A neurocomputational model for spectro-temporal phonetic abstraction

Dario J. Dematties[1]

[1]Instituto de Ingeniería Biomédica, Facultad de Ingeniería, Universidad de Buenos Aires, Ciudad Autonoma de Buenos Aires, Buenos Aires, Argentina

August 27, 2017

Executive Summary

Basic linguistics units -such as vowels, consonants, syllables, etc- are extracted and robustly classified by humans and other mammals from complex acoustic streams in speech data. Cortical structures -at different levels in the auditory pathway as well as at higher levels- respond selectively to phonetic features embeded in acoustic stimuli. In this research proposal we present a neurocomputational, completely unsupervised and biologically plausible model which establishes a new approach for deep feature extraction architectures in order to assist supervised phonetic classification techniques. The model we present here is entirely parallelized and scalable. In coordination with High Performance Computing (HPC) assistance from Argonne National Laboratory we will execute experimental tests in HPC facilities at Argonne. In this context, I will attend Message Passing Interface (MPI) Tutorials to be held at the Argonne Theory and Computing Sciences (TCS) Center, whose tutors belong to the group that invented and currently maintain MPI. Besides of the benefits this experience could bring to my project, this would also favors my future application as a candidate for the next Argonne Training Program on Extreme-Scale Computing (ATPESC) edition. We also plan to work with Assistant Computer Scientist at Argonne in data analysis and visualization to explore visualization techniques in order to inspect closely and thoroughly the evolution of training and testing stages in our systems. This project will receive mentoring from the Department of Neurobiology at University of Chicago from which we will receive state-of-the-art techniques in order to create complete and comprehensive maps of the brain. Our research will also receive mentoring in language processing from Departmental of Computing at Loyola University Chicago. The objective of our research is to develop novel deep phonetic feature extraction techniques based on relevant neurophysiological cortical mechanisms.

With the implementation of these new approaches we expect to perform at the level of state-of-the-art deep learning architectures. With these results we hope to direct the attention of new researches towards neurophysiological characteristics which present relevance for information processing in perception.

1 Introduction

It is well known that humans have the ability to discriminate phonemes as well as other linguistics units reliably, categorizing them, despite considerable variability across different speakers with different pitches, prosody, in noisy and reverberant environments. On the other hand, trained animals have also been shown to discriminate phoneme pairs categorically and to generalize to novel situations [23, 24, 26, 14, 29, 19, 40].

To understand how phonetic categories and word-like units are acquired, many computational theories have been developed. In the context of such theories, the main idea has been to try to explain relevant aspects but not to give details in order to show how the brain might provide such computations [36].

Lack of invariance phenomenon in speech perception [18], seems to be one of those scientific problems which cannot be solved by spontaneous human reasoning, given the immense amount of interrelated variables involved in phonetic categorization processes.

In that sense, deep learning architectures have shown unprecedented performance assisting conventional machine learning techniques which for decades required careful engineering in order to reach an effective feature extraction design [48].

On the other hand, artificial neural networks do not take into account remarkable biological aspects discovered during the last years in the area of neuroscience. Certain biological principles could be key in terms of information processing in the brain and they could provide us with matchless strategies in order to extract relevant information from a raw set of stimulus during perception.

The approach in this research is to gather potentially relevant biological aspects which could be significant in terms of information processing in the mammalian auditory cortex. We plan to test those principles with computational models which could perform in similar levels to state-of-the-art pattern classification techniques.

The aim in this work is not to replicate neuro-physiological mechanisms with precision nor to get a detailed reproduction of human cortical tissue. That is beyond the scope of our research. In contrast, we propose the parsimonious incorporation of cortical neurophysiological mechanisms letting the models speak by themselves. The biological mechanisms whose feature extraction properties show significant performance for invariant phonetic classification tasks, will be highlighted for future observation in upcoming researches.

In this work we seek novel computational solutions to automate feature extraction processes as a mean of assistance in spectro-temporal phonetic classification.

