PITCH-SENSITIVE COMPONENTS EMERGE FROM HIERARCHICAL SPARSE CODING OF NATURAL SOUNDS

Engin Bumbacher[1] and Vivienne Ming[2]
[1]*School of Computer and Communication Science, EPFL, Lausanne, Switzerland*
[2]*Socos LLC, San Francisco & Visiting Scholar, Redwood Center for Theoretical Neuroscience U. C., Berkeley, U.S.A.*

Keywords: Pitch perception, Gist, Sparse coding, Generative hierarchical models, Gaussian mixture models, Bayesian inference, Auditory processing, Speech processing.

Abstract: The neural basis of pitch perception, our subjective sense of the tone of a sound, has been a great ongoing debates in neuroscience.Variants of the two classic theories - spectral *Place* theory and temporal *Timing* theory - continue to continue to drive new experiments and debates (Shamma, 2004). Here we approach the question of pitch by applying a theoretical model based on the statistics of natural sounds. Motivated by gist research (Oliva and Torralba, 2006), we extended the nonlinear hierarchical generative model developed by Karklin et al. (Karklin and Lewicki, 2003) with a parallel *gist pathway*. The basic model encodes higher-order structure in natural sounds capturing variations in the underlying probability distribution. The secondary pathway provides a fast biasing of the model's inference process based on the coarse spectrotemporal structures of sound stimuli on broader timescales. Adapting our extended model to speech demonstrates that the learned code describes a more detailed and broader range of statistical regularities that reflect abstract properties of sound such as harmonics and pitch than models without the gist pathway. The spectrotemporal modulation characteristics of the learned code are better matched to the modulation spectrum of speech signals than alternate models, and its higher-level coefficients capture information which not only effectively cluster related speech signals but also describe smooth transitions over time, encoding the temporal structure of speech signals. Finally, we find that the model produces a type of pitch-related density components which combine temporal and spectral qualities.

1 INTRODUCTION

Pitch is the subjective attribute of a sound's fundamental frequency that is related to the temporal periodicity of the waveform. As such, it refers to several distinct percepts which include *spectral pitch* (evoked by a single tone), *periodicity pitch* (evoked by harmonic complex tones that are spectrally resolved by the cochlea) and *residue pitch* (a low pitch associated with the periodicity of the total waveform of a group of high harmonics - the residue - that are spectrally unresolved by the cochlea) (Shamma, 2004). Both the periodicity and the residue pitch do not require energy at the fundamental frequency of the complex tones (phenomenon of the *missing fundamental*). There has long been a debate about the mechanisms that give rise to these different pitch percepts, with a classical distinction between the Place and Timing theories (Griffiths et al., 1998). The traditional place theories explain pitch perception in terms of the pattern of excitation produced along the tonotopically organized basilar membrane. Pitch could then be computed via template matching (Shamma, 2004). On the other hand, time theories promote the idea that pitch is related to the time pattern of neural activity across the auditory nerve. A global pitch percept emerges from the dominant periodicity computed from the activity of the cochlear neurons phase-locked to the corresponding individual harmonics of the sound complex (Griffiths et al., 1998). In the case of periodicity pitch, both theories are able to explain how the frequencies of the harmonics are determined. When it comes to residue pitch, the place theory fails to identify the pitch of complex tones when there is no well-defined spectral structure or when all the harmonics are unresolved (Griffiths et al., 1998), as opposed to the time theory. In the course of time, physiological and psychophysical research has collected evidence and described phenomena supporting both theories. Oxenham et al. (Oxenham et al., 2004) have recently con-

Bumbacher E. and Ming V. (2012).
PITCH-SENSITIVE COMPONENTS EMERGE FROM HIERARCHICAL SPARSE CODING OF NATURAL SOUNDS.
In *Proceedings of the 1st International Conference on Pattern Recognition Applications and Methods*, pages 219-229
DOI: 10.5220/0003786802190229

ducted experiments whose results indicate that complex sounds with identical temporal regularity could produce different pitch percepts, which is strongly in favor of the place theory. By contrast, Shannon et al. (Shannon et al., 1995) showed that speech recognition is possible with only temporal cues, and work by Griffiths (Griffiths et al., 1998) and Patterson (Patterson et al., 2002) indicates that pitch can be produced without a set of harmonically related peaks in the internal spectrum.

As an alternative approach to analysing neural response to stimuli, we looked directly at the statistical structure of naturally occurring sounds, such as human speech, by means of an extended version of the generative hierarchical model developed by Karklin and Lewicki (2003, 2005) - the so-called density component model. This model is a generalization of linear efficient coding methods such as ICA (Bell and Sejnowski, 1995) and sparse coding (Olshausen and Field, 1996) in which the coefficients of the linear filters are no longer assumed to be independent. Karklin and Lewicki (2003, 2005) have shown that their model captures higher-order statistical regularities that reflect more abstract, invariant properties of the signal. However, these statistical models are generally implemented such that the inference process is based on randomly initialized stochastic gradient descent methods of the maximum a posteriori approximations of the probability distributions. As the space of the posterior probability distribution is highly nonlinear, the inference process is significantly affected by its initialization, not only in terms of stability but also in terms of the information captured by the inferred coefficients. Hence, random initializations introduce a systemic bias to the encoding stage, affecting the learning process of the density components.

Here, we have extended the density component model in order to address these shortcomings. First, we replaced the lower layer with an overcomplete sparse coding model, in line with studies of Lewicki and Sejnowski (2000) showing that overcomplete representations increase the efficiency of the code and its flexibility to encode various signal structures. Secondly, we softened the pure bottom-up approach of the encoding process by incorporating a pathway that serves as the initialization step of the inference process of the higher layer of the density component model based on a coarse *gist representation* of the respective sound segment. We refer to this pathway as the *gist pathway*. As such, the gist pathway moves the system into an appropriate - in terms of the gist representation - region of the posterior probability distribution and thus eliminates the randomness of the former initialization process. In other words, this pathway acts as a predictive mapping of the sound segment into the space of higher-order structural features, by means of the gist information extracted from the segment itself.

