SELF-ORGANIZATION OF SPEECH SOUND INVENTORIES

IN THE FRAMEWORK OF COMPLEX NETWORKS

Synopsis of the Thesis to be submitted
in Partial Fulfillment of the Requirements
for the Award of the Degree of

Doctor of Philosophy

by

Animesh Mukherjee

Under the supervision of

Dr. Niloy Ganguly
and
Prof. Anupam Basu

Department of Computer Science and Engineering

Indian Institute of Technology, Kharagpur

July 2009

Synopsis

Curiosity about the world as well as ourselves is one of the most important traits of the human race. It is perhaps in order to relinquish this curiosity about each other that thousands of years ago our forefathers invented an extremely sophisticated medium of communication – *language*. Linguistic ability not only makes us different from the rest of the animal kingdom but is also central to the sense of identity that unites nations, cultures and ethnic groups. The same curiosity about ourselves gives birth to numerous important as well as interesting questions about this ability itself. Some of the most intriguing ones among these are "How did language evolve?", "Why do languages change over time?", "What are the universal characteristics of the thousands of mutually unintelligible languages that co-exist at a given time?" and "How does a child almost effortlessly acquire all the intricacies of language in the very early stages of development?" Various disciplines have joined in the search for the answers to the above questions in a collaborative and systematic approach. As a consequence of this collective effort, one argument that has gained enormous popularity in recent times is that language is a *complex adaptive system*, which has evolved through the process of self-organization in order to serve the purpose of human communication needs [25]. In fact, the main premise of *synergetic linguistics* [14] is that language is a self-organizing and self-regulating system and its (a) existence, (b) properties, and (c) change can be successfully explained within this framework. The symmetries observed across languages are therefore, primarily an outcome of the dynamic interdependence of the structure and the functions of a language [14]. The aforementioned efforts have repeatedly pointed to the fact that the emergent complexity of a linguistic phenomenon can be understood by treating language as a physical system.

Like any physical system, a linguistic system (i.e., language) can be viewed from three different perspectives [3]. At one extreme, it is a collection of *utterances* that are produced and perceived by the speakers of a linguistic community during the process of communication with the other speakers of that community. This is analogous to the *microscopic* view of a thermodynamic system, where every utterance and its corresponding context together render the identity of the language, that is, its grammar. At the other extreme, a language can be described by a lexicon and a set of grammar rules. This is equivalent to the *macroscopic* view of a thermodynamic system. Sandwiched between these two extremes, one can also conceive of a *mesoscopic* view of language, where the different linguistic entities such as phonemes, syllables, words or phrases form the basic units of the system and the grammar is an emergent property resulting from the complex interactions among these units.

In the recent years, *complex networks* have proved to be an extremely suitable framework for modeling the structure and dynamics of various large-scale systems primarily at the level of mesoscopy. Examples of well-studied naturally-occurring networks include biological, ecological and social networks such as metabolic networks [18], gene regulatory networks [6], protein interaction networks [29], food webs [28], scientific collaboration networks [19] and movie-actor networks [21]. Similarly, there have also been a lot of studies on man-made networks, which mainly include communication networks and transportation infrastructures such as the Internet [10], WWW [1], p2p networks [2, 17], airlines [11] and railway networks [22].

Since human language is one of the most appropriate examples of a complex system, principles of network theory have proved to be extremely suitable for modeling as well as explaining the emergence of various intricate linguistic phenomena. In fact, it is due to this reason that within a very short period of time the study of linguistic networks, in order to understand the structure and the evolutionary dynamics of language, has gained a lot of momentum (see [5,12,23,24] for references). The primary motivation for the computational methodology adopted in this work is the burgeoning success of the aforementioned studies that not only investigate but also substantiate numerous linguistic properties within the framework of complex networks. More precisely, we show how this computational framework helps us in addressing one of the most important problems in phonology that involves modeling and explaining the structure, dynamics and emergence of human speech sound inventories across the languages of the world.

In the next section, we present examples of some popularly studied linguistic networks, which have been the principal source of inspiration for the work presented here (section 1). This is followed by a brief history of the problem addressed in the dissertation (section 2). Section 3 outlines the main objectives of the thesis and section 4 summarizes the salient contributions of the work. The organization of the thesis is presented in section 5.

1 Linguistic Networks

The study of linguistic networks at the level of mesoscopy can be broadly classified into three different categories based on the purpose of construction of these networks. These categories are

(i) *Lexical networks* that are constructed to explore the organization of the "mental lexicon" (i.e., the repository of word forms, which are assumed to reside in the human brain). The inherent complexity of the problem has motivated a lot of researchers to investigate it in the framework of complex systems and more specifically, complex networks where each node in the network corresponds to a word form and the interconnections are based on certain types of similarities between pairs of word forms. Analysis of such networks reveal various interesting properties about the structure and evolution of the mental lexicon (see [12, 23] for references).

