Multiresolution Elementary Tonotopic Features for Speech Perception

presented at the 1997 International Conference on Neural Networks, Houston, Texas

Elaine Y.L. Tsiang
et@monowave.com

Abstract

We define multiresolution elementary tonotopic features (ETFs) in general, and present specific functions and decompositions for computing them. Such decompositions, when cast in the form of local, fixed-weight FIR neural networks, have definite architectures. Results of their use as front-end inputs to a speaker-independent continuous-speech phoneme recognizer are encouraging. We analyze the dependence of the recognition performance on the various ETFs at different levels of resolution.

1. Introduction

There exist intriguing circumstantial evidence that the neural computations for vision and audition may be fundamentally similar [1][2]. Recent results from disjoint-band perceptual experiments [3] further emphasize the local, multichannel and redundant nature of the early feature extraction stage, which seems to enable a later, phoneme formation stage to preserve (or restore) phoneme perception against additive and convolutive noise[4]. We propose a model of auditory feature extraction as the decomposition of an initial spectral representation in tonotopic and time into elementary features at multiple resolutions (or scales). By tonotopy, we mean any monotonically increasing function of frequency, inclusive of the simple logarithm, or the Mel scale. The original inspiration came from the intuition that Nature could do no less than evolve an analysis as complete and comprehensive as wavelets. Recent neurophysiological data suggest that this may indeed be the case [5].

2. Elementary tonotopic features

A tonotopic feature (TF) is a family of self-similar FIR filters over tonotopy-time, indexed by tonotopy. An elementary tonotopic feature is a TF that detects the presence of tones, their onset, rise or fall. A multiresolution ETF is a family of ETFs obtained by dilations in both tonotopy and time[6]. We refer to the output from a TF as a map, and a logical grouping of maps at a particular resolution as a layer.

One possible realization of ETFs is as Gabor transforms [7]. We generalize the notion of “ripple” as periodic variation of a spectrum along tonotopy, X, to local periodic variation along any direction in the 2-D space of tonotopy and time, $X \otimes T$. Let (u, v) be the local co-ordinate system at any point in $X \otimes T$. Given a ripple resolution of δ in u, and ε in v, the transforms are, to within a normalizing factor:

$$g^{odd}_{\delta,\varepsilon,\Theta,X}(u, v) = e^{-\frac{1}{2}\cdot\frac{u^2}{\delta^2}} \cdot e^{-\frac{1}{2}\cdot\frac{v^2}{\varepsilon^2}} \cdot \sin(Ku)$$

(eq. 1)

$$g^{even}_{\delta,\varepsilon,\Theta,X}(u, v) = e^{-\frac{1}{2}\cdot\frac{u^2}{\delta^2}} \cdot e^{-\frac{1}{2}\cdot\frac{v^2}{\varepsilon^2}} \cdot \cos(Ku)$$

where K is the ripple frequency along u. Let (x, t) be the local-co-ordinate system parallel to the global co-ordinate system. (u, v) is obtained by rotating (x, t) through Θ:

$$u = x\cos\Theta - t\sin\Theta$$

(eq. 2)

$$v = x\sin\Theta + t\cos\Theta$$

The $u = 0$ trajectory gives the local frequency modulation:

$$\frac{x}{t} = \tan\Theta \qquad \text{(eq. 3)}$$

When $\Theta = 0$, the transform detects a locally stationary tone at X. $\Theta > 0$ detects a locally rising tone, and $\Theta < 0$ detects a locally falling tone. $\Theta = \frac{\pi}{2}$ detects onset or