Having applied the model to human speech signals, we show that the learned higher-level representations are strikingly different when the gist pathway is implemented. They are significantly better adapted to the modulation spectrum of the speech signals. Furthermore, these higher-level representations incorporate several types of components that account for pitch-encoding that have not been reported previously. These types encompass both harmonic templates and units that combines both temporal and spectral qualities. Derived from information theoretic approaches alone, these units shed a new light on the debate about the different mechanisms that give rise to pitch perception. Finally, the inferred coefficients not only enable intuitively meaningful clustering of speech signals but also exhibit smooth transitions over time, which can be used for further structural encoding.

2 EXTENDED DENSITY COMPONENT MODEL

The extended density component model is an hierarchical generalization of the sparse coding model (Olshausen and Field, 1996). It builds on the density component model of Karklin and Lewicki (2005), but further incorporates an additional pathway as described in section 2.1. Likewise, the data is assumed to be generated as a combination of a set of linear basis functions $\mathbf{a}_i$. In matrix form,

$$\mathbf{x} = A\mathbf{u} + \varepsilon, \tag{1}$$

where the $\mathbf{a}_i$ are the columns of A, and $\mathbf{u}$ are the basis function coefficients. Assuming the noise ε to be Gaussian, we get

$$p(\mathbf{x}|A,\mathbf{u}) \propto \exp\left(-\sum_i \frac{1}{2\sigma_\varepsilon^2}(x_i - \sum_j A_{ij}u_j)^2\right). \tag{2}$$

In our case, $\mathbf{x}$ are sound pressure waveforms of human speech.

The standard efficient coding models assume the basis function coefficients independently follow generalized Gaussian distributions with equal variances λ,

$$p(\mathbf{u}) = \prod_i z \exp\left(\left|\frac{u_i}{\lambda}\right|^q\right), \tag{3}$$

where $z = q/(2\lambda\Gamma(1/q))$ is the normalizing constant and λ is usually is fixed to one. However, the density component model goes a step further by capturing

the dependence among the linear coefficients through their respective variances, thus accounting for local deviations from the unit variance assumed by the standard models. It is assumed that the set of λ values can be modeled with a linear combination of density components B_{dc} and coefficients $\mathbf{v}_{dc}$ as follows:

$$\lambda = c\exp\left(B_{dc}\mathbf{v}_{dc}\right). \tag{4}$$

We set $q_i = 1$ for all i to model the linear coefficients to be sparse; and with $c = \sqrt{\Gamma(1/q_i)/\Gamma(3/q_i)} = 1$ and equation (4), equation (3) becomes

$$p(\mathbf{u}|B_{dc},\mathbf{v}_{dc}) = \prod_i \frac{1}{2\lambda_i}\exp\left|\frac{u_i}{\lambda_i}\right|. \tag{5}$$

Thus, the logarithm of the joint prior distribution of the coefficients $\mathbf{u}$ can be written as

$$-\log p(\mathbf{u}|B_{dc},\mathbf{v}_{dc}) \propto \sum_{ij} B_{dc,ij}v_{dc,j} + \\ + \sum_i \left|\frac{u_i}{\exp\left([B_{dc}\mathbf{v}_{dc}]_i\right)}\right| \tag{6}$$

(for derivation see appendix). Placing a sparse factorable prior on the latent variables $\mathbf{v}_{dc}$,

$$p(\mathbf{v}_{dc}) = \prod_i p(v_{dc,i}) = \prod_i \exp\left(\left|\frac{v_{dc,i}}{\mu_{dc}}\right|\right), \tag{7}$$

constrains them to independence, while independence of the coefficients $\mathbf{u}$ is now conditioned on these higher-level coefficients.

The probability density function for the linear model (1) is obtained by marginalizing over the coefficients

$$p(\mathbf{x}|A,B_{dc}) = \int p(\mathbf{x}|\mathbf{u},A)p(\mathbf{u}|B_{dc})d\mathbf{u} \\ = p(\mathbf{u}|B_{dc})/|\det A| \tag{8}$$

with

$$p(\mathbf{u}|B_{dc}) = \int p(\mathbf{u}|B_{dc},\mathbf{v}_{dc})p(\mathbf{v}_{dc})d\mathbf{v}_{dc}. \tag{9}$$

As calculating these integrals is computationally intractable, we approximate them by their maximum a posteriori (MAP) estimations which are calculated by means of gradient descent algorithms.

To distinguish between the matrices A and B, we will refer to the columns of A linear features or *sparse components* (SC) and the columns of B_{dc} *density components* (DC).

2.1 Gist Pathway

As described in the introduction, the space of the posterior probability distribution is highly nonlinear, characterized by a vast number of local extrema. Due to this nonlinearity, working with the maximum a posteriori estimates of the coefficients makes the inference process very sensitive to the initialization, as the gradient descent algorithm inherently only finds local extrema. Thus, different randomly initialized inference runs for the same sound segment lead to different inferred coefficients.

We have incorporated an initialization step of the inference process of the density component coefficients that is entirely data-driven and hence deterministic. Motivated by research on gist (Oliva and Torralba, 2006) (Harding et al., 2007), we refer to the initialization step as the gist pathway. While the purpose of the sparse component layer is to establish an accurate sparse representation of the initial signal itself, the gist pathway is designed to be a fast processing step that extracts globally meaningful information (*gist*) about the coarse spectrotemporal structure of the signal. For example, in the case of a signal with predominant power in the high frequencies, the gist pathway does not determine the single contributions of the sparse components, but rather captures the overall characteristic - high pitch - and thus initializes the inference process of the density component coefficients $\mathbf{v}_{dc}$ by favoring the density components that best capture the corresponding frequency properties of the signal.

Thus, in order to determine the gist of a given sound pressure waveform $\mathbf{x}$, the spectrogram of the sound encompassing this segment and parts of the preceding and the subsequent signal is computed and projected into the space of the significant principal components. This provides a coarse, phase-invariant representation $\mathbf{u}_G$ of the sound of interest. This works well as the spectrogram is an estimate of the local spatiotemporal power in a sound, which is therefore related to the variance variables λ in the density component model.

