Fine-grained statistical structure of speech

François Deloche

Fine-grained statistical structure of speech

François Deloche*

EHESS, PSL University
francois.deloche@ehess.fr

March 10, 2019

Abstract

In spite of its acoustic diversity, the speech signal presents statistical regularities that can be exploited by biological or artificial systems for efficient coding. Independent Component Analysis (ICA) reveals that on small time scales of about 10 ms, the overall structure of speech is well captured by a time-frequency representation whose frequency selectivity follows a power law in the high frequency range 1-8kHz, reproducing cochlear filtering. Further adaptation to phonetic categories can be reached by variations of the exponent, i.e. different time-frequency trade-offs. Here a parametric approach is proposed instead of ICA, based on a measure that reflects the sparsity of decompositions in a set of Gabor dictionaries whose atoms are Gaussian-modulated sinusoids. The variations of the exponent associated with the best decomposition are examined, first at the level of phonemes, then at an intra-phonemic level. The acoustic properties that affect the exponent can be inferred from this detailed analysis. A key result is that release bursts lower the exponent of stops concealing higher values during the opening phase. The analysis further suggests that the statistical structure of speech may be congruent with nonlinear peripheral auditory coding in mammals.

Keywords: *Independent Component Analysis, efficient coding hypothesis, sparse coding, Gabor dictionaries, acoustic phonetics, auditory coding*

Introduction

Shannon's information theory[1] provides an abstract framework for evaluating a speech coding system. One key information theoretic criterion for multichannel coding is redundancy reduction, corresponding to the idea that two channels should not waste energy coding for the same information. Redundancy reduction has been proposed as a plausible principle underlying sensory messages in the brain.[2, 3, 4] The efficient coding hypothesis states that sensory systems have evolved to keep redundancy between neural channels as low as possible when natural stimuli are presented.[5] It initiated numerous studies on the properties of the visual system.[6] It is also the basis for comparable studies on the auditory system, whether on peripheral processing,[7, 8, 9] or on higher level processing (e.g. modulation filters), more recently.[10, 11, 12, 13] Efficient coding takes advantage of statistical regularities in the input data. As a matter of fact, an efficient code can be found using a method of data analysis known as Independent Component Analysis (ICA). ICA seeks a linear transformation that makes the components of high dimensional data statistically independent. When applied to a specific class of data, the representation produced by ICA reveals its overall structure (see Hyvarinen, 1999[14] for a review) and at the same time gives an insight into the optimal sensory coding scheme.[5, 15, 6] ICA applied to raw speech data results in a time-frequency representation whose frequency selectivity follows the same power law as selectivity in the mammalian cochlea with respect to center frequency, in the range of

*Centre d'analyse et de mathématique sociales, CNRS, EHESS, PSL University, 54 blvd Raspail, Paris, France

high frequencies 1-8kHz.[7] While being coherent with the hypothesis that speech is adapted to peripheral auditory processing, the products of ICA are hard to interpret in terms of signal structure. Indeed, the diversity of phones associated with a language does not allow for an interpretation of the optimal decomposition that would apply to any speech sound. In order to get an explanation based on concrete properties of the signal, one approach is to split the data into subtypes that share common acoustic features. Stilp and Lewicki applied ICA on phonetic categories (e.g. fricatives, stops, affricates, vowels) instead of the whole speech data.[16] They found that the time-frequency resolution trade-off was different depending on the class used at the input of the algorithm and therefore assumed that the time-frequency resolution is mostly explained by the relative transiency of a sound class. Rapid changes on the time axis would make the optimal filters shift towards a time representation with poorer frequency selectivity. This view, however, does not fully explain why vowels result in a representation that is more localized in time than fricatives for example. Phonemes and categories of phonemes have been extensively described by their acoustic properties, but there have been only a few attempts to explain the consequences of these properties on the statistical structure of the signal. The aim of this study is to provide a detailed view of the statistical structure of speech and to link the signal structure to its acoustic properties. A new connection between a theoretically efficient code based on a flexible representation of the data and level-dependent nonlinear cochlear processing is also proposed.

For the purposes of this analysis, a method derived from ICA is introduced, producing the same type of optimal representations while being more flexible with the input data. ICA is a non-parametric approach that requires no prior information on the optimal filters but in return requires a sufficiently large amount of data at its input. This constrains the analysis as hypotheses about the relevant classes have to be made before ICA is performed. In particular, one cannot see the evolution of the optimal code through time, whereas the acoustic properties can change even inside a phonetic unit. Another limitation is that ICA does not provide a direct measure of the intraclass variability. Independent component analysis being basically an estimate of a high dimensional vector, performing the analysis with little data is impossible in most cases. However, we do know some properties of the optimal filters in our context. When ICA is applied to sufficiently broad subclasses of speech, it always results in a bank of Gabor wavelet-like filters whose frequency selectivity depends on the center frequency. This dependence is well fitted by a power law model. In the end, statistical analyses boil down to the study of one parameter being the regression slope of the quality factor Q_{10} on center frequency f_c on a *log-log* scale.[16, 17] This motivates the use of a parametric approach instead of the original ICA framework. A set of overcomplete dictionaries whose atoms are Gabor filters - i.e. Gaussian-modulated sinusoids - and a cost function h reflecting the sparsity of decompositions are employed. The dictionary that optimally encodes the data minimizing the cost function gives an estimate of the Q_{10}/f_c regression slope. This method needs little data and makes it possible to see the changes in the optimal representation at a finer level of speech.

The results of two analyses are discussed in this article. In the first analysis, optimal representations are sought at the phonetic level and intraclass variability is analyzed for phonetic categories. It is the continuation of Stilp and Lewicki's work. In the second analysis, we take a closer look at some phonemes, especially stops and affricates. A plosive/affricate has acoustic features that evolve in time, from the release of the closure (burst) to the end of the aspiration or opening phase. The changes in the most efficient representation over time are inspected. Along with these analyses, the best representation is also sought for two kinds of artificial signals related to speech sounds - modulated noises and radiated sounds at the output of a uniform cylindrical waveguide. These simulations will support the inference of the acoustic factors that determine the exponent of Q_{10} on f_c.

Materials and methods

Gabor dictionaries

The candidates for the optimal representations are a set of 30 overcomplete dictionaries whose atoms are Gabor filters. Gabor filters are the functions that achieve the Heisenberg limit for time-frequency resolution.[18] A Gabor filter $w(t)$ is a sinusoid modulated by a Gaussian envelope (if imaginary part is ignored). It is completely described by four parameters being the time shift τ, the Gaussian width (time deviation) σ_t, the center frequency f_c, and the phase of the sinusoid ϕ at $t = 0$:

$$w(t) = C \sin(\omega t + \phi) \exp(-\frac{t-\tau}{4\sigma_t^2}) \tag{1}$$

where C is a normalization factor set so that all filters have constant squared norm. Each dictionary is composed of 600 Gabor filters uniformly distributed in time, frequency and phase. The only parameter that has to be set is the Gaussian width which also determines the frequency selectivity. Frequency selectivity is evaluated by the quality factor Q_{10} defined by the center frequency f_c divided by the 10dB-bandwidth Δf. For the optimal filters that are found with ICA, the Q-factor plotted against center frequency is well fitted by a line on a *log-log* scale. The intercept is considered redundant with the slope of the regression as most of the lines cross around the point $(f_0 = 1kHz, Q_0 = 2)$ for different speech data at the input.[16, 17] Therefore, the regression slope of Q_{10} on f_c, that we will denote β for simplicity, summarizes the representations obtained with ICA. Each one of the 30 dictionaries of Gabor filters corresponds to a value of β, going from 0.3 to 1.2 with a constant step (see Fig 1). A decomposition with $\beta = 1$ corresponds to the windowed Fourier transform (also called Gabor transform) whereas a zero value would correspond to the multi-resolution wavelet transform or constant-Q transform. This family of dictionaries is also referred as flexible Gabor-wavelets or $\alpha-atoms$ in the field of time-frequency analysis, with the correspondence $\beta = 1 - \alpha$.[19] One can see β as a way to control the time-frequency trade-off in the high frequency range: frequency accuracy at the expense of time accuracy, or inversely. Thus, the lower β, the more the representation shifts toward a time decomposition. The range [0.3, 1.2] is chosen so as to encompass all the values taken by β in the previous analyses. To ensure more diversity of the filters, some randomness is added to the Q-factor with multiplicative noise. It follows that

$$\log Q_{10}(f) = \log Q_0 + \beta(\log f - \log f_0) + 0.04\eta \tag{2}$$

where the log is taken in base 10 and η is i.i.d. noise drawn from the normal distribution. As for the other parameters which are uniformly distributed, the ranges are respectively [1 - 6.5kHz], [2 - 14ms] and [0, 2π] for center frequency, time shift and phase. The time shift does not cover the full range of the time window ($T = 16ms$) in order to avoid potential boundary effects.