2 Plan of Work During the Stay in Argonne

We have implemented a neurocomputational model whose biological plausibility allows us to test neurophysiological hypotheses incorporated in the algorithms. With the desired number of layers, our model abstracts phonological features in a completely unsupervised fashion . The input is composed by a series of words which are procesed by another algorithm in order to feed the model. The phonetic features extracted by the model have the fucntion of inproving the performance of supervised pattern classification techniques whose main objective is to test the level of invariance achieved by the model's layers.

The objectives settled in this work impose two kind of challenges. First, the correct identification and pertinent selection of those neurophysiological mechanisms in the auditory pathway, and the way in which they must be incorpotrated in the algorithms. Regarding this issue, in the stay in Argonne National Laboratory (https://www.anl.gov/), this project will count with the research mentoring of Narayanan Kasthuriwith's lab from the Department of Neurobiology at University of Chicago. This group is currently facing one of the biggest challenges in the world creating a complete and comprehensive map of the mammalian brain. On the other hand, in order to achieve a correct algorithmic formulation and implementation, our project will also be mentored by George K. Thiruvathukaland from the Department of Computer Science at Loyola University Chicago. Professor Thiruvathukaland's specialty area includes parallel and distributed systems, software engineering, programming languages, operating systems. Dr. Thiruvathukal's early research involved object-oriented approaches to parallel programming and the development of object models with parallel programming, mostly based on C and C++on Unix platforms.

The second challenge comes from the fact that unsupervised training and testing phases for model instances of moderate to small sizes could take between 12 and 24 hours. Input computation can take up to 1 hour. Memory is a critical factor in such computation, even for very small corpora of 500 words. In the same manner, supervised training phase for Library for Support Vector Machine (LIBSVM) can take 2 or more hours. Memory and computational capability shortages obstruct our possibilities of producing the needed amount of tests over different model configurations in order to find correct parameter combinations. The size of the model instances as well as the experiments that we can execute are remarkably limited with our computational resources. During the stay in Argonne we will have access to high performance computing capabilities of unrivalled level in the world. Among the resources availability we will have access to supercomputers ranked at the 9th (Mira) and 16th (Theta) positions in the Top500 (https://www.top500.org/list/2017/06/?page=1). In this context, we will also count with the mentoring advice of Silvio Rizzi who is Assistant Computer Scientist at the Argonne Leadership Computing Facility. Dr. Rizzi will assist our project with HPC technical support. In this way, we will be able to test model configurations with dimensionalities and sizes which we have no access to in our current context. Emergent properties could arise from the enormous amount of parameter combinations in the design of those new instances. Surrounded by this framework, we plan to execute an ensemble of jobs, supervised by a genetic algorithm in which we will run several model instances in order to automate the process of parameter selection guided by the classification performance of the instances.

Beside the orders of magnitude in the number of advantages and variants for model parametrization and scalability, we have to take into account the new experimental possibilities offered by such working conditions. After getting the correct combination of parameters for the model configurations we will be able to test the model performance in standardized corpora like Acoustic-Phonetic Continuous Speech Corpus (TIMIT) (https://catalog.ldc.upenn.edu/ldc93s1), whose experimental tests are inviable with our resources.

Given the temporal requirements, we think we are in the best conditions to obtain considerable benefits from the time spent at Argonne. We count on a model implementation completely parallelized. The parallelization in OpenMP API does not affect the original sequential implementation of the code. This parallelization approach allows its automatic scalability, even a completely serialized compilation is possible if it were necessary.

Such applications will return feature vectors with orders of magnitude larger than with our current tests. In this respect, we plan to pursue extensions to the LIBSVM soft -as Multi-core Support Vector Machine (LIBLINEAR)- which can be implemented in shared-memory systems to reduce the training time dramatically.

The software used to manipulate all the data processed and produced by the model is implemented under GNU Octave (https://www.gnu.org/software/octave/). For cases of vector with high dimensionality -as the cited above- we will have to parallelize such code. GNU Octave offers packages which enable parallelization in shared memory systems.

