Segregating information about the size and shape of the vocal tract using a time-domain auditory model: The stabilised wavelet-Mellin transform

Toshio Irino a,*, Roy D. Patterson b

a ATR Human Information Processing Research Laboratories, 2-2 Hikaridai, Seika-cho, Soraku-gun, 619-0288 Kyoto, Japan
b Centre for Neural Basis of Hearing, Department of Physiology, University of Cambridge, Downing Street, Cambridge CB2 3EG, UK

Received 20 August 1999; received in revised form 31 August 2000; accepted 29 September 2000

Abstract

We hear vowels pronounced by men and women as approximately the same although the length of the vocal tract varies considerably from group to group. At the same time, we can identify the speaker group. This suggests that the auditory system can extract and separate information about the size of the vocal-tract from information about its shape. The duration of the impulse response of the vocal tract expands or contracts as the length of the vocal tract increases or decreases. There is a transform, the Mellin transform, that is immune to the effects of time dilation; it maps impulse responses that differ in temporal scale onto a single distribution and encodes the size information separately as a scalar constant. In this paper we investigate the use of the Mellin transform for vowel normalisation. In the auditory system, sounds are initially subjected to a form of wavelet analysis in the cochlea and then, in each frequency channel, the repeating patterns produced by periodic sounds appear to be stabilised by a form of time-interval calculation. The result is like a two-dimensional array of interval histograms and it is referred to as an auditory image. In this paper, we show that there is a two-dimensional form of the Mellin transform that can convert the auditory images of vowel sounds from vocal tracts with different sizes into an invariant Mellin image (MI) and, thereby, facilitate the extraction and separation of the size and shape information associated with a given vowel type. In signal processing terms, the MI of a sound is the Mellin transform of a stabilised wavelet transform of the sound. We suggest that the MI provides a good model of auditory vowel normalisation, and that this provides a good framework for auditory processing from cochlea to cortex. © 2000 Elsevier Science B.V. All rights reserved.

Keywords: Auditory pathway; Mellin transform; Wavelet transform; Stabilised auditory image; Size-shape image; Gammachirp auditory filter

1. Introduction

There is a complex relationship in speech between the formant frequencies associated with a given vowel, on the one hand, and the size and sex of the speaker, on the other hand. A comparison of recent population studies on vocal-tract length (VTL) (Fitch and Giedd, 1999) and formant fre-
quencies (Huber et al., 1999) confirms several traditional assumptions: (1) As a child grows between the ages 4 and 12, there is a steady increase in VTL and a concomitant decrease in formant frequencies that can largely be explained by the increase in VTL. (2) During this time, the differences in VTL and formant frequencies between male children and female children of the same size (height or weight) are small. (3) As females grow beyond age 12 to their mature size, the increase in VTL and the decrease in formant frequencies proceed with size. The decrease in formant-frequency between age 4 and maturity is, on average, about 20% for females. (4) As males mature beyond age 12, there is a relatively rapid increase in VTL that exceeds that explained by the growth in size, so that mature males have significantly longer vocal tracts and lower formant frequencies than females. The lowering of the formants is largely proportional to VTL (Huber et al., 1999). For males, the decrease in formant-frequency between age 4 and maturity is, on average, about 32%.

Despite these relatively large differences in VTL and formant-frequency across the population of speakers, we hear individual vowels pronounced by men, women and children as tokens of the same vowel type, as if the auditory system had some means of normalising vowels for VTL prior to categorisation. Circumstantial evidence in support of this hypothesis comes from several sources: Hashi et al. (1998) normalised vowel posture data for 14 male and 14 female speakers and showed that both groups position the tongue to maintain the appropriate vocal-tract shape within quite small limits, indicating that normalisation would be useful if it were used. Similarly, Wakita (1977) has demonstrated that VTL normalisation by formant-frequency normalisation improves vowel identification. Finally, we note that phonetic textbooks describe the relationship between vowel sounds and the shape of the vocal tract in terms of the tongue position using vowel quadrilaterals like that shown in Fig. 1 (Fant, 1970; Pullum and Ladusaw, 1986). The quadrilateral is typically presented without indication of scale to emphasise that vowel type is determined by the shape of the vocal tract rather than by its absolute length.

It is also the case that we can identify the group that the speaker represents (man, woman, or child) which suggests that we extract size information as well from the vowel sound. There is, of course, correlated pitch information in the sound which has about the same value for identification of speaker size as VTL (Bachorowski and Owren, 1999). So this ability is not entirely based on VTL information. It remains the case, however, that VTL information is a more reliable measure of size than pitch because there is a close correlation between VTL and size (Fitch and Giedd, 1999), whereas pitch can vary more than an octave as it is used to make prosodic distinctions. Taken together, our ability to identify vowels despite formant-frequency differences and our ability to identify speaker group, suggest that the auditory system may use some form of normalisation process to isolate information about the shape of the vocal tract and segregate it from information about VTL.

The question then arises as to how the auditory system might do this? On the one hand, it may be that this apparent normalisation is the by-product of a complex process that develops with age and involves learning the statistics of the features of vowel sounds and combining this statistical information with knowledge of the speaker and the
contextual nature of communication. In this case, it would probably be decades before we gained a useful understanding of this central normalisation processes. There is, however, an alternative in the form of the Mellin transform (Titchmarsh, 1948; Cohen, 1993) which indicates that size is a physical attribute of sound, much like time and frequency, and that it can be segregated from other properties in a mechanical way. This raises the possibility that the auditory system normalises all sounds automatically at a relatively early point in the system with something like a Mellin transform to assist the identification of the object producing the sound. Physiological normalisation of sounds, prior to perception, could be expected to assist learning and recognition considerably by removing size variability from the learning and recognition problems. In this paper, we describe the Mellin transform as it applies to auditory processing and argue that some transformation of this form may well be used in speech perception.

1.1. Impulse response of the vocal tract

The vocal tract can be modelled as a loss-less acoustic tube (Flanagan, 1972), and given the cross-area function, it is possible to compute sound waveforms from such a model (Rabiner and Schafer, 1978). The waveform \( 'a_m' \) in Fig. 2(a) shows the impulse response produced by a typical vocal-tract model with the cross-area function of the Japanese vowel [a] for a specific male speaker (Yang and Kasuya, 1995). The waveform \( 'a_i' \) is the impulse response produced by a vocal tract having the same cross-area function but compressed to two-thirds of the tract length. The waveform \( 'a_{sp}' \) is a version of waveform \( 'a_m' \) that has been compressed by spline interpolation and down sampling to two-thirds of its original length. A copy of \( 'a_i' \) is superimposed on \( 'a_{sp}' \) using a thin line to emphasise the similarity. (b) The Mellin magnitude distributions of the waveforms in (a) using \( p = -jc + (1/2) \). The magnitude is normalised and the curves have been displaced vertically for clarity.