(ii) *Word co-occurrence networks* where the nodes represent words and two nodes are connected by an edge if they co-occur in a language in certain context(s). Most of the studies here attempt to explore the evolution of the syntactic structure of a linguistic system through the analysis of various statistical properties of these networks [5, 9].

(iii) *Phonological networks* that are built from different sub-lexical units (e.g., syllables and phonemes) so as to determine the universal properties of the sound structure of linguistic systems (see [24] for a reference).

2 The Problem of Consonant Inventories

The most basic units of human languages are the speech sounds. The repertoire of sounds that make up the sound inventory of a language are not chosen arbitrarily, even though the speakers are capable of perceiving and producing a plethora of them. In contrast, the inventories show exceptionally regular patterns across the languages of the world, which is arguably an outcome of the self-organization that goes on in shaping their structure [20]. Earlier researchers have proposed various functional principles in order to explain this self-organizing behavior of the sound inventories. The most important among these are as follows.

(i) *Maximal perceptual contrast* [16], which implies that the phonemes as well as the other linguistic units (e.g., syllables, words) of a language should be maximally distinct from each other, because this facilitates proper perception of the individual linguistic units in a noisy environment.

(ii) *Ease of articulation* [8,16], which states that the structure of a language should facilitate expression and dissemination of information at the expense of minimal energy spent on the part of the speaker. Some of the general implications of this principle are: frequent words are shorter in length; the sound systems of all languages are formed of certain universal (and highly frequent) sounds that do not use complicated articulatory gestures, etc.

(iii) *Ease of learnability* [8], which states that a language should be easily learnable in order to propagate through the generations. Consequences of this principle include facts that linguistic structures are mostly regular and irregularities, if any, are observed for only extremely frequent linguistic units (e.g., some very frequent verbs in English are irregular).

These principles are applied to language as a whole, thereby, viewing it from the macroscopic level. In fact, the organization of the vowel inventories across languages has been quite satisfactorily explained in terms of the single principle of maximal perceptual contrast through linguistic arguments [27], numerical simulations [15] as well as genetic algorithms [13]. With the advent of highly powerful computers, it has also been possible to model the micro-level dynamics involving a group of (robotic) speakers and their interactions and this in turn has proved to be highly successful in explaining how the vowel inventories originated and self-organized themselves over the linguistic generations [8].

Right from the beginning of the 20^{th} century, there have been many linguistically motivated attempts [4,7,26] in order to explain the emergence of the regularities that are observed across the consonant inventories. However, unlike the case of vowel inventories, majority of these efforts are limited to the investigation of certain specific properties primarily because of the inherent complexity of the problem. The complexity arises from the fact that (a) consonant inventories are usually much larger in size and are characterized by much more articulatory/acoustic features than the vowel inventories, and (b) no single force is sufficient to explain the organization of these inventories; rather a complex interplay of forces collectively shape their structure. Thus, a versatile modeling methodology, which is hitherto absent in the literature, is required so that the problem can be viewed and solved from an alternative perspective.

3 Objectives

The primary objective of the thesis is to develop a computational framework for simulating the structure and dynamics of the consonant inventories of the world's languages. More specifically, we model the self-organization of these inventories through a complex network approach. Some of the typical questions that we would like to answer in the course of the thesis are as follows.

(i) *Representation of the Inventories*: The first question that one needs to answer is how can the structure of the consonant inventories be accurately represented within the framework of complex networks. This is indeed a very important problem, because all the results obtained as well as the predictions made can be heavily influenced by the underlying scheme of representation.

(ii) *Analysis of the Inventory Structure*: Once a suitable representation scheme is chosen, the next crucial question is how to conduct the analysis in order to extract meaningful results. In particular, one needs to answer (a) which statistical properties of the network(s) should be studied in order to discover the different cross-linguistic patterns that manifest across the consonant inventories, (b) what are the basic principles that could be responsible for the formation of these patterns, and (c) how can these principles be systematically quantified in order to figure out the extent to which they drive the origins of these patterns.

(iii) *Synthesis of the Inventory Structure*: A third and an equally important problem is to explain the emergence of the different statistical properties (obtained from the analysis) by means of generative mechanisms that are usually based on various models of network growth. The typical questions that one needs to answer here are (a) what can be a suitable synthesis model for explaining the statistical properties of the network, (b) how can such models be analytically solved to have a better understanding of the dynamics, (c) what are the linguistic correlates of each of these models with reference to the consonant inventories, and (d) what is the physical significance of the parameters involved (if any) in each of these models.

