The effect of spectral tilt on size discrimination of voiced speech sounds

Toshie Matsui[1], *Toshio Irino*[1], *Kodai Yamamoto*[1], *Hideki Kawahara*[1], *and Roy D. Patterson*[2]

[1]Graduate School of Systems Engineering, Wakayama University, Japan
[2]Department of Physiology, Development, and Neuroscience, University of Cambridge, UK
[1]{tmatsui, irino, kawahara}@sys.wakayama-u.ac.jp, [2]rdp1@cam.ac.uk

Abstract

A number of studies, with either voiced or unvoiced speech, have demonstrated that a speaker's geometric mean formant frequency (MFF) has a large effect on the perception of the speaker's size, as would be expected. One study with unvoiced speech showed that lifting the slope of the speech spectrum by 6 dB/octave also led to a reduction in the perceived size of the speaker. This paper reports an analogous experiment to determine whether lifting the slope of the speech spectrum by 6 dB/octave affects the perception of speaker size with voiced speech (words). The results showed that voiced speech with high-frequency enhancement was perceived to arise from smaller speakers. On average, the point of subjective equality in MFF discrimination was reduced by about 5%. However, there were large individual differences; some listeners were effectively insensitive to spectral enhancement of 6 dB/octave; others showed a consistent effect of the same enhancement. The results suggest that models of speaker size perception will need to include a listener specific parameter for the effect of spectral slope.

Index Terms: speaker size perception, voiced speech sounds, speech spectrum slope

1. Introduction

We can recognize the vowels pronounced by children, women, and men despite the large differences in their heights and despite the differences in vowel waveforms that follow from the height differences. Irino and Patterson [1] explained this apparent scale invariance using a normalized version of the Auditory Image Model [2]. The vocoder STRAIGHT [3, 4] was used to manipulate the perceived Vocal Tract Length (VTL) of several speech databases and demonstrate that listeners can make accurate judgements of differences in speaker size based on STRAIGHT's manipulation of VTL, and they can do so for both voiced and unvoiced speech sounds [5, 6, 7]. Over the course of these studies, it was noticed that the whispered speech of a given speaker gave the impression of a smaller speaker than the voiced speech of the same speaker [7]. Whispered speech has proportionately more high frequency energy than voiced speech (approximately +6 dB/octave) because of the way turbulent noise excites the vocal tract [8]. These observations suggested that spectral tilt will affect the perception speaker size for a given vowel with a fixed set of formant frequencies.

Yamamoto et al. [9] performed a perceptual study to document the effect of spectral tilt on speaker-size perception using synthetic, unvoiced speech sounds. STRAIGHT spectrograms were extracted from a database of voiced words and then resynthesized using two noise sources, one with a flat spectrum and one where the higher frequencies were enhancement +6 dB/octave. The former were referred to as "unvoiced" words and the latter as "whispered" words because it was generally agreed that the latter sounded more like whispered speech. On average, listeners judgments indicated that they did hear the whispered words as coming from a smaller speaker than the unvoiced words. However, there were large individual differences; while the psychometric functions of some listeners were shifted to smaller VTLs by the spectral tilt, those of other listener were not. Yamamoto et al. [9] constructed a computational model of size discrimination based on the dynamic, compressive gammachirp (dcGC) auditory filterbank and showed that it could explain the effect of spectral tilt on the perception of speaker size with unvoiced and whispered speech sounds.

This paper reports a perceptual experiment designed to determine whether the effect of spectral slope observed by Yamamoto et al. extends to voiced speech sounds, that is, whether enhancing the spectrum of voiced words by +6 dB/octave leads to a reduction in the perception of the speaker size.

2. Size discrimination experiment

The effect of spectral tilt on size discrimination with voiced speech was measured for eight listeners using a two-alternative, forced-choice procedure (2AFC) with the method of constant stimuli. Psychometric functions were fitted to the size discrimination data to determine the point of subjective equality (PSE) and the just noticeable difference (JND) for speaker size, which were, in turn, used to evaluate the effect of spectral tilt size discrimination.