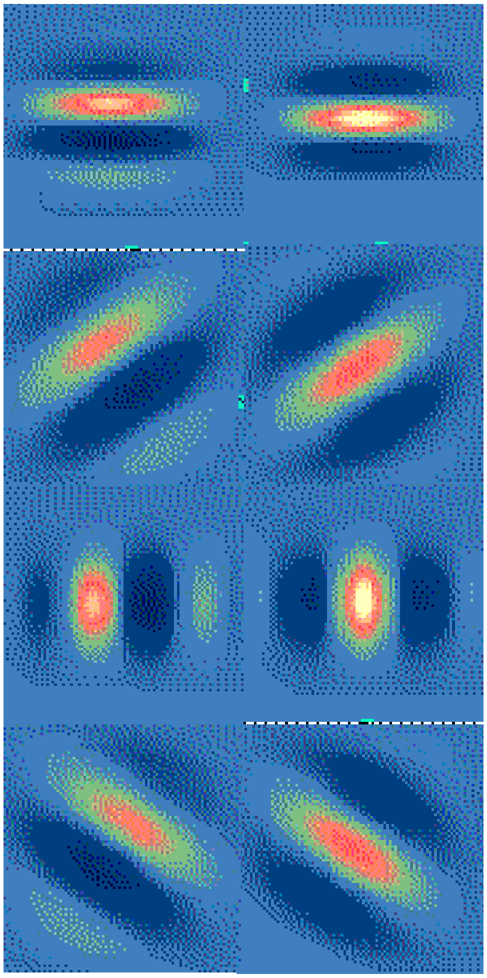

Figure 1. Receptive fields of ETFs.

offset of tones within a bandwidth of δ around X. Figure 1 shows these receptive fields with $\tan\Theta = \pm\frac{2}{3}$, $\frac{1}{\delta \cdot K} = \sqrt{2} - 1$ and aspect ratio $\frac{\delta}{\varepsilon} = \frac{2}{3} \cdot \frac{1}{\delta}$ determines the ripple bandwidth, and $\frac{1}{\varepsilon}$ the frequency modulation, or Θ, bandwidth[7].

Wang and Shamma give an alternative, more neurophysiologically motivated, realization using a 1-D exponentially decaying, rather than a 2-D Gaussian, envelope for the transforms[9].

3. Network architecture for multiresolution decomposition

The Gabor transforms are a set of non-orthogonal basis functions. In practice, a reduced set such as (eq. 1) is sufficient for good approximations. Moreover, for a fixed Θ, with sufficiently dense sampling and sufficiently large overlap, each half-resolution Gabor transform may be approximated by a weighted superposition of self-similar full-resolution transforms centered at the sampling points. The weights for such superpositions may be found by an optimal reconstruction procedure[8]. They constitute TFs in themselves. Let **c−b** denote the full-resolution ETFs, and **c** the output from applying **c−b**. **c** consists of 8 maps, from applying the odd and even transforms for each of onset, rise, tone and fall. Let **l**1, **l**2 and **l**3 denote half, quarter and eighth resolution outputs. The superposition TFs may now be denoted by **l**1−**c**. The **l**1− ETFs may be approximated by first applying **c**−, then **l**1−**c**. One additional application of **l**1−**c** yields output **l**2, yet one more yields **l**3. In this decomposition, each layer is fully redundant with the next. This redundancy can be removed by splitting each full-resolution Gabor transform into two parts, a low-pass part congruent with the half-resolution transforms, and a high-pass one orthogonal to it. We denote the orthogonal part by **h**. A set of TFs, **h**1−**c**, analogous to **l**1−**c**, may be constructed for **h**. The layers **h**1, **h**2, **h**3 and **l**3 contain non-redundant information. This decomposition is a partially orthogonal wavelet decomposition [10].

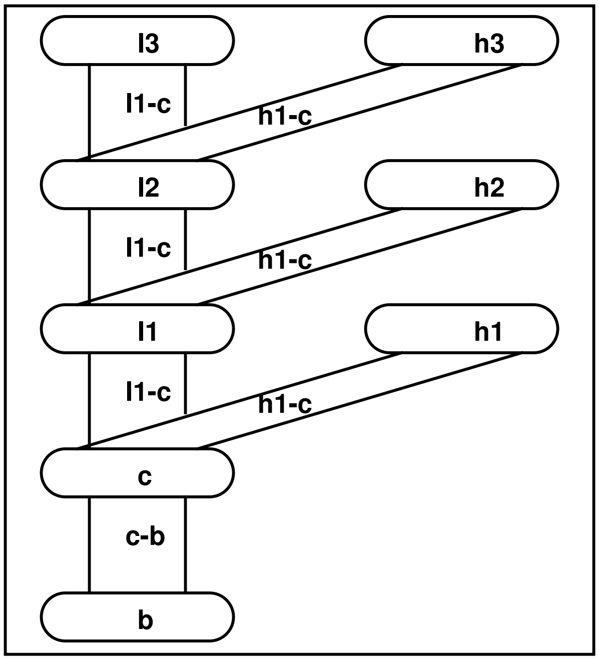

Figure 2. Network architecture

4. Multiresolution ETF maps

Figure 3 and Figure 4 show some of the outputs from a 4-resolution decomposition of a speech signal for the utterance “castle left”. We sample the signal at 13.3KHz, and analyze it with a 103-channel cochlea filter bank spanning 400Hz to 6KHz[6], decimating the output down to 1KHz. We analyze the cochlea output, **b**, with the full-resolution ETFs to obtain the 8-map layer **c**, each with 45 channels, sampled at 166.7 Hz. We then successively apply the **l**1−**c** and **h**1−**c** TFs to obtain the layers at the lower resolutions. These layers have 21 channels, and a sampling rate of 83.3 Hz.