In the next step, the sound is further processed by applying a sparse coding model on these projections $\mathbf{u}_G$, based on the assumption that they can be written as a linear superposition of gist basis functions B_G:

$$\mathbf{u}_G = B_G\mathbf{v}_G \tag{10}$$

with a sparse set of coefficients $\mathbf{v}_G$. Thus, within the standard sparse coding approach, under the assumption of additive Gaussian noise, the cost function to be minimized is given by the log-posterior probability of the coefficients (similar to (2))

$$\log p(\mathbf{v}_G|\mathbf{u}_G,B_G) = \log p(\mathbf{u}_G|\mathbf{v}_G,B_G) + \log p(\mathbf{v}_G) \\ \propto -\frac{1}{2\sigma_\varepsilon^2}\|\mathbf{u}_G - B_G\mathbf{v}_G\|^2 - \sum_k \left|\frac{v_{G,k}}{\mu_G}\right| \tag{11}$$

Here, the distribution of the coefficients $\mathbf{v}_G$ is modeled as a Laplacian with uniform variance, as in equation (7). By setting the number of linear features (columns of the matrix $B_G = [\mathbf{b}_{G,1}, \mathbf{b}_{G,2}, ..., \mathbf{b}_{G,M}]$) equal to the number of density components, the inferred coefficients $\mathbf{v}_G$ serve as the initialization to the inference process of the higher-order coefficients B_{dc}. As the gist basis functions encode the activity of principal components of the spectrogram of the sound on a broader timescale than density components, the gist pathway turns out to provide the density components with additional information about the sound.

We predict that such a gist-modulated prior on the $\mathbf{v}_{dc}$ enables the second layer of the density component model to encompass a broader range of structures of the sound, as the gist pathway provides a robust representation of the broad-scale sound, in addition to the sparse components, and as such drives the system towards a more likely representation of the sound signal.

Details to the inference process can be found in the appendix.

3 METHODS

We provide results and analyses for the TIMIT speech corpus (Garofolo et al., 1990), which includes a diverse group of native English speaker reading phonetically diverse English sentences. The sampling rate has been converted to 8kHz.

We have a two times overcomplete[1] set of sparse components A, and 50 density components. The length of the sound extracts T was set to be 20 ms. This time length is on the same order of magnitude as the temporal extent of formants and formant transitions (Harding et al., 2007) (Turner and Sahani, 2007). In order to remove second-order correlations, the set of sounds has been whitened.

The MAP posteriors are estimated by means of conjugate gradient descent software (Olshausen and Field, 1997).

4 RESULTS

4.1 Density Components of the Fully Extended Model

The density components presented in this section are the results of training the fully extended density component model. In order to interpret the weights of a density component, we first characterize each sparse component as ellipses in the spectrotemporal domain. Each ellipse is centered around the center of frequency and center of the temporal envelope of the corresponding sparse component. The height and width of each ellipse corresponds to the bandwidth and temporal envelope of the component. We then color each ellipse based on the weight of a given density component. Patterns in the organization of the sparse components revealed by this visualization show how each density component captures meaningful dependencies among the sparse components in the time-frequency domain. The results are illustrated in figure 1.

We find that some of the density components shown in figure 1 bear resemblance to the ones Karklin and Lewicki (2005) reported (e.g. top row). However, our use of a logarithmic frequency axis, in accordance to the tonotopic organization of the cochlea (Shamma, 2001), reveals more spectrotemporal interdependencies between the sparse components. Such as the very specialized type of density components that encodes the phase-locking relationship between the amplitude modulation of mid- and high-frequency linear features and very few low-frequency linear features (figure 1 bottom row). The possible role of these types of density components is further discussed in the following subsection.

Representing the density components by the centers of mass of the modulation spectra of each density components allows us to characterize the population of density components. Thus, we can compare the signal structure encoded by the population of density components to the modulation spectra of speech signals (Singh and Theunissen, 2003). The spectrogram of a density component from which the modulation spectrum is estimated was generated by summing the spectrograms of each sparse component, weighted by the corresponding weights of the higher-level unit. The modulation spectrum for speech used for the comparison is obtained from (Singh and Theunissen, 2003). Singh et al. generated the spectrograms for human speech differently, but it is assumed that this does not affect the conclusions that can be drawn from a comparison. This allows to compare the impact of different initializations of the inference process.

In figure 2b, we overlaid the modulation spectrum for speech[1] with the set of centers of mass of the modulation spectra for the density components, both for

[1]Overcomplete with respect to the number of sample points.

[1]We constrained ourselves to positive modulation frequencies as a trade-off between resolution of the image and completeness of information.

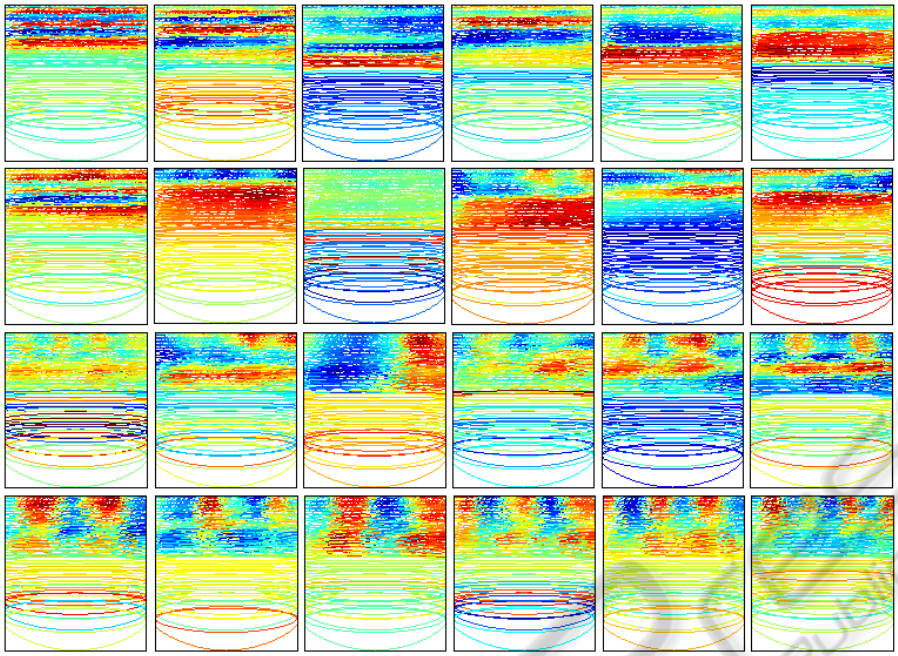

Figure 1: Subset of density components optimized for speech. Each squared figure corresponds to one of 24 density components. Within such a square, each ellipse represents a sparse component in the spectrotemporal domain. The temporal envelope width is divided in half for illustrational purposes. The ellipses are colored according to the weights in the particular density component. Red corresponds to significantly positive weights, and blue to significantly negative weights. Green shaded colors stand for values close to zero. The time window is 20 ms, and the frequency axis encompasses 4 kHz on a logarithmic scale. The density components are ordered according to their spectrotemporal modulation characteristics.