Data

The speech data was retrieved from the TIMIT database.[20] It provides audio examples of sentences in American English as well as information on their phonetic content by segment. Slices of 16 ms of speech were considered, representing 256 samples at $f_s = 16kHz$. The examples were preprocessed with filtering and then normalization. The filtering was done with a high pass Butterworth filter of order 8 and a cut-off frequency at $1.5kHz$. The use of a high cut-off frequency is not common for speech analysis as much of the phonetic information is in the low-frequency part, but the focus of this study is only on the high frequency region, where the power law model has been found to be valid in previous work. Frequencies below 1kHz play no role in

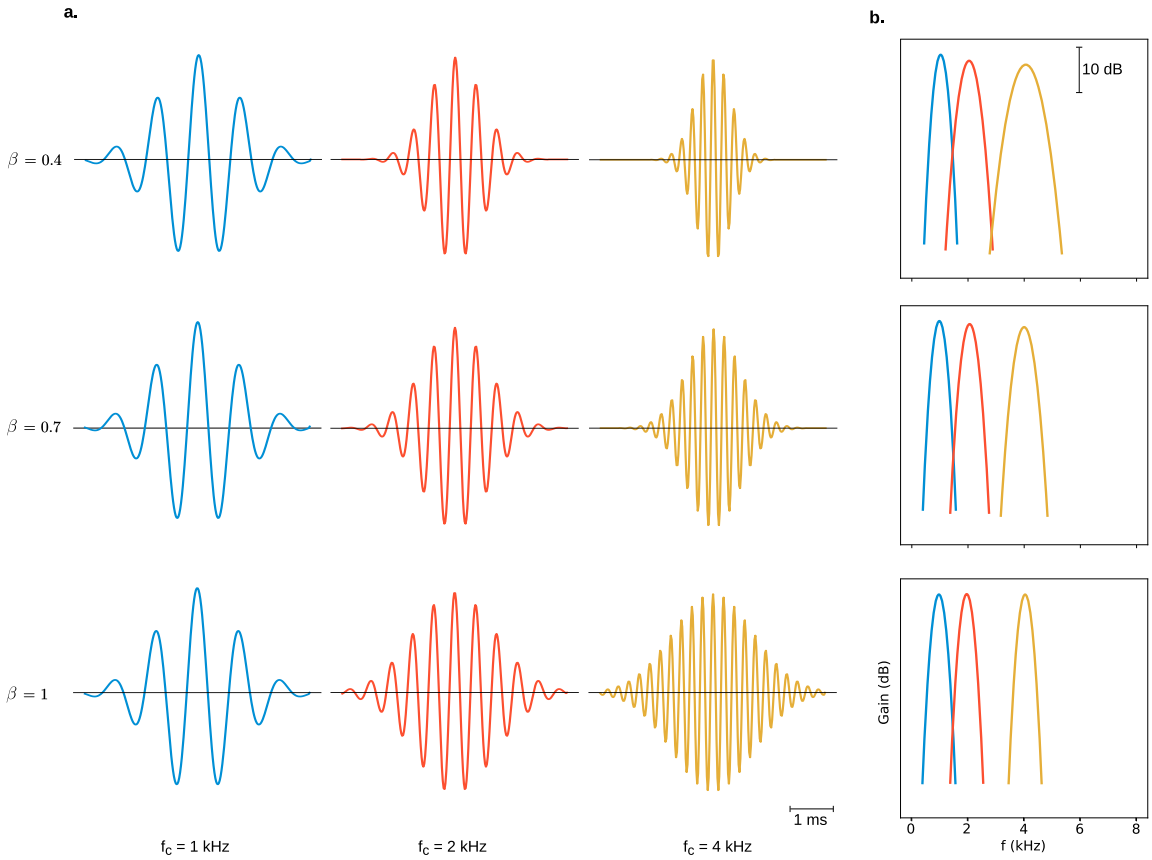

Figure 1: Examples of filters that compose the Gabor dictionaries indexed by β. *These dictionaries are the candidates for the best decomposition of the speech samples. The filters are Gabor wavelets (sinusoids modulated by Gaussians). They are all the same at $f_c = 1kHz$ but then the quality factor increases as f^{β}. β controls the time-frequency trade-off in the high frequencies: a low value of β causes the filters to be time-localized, while a high value causes the filter to be frequency selective.* ***a.*** *Time waveforms for three values of β and three values of f_c .* ***b.*** *Frequency responses (gain in dB).*

the decompositions of the speech samples, moreover the dictionaries are all the same at 1kHz, meaning that the region of discrimination is in even higher frequencies. The normalization was done by dividing each slice by its root mean square (RMS value). The TIMIT database indicates the time of releases for stops and affricates. This information was used several times in the analyses, in particular the closure part, which contain no high frequency information, was always ignored.

Cost function and relation to Independent Component Analysis

Given some speech data, a measure of the decomposition goodness is needed to select the optimal representation among the Gabor dictionaries. This paragraph explains the choice of the cost function from a theoretical point of view and the next paragraph describes how it is used in practice for the choice of the best dictionary.

$n-$dimensional vectors X are generated from slicing of speech data. Independent component analysis tries to find a set of filters $W = (W_1, ..., W_m)$ such that if we denote $Y = W^T X$, the components of the output vector Y are statistically independent. If so, $Y = (Y_1, ..., Y_m)$ would be a representation of the input X that minimizes redundancy between its components Y_i, providing the basis for an efficient code. In the ideal case of strict independence, the entropy $H(Y)$ is the

sum of the entropy of each channel:

$$\sum_i H(Y_i) = H(Y) \ . \tag{3}$$

By contrast, the worst case for independence would be to have all the output channels sharing the same information equal the total entropy. ICA looks for W that minimizes the redundancy

$$\sum_i H(Y_i) - H(Y). \tag{4}$$

One needs a probabilistic model to estimate the entropy terms. For speech, the distributions of the output channels for various time-frequency decompositions are well fitted by Laplacian distributions $\log p(y) \propto -|y|$.[21] This prior encourages sparsity of the decomposition as most of the values of Y_i are around zero under this model. It has been used multiple times for ICA on speech.[8, 16] With this prior, minimizing the sum of channels entropy comes back to minimizing the L_1 norm of the vector Y (up to some weights, see next subsection):

$$h(Y) = ||Y||_1 = \sum_i |Y_i|. \tag{5}$$

This view is very similar to sparse coding,[22, 23] with the difference that there is no attempt here to *reconstruct* the input signal from the output. The second term $-H(Y)$ is equal to the entropy of the input $-H(X)$ in case of an orthonormal decomposition and has not to be estimated. In practice this term is replaced by a penalty term that ensures that the column vectors of W represent all the directions of the n-dimensional space to avoid the collapse of filters. For a square matrix ($m = n$) the penalty term can be $-|\det W|$,[7] but for overcomplete bases ($m > n$), it has no natural expression.[24] For the method presented here, all bases are sets of Gabor filters distributed uniformly in time, frequency and phase. These bases are considered to be equivalent and the penalty term is ignored. Hence, the cost function is limited to h only (Eq 5). As such, this method yields optimal decompositions which are consistent with previous work.