2.1. Stimuli

The speech sounds were words drawn from a well known database of four-morae Japanese words (FW03) [10]. Morae are subunits of words somewhat similar to syllables [11]. The words in the database are controlled with respect to both word familiarity and phonetic balance and they were spoken naturally. The words for the experiment were selected from recordings of a male speaker (mya) who had an average glottal pulse rate (GPR) close to 150 Hz and a (geometric) mean formant frequency (MFF) of 1278 Hz. The words were chosen from a list with high familiarity ratings (list Nos. 3 and 4, each containing 1000 words).

The size information in the words was scaled by TANDEM-STRAIGHT [4]. There are three stages to the vocoding process: (1) analysis of the original utterance into a TANDEM-STRAIGHT, smoothed power spectrogram, (2) scaling of the frequency dimension of the spectrogram (which alters the effective VTL, and inversely the MFF), and (3) resynthesis of the utterance with the desired MFF. During resynthesis the spectro-temporal envelope was excited with a regular stream of glottal pulses. These resynthesized voiced words are designated "original" (Or) words. The entire set was, then, differentiated in time to produce "emphasized" (Em) versions of the words whose spectral slope was elevated by 6 dB/octave.

2.2. Procedure

On each trial, the listener was presented with two intervals both of which contained two 4-morae words randomly selected from the database (without replacement). The MMF was the same for the two words of a given interval. The first interval always contained Or words; the second interval contained either Or or Em words with equal probability. In Or-Or trials, between one interval and the next, the listeners were comparing voiced words with the same spectral tilt; in Or-Em trials, they were comparing voiced words with different spectral tilts. The listener's task was to choose the interval with the smaller speaker in all cases.

The MFF was varied between the first and second intervals to generate psychometric functions that show the effect of MFF on the perception of speaker size (see Fig. 1). Or-Or and Or-Em psychometric functions were generated for five "reference" speakers whose words had the combinations of GPR and MFF shown by the green X's in Fig. 1a; the GPR and MFF values are expressed as ratios of the original speaker's average GPR (150 Hz) and average MFF (1278 Hz). The MFF ratios used to simulate speakers with different sizes had the MFF ratios shown by the small circles in Fig. 1a. For the five reference speakers, the combinations of MFF ratio and GPR ratio were (1) 0.84 and 0.5, (2) 0.84 and 2.0, (3) 1.12 and 1.0, (4) 1.5 and 0.5, and (5) 1.5 and 2.0). In these units, the MMF and GPR ratios for the original speech sounds are 1.0. To generate the psychometric functions for each reference speaker, we prepared six comparison speakers whose MFF ratios were set to $2^{-5/12}, 2^{-3/12}, 2^{-1/12}, 2^{1/12}, 2^{3/12}, 2^{5/12}$ (open circles above and below each green X in Fig. 1a.

The trials required to generate the 5 Or-Or and 5 Or-Em psychometric functions that together constitute the complete experiment were presented interleaved, so the perception of speaker size was varying over the full range throughout the experiment. A 0.5 s silence was inserted between the two intervals on each trail. There was no feedback during the main experiment.

The total number of trials per listener was 1200 (5 reference speakers $\times$ 6 comparison speakers $\times$ 2 counterbalancing orders of presentation $\times$ 2 types of spectral tilt $\times$10 replications). Since there were two words per interval, a total of 4800 words were presented to each listener in the experiment. They were selected at random without replacement from the lists in the FW03 database with the highest familiarity (list Nos. 3 and 4), each of which contained 1000 words.