Fig. 2. (a) Relationship between vocal-tract length and the impulse response. The waveform \( 'a_m' \) is the impulse response produced by a typical vocal-tract model (Rabiner and Schafer, 1978) with the cross-area function of the Japanese vowel [a] for a specific male speaker (Yang and Kasuya, 1995). The waveform \( 'a_i' \) is the impulse response produced by a vocal tract having the same cross-area function but compressed to two-thirds of the tract length. The waveform \( 'a_{sp}' \) is a version of waveform \( 'a_m' \) that has been compressed by spline interpolation and down sampling to two-thirds of its original length. A copy of \( 'a_i' \) is superimposed on \( 'a_{sp}' \) using a thin line to emphasise the similarity. (b) The Mellin magnitude distributions of the waveforms in (a) using \( p = -jc + (1/2) \). The magnitude is normalised and the curves have been displaced vertically for clarity.
identical cross-area function. The reduction in VTL reduces the length of the impulse response but the shape of the wave is similar. To provide a direct comparison, the original wave \( a_m \) was reduced to two-thirds of its duration by spline interpolation. The resulting wave, \( a_{op} \), is plotted at the top of Fig. 2(a), and it reveals that \( a_t \) and \( a_{op} \) are virtually identical. This shows that extension or reduction of the acoustic tube produces a proportional dilation or contraction of the impulse response in time. The similarity that we perceive between men’s and women’s vowels leads us to assume that the auditory system can compare sounds to determine if one is a scaled version of the other, and in this way recognize similarities in the shape of the vocal tract, while at the same time noting differences in its size.

In speech research, the process of scaling vowel tokens to improve recognition of vowel type is referred to as vowel normalisation, and numerous techniques have been developed over the past two decades based on the location of formants in the magnitude spectrum (for example, Wakita, 1977; Davis and Mermelstein, 1980; Imai, 1983; Umesh et al., 1999). There is not, however, any general agreement on the best technique. They are all based on one form of magnitude spectrum or another, derived by linear predictive analysis, the Fourier transform or similar transforms, and in each case, they ignore the phase spectrum which contains half of the vocal-tract information (Appendix B). Contrary to common belief, the auditory system is not phase deaf (see Patterson, 1987, for a review), and phase information has been shown to play an important role in the perception of sound quality (e.g., Patterson, 1994a,b). This is one reason why the magnitude spectrum is not used in auditory models to simulate auditory spectral analysis. In this paper, to avoid the limitations of spectral vowel normalisation, we examine the problem of vowel normalisation in the time domain beginning with a simple waveform example and then proceeding to expand the principles to operate on vowels as they are represented in the auditory system.

1.2. The Mellin transform

There is a mathematical transform, the Mellin transform, that can convert the two waveforms \( a_m \) and \( a_t \) into the same magnitude distribution (Titchmarsh, 1948; Bertrand et al., 1996). The Mellin transform of a signal, \( s(t) \) \((t > 0)\), is

\[
S(p) = \int_{0}^{\infty} s(t) t^{p-1} dt = \int_{0}^{\infty} s(t) e^{(p-1)\ln t} dt,
\]

where \( p \) is a complex argument. It is a property of this transform that when the sound, \( s(t) \), is scaled in time, it does not affect the shape of the transformed distribution, it simply results in a scalar being applied to the distribution. In transform notation,

\[
\text{if } s(t) \Rightarrow S(p), \text{ then } s(at) \Rightarrow a^{-p}S(p),
\]

where the arrow, \( \Rightarrow \), indicates “is transformed into” and \( a \) is the dilation constant which is real. So, the magnitude distribution \(|S(p)|\) is not affected by scaling of the sound in time, and the constant \(|x^{-p}|\) specifies the scale of the current version of the sound. When \( p = jc \), Eq. (1) becomes

\[
S(c) = \int_{0}^{\infty} s(t) e^{(jc-1)\ln t} dt = \int_{0}^{\infty} s(t) e^{jc\ln t} dt = \int_{0}^{\infty} s(t) e^{jc\ln t} d(\ln t).
\]
This is a form of Fourier transform defined on a logarithmic time dimension, \( \ln t \). The complex variable, \( c \), plays the role of frequency on \( \ln t \) axis. The Mellin transforms of \( 'a_m' \) and \( 'a_f' \) are presented in Fig. 2(b) and the abscissa is \( c \). It is clear that the two distributions are the same except at very low values of \( c \). Therefore, this simple one-dimensional Mellin transform segregates the shape information of the vocal tract from the size information, and solves the normalisation problem for these vowel impulse responses. A comparison of the magnitude and phase distributions of the Mellin transform with those of the Fourier transform is presented in Appendix B. The one-dimensional Mellin transform is not a good model of auditory vowel normalisation because the auditory system performs a spectral analysis before the proposed normalisation process takes place. It is the case, however, that in some computational models of auditory processing, [a] vowels from vocal tracts with different lengths produce two-dimensional images that are scaled versions of each other, except that now they are scaled images rather than scaled waves. This led us to develop a two-dimensional version of the Mellin transform that can perform normalisation on vowels as they appear in the auditory image. The organisation of this section is illustrated in Fig. 4, which describes an algorithm for the calculation of the auditory Mellin transform.

2. An auditory Mellin transform

The important aspects of auditory processing for present purposes are the spectral analysis performed by the basilar membrane in the cochlea and the analysis of the time-intervals between basilar membrane peaks which is probably performed in the brainstem. The former introduces a log-frequency dimension into the internal representation of the sound, while the latter introduces a time-interval dimension into the internal representation of the sound. In the auditory image model (AIM) of perception (Patterson et al., 1995), it is assumed that these processes produce the first internal representation of sound of which we are consciously aware of; the representation is referred to as a stabilised auditory image (SAI). The SAIs of the two vowels produced when the vocal tracts associated with \( 'a_m' \) and \( 'a_f' \) are excited by a 100-Hz stream of glottal pulses are presented in Fig. 3(a) and (b). The vertical ridges are the glottal pulses in this representation; the rightwards-pointing triangles are the formants of the vowels in this representation (marked by arrows). The overall shapes of the patterns are quite similar. Relative to \( 'a_m' \), the formants of \( 'a_f' \) are shifted up along the log-frequency dimension, and they are shorter in the time-interval dimension (most evident for the second formant in Fig. 3). Scaling the vocal tract produces these changes in the SAI. In this paper, we develop a two-dimensional version of the Mellin transform that can normalise the SAI and separate the shape and size information of vowels as they appear in the auditory image. The organisation of this section is illustrated in Fig. 4, which describes an algorithm for the calculation of the auditory Mellin transform.

2.1. The stabilised auditory image

In the AIM, the initial spectral analysis that creates the frequency dimension of the image is performed by an auditory filterbank. The bandwidth of the filter is a constant proportion of the ‘centre’, or ‘peak’, frequency of the filter in the region above about 500 Hz (Glasberg and Moore, 1990). Thus, it is essentially a ‘constant-Q’ filterbank and the internal representation has a quasi-log frequency dimension as shown in Fig. 3. The impulse responses of the individual filters have gamma envelopes (de Boer and de Jongh, 1978)
and chirping carriers (Carney et al., 1999); a mathematical representation of this ‘gammachirp’ auditory filter is derived in (Irino and Patterson, 1997) as

\[ g_c(t) = a t^{-1} \exp \left( -2\pi b \text{ERB}(f_r) t \right) \times \exp \left( j2\pi f_r t + j\phi \ln t + j\phi \right) \quad (t > 0), \]

where \( a, b, \) and \( c \) are parameter values, \( \phi \) is phase, \( f_r \) is the asymptotic frequency, and \( \text{ERB}(f_r) \) is the equivalent rectangular bandwidth. Together, these findings mean that auditory spectral analysis is basically a wavelet transform (Combes et al., 1989) in the region above 500 Hz, with a gammachirp kernel (Irino and Patterson, 1997) whose parameter values are set to simulate cochlear filtering. We introduce an additional parameter, \( x \), to specify the peak-frequency of the gammachirp filter, \( x f_0 \), in terms of the base kernel frequency, \( f_0 \), and the kernel dilation factor which is \( x \). If the signal is \( s(t) \), the output of the filterbank is

\[ S_w(xf_0, t) = \int_0^\infty g_c(xf_0, \tau_1)s(t - \tau_1) d\tau_1. \]

