INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

An auditory model of speaker size perception for voiced speech sounds

Toshio Irino[1], *Eri Takimoto*[1], *Toshie Matsui*[1], *Roy D. Patterson*[2]

[1]Faculty of Systems Engineering, Wakayama University,
[2]Department of Physiology, Development, and Neuroscience, University of Cambridge
[1]{irino,s185063,tmatsui}@sys.wakayama-u.ac.jp, [2]rdp1@cam.ac.uk

Abstract

An auditory model was developed to explain the results of behavioral experiments on perception of speaker size with voiced speech sounds. It is based on the dynamic, compressive gammachirp (dcGC) filterbank and a weighting function (SSI weight) derived from a theory of size-shape segregation in the auditory system. Voiced words with and without high-frequency emphasis (+6 dB/octave) were produced using a speech vocoder (STRAIGHT). The SSI weighting function reduces the effect of glottal pulse excitation in voiced speech, which, in turn, makes it possible for the model to explain the individual subject variability in the data.

Index Terms: size perception, gammachirp auditory filterbank, stabilized wavelet-Mellin transform, size-shape image

1. Introduction

We hear vowels pronounced by men, women, and children as approximately the same although the vocal tract length (VTL) varies considerably from group to group. At the same time, we can identify the speaker group. Irino and Patterson [1] proposed a computational theory and an algorithm to explain how the auditory system might segregate the acoustic features in speech sounds associated with vocal tract shape from those associated with VTL, and thereby produce an internal representation of speech sounds that is speaker-size invariant. The theory is based on the Stabilized-Wavelet Mellin Transform (SWMT) which is a cascade of a Wavelet transform, image Stabilization, and finally, a Mellin Transform.

STRAIGHT[2, 3] was then used to manipulate the VTL features of natural speech sounds, that is, the (geometric) mean formant frequency, MFF (e.g., Fig.2). It was demonstrated that humans are very good at discriminating speaker size using either voiced[4, 5] or unvoiced[6] speech. The experiments showed that the just noticeable difference (JND) for speaker size is about 7% of VTL or MFF for vowels[4] and about 5% for syllables or words; for comparison, the JND for loudness is about 11%. Two of these studies [4, 6] also showed that speech recognition performance was largely unaffected by speaker size even when it was extended well beyond the normal range.

Recently, Yamamoto et al. [7] performed an experiment to clarify the effect of spectral tilt on size perception using two versions of noise vocoded speech: one had the same spectral profile as the original voiced speech, the other had high-frequency enhancement with a spectral tilt of +6 dB/octave relative to the original speech. The two forms were referred to as "unvoiced" and "whispered" speech sounds since the latter version sounds more like whispered speech than the former. The psychometric functions for speaker-size judgements (e.g., Fig.3) with unvoiced and whispered speech revealed that the effects of the spectral tilt were dependent on the listener. Some listeners' psychometric functions were shifted by the spectral uplift while

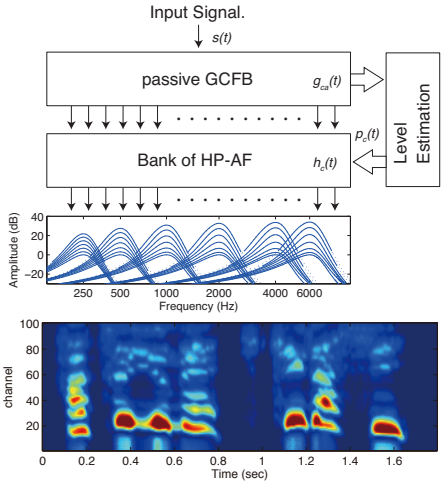


Figure 1: *Auditory spectrogram derived from the dcGC-FB. The variation in filter shape and gain with stimulus level in the the dcGC-FB is illustrated by the six responses associated with each channel; filter gain increases as level decreases from 90 to 30 dB SPL in 10 dB steps.*

others were not. To explain the results, Yamamoto et al. [7] constructed a computational model of size discrimination based on the wavelet stage of the SWMT [1], which was implemented with the dcGC filterbank [8, 9, 10, 11] and used to produce auditory spectrograms of the speech sounds. A block diagram of the dcGC filterbank is shown in the upper part of Fig. 1; it consists of a passive GC filterbank, a bank of high-pass asymmetric filters, and a bank of level estimation units that control the positions of their respective HP-AFs. The frequency responses of the composite filters are shown for six channels and seven input SPLs in the middle panel. The filterbank had 100 channels equally spaced along the quasi-logarithmic $\mathrm{ERB_N}$ axis between 100 Hz and 6000 Hz. The rms value of the output of each channel was calculated with a 25-ms hamming window and the frame step was 5 ms to derive auditory spectrogram in the bottom of Fig. 1. The spectrograms of the vowels show that there is more high-frequency energy in the emphasized speech, and it was argued that this affects the decision of some listeners. The discrimination process is described in Section 3.