The words were presented over headphones (Sennheiser HD-580) at a 48-kHz sampling rate to listeners seated in a sound attenuated room. The rms level was A-weighted SPL 70 dB on average. The headphones were calibrated by sound level meter Type 2250-L (Bruël & Kjær) and artificial ear Type 4153 (Bruël & Kjær). The sound level of the individual words was roved (or varied randomly) over a 3-dB range to discourage listeners from basing their size judgments on sound level.

2.3. Training

In order to familiarize the listeners with size judgment, the listeners were given extensive training which began with trials in which the MFF difference between the intervals was large and all of the speakers were limited to the set associated with one reference speaker. The training sessions were based on words from list Nos. 3 and 4 uttered by a different male speaker (mis) in the FW03 database. Training with large MFF differences continued until performance reached a criterion level of 90% correct. Then the MFF difference was gradually reduced and finally in the session with the smallest MFF differences training continued performance reached a criterion level was 80% correct. Each training session contained 20 trials constructed with 2 comparison speakers $\times$ 2 counterbalanced orders of presentation $\times$ 5 trials. Finally, there were two training runs that had the same structure as the runs from the main experiment. These sessions included 60 trials with 5 reference speakers $\times$ 6 comparison speakers $\times$ 2 counterbalanced orders of presentation. The final training run continued until performance reached a criterion level of 85% correct for two successive sessions. Throughout the training, feedback was provided as to whether they had correctly identified the smaller speaker. The duration of the training session was 5 hours on average.

2.4. Listeners

Eight Japanese listeners (two male and six females between 21 and 24 years of age) participated in the recognition experiment after giving informed consent. They all had normal hearing thresholds between 125 and 8000 Hz. This experiment was approved by the local ethics committee of Wakayama University and was in accordance with the ethical standards stated in the Declaration of Helsinki.

3. Results and discussion

Or-Or (red) and Or-Em (blue) psychometric functions for the average data of the eight listeners are presented in Fig. 1(b) with a separate panel for each of the five reference speakers. The layout of the figure mirrors that of panel (a) used to define the stimulus conditions. The abscissa for the psychometric function is MFF ratio relative to that of the original speaker; the ordinate is the percentage of trials on which the comparison interval was identified as having the smaller speaker. The error bars show ± 1 standard deviation across listeners. The psychometric functions are cumulative Gaussians fitted to the data with a bootstrap method [12]; these same functions were used to calculate the JND, which is the rise in MFF ratio for a 26% increase in performance from 50% to 76%. The value is shown by the inset in each panel of the figure; it corresponds to a d' of 1 in this 2AFC task.

3.1. Point of Subjective Equality (PSE)

The PSE of the psychometric functions were calculated as the MMF ratio where the percentage of "test speaker chosen" is 50%. The PSE for the Or-Or condition was invariably close to the MFF ratio of the reference speaker as illustrated by the vertical dotted line in each panel. That is, there is no systematic bias in this judgement—a finding that is consistent with previous studies [5, 6, 7]. The psychometric functions for the Or-Em conditions are located to the left of those of the Or-Or conditions, which shows that the speech sounds with a spectral tilt of +6 dB/octave are typically heard to emanate from smaller speakers than those without a spectral tilt. The average Or-Em PSE was 4.9% smaller than the average Or-Or PSE.

The PSEs of individual listeners were calculated and submitted to an ANOVA with two factors: spectral tilt of the test speaker (Or or Em) and MFF-GPR region of the reference speaker. There was a significant effect for spectral tilt ($F(1, 70) = 46.58, p < 0.0001$) but not for MFF-GPR region ($F(4, 70) = 2.13, p = 0.086$); there was also an interaction between the two factors MFF-GPR ($F(4, 70) = 3.10, p = 0.021$). Multiple comparisons with the Tukey-Kramer HSD test ($\alpha = 0.05$) revealed that PSE does not depend on the MFF-

