Proceedings of Meetings on Acoustics

Volume 23 http://acousticalsociety.org/

169th Meeting of the Acoustical Society of America

Pittsburgh, Pennsylvania
18-22 May 2015

Speech Communication: Paper 2pSC21

Languages across the world are efficiently coded by the auditory system

Christian E. Stilp
Department of Psychological and Brain Sciences, University of Louisville, Louisville, KY; christian.stilp@louisville.edu

Ashley A. Assgari
Department of Psychological and Brain Sciences, University of Louisville, Louisville, KY; ashley.assgari@louisville.edu

Independent Component Analysis (ICA) is a powerful method for uncovering statistical structure in natural stimuli. Lewicki (2002 *Nature Neuroscience*) used ICA to examine statistical properties of human speech. Filters that optimally encoded speech were an excellent match for frequency tuning in the cat auditory nerve, leading to suggestions that speech makes efficient use of coding properties in the mammalian auditory system. However, Lewicki only examined American English, which is neither normative nor representative of the world's languages. Here, fifteen languages were examined (Dutch, Flemish, Greek, Javanese, Jul'hoan, Mandarin Chinese, Norwegian, Swedish, Tagalog, Tahitian, Urhobo, Vietnamese, Wari', Xhosa, Yeyi). Each recording contained speech tokens from native speakers without any background noise for at least seven minutes. Maximum likelihood ICA was used to create statistically optimal filters for encoding sounds from each language. These filters were then compared to the same physiological measures analyzed in Lewicki (2002). Languages produced a range of ICA solutions, as expected, but were highly consistent with both statistically optimal filters for American English and physiological measures. Results significantly extend Lewicki (2002) by revealing agreement between response properties of the auditory system and speech sounds in a wide range of languages.

Published by the Acoustical Society of America

 [DOI: 10.1121/2.0000064]
Received 23 June 2015; Published 22 July 2015

INTRODUCTION

Sensory systems adapt and evolve to accommodate stable or regular inputs in the environment. This is evident in the close relationship between sensory system response properties and natural signal statistics. Response properties of the visual system are well-tuned to the statistics of natural images (Field, 1987; Bell & Sejnowski, 1995; Olshausen & Field, 1996; Geisler, 2008), and recent efforts reported correspondences between auditory system response properties and natural sound statistics as well. Using Independent Components Analysis (ICA), Lewicki (2002) derived filters that were statistically optimal for encoding human speech sounds. Considerable similarity existed between these filters and frequency tuning curves measured in the cat auditory nerve, leading to the proposal that speech might be optimally adapted to the coding capacity of the auditory system (Smith & Lewicki, 2006).

However, these analyses suffer from a serious limitation in that they only analyzed American English. English is neither a normative nor representative language, given its finite sampling of roughly 40 phonemes from over 800 phonemes observed across over 5,000 documented languages. The statistical structure of one language is not guaranteed to hold for (all) other languages. The statistics of natural stimuli can vary from one stimulus class to the next (*e.g.*, Lewicki, 2002; Torralba & Oliva, 2003; Lewis *et al.*, 2012; Stilp & Lewicki, 2014). This holds true for approaches using ICA, as statistically optimal filters differed for encoding animal vocalizations, environmental sounds, or human speech (Lewicki, 2002), and differed across individual categories of speech sounds (English consonant sounds arranged by manner of articulation; Stilp & Lewicki, 2014). At issue is whether the statistics of human speech are consistent across languages, thereby validating claims that speech is efficiently coded by the auditory system, or if stimulus statistics vary by language (*i.e.*, by stimulus class), challenging the generalizability of Lewicki's (2002) claims.

Here, we investigated the statistical structure of each of a wide range of languages. Comparisons are made to assess the similarity of language statistics to each other, to American English, and most importantly to physiological measures of the mammalian auditory system.

METHODS

1. Stimuli

Language recordings were collected from multiple online resources including: Global Recordings (http://globalrecordings.net/en), the Endangered Languages Archive (http://elar.soas.ac.uk/), the University of Minnesota Center for Advanced Research on Language Acquisition (http://www.carla.umn.edu/), Loyal Books (http://www.loyalbooks.com/), and Speech Ocean (http://speechocean.com/). All recordings were clearly spoken and free of any audible background noise. Recordings from different talkers were collected whenever possible (see Table I). Individual recordings for each language were concatenated to meet a total duration of approximately ten minutes (with the exception of Tahitian, which was seven minutes in duration). Concatenated recordings were then resampled at 16 kHz.

