ABSTRACT

We hear phonemes pronounced by men, women and children 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 we extract and separate the size and shape information of sound sources. The impulse response of the vocal tract is compressed or expanded in time when the length of the vocal tract is compressed or expanded proportionally with the same cross-area function. The compressed and dilated versions of the impulse response can be converted into the same distribution using the Mellin transform. In this paper we show that the Mellin transform can be applied to the stabilised wavelet transform that forms the basis of the Auditory Image Model (AIM) of processing in the auditory pathway. The combined processing normalises source size information and produces a new, fruitful representation of source shape information, referred to as the “Mellin Image.” This “Stabilised Wavelet-Mellin Transform” (SWMT) also provides the mathematical framework for the derivation of the gammachirp auditory filterbank and the signal synchronous analysis in AIM.

1. INTRODUCTION

We hear phonemes pronounced by men, women and children as approximately the same although the length of the vocal tract varies considerably from group to group and from speaker to speaker. At the same time, we can identify the speaker group easily. Similarly, we hear the violin, cello, and viola as members of the same family of instruments while also identifying the approximate sizes of the instruments. Sound sources in the world have size and shape and it would appear to be fairly natural to map physical size and shape to the size and shape of structures in auditory representations of sources.

The impulse response of a vocal tract, or a loss-less acoustic tube, is compressed or expanded in time when the length of the vocal tract is compressed or expanded proportionally with the same cross-area function. The dilated versions of the impulse response can be converted into the same distribution using the Mellin transform [1]. Recently, a Wavelet-Mellin framework of the auditory processing has been proposed [2] to illustrate that a "gammachirp" auditory filter can be derived as a function of satisfying the minimal uncertainty constraint [3] in terms of the Mellin transform. In addition, the optimal spectral preprocessor for the Mellin transform is a wavelet transform [4] which agrees well with the current understanding of the auditory filterbank, i.e., constant-Q above 500 Hz [5]. The Mellin transform represents signals in terms of the size and shape of the signal in contrast to the Fourier transform which specifies the energy in a physical frequency band.

The Mellin transform is, however, time variant and so it cannot be applied to the wavelet transform of the signal directly. Briefly, there must be a process between the wavelet and Mellin transforms to identify the appropriate start point for the Mellin transform at all moments in time. The Strobed Temporal Integration (STI) process in the Auditory Image Model (AIM) identifies the start point of each cycle of periodic sounds in order to construct the Stabilised auditory Image (SAI)[6], and in so doing can provide the necessary start point for the Mellin transform. This suggests that pre-cortical auditory processing may be regarded as the Mellin transform of the stabilised wavelet transform of a sound, designed to extract size and shape information about the source of the sound. In this paper we describe a computational version of the auditory Mellin transform (see also [7][8] for more detail).

2. AUDITORY MELLIN TRANSFORM

2.1 The Mellin transform of the SAI

In the AIM, incoming sounds are initially analysed by an auditory filterbank; the output is converted into a neural activity pattern (NAP). The activity in each channel is monitored to identify local maxima in the activity, which are used to control temporal integration. The local maxima occur regularly when the signal is periodic or quasi-periodic as in the voiced parts of speech and sustained musical notes. Temporal integration is strobed on each of the local maxima and temporal integration consists of taking the current segment of the neural activity pattern (about 35 ms in duration) and adding it into the corresponding channel of the auditory image, with whatever is currently in that channel. The process is referred to as the Strobed Temporal Integration (STI) and the output is the Stabilised Auditory Image (SAI).

The SAI's for the click train and two versions of the vowel 'a.' are shown in Figs. 1(a), 1(b), and 1(c). One period of the pattern in the SAI corresponds to one period of the sound and is referred to as the Auditory Figure (AF). The AF represents resonance information of the vocal tract as shown in Figs. 1(b) and 1(c). When the AF is designated \( A_f(\alpha_j, \tau) \), where \( \alpha_j \) is the best-frequency of one auditory filter and \( \tau \) is the time-interval axis of the SAI, the Mellin transform of the AF is...
\[ M(h, c) = \int_0^\infty A_\mu(\alpha f_s, \tau) e^{-(\mu + \frac{1}{2})h \tau^2} d\tau. \]  

where \( t_p \) is the pitch period if the sound is periodic, and \( h \) is a constant to be defined shortly.