The output of the auditory filterbank is converted into a neural activity pattern by hair cells along the edge of the basilar membrane. The process includes half-wave rectification, compression, and adaptation, which together enhance the onset of the signal and sharpen features in the filterbank output. The compression and sharpening can be ignored for present purposes; they are described in (Irino and Patterson, 1999a; Irino and Patterson, 1999b; Irino and Patterson, 1999c). Examples of the responses of these first two stages of auditory
processing to the [æ] vowel in the word ‘hat’ are presented in Patterson et al. (1995), (Fig. 2(a) and (b)). The output of primary nerve fibres innervating the inner hair cells that transduce basilar membrane motion is phase-locked to the motion up to frequencies as high as 5 kHz and so the neural activity pattern has detailed temporal information about the decay of formant resonances within glottal periods. The brainstem performs a time-interval analysis to reveal the temporal structure, and in AIM, the process is simulated by a form of strobed temporal integration (STI) and it is this processing that converts the neural pattern into an auditory image. Briefly, the activity in each neural channel is monitored to identify local maxima in the neural activity, and these local maxima are used to control temporal integration. The process operates on the envelope of the activity, specifically on the derivative of the envelope, referred to as ‘delta gamma’ (Irino and Patterson, 1996; Patterson and Irino, 1998). The local maxima occur regularly when the signal is periodic or quasi-periodic, as in the voiced parts of speech. Temporal integration is strobed on each of the local maxima and temporal integration consists of taking the current segment of the neural activity image with whatever is currently in this channel, point by point. Mathematically, if the signal is \( s(t - ktp) \), where \( t_p \) is the period and \( k \) is an integer, then the SAI has the form

\[
A_I(\alpha f_0, \tau) = \sum_{k=0}^{\infty} S_n(\alpha f_0, \tau + ktp) e^{-\xi \tau} e^{-\eta t_p},
\]

where \( \alpha f_0 \) is the peak-frequency of one auditory filter and \( \tau \) is the time-interval axis of the SAI. The exponential scalars, \( \xi \) and \( \eta \), delimit the maximum time-interval and the persistence of the auditory image. So each channel of the SAI is the decaying sum of a sequence of segments of neural activity that have been shifted by one pitch period. Physiologically, the process is similar to measuring time-intervals from moments of peak activity to peaks that are nearby in time and making a dynamic interval histogram of the results, that is, an interval histogram with a continuous decay appropriate to the decay of auditory perception (with a 30-ms half-life).

STI converts the time dimension of the neural activity pattern into the time-interval dimension shown in Fig. 3. At the same time, it removes the propagation lag associated with auditory filtering; it aligns the peaks of the responses to the glottal pulse across frequency and assigns peaks from the most recent glottal pulse to the origin of the time-interval dimension. STI is applied separately to each channel of the filterbank output; the SAI is the array of stabilised neural patterns for all the channels in the auditory filterbank. The combination of strobing and a 30-ms half-life stabilises the vowel pattern for as long as the sound is stationary.

The problem at this point in the process is to extend the Mellin transform to operate on representations like the SAI.

### 2.2. An auditory version of the Mellin transform

The version of the Mellin transform used to normalise impulse responses in Section 1.1 (Fig. 2) can be extended to auditory images by substituting the equation of the SAI (Eq. (6)) for \( s(t) \) in Eq. (1). The Mellin transform of the SAI is thus,

\[
M(h, p) = \int_0^T A_I(\alpha f_0, \tau)e^{(p-1)\ln \tau} \, d\tau.
\]

In order to use the scale-invariance property of the Mellin transform to normalise sounds for size, we must apply a constraint, represented by the parameter \( h \) on the left-hand side of Eq. (7), to the integration on the right-hand side of Eq. (7). Specifically, the requirement, that auditory images of waves that are expanded or compressed in time must all be transformed into a fixed distribution, implies a particular relationship between the frequency variable of the auditory image, \( \alpha f_0 \) and time-interval, \( \tau \). The relationship is illustrated by the motion of the second formant in the auditory images of Fig. 3. When the vocalic waveform is compressed to two-thirds its original duration, the frequency of the second formant goes up by 50% from 1100 to 1650 Hz, and the time-interval occupied by the formant decreases by a third from 9 to 6
ms. That is, the product of time-interval and peak-frequency in the auditory image is a constant. This means that to achieve scale invariance, the integration must be performed along lines in the auditory image where the product of time-interval and peak-frequency satisfies

$$\tau \cdot f_0 = h. \quad (8)$$

This in turn defines the variable $h$ in Eq. (7). The representation produced by the Mellin transform (Eq. (7)) with this integration constraint is referred to as the MI, and this is the desired auditory Mellin transform (see Appendix A for an extended description).

2.3. Size-shape image

In this section, we present a detailed illustration of the steps involved in the transformation of the SAI into an MI using the simplest of periodic sounds, a click train; MIs of vowel sounds are presented in Section 3. Fig. 5(a) shows the SAI of a click train with a click rate of 100 Hz. The click train produces a pitch like that of a male speaker with a ‘deep’ voice. The ordinate is channel peak-frequency in Hertz scaled on a quasi-logarithmic frequency axis. The abscissa is the time-interval in milliseconds from the local maximum that initiated temporal integration; it is a linear axis in this representation. The main verticals are spaced at the period of the original wave. The zero on the abscissa is the point to which the local maxima of the neural patterns are mapped during temporal integration.

2.3.1. Alignment for wavelet origin

The Mellin transform is “shift-varying” (Appendix B), so it is necessary to specify the origin of the transform at all points in time. As a result, the starting points of the individual impulse responses of the wavelet filterbank should all be aligned to zero to ensure that they are all properly a part of the same Mellin distribution. STI introduces a negative phase shift to the low-frequency channels because the strobing is initiated on local maxima. The misalignment is illustrated by the curve of pulses to the left of each main vertical in the auditory image in Fig. 5(a). The SAI can be realigned as required by the Mellin transform simply by shifting each channel to the right by one period of the carrier ($1/zf_0$) for a given auditory filter, $g_c(zf_0, t)$. The aligned SAI for Fig. 5(a) is presented in Fig. 5(b). The main verticals now provide a very good approximation to the correct starting point for each channel in this two-dimensional Mellin transform.

2.3.2. Integration limit for periodic sounds

The analysis of resonator-shape information is limited to one cycle of the auditory image because the decaying activity produced by one glottal pulse is disrupted and largely overwritten by the activity from the next pulse. Thus, for periodic sounds, $s(t – ktp)$, the limit of the Mellin integral, $T$, in Eq. (7) is set to $tp$, the period of the wave. The pattern of activity within one cycle of the auditory image is referred to as an auditory figure (AF).

