INTERSPEECH 2009 BRIGHTON

The role of glottal pulse rate and vocal tract length in the perception of speaker identity

Etienne Gaudrain[1], *Su Li, Vin Shen Ban, Roy D. Patterson*[2]

Centre for the Neural Basis of Hearing,
Department of Physiology, Development and Neuroscience,
University of Cambridge, United-Kingdom
[1]epg22@cam.ac.uk, [2]rdp1@cam.ac.uk

Abstract

In natural speech, for a given speaker, vocal tract length (VTL) is effectively fixed whereas glottal pulse rate (GPR) is varied to indicate prosodic distinctions. This suggests that VTL will be a more reliable cue for identifying a speaker than GPR. It also suggests that listeners will accept larger changes in GPR before perceiving speaker change. We measured the effect of GPR and VTL on the perception of a speaker difference, and found that listeners hear different speakers given a VTL difference of 25%, but they require a GPR difference of 45%.

Index Terms: speaker identity, glottal pulse rate, vocal tract length

1. Introduction

Glottal pulse rate (GPR) and vocal tract length (VTL) are two fundamental voice characteristics. Recent studies have shown that VTL and GPR largely determine the perceived size of a speaker [1] and whether the speaker is a man, woman or child [1, 2]. Thus, these vocal parameters play an important role in determining speaker identity, and listeners are likely to track them in multi-speaker environments.

Ives *et al.* [3] have observed that the just noticeable difference (JND) for a change in VTL is about 5% (for consonnant-vowel sequences). For GPR, Smith *et al.* [4] found that the JND is about 2% (for sequences of vowels). So, the JND for VTL is more than double that for GPR. Darwin *et al.* [5] reported results consistent with this difference for a task involving identification of words in concurrent sentences using the Coordinated Response Measure task[1]. They found similar levels of performance when the changes in VTL were about 1.4 times larger than the changes in GPR. More recently, Vestergaard *et al.* [6] tested the effect of changes in GPR and VTL on the identification of concurrent syllables. They found that to yield the same syllable identification performance, the changes in VTL have to be about 1.6 times larger than the changes in GPR. The tasks employed in the latter two papers involve energetic masking and the results are correlated with the salience of the GPR and VTL cues as might be expected.

However, the relative value of GPR and VTL cues in identification tasks contrasts with the variations of these two parameters within the speech of a single speaker. Kania *et al.* [7] observed, in natural speech, that the standard deviation for GPR is about 3.7 semitones. The VTL, on the other hand, only varies by about one JND, *i.e.* about 1 semitone [*e.g.* 8]. Thus, in speaker differentiation VTL is a more reliable cue than GPR, although listeners are more sensitive to a difference in GPR. Kuwabara and Takagi [9] reported that listeners were able to recognise a familiar speaker in 50% of the trials when the GPR was changed by 4.5 semitones. A change of less than one semitone in VTL was enough to yield the same recognition score.

The purpose of the present study was to compare the roles of GPR and VTL in the definition of speaker identity, and to determine listeners' expectations about the variation in GPR and VTL for a single speaker. To ensure that the listeners were making a speaker similarity judgement, as opposed to simple discrimination, they were presented two short sequences of syllables and asked: "Is it possible that both sequences were uttered by the same speaker?"

2. Method

2.1. Participants

Five male and five female students, aged 20-21, were paid an hourly wage to take part in this experiment. Their hearing threshold was measured at 0.5, 1 and 4 kHz, and were invariably found to be below 15 dB-HL.

2.2. Stimuli

The stimuli were 65 CV syllables; 5 vowels /a, e, i, o, u/ paired with 13 consonants /b, d, f, g, h, k, l, m, n, p, r, s, t/. The syllables were spoken by a single speaker in a steady-state manner, *i.e.* with limited fluctuation in GPR and intensity. The original syllables had a duration of about 500 ms; they were trimmed to be 200 ms preserving the onset and offset shapes. The RMS level of these sounds was adjusted to be the same for all the syllables. The syllables were then processed with STRAIGHT [10] to set their duration to 250 ms, and to effect changes in GPR and VTL. The resulting syllables were presented in sequences of three, separated by 50 ms silence. In a sequence, the GPR and the VTL followed a contour randomly selected from among the following option: rising, falling, down-across-up, up-across-down. The step-size for the contours was 0.5 semitone for both GPR and VTL. The contours were centred around zero semitone. In the experiment the sequences were presented in pairs, in a random order: a sequence from a reference speaker and a sequence from a comparison speaker. The two sequences always had different contours and different syllables. The two intervals were separated by a silence of 500 ms.

