The gammachirp auditory filter and its application to speech perception

Toshio Irino¹,* and Roy D. Patterson²,†

¹Faculty of Systems Engineering, Wakayama University, 930 Sakaedani, Wakayama, 640–8510 Japan
²Department of Physiology, Development and Neuroscience, University of Cambridge, Downing Street, Cambridge CB2 3EG, United Kingdom

Abstract: We review the gammachirp (GC) auditory filter and its use in speech perception research. The GC was originally developed to explain the asymmetric, auditory filter shapes derived in notched-noise (NN) masking studies, and the strongly compressive input-output function observed in the mammalian cochlea. This compressive GC was fitted to a very large collection of notched-noise (NN) masking thresholds measured with a wide range of stimulus levels and center frequencies. The fit showed how the GC auditory filter could explain NN masking throughout the domain of human hearing with a relatively small number of parameters, only one of which was level dependent. Subsequently, a dynamic, compressive GC filterbank (dcGC-FB) was developed to simulate time-domain cochlear processing. This dcGC-FB has been used to cancel the peripheral compression of normal hearing and thereby simulate the most common forms of hearing loss. This simulator allows normal hearing listeners to experience the difficulties of hearing impaired listeners. It has been used in training courses for speech-language-hearing therapists and psychoacoustic experiments. The dcGC-FB has also been used for modeling speaker size perception and predicting speech intelligibility with GEDI (the gammachirp envelope distortion index).

Keywords: Auditory model, Cochlea, Hearing impairment simulator, Speech perception, Size perception, Speech intelligibility

PACS number: 43.66.x  [doi:10.1250/ast.41.99]

1. INTRODUCTION

Speech communication is one of the most important factors in maintaining quality of life over the lifespan. To develop support apparatus for communication, it is essential to understand how speech sounds are processed in the auditory system and to construct computational models of speech perception. Since the human speech processing begins with a cochlear time-frequency analysis, it is essential to begin any model of speech processing with a reliable auditory filter bank. The gammachirp (GC) auditory filter is a unique peripheral model based jointly on psychoacoustical and physiological knowledge. This paper provides an overview of the GC filter and its use in speech perception research. First, we briefly describe the history, architecture, and essential parameters of the GC filter [1–3]¹ and the GC filterbank [4]. Then we describe applications of the GC filterbank (GC-FB) in speech perception studies. The GC architecture enables us to construct a hearing impairment simulator which allows normal hearing listeners to experience the difficulties of hearing impaired listeners [5,6]. The GC has also been used to model speaker size perception [7,8] and speech intelligibility [9–13].

2. THE GAMMACHIRP AUDITORY FILTER

This section summarizes the theoretical background of the GC auditory filter. Psychoacoustic notched-noise experiments, which provide the basis of the GC, are reviewed in [1–3,14]. The historical background including the physiological studies is reviewed in [3].

2.1. The Gammatone Filter

In time-domain auditory models, the spectral analysis performed by the basilar membrane is often simulated by a bank of gammatone auditory filters (e.g., [15]). The impulse response of the gammatone is

\[ g(t) = a t^{n-1} \exp(-2\pi b \text{ERB}_N(f_c) t) \cos(2\pi f_c t + \phi) \]  (1)

where \( a, b, n, f_c, \) and \( \phi \) are parameters. Time \( t \) is positive. \( \text{ERB}_N(f_c) \) is the equivalent rectangular bandwidth of the
filter for normal hearing listeners at moderate sound levels. ERB_N(f) = 24.7 + 0.108f in Hz [16]. The filter gets its name from the fact that the envelope formed by the power function and the exponential is a gamma distribution function, and the cosine carrier is perceived as a tone when it is in the auditory range.

The amplitude spectrum of the gammatone filter is essentially symmetric on a linear frequency scale, so it has difficulty representing the level-dependent asymmetry of the auditory filter. The “roex” auditory filter [16] can handle asymmetry, however, it does not have a well-defined impulse response which largely precludes its use in auditory filterbanks. Historically, more complicated physiological models of cochlear mechanics (for example, [15]) could not be used to fit human masking data because they were not sufficiently fast to support iterative fitting procedures. Recently, an efficient transmission line model has been successfully used to fit human masking data [17].