2.3.3. The path of integration

The activity in the AF of the click train (Fig. 5(b)) falls along curved ridges associated with successive cycles in the impulse responses of the auditory filters. Along these impulse-response lines the product of time-interval and peak-frequency is constant (Eq. (8)). These ridges of activity are straightened by plotting the AF on a log-time axis – a transformation which reflects the operation of the In $t$ term in the kernel function of the Mellin transform. This operation converts the curved impulse response lines of the AF (Fig. 5(b)) into parallel, regularly spaced lines (Fig. 5(c)) that are essentially straight in the region above 500 Hz. The dashed lines in Fig. 5(c) show integer multiples of $h$. The activity associated with the ringing of the auditory filters in response to a click falls along the dashed lines in this representation of the AF. The slope of the lines is that of the negative diagonal across the AF frame. The dimensions of the representation are log-time-interval and log-frequen-

\[^2\] Fig. 5(c) was produced from Fig. 5(b) using spline interpolation. Fig. 5(e) is actually produced directly from Fig. 5(b); the sample points in each channel are re-sampled in proportion to the peak-frequency of the channel. This is the reason why activity reappears on the line $h = 0$ in Fig. 5(e).
Fig. 5. Auditory Mellin transform of a click train: (a) stabilised auditory image (SAI), (b) filter response alignment, (c) log-transform of the time-interval axis, (d) impulse response alignment, (e) size-shape image (SSI), and (f) Mellin image (MI). Note that the impulse response is reduced to a band of activity at low spatial frequencies, on integer values of $h$. 
cy, and this is the form that facilitates calculation of the Mellin transform.

The calculation of the Mellin transform is more easily understood if the AF with log-time-interval axis (Fig. 5(c)) is time-interval aligned to render lines of constant $h$ in vertical orientation as shown in Fig. 5(d). Each channel is time-interval shifted by an amount equal to the log of the peak-frequency of the channel. The vertical that terminates the first AF in Fig. 5(c) now appears as a curved lower bound in Fig. 5(d). The abscissa in this representation of the AF is the logarithm of the Mellin variable $h$, which is the product of time-interval and peak-frequency (Eq. (8)). The ordinate remains peak-frequency on a logarithmic frequency scale. The left-most dashed vertical shows points in the AF where $h$ is unity when the carrier of the impulse response of the auditory filter is a sinusoid. The activity produced by a click train is concentrated on verticals at integer multiples of $h$. This is emphasised in Fig. 5(e) which is a version of Fig. 5(d) with a linear $h$ axis; the response in each channel is a transformed and aligned version of the gammatone wavelet kernel. The solid curve is the upper boundary of the AF. The shapes of AFs in this representation do not change with the size of the resonator; they just move up or down the verticals as resonator size decreases or increases, respectively. Accordingly, this representation is referred to as the size-shape image (SSI). It is particularly useful for visualising the shapes associated with the AFs of vowel sounds, as will be illustrated in Section 3.3. The AF in Fig. 5(e) is derived from the left-most frame in the SAI of Fig. 5(a), but this need not be the case. The origin for the AF in the SSI can be the starting point of any of the AFs in the SAI.

2.4. Mellin image

The impulse responses of the auditory filters dominate the left half of the SSI in Fig. 5(e). This is also the case for vowel sounds where the excitation is click-like (see, for example, Fig. 6(c)). The auditory system needs to extract information about resonator shape, and for this it is useful to separate the impulse response information from other information in the SSI. Fig. 5(d) and (e) show that the peaks on verticals of constant integer, $h$, are all similar in height, indicating that the distribution of peak heights for a click train is largely flat across peak-frequency. This suggests that a form of deconvolution can be achieved by calculating the spatial Fourier transform (FT) of the SSI along the vertical line of constant $h$. In this case, the majority of the impulse response activity will be concentrated at low spatial frequencies, leaving the high spatial frequencies to present information about resonance spacing across frequencies. Many animal calls are based on pulse-resonator sound production, and auditory segregation of the pulsive excitation from the resonance information would be appropriate for auditory processing in general. In this transformation of the SSI, each vertical vector is replaced by the magnitude of the Fourier transform of the activity on the corresponding SSI vector. Specifically, it is the integration of the SSI, $A_{SSI}(z\hat{f}_b, h)$, with a kernel function that is a complex sinusoid on a log-peak-frequency axis, $\ln z\hat{f}_b$; that is,

$$M_I(h, c) = \int_{z_{\min} f_b}^{z_{\max} f_b} A_{SSI}(z\hat{f}_b, \tau) e^{-je \ln z\hat{f}_b} d(z\hat{f}_b).$$

The result is another two-dimensional image in which each vertical is the magnitude spectrum for the corresponding line of the SSI. It is this representation that is referred to as the MI. It has the same abscissa, $h$, as the SSI, but the ordinate is a new variable, $c/(2\pi)$; it shows the spatial frequencies present along the vertical line associated with a given value of $h$ in the SSI. The unit is cycles/peak-frequency range, which means that an ordinate value of unity in the MI corresponds to the spatial frequency in the SSI whose period is the full frequency range of the SSI ordinate from 100 to 6000 Hz. The vertical position of an AF in the SSI is converted into phase information in the Fourier transform, and as such, does not appear in the magnitude distribution. Thus, the MI version of the AF presents shape information about the

---

3 The gammatone is a special case of the gammachirp obtained by setting $c = 0$ in Eq. (3). The carrier is a sinusoid in this case.
Fig. 6. Stabilised auditory images (SAIs) (a) and (b), Size-shape images (SSIs) (c) and (d), and Mellin images (MI) (e) and (f) of two damped sounds. The damped sound on the left has a single sinusoid carrier of 1100 Hz, with a 2-ms half-life and a 100-Hz repetition rate. The damped sound on the right is composed of two damped sinusoids, one having a 1100-Hz carrier with a 2-ms half-life and the other having a 2600-Hz carrier with a 1.5-ms half-life; the repetition rate is 100 Hz in both cases.
resonator in a form that does not change with the size of the resonator when the scaling is uniform.

The MI version of the AF for the click train (Fig. 5(e)) is presented in Fig. 5(f). The click response is converted into activity at low spatial frequencies, as expected, because the response in the SSI on any vertical line is essentially flat. In point of fact, the magnitude of the response in the SSI rises slowly with peak-frequency because auditory-filter bandwidth increases with peak-frequency; otherwise the click response would be even more restricted in the MI. The repetition rate of the sound affects the upper limit of the frame of the AF, but the form of the MI is unchanged and the activity is little affected for speech sounds.

