Proceedings of 20th International Congress on Acoustics, ICA 2010

23-27 August 2010, Sydney, Australia

Time to judge sex of speaker: effect of glottal-pulse rate and vocal-tract length

David R. R. Smith

Dept Psychology, University of Hull, Cottingham Rd, Hull, HU6 7RX, United Kingdom; d.r.smith@hull.ac.uk

PACS: 43.71.Bp

ABSTRACT

When listening to someone's voice: what stimulus duration is required to tell whether the person speaking is a man or a woman; what are the acoustic cues in speech that influence such judgements; and how does manipulations in these acoustic cues influence such judgements? The vowels of five men and five women were recorded and played at a number of brief durations. The vowels were either unmodified and thus differed in both glottal-pulse rate and vocal-tract length (Expt 1), or had their glottal-pulse rate modified to be the same (Expt 2), or had their simulated vocal-tract length modified to be the same (Expt 3). Listeners were required to indicate whether the vowels were spoken by a man or woman. Results show that correct speaker-sex judgement requires only brief duration stimuli (about 10 ms), and that the removal of either the glottal-pulse rate or the vocal-tract length cue leads to reduced performance in judging the sex of the original speaker.

INTRODUCTION

The voices of men and women sound different from each other. Key perceptual differences are caused by differences in the length and mass of the vocal folds (Titze, 1989) and the length of the vocal tract (Fant, 1970; Fitch and Giedd, 1999). If we take the recorded voice of a man (or woman) we can alter whether we hear a woman (or man) by manipulating the simulated glottal-pulse rate and vocal-tract length (e.g., Smith, Walters and Patterson, 2007).

This paper investigates how long it takes to acquire information about the sex of a speaker and what acoustic cues are important for making that decision. Specifically, when listening only to someone's voice: what vowel duration is required to reliably tell whether the person speaking is a man or a woman; what are the acoustic cues in speech that influence such judgements, and how does manipulations in acoustic cues influence such judgements?

METHOD

Listeners were presented isolated vowels recorded from ten different speakers (five adult men and five adult women). The vowels were either not touched (Expt 1), had their glottal-pulse rate (GPR) modified to be the same (=geometric mean of the mens' and womens' GPR, Expt 2), or had their simulated vocal-tract length (VTL) modified to be the same (=geometric mean of the mens' and womens' estimated VTL, Expt 3). The vowels were presented at six different durations (8, 12, 18, 27, 40 and 60 ms).

A. Stimuli

Examples of the five English vowels (/a/, /e/, /i/, /o/, /u/) of five adult men and five adult women were recorded using a high-quality microphone (Shure SM58-LCE), with a sampling rate of 48 kHz and a 16-bit amplitude resolution. The microphone was connected to a preamp (Xenyx Behringer 502) to boost the signal before recording through the PC soundcard. Speakers were required to utter the vowels at a regular relaxed rate at a comfortable effort level. For each speaker, one example best vowel was selected of each type. Details of the physical and acoustic characteristics of the speakers are shown in Table 1.

Table 1. Physical and acoustic variables of ten 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	20	179	106	15.90
Man 5	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	165	180	14.16
Woman 5	21	160	182	13.78

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

Walters, Monaghan and Patterson, 2009)

The scaling of the vowels was performed by STRAIGHT (Kawahara, Masuda-Kasuse and de Cheveigne, 1999). See Smith et al (2005) for a description of how STRAIGHT is used to scale vowels to simulate different speakers, and Kawahara and Irino (2004) for the underlying principles. In Expt 1, the GPR and simulated VTL of the five men and five women's vowels were not manipulated. In Expt 2, the GPR of the men and women's vowels was modified to be the same (= geometric mean of the men and women's GPR). Thus, the vowels of man 1 and woman 1 were set to have a GPR of 119 Hz (= √(95•150)). Similar manipulations were performed for the GPR of the other four men and women pairs. The simulated VTL was not manipulated in Expt 2. In Expt 3, the simulated VTL of the men and women's vowels was modified to be the same (=geometric mean of the men 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)). Similar manipulations were performed for the simulated VTL of the other four men and women pairs. The GPR was not manipulated in Expt 3. Figure 1 shows a schematic of these three types of manipulation.