2.2. The Gammachirp Filter

In 1995, it was demonstrated that an analytic relative of the gammatone function, referred to as the “gammachirp” (GC) is a theoretically optimum auditory filter, in the sense that it leads to minimal uncertainty in a joint time and scale representation of auditory signal analysis [1]. The GC auditory filter is the real part of the analytic GC function; that is,

\[ g_c(t) = a t^{n-1} \exp(-2\pi b t \cdot \text{ERB}_N(f_r t)) \times \cos(2\pi f_r t + c_1 \ln t + \phi) \]  

(2)

The fundamental difference between it and the gammatone Eq. (1) is the term \( c_1 \ln t \); \( c_1 \) is an additional parameter; \( \ln \) is the natural logarithmic operator. The filter has a monotonically frequency-modulated carrier (a chirp) with an envelope that is a gamma distribution function, and hence the name “gammachirp.”

2.2.1. Frequency response

The Fourier magnitude spectrum of the complex form of the GC filter Eq. (2) is

\[ |G_C(f)| = \frac{a |\Gamma(n_1 + jc_1)|}{2\pi \sqrt{(b_1 \cdot \text{ERB}_N(f_r))^2 + (f - f_r)^2}^\alpha} \cdot \exp(c_1 \theta_1) \]

\[ = a \cdot |G_T(f)| \cdot \exp(c_1 \theta_1) \]  

(3)

\[ \theta_1 = \arctan \left( \frac{f - f_r}{b_1 \cdot \text{ERB}_N(f_r)} \right) \]  

(4)

The scalar values \( a, b_1, c_1, f_r, \) and \( n_1 \) are the amplitude, bandwidth, chirp factor, asymptotic frequency, and envelope factor of the filter. \( \text{ERB}_N(f_r) \) is the equivalent rectangular bandwidth at \( f_r \). It is clear that the frequency response of the GC is the product of that of the gammatone and the asymmetric function, \( \exp(c_1 \theta_1) \). This form of the GC can explain the level-dependent asymmetry of the auditory filter if the parameter \( c_1 \) is associated with stimulus level, and it is this version that is used to fit notched-noise masking data.

2.2.2. Remaining problems

There were, however, several problems with this form of the GC filter as a representation of cochlear filtering. It was difficult to explain both the compressive gain and the level-dependent shape of the auditory filter simultaneously. Moreover, when the chirp parameter, \( c_1 \), is level dependent, the trajectory of the chirp changes with level as illustrated by Eq. (2). A physiological study of the cochlear filter of the cat [18] had shown that, although there is a chirp in the impulse response, it does not vary with level as it does in the analytic GC.

2.3. The Compressive Gammachirp Filter

To solve these problems, Irino and Patterson [2] decomposed the asymmetric function, \( \exp(c_1 \theta_1) \), in Eq. (3) into separate low-pass and high-pass asymmetric functions. The resulting “compressive” GC filter, \( G_{CC}(f) \), is

\[ |G_{CC}(f)| = |G_T(f)| \cdot \exp(c_1 \theta_1) \cdot \exp(c_2 \theta_2) \]

\[ = |G_{CP}(f)| \cdot \exp(c_2 \theta_2) \]  

(5)

\[ \theta_1 = \arctan \left( \frac{f - f_r}{b_1 \cdot \text{ERB}_S(f_r)} \right) \]  

(6)

\[ \theta_2 = \arctan \left( \frac{f - f_r}{b_2 \cdot \text{ERB}_S(f_r)} \right) \]  

(7)

Conceptually, this compressive GC (cGC) filter is composed of a level-independent, “passive” GC (pGC) filter, \( |G_{CP}(f)| \), that represents the passive basilar membrane, and a level-dependent, high-pass asymmetric function (HP-AF; \( \exp(c_2 \theta_2) \)), that represents the active mechanism in the cochlear partition. Figure 1 shows the frequency responses of the pGC, the HP-AF, and the cGC. It illustrates how a level-dependent set of cGC filters (upper set of solid lines; left ordinate) can be produced by cascading a fixed pGC filter (lower solid line; right ordinate) with a set of HP-AFs (the set of dashed lines; right ordinate). When the leftmost HP-AF is cascaded with the pGC, it produces a cGC filter with much more gain than when the rightmost HP-AF is cascaded with the pGC. As a result, at the peak frequency of the compressive GC, gain decreases as level increases. The ordinate is filter gain and the gain is normalized to the peak value of the filter associated with the lowest probe level, which in this case is 30 dB. The five filter shapes have been calculated for probe levels of 30, 40, 50, 60, and 70 dB. The filter is referred to as a “compressive” GC because the compression around the peak frequency is incorporated into the filtering process itself. The HP-AF also makes the passband of the cGC more symmetric at lower levels.