2.5. Equivalence of frequency domain and time-interval domain versions of Mellin transform

Finally, consider the relationship between the MI presented in Eq. (9), which is written in terms of frequency-domain integration, and the Mellin transform presented in Eq. (7), which is written in terms of time-interval-domain integration. The time-frequency constraint presented in Eq. (8) can be rewritten in logarithmic terms as

$$\ln \tau + \ln \alpha f_0 = \ln h.$$  \hspace{1cm} (10)

Since $h$ is a constant, the derivative of this equation is

$$\frac{1}{\alpha \frac{d\tau}{d\alpha}} = -\frac{1}{\tau}. \hspace{1cm} (11)$$

Substituting Eq. (11) into Eq. (9) and replacing the SSI, $A_{SSI}(z f_0, h)$, with the original AF, $A_{F}(z f_0, \tau)$, leads to

$$M_{I}(h, c) = \int_{\alpha_{\min}}^{\alpha_{\max}} A_{SSI}(z f_0, h) e^{j c \ln \alpha f_0} dz$$

$$= \int_{\alpha_{\min}}^{\alpha_{\max}} A_{SSI}(z f_0, h) e^{(j c - 1) \ln \alpha f_0 (1/\alpha)} dz$$

$$= \int_{0}^{t_p} A_{I}(z f_0, \tau) e^{(j c - 1) \ln h - \ln \tau} (-1/\tau) d\tau$$

$$= \left( -e^{(j c - 1) \ln h} \right) \int_{0}^{t_p} A_{I}(z f_0, \tau) e^{(-j c - 1) \ln \tau} d\tau. \hspace{1cm} (12)$$

When $p = -jc$, Eqs. (9) and (7) are the same except for a constant. It is also the case when the complex argument $p$ has a real part; it simply corresponds to a real exponential term of the kernel function in Eq. (9).

3. Characteristics of the SSI and the MI

In this section, we describe how the vowel sounds are represented in the SSI and the MI and the separation of size information from shape information. We show the SSIs and MIs for: (1) damped sinusoids to illustrate the basic characteristics of these images (Sections 3.1 and 3.2), (2) the synthetic male vowel, ‘$a_{m}$’, and female vowel, ‘$a_{f}$’, to illustrate the segregation of size and shape information (Sections 3.3 and 3.4), and (3) Synthetic Japanese vowels [e], [i], [o], and [u] to contrast the differences between their images (Section 3.4).

3.1. Single-damped sinusoid

A sinusoid with a decaying exponential envelope was repeated every 10 ms to produce what is referred to as a damped sinusoid (Patterson, 1994a). It is like a single format vowel. The frequency of the sinusoid was 1100 Hz, the half-life of the damped envelope was 2 ms, and the repetition rate was 100 Hz. The frequency of the carrier and the degree of damping are similar to those of the second formant of the vowel ‘$a_{m}$’ in Figs. 2(a) and 3(a). The SAI of the damped sinusoid is shown in Fig. 6(a). The repeating onsets of the damped sinusoid produce a click-like response at frequencies away from 1100 Hz; the spacing of the verticals shows the period of the sound. In the region of 1100 Hz, the impulse response is accentuated and lengthened by the resonance associated with the decaying exponential envelope. This is a common feature of natural sounds, including speech. The SSI of the AF of the damped sinusoid is presented in Fig. 6(c) and much of the activity in channels away from 1100 Hz is just like that of the click train (Fig. 5(e)). Around the 1100 Hz channel, however, the activity is extended to higher $h$ values. When $h$ is greater than or equal to 3, the activity tilts from the vertical because the filterbank
is driven by the carrier of the damped sinusoid and the instantaneous frequency is essentially constant in the 1100-Hz region. The alignment of channels in the SSI makes it easier to determine when the impulse response gives way to the resonance.

The MI for the damped sinusoid is presented in Fig. 6(e). The activity associated with the repeating onset of the damped sinusoid appears as activity on and near the abscissa in the same position as activity produced by a click train; compare the activity at values of \( c/(2\pi) \) less than 3 in Figs. 6(e) and 5(f). The activity associated with the resonance in the 1100-Hz region of the SSI introduces a weak, diagonal response at \( c/(2\pi) \) values from 5 to 15 in the MI. For values of \( h \) greater than 5, the vertical bands in the MI become progressively wider because, in the SSI, the lines in the fine structure of the feature tilt progressively more as \( h \) increases. This is the characteristic of a single resonance or formant. The banding in the MI of other damped sinusoids with a single carrier component is essentially the same, independent of the frequency of the carrier, the half-life of the envelope, and the repetition rate of the sound. The level and extent of the vertical bands increase slowly with the half-life of the damped sinusoid.

3.2. Double-damped sinusoid

The addition of a second damped sinusoid illustrates the distinction between the SAI, the SSI and the MI representations. The second damped sinusoid has a carrier frequency of 2600 Hz and a damped half-life of 1.5 ms. The repetition rate remains 100 Hz. The carrier frequency and the degree of damping are almost the same as for the third formant of the vowel ‘\( a_m \)’. The SAI for the double-damped sinusoid is presented in Fig. 6(b), and it shows activity in channels near 1100 and 2600 Hz; the lower formant extends further than the higher formant along the time-interval dimension. The SSI of the AF of the double-damped sinusoid is presented in Fig. 6(d), and it shows that in units of constant time-interval/peak-frequency product (i.e., \( h \)), the activity in the 2600-Hz region extends at least as far as that in the 1100-Hz region.

The MI for the double-damped sinusoid is presented in Fig. 6(f). The main difference between this MI and the one for the single-damped sinusoid (Fig. 6(e)) is that the bands of horizontal activity between \( h \) values of 3 and 7. It is this horizontal banding in the MI that is characteristic of vowel sounds. It arises from the interaction of the parallel bands of activity that formants produce in the SSI. For double-damped sinusoids, the bands are centred on harmonics of the spatial frequency corresponding to the distance between formants in the SSI. In this case, this distance is a little less than one-quarter of the total height of the SSI and so the bands occur at \( c/(2\pi) \) values of 4.5 and its harmonics. The banding in the MI of double-damped sinusoids is essentially fixed in position as well as shape as long as the ratio of the carrier frequencies is the same. When the ratio of the carrier frequencies decreases, the spacing between the bands in the MI decreases, and vice versa. This is a basic property of activity in the MI. In the next subsection we show that this property can be used to normalise [a] vowels synthesised by vocal tracts with the same cross-area functions and different lengths.

The activity in the MI at \( h \) values less than 3 shows that the activity in the SSI is dominated by the glottal pulse and there is very little information about the formants in this region. The SSI also shows that the energy in this region is large relative to the energy of the formants which are relatively stronger at higher \( h \) values. In traditional spectrographic analysis, activity is averaged across the glottal cycle, thus averaging the glottal-pulse response and the formant response. In the MI the two components are separated, which should facilitate the extraction of formant information.

3.3. Segregation of resonator-shape information

Two synthetic [a] vowels were constructed to illustrate how resonator-shape information can be extracted from the SSI and MI, independent of VTL and glottal-pulse rate. One of the vowels is the same as the vowel ‘\( a_m \)’ used in Fig. 3(a) where the glottal pulse rate is 100 Hz. The SAI is presented in Fig. 7(a), and it is the same as Fig. 3(a). The other [a] vowel is produced with the com-
Fig. 7. Stabilised auditory images (SAIs) (a) and (b), Size-shape images (SSIs) (c) and (d), and Mellin images (MIs) (e) and (f) for synthetic vowels ‘a_m’ and ‘a_f’. The vowel on the left is the same as ‘a_m’ in Fig. 3(a). The vowel on the right was produced from the same vocal tract as ‘a_f’ in Fig. 3(b) but with a repetition rate of 160 Hz.
pressed vocal tract used for ‘a’ in Fig. 3(b) but with a glottal pulse rate of 160 Hz. The SAI is presented in Fig. 7(b). In Fig. 7(a), the second and third formants of the vowel have centre frequencies of approximately 1100 and 2600 Hz, respectively. The spacing of the second and third formants is the same in the two figures, but the absolute positions have moved up by a factor of 50% in Fig. 7(b) to about 1600 and 3900 Hz, due to the shortening of the vocal tract. The spacing of the main verticals in Fig. 7(b) is closer together than in Fig. 7(a), reflecting the increased glottal rate.