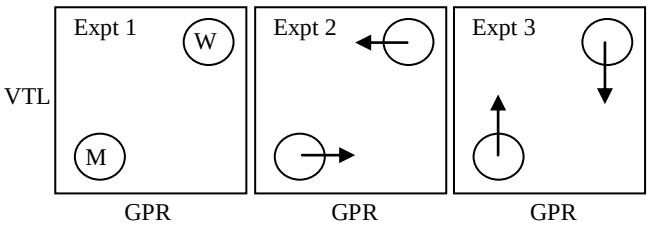

Figure 1. Schematic of GPR and VTL manipulations. The men speakers (M) occupy the bottom-left corner (low GPR and long VTL) and the women speakers (W) occupy the top-right corner (high GPR and short VTL) of the GPR-VTL plane. Arrows indicate manipulations of GPR and VTL. Expt 1, no manipulation. Expt 2, the GPR of the men and women's vowels were modified to be the same (= geometric mean). Expt 3, the simulated VTL of the men and women's vowels were modified to be the same (= geometric mean).

The duration of all vowels was adjusted to have six different durations (8, 12, 18, 27, 40 and 60 ms) within STRAIGHT. Finally, all the vowel sounds were normalized to the same rms level.

The stimuli were played by a 24-bit sound card and presented to the listener diotically over Sennheiser HD600 headphones. The sound level of the vowels at the headphones was ~70 dB SPL.

B. Procedures

The experiments were performed using a single-interval, single-response paradigm. The listener heard a vowel of a given duration and had to indicate whether a man or a woman had spoken the vowel. There was a 50% chance that either a man or a woman had spoken the original vowel. The judgement of the sex of the speaker was made by selecting the appropriate button on a visual display. The order of the 'man' and 'woman' buttons was pseudo-randomly switched at the beginning of any run.

Listeners were first given a practice run of 20 trials with a set duration of 100 ms. Half of the trials were vowels spoken by men and half of the trials were vowels spoken by women. Listeners found it an easy task to judge the sex of the speaker at this duration. Listeners were then given a run of 300 trials, consisting of six durations (8, 12, 18, 27, 40, 60 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). The durations were presented in a pseudo-random order. Which of the five men, or five women, or five vowels, was used in any one trial was pseudo-randomly determined. There was no feedback. Each run of 300 trials took approximately 15 min to complete.

There were two sets of listeners. One set of listeners did Expt 1 and Expt 2 (counterbalanced for order). A second set of listeners did Expt 1 and Expt 3 (counterbalanced for order). Each listener did their two experiments in one session lasting approximately 45 min. Ten listeners did Expt 1 and Expt 2. Twelve listeners did Expt 1 and Expt 3. All listeners were volunteers. All reported normal hearing.

RESULTS

The results averaged across each set of listeners are presented in Figs 2 and 3. Fig. 2 shows percent correct judgement of original speaker sex as a function of duration of the vowel when both GPR and VTL cues are available (Expt 1), and when the GPR cue is modified to be the same (Expt 2), leaving VTL as a cue to speaker sex. The results are pooled across both men and women speakers. Each datum point for each duration is based on 500 trials [(25 men + 25 women) x 10 listeners].

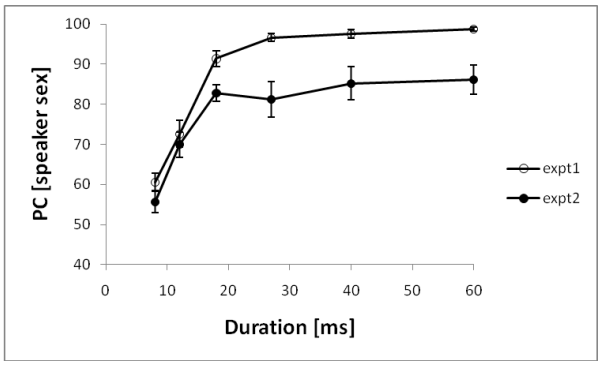

Figure 2. Percent correct judgement of speaker sex (collapsed across correct judements of both men and woman speakers), as a function of duration, when both GPR and VTL cues are present [Expt 1 open circle symbols] and when GPR has been equalized to be the same (= geometric mean) [Expt 2 solid circle symbols]. Error bars are standard error of the mean across the ten listeners.