Acoust. Sci. & Tech. 41, 1 (2020)
shows the decrement of cGC gain as level increases. Filter bandwidth increases as level increases, which agrees, at least qualitatively, with the physiological data. It is also the case that the chirp in the impulse response of this cGC does not vary with level, as required by the physiological data [18].

The HP-AF, \( \exp(c_2\theta_2) \), does not have an analytic impulse response. An asymmetric compensation filter was developed to simulate the HP-AF (see Appendix B of [2]). This enabled the simulation of the cGC impulse response as follows:

\[
g_{cc}(t) = a_c \cdot g_{ca}(t) \ast h_c(f),
\]

where \( a_c \) is a constant, \( g_{ca}(t) \) is the pGC impulse response as in Eq. (2), and \( h_c(t) \) is the impulse response of the asymmetric compensation filter.

## 2.4. The Dynamic, Compressive GC Filterbank

A dynamic, compressive gammachirp filterbank (dcGC-FB) was developed to simulate cochlear processing. Figure 3 presents a block diagram of the system which includes a resynthesis option. The analysis stage consists of a bank of linear, passive GC filters, a bank of asymmetric compensation filters, and a level estimation circuit. The synthesis stage consists of a bank of inverse HP-AFs and a bank time-reverse, passive GC filters. This stage was made possible by the fact that the asymmetric compensation filter (AF), \( h_c(t) \), in Eq. (11) was implemented with a minimum phase IIR filter, which makes it invertible. Moreover, since it is a time-varying linear filter, it is possible to invert the signal even when the filter coefficients are time-varying, provided a history of the coefficients from the analysis stage is preserved and applied appropriately in the resynthesis stage.

Between the analysis and resynthesis stages, it is possible to include a very wide range of signal processing algorithms including ones previously developed with linear systems applications. The complete system makes it possible to resynthesize sounds from the output of the
The nonlinear compression and suppression behavior observed in human psychophysics indicates that the rate of adaptation to level changes in the cochlea is very rapid; the time constant is on the order of 1 ms. In the dcGC model, fast acting level estimation and parameter updating is achieved locally for each auditory filter using the control circuit shown in Fig. 4. It has separate level-estimation and signal-flow paths, both of which include a pGC filter and an HP-AF. In the signal path, the pGC filter has parameters $b_1$, $c_1$, $f_{p1}$, and the HP-AF has parameters $b_2$, $c_2$, $f_{r1}$ (=$f_{rat}$ : $f_{p1}$). This combination of pGC and HP-AF results in the compressive GC (cGC) defined in Eq. (5) with peak frequency $f_{p1}$. The parameter values are the same as described in Sect. 2.3. In the level-estimation path, the pGC filter has parameters $b_1$, $c_1$, $f_{pL}$ and the HP-AF has parameters $b_2$, $c_2$, $f_{rL}$ (=$f_{rat}$ : $f_{pL}$), which are essentially the same as for the signal path. The difference is in the level-independent frequency ratio, $f_{rat}$. To simulate the asymmetric suppression observed in humans, the peak frequency $f_{pL}$ of the pGC in the level-estimation path is required to satisfy the relationship

$$\text{ERB}_N \text{ number}(f_{pL}) \rightarrow \text{ERB}_N \text{ number}(f_{p1}) + r_{EL}, \quad (12)$$

where ERB$_N$ number($f$) is the ERB$_N$ rate [16] at frequency $f$, and $r_{EL}$ describes the frequency separation between the two pGC filters on the ERB rate axis. The output of the level-estimation path is used to control the level-dependent parameters of the HP-AF in the signal path. In order to account for the different rates of growth of suppression in the upper and lower suppression regions, it was necessary to use, not only the level at the output of the pGC, but also the level of the output of the HP-AF. The level was estimated in decibels on a sample-by-sample basis and used to control the level in the signal path. The detail of the level calculation is described in [4].