The significance of visual observation in scientific data becomes apparent when we enumerate the overwhelming amount of examples in this respect. Besides all the evidence in favour of this policy, we will cite a compelling example from fractional calculus notation [34]. From this example it should be clear that, as visual animals, visual observation brings us unparalleled advantages.

“The choice of a precise notation for the fractional calculus cannot be minimized. For as we shall see, some of the power and elegance of the fractional calculus rests in its simplified notation. The abridged manner of representing these defining integrals may seem to be a trivial matter; but the advantage of a simple notation has been the source of many profound discoveries not obvious by other means.”

The way in which we visually represent our data is crucial for the success of our project. Regarding this aspect, we will be assisted by Argonne Assistant Computer Scientists in order to get the best options for the visual representation of our data. We plan to visualize the training and processing evolution of hundred of thousand of units with millions of connections through the use of large tiled displays. Figure 1 shows a large-scale visualization of the connectomes. The image technique is based on X-Ray extended tomography (or Mosaic Tomography) with 1 micron resolution done at the beamline 32-ID-C on the Advanced Photon Source. The segmentation is tensor flow based in order to extract features like cell bodies, myelinated axons and blood vessels. The image resolution is obtained through multiple images acquired and stitched together. Image on Figure 1 comes from downsampled segmented dataset at 2Kx1Kx2K. Full resolution is about 20Kx20K pixels per slice.

```
045885  avgFPS: 30.5647954638
BMTool: 0 app timestamp: 1493316894.8 cachedFPS: 30.305
705962  avgFPS: 30.5487278879
BMTool: 0 app timestamp: 1493316894.8 cachedFPS: 30.304
998909  avgFPS: 30.5379775963
BMTool: 0 app timestamp: 1493316894.9 cachedFPS: 30.291
674384  avgFPS: 30.576925353
BMTool: 0 app timestamp: 1493316894.9 cachedFPS: 30.291
384646  avgFPS: 30.5767913168
BMTool: 0 app timestamp: 1493316894.9 cachedFPS: 30.28
159317  avgFPS: 30.564210257
```

Figure 1: Large-Scale Computing and Visualization on the Connectomes of the Brain.

3 Expected Outcomes and Benefits

We are convinced that, in order to get truly general artificial intelligent agents, we need to reverse-engineer -at least- those physiological brain mechanisms of relevance for information processing in perception. The objectives sought in the proposed research determines unexplored challenges in the automatic deep feature extraction world. We seek unprecedented techniques to get phonetic classification performance improvements through the use of neurophysiological data that has not been used in current deep learning technologies.

Invariance is the holy grail of pattern classification, and tiny improvements -even in some small fractions of percentage- are highly appreciated in this area. We are seeking a completely new approach in the world of artificial phonetic perception, but besides that, the contributions that we could obtain from our experimental results could stablish new directions in terms of the relevance that scientists impute to certain physiological structures for information processing in the brain. Furthermore, the relevance of our future discoveries could cross the frontiers of auditory perception and make influence on other modalities as visual and somatosensory.

Having the possibility to work with professor Kasthuriwith from the Department of Neurobiology, professor Thiruvathukal from the Loyola University of Chicago and Dr. Silvio Rizzi from Argonne National Laboratory and University of Chicago, and given the favorable relationships available through the INTERNATIONAL RESEARCH COOPERATION BETWEEN THE MINISTRY OF SCIENCE, TECHNOLOGY AND PRODUCTIVE INNOVATION OF THE ARGENTINE REPUBLIC AND THE UNIVERSITY OF CHICAGO, represents incomparable prospects for the evolution of this work and for my personal career, today and for lasting scientific cooperation relationships in the future.

4 Previous Research

Among the computational theories developed to understand human phonetic acquisition, some models, bypass the initial speech signal processing and instead of dealing with the complexity and variability of real speech at the prelexical level, they use an artificial, often hand-crafted, idealized discrete (prelexical) representation of the acoustic signal as input to the lexical level [37]. In other works [12], although some biological observations are made, the input components are syllable representations from specific corpora.