Matsui et al. [12] then performed an analogous experiment using voiced speech sounds with and without a spectral tilt of +6 dB/octave, and once again, the results showed that the effect of spectral tilt depended on the listener. However, when the computational model of size discrimination developed for unvoiced and whispered speech sounds was applied to the voiced data, it failed to explain the form of individual listeners' psychometric functions. The problem appeared to be that the stream of glottal pulses in the voiced speech produces harmonic peaks in

http://dx.doi.org/10.21437/Interspeech.2017-196

the auditory spectrum that are large relative to the higher formant peaks and which are somewhat unstable. In this paper, we extend the size perception model of Yamamoto et al. [7] to incorporate aspects of the second, image Stabilization, stage of the SWMT model [1], and we present an extended size-discrimination experiment designed to test the voiced speech version of the size perception model. ASR studies of VTL normalization have related problems with voiced speech sounds, and the solution proposed for the size perception model in this paper would appear to be applicable to traditional VTL normalization algorithms.

2. Size discrimination experiment

The new experiment on size discrimination of voiced speech sounds is similar to those of previous studies [6, 7]; the details of the procedure are described in [12]. The spectral weighting of the stimuli is the same as in the experiments on unvoiced and whispered speech[7]. The main difference in the new experiment is the inclusion of a condition where voiced speech sounds with and without the spectral tilt (+6 dB/octave) were compared within a two-alternative, forced-choice trial. The speech sounds without and with the uplift are hereafter referred to as "original" (Or) and "emphasized" (Em) speech sounds, respectively. The reference speaker was always Or, so the two forms of trial are designated Or-Or and Or-Em.

2.1. Manipulation of MFF and GPR

The size information in the words was scaled using TANDEM-STRAIGHT [13]. There were three stages to the vocoding process: (1) analysis of the original utterance into a TANDEM-STRAIGHT smoothed spectrogram, (2) scaling of the frequency dimension of the spectrogram to manipulate the MFF, and (3) resynthesis of the speech by excitation of the spectro-temporal envelope using one of three glottal pulse rates (GPRs) — 0.5, 1.0, or 2.0 times the original GPR. The experiment was performed in the five regions of GPR-MFF space shown in Fig. 2.

2.2. Results

Average Or-Or and Or-Em psychometric functions for the eight listeners are presented separately in Fig.3 for the five regions of GPR-MFF space. The abscissa for the psychometric functions is "MFF ratio" relative to that of the original speaker; the ordinate is the percentage of trials on which the test interval was identified as having the smaller speaker. A cumulative Gaussian function was fitted to the data of all of the subjects in each condition [14]. The Or-Or psychometric functions (red) are steep, symmetric, and unbiased (that is, centered horizontally on relative MFF=1); the Or-Em psychometric functions (blue) are slightly less steep, effectively symmetric and the mid-point is biased toward lower MFF values than their counterparts. So, at the lower MFF ratios, the test sound with high-frequency enhancement is heard as the smaller speaker, as in previous experiments. The just noticeable difference (JND) in MFF is defined as the interval of MFF ratio over which the ordinate rises from 50% to 76%, expressed as a percentage. For the Or-Or data (red), the JND ranged from 5.2% to 5.8%, and the average was 5.5%, which is effectively the same as that reported in the previous study on voiced speech [6]. For the Or-Em data (blue), the JND values are between 4.5% and 7.9%, and the average is 5.9%, somewhat larger than the average Or-Or value. The between-listener variability in the Or-Em data is greater than in Or-Or data because, for some listeners, the psychometric function was largely unaffected by the presence of the spectral uplift, while for others, the mid-point was clearly shifted by the presence of the spectral uplift. This is similar to the effect of the spectral uplift in the experiment with unvoiced and whispered sounds[7].