### 3. HEARING IMPAIRMENT SIMULATOR

In 2013, Irino et al. [5,6] proposed a new hearing impairment simulator (HIS) based on the cGC auditory filter. Briefly, the compression applied by the auditory filter of a normal hearing (NH) individual is inverted and used to cancel the compression of that individual, and then simulate the perception of a given hearing impaired (HI) listener for the original NH listener.

#### 3.1. Compression in NH and Its Cancellation

The strategy is illustrated in Fig. 5 using the input-output (IO) functions of a NH individual (circles) and a HI individual (asterisks) with a severe sensory-neural hearing loss. The shallow, compressive section of the NH IO function is missing from the IO function of the HI person. What is required is a sound processor with inverse compression that can convert the IO function of the NH person into that of the HI person. The third IO function in the figure (squares) illustrates what is required. This section presents a method for calculating the inverse compression function required to simulate sensory-neural hearing loss. A procedure for including conductive losses and/or losses due to dysfunctional neural transduction is described in Sect. 3.1.3.

#### 3.1.1. Compression in the cGC and its inverse

We begin by defining the relationship between the ratio, $f_{rat}$, and the total stimulus level at the output of the pGC, $P_{emp}$, as shown in Eq. (10). The slope of the ratio, $f_{rat}$, is limited to positive values. Figure 1 shows how the HP-AF normally shifts up in frequency (rightward arrow).
causing the peak gain of the cGC to decrease (downward arrow) as input sound level increases.

Inverse compression is produced by reversing the arrows in Fig. 1, thereby inverting the sign of \( f^{(rev)}_{rat} \). The inverted frequency ratio, \( f^{(rev)}_{rat} \), can be written as

\[
f^{(rev)}_{rat} = f^{(0r)}_{rat} + f^{(1r)}_{rat} \cdot P_{\text{gcp}}
\]

(13)

where \( f^{(1r)}_{rat} \) is negative. This equation has the same form as Eq. (10) but different coefficients. To simulate the full range of hearing loss, we introduced a control parameter, \( \alpha \) (0 ≤ \( \alpha \) ≤ 1), which represents the ‘cochlear health’ of the HI person relative to that of the NH person.

\[
f^{(rev)}_{rat} = f^{(0r)}_{rat}(\alpha) + (1 - \alpha)f^{(1r)}_{rat} \cdot P_{\text{gcp}}.
\]

(14)

Full cancellation occurs when \( \alpha = 0 \) and no cancellation occurs when \( \alpha = 1 \). Full cancellation is the situation illustrated in Fig. 5. In the current version, the control level, \( P_{\text{gcp}} \), in Eq. (14) is the output of the level estimation circuit in Fig. 4.

3.1.2. The IO function and hearing level

The standard measure of hearing loss is hearing level (HL). The HL of an individual at 1,000 Hz can be calculated from the IO function of that individual at 1,000 Hz, as illustrated in Fig. 5. The horizontal “Threshold” line shows the output level of the 1,000-Hz auditory filter that will create activity in the auditory nerve that is just detectable. The intersection of an individual’s IO function with the threshold line specifies the SPL of their absolute threshold at 1,000 Hz. Their HL is their absolute threshold relative to that of a typical NH listener, which at 1,000 Hz is 7 dB SPL. In Fig. 5, this specific “NH” listener has a HL of 5 dB, while the simulated HI person has a HL of 41 dB.

3.1.3. The audiogram and hearing loss

The standard summary of an individual’s hearing is their “audiogram” which is a plot of the individual’s HL across a range of standard frequencies. For clinical applications, it is most convenient to use a patient’s audiogram to set up the inverse compression of the HI simulator for each frequency channel. This is how the impairment of an individual is tailored for presentation to a NH person. The audiogram of the HI person in our example is shown by the bottom line (black) in the main panel of Fig. 6. The upper line (green) shows the audiogram of the standard NH listener calculated with the cGC filter. The middle line (magenta) shows the portion of the hearing loss of our HI person due to their loss of compression, as a function of signal frequency. This is commonly associated with a reduction in the performance of their outer hair cells (OHCs). The difference between this magenta line and the black line corresponds to the loss associated with a reduction in the performance of their inner hair cells (IHCs). Together these two losses are referred to as “sensorineural” hearing loss. The total loss is assumed to be

\[
\text{Loss}^{(\text{Total})} = \text{Loss}^{(\text{OHC})} + \text{Loss}^{(\text{IHC})}
\]

(15)
on a dB scale; this is Eq. (5) of [19].