Five different reference speakers (shown in Fig. 1) were used to cover the part of the GPR-VTL plane occupied by hu-

[1]Two sentences of the form "Ready *call sign* got to *color number* now" are presented simultaneously. The task is to report the color and number pair for the sentence containing the "Baron" call sign.

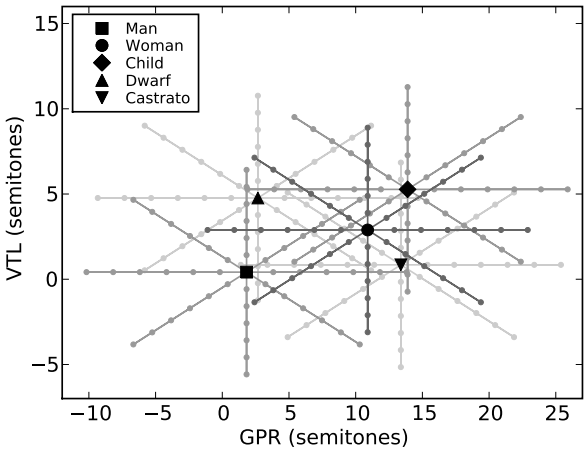

Figure 1: *The five reference speakers that were created, represented in the GPR-VTL plane. The axes are in semitones* re *the original speaker. Each reference speaker is surrounded by its comparison speakers along eight spokes.*

mans: Man, Woman, Child, Dwarf and Castrato. Both dimensions are expressed in semitones relative to the average value for the original speaker: GPR *re* 120 Hz, and VTL *re* 155.4 mm. The reciprocal of VTL was used so that shorter VTLs produce a greater numerical value in semitones, consistent with an increase in frequency. Each reference speaker was surrounded by 48 comparison speakers arranged along eight spokes inscribed in an ellipse that had a radius of 12 semitones along the GPR axis, and a radius of 6 semitones along the VTL axis (these values were chosen after pre-testing on a wider range). Each reference speaker was compared to the 48 comparison speakers and itself ten times, yielding a total of 2450 comparisons per subject. To prevent the participants becoming too familiar with the reference speakers, a rove was applied to the spoke pattern for each trial. The rove magnitude was determined with a Gaussian noise having a standard deviation 0.5 semitone for the GPR, and 1 semitone for the VTL.

2.3. Apparatus

The sequences for the trials were generated off-line for each listener separately, using Matlab and STRAIGHT, and stored in PCM format (24 kHz, 16 bit). The sounds were presented through a SoundBlaster Audigy 2 sound-card and AKG K240DF headphones, while the subject was seated in a double-walled IAC (Winchester, UK) sound-attenuated booth.

3. Results and discussion

3.1. General description

Performance is presented in terms of "the proportion of trials on which listeners judge the two intervals as being uttered by the same speaker". The data points in Fig. 2 show average performance over listeners as a function of the *radial distance* between the comparison speaker and the reference speaker. Since both axes are expressed in semitones, the radial distance is the Euclidean distance:

$$d = \sqrt{\Delta\mathrm{GPR}^2 + \Delta\mathrm{VTL}^2} \quad (1)$$

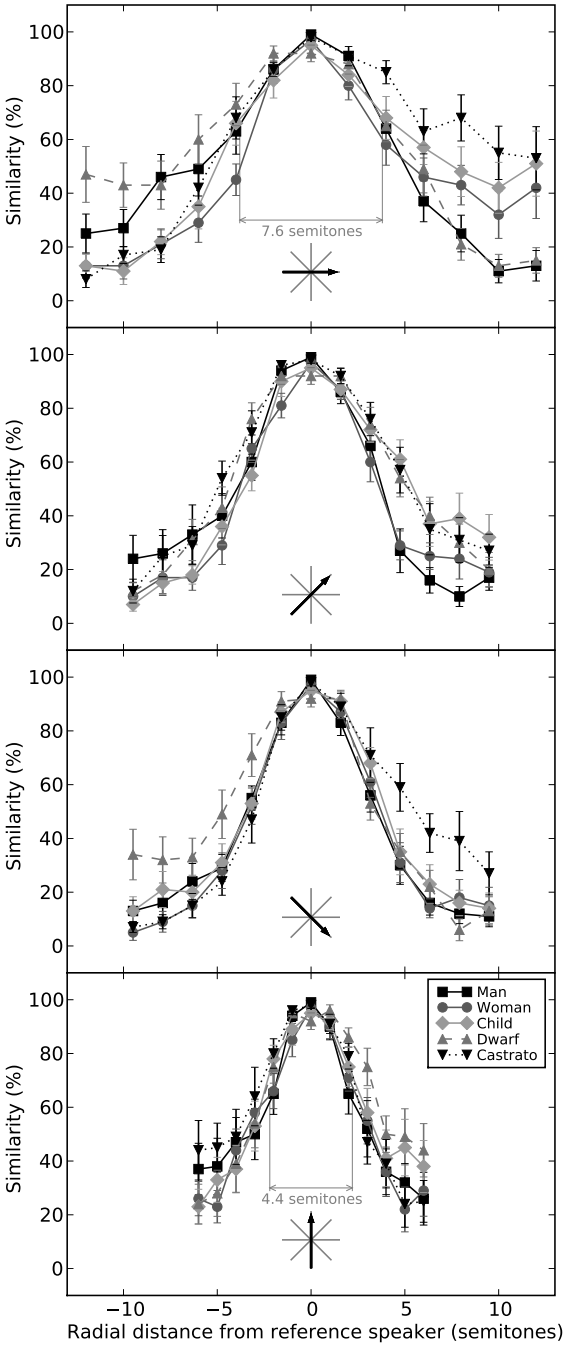

Figure 2: *Similarity judgement as a function of the radial distance from reference speaker. Each panel shows the average proportion of "Yes" answers for a pair of aligned spokes (as depicted by the arrow). Each curve represents a different speaker. For the top and bottom panels, the width of the distribution (as defined by its standard deviation) is indicated by a horizontal arrow. The error bars represent the inter-subject standard error.*

where ΔGPR and ΔVTL are the differences between the speakers along the two dimensions, in semitones. The step-size can also be expressed relative to the just-noticeable difference: a 1 semitone change in GPR is about 2.9 times the JND, a 1 semi-

tone change in VTL is about 1.2 times the JND.

The main result is that the acceptable difference for GPR is about 3.8 semitones (11.1 times the JND) while that for VTL is only 2.2 semitones (2.6 times the JND). Thus, in this experiment, it is experience that governs the decision rather than the JND. Comparable measurements for the "individuality" distribution of Kuwabara and Takagi [9] yield a value of 3.9 semitones for the GPR difference and 0.7 semitones for the VTL difference. In the study of Kania *et al.* [7], the distribution of GPR in running speech (measured by electroglottography) had a standard deviation of 3.7 semitones. Little is known about the natural variation of VTL in running speech but it cannot be very large.

A linear mixed model was used to analyse the results (transformed into rationalized arcsine units [11]). The fixed effects set up in the model were radial distance, spoke and reference speaker and their interactions. The analysis revealed significant effects of radial distance $[F(1, 2311) = 3227.7,\ p < 0.001]$, spoke $[F(7, 2311) = 37.2,\ p < 0.001]$ and reference speaker $[F(4, 2311) = 25.4,\ p < 0.001]$, as well as significant interactions between reference speaker and spoke $[F(28, 2311) = 9.4,\ p < 0.001]$, spoke and radial distance $[F(7, 2311) = 31.9,\ p < 0.001]$, and a three way interaction between all of the factors $[F(28, 2311) = 2.7,\ p < 0.001]$. The interaction between reference speaker and radial distance was not significant $[F(4, 2311) = 0.8,\ p = 0.52]$.

The effect of reference speaker indicates that the participants had somewhat different expectations for different speakers. The interaction with spoke indicates that the expectation did not differ equally on all the spokes. Moreover, the three way interaction indicates that the difference varies along the spoke. As shown in Fig. 2, the interaction with radial distance is probably due to the differences in the tails of the curves. Using pairwise t-tests, the different reference speakers were compared within each spoke. The results of this post-hoc analysis shows that most of the differences between reference speakers take place on spokes pointing out from the centre of the GPR-VTL plane.

3.2. Trading relationship between GPR and VTL

In the concurrent speech study of Darwin *et al.* [5], when both dimensions change, the effect on speech segregation is more than the sum of the effects for the same changes when they occur separately. In the current study, the additivity can be evaluated by comparing the diagonals to the vertical and horizontal axes. The comparison between the data observed along the diagonal and what it would be if it were strictly cumulative is shown in Fig. 3. It shows that the similarity judgements on the diagonal are close to the sum of the GPR and VTL component judgements.