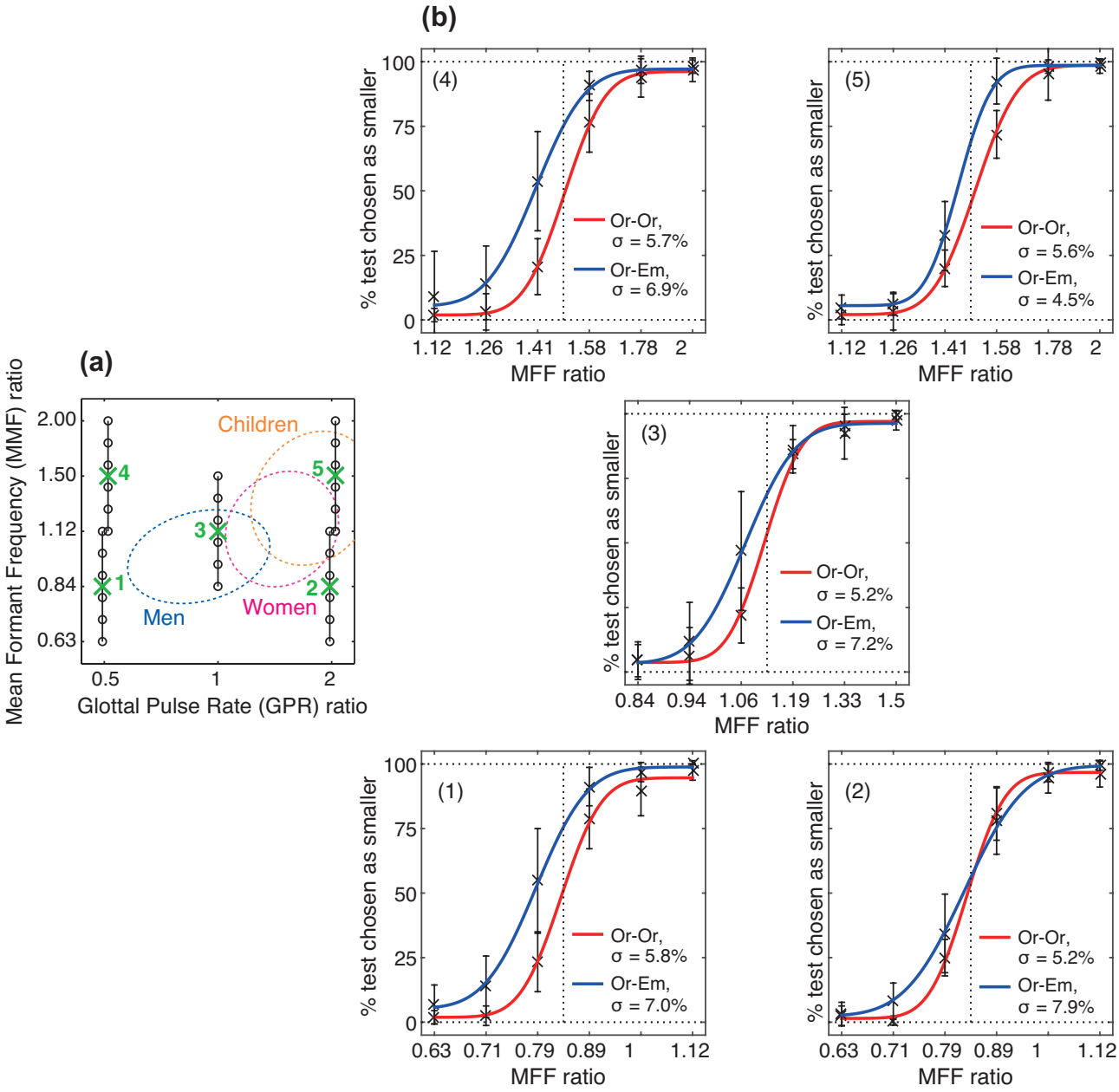

Figure 1: *(a) The MFF-GPR combinations for the stimuli in the MFF discrimination experiment. An MFF ratio of 1 and a GPR ratio of 1 corresponds to the MFF and GPR of the original words (1278 Hz on average, 150 Hz on average, respectively). Five reference speakers are shown by the numbered green crosses; the test speakers are shown by the small open circles. The three dotted ellipses show approximate distributions of MFF and GPR values in normal speech for men, women, and children. Although the lines of (1)(4) and (2)(5) are slightly shifted on the GPR axis in order to make the overlapping part visible, the actual GPR ratios are 0.5 and 2.0, respectively. (b) Psychometric functions for the data associated with the five reference speakers were obtained by fitting cumulative Gaussian distributions to the average data of all eight listeners. The ordinate is the percentage of trials on which the test speaker was chosen as the smaller speaker. The abscissa is the MFF ratio of the comparison speaker. The dashed vertical line shows MFF ratio of the reference speaker. Red curves: Or-Or psychometric functions. Blue curves: Or-Em psychometric functions. Crosses: response rates averaged across the eight listeners. Error bars: ± 1 standard deviation across listeners. σ: JND of MFF in %.*

GPR region in the Or-Or conditions, as would be expected. PSE does depend on MFF-GPR region in the Or-Em condition for regions 1, 3, and 4, but not for regions 2 and 5. Some of the speech sounds in regions 2 and 5 have combinations of GPR and MFF that are rarely encountered in the normal population of speakers. This may have affected the results, but it is not clear solely from the analysis of the PSE values.