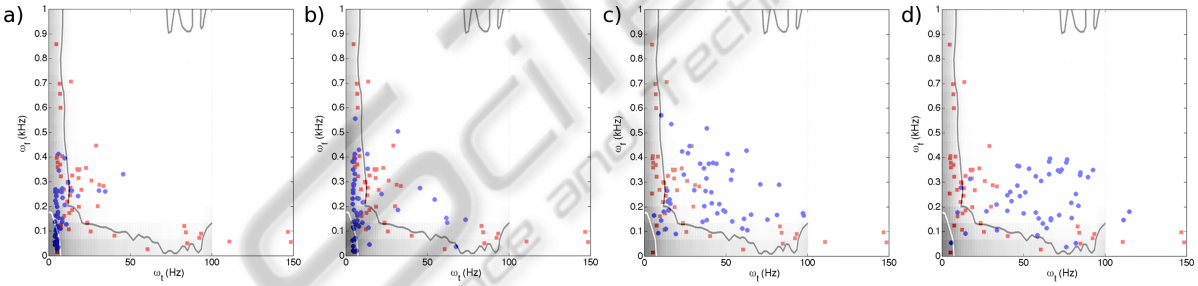

Figure 2: Population of modulation spectra of density components for different initializations. The centers of mass were overlaid with the spectrograms of speech calculated by Singh and Theunissen (2003). The red squares are the population of modulation spectra of density components for gist initialization, as a comparison. a) - c) show the results of optimizing the model to speech when initializing the encoding process of the higher-order coefficients with random gaussian noise with different variances σ. a) $\sigma = 0.1$. b) $\sigma = 0.5$. c) $\sigma = 1.0$. d) The coefficients were initialized based on the estimated local variance of the projections of the data onto the linear features. See text for further information.

the hierarchical models with the gist pathway (red) and without it (blue). We conclude that the density components emerging from the hierarchical model incorporating the gist pathway are better adapted to the spectrotemporal structure of speech. The blue dots correspond to density components learned by initializing the inference process either with random Gaussian noise or with the local variance structure of the linear filter outputs. First, the centers of mass are concentrated on regions of the speech modulation spectrum of high power. The frequency modulation frequencies are smaller than 0.9 cycles/kHz, and those of amplitude modulation are within the range of up to 100Hz and beyond. Secondly, in speech signals as well as the learned density components, high spectral and high temporal modulations are unlikely to occur at the same time, reflected by the star-shaped pattern of the modulation spectrum. The modulation characteristics of the other density components derived from models without gist initialization are less well adapted to the range of speech structures. They show significantly more redundancy than the ones with the gist

initialization or the ones based on initialization with the local variance structure.

How well the density components are able to generalize across natural signals can be assessed by looking at how well the speech signals are clustered based on the response patterns of the density components. In figure 3, we apply Locally Linear Embedding (LLE)[2] to the raw speech signals, the SC and the DC coefficients. This method discovers structures of high-dimensional data by assuming that it is sampled from a smooth manifold, and thus creates natural clusters of the data. As opposed to the raw data itself and sparse components, the output of the density components captures similarities of sound segments and separates distinct sound regions. The number of neighbors chosen for the LLE algorithm did not affect the quality of the results.

Furthermore, the clustering of the density component coefficients within the two-dimensional projection additionally reflects the temporal structure of the speech signal (not visualized in figure 3): Coefficients representing samples at the beginning of the sound segment are projected onto the left-hand side of the projection space while those coding the signal at later times are projected onto the right-hand side. The coefficients representing the region of transition are scattered in-between the two clusters, slowly transitioning from left to right. The temporal course of the higher-level coefficients reflect the temporal structure of their stimuli through smooth changes. This is also indicated by figure 4. We have applied a sliding window to the same sound extract as used in figure 3 and inferred the coefficient values $\mathbf{v}_{dc}$ at each step. The population of these inferred coefficients $\mathbf{v}_{dc}$ has been projected into the joint space of three specific density components, as shown in figure 4. Figure 4 reveals that the coefficient values are not scattered randomly across the three-dimensional space but are located on a clearly oval-like manifold. Furthermore, when watching an animation that illustrates how the coefficient values of these three components evolve in the joint space when sliding the window across the sound extract, one can observe that the coefficients change smoothly over time.

4.1.1 Pitch Sensitivity

We found three types of pitch-related density components. One set of components represents harmonic relations among the sparse components as frequency modulations, in accordance with the previous research (Klein et al., 2003), favoring the place theory. Within a set of 50 density components, these *harmonicity units* made up about 1/4 of the whole set. The second set of components encodes pitch by amplitude modulation across the mid- and high-frequencies, with no distinct activation pattern in the lower frequencies (*type-I AM units*), similar to the periodicity sensitive units from (Ming et al., 2009). The units of Ming et al. (2009) emerged from applying a sparse coding algorithm to the output of a pitch-based auditory image model. The third type of density components encodes the amplitude modulation across the mid- and high-frequency sparse components phase-locked to a low-frequency sparse component with a center frequency matched to the corresponding modulation frequency (*type-II AM units*), as an analogue to residue pitch. This is illustrated in figure 5.

Shown are three type-II AM units which capture the relationship between the fundamental frequency and the amplitude modulation in three different ways. In figure 5a, the waveform of the linear feature with center frequency 108.5 Hz alone synchronizes with the phase-locked activity of the higher frequency-units. Therefore, the density component assigns a significantly positive weight to this low-frequency unit, while all the neighboring units have negative weights. Whenever the amplitude modulation frequency and the phase of the modulation do not have a single counterpart within the low-frequency sparse components, the density component tries to encode the fundamental frequency by means of a combination of the low-frequency units, as seen in figure 5b and c. The higher-order unit of 5b has big positive weights on the two low-level units with the smallest frequencies and weights close to zero on their neighbors. The sum of the waveforms of the two units, weighted accordingly, matches the amplitude modulation. Similarly, the density component in figure 5c has a big negative weight on the low-level unit with a center frequency of 108.7 Hz, and a significantly negative but smaller weight on the unit with the next higher frequency, in order to elevate the average frequency closer to the modulation frequency. These negative weights introduce a phase shift of 180°. Generally, the resulting waveform of the relevant low-level units are slightly phase-shifted (see figure 5b and c). In this sense, the type-I and type-II AM units are sensitive also to a particular phase, as they are not phase-invariant. It is important to note that the fundamental frequencies found are within the range of fundamental frequencies of voices, i.e. between 90 and 250 Hz.