Choice of the optimal representation

Let Y_β be the coefficients vector obtained at the output of one Gabor filter bank indexed by β for some input X. In accordance with the previous paragraph, the decomposition is evaluated with

$$h_\beta(X) = ||Y_\beta||_1 = \sum_i |Y_{\beta,i}|. \tag{6}$$

Some weights were included to avoid the medium frequencies taking advantage over the high frequencies because of the natural decrease in energy along the frequency axis. h_β becomes

$$h_\beta(X) = \sum_i \gamma(f_i)|Y_{\beta,i}| \tag{7}$$

where f_i is the center frequency of the filter i and $\gamma(f)$ is an increasing function of the frequency. If γ is set proportional to the inverse of the average spectral power density (+5dB/octave), the weighting is almost equivalent to whitening of the data, a usual preliminary step for ICA. The results shown in this article were obtained with a slighter gain of +2dB/octave as the values of $\beta^\star$ are unchanged simultaneous with more consistent values of h. Finally, the sum was arbitrarily normalized with h_β set to 1 for the less sparse signals.

h_β was computed for each Gabor dictionary over a set of data. The dictionary that minimizes the cost function was chosen as the best representation of the data. The best dictionary is associated with a slope $\beta^\star$.

The cost defined as above measures the lack of structure for some data. Average values of h over the full set of Gabor dictionaries were considered simultaneously with $\beta^\star$ in our analyses. $\beta^\star$ describes the optimal representation whereas h quantifies *how easily* the signal is decomposed. Low values of h characterizes sounds that present structure, typically vowels. On the contrary, noise-like obstruent sounds (fricatives, stops) are expected to yield high values of h.

Data analysis

Analysis 1: The goal of the first analysis is to estimate $\beta^\star$ for different classes of speech sounds. For each class of speech sounds, 400 occurrences (for single phonemes) or 800 occurrences (for broad phoneme classes: fricatives, stops, vowels...) were retrieved from the TIMIT database, randomly sampled from throughout the database. There are fewer examples for phonemes as data can be limited for some phonemes and we want confidence intervals to be comparable between all phonemes. A 16 ms slice was selected at random from each occurrence. The cost functions h_β were computed and summed over the slices for each value of β. They were smoothed with a Gaussian filter ($\sigma = 0.03$) along the β axis. The minimal score was obtained at $\beta = \beta^\star$. From the same 400 or 800 slices, the estimation of $\beta^\star$ was repeated 3 000 times with re-sampled versions of the slices with repetitions. This procedure gives a bootstrap distribution of $(\beta^\star, h)$ whose average and 70% bootstrap confidence interval are reported on the (β, h) plane. Alternatively, the bootstrap distribution can be represented by box plots as done in Fig 2. Some histograms of the bootstrap distributions are shown in Supplementary Material. Results of Analysis 1 on broad categories and on phonemes are shown in Fig 2 and Fig 3 respectively. Phonemes are divided into vowels, stops, fricatives, affricates, laterals/glides (semivowels) and nasals, as in Stilp and Lewicki 2013 (see Table 1 in Ref 16). Confidence intervals on h are not represented because variations are small (standard deviation of order 0.01).

Analysis 2: The second analysis investigates variations in $\beta^\star$ on a finer time scale. The motivation behind Analysis 2 is that some phonemes like affricates or stops are subject to acoustic changes even within an occurrence. Thus, time patterns on $\beta^\star$ can be expected inside some phonemic units. Similarly to Analysis 1, 400 occurrences of a phoneme were retrieved from the TIMIT database. This time, eight 16ms slices at regular intervals were considered for each example, possibly with some overlap, instead of a single slice by occurrence. As the occurrences do not have the same duration, the eight steps represent relative time rather than absolute time ($1 = START$, $8 = END$). The estimation of $\beta^\star$ was the same as described for Analysis 1. At the end of the procedure, one has a $\beta^\star$ value at each step (1-8) and can see its evolution through time. Figure 6 shows some of these time courses for several phonemes. Data for other phonemes and other information like duration histograms are included in Supplementary Material.

Artificial signals

In addition to actual speech data, two kind of signals were generated to supplement the analyses. The first signals are based on some noise and relate to consonants. The second generated signals are vowel-like sounds.

Simulation 1: First kinds of simulated signals are noises modulated by Gaussians in time or frequency. 200 samples were generated, each lasting 16ms. At first, they were all Gaussian noise filtered by a low pass filter of order 3 at 4kHz. It is with this setting that a symmetrical pattern was obtained in Fig 4. Each version is associated with a parameter u going from 0 (first sample) to 1 (last sample) which controls the time/frequency modulations. At $u = 0$, noise is windowed - multiplied - by a Gaussian of time deviation $\sigma_t = 0.01 \times T = 0.16ms$. At $u = 1$, noise is convolved by a Gaussian filter of frequency deviation $\sigma_f = 0.01 \times f_s/2 = 80Hz$. The sounds of intermediate values are shifts of these two configurations. From 0 to 0.5, they undergo time

modulation essentially with σ_t increasing. From 0.5 to 1, they go through frequency modulation with σ_f decreasing. $u = 0.5$ is not very different from non-modulated white noise. Some values for the modulation widths are given on Table 1. See Supplementary material for more details on the generation and a sound file that demonstrates the transition between all the samples.

Table 1: Correspondence between the control parameter u and modulation widths (Simulation 1)/aperture radius (Simulation 2).

u	σ_t (ms)	σ_f (Hz)	r (cm)
0	0.2	$\star$	0.20 ([ʊ])
0.2	1.0	$\star$	0.42 ([u], [ɔ])
0.4	4.5	$\star$	0.64 ([ʌ], [ɪ])
0.6	$\star$	2000	0.86 ([i], [ɛ])
0.8	$\star$	470	1.08 ([æ])
1	$\star$	80	1.30 ([ɑ])

For guidance, we gave some examples of vowels that have comparable sectional area in the hypothesis of a circular aperture (see Story, 1996[25] for more accurate data).

Simulation 2: Second kinds of simulated signals are radiated sounds at the output of a uniform cylindrical waveguide with different values for the termination impedance. The aim is to synthesize vowel-like sounds with various bandwidths. The termination impedance was chosen as for a radiating half-sphere of radius r[26] - only resistive part was considered:

$$Z(k) = \rho_0 c \frac{(kr)^2}{1 + (kr)^2} \tag{8}$$

where $k = \omega/c$, ρ_0 and c are resp. air density and the speed of sound. To account for other surface losses in the waveguide, k was substituted with $k = \omega/c - j\alpha$, $\alpha = 1.2e - 5\sqrt{\omega}/0.01$ as in Hanna and al. 2016.[27] The waveguide length is similar to the length of the vocal tract (in average 16.5 cm). 200 samples were created. This time the parameter $u \in [0, 1]$ controls linearly the radius r of the cylinder, from $r = 0.2cm$ ($u = 0$) to $r = 1.3cm$ ($u = 1$). Some values of r are given on Table 1. See Supplemetary Material for more details on the generation and a sound file showing the transition between all the samples.

For Simulations 1&2, 200 values of $\beta^\star$ were computed as for the analyses on real data. More precisely, after each sample is decomposed in the dictionaries indexed by β, a 200×30 score matrix $h(u, \beta)$ was obtained. A Gaussian filter ($\sigma = 1$) was applied on the u-axis, then $\beta^\star$ was computed on each row. These values are plotted in Fig 4.

Results

In the following of the article, β refers to $\beta^\star$, that is the exponent that offers the best decomposition for the data under consideration.

The distinction between structured and non-structured sounds

The data obtained for broad categories (Fig 2) is similar to the distribution of exponents found by Stilp and Lewicki (see). The addition of the parameter h - as a reference to entropy - demonstrates the existence of two separate types of sounds being structured sounds ($h < 0.7$: semivowels, vowels, nasals) and non-structured sounds ($h > 0.7$: stops, affricates, fricatives). The latter are characterized by poor time and/or frequency structure. It does not mean that consonants have

no structure at all on a "macroscopic" scale (e.g. stops have a clear time pattern closure - burst - opening phase). What it means, however, is that non-structured sounds can be related to noise on small times scale of about 10 ms. The distinction between structured and non-structured sounds is relevant to our analysis because the factors determining β are different for each type. They are described separately and in detail in the following paragraphs. Most sparse signals are the approximant [ɹ] ($\beta = 0.82$) and the related vowels [ɚ] ($\beta = 0.85$) and [ɝ] ($\beta = 0.88$) with $h = 0.47$ for the three phonemes, not represented in Fig 3. h is more than twice larger for the least sparse sounds being the fricatives [f] ($h = 1.00$) and [ʃ] ($h = 0.95$).