Language	Family	# Talkers
Dutch	West Germanic	1
Greek	Hellenic	2
Javanese	Western Malayo-Polynesian brand of the Austronesian languages	1
Ju'|hoan	Khoisan Language, !Kung Family	3
Mandarin Chinese	Sinitic branch of Sino-Tibetan	87
Norwegian	North Germanic	5
Swedish	North Germanic	1
Tagalog	Central Philippine group of the Philippine subgroup of the Western-Malayo-Polynesian branch of the Malayo-Polynesia subfamily of the Austronesian language family	1
Tahitian	Polynesian Languages, Austronesian	3
Urhobo	Niger-Congo	5
Vietnamese	Muong-Vietnamese subgroup of the Mon-Khmer subfamily of the Austro-Asiatic family	1
Vlaams (Flemish)	West Germanic	6
Wari	Chupacura, Madeira	3
Xhosa	Nguni group of the Bantu sub branch of the Benue-Congo brand of the Niger-Congo subfamily of the Niger-Khordofanian family	11
Yeyi	Bantu	5

TABLE I. List of languages analyzed, family of origin, and number of unique talkers.

Language databases were constructed following the methods of Lewicki (2002) and Stilp and Lewicki (2014). Concatenated recordings for a given language were high-pass filtered at 125 Hz (100-coefficient finite impulse response filter), set to zero mean and unit variance, then divided into 8-ms segments.

2. ICA

Details of the basic ICA algorithm for deriving statistical structure in natural stimuli have been provided by Bell and Sejnowski (1995). Briefly, ICA assumes that the observed data **x** are the result of linear combinations of **s**:

$$\mathbf{x} = \mathbf{A}\mathbf{s} \quad [1]$$

where **A** is a mixing matrix whose columns constitute basis functions, and **s** is an independent source vector with components s_i that are statistically independent from each other. **A** and **s** are unknown, so ICA estimates them according to Equation 2:

$$\mathbf{y} = \mathbf{W}\mathbf{x} \quad [2]$$

W is an unmixing matrix of the same dimensionality as **A** ($\mathbf{W} = \mathbf{A}^{-1}$), making the output **y** the recovered source vector which approximates **s** up to scaling and permutation. Thus, the rows of **W** are statistically optimal filters for recovering source signals **s** from the observed mixtures **x**.

Each language was analyzed using maximum likelihood ICA (Pearlmutter & Parra, 1997). The natural gradient extension was used to facilitate convergence (Amari *et al.*, 1996). A Laplacian prior was used to model the distribution of source signals in **s** and correspondingly in **y**

(Gazor & Zhang, 2003). **W** was iteratively updated by stochastic gradient descent, resulting in the learning rule in Equation 3:

$$\Delta \mathbf{W} = [\mathbf{I} - \text{sign}(\mathbf{y})\ \mathbf{y}^T]\mathbf{W} \quad [3]$$

where **I** is the identity function, sign(·) is the sign function, and $\mathbf{y}^T$ is the transpose of **y**. **W** is initialized to the identity matrix, and Δ**W** is the change in the unmixing matrix that is added to **W** at each iteration. ICA simulations were conducted for 20,000 iterations, with the learning rate set to .01 for the first 16,000 iterations and reduced to .001 for the final 4,000. Different batches of 500 8-ms samples were randomly selected at each iteration for analysis. This process was repeated ten times for each language. ICA results were highly stable across simulations of a given language, thus representative results are shown.

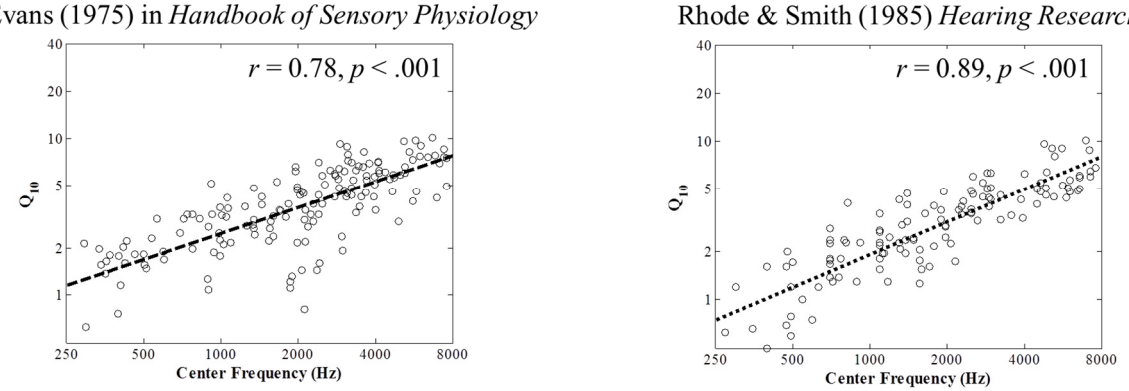

FIGURE 1: Filter quality measures for auditory nerve fibers in the cat auditory system. Each plot depicts Q_{10} measures as a function of filter center frequency measured up to 8 kHz. Each circle depicts tuning for a single auditory nerve fiber. Dotted lines are linear regressions fit to the data, with correlation coefficients listed at the top of each plot. These are the datasets used by Lewicki (2002) to compare physiological measures to statistically optimal encoding of American English.