[2]"An unsupervised learning algorithm that computes a low dimensional, neighborhood preserving embedding of high dimensional data" (Saul and Roweis, 2000)

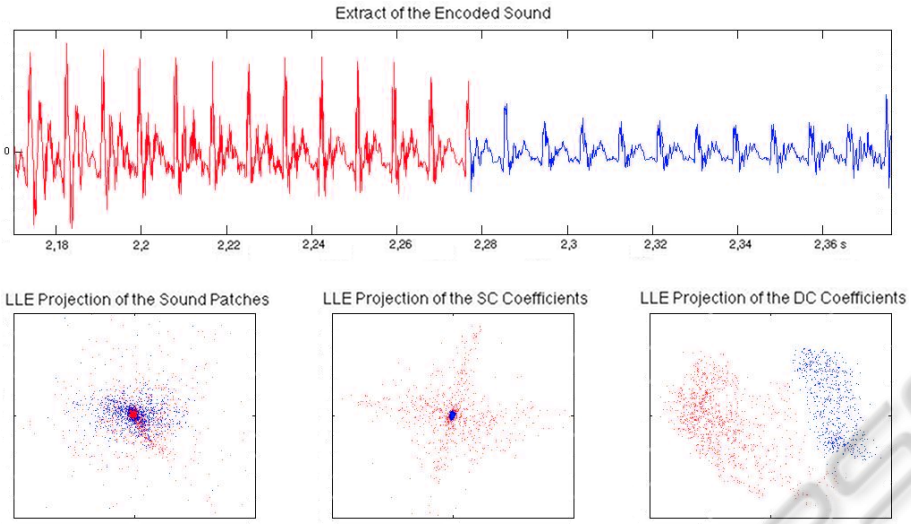

Figure 3: LLE Projections of the different representations. The colors have been assigned to the extracted sound segment based on the phonemic code. At each time point of the segment, the lower- and higher-level coefficients $\mathbf{u}$ and $\mathbf{v}_{dc}$ have been inferred. The resulting coefficient vectors then have been projected onto a two-dimensional space, using the standard LLE algorithm, with the coefficient vectors colored according to the sample window of the segment they are representing. The figure on the lower left shows the projection of the 160-dimensional sound patches themselves. The one in the middle shows the projection of the 320-dimensional linear feature coefficients and the figure in the lower corner on the right illustrates the projection of the 50-dimensional density component coefficients.

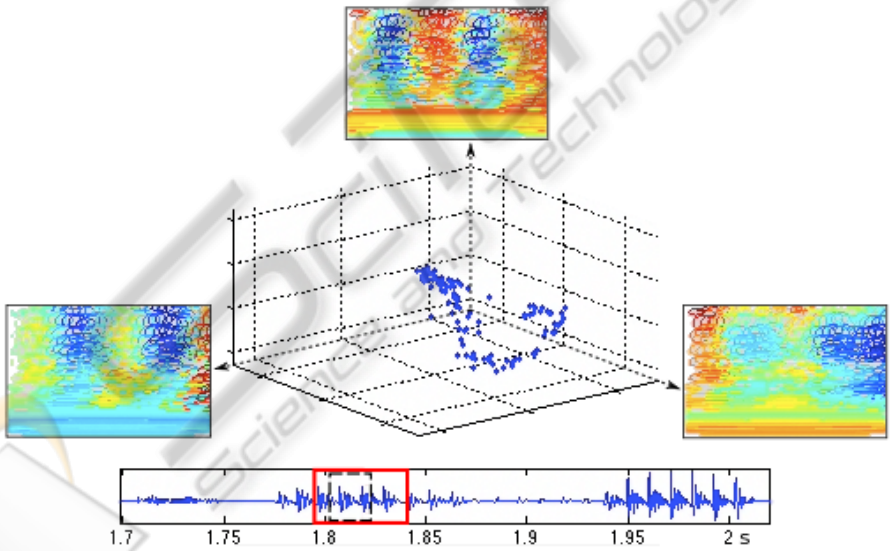

Figure 4: Temporal dynamics of the density components for a time-varying signal. The lower panel shows an extract of about 1.3 seconds of a speech sample. A sliding window (black dotted rectangle) was applied to the speech segment within the red rectangle. The sliding window was shifted by one sample at a time, and each time, the density coefficients $\mathbf{v}_{dc}$ were inferred. The upper panel shows the coefficient values of three specific density components (as illustrated by the three subplots at the axes), plotted in their joint space. Each blue dot corresponds to one set of inferred density component coefficients at a specific time step of the sliding window.

5 DISCUSSION

We have extended an existing probabilistic model - the density component model - for learning higher-order structures in natural signals and analyzed the statistical regularities it captured when applied to speech signals. The results from these models and the effects of the modifications allow us to draw con-

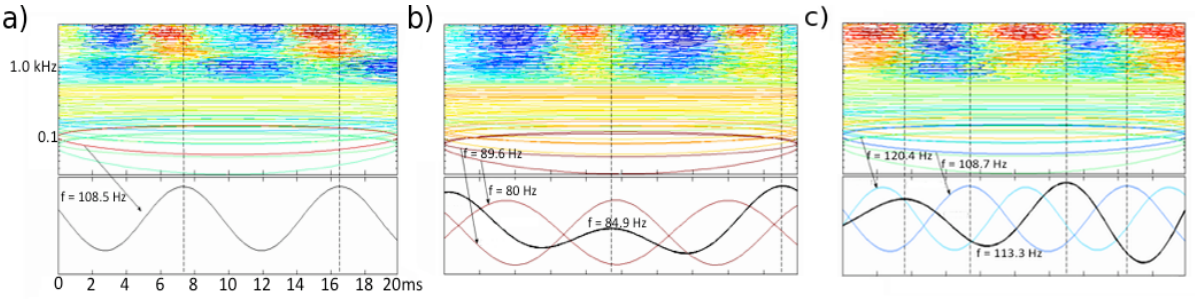

Figure 5: Type-II AM units. a) - c) show the spectrogram representations of examples of such components. The raw amplitude representation of the low-level functions which have significantly nonzero weights are plotted in the bottom panels, with their center frequencies displayed. The dotted lines illustrate the phase-locked weight pattern in the higher-frequencies, synchronized by the lower frequency. In b) and c), the two most relevant linear features are shown at the bottom, colored according to their weights. The bold black curve is the sum of the two functions, weighted by the corresponding density component values. b) The density component combines the two sparse components with the lowest center frequencies, 80 Hz and 89.6 Hz, to represent the amplitude modulation of about 85 Hz by their weighted sum, revealing a corresponding center frequency. c) A phase shift of 180° can be induced by assigning strongly negative weights to the low-frequency units. The sparse component with frequency 108.7 alone does not align well enough with the amplitude modulation (see dotted lines).