5. Application to phoneme recognition

We have used the multiresolution ETFs as a front-end to a neural-network-based speech recognizer[6]. A full description of this upper net is beyond the scope of this paper. However, its architecture is consistent with the lower net: local, independent filtering at multiple channels, and hierarchical. It takes as input the full complement of ***h***1, ***h***2, ***h***3, ***l***3 layers. We report here the performance on a limited-vocabulary speaker-independent, continuous-speech phoneme recognition experiment.

5.1 Speech data

We collected the samples in an office environment using a Nady MCM-400 headset microphone, a Radio Shack SSM-100 mixer. The vocabulary consists of 25 words for playing chess. These 25 words require almost the full set of phonemes, missing 1 labial nasal, 2 voiced stops - dental and velar, and 3 voiced fricatives - dental, alveolar and palatal. The training set consists of 910 utterances of 88 chess commands from 84 male and 62 female speakers. The test set consists of 492 utterances from 33 male and 20 female speakers.

5.2 Performance

We scored the phoneme recognition on a per-frame basis at the sampling rate of 83.3Hz. The best recognition rate we obtained on the test set was 87.2% accuracy. Although no direct comparison can be made with other current phoneme recognition results, this result is encouraging in view of the limited amount of training data used, and the stringent scoring. The phoneme recognizer is currently deployed in a speech-driven chess playing demonstration

5.3 Influence of ETFs

We analyzed the dependence of the phoneme recognition accuracy on the ***h***1, ***h***2, ***h***3 and ***l***3 ETF maps. Table 1 gives the accuracy deficit when the corresponding ETFs are excised from the system. The accuracy deficit may be used as a relevance measure. Table 1 shows an interesting pattern: stationary tone information is relevant only at the coarser resolutions, but spectral dynamics (the onset, rise and fall of tones) are relevant at all resolutions. This simply confirms the non-stationary nature of the speech signal. The greater relevance of the coarser ETFs also suggests this hypothesis: that the ETFs at each resolution provide acoustic context for the higher-resolution ETFs. It follows that the coarser ETFs are more often involved in the formation of complex features, and are therefore more relevant. Note also how very localized onsets become less important relative to frequency modulations in ***h***1.

TABLE 1. Relevance of the different ETFs

	onset	fall	rise	tone
h1	1.0	3.0	4.5	0.7
h2	9.1	5.1	4.0	0.9
h3	12.0	11.4	7.8	2.9
l3	21.1	19.3	11.7	7.3

6. Conclusion

Multiresolution ETFs exhibit the characteristics of being local, multichannel, and redundant. Our success in using them as a front-end to a speaker-independent, continuous-speech recognizer trained on a very modest amount of data indicates that this model is capable of capturing the critical information about the dynamics of the speech signal.

Acknowledgments

This research was supported by Monowave Partners, L.P. and Steven A. Reisler.

References

[1] A.R. Luria. *The Mind of a Mnemonist*, Harvard University Press, 1987.

[2] D.D.M. O'Leary, "Do cortical areas emerge from a protocortex?", *Trends Neuronscience*, vol. 12, no. 10, pp. 400-406, 1989.

[3] R.P. Lippmann, "Accurate consonant perception without mid-frequency speech energy", *IEEE Trans. Speech and Audio Processing*, vol. 4, no. 1, pp. 66-69, 1996.

[4] J.B. Allen. "How do humans process and recognize speech?", *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 4, pp. 567-577, 1994.

[5] C.E. Schreiner and M.L. Sutter, "Functional topography of cat primary auditory cortex: distribution of integrated excitation", *J. Neurophysiology*, vol. 64, no. 5, pp. 1442-1459, 1990.

[6] E.Y.L. Tsiang, "System for recognizing speech", U.S. Patent 5377302, 1994.