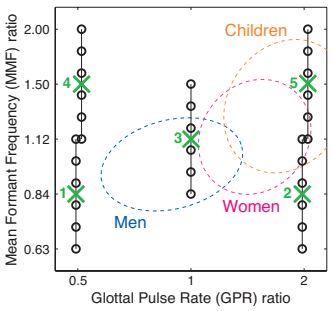


Figure 2: *The GPR-MFF combinations for the voiced words presented in the discrimination experiment. The five reference speakers are shown by the crosses; the six test speakers associated with each reference speaker are shown by the solid circles. Dotted ellipses show normal speech ranges.*

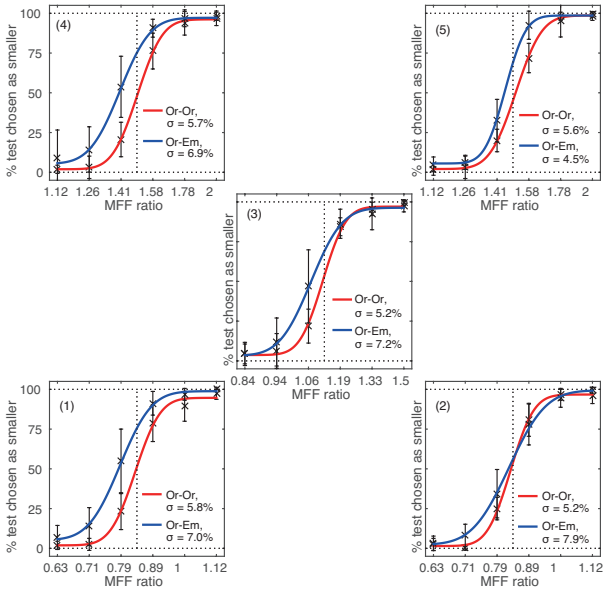


Figure 3: *The psychometric functions show the average data of the eight listeners, separately for the five reference speakers. The original-original (Or-Or) and original-emphasized (Or-Em) conditions are presented by red and blue lines, respectively. The error bars represent ±1 standard deviation between listeners. The just noticeable difference (JND or σ) is relative MFF increment between the ordinate values 50% and 76%.*

3. Auditory model of size perception

The model of size perception developed for unvoiced and whispered speech [7] can explain the psychometric functions for the voiced speech sounds of the current experiment in conditions where the GPR value was 0.5 (GPR-MFF regions 1 and 4 in Fig. 2) but it cannot explain the results when the GPR value is 1.0 or 2.0 (GPR-MFF regions 2, 3 and 5 in Fig. 2). This prompted us to include the image-stabilization stage of the SWMT to the model of size perception, and examine the size-shape-images

(SSI) of speech that it produces.

3.1. Modification of the speech spectrogram

The image stabilization is a form of pitch-synchronous temporal integration that emphasizes the repeating time-interval patterns that voiced sounds generate at the output of the wavelet transform. The details of image stabilization are beyond the scope of the current paper; the SSI of an /o/ vowel is presented in Fig. 4(a) and it will suffice to illustrate the problem with the original auditory spectrum and the proposed solution. To make the SSI "scale-shift covariant," the abscissa of the stabilized image is converted from "time interval within the glottal period" to the product of time interval and the peak frequency of the filterbank channel — a variable designated, h in the SWMT model. This operation is restricted to one glottal period and the remaining periods are removed from the SSI. Thus, the bottom right-hand boundary of the image varies with GPR and there is no activity below this boundary. The area of the image that facilitates the processing of size information falls above this curved boundary. The spectrograms that form the basis of the size perception models in Yamamoto et al. [7] do not take the exclusion of activity below the boundary into account, and this is why these versions of the size perception model do not work for relatively high GPRs.

These observations prompted us to develop a weighting function to reduce activity in the region of the spectrogram associated with the lower harmonics of the GPR. The function is shown in Fig. 4(b). When the glottal pulse rate is F_0, the glottal period is $1/F_0$, and the weighting function ("SSI weight") w_{SSI}, has the form

$$w_{SSI}(e_f) \quad = \quad \frac{\min(f_p(e_f)/F_0, h_{max})}{h_{max}}. \qquad (1)$$

In this expression, $f_p(e_f)$ is the peak frequency of the dcGC auditory filter at e_f on the $\mathrm{ERB_N}$ axis, and h_{max} is the maximum value of the time-interval, peak-frequency product, h, which specifies the horizontal range of the SSI. For simplicity of calculation, F_0 was fixed at a value, $F_0^{(lim)}$, which determines the limit of the area that is modified by w_{SSI}. Figure 5 shows two spectrograms of Japanese words when w_{SSI} is included in the calculation of the spectrogram. The upper and lower spectrograms show words where there is and is not spectral uplift; a comparison of the spectrograms reveals that there is more high frequency activity in the upper spectrogram with the uplifted speech.