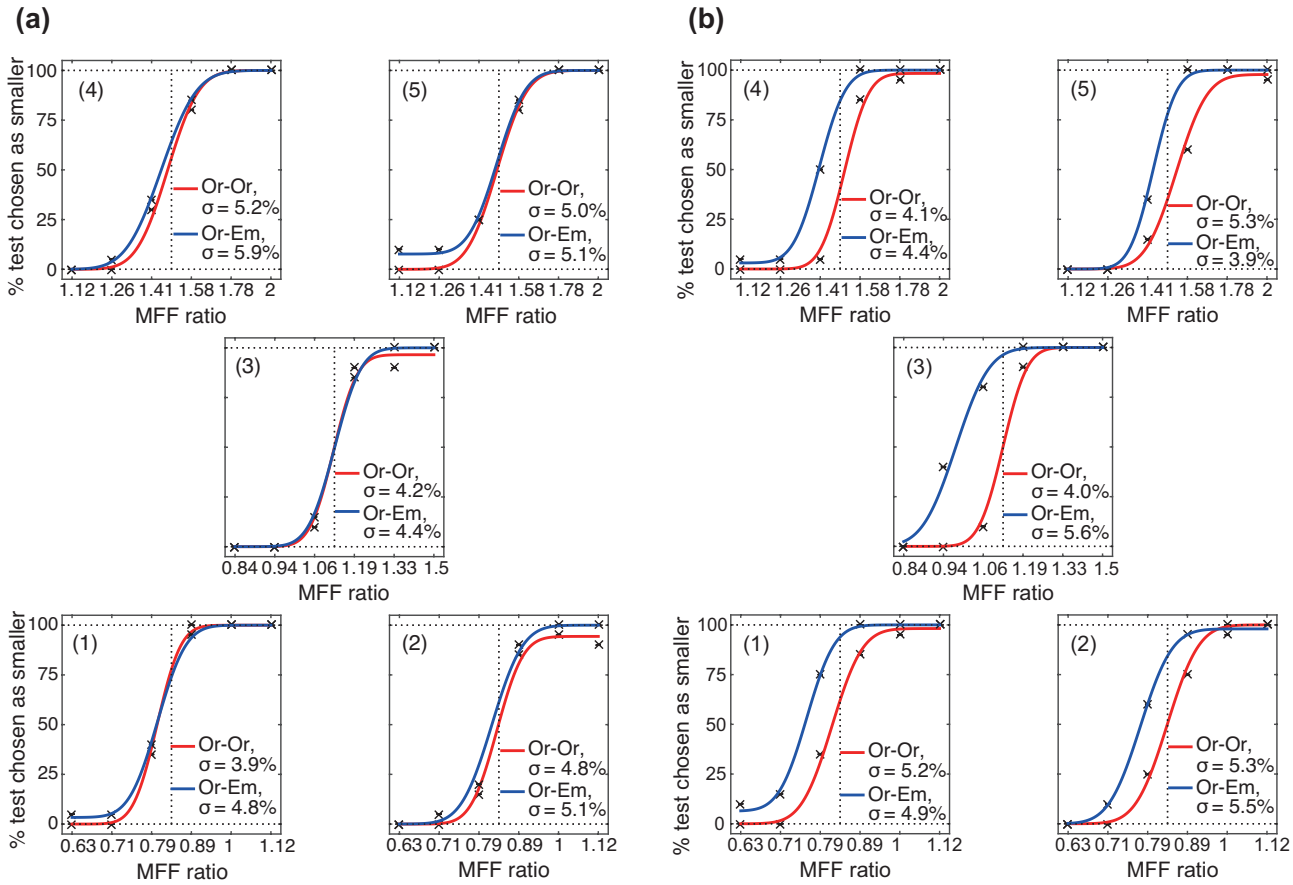

Figure 2: *(a) Listener ET's psychometric functions where PSE is not affected by spectral tilt. (b) Listener HT's psychometric functions where the PSE of the Or-Em psychometric function is shifted toward lower MFFs by spectral tilt. Panels (a) and (b) have the same format as Fig. 1(b).*

3.2. Just Noticeable Difference (JND)

The Or-Or JNDs (σ values for the red lines in Fig. 1b) range from 5.2% to 5.8%; the average is 5.5%, which is only slightly larger than in the previous study [9] where the average was 5.0% and the range was from 4.2% to 5.4%. The Or-Em JNDs (σ values for the blue lines in Fig. 1b) range from 4.5% to 7.9%, with an average value of 5.9%. A two-way ANOVA was performed on the JND values like that performed on the PSE values. Neither factor (MFF-GPR region or spectral tilt) produced a significant effect (MFF-GPR region, $F(4, 70) = 0.88, p = 0.48$; spectral tilt, $F(1, 70) = 1.51, p = 0.22$).

3.3. Effect of spectral tilt for individuals

There were some large individual differences in the shift of the Or-Em psychometric function relative to the Or-Or psychometric function, as in the previous study [9]. The psychometric functions for listener's ET and HT in Figs 2a and 2b illustrate the difference. Whereas there is effectively no effect of spectral tilt on the Or-Em psychometric functions of ET (Fig. 2a), there is a consistent shift of the Or-Em psychometric functions for listener HT (Fig. 2b). This pair of examples is generally indicative of the eight listener's in this experiment; they are either insensitive to the spectral tilt or they show a consistent effect of spectral tilt across the five reference speakers. Thus the effect of spectral tilt with voice speech sounds is similar to that revealed in the previous study [9] using unvoiced and whispered words. This results suggest that models of speaker size perception of voiced speech will also need to include a listener specific parameter for the effect of spectral slope [13].

4. Conclusions

An experiment was performed to determine the effect of high-frequency spectral emphasis on speaker size discrimination with voiced speech. In the average data, spectral enhancement of +6 dB/octave prompted a shift to smaller speakers. However, the effect was found to be listener dependent; some listeners were effectively insensitive to spectral tilt while others showed a consistent shift toward small speaker choices in the presence of spectral enhancement. The results make it clear that any computational model of speaker size perception will need to include a listener specific parameter for spectral slope.

5. Acknowledgements

We thank Aya Ozaki for collecting the experimental data. This research was supported by the JSPS KAKENHI, JP25280063, JP16H01734, JP15H02726, and JP16K12464.