Looking at Fig. 2, we can see that at low durations (8 and 12 ms), there is little difference between performance between Expt 1 and Expt 2. However, at durations of 18 ms and longer there is a reduction in performance when having only one cue (Expt 2), as compared to when having two cues (Expt 1).

Fig. 3 shows percent correct judgement of original speaker sex as a function of duration of the vowel when both GPR and VTL cues are available (Expt 1), and when the VTL cue is modified to be the same (Expt 3), leaving GPR as a cue to speaker sex. The results are pooled across both men and women speakers. Each datum point for each duration is based on 600 trials [(25 men + 25 women) x 12 listeners].

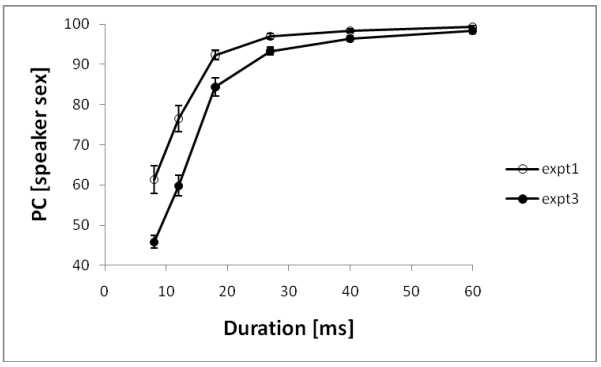

Figure 3. Percent correct judgement of speaker sex (collapsed across both correct judements of men and woman speakers), as a function of duration, when both GPR and VTL cues are present [Expt 1 open circle symbols] and when VTL has been equalized to be the same (= geometric mean) [Expt 3 solid circle symbols]. Error bars are standard error of the mean across the twelve listeners.

Looking at Fig. 3, we can see that at low durations upto 27 ms, there is a reduction in performance when having only one cue (Expt 3), as compared to when having two cues (Expt 1). However, at longer durations of 40 and 60 ms there is little difference between performance.

The data from Expt 1 was collapsed across the two sets of listeners (cf. Figs 2 and 3), giving 1100 trials at each duration [(25 men trials + 25 women trials) x 22 listeners]. Best-fitting psychometric functions were fitted to the percent correct speaker-sex scores in Expt 1 using a maximum-likelihood method (Wichmann and Hill, 2001). The point at which listeners can reliably tell whether a man or woman spoke – the duration threshold for reliable discrimination – was taken to be the 75% point on this fitted curve. When listeners have access to unmodified voices as in Expt 1, the vowel duration needs to be around 11 ms, before the listener can reliably tell whether a man or woman spoke the original vowel.

Best-fitting psychometric functions were also fitted to the data from Expt 2 and Expt 3. 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 (=geometric mean), but still leaving VTL as a potential cue (Expt 2, cf. Fig. 1 and Fig. 2), the listener needs a vowel duration of around 14 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 (=geometric mean), but still leaving the GPR as a potential cue (Expt 3, cf. Fig. 1 and Fig. 3), the listener needs a vowel duration of around 15 ms to reliably tell whether a man or woman spoke the original vowel.

DISCUSSION

When listeners can access GPR and VTL cues to speaker sex they can tell whether a man or woman spoke a vowel at stimulus durations as short as 11 ms (Expt 1, cf. Fig. 1 for schematic and Figs 2-3 for data). This is in close agreement with Harding and Cooke (2008) and Owren, Berkowitz and Bachorowski (2007), who variously estimate it at between about 10 and 15 ms for experiments similar to Expt 1. It is clear that the acoustic information in speech relating to speaker sex can be extracted from very short duration stimuli indeed. This agrees with the idea that many characteristics of the auditory scene are extracted very rapidly (Harding, Cooke and König, 2008). At stimulus durations shorter than about 10 ms, performance approaches chance levels of 50%. At stimulus durations greater than about 25 ms, performance reaches 100%.

Reducing the number of available cues to speaker sex by either equalizing GPR (Expt 2, cf. Fig. 1), or equalizing VTL (Expt 3, cf. Fig. 1), leads to reduced performance in judging the sex of the original speaker. Compared to when having both GPR and VTL (Expt 1), the stimulus duration required to reliably tell whether the original speaker was a man or woman increases to about 15 ms.

