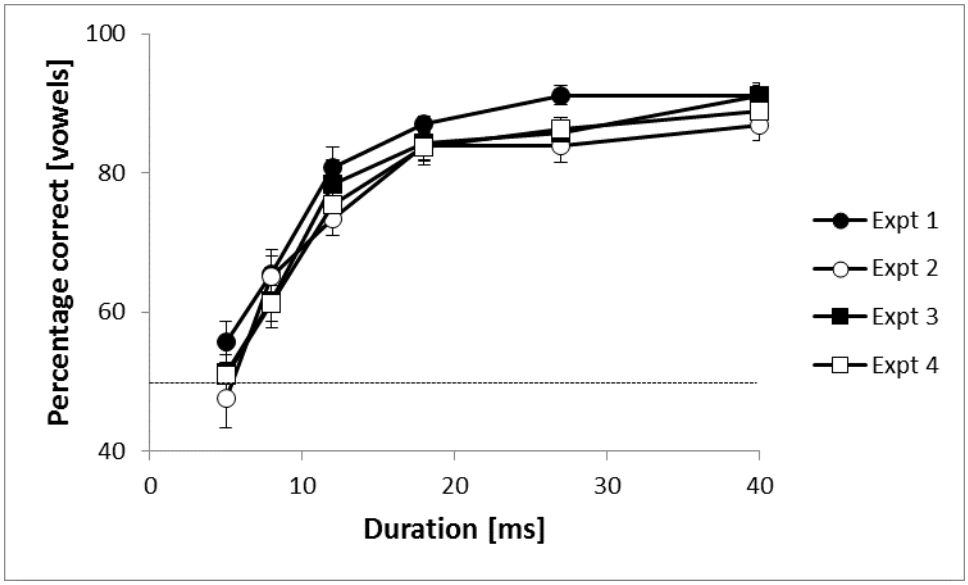


FIG. 4. Percentage correct vowel recognition, as a function of vowel duration, for Experiments 1—4. See Fig. 1 for the schematic representation of the experimental manipulations for each experiment. Data collapsed across correct judgements of both men and women speakers, and across all five vowels. Each point shown for each duration is based on 600 trials [(25 men + 25 women speaker repetitions) X 12 listeners]. Error bars are standard error of the mean across the twelve listeners. The dotted line at 50% shows the threshold point for reliable vowel recognition (d'=1 in a 5AFC task)