[7] J.G. Daugman, "Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters", *J. Optical Society America*, vol. 2, no. 7, pp. 1160-1169, 1985.

[8] J.G. Daugman, "Complete discrete 2-D Gabor transforms by neural networks for image analysis and compression", *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 36, no. 7, pp. 1169-1179, 1988.

[9] K. Wang and S.A. Shamma, "Spectral shape analysis in the central auditory system", *IEEE Trans. Speech Audio Processing*, vol. 3, no. 5, pp.382-395, 1995.

[10] S.G. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 11, no. 7, pp.674-693, 1989.

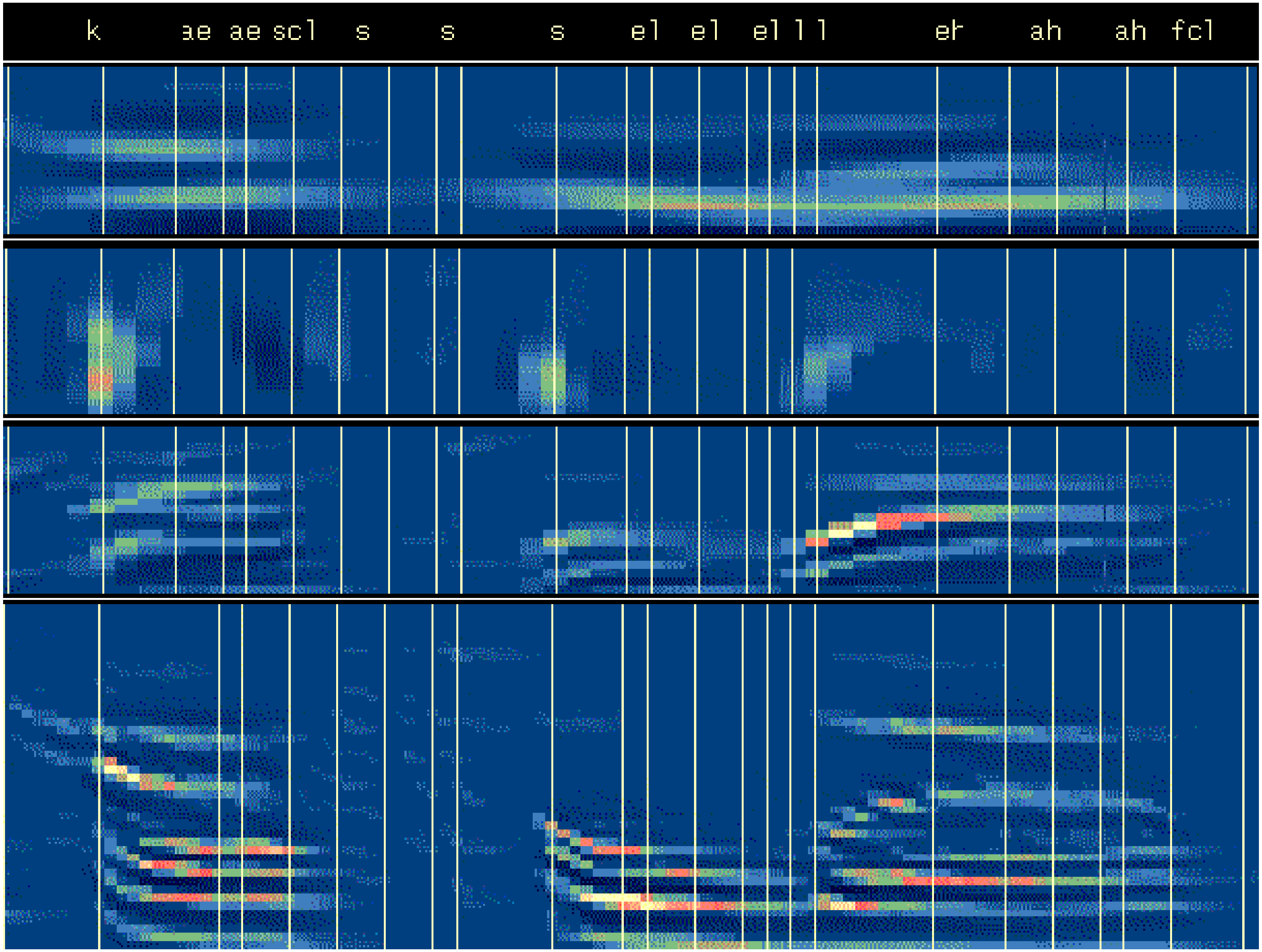

Figure 3. Maps for $\mathcal{I}3$ tone,$\mathcal{I}2$ onset,$\mathcal{I}1$ rise and $\mathcal{C}$ fall