3.2. Fast Signal Processing

The original version of the HI simulator [5] was implemented with the analysis/synthesis dcGC-FB shown in Fig. 3. The processing speed was several tens of times slower than real-time because the calculation of the dcGC-FB was done on a sample-by-sample basis. It was too slow to support the interactive system with the attendant GUI.

3.2.1. Frame-based time-varying filter

To increase processing speed, we developed a frame-based system to simulate the nonlinear, time-varying filtering between the input and output of the HI simulator. Figure 7 presents a schematic of the system. The input signal is divided into frames with a square-root hanning window, \( w(t) = \sqrt{0.5 + 0.5 \cos(2\pi/T)\tau} \).

Fig. 6 GUI for the HI simulator, with a main audiogram panel and a column of controls.

Fig. 7 Architecture of the fast HI simulator with frame-based time-varying filtering.
The minimum phase filter is calculated from the excitation pattern of the dcGC-FB. The input signal is analyzed by the passive GC filterbank and the signal level is estimated for each channel by the circuit shown in Fig. 4. The estimated level is averaged over 20-ms frames which are aligned with the frames in the time-varying filter described above. This average level is used to determine the frequency ratio for the HP-AF, \( f_{rat} \), in Eq. (10) and, subsequently, the gain of the “inverse” HP-AF. The gain vector for the filterbank channels is used as the “inverse” excitation pattern. On the ERBN axis, this excitation pattern is squared and warped onto a linear frequency axis to derive the power spectrum of the signal. The minimum phase filter is calculated from this power spectrum using a cepstral method.

### 3.2.2. Improved processing speed

The MATLAB version of this algorithm runs in 2–3 times real time and the C version runs in real time [20]. It is this frame-based system that supports the GUI version of the HI simulator described below. Note that the speed problem can also be solved using the massive computational power of a GPU programmed with python. Recently, Grimault et al. [21] used this approach to produce a real-time HI simulator based on the original analysis/synthesis dcGC-FB.

### 3.3. Simulator GUI and Implementation

Figure 6 is a snapshot of the graphical user interface (GUI) developed for the HI simulator; it is implemented in MATLAB. In addition to the main audiogram panel, there are several sets of control buttons. After calibration of sound levels, the user chooses an audiogram and the healthiness-of-compression value. Then the user can record a sentence or more of speech or load some prerecorded speech. The user can then listen to the version of the speech that the HI person would hear by pressing the button “Processing.”

Recently, a web server-client version has been developed where the GUI is implemented with HTML5 and the processing engine is implemented with C code [20]. This version of the simulator makes it possible to provide access to the system from any internet terminal without the need to download and install the software. A demonstration version is located at [22]. A VST (Steinberg’s Virtual Studio Technology) plugin [20] of the HI simulator is also being developed for desktop music (DTM) to illustrate the effect of a patient’s hearing impairment on their perception of music that they play.

### 3.4. Simulator Applications

The GUI version of the HI simulator has been used for three years in the training of speech-language-hearing therapists. In the class, the students get a chance to experience the difficulties of HI listeners.

Another application involves psychoacoustic experiments. HI people often have difficulty understanding speech in multi-speaker, or noisy, environments. With HI listeners, it is often difficult to specify which stage, or stages, of auditory processing are responsible for the deficit. Moreover, there is huge inter-subject variability even in the audiogram. There might also be cognitive problems associated with aging. This means that it is not easy to perform psychoacoustic experiments in uniform, well controlled conditions with HI listeners. The simulator allows us to substitute young NH listeners for HI listeners to isolate the deficit associated with a specific loss of compression on speech perception; the young NH listeners would listen to sounds from the HI simulator set to produce the given degree of compression loss. For example, Matsui et al. [23] used the HI simulator to measure the effect of a loss of compression on syllable recognition. It is difficult to do this kind of experiment with HI listeners due to the extreme variability of their audiograms and the difficulty of specifying the OHC and IHC components of their losses. After the NH listeners were converted into a uniform set “HI listeners” using the HI simulator, they were required to identify the second syllable of a three-syllable “nonsense word” on each trial of the experiment, and between trials, the relative level of the second syllable was varied, or the level of the entire sequence was varied. The difference between the Speech Reception Threshold (SRT) in these two conditions reveals the effect of compression on speech perception. The HI simulator adjusted the NH listener’s HL to that of a typical 80-year old with either normal compression or complete loss of compression. A reference condition was included where the HI simulator applied a flat 30-dB reduction in stimulus level. The results showed that the loss of compression has its largest effect on recognition when the second syllable was attenuated more than those before and after it.