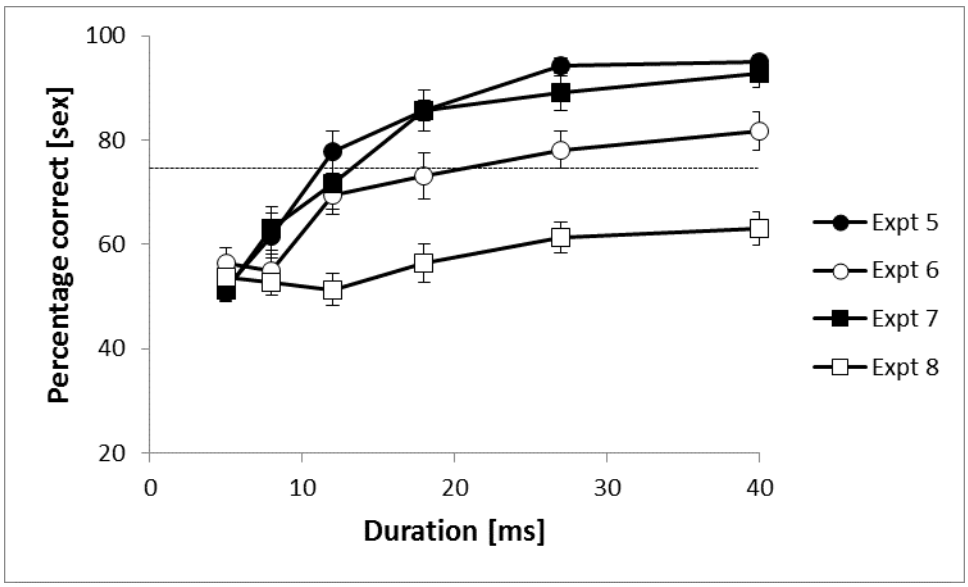


FIG. 5. Percentage correct judgement of original speaker sex, as a function of vowel duration, for Experiments 5—8 (addition of offset noise mask). Each point shown for each duration is based on 300 trials [(15 men + 15 women speaker repetitions) X 10 listeners]. For all other details see Fig. 3.

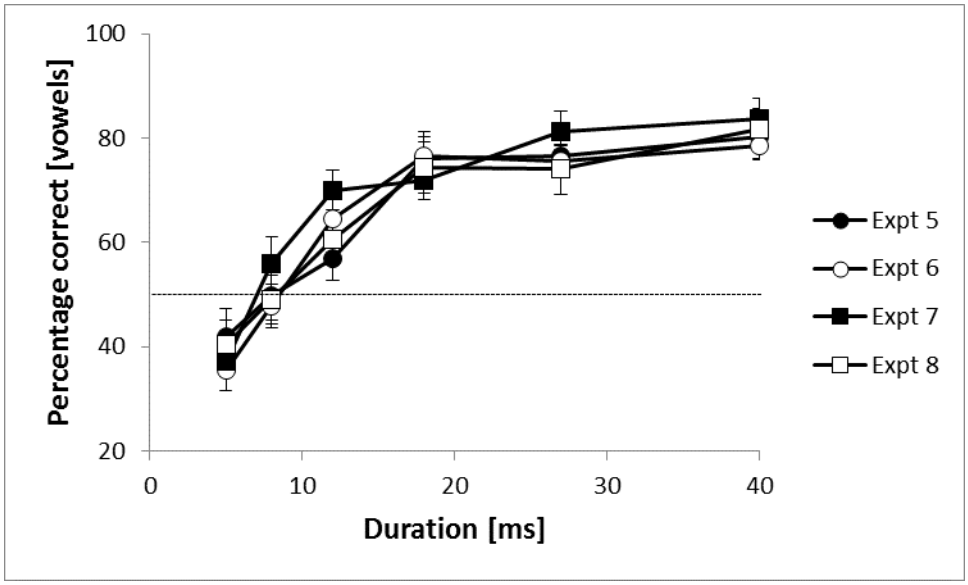


FIG. 6. Percentage correct vowel recognition, as a function of vowel duration, for Experiments 5—8 (addition of offset noise mask). Each point shown for each duration is based on 300 trials [(15 men + 15 women speaker repetitions) X 10 listeners]. For all other details see Fig. 4.