The SSIs for the two vowels shown in Fig. 7(a) and (b) are presented in Fig. 7(c) and (d), respectively. The distinctions between the responses to the glottal pulses towards the left of the AF and the formants towards the right of the AF are enhanced in these AFs. The pattern of activity produced by the first four formants of the two vowels is very similar in the two SSIs (Fig. 7(c) and (d)); the main difference is that the pattern is shifted up together as a unit for the shorter vocal tract (Fig. 7(d)). The fifth and sixth formants shown in Fig. 7(c) also shift up with the pattern when the vocal tract is shortened; they are not visible simply because of the limited frequency range of the analysis. The other difference is the right-hand boundary of the AF, which is determined by the repetition rate of the wave, and so is more limited for the vowel with the higher pitch (Fig. 7(d)). Aside from the position of this boundary, the pitch information is absent from the SSI.

The MIs of the two vowels are presented in Fig. 7(e) and (f). The magnitude values in the MI associated with a specific value of $h$ show the spatial-frequency components of the distribution of activity in the corresponding column of the SSI. For the first few integer multiples of $h$ in the SSI of the vowel [a], the activity in the SSI is broadband in response to the glottal pulse. As a result, the activity is primarily at $c/(2\pi)$ values below about 4 in the MI. As the value of $h$ increases from 3 to 6, the formants appear as separate bands in the SSI in a somewhat more complicated pattern due to the presence of the extra formants. Then, the activity in the MI appears centred about $c/(2\pi)$ values of 3, 6, 9, 14, and 18. The addition of the first formant imposes a ripple on the activity in low-frequency channels of the SSI, which leads to a characteristic trough in the MI near $c/(2\pi)$ values of 4; this notch distinguishes these vowels from the double-damped sinusoid. These features appear in the same vertical positions in the MIs for the two vowels (Fig. 7(e) and (f)). The upper three bands at $c/(2\pi)$ values of 9, 14, and 18 are essentially the same as for the double-damped sinusoid (Fig. 6(b)). Thus, the pattern of activity produced by the vowel is horizontal bands with fixed, but irregular spacing. The fact that the general position of the pattern is fixed demonstrates the value of the MI as a means of normalising for VTL.

3.4. Other Japanese vowels [e], [i], [o] and [u]

The remaining four of the canonical Japanese vowels, [e], [i], [o], and [u] were analysed to compare with [a] and illustrate how vowel differences appear in the SSI and the MI. All five vowels were synthesised with the vocal-tract model of one male Japanese speaker but with different cross-sectional area functions for the different vowels (Yang and Kasuya, 1995). The VTLs were all about 18 cm and the glottal excitation rate was 100 Hz. The SAIs for the four vowels [e], [i], [o], and [u] are presented in Figs. 8(a), (b) and 9(a), (b), respectively. The corresponding SSIs for these vowels are presented in Figs. 8(c), (d) and 9(c), (d); the corresponding MIs are presented in Figs. 8(e), (f) and 9(e), (f).

The first formant in the SAI of [e] (Fig. 8(a)) is centred near 500 Hz, some distance from the higher formants, which are all above 1500 Hz. The higher formants of [e] in the SSI (Fig. 8(c)) are more regularly spaced than those for [a] (Fig. 7(c)), and this regularity enhances the band in the MI in the region centred on a $c/(2\pi)$ value of 12 (Fig. 8(e)). Indeed, the positions of the peaks and notches in the activity are different from those in the MI of [a]; the activity does, however, fall between $h$ values of 3 and 8 in both cases.

Compared to the SSIs of [a] and [e] (Figs. 7(c) and 8(c)), the SSI of [i] (Fig. 8(d)) has more closely clustered higher formants. The clustering leads to activity at $c/(2\pi)$ values of 13–20 in the MI (Fig.
Fig. 8. Stabilised auditory images (SAIs) (a) and (b), Size-shape images (SSIs) (c) and (d), and Mellin images (MIs) (e) and (f) of the Japanese vowels [e] (left panels) and [i] (right panels). The vowels were synthesised from a model of the vocal tract with cross-area functions when the same male speaker pronounced [e] and [i].
Fig. 9. Stabilised auditory images (SAIs) (a) and (b), Size-shape images (SSIs) (c) and (d), and Mellin images (MIs) (e) and (f) of the Japanese vowels [o] (left panels) and [u] (right panels). The vowels were synthesised from a model of the vocal tract with cross-area functions when the same male speaker pronounced [o] and [u].
8(f)), the horizontal banding is less regular than that of [a] and [e].

The SSI of the vowel [o] (Fig. 9(c)) reveals a large frequency separation between the first and second formants and the remaining upper formants. As a result, there is less activity in the MI (Fig. 9(e)) at \( c/(2\pi) \) values below 4, for \( h \) values greater than 3. The first formant is present in the SSI for \( h \) values up to 5 (Fig. 9(c)), and in the MI (Fig. 9(e)) there is activity at \( c/(2\pi) \) values of 5, 8, and 12 reflecting the spacing between the first and second formants in this region. As the first formant dies away, a set of narrow, closely spaced bands appear between \( c/(2\pi) \) values of 11 and 21. The ringing of the high formants in [o] (Fig. 9(c)) also leads to diffuse activity at low spatial frequencies extending out to high \( h \) values (Fig. 9(d)), which distinguishes this vowel from all the others.

The SSI and MI for the vowel [u] (Fig. 9(d) and (e)) are simpler because the bandwidths of the formants are relatively wide; as a result, the formants do not extend very far into the SSI or MI. There are, however, distinctive features at small \( h \) values: in the range 2–5, there is strong activity at \( c/(2\pi) \) values around 7, and in the range 4–5, there is strong activity at \( c/(2\pi) \) values around 14.

In summary, the frequency spacing and duration of the formants are encoded in the MI, in a form that is independent of voice pitch and VTL; that is, in the form of a template that is fixed in position and size in the MI.

4. Summary

We have described a Mellin transform of the stabilised-wavelet transform referred to as the auditory image and discussed how the auditory system might use some form of this stabilised wavelet-Mellin transform (SWMT) to extract the shape information associated with a given vowel class from vocal tracts of varying length. This involved: (1) specifying the role of existing auditory processes and representations, such as the auditory filterbank and auditory image within the larger SWMT framework, (2) developing a two-dimensional form of the Mellin transform that could be applied to auditory images to complete the SWMT, and (3) illustrating the relationship between the auditory image, the SSI, and the MI for vowel sounds.

The MI normalises for vocal-tract size and so the image presents vocal-tract shape information independent of size and glottal pulse rate. The size information is represented as the vertical position of the pattern of resonance activity in the SSI (or as the phase components of the Mellin transform). Thus, the size information is also available, and compactly represented within the SWMT framework.

We used a small number of synthetic vowels to demonstrate that the MI normalises vowels of one type over vocal-tract length while preserving the distinction between vowel types. The next step would be to evaluate the MI as a preprocessor for real vowels in the context of automatic speech recognition, but this is beyond the scope of the current paper.