### 4. AN AUDITORY MODEL OF SIZE PERCEPTION

We hear vowels pronounced by men, women, and children as approximately the same, although vocal tract
length (VTL) varies considerably between these groups. At the same time, we can identify the speaker group. This suggests that the auditory system segregates information about vocal tract length from that about vocal tract shape at an early point in auditory processing. Irino and Patterson [24] proposed a model involving the dcGC auditory filter to explain how the auditory system might segregate the acoustic features in speech sounds associated with vocal tract shape and vocal tract length, and thereby produce an internal representation of speech sounds that is speaker-size invariant.

4.1. Theory

The theory is based on the Stabilized-Wavelet Mellin Transform (SWMT) which is a cascade of a wavelet transform, image stabilization, and a Mellin transform. The wavelet transform is based on the GC auditory filter which was shown to satisfy minimum uncertainty in a time-scale representation derived using the Mellin transform. This suggests that the dcGC-FB is a suitable front end for models of speaker normalization and speaker size perception.

4.2. Experiments

A series of behavioral experiments was performed to determine the accuracy of human size perception. Speech sounds were scaled with the high-quality speech vocoder referred to as STRAIGHT [25] to manipulate the vocal tract length (VTL) features of natural speech sounds.

Smith et al. [26] performed a size discrimination experiment using natural vowels and showed that the just noticeable difference (JND) in VTL for speaker size is about 7%. It also showed that phoneme recognition performance remains high over very wide ranges of VTL and glottal pulse rate — ranges that went well beyond those exhibited by the human population. Subsequently, Ives et al. [27] showed the JND improved to about 5% when the speech sounds are sequences of CV and VC syllables. Irino et al. [28] showed the JND is also about 5% when the stimuli are unvoiced words created from voiced words with STRAIGHT.

Yamamoto et al. [7] measured the effect of spectral slope on the perception of speaker size. STRAIGHT was used to generate “unvoiced” words with the spectral slope of the original voiced words, and “whispered” words with a 6-dB/octave spectral tilt. Examples of the psychometric functions they measured are shown in the left column of Fig. 8. The results show that the whispered words with the spectral tilt were perceived to be from smaller speakers. However, there seemed to be two classes of listeners — some who were effectively insensitive to the 6-dB/octave tilt and others who reported a consistent reduction in perceived size for the same spectral tilt. Matsui et al. [29] performed a similar experiment with voiced words and got similar results [7]. Together the experiments suggest that models of speaker size perception will need to include a listener specific parameter for the effect of spectral slope.

4.3. Implementation of the Size Perception Model

A dcGC software package was developed to explain the results of size discrimination experiments with voiced and unvoiced speech sounds. The model was based on auditory spectrograms produced with the dcGC-FB [4]. They were constructed using the rms levels output by individual channels of the dcGC-FB. Examples for whispered and unvoiced speech sounds are shown in Fig. 9. There is more high-frequency energy in the auditory spectrogram of the whispered speech.

The model was set up to “listen” to the trials of the two-alternative, forced-choice (2AFC) perception experiment performed by human listeners, and to specify, on each trial, which of the two intervals had the smaller speaker, just as the human listeners did. This procedure makes it possible to construct psychometric functions for the model’s performance that are comparable to those used to analyze the performance of the human listeners. Figure 9 illustrates the procedure. The excitation pattern, $E_p$, was calculated for all segments of the Japanese vowels, $v = \{/a/, /e/, /i/, /o/, /u/\}$. The $E_p$’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 (illustrated for the /a/ vowels in Fig. 9). Then an overall cross-correlation function for the trial was calculated for all vowel types (as shown in the right-most distribution). The peak shift in this summary cross-correlation function was used to generate psychometric functions for the 2AFC experiment.