In the works of de Boer and Kuhl [11] and Vallabha, McLelland, Pons, Werker and Amano [15], the models classify some vowels through statistical mechanisms which take into account formant components and Vowel Length (VL). In Toscano and McMurray [8], statistical methods are used to classify consonantal phonetic characteristics, by means of Voice Onset Time (VOT), VL, pitch and first formant onset frequency.

The statistical methods used in these works interrelate different features extracted from acoustic speech signals. Those features are carefully considered and evaluated through human engineering and considerable domain expertise which evaluate their relevance in order to include them in the computations. Some features, as VL and VOT, make reference to highly abstract dynamic characteristics which are taken as available parameters without any previous natural processing.

We concede the possibility of the existence of other features which can scape human expertise. Furthermore, some hidden features could be constituent part of those abstract features evaluated by humans.

In this respect, deep learning approaches have gained significant interest as a way of building hierarchical representations from unlabeled data. Convolutional deep belief networks have been applied to audio data and empirically such architectures have been evaluated on various audio classification tasks. In the case of speech data, it was shown that the learned features corresponded to phones/phonemes [17].

In Kouki et al. [22], the use of Mel Frequency Cepstral Coefficients (MFCC) strategy presupposes a more biologically accurate input stream, though the cepstrality in such coefficients does not reflect -under our point of view- the responsive air cells in front of cochlear vibration. In a posterior work, Kouki et al. [30], designed a method to separate "stable" and "dynamic" speech patterns.

All above-cited works lack biological plausibility. Regarding this point, in the last few years a compelling theoretical framework has been developed. In this theory, plausible hypotheses about the role of the neocortex in the mammalian brain are given in an approach called Memory-Prediction Framework (MPF) [21]. This approach is based on evidence which supports the idea that there are fundamental mechanisms which underly a common neocortical structure and its connectivity. In a recent work it was shown that a neuron with several thousand synapses could recognize hundreds of independent patterns of cellular activity even in the presence of large amounts of noise and pattern variation. A neuron model was proposed in which by means of the combination of proximal and distal dendrites it could predict its activation in hundreds of independent contexts. Through simulation procedures, it was shown a network which scaled well and operated robustly over a wide range of parameters as long as this used a sparse distributed code of cellular activations. It was concluded that pyramidal neurons with thousands of synapses, active dendrites, and multiple integration zones created a robust and powerful sequence memory [20].

We propose to take an approach similar to the one taken in [20]. We support the idea that in order to design truly powerful machine learning techniques, it is necessary to gather those biological characteristics which are relevant -regarding information processing- as to get highly robust invariant pattern representation capabilities.

5 Theoretical Framework and Hypotheses to be Tested

Cortical cells are aligned into restricted domains for common receptive field locations, which represent different sensory modalities and are composed by neural cells of identical salient physiological characteristics. V. Mountcastle proposed such structures as elementary units for structural organization in the somatic cortex and called them cortical columns [27, 45]. The first confirmatory researches for this phenomenon came from Hubel and Wiesel's discoveries [9, 10]. Margins in column diameter are between 300 and 600 μm; even among different species whose brains differ in volume by a factor of 10^3. The evolutionary cortical brain expansion is achieved through the expansion in cortical surface area by means of an increase in the number of cortical columns and not by the increase in individual column size [39]. These facts suggest an uniform and modular structure in cortical tissue organization. In accordance, Mountcastle suggested in 1978 that there could be a unique cortical algorithm replicated through all the neocortex [7].