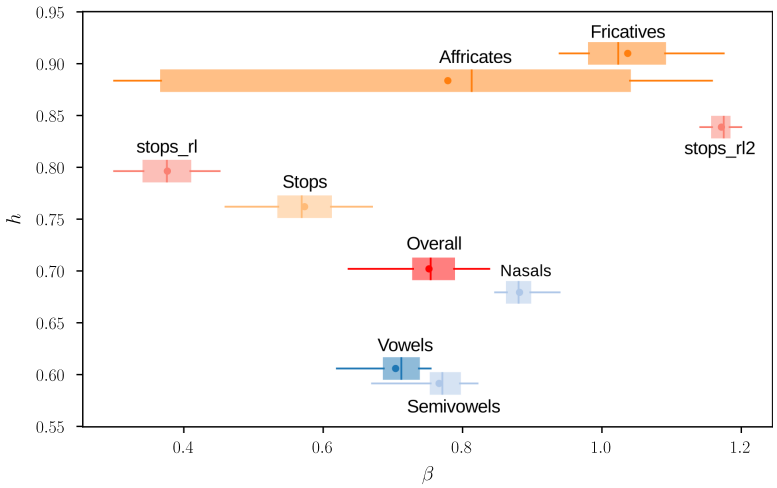

Figure 2: Distribution of American English broad phonetic categories in the (β, h) plane. *Box plots show quartiles Q1, Q2, Q3, [5%, 95%] percentiles (whiskers) and mean (dot) of bootstrap distributions based on 800 occurrences for each category. stops_rl and stops_rl2 are for first parts and second parts of stop releases (see text).*

Phonetic categories are unequal as for class variability. Confidence intervals on β are larger for consonants. They are extreme for affricates as almost the full range of values is covered. It comes as no surprise because affricates borrow acoustic features from both stops and fricatives. The parametric method introduced in this paper allows us to inspect the variations of β at a finer level of speech than broad phonetic categories. The more detailed figure on phonemes (Fig 3) shows that variability can sometimes be explained by contrasted values of β within a class. Most of the fricatives are in the region $\beta > 1$, but others ([v], [ð], [f]) are found in the opposite region $\beta < 0.5$. Other times, the same variability is found again on the level of phonemes, meaning that intra-phonetic variability does exist. For example, the affricate [dʒ], the fricative [θ] or the stops [t] and [k] have large confidence intervals. Although there is some scattering among the fricatives and stops, phonetic categories form consistent groups. Most of the time, phonemes that are close in the acoustic space are a also close in the (β, h) space. However, the phonetic categories that we use in Figure 2 do not always offer the best clustering of the data for statistical structure. Some phonemes seem to belong to a cluster different from their attributed category. Some examples are the aspirant [h] with the cluster of fricatives, the fricatives [v] and [ð] with stops, the stop [g] with the laterals [j] and [l], the *flap* [ɾ] with approximants.

Another measure of the significance of β is *contrast*, defined by the relative difference between h_{max} and h_{min} over the values of β:

$$h_{min} = \min_\beta h_\beta,\ h_{max} = \max_\beta h_\beta,$$

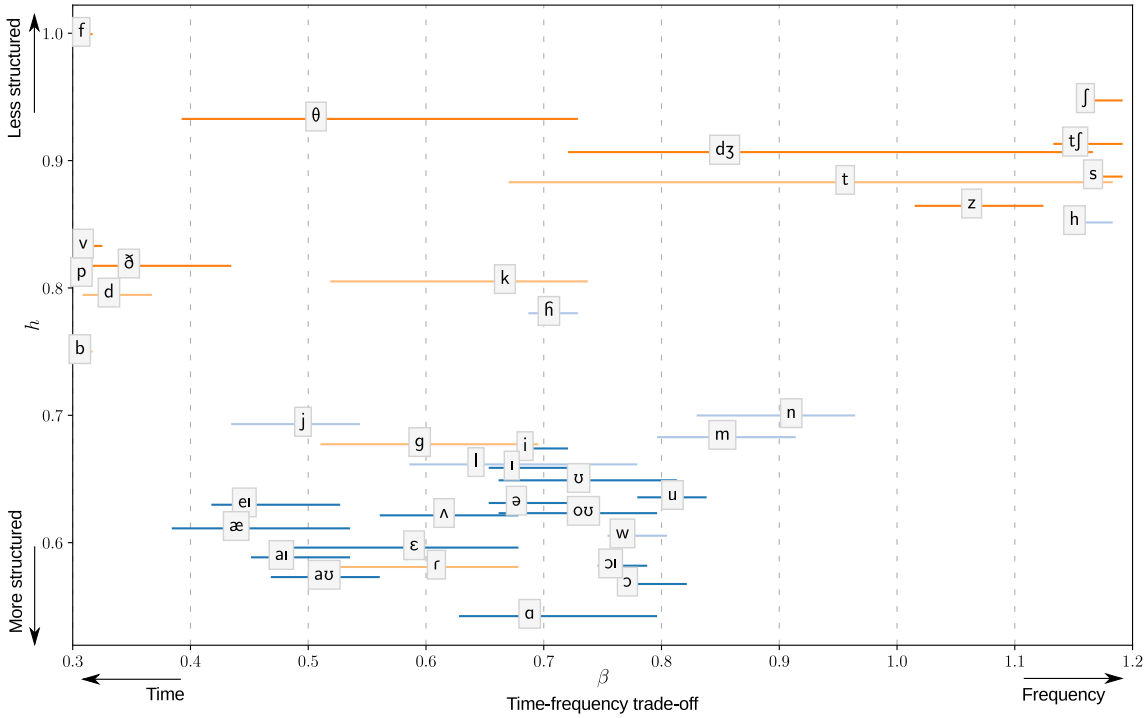

Figure 3: Distribution of American English phonemes in the (β, h) plane. *β is the exponent of the power law satisfied by frequency selectivity with respect to center frequency for the best representation of the data. h measures the lack of structure. The labels are positioned on the bootstrap distribution averages, the lines represent the 70% bootstrap confidence intervals. Bootstrap distributions are based on 400 occurrences for each phoneme and 3000 repetitions.*

$$c = \frac{h_{max} - h_{min}}{h_{max}}.$$

Contrast c is small for a flat score function and big when the score function has a clear minimum. $c = 1\%$ for speech as a whole when the scores are averaged over all the samples. The phonetic categories are in increasing order of c: affricates (0.4%), stops (0.7%), fricatives (1.7%), vowels and approximants (1.8%), and nasals (2.1%). Contrast again indicates a strong variability for stops and affricates which requires to be examined at a fine level.

Stops, fricatives, and affricates

A rough model for stops and fricatives is Gaussian noise modulated in time or in frequency. Gaussian white noise maximizes channel entropy for fixed output power. In the case of non-structured sounds, the coding strategy to reduce channel entropy amounts to try to reduce the mean amplitude value at the output of each channel. Hence efficient coding for non-structured sounds is more to find the *least bad* decomposition than to find the best one, trying to lessen the noise in the output channels. In the ideal case of an orthonormal decomposition which has the property to conserve total power along the transformation, the most efficient strategy is to have filters with an almost zero output and filters at full output. More generally, the optimal decomposition for noisy sounds shifts toward a time (resp. frequency) decomposition if it has a sharp power increase/decrease in the time (resp. frequency) domain. Simulation 1 on modulated noises is an illustration of this fact (Fig 4A). β takes the lowest value (time representation) when

the noise is multiplied by a Gaussian function localized in time. Then, β increases up to a median value as the Gaussian expands. At the same time, h increases because any structure is lost. Halfway through the simulation, at $u = 0.5$, generated samples are like white noise. h is at its peak while β has a rather erratic behavior. This is because the score function becomes flat as indicated by the low contrast value. The symmetrical pattern occurs when the simulation goes on frequency modulations. At $u = 1$, β takes the highest value (frequency representation). For a stop, the modulation function can be thought as a gate function in the time domain with random time for the *burst*. This is close to simulation with u=0 when the modulated noises show a rapid increase in intensity for a short amount of time. Under this extremely simplified model, the time representation is expected to be optimal. Stops are indeed associated with low β values, but the inner variability indicates a more complex behavior described further on.