### 2.2 Constraint for the Mellin transform

The coefficient \(-jc + (\mu + \frac{1}{2})\) of the Mellin transform in Eq. 1 can be rewritten in terms of an operator which is the product of the time and frequency operators \([9][3][2]\). The product of time and frequency is essential to the concept of the Mellin transform. Since we need a representation that is invariant to scale change, we select the path of integration in Eq. 1 to be along lines of constant 'time-interval' - 'channel-frequency' product, i.e.,

\[ \alpha f_s \cdot \tau = h. \]  

The representation \( M(h, c) \) obtained from Eq. 1 under this constraint is referred to as "Mellin Image". In the following sections, we describe how the SAI is converted first into a size-shape image and then into a Mellin Image (MI), which is the output of the Auditory Mellin Transform.

### 2.3 Size-Shape Image

An example of a stabilised auditory image from the standard AIM is presented in Fig. 1(a). It shows just under three cycles of the pattern produced in the auditory image by a click train with a click rate of 100 Hz. The SAI is most often presented in a rectangular form with a linear time-interval axis oriented horizontally. As described in the previous section, the Auditory Mellin Transform is derived under a constraint for the integral path along a constant product of time-interval and best-frequency, \( h \). But the constant lines are curves of the impulse response in Fig. 1(a). To straighten the curves to represent the integral path more effectively, the sample points in each channel are re-sampled proportional to the best frequency of the channel. The result from the leftmost auditory figure (AF) in Fig. 1(a) is shown in Fig. 1(d) where the abscissa is \( h \).

The response in each channel is a transformed and aligned version of the wavelet kernel and has the same impulse-response length. The solid curve is the boundary of the AF. This representation is particularly useful for visualising the
shapes associated with the AF's of vowel sounds, as will be illustrated in the next subsection. The shape of an AF does not change with the size of the source in this representation; the AF just moves up or down the verticals as source size decreases or increases, respectively. Accordingly, this representation is referred to as the "Size-Shape Image", or SSI.

2.4 Mellin Image

The SSI facilitates the Mellin transform of the AF. Specifically, it is the integration of the SSI, $A_{SSI}(\alpha_f, h)$, with a kernel function that is a complex sinusoid on the log channel-frequency; that is,

$$M(f, z) = \int_{0}^{\infty} A_{SSI}(\alpha_f, h) e^{-j\omega(h/2\pi)} dh$$

The result is another two-dimensional image in which each vertical is the magnitude spectrum for the corresponding line of the SSI. The new representation is referred to as the "Mellin Image" or MI. The MI of the AF of the click train (Fig. 1(a)) is presented in Fig. 1(g). It has the same abscissa, $h$, as the SSSI, but the ordinate is a new variable, $c/2\pi$, which is the spatial frequency of the spatial Fourier transform defined on log-frequency axis. Note that it is easy to show the equivalence between Eqs. 1 and 3 by using the constraint of Eq. 2 [7][8].

The click response is restricted to the lowest spatial frequencies in the MI, because the response in the SSSI on any vertical line is essentially flat. In point of fact, the amplitude of the response in the SSSI rises slowly with best frequency because auditory-filter bandwidth increases with best frequency; otherwise the click response would be even more restricted. The repetition rate of the sound affects the upper limit on the frame width of the auditory figure, but the form of the MI is virtually unchanged and the activity is little affected for speech sounds.

The vertical position of an auditory figure in the SSSI is converted into phase information in the spatial Fourier transform, and as such does not appear in the magnitude spectrum. Thus, the MI version of the auditory figure presents shape information about the source in a form that does not change with the size of the source or the repetition rate of the excitation of the source.

3. CHARACTERISTICS OF SSI AND MI

3.1 Two versions of vowel ‘a’

Two synthetic ‘a’ vowels were constructed to illustrate the invariance properties of the size-shape image (SSI) and the Mellin image (MI). The vowels were produced using a typical vocal tract model with cross-area functions from one specific male speaker [10]. In the case of speech, the goal of the SSSI and MI is to characterise vocal tract shape independent of length and glottal pulse rate. One of the vowels had the original vocal tract length and was excited by glottal pulses at a 100-Hz rate. The stabilised auditory image (SAI) of this vowel is presented in Fig. 1(b) which shows how vocal tract resonances, i.e., formants [11], extend the impulse responses in the frequency regions of the resonances. The second and third formants of the vowel have centre frequencies of approximately 1100 and 2500 Hz, respectively. The other ‘a’ vowel was produced using a vocal tract with the same cross-area function but the length was shortened by 1/3, and the glottal rate was increased to 160 Hz. The auditory image of this vowel is presented in Fig. 1(c). The relative positions of the second and third formants are the same in the two figures, but the absolute positions have moved up by a factor of 3/2 in Fig. 2(b) to about 1600 and 3800 Hz, respectively, due to the shortening of the vocal tract. The spacing of the main verticals in Fig. 1(c) is closer together than in Fig. 1(b), reflecting the increased glottal rate.