Linden and Schreiner proposed that although auditory cortical circuits have some unique characteristics, their similarities with other sensory regions -such as visual or somatosensory cortex- seem to be much more categorical [13]. They proposed a series of analogies. First, at the sensory level, the cochear one-dimensional frequency map could be analogue to the two-dimensional spatial maps which are found in the retina or body surface. Second, the tonotopic maps found in the auditory system could be analogue to the retinotopic and somatotopic organization found in visual and somatosensory cortices, respectively. Frequency tuning curves in the auditory system could correspond to inhibition of spatial surrounding boundaries in visual and somatosensory receptive fields. A correspondence could be drawn between amplitude modulation rate in the auditory system and flicker sensitivity in the visual system, or whisker vibration sensitivity in the somatosensory system. Finally, auditory receptive fields tuned for frequency-sweep, could be analogous to visual and somatosensory motion sensitivity.

Primary Auditory Cortex (A1) shares common structural characteristics with other sensory cortices [1, 3, 46, 2, 35]. Thalamo-cortical circuits rewired to receive visual signals in live ferret auditory cortex, show how this structure can support thalamo-cortical and intracolumnar transformations seen in other modalities. When retinal inputs are routed into the auditory thalamus, auditory cortical cells develop visual response properties such as direction selectivity, orientation preference and complex and simple receptive fields [47, 28, 32]. Retinotopic maps, in terms of orientation tuning with lateral connectivity between orientation domains, emerge in superficial layers of the rewired auditory cortex [31, 33]. These data suggest the existence of neuronal circuitry with similar processing capabilities for different modalities.

In the context of perceptual capabilities in the auditory pathway, neuronal responses to continuous speech in A1 of naive ferrets, revealed the existence of a spectro-temporal tuning in that area with the capacity of supporting discrimination of many American English phonemes [5], even when stimuli were distorted by additive noise and reverberance [6].

In this proposal, we present a computational theory which incorporates relevant characteristics present in cortical pathways commonly found in mammalian microcircuits. The theories behind the model's algorithms conceive complex auditory linguistic stimuli as signals with an intrinsic dynamic statistical structure.

The implementation consists of three main sections. First, we process the sound waves with an algorithm that takes some guidelines from the technique elaborated by Chi T. et al. [38]. In this work, accumulating experimental findings from the central auditory system were exploited demonstrating its applications in the objective evaluation of speech intelligibility. As the authors pointed out, the model was not biophysical in spirit, but rather it abstracted from the physiological data an interpretation that was likely to be relevant in the design of sound engineering systems.

For the following section, we implemented a structure called encoder. The function of the encoder is to transduce a multidimensional array of real numbers into a multidimensional sparse and distributed representation of active cells. Empirical evidence suggests the neocortex represents information using sparse distributed patterns of activity [4]. The encoder is composed of a set of Self Organizing Maps (SOMs) [43, 25] and incorporates neurophysiological phenomena as columnar organization, proximal and distal dendritic arborization, afferent, apical and lateral intercolumn interaction, proximal lateral intracolumn inhibition, Long-Term Potentiation (LTP), Long-Term Depression (LTD) and Spike-timing dependent plasticity (STDP).

The last section is called regular layer, this structure has the capacity of processing afferent Sparse Distributed Representations (SDRs) and in addition to the neurophysiological mechanisms present in the encoder layer, the regular layer also incorporates activation and synaptic homeostatic regulations. As it happens in cortical tissue, neural circuits in our implementation must maintain stable function in which incoming information can be distributed across all the units in the structure. Recent work has shown that these destabilizing influences present in neural tissue are counterbalanced by homeostatic plasticity mechanisms that act to stabilize neuronal and circuit activity [16].

In other works, it has been discussed that the activation of several distal synapses can lead to a local dendritic N-Methyl-D-aspartic acid (NMDA) spike and consequently a significant and sustained depolarization of the soma. Novel computational theories have drawn a possible explanation about the role of distal synapses in relation with NMDA phenomenon [20] combining it with SDRs [41]. In this approach, dendrite branches are taken as active processing elements. We also incorporate these properties in our computational models.