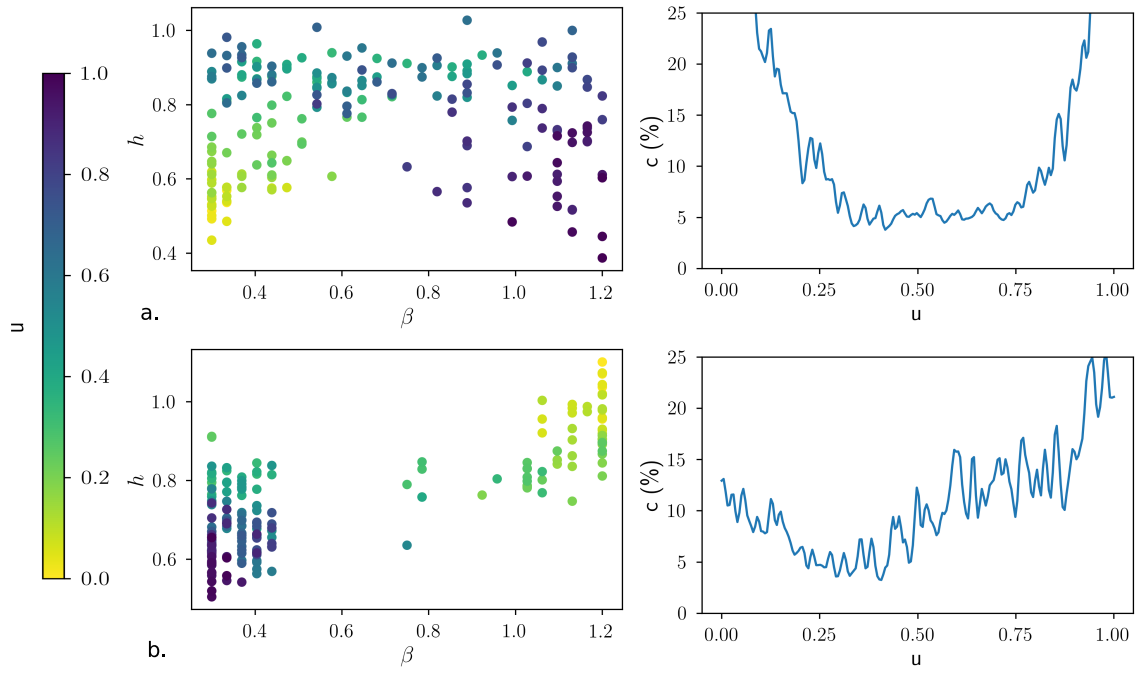

Figure 4: Best decomposition of artifical signals. *Scatter plot, left: best decomposition of simulated signals represented in the (β, h) plane. Right : contrast against the control parameter u.* ***a. Simulation 1:*** *Modulated noises from time modulation ($u = 0$) to frequency modulation ($u = 1$), passing by white noise ($u = 0.5$).* ***b. Simulation 2:*** *Radiated sounds at the output of a uniform cylindrical waveguide with different apertures, from $r = 0.2cm$ ($u = 0$) to $r = 1.3cm$ ($u = 1$).*

Fricatives are more explicit for now with Analysis 1 because they can be well approximated by stationary processes. Fricatives are the result of turbulent airflow occurring at a constriction in the vocal tract.[28] Some noise is produced and then filtered by the vocal tract, similar to simulated sounds passing through frequency modulations. Most fricatives yield values of β close to 1 consistent with this frequency description (Fig 3). It is at least true for the sibilant fricatives. These are filtered with a short cavity after the alveolar ridge and therefore present sharp rise/decay in the high frequency range. It is a clear trend for the hissing alveolar fricatives [s] ($c = 4\%$) and [z] ($c = 3\%$) and remains valid for the hushing post-alveolar fricative [ʃ] ($c = 1\%$). On the other side, labial or dental fricatives [f], [θ], which are less affected by vocal filtering resulting in wide-band noise, are associated with lower β values ($c = 1\%$, $c = 0.6\%$ resp.). Voiced fricatives are an interesting intermediary case for what has been seen as they are affected by both time and

frequency modulations. In addition to vocal filtering, the sound intensity follows the repeated openings and closures of the glottis. The coincidence of time and frequency events is likely to explain why the points move to bottom-left on the (β, h) plane when replacing the unvoiced versions of the fricatives by the voiced ones (compare [s], [h], [θ], [f] with [z], [ɦ], [ð], [v] resp.). This fact is also true for stops to some extent (compare [t], [k], [p] with [d], [g], [b]).

The description of stops and affricates must be refined to take into account dynamic aspects. Stops and affricates have the following temporal pattern: closure - release burst - release transition. During the closure, no sound is emitted (or a low frequency sound only as in [b]). Closures are always ignored in the analyses because they do not contain high frequency information. Then, the release can be decomposed into two phases. The first phase is the burst following the instant of the occlusion release. During this phase of small duration (few milliseconds), the intensity increases and decreases rapidly and the power spectrum is almost flat. The second phase, the release transition or opening/aspiration phase, is similar to a fricative sound as there is still some obstruction at the place of articulation and/or aspiration at the glottis.[28] Figure 5 demonstrates that the dual nature of stops and affricates after the occlusion is also revealed by β. For this particular analysis, the releases were separated into two parts of same duration. First and second parts of the releases are respectively indexed by the suffixes *_rl* and *_rl*2. Whereas the *_rl* parts containing the bursts yield minimal values of β, the *_rl*2 parts yield higher values close to 1. Values for stops as a category are also reported in Figure 2. These figures demonstrate that the opening phase of stops and fricatives are in reality alike with regard to statistical structure. Analysis 2 makes it possible to go further into the temporal description of stops and affricates. Figure 6 describes the time evolution of β for some phonemes. β is stable for vowels or nasals - apart from diphthongs. But it increases during the occurrences of stops and affricates, joining the extreme values. However this transition occurs more or less steeply depending on the nature of the opening phase. The stop [t] whose tail is similar to the sibilant fricative [s] has a fast transition after the burst. In contrast, the stop [p] has a more gradual transition. Indeed, the opening phase is similar to the low β fricative [f] on which some formant structure appears gradually when the back cavity plays a role again (as for the high β fricative [h]). Note that although the change is less pronounced, fricatives have also an upward shift of β at their onset.

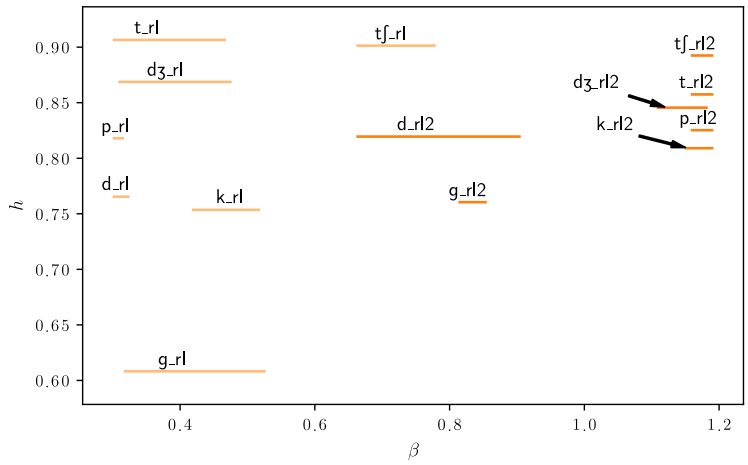

Figure 5: Detailed distribution of stops and affricates in the (β, h) plane. *When stop or affricate releases are separated into two parts of same duration, first parts _rl (including the bursts) are best represented in a dictionary of low β value but second parts _rl2 are best represented in a dictionary of high β value (plotted: distribution averages and 70% bootstrap confidence intervals).*

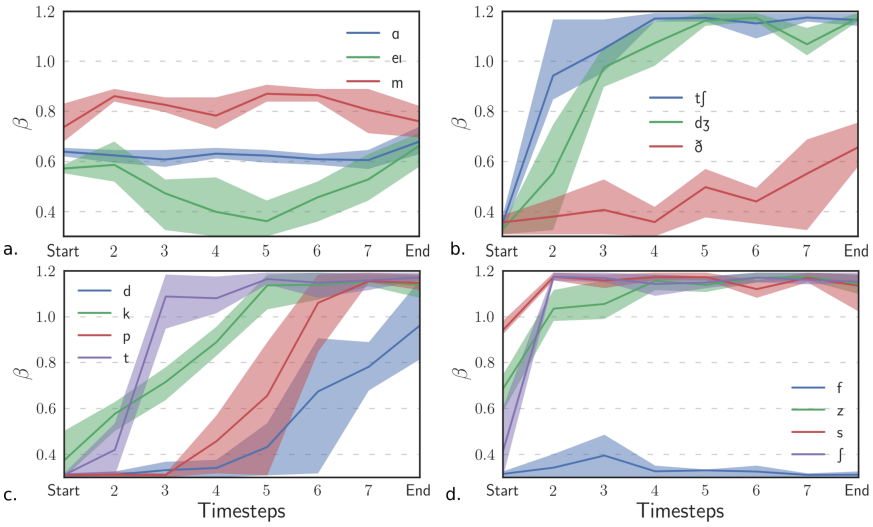

Figure 6: Temporal evolution of β for some phonemes. *The timesteps represent the relative times between the beginning (Start) and end (End) of the occurrences at regular intervals. The region filled represents 70% confidence intervals.* ***a.*** *Vowel:* ɑ, *diphtong:* eɪ, *nasal:* m ***b.*** *Affricates:* tʃ, dʒ, *fricative:* ð ***c.*** *Stops:* d, k, p, t ***d.*** *Fricatives:* f, z, s, ʃ.