3.2. Size discrimination procedure

The size perception model plays the role of the listener in the 2AFC experiment performed by the human listeners. On each trial, it is required to specify which interval had the smaller speaker, after processing precisely the same words as presented to the human listeners. This procedure makes it possible to construct psychometric functions for the model's performance like those shown in Fig. 3.

3.2.1. The spectrogram cross-correlation function

Figure 5 illustrates the processing used to make the size judgement. Weighted auditory spectrograms of the words in the two intervals of a trial were calculated. Then 50-ms segments were extracted from the center of each vowel using an automated HMM recognizer [15]. The spectral profile of each vowel's 50-ms spectrogram is referred to as its excitation pattern, Ep, and the Ep was calculated for all of the vowel types,

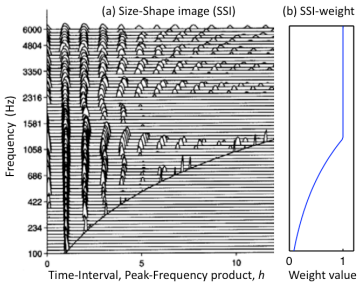


Figure 4: *The SSI weighting function: (a) a Size-Shape Image of the vowel /o/ (modified from [1]), (b) the weighting function based on the active region of the SSI.*

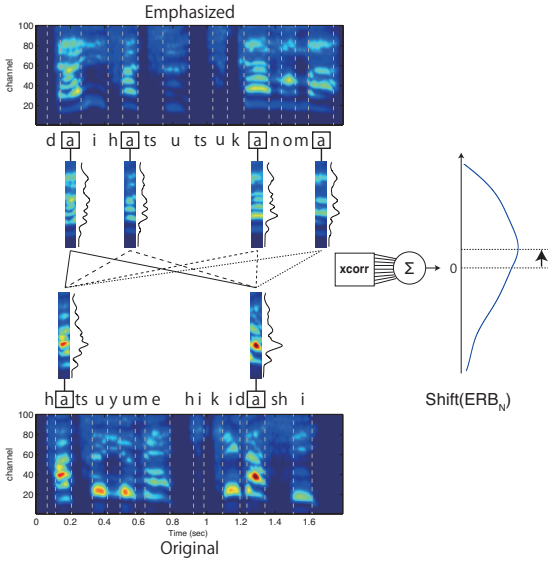


Figure 5: *Cross-correlation model based on the auditory spectrograms of the sounds presented in the two intervals of a trial. In the example, the first interval (top panel) contains two emphasized words, "daihatsu" and "tsukanoma"; the second interval (bottom panel) contains two original words, "hatsuyume" and "hikidashi."*

$v = \{/a/, /e/, /i/, /o/, /u/\}$ in all of the words.

The Ep's for all of the vowels of one type in the first interval were individually cross-correlated with all of the vowels of the same type in the second interval (as illustrated for the /a/ vowels in Fig. 5). The cross-correlation function for a single cross-interval comparison is

$$R_{Ep12}(e_s) \quad = \quad \sum_{i=-N+1}^{N-1} Ep_1(e_{f_i}) Ep_2(e_{f_i} + e_s) \qquad (2)$$

where e_{f_i} is the $\mathrm{ERB_N}$ number, e_s is the shift value on the $\mathrm{ERB_N}$ axis, and N is the number of channels. In cases where the test speaker is smaller than the reference speaker, the peak in this cross correlation function will, on average, shift to values above zero, and so the "peak shift" seems a reasonable variable for making decisions about speaker size. There are often several tokens of one vowel type in an interval, so after the cross-correlation functions were calculated for all individual pairings (index k) and all vowel types (index v), an overall cross-correlation function for the trial was calculated as follows

$$R_{Ep_{v,k}}(e_s) \quad = \quad \frac{R_{Ep12,v,k}(e_s)}{\sqrt{R_{Ep11,v,k}(0) \cdot R_{Ep22,v,k}(0)}}, \qquad (3)$$

$$\Phi(e_s) = \sum_{v=/a/}^{/u/} \sum_{k} R_{Ep_{vk}}(e_s). \quad (4)$$

An example of the distribution of $\Phi(e_s)$ is shown in right hand side of Fig. 5. The peak is shifted up from the center of the correlation axis; the shift unit is $\mathrm{ERB_N}$. This peak shift measure was used to "predict" the human response to the stimuli in each trial of the experiment, and thereby generate psychometric functions for the model for comparison with those of the individual listeners. The peak shift value, S_{Ep}, is calculated as

$$S_{Ep} = \arg\max_{e_s} \Phi(e_s). \quad (5)$$

The interval toward which S_{Ep} is shifted was selected as the "smaller speaker." When $S_{Ep} = 0$, the interval was randomly selected.