Yamamoto et al. [7] calculated the cross-correlation function of the gradient $Ep$, $VEp$, and introduced a weighting constant $w$ to explain listener variability. The right-hand panels of Fig. 8 show the simulation results. The patterns of the psychometric functions are similar to those observed with the humans. This was not the case, however, in a follow-up experiment with voiced speech [29]. Irino et al. [8] suggested that this was because the resolved harmonics of the voiced vowels added variability to the spectral envelopes used to estimate vocal tract length. Using slope-compensated $Ep$’s and a weighting function extracted from the Size-Shape Image (SSI) of the SWMT (in Sect. 4.1), they effectively suppressed the harmonic interference and simulate the human psychometric functions for voiced, unvoiced and whispered speech. The results provide support for the SWMT model of size perception.

5. CONCLUSION

The first half of this paper reviewed the development of the gammachirp (GC) auditory filter from its introduction as a theoretically optimum form of auditory filter through to the dynamic, compressive GC filter and its associated time-domain filterbank, the dcGC-FB. The second half illustrated the role of the dcGC-FB in several areas of speech research: simulation of compressive hearing impairment and modelling of speaker size perception with voiced, unvoiced and whispered speech.

The dcGC-FB was also used in models for speech intelligibility prediction [9–13]. A new model referred to as GEDI (the gammachirp envelope distortion index) [10,11] predicted the intelligibility of speech sounds processed with non-linear enhancement algorithms better than other recent indexes like STOI, CSII, and HASPI [12,13].

The results suggest that the dcGC would be a good frontend for any model of human speech perception.

ACKNOWLEDGEMENTS

Our gammachirp research has been continuously supported by JSPS KAKENHI Grants: recently, by grants JP25280063, JP15H02726, JP16K11196, and JP16H01734. The authors wish to thank all collaborators on our projects.

REFERENCES

[1] T. Irino and R. D. Patterson, “A time-domain, level-dependent auditory filter: The gammachirp,” J. Acoust. Soc. Am., 101, 412–419 (1997).
[2] T. Irino and R. D. Patterson, “A compressive gammachirp auditory filter for both physiological and psychophysical data,” J. Acoust. Soc. Am., 109, 2008–2022 (2001).
[3] R. D. Patterson, M. Unoki and T. Irino, “Extending the domain of center frequencies for the compressive gammachirp auditory filter,” J. Acoust. Soc. Am., 114, 1529–1542 (2003).
[4] T. Irino and R. D. Patterson, “A dynamic compressive gammachirp auditory filterbank,” IEEE Trans. Audio Speech Lang. Process., 14, 2222–2232 (2006).
[5] T. Irino, T. Fukawatase, M. Sakaguchi, R. Nisimura, H. Kawahara and R. D. Patterson, “Accurate estimation of compression in simultaneous masking enables the simulation of hearing impairment for normal hearing listeners,” in Basic Aspects of Hearing, Advances in Experimental Medicine and Biology, 787 (Springer, New York, Heidelberg, Dordrecht, London, 2013), pp. 73–80.
[6] M. Nagae, T. Irino, R. Nisimura, H. Kawahara and R. D. Patterson, “Hearing impairment simulator based on compressive gammachirp filter,” Proc. APSIPA ASC 2014 (2014).
[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,” Proc. Interspeech 2015, pp. 1556–1560 (2015).
[8] T. Irino, E. Takimoto, T. Matsui and R. D. Patterson, “An auditory model of speaker size perception for voiced speech sounds,” Proc. Interspeech 2017, pp. 1153–1157 (2017).
[9] K. Yamamoto, T. Irino, T. Matsui, S. Araki, K. Kinoshita and T. Nakatani, “Speech intelligibility prediction with the dynamic compressive gammachirp filterbank and modulation power spectrum,” Acoust. Sci. & Tech., 40, 84–92 (2019).
[10] K. Yamamoto, T. Irino, S. Araki, K. Kinoshita and T. Nakatani, “Predicting speech intelligibility using a gammachirp envelope distortion index based on the signal-to-distortion ratio,” Proc. Interspeech 2017, pp. 2949–2953.
[11] K. Yamamoto, T. Irino, N. Ohashi, S. Araki, K. Kinoshita and T. Nakatani, “Multi-resolution gammachirp envelope distortion index for intelligibility prediction of noisy speech,” Proc. Interspeech 2018, pp. 1863–1867 (2018).