Vowels, semivowels, and nasals

The above reasoning for non-structured sounds does not apply to vowels. We have to look for the acoustic properties relevant to signal structure. The structure of vowels can be seen both in time and in frequency. Along the frequency axis, vowels are characterized by a few spectral peaks arranged at almost regular intervals ($\sim$ 1kHz): these are called the formants and correspond to the resonances of the vocal tract. On the time axis, the signal presents peaks of intensity at the instant of glottal closure remaining true if the signal is band-passed around formants. The latter statement stands at least for the first formants as higher formants can be excited at other instants, especially at the glottal opening.[29] On one glottal cycle, a naive image of the underlying structure in the time-frequency plane can be a comb shape whose *teeth* represent the formants. The complete structure can not be perfectly covered by a Gabor filter bank: a representation is always a compromise between time and frequency favoring either the glottal pulse or the tails of the formant oscillations. From a frequency point of view, it means that either the wideband parts of the signal associated with low response and low group delay or the narrow bands associated with the formants, are neglected. This competition between the pulse and the formant oscillations has a visible effect on the quality factor of the efficient coding filters at medium frequencies, at 1kHz and up to 5kHz. When ICA is performed on the front vowels which have high second formants, the quality factor goes from 1.3 at 1kHz to 3 at 1.7kHz. On back vowels, the quality factor increases at 1kHz more steeply (see Fig 1 of ref. 16).

The optimal representation, anyhow, is expected to be related to formant bandwidths. Formant bandwidths are associated with damping level and acoustic losses, in particular wall losses at low frequencies and radiation losses at high frequencies.[30, 31] Higher formants play a prominent role in the determination of β, therefore radiation at the lips has good chance to be one key factor for statistical structure of vowels. The resistive part of the termination impedance increases with frequency and mouth aperture.[26, 32] Simulation 2 demonstrates the impact of aperture radius on β with synthetic vowels generated by an uniform cylindrical waveguide (Fig 4B). A small aperture ($u = 0$) is associated with low damping, narrow bandwidths and β at maximum. We get the opposite for a large aperture ($u = 1$). The two extremes are separated by a steep phase transition

occurring at $u = 0.25$ ($r = 0.7cm$). Note that the correction of the wave number that is used in the simulation corresponds to a low estimation of surface losses. Hanna and al. proposed instead to increase this correction by a factor 5 to be closer to reality:[27] this change makes the transition to be very close to $u = 0$. The estimation of the degree of acoustic radiation with the mouth aperture radius is a rough approximation, in particular it does not take into account inner reflexions.[30] Still, the same trend can be observed for real data although the transition is more gentle and less apparent. The vowels form a tight cluster around $\beta = 0.6 \pm 0.2$, however the unrounded vowels and diphthongs [æ], [ɛ], [eɪ], [aʊ], and [aɪ] give smaller values than the rounded vowels [u], [ʊ], and [ɔ] (Fig 3). The intermediate sounds [ə], [ʌ], [i], and [ɪ] lie in between. The vowel [ɑ], however, does not fit the rest of the distribution. The simulation reveals the parallel trend that h decreases along with β. The most likely reason for this phenomenon is that narrow bandwidths (high β values) fill the time-frequency domain with longer tails whereas damped signals (low β values) are localized in time, therefore sparser. But it is not very clear whether this rule can apply to real data.

In the continuity of vowels are nasals with higher values of β and h, meaning that nasals are better described in the frequency space. It can be explained by the presence of antiresonances surrounding the formants which have the effect to cut their bandwidths. Note that it is rather contrary to the known fact that nasals have wider bandwidths because of greater surfaces losses. This is not contradictory to the extent that the region of interest is in the high frequency range plus the values of wide bandwidths (e.g -10dB bandwidths) here are more significant than the -3dB narrow bandwidths. Semivowels appear to yield the same range of values of β and h as vowels. The rhotic approximant and r-colored vowels occupy the lower right part of the cluster ($\beta = 0.8, h = 0.47$) in the (β, h) plane. In fact, it seems that all the vowels that can relate to the [r] sound are pushed to the bottom right. This includes the back vowels [ɑ] and [ɔ]. This phenomenon could offset the effects demonstrated by the simulation for back vowels.

Discussion

Distribution of the exponent β

The parameter β is the exponent of the power law satisfied by frequency selectivity against frequency for the optimal time-frequency decomposition of the signal. The overall distribution of β for the broad phonetic categories is in agreement with the regression slopes found by Stilp and Lewicki (compare Fig 2 with Fig 2&3 in Ref 16). Especially, β is found between 0.7 and 0.8, with both ICA and the current method, for speech data as a whole. The most noticeable gap is for semivowels. Semivowels are special because they are often associated to a low score h and the value of β is not necessarily very significant in this case. [ɹ] sounds especially present a strong frequency decrease in the high frequencies, hence the underlying structure for the high-passed filtered signal is essentially a prominent peak in the frequency domain near 1kHz. In this situation it is hard to draw any conclusion from the values taken by β, but one explication for a high β value specific to the method based on dictionaries could be that higher frequency filters try to avoid the intensity peak at 1kHz and that higher frequencies are too weak to play a role. The last comments could also apply to the r-colored vowels, including the low back vowels. The detailed distribution at the level of phonemes (Fig 3) shows consistency of the phonetic categories, at least after the division of phonetic units is refined. Phonemes that are close acoustically are found together in the (β, h) plane. The results of Stilp and Lewicki on the impact of vowel frontness on β (Fig 4 in Ref 16) could not be replicated, but the figure on single phonemes (Fig 3) tends to highlight other vowel features that have much more pronounced effects.

Relation between β and acoustic features

Intraclass variability and even intra-phonetic variability give an indication of the acoustic properties relevant to β. Some of these properties which were put forward in are in agreement with previous proposals, but others are new or clarify some previous ideas.

In 2002, Lewicki examined whether the spectral tilt could explain the efficient coding filter properties.[7] The spectral tilt is the $1/f$ law satisfied by the power spectrum density. His conclusion was that there is no relation between the two. The average spectrum density has indeed a small impact on signal structure, because the efficient coding filters are localized in frequency - it has an effect on the weighting between midrange and high frequencies but not on the atomic components. An exception is that the addition of a decrease or increase in the frequency power spectrum leads to the emergence of frequency structure in the case of non-structured sounds. We have seen this phenomenon on fricatives: high-pass filtered hissing sounds [s] and [z] are associated with a higher value of β and a lower value of h compared to broadband noise [f]. In 2013, Stilp and Lewicki listed three others acoustic factors impacting β: harmonicity, acoustic transience and bandwidths.[16] We argue that the *F0* periodicity plays little if no role because the efficient coding filters are shorter than the period length. More generally, acoustic changes of characteristic time greater than the duration of a glottal cycle (e.g. coarticulation, formant transitions) may not have a significant impact as such on the efficient coding filters. However the fact that voiced sounds are characterized by scarce time-localized excitations has the effect to reduce h and to enhance time localization, thus to decrease β at the same time. Vowels have been shown to be associated with relatively small values of β, a result that may be counterintuitive to some. Vowels are sustained sounds that are often believed to be better captured by a frequency representation. This view might be biased by the source-filter model that focuses on the resonances in the frequency space and makes extensive use of Fourier analysis. On the contrary, this paper suggests that a time representation would be more appropriate for the efficient coding of vowels. The fine-grained analysis of statistical structure supports the hypothesis that β relates to the formant bandwidths in the high frequencies for vowels and nasals. It suggests lip radiation and the existence of antiresonances as key acoustic factors. Finally, we agree with Stilp and Lewicki on *transiency* being the most prominent acoustic factor for speech statistical structure since β reaches its low point on stop bursts.