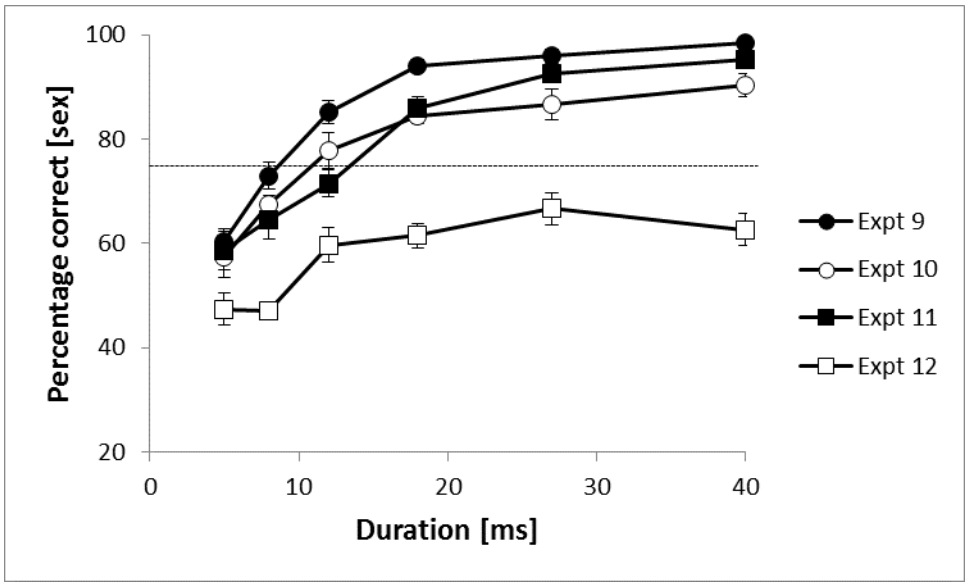


FIG. 7. Percentage correct judgement of original speaker sex, as a function of vowel duration, for Experiments 9—12 (increased variability of men—women pairings). Each point shown

for each duration is based on 270 trials [(15 men + 15 women speaker repetitions) X 9 listeners]. For all other details see Fig. 3.

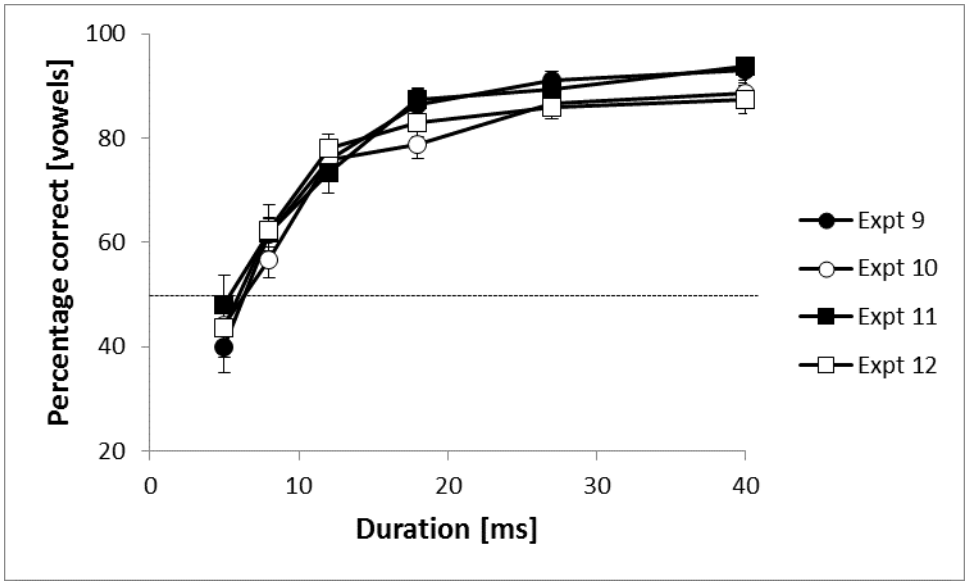


FIG. 8. Percentage correct vowel recognition, as a function of vowel duration, for Experiments 9—12 (increased variability of men—women pairings). Each point shown for each duration is based on 270 trials [(15 men + 15 women speaker repetitions) X 9 listeners]. For all other details see Fig. 4.

TABLE I. Physical and acoustic variables of the eight speakers.

Speaker	Age (yr)	Height (cm)	GPR[a] (Hz)	VTL[b] (cm)
Man 1	21	185	95	16.32
Man 2	22	175	94	15.50
Man 3	21	176	103	15.58
Man 4	29	175	99	15.54
Woman 1	35	169	150	14.57
Woman 2	21	163	223	14.03
Woman 3	21	157	166	13.48
Woman 4	21	160	182	13.78

[a]Average across the five vowels. [b]Estimated using VTL averages for men and women from Fitch and Giedd (1999), scaled by known average adult heights for men of 1750 mm and women of 1612 mm (NHS Health Survey England, 2004), assuming linear scaling between VTL and height (Turner, Walters, Monaghan and Patterson, 2009).