Although the thrust of this work is on consonant inventories, we also aim to investigate certain well-known properties of the vowel inventories within the same computational framework. The objective of this is twofold – (a) to show that the formalism proposed here is generic and is useful in studying the evolution and emergence of human speech sound inventories, and (b) to report interesting new observations about the vowel inventories apart from validating the results presented by the earlier researchers.

4 Contributions

In this work, we show how the structure of the consonant inventories can be represented, analyzed as well as synthesized within the framework of complex networks. For this purpose, we construct two networks, one of which is based on the occurrence of consonants across languages while the other is based on co-occurrence of the consonants across languages. A brief report on the studies of these two networks and the results obtained thereby, are presented below.

Occurrence Network of Consonants

We represent the inventories as a bipartite network in which one of the partitions consists of nodes corresponding to the languages while the other partition consists of nodes corresponding to the consonants. There is an edge between the nodes of these two partitions if a particular consonant occurs in a particular language. An exhaustive study of this network reveals various interesting results as follows.

(i) The size of the consonant inventories (indicated by the distribution of the degrees of the language nodes) follow a β-distribution.

(ii) The distribution of occurrence of the consonants over languages (i.e., the degree distribution of the consonant nodes in the network) follow a well-behaved probability distribution.

(iii) A synthesis model based on preferential attachment (i.e., a language node attaches itself to a consonant node depending on the current degree (k) of the consonant node) coupled with a tunable randomness component can explain the emergence of the degree distribution of the consonant nodes.

(iv) The emergent degree distribution obtained from the synthesis model can be analytically shown to approach a β-distribution in the asymptotic limits.

Co-occurrence Network of Consonants

After studying the properties of occurrence of consonants, the next apparent step is to investigate their co-occurrence properties. For this purpose, we construct a network in which the nodes are the consonants and an edge between two nodes (read consonants) signifies their co-occurrence likelihood across languages. Some of the important findings from this study are summarized below.

(i) The co-occurrence distribution of the consonants across languages (i.e., the degree distribution of the consonant nodes in the co-occurrence network) is again found to follow a well-behaved probability distribution.

(ii) The clustering coefficient of the co-occurrence network is very high, a property commonly observed in social networks [19, 21] that is indicative of the presence of a large number of densely connected neighborhoods (formed by groups of consonants).

(iii) Community structure analysis of this network reveals strong patterns of co-occurrence of consonants that are prevalent across the languages of the world.

(iv) Languages exhibit an economic behavior by using a small number of articulatory/acoustic features and maximizing the combinatorial possibilities of these features in order to generate a large number of consonants. This behavior, often termed as *feature economy* [4, 7], leads to the formation of the consonant communities. An information theoretic quantification of this principle further shows the extent to which it is responsible for the community formation.

(v) The emergent degree distribution of the co-occurrence network can be shown to be sensitive to the distribution of the consonant inventory sizes even though the degree distribution of the occurrence network does not depend on the same.

(vi) The clustering coefficient of the co-occurrence network can be explained through a synthesis model that is based on both preferential attachment and *triad* (i.e., fully-connected triplet) formation. This process of triad formation actually imposes a large number of triangles onto the generated network thereby creating many densely connected neighborhoods and increasing the clustering coefficient.

Apart from exploring various significant properties of the consonant inventories, we also employ our computational methodology to study the structure of the vowel inventories. Some of our observations are

(i) The topological properties of the occurrence and co-occurrence networks constructed from the vowel inventories are found to be largely similar to that of the consonant inventories. In particular, preferential attachment plays the key role in the emergence of their structure.

(ii) Community analysis of the co-occurrence network of vowels indicate that the small size vowel inventories tend to be organized based on the principle of maximal perceptual contrast – an observation that is in agreement with those reported by the earlier researchers [8, 15].

(iii) On the other hand, the larger vowel inventories reflect a considerable extent of feature economy – an observation that has been made by a school of linguists earlier [4, 7], but quantitatively substantiated here.

(iv) Co-occurrences based on implications (one vowel implying the presence of another) are prevalent across the vowel inventories and their presence is again a consequence of feature economy. This property has also been noted by linguists earlier; however, it has been quantitatively established here.

Finally, to summarize the contributions of the thesis in a single sentence, we have shown that the self-organization and the emergence of the structure of human speech sound inventories can be successfully studied within the framework of complex networks. Thus, we believe that in future, this computational framework can serve as an extremely powerful tool in modeling the structure and dynamics of several linguistic phenomena, which are as complex as the one presented here and for which no satisfactory explanation exists.