The fact that β is bound to a few acoustic properties could mean that the analysis conducted here is not only specific to speech. The reasoning on consonants and non-structured sounds would probably equally well apply to many environmental sounds. The fact that β is negatively correlated with mouth aperture could be valid as well for other animal vocalizations.

Congruency with the efficient coding hypothesis and cochlear signal processing

Early in the study of efficient coding of speech, Lewicki drew a parallel between the theoretical optimal representation of speech and the properties of the mammalian cochlea.[7] Physiological measures of the cochlear frequency selectivity in cats based on tuning curves also satisfy a power law in the high frequency range 1-8kHz (see Fig 3 of Ref 33). The exponent is about the same for ICA filters, although slightly lower (0.6 compared to 0.7-0.8). This comparison is a reproduction in the auditory field of various analyses done previously in the vision domain. ICA or sparse coding of natural image scenes are demonstrated to yield oriented Gabor wavelet-like filters analogous to the visual receptive fields in the primary visual cortex.[22, 34] Together these are pieces of evidence that help to show some confidence for the efficient coding hypothesis which states that sensory systems have adapted to natural stimuli by reducing redundancy.[3] Speech, however, is special because it is a human-controlled stimulus, even though it is still subject to acoustic constraints. There is some ongoing debate on the specificity of human auditory tuning,[35] especially at low

input levels,[36] but it is generally agreed than humans are not very different from unspecialized mammals regarding auditory tuning. As speech emerged lately compared to the time of the evolution of the cochlea, Lewicki hypothesized that speech evolved to be optimally coded by the mammalian auditory system. The diversity between transient and sustained sounds was proposed as an explanation for the median β value - the same agreement with physiological data was obtained with a mixture of environmental sounds and animal vocalizations.[7] However, the scattering of β when ICA is performed on subclasses of speech could imply a more efficient coding scheme. Stilp and Lewicki's proposal is that it is congruent with the diversity of time-frequency trade-offs found in the characteristic responses of the neurons in the cochlear nucleus, but they admit that these observations stand for single tones and maybe not for complex stimuli.[16] It can be added that the recombination of filters is a compute-intensive task that could not be included in an efficient coding scheme easily.

Instead, we argue that if an efficient coding strategy is implemented physiologically to adapt the neural representation to subparts of speech, it is at the level of the peripheral auditory system. The assumption that the auditory filters are fixed and independent of the input is wrong. The active mechanism in the cochlea makes auditory filtering highly nonlinear. In particular, the shape of auditory filters changes with the input and more specifically with sound intensity level as first approximation. Filters expand with sound intensity level, because the cochlear amplifier, thanks to which the ear is highly frequency selective, is compressive. The strength of this nonlinearity increases with frequency,[37, 36, 38] consistent with the starting assumption that the optimal representation does not change much at 1kHz while it has great variations at frequencies close to 8kHz. The proposition that peripheral auditory processing matches the fine-grained structure of speech would make sense if β is negatively correlated to sound intensity. This appears to be the case, at least if the transient parts of stops and affricates are ignored (Fig 7). The testing of this hypothesis is not the purpose of this article, but the current statistical study allows us to outline an efficient coding scheme possibly in agreement with cochlear nonlinearities. Low-level sounds are mainly consonants (stops, affricates, fricatives) which are better decomposed in high β dictionaries - i.e. frequency selective filters - if the transient parts are removed. Transients, however, are more efficiently decomposed with a low value of β. The special processing of transients is physiologically plausible as the active system may not be at its full strength immediately at the onset of a sound, meaning that the burst part could be analyzed with a broader tuning.[39] When the intensity level increases, auditory filters broaden as cochlear compression begins. Coherently, high intensity sounds are essentially vowels and formant bandwidths increase with jaw opening, as does intensity level. Synthetic vowels from Simulation 2 present the same decrease in β with respect to intensity as observed in Fig 7, although the transition is more severe and steep (see fig. in supporting information). In figure 7, the curve with burst parts removed seems to present a knee at -30/-20dB, potentially in accordance with the same kind of pattern for cochlear compressive nonlinearity. Contrast is under 1% before the knee and above after. Again, this coding strategy could not only be relevant to speech sounds but to other natural sounds as well.

One of the limitations of comparisons with the auditory system based on Gabor filters is that the cochlear filters are actually not symmetric.[40] Symmetry is a characteristic shared by the filters at the output of standard ICA, but asymmetric filters can be obtained if sparse response patterns are reinforced by more complicated coding techniques.[9] This may be the mark of extra coding principles implemented peripherally or centrally in the auditory system. ICA can depart from the power law model when being applied to specific classes of speech sounds. The parametric model is still convenient because speech sounds can be compared at a fine level of detail with a single parameter. Despite its simplicity, the model is well suited for many classes of speech sounds and extends quite well to others. It is also the easiest way to realize a flexible representation of the input. Including the intensity level as a control parameter would provide a simple and effective coding strategy because intensity level is an immediate indicator contrary to the phonetic class to

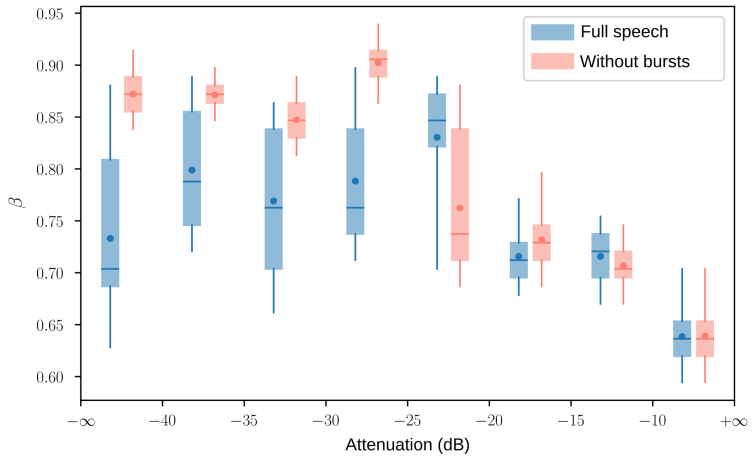

Figure 7: Exponent β with respect to intensity level. *The exponent β associated with the best decomposition with respect to intensity level in dB (ref:max) by intervals of 5dB.* ***In blue:*** *Full speech,* ***in red:*** *same but with the first parts of stop and affricate releases removed. Box plots show quartiles, [5%, 95%] percentiles (whiskers) and mean (dot) of bootstrap distributions based on 2 500 16ms-slices of speech.*

which the sound belongs. The parametric approach is a tool that comes not in replacement but in complement of standard ICA and other machine learning techniques in the study of the statistical regularities of speech. In particular, the correlations that intervene in the determination of β are under 10 ms, regularities at higher time scales have also to be exploited for an efficient speech coding system to be complete.

Conclusion

This work demonstrates that a parametric approach, based on dictionaries of Gabor filters and a sparsity score, yields the same power laws for frequency selectivity with respect to frequency as standard ICA applied to speech. The statistical structure was analyzed at a fine level of speech by means of the parametric approach (Fig 3). This allowed the power laws to be linked to acoustic features enumerated according to the dichotomy between structured and non-structured sound. Among non-structured sounds, stops and affricates have been shown to be *biphasic* after the closure: a transient part best captured by a time representation followed by a fricative-like sound best captured by a frequency representation. For structured sounds, mainly vowels, the power laws relate to formant bandwidths partly determined by the degree of acoustic radiation at the lips. The frequency selectivity of cochlear filters also follows a power law whose exponent decreases with sound intensity level. The detailed analysis of statistical structure shows that with a few restrictions the exponent is negatively correlated with sound intensity for the efficient coding filters; hence, the present study suggests a connection between nonlinear cochlear signal processing in mammals and the statistical structure of speech. Further analysis should be carried out to assess whether the efficient coding hypothesis can be pushed forward for auditory coding.