It has been shown that overfitting can be greatly reduced with stochastic properties in training procedures applied of neural networks (dropout) [42]. In this regard, we have incorporated stochastic characteristics to the encoder and regular layers in the training stage. Hence, the evolution of the network during training does not determine a neuron to fire but bias its probability of doing that. Additionally, afferent dendritic arborizations in the encoder layer receive random information whose boundary values are established by learning.

6 Materials and Methods

The algorithms in this work have been implemented under the standard C++11 through a set of classes related by inheritance and composition. The scalability of the classes allows every cortical layer and column to be generated with the desired dimensionality and number of units. Connectivity among cortical columns as well as cortical layers are randomly auto-generated with the guide of arguments passed to the object constructors. In order to handle the data produced by the model, a library has been implemented to save the data in Octave/Matlab compatible (.mat) file formats. Every class in the implementation has been parallelized by means of the OpenMP API.

The inputs to feed the model are computed with spectral analysis via the FFTW library (http://www.fftw.org/). The algorithms are based on Mel filterbanks and multi-resolution spectrotemporal analysis of complex sounds [44]. These algorithms have been parallelized with the OpenMP API.

Corpora audio files are generated via Festival Text to Speech
(http://www.cstr.ed.ac.uk/projects/festival/).

Classification performance of the cortical features extracted by the model are tested with Support Vector Machine (SVM) techniques using the LIBSVM library (https://www.csie.ntu.edu.tw/ cjlin/libsvm/).

We have implemented an instance of the model with a bidimensional input of 128 by 5 components and an encoder layer with 81 cortical columns of 225 neurons each. This instance presented proximal afferent connections from the input and distal lateral connections from neighboring columns. In

Table 1: Classification Performance

Input	Encoder	Procedure
98%	97.2%	Training/Cross-Validation
27,4% 137/500	58% 290/500	Testing/Noise 0.02
62% 310/500	69,6% 348/500	Testing/Reverberation 30%
56.6% 283/500	57.8% 289/500	Testing/Pitch +30

order to feed the model, we generated a corpus with 500 words from a vocabulary of 5 words uttered by 10 different speakers (8 males and 2 females) available from the synthesizer. The organization of the corpus had certain rules and restrictions. The speakers were sequentially chosen at random with the restriction that no speaker could utter a second time until all the speakers had uttered in their turns. Every speaker uttered two words per turn and every word uttered by a speaker could not be repeated until all the words were used by such speaker.

Once the corpus was generated, the encoder layer was trained with a complex procedure in which certain parameters -as learning rate and neighborhood interaction- were reduced progressively as training progressed. The model then processed the original corpus, the corpus affected by noise, reverberation and pitch variations. All the variations applied to the corpora were grenerated with Audacity® free, open source, cross-platform audio software for multi-track recording and editing (http://www.audacityteam.org/home/). The SVM classifier was trained and tested with cross-validation with the output of the model in response to the original corpus. The boundaries of the words were marked and the values between the marks were accumulated in order to compose condensed vector with which train the classifier. Then, the vectors were scaled -as the LIBSVM documentation suggest- in order to improve the classification performance. LIBSVM was configured to use a linear kernel with one parameter C which was swept in order to find the best trained model for the classifier. Then, the classifier tested the invariance in the responses of the model to the altered corpora. Table 1 shows the performances.

References

[1] Huang C. L., Winer J. A. Auditory thalamocortical projections in the cat: laminar and areal patterns of input. *J. Comp. Neurol.*, 427:302–331, 2000.

[2] Mitani A. and Shimokouchi M. Neuronal connections in the primary auditor y cortex: an electrophysiological study in the cat. *J. Comp. Neurol.*, 235:417–429, 1985.

[3] Winer J. A. The functional architecture of the medial geniculate body and the primary auditory cortex, in: The mammalian auditory pathway: neuroanatomy. *New York: Springer-Verlag.*, pages 222–409, 1992.

[4] Barth A.L. and Poulet J.F. Experimental evidence for sparse firing in the neocortex. *Trends Neurosci.*, 35:345–55, 2012.