Vestergaard *et al.* [6] also considered the question of additivity and concluded that the trading relationship between GPR and VTL was best represented by the Euclidian distance:

$$\delta_\xi = \sqrt{\xi^2\ \Delta\mathrm{GPR}^2 + \Delta\mathrm{VTL}^2} \qquad (2)$$

The judgement in their experiment was dominated by the JNDs for GPR and VTL and they observed a trading ratio of 1.6, that is, a semitone of GPR had the same effect as 1.6 semitones of VTL. To evaluate the trading ratio ξ for the current experiment, the data were modelled with a two-dimensional cumulative gamma distribution similar to that described by Vestergaard

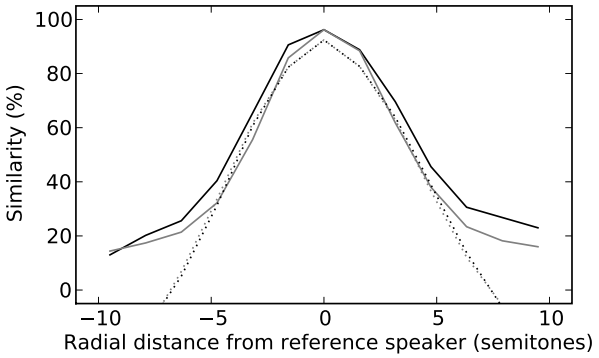

Figure 3: *Similarity judgement along the diagonal spokes (solid lines) and the cumulative judgement for changes in GPR only and VTL only (dotted lines). The cumulative judgement is the sum of the interpolated similarity judgement at the GPR only and VTL only values corresponding to those of the diagonal. The black lines are for the diagonal pointing toward the upper-right corner of the GPR-VTL plane, and the grey lines are for the diagonal pointing toward the lower-right corner.*

et al. [6]. In this model, the similarity judgement, $p_r(\delta_\xi)$, depends solely on δ_ξ and has the form:

$$p_r(\delta_\xi) = \alpha + \frac{1-\alpha}{\beta} \int_0^{\delta_\xi} x^{k-1}\, \frac{\mathrm{e}^{-x/\theta}}{\theta^k \Gamma(k)}\, dx \qquad (3)$$

The similarity judgement is approximately cumulative when k is close to 3. The model was fitted, using the least mean squares method, to evaluate the five parameters: the offset and scaling parameters α and β, the two parameters of the gamma distribution k and θ, and most importantly the GPR-VTL trading ratio ξ. The best fitting model had a trading ratio $\xi = 0.60$ and a $k = 3.4$. The trading ratio means that, for the voice similarity judgement, a change of 1 semitone in GPR is equivalent to a change of 0.6 semitone in VTL – virtually the opposite of what was observed for the syllable recognition judgement.

3.3. Influence of the boundaries of the GPR-VTL plane

The residues of the radial-scale model are displayed in Fig. 4. The parabolic shape of the residue with respect to both axes suggests that the judgement is also affected by the relative position of the pair of speakers in the plane. This hypothesis is supported by observation that the similarity judgement is closer to indecision on the edges of the GPR-VTL plane. A complementary model involving the *absolute* GPR and VTL values can be layered on top of the current model as follow:

$$p_a(\mathrm{GPR}, \mathrm{VTL}) = \lambda_G(\mathrm{GPR} - \mu_G)^2 + \lambda_V(\mathrm{VTL} - \mu_V)^2 \qquad (4)$$

$$p(\delta_\xi, \mathrm{GPR}, \mathrm{VTL}) = p_r(\delta_\xi) + p_a(\mathrm{GPR}, \mathrm{VTL}) \qquad (5)$$

When this more complete model is fitted, $\xi = 0.61$, $k = 2.7$ and the correlation coefficient, r, rises from 0.90 to 0.94.