Acknowledgments

I gratefully acknowledge PSL University for the scholarship of the PSL-Maths program. I thank JP. Nadal, L. Bonnasse-Gahot, C. Lorenzi, J. Gervain and R. Guevara Erra for their valuable help and advise. This work has been partly supported by the project “SpeechCode” of the French National

Research Agency, the ANR, contract ANR-15-CE37-0009-03.

References

[1] Claude E Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(July 1928):379–423, 1948.

[2] Fred Attneave. Some informational aspects of visual perception. *Psychol. Rev.*, 61(3):183–193, 1954.

[3] H B Barlow. Possible principles underlying the transformations of sensory messages. In W A Rosenblith, editor, *Sensory Communication*, pages 217–234. MIT Press, Cambridge, MA, 1961.

[4] Horace Barlow. Redundancy reduction revisited. *Network: Computation in Neural Systems*, 12(3):241–253, 2001.

[5] Joseph J. Atick. Could information theory provide an ecological theory of sensory processing? *Network: Computation in Neural Systems*, 3(2):213–251, 1992.

[6] Eero P Simoncelli and Bruno A Olshausen. Natural Image Statistics and Neural Representation. *Annu. Rev. Neurosci.*, 24(1):1193–1216, 2001.

[7] Michael S Lewicki. Efficient coding of natural sounds. *Nat. Neurosci.*, 5(4):356–363, 2002.

[8] Jong-Hwan Lee, Te-Won Lee, Ho-Young Jung, and Soo-Young Lee. On the Efficient Speech Feature Extraction Based on Independent Component Analysis. *Neural Process. Lett.*, pages 235–245, 2002.

[9] Evan C Smith and Michael S Lewicki. Efficient auditory coding. *Nature*, 439(7079):978–82, 2006.

[10] Nicholas A. Lesica and Benedikt Grothe. Efficient temporal processing of naturalistic sounds. *PLoS ONE*, 3(2):e1655, feb 2008.

[11] F. A. Rodriguez, C. Chen, H. L. Read, and M. A. Escabi. Neural Modulation Tuning Characteristics Scale to Efficiently Encode Natural Sound Statistics. *J. Neurosci.*, 30(47):15969–15980, nov 2010.

[12] Nicole L Carlson, Vivienne L Ming, and Michael Robert DeWeese. Sparse codes for speech predict spectrotemporal receptive fields in the inferior colliculus. *PLoS Comput. Biol.*, 8(7):1002594, 2012.

[13] Wiktor Młynarski and Josh H. McDermott. Learning Midlevel Auditory Codes from Natural Sound Statistics. *Neural Comput.*, 30(3):631–669, mar 2018.

[14] Aapo Hyvärinen. Survey on Independent component analysis. *Neural Comp. Surveys*, 2:94–128, 1999.

[15] Jean-Pierre Nadal and Nestor Parga. Nonlinear neurons in the low-noise limit: a factorial code maximizes information transfer. *Network: Computation in neural systems*, 5(4):565–581, 1994.

[16] Christian E Stilp and Michael S Lewicki. Statistical structure of speech sound classes is congruent with cochlear nucleus response properties. In *Proc. Meet. Acoust. 166ASA*, volume 20, page 050001, nov 2013.

[17] Ramon Guevara Erra and Judit Gervain. The efficient coding of speech: Cross-linguistic differences. *PLoS ONE*, 11(2):e0148861, feb 2016.

[18] Karlheinz Gröchenig. *Foundations of Time-Frequency Analysis*. Applied and Numerical Harmonic Analysis. Birkhäuser Boston, Boston, MA, 2001.

[19] Hans G Feichtinger and Massimo Fornasier. Flexible Gabor-wavelet atomic decompositions for L2-Sobolev spaces. *Annali di Matematica Pura ed Applicata*, 185(1):105–131, 2006.

[20] John S. Garofolo, Lori F. Lamel, Wiliam M. Fischer, Jonathan G. Fiscus, David S. Pallett, and Nancy L. Dahlgren. The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CD-ROM. *NIST*, pages 1–94, 1986.

[21] Saeed Gazor and Wei Zhang. Speech probability distribution. *IEEE Signal Process. Letters*, 10(7):204–207, jul 2003.

[22] Bruno A Olshausen and D Field. Emergence of Simple-Cell Receptive Field Properties by Learning a Sparse Code for Natural Images. *Nature*, 381:607–609, 1996.

[23] Michael S Lewicki and Bruno A Olshausen. Probabilistic framework for the adaptation and comparison of image codes. *JOSA A*, 16(7):1587–1601, 1999.

[24] Aapo Hyvärinen and Mika Inki. Estimating overcomplete independent component bases for image windows. *J. Math. Imaging Vision*, 17(2):139–152, 2002.

[25] Brad H Story, Ingo R Titze, and Eric A Hoffman. Vocal tract area functions from magnetic resonance imaging. *J. Acoust. Soc. Am.*, 100(1):537–554, 1996.

[26] Mario Fleischer, Silke Pinkert, Willy Mattheus, Alexander Mainka, and Dirk Mürbe. Formant frequencies and bandwidths of the vocal tract transfer function are affected by the mechanical impedance of the vocal tract wall. *Biomech. Model. Mechanobiol.*, 14(4):719–733, aug 2015.

[27] Noel Hanna, John Smith, and Joe Wolfe. Frequencies, bandwidths and magnitudes of vocal tract and surrounding tissue resonances, measured through the lips during phonation. *J. Acoust. Soc. Am.*, 139(5):2924–2936, 2016.

[28] Keith Johnson. *Acoustic and Auditory Phonetics*. Wiley-Blackwell, 2003.

[29] Helen M Hanson and Erika S Chuang. Glottal characteristics of male speakers: Acoustic correlates and comparison with female data. *J. Acoust. Soc. Am.*, 106(2):1064–1077, 1999.

[30] Gunnar Fant. Vocal tract wall effects, losses, and resonance bandwidths. *STL-QPSR*, 13:28–52, 1972.

[31] Kenneth N. Stevens. *Acoustic phonetics*. MIT Press, 1998.

[32] Marc Arnela and Oriol Guasch. Finite element computation of elliptical vocal tract impedances using the two-microphone transfer function method. *J. Acoust. Soc. Am.*, 133(6):4197–4209, jun 2013.

[33] Roger L Miller, John R Schilling, Kevin R Franck, and Eric D Young. Effects of acoustic trauma on the representation of the vowel /ϵ/ in cat auditory nerve fibers. *J. Acoust. Soc. Am.*, 101(6):3602–3616, 1997.

[34] J Hans van Hateren and Dan L Ruderman. Independent component analysis of natural image sequences yields spatio-temporal filters similar to simple cells in primary visual cortex. *Proc. Roy. Soc. London B: Biological Sciences*, 265(1412):2315–2320, 1998.

[35] Geoffrey A Manley. Comparative Auditory Neuroscience: Understanding the Evolution and Function of Ears. *J. Assoc. Res. Otolaryngol.*, 18(1):1–24, feb 2017.

[36] Andrew J Oxenham and Andrea M Simonson. Level dependence of auditory filters in nonsimultaneous masking as a function of frequency. *J. Acoust. Soc. Am.*, 119(1):444–453, jan 2006.

[37] Xuedong Zhang, Michael G. Heinz, Ian C. Bruce, and Laurel H. Carney. A phenomenological model for the responses of auditory-nerve fibers: I. Nonlinear tuning with compression and suppression. *J. Acoust. Soc. Am.*, 109(2):648–670, feb 2001.

[38] Eric Verschooten, Luis Robles, Damir Kovačić, and Philip X Joris. Auditory nerve frequency tuning measured with forward-masked compound action potentials. *J. Assoc. Res. Otolaryngol.*, 13(6):799–817, 2012.

[39] Mario A Ruggero, Nola C Rich, Alberto Recio, S Shyamla Narayan, and Luis Robles. Basilar-membrane responses to clicks at the base of the chinchilla cochlea. *J. Acoust. Soc. Am.*, 103(4):1972–89, apr 1998.

[40] Laurel H. Carney and Tom C T Yin. Temporal coding of resonances by low-frequency auditory nerve fibers: single-fiber responses and a population model. *J. Neurophysiol.*, 60(5):1653–1677, 1988.