The SSSI's for the two vowels shown in Figs. 1(b) and 1(c) are presented in Figs. 1(e) and 2(f), respectively. The distinction between the response to the glottal pulses towards the left of the auditory figure (AF) and the formants towards the right of the AF is enhanced in the SSSI's more than in the SAI's. So, the channel-based resampling in the SSSI's changes the emphasis of the formants. For example, in the SAI in Fig. 1(b), the second formant is about three times the duration of the fourth formant, but in the SSI in Fig. 1(e), the two have about the same extent on the frequency-weighted, time-interval dimension, $h$. Without the re-sampling, the higher formants would have very little affect on form of the MI. The alignment of channels in the SSSI also makes it easier to determine when the impulse response gives way to the resonance properties of the source.

The pattern of activity produced by the first four formants of the two vowels is very similar in the two SSSI’s shown in Figs. 1(e) and 1(f); the main difference is that the pattern is shifted up as a unit for the shorter vocal tract (Fig. 1(f)). The fifth and sixth formants shown in Fig. 1(e) 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. 1(f)).

The Mellin Images of the two vowels are presented in Figs 1(h) and 1(i). The ordinate of the MI is the Mellin coefficient, $c/2\pi$. The unit is cycles/best-frequency-range which means that an ordinate value of unity in the MI corresponds to a spatial frequency in the SSSI whose period is the full frequency range of the SSSI ordinate from 100 to 6000 Hz. The abscissa of the MI is the product of time-interval and best-frequency, $h_c$, as in the SSSI. The amplitude 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 SSSI. 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 spatial frequencies below about $c/2\pi$ values of 4 in the MI. As the value of $h$ increases from 3 to 6, the formats begin to appear as separate bands in the SSSI and activity appears in the MI centred about spatial frequencies of 6, 10, and 14. These features appear in the same vertical position in the
two MI’s (Figs 1(h) and 1(i)), demonstrating the value of the MI as a means of normalising for vocal-tract length.

3.2 Two Japanese vowels 'i', 'u'

Two Japanese vowels, 'i' and 'u', were selected to illustrate how vowel differences appear in the SSI and MI. These vowels are distant from 'a' in the traditional vowel quadrilateral [11] (See [7] for the Japanese vowels 'e' and 'o'). All vowels were synthesised with the vocal tract model of one male speaker but with different cross-area values for the different vowels [10]. The vocal tract length was fixed as was the glottal excitation rate which was 100-Hz. The SSI’s for the two vowels 'i' and 'u' are presented in Figs. 2(a) and 2(b), respectively; the corresponding MI’s are presented in Figs. 2(c) and 2(d).

Compared to the SSI of 'a', the SSI of 'i' has more closely clustered higher formants (compare Figs. 2(a) and 1(e)). The formants also extend to higher \( h \) values in the SSI of 'i'. The clustering leads to activity at \( c/2\pi \) values of 13 - 20 in the MI (Fig. 2(c)) extending to \( h \) values of 10, whereas the activity in the MI of 'a' (Fig. 2(e)) is centred on \( c/2\pi \) values of 10, 14, and 18 and is limited to \( h \) values of about 8.

The SSI and MI for the vowel 'u' (Figs 2(b) and 2(d)) are simpler because the bandwidths of the formants are wider; as a result, they do not extend far out 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 temporal range of the formants is directly reflected in the MI, and so it provides unique templates for the vowels that are independent of voice pitch and vocal tract length; that is, the template is fixed in position in the MI.

4. SUMMARY

We have described a "Stabilised Wavelet-Mellin Transform" (SWMT) for extracting source information from waveforms and an auditory version of the SWMT for extracting vowel information from speech sounds. This involved 1) specifying the role of existing auditory processes and representations, 2) developing a form of the Mellin transform that could be applied to stabilised auditory images (SAI’s) to complete the SWMT, and 3) illustrating the relationships between the SAI, the size-shape image (SSI), and the Mellin image (MI) for vowel sounds. This process normalises for source size and so the MI presents source shape information independent of source size. The size information is also available from the vertical position in the SSI or the phase components in the Mellin transform. The SWMT also provides the mathematical framework for the derivation of the gammachirp auditory filterbank [2] and the signal synchronous analysis in AIM [6].

ACKNOWLEDGEMENTS

This work is partially supported by CREST.

REFERENCES