[5] Mesgarani N., David S. V., Fritz J. B., and Shamma S. A. Phoneme representation and classification in primary auditory cortex. *J. Acoust. Soc. Am.*, 123:899–909, 2008.

[6] Mesgarani N., David S. V., Fritz J. B., and Shamma S. A. Mechanisms of noise robust representation of speech in primary auditory cortex. *PNAS.*, 123:899–909, 2014.

[7] Mountcastle V. B. An organizing principle for cerebral function: The unit model and the distributed system. *Cambridge, MA.*, 1978.

[8] Toscano J. C. and McMurray B. Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics. *Cognitive Scienc.*, 34:434–464, 2010.

[9] Hubel D. and Wiesel T. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J Physiol.*, 160:106–154, 1962.

[10] Hubel D. and Wiesel T. Receptive fields and functional architecture of monkey striate cortex. *J Physiol.*, 195:215–243, 1968.

[11] de Boer B., and Kuhl P. Investigating the role of infant-directed speech with a computer model. *Acoustics Research Letters Online.*, 4:129–134, 2003.

[12] Dominey P. F. and Ramus F. Neural network processing of natural language: I. sensitivity to serial, temporal and abstract structure of language in the infant. *Language and Cognitive Processes.*, 15:87–127, 2000.

[13] Linden J. F. and Schreiner C. E. Columnar transformations in auditorycortex? a comparison to visual and somatosensory cortices. *Cerebral Cortex*, 13:83–89, 2003.

[14] Pons F. The effects of distributional learning on rats sensitivity to phonetic information. *J. Exp. Psychol. Anim. Behav.*, 32:97–101, 2006.

[15] Vallabha G. K., McLelland J. L., Pons F., Werker J. F. and Amano S. Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of National Academy of Sciences.*, 104:13273–13278, 2007.

[16] Turrigiano G. Homeostatic synaptic plasticity: Local and global mechanisms for stabilizing neuronal function. *Cold Spring Harb Perspect Biol.*, 4, 2012.

[17] Lee, Honglak, Largman, Yan, Pham, Peter, and Ng, Andrew Y. Unsupervised feature learning for audio classification using convolutional deep belief networks. In *Proceedings of the 22Nd International Conference on Neural Information Processing Systems*, NIPS'09, pages 1096–1104, USA, 2009. Curran Associates Inc.

[18] Appelbaum I. The lack of invariance problem and the goal of speech perception. *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, 3, 1996.

[19] Dent M. L., Brittan-Powell E. F., Dooling R. J., and Pierce A. Perception of synthetic /ba/-/wa/ speech continuum by budgerigars (melopsittacus undulatus). *J. Acoust. Soc. Am.*, 102:1891–1897, 1997.

[20] Hawkins J. and Ahmad S. Why neurons have thousands of synapses, a theory of sequence memory in neocortex. *Frontiers in Neural Circuits.*, 10, 2016.

[21] Hawkins J. and Blakeslee S. On intelligence. *Times Books.*, 2004.

[22] Kouki M., Hideaki K. and Reiko M. Unsupervised learning of vowels from continuous speech based on self-organized phoneme acquisition model. *Interspeech.*, 2010.

[23] Kuhl P. K. and Miller J. D. Speech perception by the chinchilla: Voiced voiceless distinction in alveolar plosive consonants. *Science.*, 190:69–72, 1975.

[24] Kuhl P. K. and Padden D. M. Enhanced discriminability at the phonetic boundaries for the place feature in macaques. *J. Acoust. Soc. Am.*, 73:1003–1010, 1983.

[25] T. Kohonen. *Self-organization and Associative Memory: 3rd Edition.* Springer-Verlag New York, Inc., New York, NY, USA, 1989.

[26] Kluender K. R., Lotto A. J., Holt L. L., Bloedel S. L. Role of experience for language-specific functional mappings of vowel sounds. *J. Acoust. Soc. Am.*, 104:3568–3582, 1998.