6. References

[1] T. Irino and R. D. Patterson, "Segregating information about the size and shape of the vocal tract using a time-domain auditory model: The stabilised wavelet-Mellin transform," *Speech Communication*, vol. 36, no. 3–4, pp. 181–203, 2002. [Online]. Available: http://dx.doi.org/10.1016/S0167-6393(00)00085-6

[2] R. D. Patterson, M. H. Allerhand, and C. Giguère, "Time-domain modeling of peripheral auditory processing: A modular architecture and a software platform," *The Journal of the Acoustical Society of America*, vol. 98, no. 4, pp. 1890–1894, 1995. [Online]. Available: http://dx.doi.org/10.1121/1.414456

[3] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds," *Speech communication*, vol. 27, no. 3, pp. 187–207, 1999. [Online]. Available: http://dx.doi.org/10.1016/S0167-6393(98)00085-5

[4] H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino, and H. Banno, "TANDEM-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0 and aperiodicity estimation," *Proc. IEEE ICASSP 2008*, pp. 3933–3936, 2008. [Online]. Available: http://dx.doi.org/10.1109/ICASSP.2008.4518514

[5] D. R. Smith, R. D. Patterson, R. Turner, H. Kawahara, and T. Irino, "The processing and perception of size information in speech sounds," *The Journal of the Acoustical Society of America*, vol. 117, no. 1, pp. 305–318, 2005. [Online]. Available: http://dx.doi.org/10.1121/1.1828637

[6] D. T. Ives, D. R. Smith, and R. D. Patterson, "Discrimination of speaker size from syllable phrases," *The Journal of the Acoustical Society of America,* vol. 118, no. 6, pp. 3816–3822, 2005. [Online]. Available: http://dx.doi.org/10.1121/1.2118427

[7] T. Irino, Y. Aoki, H. Kawahara, and R. D. Patterson, "Comparison of performance with voiced and whispered speech in word recognition and mean-formant-frequency discrimination," *Speech Communication*, vol. 54, no. 9, pp. 998–1013, 2012. [Online]. Available: http://dx.doi.org/10.1016/j.specom.2012.04.002

[8] H. Fujisaki and T. Kawashima, "The roles of pitch and higher formants in the perception of vowels," *IEEE Transactions, Audio and Electroacoustics*, vol. 16, no. 1, pp. 73–77, 1968. [Online]. Available: http://dx.doi.org/10.1109/TAU.1968.1161952

[9] K. Yamamoto, T. Irino, R. Nisimura, H. Kawahara, R. D. Patterson, "How the slope of the speech spectrum affects the perception of speaker size," in *Interspeech 2015 – 16th Annual Conference of the International Speech Communication Association Sept., 6–10, Dresden, Germany, Proceedings*, 2015, pp. 1556–1560.

[10] S. Sakamoto, N. Iwaoka, Y. Suzuki, S. Amano, and T. Kondo, "Complementary relationship between familiarity and SNR in word intelligibility test," *Acoustical science and technology*, vol. 25, no. 4, pp. 290–292, 2004. [Online]. Available: http://dx.doi.org/10.1250/ast.25.290

[11] N. Tsujimura, *An Introduction to Japanese Linguistics*, Wiley-Blackwell, Malden, MA, 2007.

[12] F. A. Wichmann, and N. J. Hill, "The psychometric function: I. Fitting, sampling and goodness-of-fit," *Perception and Psychophysics*, vol. 63, no. 8, pp. 1293–1313, 2001. [Online]. Available: http://dx.doi.org/10.3758/BF03194544

[13] T. Irino, E. Takimoto, T. Matsui, and R. D. Patterson, "An auditory model of speaker size perception for voiced speech sounds," in *Interspeech 2017 – 18th Annual Conference of the International Speech Communication Association Aug., 20–24, Stockholm, Sweden, Proceedings*, 2017. Submitted.