5 Organization of the Thesis

The thesis is organized into seven chapters.

Chapter 2 presents the history of the problem of sound inventories. It describes the human articulatory apparatus and the representation scheme for a phonological system. This is followed by a concise review of the different linguistic and computational studies pertaining to the organization of the sound inventories, which together form the basic motivation for this work.

Chapter 3 centers around the study of the occurrence network of consonants. It outlines the construction procedure for the network and describes the data source used for this construction. This is followed by an analysis of some of the interesting topological properties of the network. A synthesis model is then proposed and analytically solved to explain these properties. Finally, we employ this model to investigate the dynamics within and across five major language families of the world.

Chapter 4 investigates in detail the properties of the co-occurrence network of the consonants. It begins with a study of some of the important topological properties of this network. Suitable refinements of the synthesis model presented in Chapter 3 are also proposed in order to explain the emergence of these topological properties.

Chapter 5 presents the community structure analysis of the co-occurrence network of consonants. It discusses the algorithm for community detection and identifies the role of feature economy in the formation of the communities. Furthermore, it describes an information theoretic approach to quantify feature economy so as to determine the extent to which this factor governs the community formation in consonant inventories.

Chapter 6 presents a detailed study of the topological properties of the occurrence and co-occurrence network of vowels. Furthermore, it outlines the community analysis of the co-occurrence network and reports various interesting as well as important observations about the vowel inventories apart from validating the results that are already documented in the literature.

Chapter 7 concludes the thesis by summarizing the contributions and pointing to a few topics of future research that have been opened up from this work.

Bibliography

[1] L. A. Adamic and B. A. Huberman. Power-law distribution of the world wide web. *Science*, 287:2115, 2000.

[2] L. A. Adamic, R. M. Lukose, and B. A. Huberman. Local search in unstructured networks. In S. Bornholdt and H. G. Schuster, editors, *Handbook of Graphs and Networks*. Wiley-VCH, 2003.

[3] P. Arhem, H. A. Braun, M. T. Huber, and H. Liljenstrom. *Micro-Meso-Macro: Addressing Complex Systems Couplings*. World Scientific, 2004.

[4] P. Boersma. *Functional Phonology*. The Hague: Holland Academic Graphics, 1998.

[5] R. Ferrer-i-Cancho and R. V. Solé. The small world of human language. *Proceedings of The Royal Society of London. Series B, Biological Sciences*, 268(1482):2261–2265, 2001.

[6] C. Christensen, A. Gupta, C. D. Maranas, and R. Albert. Inference and graph-theoretical analysis of Bacillus Subtilis gene regulatory networks. *Physica A*, 373:796–810, 2007.

[7] G. N. Clements. The role of features in speech sound inventories. In E. Raimy and C. Cairns, editors, *Contemporary Views on Architecture and Representations in Phonological Theory*. Cambridge, MA: MIT Press, 2008.

[8] B. de Boer. *Self-Organisation in Vowel Systems*. PhD Thesis, AI Lab, Vrije Universiteit Brussel, 1999.

[9] S. N. Dorogovtsev and J. F. F. Mendes. Language as an evolving word web. *Proceedings of the Royal Society of London B: Biological Sciences*, 268(1485):2603–2606, 2001.

[10] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. *Computer Communications Review*, 29:251–262, 1999.

[11] R. Guimerà, S. Mossa, A. Turtschi, and L. A. N. Amaral. The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles. *PNAS*, 102:7794–7799, 2005.

[12] V. Kapatsinski. Sound similarity relations in the mental lexicon: Modeling the lexicon as a complex network. *Speech Research Lab Progress Report, Indiana University*, 2006.

[13] J. Ke, M. Ogura, and W. S.-Y. Wang. Optimization models of sound systems using genetic algorithms. *Computational Linguistics*, 29(1):1–18, 2003.

[14] R. Köhler. Synergetic linguistics. In R. Köhler, G. Altmann, and R. G. Piotrovskií, editors, *Quantitative Linguistics. An International Handbook*, pages 760–775. Walter de Gruyter, Berlin, New York, 2005.

[15] J. Liljencrants and B. Lindblom. Numerical simulation of vowel quality systems: the role of perceptual contrast. *Language*, 48:839–862, 1972.

[16] B. Lindblom and I. Maddieson. Phonetic universals in consonant systems. In M. Hyman and C. N. Li, editors, *Language, Speech, and Mind*, pages 62–78, 1988.