[27] Mountcastle V. B., Berman A. L. and Davies P. W. Topographic organization and modality representation in first somatic area of cat's cerebral cortex by method of single unit analysis. *Am. J. Physiol.*, 183, 1955.

[28] Angelucci A., Clasca F., Sur M. Brainstem inputs to the ferret medial geniculate nucleus and the effect of early deafferentation on novel retinal projections to the auditory thalamus. *J. Comp. Neurol.*, 400:417–439, 1998.

[29] Hienz R. D., Aleszczyk C. M., and May B. J. Vowel discrimination in cats: Acquisition, effects of stimulus level, and performance in noise. *J. Acoust. Soc. Am.*, 99:3656–3668, 1996.

[30] Kouki M., Hideaki M., Hideaki K., Reiko M. The multi timescale phoneme acquisition model of the self-organizing based on the dynamic features. *Interspeech.*, 2011.

[31] Roe A. W., Pallas S. L., Hahm J. O., Sur M. A map of visual space induced in primary auditory cortex. *Science.*, 250:818–820, 1990.

[32] Roe A. W., Pallas S. L., Kwon Y. H., Sur M. Visual projections routed to the auditory pathway in ferrets: receptive fields of visual neurons in primary auditory cortex. *J. Neurosci.*, 12:3651–3664, 1992.

[33] Sharma J., Angelucci A., Sur M. Induction of visual orientation modules in auditory cortex. *Nature.*, 404:841–847, 2000.

[34] Kenneth S. Miller and Bertram Ross. *An Introduction to the Fractional Calculus and Fractional Differential Equations.* A Wiley-Interscience Publication, Printed in the United States of America, first edition, 1993.

[35] Mitani A., Shimokouchi M., Itoh K., Nomura S., Kudo M., Mizuno N. Morphology and laminar organization of electrophysiologically identified neurons in the primary auditory cortex in the cat. *J. Comp. Neurol.*, 235:430–447, 1985.

[36] Räsänen O. Computational modeling of phonetic and lexical learning in early language acquisition: Existing models and future directions. *Speech Communication.*, 54:975–997, 2012.

[37] Scharenborg O. and Boves L. Computational modelling of spoken-word recognition processes. *John Benjamins Publishing Company.*, 2010.

[38] Chi T., Ru P. and Shamma S.A. Multiresolution spectrotemporal analysis of complex sounds. *J. Acoust. Soc. Am.*, 118:887–906, 2005.

[39] Rakic P. Radial versus tangential migration of neuronal clones in the developing cerebral cortex. *Proc Natl Acad Sci USA.*, 92, 1995.

[40] Lotto A. J., Kluender K. R., and Holt L. L. Perceptual compensation for coarticulation by japanese quail (coturnix coturnix japonica). *J. Acoust. Soc. Am.*, 102:1134–1140, 1997.

[41] Ahmad, S., and Hawkins, J. How do neurons operate on sparse distributed representations? a mathematical theory of sparsity, neurons and active dendrites. *arXiv:1601.00720 [q–bio.NC]*, 2016.

[42] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958, 2014.

[43] Kohonen T. Self-organized formation of topologically correct feature maps. *Biological Cybernetics.*, 43:59–69, 1982.

[44] Powen Ru Taishih Chi and Shihab A. Shamma. Multiresulution spectrotemporal analysis of complex sounds. *J. Acoust. Soc. Am.*, 118:887–906, 2005.

[45] Mountcastle V. Modality and topographic properties of cat's somatic sensory cortex. *J. Neurophysiol.*, 20:408–434, 1957.

[46] Rockel A. J., Hiorns R. W. and Powell T. P. The basic uniformity in the structure of the neocortex. *Brain*, 103:221–244, 1980.

[47] Sur M., Garraghty P. E., Roe A. W. Experimentally induced visual projections into auditory thalamus and cortex. *Science.*, 242:1437–1441, 1988.

[48] LeCun Y., Bengio Y. and Hinton G. Deep learning. *Nature*, 521:436–444, 2015.