3.2.2. *Individual differences in the effect of the spectral emphasis*

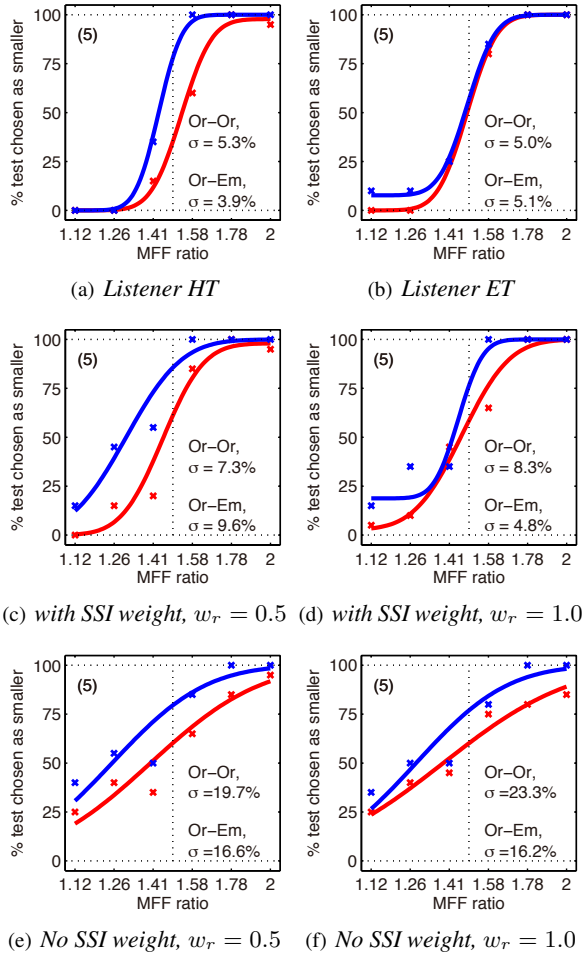
There were relatively large individual differences in the degree of shift of the psychometric function in the Or-Em condition. It seems that some listeners either ignore or repress the effect of the high- frequency spectral enhancement. To characterize the variation of individuals in this regard, we developed a version of the Ep based on the slope of the Ep. When the slope is α_r , the personal pattern, Ep^c, is

$$Ep^c(e_f) = Ep(e_f) - w_r \alpha_r e_f \quad (6)$$

where the weight, w_r, describes the degree to which the listener's judgments are influenced by the spectral enhancement. We calculated S_{Ep} in Eq. 5 from this modified form of the Ep^c. In passing it was noted that including 2nd and 3rd order regression components had little effect on the results.

3.3. Results

Figure 6 shows example psychometric functions in GPR-MFF region 5 for two listeners (HT and ET; top row); along with simulations of the psychometric functions from the SWMT model when it includes w_{SSI} (middle row) and when it does not (bottom row). The top panel shows that the psychometric function for the Or-Em condition (blue) is shifted to the left of that for the Or-Or condition for listener HT (a) but not for listener ET (b). The JNDs (σ) are around 5%. The psychometric functions from the SWMT model that includes w_{SSI} (middle row) show a similar leftward shift for HT (c) but not for ET (d). The JNDs are 2-4% larger in the Or-Or conditions and in the Or-Em condition for listener HT. The psychometric functions simulated without w_{SSI} (lower row) are like those originally developed for unvoiced speech [7], and they are considerably shallower than the psychometric functions derived directly from the data (top row). The JNDs are more than 16% which is far greater than the 5% derived directly from the listeners' data (top row). This is because the glottal pulses of these voiced sounds impart peaks to the excitation pattern which are not distinguishable from formant peaks. The weighting function, w_{SSI} , derived from the SSI of the SWMT reduces the effect of the glottal pulses in the excitation pattern. The $F_0^{(lim)}$ in Eq. 1 was fixed at 150 Hz in the weighted simulations shown in the middle row of Fig. 6. It is the case, however, that when $F_0^{(lim)}$ was increased to 200 Hz, or even 300 Hz, the psychometric functions were largely unaffected. This was true even for the psychometric functions in regions 2 and 5 where the GPR ratio is 2 and the average F_0 is 300 Hz. This indicates that w_{SSI} does not require precise F_0 estimation, making it a simple function for size estimation with voiced speech sounds.