Though judgement of original speaker-sex performance is impaired with the loss of either GPR or VTL cue, the pattern of impairment across duration is different. With the equalization of GPR (Expt 2, Fig. 2) performance is impaired at durations of 18 ms and longer. For the 8 and 12 ms duration there is no difference between having both GPR and VTL cues (Expt 1), and having only the VTL cue (Expt 2). This is presumably because at the shortest durations there is no pitch cue available – speaker-sex performance is determined by the only cue available 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 and Patterson, 1995). The impaired performance in Expt 2 at stimulus durations of 18 ms and longer highlights the importance of GPR as a cue to speaker sex. With the equalization of simulated VTL (Expt 3, Fig. 3) performance is impaired for durations upto 27 ms but not for durations longer than this. This would suggest that performance in Expt 3 is impaired at very low durations because of the loss of VTL as a reliable cue and the relative weakness of the available GPR cue. However, at durations beyond about 30 ms, pitch arises as a strong perceptual cue. It can then be used to support speaker-sex performance levels at the same errorless level as having both GPR and VTL cues.

The suggestion is that to determine speaker sex information in very brief duration vowel sounds, the listener uses VTL-related perceptual cues (frequencies of the formants) to distinguish men from women. However, at the point at which the percept is available then the listener switches to increasingly using GPR-related perceptual cues (pitch). Speculatively, when constructuring 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 progresses. This maximises performance in a rapidly changing dynamic environment.

SUMMARY AND CONCLUSIONS

Listeners were presented with very brief duration vowels (8—60 ms) that were spoken by either men or women. Listeners were required to judge whether a man or woman had spoken the vowels. In Expt 1, the vowels were untouched. In Expt 2, the GPR of the vowels of the men and women were modified to be the same (= geometric mean). In Expt 3, the simulated VTL of the men and women were modified to be the same (= geometric mean). The results show that it takes a stimulus duration of about 11 ms to tell whether a man or woman spoke a vowel when listeners have access to both GPR and VTL cue information to speaker sex (Expt 1, Figs 2 and 3). When either GPR or VTL is equalized, performance is impaired such that a stimulus duration of about 15 ms is required to tell whether a man or woman spoke a vowel. When the GPR cue information is equalized whilst leaving VTL as a cue (Expt 2, Fig. 2), performance is impaired at stimulus durations of 18 ms and above. However, performance is unimpaired at very short stimulus durations of 8 and 12 ms. When the VTL cue is equalized whilst leaving GPR as

a cue (Expt 3, Fig. 3), performance is impaired at stimulus durations of 27 ms and below. However, performance is unimpaired at stimulus durations of 40 and 60 ms.

REFERENCES

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

Fitch, W. T. and Giedd, J. (1999) "Morphology and development of the human vocal tract: A study using magnetic resonance imaging" *J. Acoust. Soc. Am*. **106**, 1511–1522.

Harding, S. and Cooke, M. P. (2008) "Perception of speech properties from extremely brief segments" *J. Acoust. Soc. Am*. **123**, 3724.

Harding, S., Cooke, M. and 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.

Kawahara, H., Masuda-Kasuse, I. and de Cheveigne, A. (1999) "Restructuring speech representations using pitch-adaptive time-frequency smoothing and instantaneous-frequency-based F0 extraction" *Speech Comm*. **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 Speech separation by humans and machines Divenyi, P. (ed.), Kluer Academic, Massachusetts, 167–180.

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

Robinson, K. And 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.

Smith, D. R. R., Patterson, R. D., Turner, R., Kawahara, H. and Irino, T. (2005) "The processing and perception of size information in speech sounds" *J. Acoust. Soc. Am*. **117**, 305–318.

Smith, D. R. R., Walters, T. C. And Patterson, R. D. (2007) "Discrimination of speaker sex and size when glottal-pulse rate and vocal-tract length are controlled" *J. Acoust. Soc. Am*. **122**, 3628–3639.

Titze, I. R. (1989) "Physiologic and acoustic differences between male and female voices" *J. Acoust. Soc. Am*. **85**, 1699–1707.

Turner, R. E., Walters, T. C., Monaghan, J. J. M. and Patterson, R. D. (2009) "A statistical, formant-pattern model for segregating vowel type and vocal-tract length in developmental formant data" *J. Acoust. Soc. Am*. **125**, 2374–2386.

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