The post-hoc analyses above indicate that near the boundaries of the GPR-VTL plane, listeners' judgements become less confident (more variable). The improvement in the fit of the speaker similarity model with the inclusion of absolute position information confirms that the judgements were influenced by the boundaries of the GPR-VTL plane. There could be several different forms of explanation for this boundary effect involving the fidelity of stimulus manipulation or degree of experience

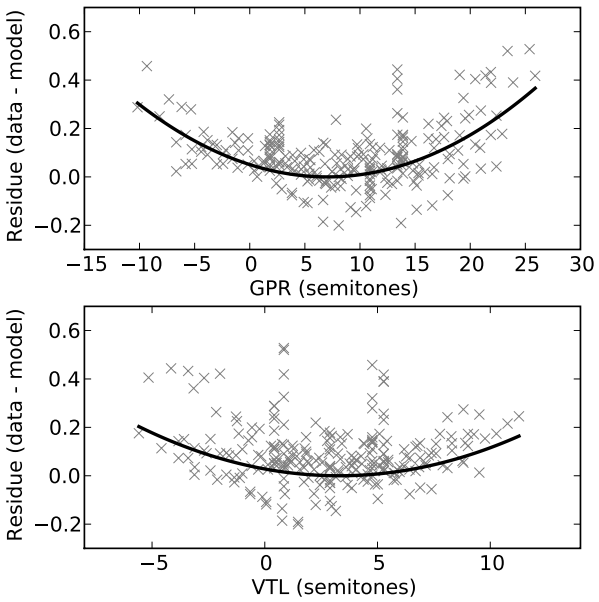

Figure 4: *Difference between the measured similarity and the one predicted by the model along the GPR axis (*upper panel*) and along the VTL axis (*lower panel*). The crosses* × *represent the difference for each data point. The black lines represent quadratic fittings of the residues as obtained in the second model* $p'(\delta_\xi)$.

with extreme GPR and VTL values. The most promising explanation, however, involves the process whereby the voices are categorized as men, women, children, etc. A speaker having a lower GPR than a large male has to be a male, and this judgement may make it more difficult conclude that the voice represents a different speaker. When the change goes in the other direction, the voice is likely to be categorized as a women, and so be more readily judged as different.

4. Conclusions

The trading ratio observed in the current experiment is effectively the inverse of the one observed in recognition and discrimination experiments [3–6]. Vestergaard *et al.* [6] found that for concurrent syllable segregation, GPR was about 1.6 times *more* effective than VTL. In the present study, GPR was found to be about 1.6 times *less* important than VTL in speaker similarity judgements, when measured on the same semitone scale. This makes it clear that the identity judgement is not determined by sensitivity to differences in GPR or VTL. Instead, the judgement seems to be based on experience concerning the variation of GPR and VTL in natural speech, namely, the fact that GPR is much more variable than VTL for a given speaker.

The listeners also seemed to extrapolate the knowledge acquired with common speakers to speakers they have not encountered previously. There is, however, a tendency for the judgement to became less confident closer to the boundaries of the GPR-VTL plane.

5. Acknowledgement

This work was supported by the UK Medical Research Council (G9900369 and G0500221).

6. References

[1] Smith, D. R. R. and Patterson, R. D. (**2005**), "The interaction of glottal-pulse rate and vocal-tract length in judgements of speaker size, sex, and age," J. Acoust. Soc. Am. **118**:3177–3186.

[2] 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.

[3] Ives, D. T., Smith, D. R. R., and Patterson, R. D. (**2005**), "Discrimination of speaker size from syllable phrases," J. Acoust. Soc. Am. **118**:3186–3822.

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

[5] Darwin, C. J., Brungart, D. S., and Simpson, B. D. (**2003**), "Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers," J. Acoust. Soc. Am. **114**(5):2913–2922.

[6] Vestergaard, M. D., Fyson, N. R. C., and Patterson, R. D. (**2009**), "The interaction of vocal tract length and glottal pulse rate in the recognition of concurrent syllables," J. Acoust. Soc. Am. **125**:1114–1124.

[7] Kania, R. E., Hartl, D. M., Hans, S., Maeda, S., Vaissiere, J., and Brasnu, D. F. (**2006**), "Fundamental frequency histograms measured by electroglottography during speech: a pilot study for standardization," J. Voice **20**:18–24.

[8] Chuenwattanapranithi, S., Xu, Y., Thipakorn, B., and Maneewongvatana, S. (**2008**), "Encoding emotions in speech with the size code. A perceptual investigation," Phonetica **65**:210–30.

[9] Kuwabara, H. and Takagi, T. (**1991**), "Acoustic parameters of voice individuality and voice-quality control by analysis-synthesis method," Speech Commun. **10**:491–495.

[10] 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*, edited by P. L. Divenyi (Kluwer Academic, Massachusetts), pp. 167–180.

[11] Studebaker, G. (**1985**), "A "rationalized" arcsine transform," J. Speech Hear. Res. **28**:455–462.