[12] K. Yamamoto, T. Irino, S. Araki, K. Kinoshita and T. Nakatani, “Speech intelligibility prediction using a multi-resolution gammachirp envelope distortion index with common parameters for different noise conditions,” Semin. Brain, Hearing and Speech Sciences for Universal Speech Communication, Tohoku Univ., Sendai, Japan, 25–26 Oct. (2018).

[13] K. Yamamoto, T. Irino, S. Araki, K. Kinoshita and T. Nakatani, “GEDI: Gammachirp envelope distortion index for predicting intelligibility of enhanced speech,” arXiv preprint [arXiv:1904.02096], 3 Apr. (2019).

[14] T. Irino, K. Yokota, T. Matsui and R. D. Patterson, “Auditory filter derivation at low levels where masked threshold interacts with absolute threshold,” Acta Acust. united Ac., 104, 887–890 (2018).

[15] R. D. Patterson, M. Allerhand and C. Giguère, “Time-domain modelling of peripheral auditory processing: A modular architecture and a software platform,” J. Acoust. Soc. Am., 98, 1890–1894 (1995).

[16] B. R. Glasberg and B. C. J. Moore, “Derivation of auditory filter shapes from notched-noise data,” Hear. Res., 47, 103–138 (1990).

[17] R. F. Lyon, “Cascades of two-pole–two-zero asymmetric resonators are good models of peripheral auditory function,” J. Acoust. Soc. Am., 130, 3893–3904 (2011).

[18] L. H. Carney, J. M. Megean and I. Shekhter, “Frequency glides in the impulse responses of auditory-nerve fibers,” J. Acoust. Soc. Am., 105, 2384–2391 (1999).

[19] B. C. J. Moore, B. R. Glasberg and B. Thomas, “A model for the prediction of thresholds, loudness, and partial loudness,” J. Audio Eng. Soc., 45, 224–240 (1997).

[20] T. Matsui, H. Banno, R. Nisimura and T. Irino, “Implementation of hearing impairment simulator based on the gammachirp auditory filterbank and its educational application,” IEICE Tech. Rep., WIT SP (2018) (in Japanese).

[21] N. Grimault, T. Irino, D. Dimachki, A. Corneylie, R. D. Patterson and S. Garcia, “A real time hearing loss simulator,” Acta Acust. united Ac., 104, 904–908 (2018).

[22] https://media.sys.wakayama-u.ac.jp/hisim/ (in Japanese, Latest view: 14 Apr. 2019).

[23] T. Matsui, T. Irino, M. Nagae, H. Kawahara and R. D. Patterson, “The effect of peripheral compression on syllable perception measured with a hearing impairment simulator,” in Physiology, Psychoacoustics and Cognition in Normal and Impaired Hearing, Advances in Experimental Medicine and Biology, 894 (Springer, 2016), pp. 307–314.

[24] 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 Commun., 36, 181–203 (2002).

[25] H. Kawahara, I. Masuda-Kasuse and A. de Cheveigne, “Restructuring speech representations using pitch-adaptive time-frequency smoothing and instantaneous-frequency-based F0 extraction: Possible role of repetitive structure in sounds,” Speech Commun., 27, 187–207 (1999).

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

[27] D. T. Ives, D. R. R. Smith and R. D. Patterson, “Discrimination of speaker size from syllable phrases,” J. Acoust. Soc. Am., 118, 3816–3822 (2005).

[28] T. Irino, A. Aoki, H. Kawahara and R. D. Patterson, “Comparison of performance with voiced and whispered speech in word recognition and mean-formant-frequency discrimination,” Speech Commun., 54, 998–1013 (2012).

[29] T. Matsui, T. Irino, K. Yamamoto, H. Kawahara and R. D. Patterson, “The effect of spectral tilt on size discrimination of voiced speech sounds,” Proc. Interspeech 2017, pp. 601–605 (2017).