clusions with respect to the learning process of generative models in general, the unraveled structure of speech signals, as well as the processing of sounds.

The effect of deploying a systematized rather than random initialization for the gradient ascent step of the encoding process on the quality of the learned code is significant, illustrating the strong bias introduced by the chosen algorithm for performing Bayesian inference. This is relevant insofar as optimization of probabilistic models generally is based on stochastic gradient descent algorithms. The gist pathway has been implemented as a sparse coding algorithm on a Fourier-based spectrogram of the speech signals, which serves as a data-driven initialization for the inference process. As opposed to random initializations, the gist initialization leads to higher-order codes which capture broader and more complex structure of the speech signals and are better adapted to the spectral modulation characteristics of speech signals. This suggests that robustness of the encoding process is important for revealing structure in speech that corresponds to phase-locked activity of linear features across frequency (i.e. amplitude modulations in the signal). Furthermore, the gist pathway provides additional information to the model enabling to capture a wider variety of structures intrinsic to speech signals. As hypothesized, the gist pathway seems to move the system into a more appropriate region in the highly nonlinear space of the posterior probability distributions. This is seen when comparing the results with those obtained when initializing the coefficients according to the estimated local variance structure in figure 2: Despite the robustness of the inference process, the learned density components are significantly less well matched to the modulation spectrum of speech signals. In addition, we have found that the gist initialization increases the mean usage of the density components across ensembles of sounds and the sparseness of the coefficients which improves the speed of convergence and the efficiency of the code.

Among the density components learned in the fully extended model, we find three types of pitch-related density components, the harmonicity, the type-I AM and the type-II AM components. The latter two have not been reported in previous work (Karklin and Lewicki, 2005). We conclude that combining both the harmonicity and the AM components into one code allows a flexibility of pitch representation which might account for much of the diversity reported in pitch phenomena. This flexibility emerges because because the AM components map spectral cues around the fundamental and low-order harmonics onto periodicity cues at higher frequensies and visa-versa. These components become activated by a pure tone at its fundamental, periodic residue-like structure of a missing fundamental and any combination. We want to point out again that the model has not been hand-built, but that it is fully derived by the statistics of the speech sound population. As such, this statistically derived model reveals that the higher-order statistics of speech sounds alone show relationships between different types of pitch-related cues. Importanly, the statistical structure of speech sounds revealed by this model suggests that pitch computation is more complex and integrated than a simple harmonic or periodicity template alone. As such, this work extends on the debate about the relevance of both time and place theory by suggesting to soften this pure dichotomy. However, to make a strong argument about pitch, further work needs to show invari-

ance of the model to a variety of pitch phenomena to the model.

However, the given implementation of the model hampers its ability to robustly capture this distinct relationship between the low-frequency and the high-frequency units with higher precision because of two reasons. First, the restricted number of sparse components inherently introduces a trade-off in their frequency resolution as well as in their capacity to encode the phase of the signal. Second, the model training implements a block based approach to signal encoding, with the sound segments being randomly drawn from the set of available sentences, irrespective of the phase structure of the signals. Therefore, the representation of the higher-order structure related to temporal pitch is highly sensitive to the learning process. We plan on addressing these issues in future work.

REFERENCES

Bell, T. and Sejnowski, T. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7:1129–1159.

Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallett, D., Dahlgren, N., and Zue, V. (1990). *TIMIT Acoustic-Phonetic Continuous Speech Corpus*.

Griffiths, T., Buechel, C., Frackowiak, R., and Patterson, R. (1998). Analysis of temporal structure in sound by the human brain. *Nature Neuroscience*, 6:633–637.

Harding, S., Cooke, M., and Konig, P. (2007). Auditory gist perception: An alternative to attentional selection of auditory streams? In *Lecture Notes in Computer Science*. Springer.

Karklin, Y. and Lewicki, M. (2003). Learning higher-order structures in natural images. *Network: Computation in Neural Systems*, 14:483–499.

Karklin, Y. and Lewicki, M. (2005). A hierarchical bayesian model for learning nonlinear statistical regularities in nonstationary natural signals. *Neural Computation*, 17:397–423.

Klein, D., Konig, P., and Kording, K. (2003). Sparse spectrotemporal coding of sound. *EURASIP J. on Advances in Signal Processing*.

Ming, V., Rehn, M., and Sommer, F. (2009). Sparse coding of the auditory image model. *UC Berkeley Tech Report*.

Oliva, A. and Torralba, A. (2006). Building the gist of a scene: The role of global image features in recognition. *Progress in Brain Research*.

Olshausen, B. and Field, D. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609.

Olshausen, B. and Field, D. (1997). Sparse coding with an overcomplete basis: A strategy employed by v1? *Vision Research*, 37:3311–3325.

Oxenham, A., Bernstein, J., and Penagos, H. (2004). Correct tonotopic representation is necessary for complex pitch perception. In *Proc Natl Acad Sci USA*, volume 101, pages 1114–1115.

Patterson, R., Uppenkamp, S., Johnsrude, I., and Griffiths, T. (2002). The processing of temporal pitch and melody information in auditory cortex. *Neuron*, 36:767–776.

Saul, L. and Roweis, S. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, 22:2323.

Shamma, S. (2001). On the role of space and time in auditory processing. *Trends in Cognitive Science*, 5:340–348.

Shamma, S. (2004). Topographic organization is essential for pitch perception. *PNAS*, 5:1114–1115.