[17] B. Mitra, N. Ganguly, S. Ghose, and F. Peruani. Generalized theory for node disruption in finite-size complex networks. *Phys. Rev. E*, 78:026115, 2008.

[18] A. E. Motter, N. Gulbahce, E. Almaas, and A.-L. Barabási. Predicting synthetic rescues in metabolic networks. *Mol. Sys. Biol.*, 4(168):1–10, 2008.

[19] M. E. J. Newman. Coauthorship networks and patterns of scientific collaboration. *PNAS*, 101:5200–5205, 2004.

[20] P.-Y. Oudeyer. *Self-organization in the Evolution of Speech*. Oxford University Press, Oxford, 2006.

[21] J. J. Ramasco, S. N. Dorogovtsev, and R. Pastor-Satorras. Self-organization of collaboration networks. *Phys. Rev. E*, 70:036106, 2004.

[22] P. Sen, S. Dasgupta, A. Chatterjee, P. A. Sreeram, G. Mukherjee, and S. S. Manna. Small-world properties of the Indian railway network. *Phys. Rev. E*, 67:036106, 2003.

[23] M. Sigman and G. A. Cecchi. Global organization of the wordnet lexicon. *PNAS*, 99:1742–1747, 2002.

[24] M. M. Soares, G. Corso, and L. S. Lucena. The network of syllables in Portuguese. *Physica A*, 355(2-4):678–684, 2005.

[25] L. Steels. Language as a complex adaptive system. In *Proceedings of PPSN VI*, pages 17–26, 2000.

[26] N. Trubetzkoy. *Principles of Phonology*. University of California Press, Berkeley, 1969.

[27] W. S.-Y. Wang. The basis of speech. In C. E. Reed, editor, *The Learning of Language*, 1971.

[28] R. J. Williams and N. D. Martinez. Simple rules yield complex food webs. *Nature*, 404:180–183, 2000.

[29] S. Wuchty, A. Barabási, and M. T. Ferdig. Stable evolutionary signal in a Yeast protein interaction network. *BMC Evol. Biol.*, 60(8):1–10, 2006.

Publications from the Thesis

Publications from the work presented in the thesis: The publications are listed in chronological order.

[1] M. Choudhury, A. Mukherjee, A. Basu, and N. Ganguly. Analysis and synthesis of the distribution of consonants over languages: A complex network approach. In *Proceedings of COLING–ACL*, 128–135, 2006.

[2] F. Peruani, M. Choudhury, A. Mukherjee, and N. Ganguly. Emergence of a non-scaling degree distribution in bipartite networks: A numerical and analytical study. *Euro. Phys. Lett.*, 79(2):28001, 2007.

[3] A. Mukherjee, M. Choudhury, A. Basu, and N. Ganguly. Modeling the co-occurrence principles of the consonant inventories: A complex network approach. *Int. Jour. of Mod. Phy. C*, 18(2):281–295, 2007.

[4] A. Mukherjee, M. Choudhury, A. Basu, and N. Ganguly. Redundancy ratio: An invariant property of the consonant inventories of the world's languages. In *Proceedings of ACL*, 104–111, 2007.

[5] A. Mukherjee, M. Choudhury, A. Basu, and N. Ganguly. Emergence of community structures in vowel inventories: An analysis based on complex networks. In *Proceedings of ACL SIGMORPHON9*, 101–108, 2007.

[6] A. Mukherjee, M. Choudhury, S. Roy Chowdhury, A. Basu, and N. Ganguly. Rediscovering the co-occurrence principles of the vowel inventories: A complex network approach. *Advances in Complex Systems*, 11(3):371–392, 2008.

[7] A. Mukherjee, M. Choudhury, A. Basu, and N. Ganguly. Modeling the structure and dynamics of the consonant inventories: A complex network approach , In *Proceedings of Coling*, 601–608, 2008.

[8] M. Choudhury, A. Mukherjee, A. Garg, V. Jalan, A. Basu, and N. Ganguly. Language diversity across the consonant inventories: A study in the framework of complex networks. In *Proceedings of EACL workshop on Cognitive Aspects of Computational Language Acquisition*, 51–58, 2009.

[9] A. Mukherjee, M. Choudhury, A. Basu, and N. Ganguly. Self-organization of sound inventories: Analysis and synthesis of the occurrence and co-occurrence network of consonants. *Journal of Quantitative Linguistics*, 16(2):157–184, 2009.

[10] A. Mukherjee, M. Choudhury, and N. Ganguly. Analyzing the degree distribution of the one-mode projection of alphabetic bipartite networks ($\alpha - BiN$s). *preprint: arXiv.org:0902.0702*.