Acknowledgements

This study was initiated while the second author was an invited researcher at ATR-HIP between October and December 1998 and completed while the first author was visiting the CNBH in August 1999. The authors wish to thank H. Kawahara of Wakayama University, and Y. Ichinose and S. Katagiri of ATR-HIP for arranging this opportunity. This work is partially supported by CREST (Core Research for Evolutional Science and Technology) of the JST (Japan Science and Technology Corporation) and a grant from ATR-HIP to the CNBH.

Appendix A. Operator for the Mellin transform

In Section 2.2, we demonstrated that the product of time and frequency plays an important role in the auditory version of the Mellin transform. A
more complete explanation is produced here using the ‘operator’ methods introduced into signal processing from quantum mechanics by Gabor (1946). The time and frequency operators are defined as $\mathcal{T} = t$ and $\mathcal{W} = -j(\frac{d}{dt})$ in the time domain. The operator associated with the Mellin variable, $p$, can be rewritten in terms of the time and frequency operators (Cohen, 1993; Irino and Patterson, 1997). Cohen (1991, 1993) introduced one such Mellin operator involving the product of time and frequency and referred to it as the ‘scale’ operator,

$$C = \frac{1}{2}(\mathcal{T}\mathcal{W} + \mathcal{W}\mathcal{T}) = \mathcal{W} - \frac{1}{2}j$$

(A.1)

since $\mathcal{T}\mathcal{W} - \mathcal{W}\mathcal{T} = j$. In quantum mechanics, this form of operator is used for ‘affine variables’ (Klauder, 1980). The corresponding transform, that is, the ‘scale transform’ (Cohen, 1993) is

$$D(c) = \frac{1}{\sqrt{2\pi}} \int_{0}^{\infty} s(t)e^{-jct} dt$$

(A.2)

So, the scale transform is the Mellin transform, in which $p - 1 = -jc - (1/2)$ or $p = -jc + (1/2)$. In Eq. (A.2), the argument is restricted in range to imaginary values with a fixed real value; we can, however, extend it to cover the entire complex plane by the introduction of a real constant $\mu$ as follows: $p - 1 = -jc + (\mu - (1/2))$ or $p = -jc + (\mu + (1/2))$. This general Mellin operator can be obtained by tracing the eigenvalue problem illustrated by Cohen (1993, p. 3278), in the reverse direction. The general Mellin operator, $C_m$, is

$$C_m = C + j\mu = \mathcal{T}\mathcal{W} + j\left(\mu - \frac{1}{2}\right)$$

(A.3)

and contains the product of the time and frequency operators. So, the corresponding Mellin variable is also closely related to the product of time and frequency values.

**Appendix B. A comparison of the shift and scale properties of the Fourier and Mellin transforms**

The Fourier and Mellin transforms have been applied to three vowels to illustrate the shift and scale properties of their magnitude and phase spectra (Fig. 10). The vowels are ‘$a_m$’, the spline compressed version of ‘$a_m$’ referred to as ‘$a_{sp}$’, and a time-shifted version of ‘$a_m$’ referred to as ‘$a_{sh}$’. The impulse responses of the vowels are presented in Fig. 10(a); the scale factor is $2/3$ and the time shift is 0.5 ms.

**B.1. Magnitude distributions**

The Fourier magnitude spectra for the vowels are presented in Fig. 10(b). Comparison of the spectra of ‘$a_m$’ and ‘$a_{sh}$’ shows that they are the same, and this illustrates the very useful ‘shift-invariant’ property of the Fourier magnitude spectrum. In other words, the Fourier magnitude spectrum of a sound is the same no matter when the transform encounters the signal (provided the duration of the Fourier integral is long with respect to the period of the lowest frequency component.) Comparison of the spectra of ‘$a_m$’ and the compressed version ‘$a_{sp}$’ shows that the spectrum is expanded by a factor of 50%, illustrating that the Fourier magnitude spectrum is not scale invariant. It is this lack of scale invariance that leads to the need for vowel normalisation as it arises from resonator size differences.

The Mellin magnitude distributions for the vowels are presented in Fig. 10(c). Comparison of the distributions for ‘$a_m$’ and the compressed version ‘$a_{sp}$’ shows that they are the same. This illustrates the very useful ‘scale-invariant’ property of the Mellin magnitude distribution which provides the basis for vowel normalisation described in the text. 5

Comparison of the spectra of ‘$a_m$’ and the shifted vowel, ‘$a_{sh}$’, shows that the distribution is expanded by a factor related to the time shift, illustrating that the Mellin magnitude distribution is not shift invariant. The shift-varying property arises from the fact that the starting point for the

---

5 The relationship between VTL and dilation of the Fourier magnitude spectrum has led several groups to apply the Mellin transform (or equivalent transforms) to the Fourier magnitude spectrum to solve the ‘dilation-normalisation’ problem (Altes, 1978; Gambardella, 1979; Davis and Mermelstein, 1980; Imai, 1983; Umesh et al., 1999). These methods, however, ignore the information in the Fourier phase spectrum.
Mellin integral is always zero (Eq. (1)). Alternatively, the appropriate starting point can be specified prior to the application of the Mellin transform, and this is the option proposed in this paper. It is often difficult to specify the starting point of a speech wave in a noisy, everyday environment. Thus, the solution proposed is to apply a wavelet transform to the wave, which simplifies the modulation pattern in any one channel, and then to apply STI which locates the starting point of

Fig. 10. Comparison of the Fourier and Mellin distributions for the vowel ‘aₐₗ’, a version that is shifted by 0.5 ms ‘aₐₗ’ and a version that is compressed by one-third ‘aₐₗ’ (a) Waveforms. The Fourier magnitude spectra (b) and the Mellin magnitude distributions (c) of the waveforms in (a). The curves have been displaced vertically for clarity. The Fourier phase spectra (d) and the Mellin phase distributions (e) of the waveforms in (a). The Mellin phase functions (e) are very steep, so a scalar proportional to e has been applied to reveal the differences.
B.2. Phase distributions

The unwrapped Fourier phase spectra for the three vowels are presented in Fig. 10(d). (The unwrapping operation is one of the difficulties associated with the phase spectrum.) Comparison of the spectra of \( a_m \) and \( a_sp \) shows that time compression leads to dilation of the phase spectrum, just as it led to dilation of the magnitude spectrum, and by the same factor. Comparison of the spectra of \( a_m \) and \( a_{sh} \), however, shows that a time shift also scales the phase spectrum; that is, the Fourier phase spectrum is not shift-invariant. Thus, if there is useful phase information in speech, it will be necessary to compensate for time shifts to normalise the phase spectra — an operation that is equivalent to locating the starting point of the impulse response for the Mellin transform. Moreover, the range of the Fourier transform has to be limited for use with real sounds, and the fixed-length windowing function used to implement the short-term Fourier transform affects the phase spectrum; changes in the position of the glottal cycle relative to the window, and changes in the period of the glottal cycle relative to window duration, both affect the phase spectrum. Indeed, the magnitude spectra produced by the short-term Fourier spectrum are also affected by the position and duration of the windowing function.

The Mellin phase distributions for the three vowels are presented in Fig. 10(c). The distribution for the shifted vowel \( a_{sh} \) is different from that of \( a_m \), indicating that the Mellin phase distribution is shift variant; however, the stabilisation process used to define the starting point for the Mellin magnitude distribution also defines the required starting point for the phase distribution. The phase distribution of the compressed vowel \( a_sp \) is a vertically scaled version of the phase distribution for \( a_m \) and the rate of expansion is constant. The rate coefficient specifies the relative size of the resonator.