Shannon, R., Zeng, F., Kamath, V., Wygonski, J., and Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science*, 270.

Singh, N. and Theunissen, F. (2003). Modulation spectra of natural sounds and ethological theories of auditory processing. *J Acoust Soc Am*, 114:3394–3411.

Turner, R. and Sahani, M. (2007). Probabilistic amplitude demodulation. In *Lecture Notes in Computer Science*, volume 4666, pages 544–551.

APPENDIX

Inference in the Fully Extended Density Component Model

As described previously, the alterations of the original density component model affect the encoding and learning procedures, while the generative model itself remains the same. As the transformation from the sound to the higher-order representation vdc is fundamentally nonlinear, the optimal coefficient values for the representation cannot be expressed in closed form[2]. In order to encode a given signal, the MAP estimation of the sparse (SC) and the density component coefficients (DC) is illustrated in figure 6:

1. Choose a whitened sound extract $\mathbf{x}_w$ of length T.
2. Generate the corresponding spectrogram $S_\mathbf{x}$ of temporal length $T_S > T$ and frequency resolution F_{res}, using a logarithmic scaling of the frequencies.
3. Calculate the gist information:
   (a) Project $S_\mathbf{x}$ into the space of the first 50 principal components explaining approximately 95% of the variance:

$$\mathbf{u}_G = W^S_{pca} S_\mathbf{x}, \quad W^S_{pca} = D_S^{-1/2} E_S^T, \qquad (12)$$

[2]Closed form means that the expression can be written analytically in terms of a bounded number of certain well-known functions (i.e. no infinite series, etc.)

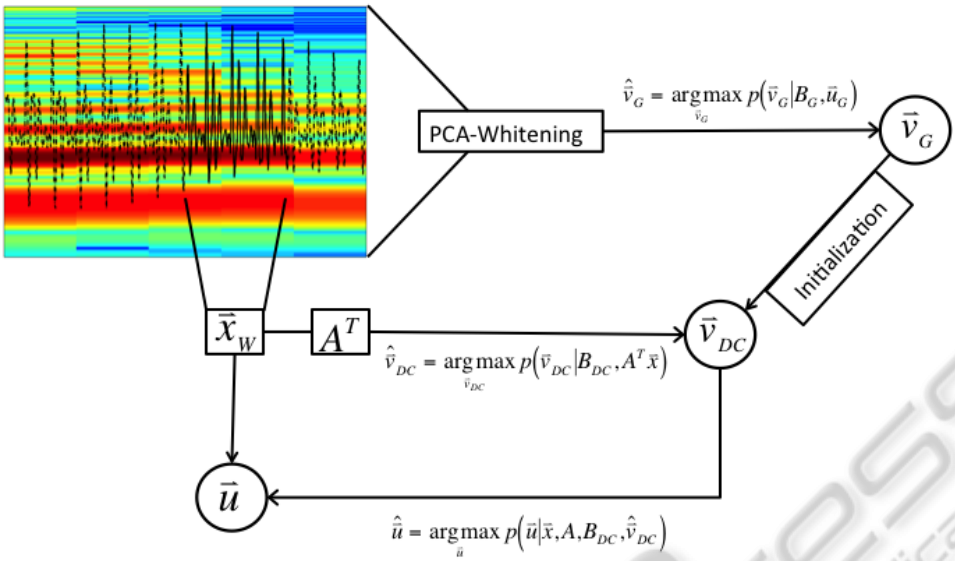

Figure 6: Encoding within the extended Density Component Model. The main stages comprise a projection of the data $\mathbf{x}_w$ onto the columns of the matrix of basis functions A, a parallel pathway which infers information about the global context of the specific data sample in order to initialize the inference of the actual DC coefficients, and finally the inference of the SC coefficients, given the DC coefficients. (*For further description see text.*)

where D_S and E_S are the eigenvalues and eigenvectors of the total set of spectrograms respectively.

(b) Calculate the MAP estimate of the gist coefficients $\mathbf{v}_G$ by performing conjugate gradient ascent on the corresponding log-posterior distribution in equation (11):

$$\begin{aligned}\hat{\mathbf{v}}_G &= \underset{\mathbf{v}_G}{\operatorname{argmax}} \log p(\mathbf{v}_G|\mathbf{u}_G, B_G) \\ &= \underset{\mathbf{v}_G}{\operatorname{argmax}} \log \left(p(\mathbf{u}_G|B_G, \mathbf{v}_G) p(\mathbf{v}_G)\right) \\ &= \underset{\mathbf{v}_G}{\operatorname{argmax}} \left(-\frac{1}{2\sigma_G^2} \|\mathbf{u}_G - B_G \mathbf{v}_G\|_2^2 - \right. \\ &\qquad \left. - \sum_{i=1}^{M} \left| \frac{v_{G,i}}{\mu_G} \right| \right) \end{aligned} \tag{13}$$

4. Use $\hat{\mathbf{v}}_G$ as the initialization to the gradient ascent algorithm which maximizes the log-posterior distribution of the DC coefficients $\mathbf{v}_{dc}$ in equation (6), given the projection of the whitened sound $\mathbf{x}_w$ onto the set of sparse components A :

$$\begin{aligned}\hat{\mathbf{v}}_{dc} &= \underset{\mathbf{v}_{dc}}{\operatorname{argmax}} \log p(\mathbf{v}_{dc}|\tilde{\mathbf{u}}_{dc}, B_{dc}) \\ &= \underset{\mathbf{v}_{dc}}{\operatorname{argmax}} \log \left(p(\tilde{\mathbf{u}}_{dc}|B_{dc}, \mathbf{v}_{dc}) p(\mathbf{v}_{dc})\right) \\ &= \underset{\mathbf{v}_{dc}}{\operatorname{argmax}} \left(-B_{dc}\mathbf{v}_{dc} - \left| \tilde{\mathbf{u}}_{dc} \diamond e^{B_{dc}\mathbf{v}_{dc}} \right| - \right. \\ &\qquad \left. - \sum_{i=1}^{M} \left| \frac{v_{dc,i}}{\mu_{dc}} \right| \right), \end{aligned} \tag{14}$$

where $\tilde{\mathbf{u}}_{dc} = A^T \mathbf{x}_w$ is the projection and $\mathbf{a} \diamond \mathbf{b} := \left[\frac{a_1}{b_1}, \frac{a_2}{b_2}, ..., \frac{a_n}{b_n} \right] \ \forall \mathbf{a}, \mathbf{b} \neq \in \mathbb{R}^n$.