(a) *Listener HT* (b) *Listener ET*

(c) *with SSI weight,* $w_r = 0.5$ (d) *with SSI weight,* $w_r = 1.0$

(e) *No SSI weight,* $w_r = 0.5$ (f) *No SSI weight,* $w_r = 1.0$

Figure 6: *Example psychometric functions in GPR-MFF region 5 for two listeners (a),(b). Simulations of the psychometric functions from the SWMT model when it includes* w_{SSI} *(c),(d) and when it does not (e),(f). Red lines for Or-Or conditions; blue lines for Or-Em conditions.*

4. Conclusions

An auditory model based on the dcGC filterbank was developed to explain speaker-size discrimination with voiced speech sounds. The dcGC filterbank is the first stage of the SWMT model of speech processing[1]. Data from a size discrimination experiment show that the low-frequency resolved harmonics of voiced speech sounds disrupt size estimation because they are confused with formant peaks. The problem was solved by incorporating a "SSI weighting" function from the second stage of the SWMT model. It reduces the effect of glottal pulse excitation on the auditory speech spectrum. Since the SSI weight is a simple spectral weighting function, the solution should be applicable to other linear, spectral analysis systems using gammatone, mel-frequency or one-third-octave filterbanks, when the processing involves estimating vocal tract length (VTL), or speaker size, with voiced speech sounds.

5. Acknowledgements

This research was partially supported by JSPS KAKENHI Grant Numbers JP25280063, JP15H02726, and JP16H01734.

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, pp. 181–203, 2002. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167639300000856

[2] 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

[3] H. Kawahara and T. Irino, *Underlying principles of a high-quality speech manipulation system STRAIGHT and its application to speech segregation*, P. Divenyi, Ed. Springer, 2005. [Online]. Available: http://link.springer.com/chapter/10.1007/0-387-22794-6_11

[4] 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

[5] 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

[6] 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://www.sciencedirect.com/science/article/pii/S0167639312000465

[7] K. Yamamoto, T. Irino, R. Nisimura, H. Kawahara, and R. D. Patterson, "How the slope of the speech spectrum affects the perception of speaker size." in *INTERSPEECH*, 2015, pp. 1556–1560.

[8] T. Irino and R. D. Patterson, "A time-domain, level-dependent auditory filter: The gammachirp," *The Journal of the Acoustical Society of America*, vol. 101, no. 1, pp. 412–419, 1997. [Online]. Available: http://dx.doi.org/10.1121/1.417975

[9] ——, "A compressive gammachirp auditory filter for both physiological and psychophysical data," *The Journal of the Acoustical Society of America*, vol. 109, no. 5, pp. 2008–2022, 2001. [Online]. Available: http://dx.doi.org/10.1121/1.1367253

[10] R. D. Patterson, M. Unoki, and T. Irino, "Extending the domain of center frequencies for the compressive gammachirp auditory filter," *The Journal of the Acoustical Society of America*, vol. 114, no. 3, pp. 1529–1542, 2003. [Online]. Available: http://dx.doi.org/10.1121/1.1600720

[11] T. Irino and R. D. Patterson, "A dynamic compressive gammachirp auditory filterbank," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 6, pp. 2222–2232, 2006. [Online]. Available: http://dx.doi.org/10.1109/TASL.2006.874669

[12] T. Matsui, T. Irino, K. Yamamoto, H. Kawahara, and R. D. Patterson, "The effect of spectral slope on size discrimintion of voiced speech sounds," in *INTERSPEECH*, 2017, p. to appear.

[13] 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," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*. IEEE, 2008, pp. 3933–3936. [Online]. Available: 10.1109/ICASSP.2008.4518514

[14] F. A. Wichmann and N. J. Hill, "The psychometric function: I. fitting, sampling, and goodness of fit," *Perception & psychophysics*, vol. 63, no. 8, pp. 1293–1313, 2001. [Online]. Available: http://link.springer.com/article/10.3758/BF03194544

[15] A. Lee, T. Kawahara, and K. Shikano, "Julius—an open source real-time large vocabulary recognition engine," in *EUROSPEECH2001: the 7th European Conference on Speech Communication and Technology, September 3-7, 2001, Aalborg, Denmark*, 2001, pp. 1691–1694. [Online]. Available: http://hdl.handle.net/10061/7954