Appendix C. The relationship between peripheral auditory processing and the stabilised wavelet-Mellin transform

In the text it is argued that the auditory processing of shape and size information can be regarded as a form of Mellin transform in which the input function is a stabilised wavelet transform (Irino and Patterson, 1999a,b). The relationship between auditory analysis, wavelet-Mellin analysis and spectrographic analysis is presented in Fig. 11.

With regard to spectral analysis, it has often been noted that auditory spectral analysis is like a wavelet transform rather than a Fourier transform (left column); the bandwidth of the auditory filter (gammatone, GT, or gammachirp, GC) is proportional to the peak-frequency (PF) in the region above 500 Hz whereas in the short-term Fourier transform it is fixed. But the reason why the auditory system should be more like a wavelet transform is never explained. We hypothesise that the auditory system combines the advantages of the Mellin transform with spectral analysis, and that in order to do this, the spectral analysis has to be of the wavelet form.

The advantage of the Mellin transform is that it converts dilated and compressed versions of a signal into the same distribution and so enables the system to recognise the relationship between similar resonators that differ mainly in size. The advantage of spectral analysis is that it removes off-frequency components that otherwise interfere with signal detection, and it reduces envelope fluctuations, which simplifies the problem of identifying the starting point of the glottal cycle. The question, then, is how to perform a spectral analysis within the Mellin-transform framework. When presented with expanded or compressed versions of a signal, the wavelet filterbank converts them into two-dimensional surfaces that have the same shape as the original; that is, when plotted on the usual linear time and log-frequency coordinates, the surfaces only differ in terms of their position in the frequency dimension and their scale.
in the time dimension. The Mellin transform converts all such shapes into the same distribution, which can be identified and used to recognise vowels in the real world. Thus, the wavelet transform performs a spectral analysis that is transparent to dilation; it does not distort shape information. It is also the only form of spectral analysis that avoids distorting shape information. The short-term Fourier transform is not scale-invariant on its linear frequency axis and the window size does not scale with centre frequency. This leads to the distortion of shape information.

The optimal form of the kernel for a wavelet transform operating within the Mellin framework has been derived using the minimal uncertainty constraint (Cohen, 1993). The result is a gamma-chirp kernel rather than the Gabor (1946) kernel of the Fourier transform. Irino and Patterson (1997, 1999c) have shown that the gamma-chirp filter simulates the manifestations of auditory filtering well as they arise both physiologically and psychophysically. This strongly supports the analogy between the auditory filterbank and the wavelet transform.

With regard to stabilisation (middle column), the Mellin transform, is not time invariant; it requires a mechanism to specify the starting point for all significant features in the output of the wavelet transform. In the AIM (Patterson et al., 1995), this task is performed by a form of STI which produces a pitch-synchronous pattern when it is applied to the output of the auditory filterbank. It illustrates the type of stabilisation process that must be inserted between the wavelet and Mellin transforms to produce a continuous form of wavelet-Mellin analysis that is effectively time invariant. There are undoubtedly other forms of start-point detection that would also satisfy the conditions of the Mellin transform. In the case of spectrographic analysis, the wave is segmented by a windowing function that is optimally a Gabor function, and the magnitude of the Fourier transform of the windowed wave becomes a frequency vector in the spectrogram. In the process, all temporal fine structures within the window range are integrated and turned into a single magnitude value for each frequency bin. This temporal averaging produces stabilisation when the duration of the window function is long relative to the pitch period, but in practice, this is often not the case.

With regard to pattern normalisation (right column), in wavelet-Mellin analysis, the shape information in the stabilised wavelet transform is recorded in terms of its spatial frequencies and normalised to a fixed distribution. It is not difficult to imagine a neural circuit in the cortex that could perform a one-dimensional Mellin transform (1-D MT) along topographic axes where the product of time-interval (TI) and best-frequency (BF) is con-
stant. Accordingly, an auditory analogue is proposed in the form of the MI. In speech recognition, the relationship between VTL and dilation in frames of the spectrogram has led to a mechanism for vowel normalisation involving the one-dimen-
sional Mellin transform of individual spectro-
graphic frames (Davis and Mermelstein, 1980; Imai, 1983; Umesh et al., 1999). In this case, in-
formation about dilation of the duration of for-
mant response within the glottal cycle is not
available to support shape analysis; that is, informa-
tion about the location of the band in the hdi-

References

Altes, R.A., 1978. The Fourier–Mellin transform and mamma-

Bachorowski, J., Owren, M.J., 1999. Acoustic correlates of
talker sex and individual talker identity are present in a
short vowel segment produced in running speech. J.

transform. In: Poularikas, A.D. (Ed.), The Transforms and
Applications Handbook. CRC Press (IEEE Press), Flori-
da.

slides in the impulse responses of auditory-nerve fibers. J.

Speech and Signal Process. 41, 3275–3292.


representations for monosyllabic word recognition in

de Boer, E., de Jongh, H.R., 1978. On cochlear encoding:
Potentials and limitations of the reverse-correlation

Mouton, Paris.

Fitch, W.T., Giedd, J., 1999. Morphology and development of
the human vocal tract: A study using magnetic resonance

Flanagan, J.L., 1972. Speech Analysis and Synthesis and

Gambardella, G., 1979. The Mellin transform and constant-Q

filter shapes from notched-noise data. Hear. Res. 47, 103–

Huber, J.E., Stathopoulos, E.T., Curione, G.M., Ash, T.A.,
Johnson, K., 1999. Formants of children, women, and

Imai, S., 1983. Cepstral analysis synthesis on the mel frequency
(ICASSP) 83, 93–96.

Irino, T., Patterson, R.D., 1996. Temporal asymmetry in the

Irino, T., Patterson, R.D., 1997. A time-domain, level-depen-
101, 412–419.

Irino, T., Patterson, R.D., 1999a. A mathematical framework
for auditory processing: A Mellin transform of a stabilised

Irino, T., Patterson, R.D., 1999b. Stabilised wavelet-Mellin
transform: An auditory strategy for normalising sound-
source size. In: Proceedings of the Eurospeech’99, Bud-
apest, Hungary.

Irino, T., Patterson, R.D., 1999c. A gammachirp perspective of
cochlear mechanics that can also explain human auditory
masking quantitatively. In: Proceedings of the Symposium
on Recent Developments in Auditory Mechanics, Sendai,
Japan.

Patterson, R.D., 1987. A pulse ribbon model of monaural phase

Patterson, R.D., 1994a. The sound of a sinusoid: Spectral

Patterson, R.D., 1994b. The sound of a sinusoid: Time-interval

Patterson, R.D., Allerhand, M., Giguière, C., 1995. Time-
domain modeling of peripheral auditory processing: A
modular architecture and a software platform. J. Acoust.
Soc. Am. 98, 1890–1894.

Patterson, R.D., Irino, T., 1998. Modelling temporal asymme-
2979.

University of Chicago Press, London.

transform of speech analysis. IEEE Trans. Speech Audio
Process. 7, 40–45.

Rabiner, L.R., Schafer, R.W., 1978. Digital Processing of


length and its application to vowel identification. IEEE

Yang, C-S., Kasuya, H., 1995. Dimension differences in the
vocal tract shapes measured from MR images across boy,