3. Filter Regression Analysis

After conducting ICA on a given language, peak (center) frequencies for each filter (row) in **W** were identified using FFT. Q_{10} was then measured for each filter when possible (some filters could not be analyzed owing to the lack of 10 dB decreases both above and below the center frequency; this occurred at lower and upper extremes of the 8 kHz signal bandwidth). A linear regression was fit to Q_{10} as a function of center frequency with both metrics on logarithmic scales. Regression functions were then visually compared to those for physiological measures (Figure 1) in order to assess similarity across datasets.

RESULTS AND DISCUSSION

ICA filters and regression functions for each language are presented in Figure 2. Regression functions for physiological data from Figure 1 are included in each panel to facilitate visual comparisons of regression slopes and intercepts. Filters that optimally encode speech sounds in a wide variety of languages generally align with tuning properties in the mammalian auditory nerve. This supports Lewicki's (2002; Smith & Lewicki, 2006) claim that the auditory system efficiently codes speech, extending it to a wide variety of languages found worldwide.

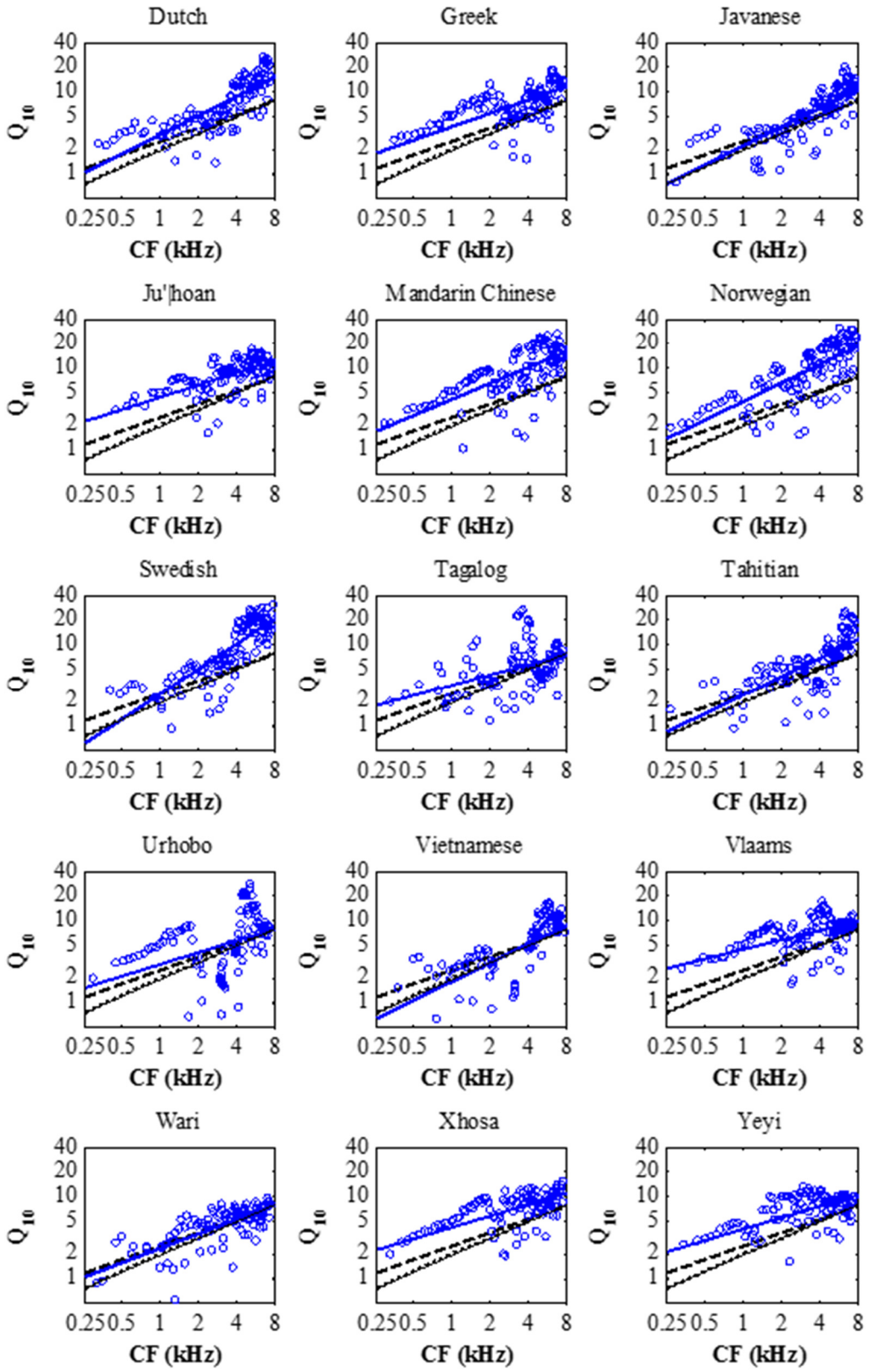

FIGURE 2: Statistically optimal filters for encoding a variety of natural languages. In each plot, each circle depicts the sharpness (Q_{10}) and center frequency of a single filter calculated by ICA.

Linear regression fit to ICA filters is depicted by the solid blue line; all regressions were statistically significant at $p < .001$. The black dashed line is the regression fit to physiological measures from Evans (1975) (see Figure 1, left); the black dotted line is the regression fit to physiological measures from Rhode & Smith (1985) (see Figure 1, right).

Regression slopes for ICA filters are highly comparable to those for auditory nerve fibers across all languages. However, several languages exhibit higher regression intercepts for ICA filters than those for physiological measures, especially languages where multiple talkers are sampled. This suggests that statistically optimal filters for encoding these stimuli may be slightly sharper than those observed in the cat auditory system. This is qualitatively consistent with recent studies suggesting that human cochlear tuning is sharper than previously considered. Shera and colleagues (2002; 2010) used stimulus-frequency otoacoustic emissions (SFOAEs) to measure cochlear tuning in guinea pigs, cats, chinchillas, and humans. They concluded that human cochlear tuning was significantly sharper than these animal models; results were corroborated by human behavioral data in a separate forward masking experiment (but see Ruggero & Temchin, 2005; Lopez-Poveda & Eustaquio-Martin, 2013). Joris and colleagues (2011) examined frequency tuning in macaque monkeys, which are phylogenetically closer to humans than to guinea pigs or cats. Auditory nerve fiber frequency-threshold tuning curves and SFOAEs revealed sharper tuning in macaques than other laboratory animals, but similarly sharp tuning as that estimated from human SFOAEs. It is important to note that these studies differ from the present investigation in terms of stimuli (narrowband tones vs. broadband speech), methodology (SFOAEs vs. ICA), and specific measures of filter sharpness (Q_{ERB} vs. Q_{10}). Further, there is ongoing debate regarding measures of auditory filter bandwidths using narrowband tones versus broadband stimuli (see de Cheveigné, 2008; Sayles & Winter, 2010 for discussions). Nevertheless, results shed light on how sharper tuning of the human auditory system is well-equipped to encode speech sounds in many different languages. Possible explanations of why some but not all languages demanded sharper tuning for optimal encoding (*i.e.*, exhibited higher regression intercepts) are discussed below.

Lewicki (2002) calculated statistically optimal filters for encoding roughly four minutes of American English (100 sentences x mean TIMIT sentence duration of approximately 2.5 seconds). The wide range of languages in the present investigation was expected to introduce substantial acoustic (and perhaps statistical) variability, so ten minutes of recordings were collected and analyzed in each case. Yet, several languages exhibit considerable variability among ICA filters (weaker correlations for Tagalog, Urhobo, and Vietnamese), and Swedish displays a notably steeper regression slope than any other language. These findings may be due to sampling a small number of unique talkers in these languages, in some cases only a single talker (Table I). Lewicki (2002) noted that analyzing speech from only one talker produced ICA filters that optimally code that talker's harmonic (formant) structure rather than general properties of speech sounds in that language. Further analyses are needed where each language contains speech from multiple talkers, not simply longer durations of a single or small number of talkers. It bears mention that languages with the most talkers (Mandarin Chinese with 87, Xhosa with 11) are well-fit by linear regressions and display higher regression intercepts than cat data. These results are expected to occur for other languages investigated here once a sufficient number of different talkers are sampled.

In all, results provide broad support for the efficient coding hypothesis (Barlow, 1961), as the auditory system has evolved to optimally encode the acoustic and statistical structure of speech sounds from a wide variety of languages.