5. Sparsify the SC coefficients $\mathbf{u}$, given $\hat{\mathbf{v}}_{dc}$ by means of conjugate gradient ascent on the log-posterior of the SC coefficients, given the data:

$$\begin{aligned}\hat{\mathbf{u}} &= \underset{\mathbf{u}}{\operatorname{argmax}} \log p(\mathbf{u}|\mathbf{x}_w, A, B_{dc}, \hat{\mathbf{v}}_{dc}) \\ &= \underset{\mathbf{u}}{\operatorname{argmax}} \log \left(p(\mathbf{x}_w|\mathbf{u}, A) p(\mathbf{u}|B_{dc}, \hat{\mathbf{v}}_{dc})\right) \\ &= \underset{\mathbf{u}}{\operatorname{argmax}} \left(-\frac{1}{2\sigma_\varepsilon} \|\mathbf{x}_w - A\mathbf{u}\|_2^2 - B_{dc}\hat{\mathbf{v}}_{dc} - \right. \\ &\qquad \left. - \left| \mathbf{u} \diamond e^{B_{dc}\hat{\mathbf{v}}_{dc}} \right| \right), \end{aligned} \tag{15}$$

For the derivation of the gradients see the following section.

Derivation of the Log-likelihood and the Gradients

The MAP estimates of the $\hat{\mathbf{u}}$ and $\hat{\mathbf{v}}_{dc}$ were obtained by maximizing the joint log posterior distributions for a given sound segment $\mathbf{x}$

$$\begin{aligned} L &= \log p(\mathbf{u}, \mathbf{v}_{dc}|A, \mathbf{x}, B_{dc}, \mathbf{v}_{dc}) \\ &\propto \log \left(p(\mathbf{x}|A, \mathbf{u}) p(\mathbf{u}|B_{dc}, \mathbf{v}_{dc}) p(\mathbf{v}_{dc})\right) \end{aligned} \tag{16}$$

with

$$p(\mathbf{x}|A,\mathbf{u}) = \frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}} \exp\left(-\frac{1}{2\sigma_\varepsilon^2}||\mathbf{x}-A\mathbf{u}||^2\right) \quad (17)$$

$$p(\mathbf{u}|B_{dc},\mathbf{v}_{dc}) \propto \prod_{i=1}^{N} z_i \exp\left(-\left|\frac{u_i}{\lambda_i}\right|\right) \quad (18)$$

$$p(\mathbf{v}_{dc}) \propto \prod_{j=1}^{K} \exp\left(-\left|\frac{\mathbf{v}_{dc,j}}{\mu_{dc}}\right|\right) \quad (19)$$

with the normalization factor $z_i = 1/(2\lambda_i)$ and the scale parameter $\lambda_i = \exp([B_{dc}\mathbf{v}_{dc}]_i)$:

$$L \propto -\frac{1}{2\sigma_\varepsilon^2}||\mathbf{x}-A\mathbf{u}||^2 + \sum_{i=1}^{N}\left[\log\lambda_i - \left|\frac{u_i}{\lambda_i}\right|\right] - \\ - \sum_{j=1}^{K}\left|\frac{v_{dc,j}}{\mu_{dc}}\right|. \quad (20)$$

The MAP estimates were calculated by means of gradient descent. Writing the element wise division of vectors as

$$\mathbf{a}\diamond\mathbf{b} := \left[\frac{a_1}{b_1}, \frac{a_2}{b_2}, ..., \frac{a_n}{b_n}\right] \ \forall \mathbf{a},\mathbf{b} \neq 0 \in \mathbb{R}^n \quad (21)$$

the gradients with respect to $\mathbf{u}$ and $\mathbf{v}_{dc}$ are[1]

$$\frac{\partial L}{\partial \mathbf{u}} = \frac{1}{\sigma_\varepsilon} A^T(\mathbf{x}-A\mathbf{u}) - \mathrm{sign}(\mathbf{u})\diamond\exp(B_{dc}\mathbf{v}) \quad (22)$$

$$\frac{\partial L}{\partial \mathbf{v}} = B_{dc}^T(|\mathbf{u}\diamond\exp(B_{dc}\mathbf{v})| - 1) - \frac{1}{\mu_{dc}}\mathrm{sign}(\mathbf{v}). \quad (23)$$

The sparse components A and the density components B_{dc} were estimated by maximizing the posterior over the sound batch containing D segments $\mathbf{x}_n$, approximated by means of the MAP estimates $\hat{\mathbf{u}}_n$ and $\hat{\mathbf{v}}_n$

$$\{\hat{A},\hat{B}_{dc}\} = \underset{A,B_{dc}}{\mathrm{argmax}} \sum_{n=1}^{D} \log[p(\mathbf{x}_n|A,B_{dc},\hat{\mathbf{u}}_n,\hat{\mathbf{v}}_n)\cdot \\ \cdot p(\hat{\mathbf{u}}_n|B_{dc},\hat{\mathbf{v}}_n)p(\hat{\mathbf{v}}_n)p(A,B_{dc})]. \quad (24)$$

Setting $p(A,B_{dc}) = p(B_{dc}) = \mathcal{N}(0,\sigma_B)$, we implement stochastic gradient ascent

$$\Delta A = \frac{1}{D}\sum_{n=1}^{D}\frac{\partial L_n}{\partial A},\ \Delta B_{dc} = \frac{1}{D}\sum_{n=1}^{D}\frac{\partial L_n}{\partial B_{dc}},$$

where L_n refers to the terms of the sum in equation (24). Using equation 20, the gradients are:

$$\frac{\partial L_n}{\partial B_{dc}} = (|\hat{\mathbf{u}}_n\diamond\exp(B_{dc}\mathbf{v}_n)| - 1)\,\hat{\mathbf{v}}_n^T - \frac{1}{2}B \quad (25)$$

$$\frac{\partial L_n}{\partial A} = \frac{1}{\sigma_\varepsilon^2}(\mathbf{x}_n - A\mathbf{u}_n)\,\mathbf{u}_n^T. \quad (26)$$

[1]We omit the index dc in the coefficients $\mathbf{v}_{dc}$ for the sake of simplicity.