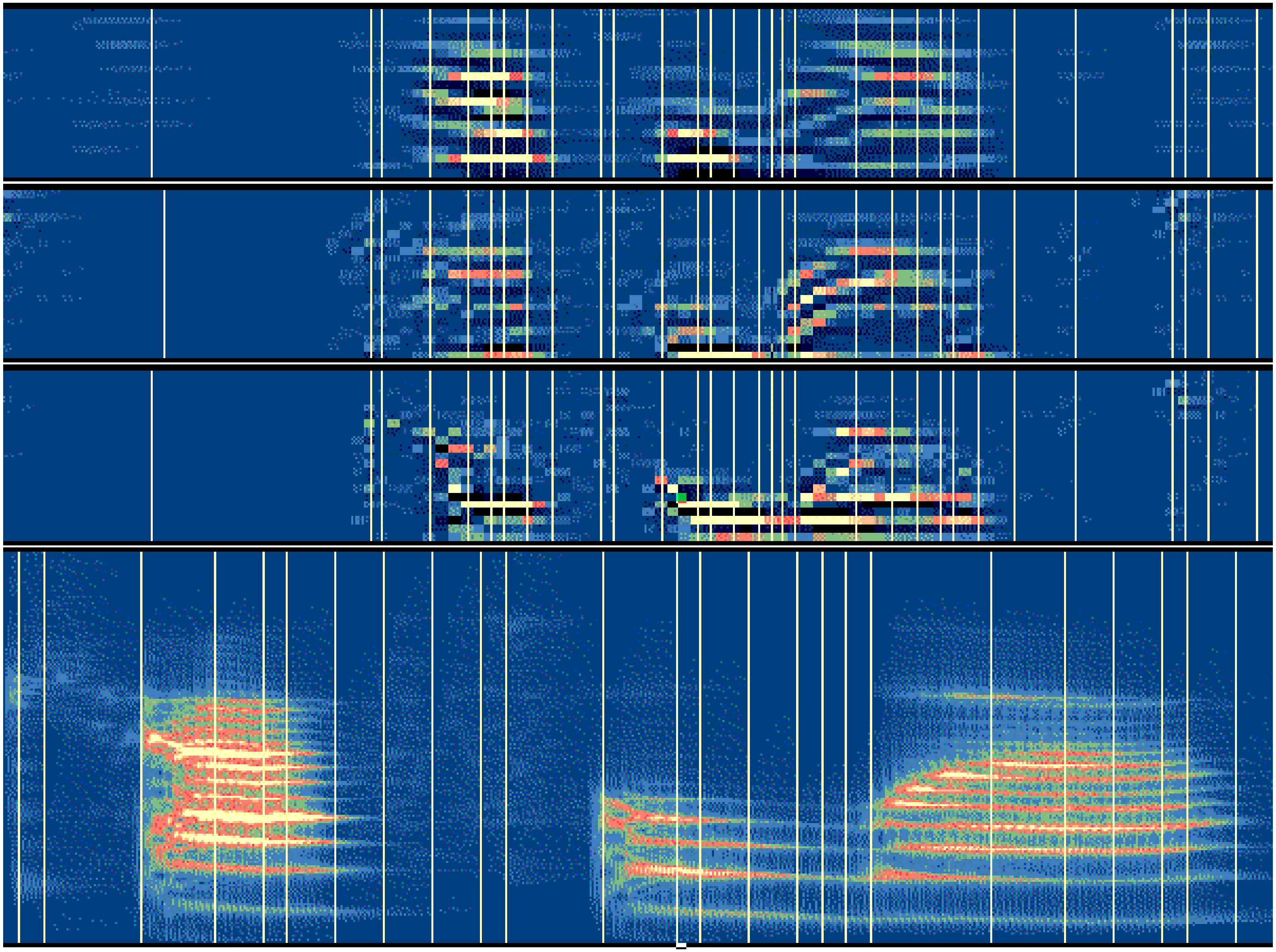

Figure 4. Maps for h3 tone, h2 rise, h1 fall and b