ACNOWLEDGEMENTS
We thank Mark Sayles for valuable comments, and McKenzie Sexton, Lillian Slaughter, Asim Mohiuddin, Almina Klanco, Caitlyn Stromatt, Orlando Madriz, and Emily Nations for their assistance in stimulus collection.

REFERENCES

Amari, S., Cichocki, A., & Yang, H. (1996). A new learning algorithm for blind signal separation. In D. Touretzy, M. Mozer, & M. Hasselmo (Eds.), *Advances in Neural and Information Processing Systems*, *8* (pp. 757–763). San Mateo, CA: Morgan Kaufmann.

Barlow, H. B. (1961). Possible principles underlying the transformation of sensory messages. In W. A. Rosenbluth (Ed.), *Sensory Communication* (pp. 217–234). Cambridge, MA: MIT Press.

Bell, A.J., & Sejnowski, T.J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Computation, 7*(6)*,* 1129-1159.

de Cheveigné, A. (2008). Pitch perception. In C. J. Plack (Ed.), *Oxford Handbook of Auditory Science – Hearing* (pp. 71-104) Oxford: Oxford University Press.

Evans, E. F. (1975). Cochlear nerve and cochlear nucleus. in W.D. Keidel & W.D. Neff (Eds.), *Handbook of Sensory Physiology*, *5*(2), (pp.1-108). Berlin: Springer.

Field, D. J. (1987). Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America A*, *4*(12), 2379-2394.

Gazor, S., & Zhang, W. (2003). Speech probability distribution. *Signal Processing Letters, IEEE*, *10*(7), 204-207.

Geisler, W. S. (2008). Visual perception and the statistical properties of natural scenes. *Annual Review in Psychology*, *59*, 167-192.

Joris, P. X., Bergevin, C., Kalluri, R., Mc Laughlin, M., Michelet, P., van der Heijden, M., & Shera, C. A. (2011). Frequency selectivity in Old-World monkeys corroborates sharp cochlear tuning in humans. *Proceedings of the National Academy of Sciences*, *108*(42), 17516-17520.

Lewicki, M. S. (2002). Efficient coding of natural sounds. *Nature Neuroscience*, *5*(4), 356-363.

Lewis, J. W., Talkington, W. J., Tallaksen, K. C., & Frum, C. A. (2012). Auditory object salience: human cortical processing of non-biological action sounds and their acoustic signal attributes. *Frontiers in Systems Neuroscience*, *6*. doi: 10.3389/fnsys.2012.00027

Lopez-Poveda, E. A., & Eustaquio-Martin, A. (2013). On the controversy about the sharpness of human cochlear tuning. *Journal of the Association for Research in Otolaryngology*, *14*(5), 673-686.

Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, *381*(6583), 607-609.

Pearlmutter, B. A., & Parra, L. C. (1997). Maximum likelihood blind source separation: A context-sensitive generalization of ICA. In M. Mozer, M. Jordan, & T. Petsche (Eds.), *Advances in Neural and Information Processing Systems*, *9*. San Mateo, CA: Morgan Kaufmann.

Rhode, W. S. & Smith, P. H. (1985). Characteristics of tone-pip response patterns in relationship to spontaneous rate in cat auditory nerve fibers. *Hearing Research, 18*, 159–168.

Ruggero, M. A., & Temchin, A. N. (2005). Unexceptional sharpness of frequency tuning in the human cochlea. *Proceedings of the National Academy of Sciences*, *102*(51), 18614-18619.

Sayles, M., & Winter, I. M. (2010). Equivalent-rectangular bandwidth of single units in the anaesthetized guinea-pig ventral cochlear nucleus. *Hearing Research*, *262*(1), 26-33.

Shera, C. A., Guinan, J. J., & Oxenham, A. J. (2002). Revised estimates of human cochlear tuning from otoacoustic and behavioral measurements. *Proceedings of the National Academy of Sciences*, *99*(5), 3318-3323.

Shera, C. A., Guinan, J. J., & Oxenham, A. J. (2010). Otoacoustic estimation of cochlear tuning: validation in the chinchilla. *Journal of the Association for Research in Otolaryngology*, *11*(3), 343-365.

Smith, E.C., & Lewicki, M.S. (2006). Efficient auditory coding. *Nature*, *439*, 978-982.

Stilp, C. E., & Lewicki, M. S. (2014). Statistical structure of speech sound classes is congruent with cochlear nucleus response properties. *Proceedings of Meetings on Acoustics*, *20*(1), 50001.

Torralba, A., & Oliva, A. (2003). Statistics of natural image categories. *Network: Computation in Neural Systems*, *14*(3), 391-412.